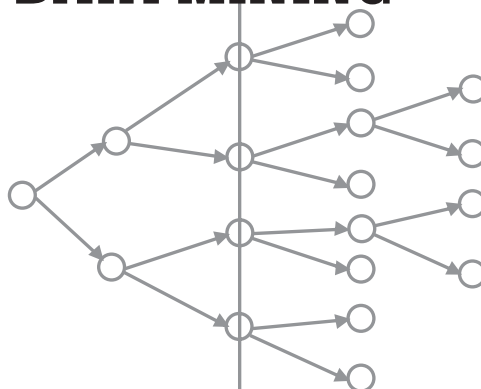


# Final Program and Abstracts

## 2014 SIAM International Conference on **DATA MINING**

April 24-26, 2014



**Sheraton Society Hill Hotel  
Philadelphia, Pennsylvania, USA**

*Sponsored by the SIAM Activity Group on  
Data Mining and Analytics (SIAG/DMA)*

The purpose of the SIAM Activity Group on Data Mining and Analytics (SIAG/DMA) is to advance the mathematics of data mining, to highlight the importance and benefits of the application of data mining, and to identify and explore the connections between data mining and other applied sciences. The activity group organizes the yearly SIAM International Conference on Data Mining (SDM), organizes minisymposia at the SIAM Annual Meeting, and maintains a membership directory and electronic mailing list.

*This conference is held in cooperation with the  
American Statistical Association.*



Society for Industrial and Applied Mathematics  
3600 Market Street, 6th Floor  
Philadelphia, PA 19104-2688 USA

Telephone: +1-215-382-9800 Fax: +1-215-386-7999

Conference E-mail: [meetings@siam.org](mailto:meetings@siam.org)

Conference Web: [www.siam.org/meetings/](http://www.siam.org/meetings/)

Membership and Customer Service:

(800) 447-7426 (US & Canada) or +1-215-382-9800 (worldwide)

**[www.siam.org/meetings/sdm14](http://www.siam.org/meetings/sdm14)**

## Table of Contents

Program-at-a-Glance.....Fold out section	
General Information.....	2
Get-togethers.....	7
Invited Plenary Presentations .....	9
Tutorials .....	11
Workshops and Minisymposium .....	12
Program Schedule .....	13
Poster Session .....	18
Abstracts .....	27
Organizer & Speaker Index .....	59
Conference Budget .....	Inside Back cover
Hotel Meeting Room Map .....	Back cover

## Organizing Committee

### Steering Committee Chairs

**Chandrika Kamath**  
Lawrence Livermore National Laboratory, USA

**Srinivasan Parthasarathy**  
Ohio State University, USA

### Conference Chairs

**Zoran Obradovic**  
Temple University, USA

**Mohammed Zaki**  
Qatar Computing Research Institute (QCRI),  
Qatar and Rensselaer Polytechnic Institute  
(RPI), USA

### Program Chairs

**Pang Ning Tan**  
Michigan State University, USA

**Arindam Banerjee**  
University of Minnesota, USA

### Workshop Chairs

**Tina Eliassi-Rad**  
Rutgers University, USA

**Amol Ghoting**  
IBM T.J. Watson Research Center, USA

### Tutorial Chair

**Suresh Venkatasubramanian**  
University of Utah, USA

### Publicity Chairs

**Xiaoli Fern**  
Oregon State University, USA

**Jing Gao**  
State University of New York, Buffalo, USA

### Sponsorship Chairs

**Hui Xiong**  
Rutgers University, USA

**Matthew Otey**  
Google, USA

### Local Chair

**Slobodan Vucetic**  
Temple University, USA

### Senior Program Committee

**Naoki Abe**  
IBM T.J. Watson Research Center, USA

**Roberto Bayardo**  
Google, USA

**Longbing Cao**  
University of Technology Sydney, Australia

**Varun Chandola**  
Oak Ridge National Laboratory, USA

**Nitesh Chawla**  
University of Notre Dame, USA

**Sanjay Chawla**  
University of Sydney, Australia

**Ian Davidson**  
University of California, Davis, USA

**Carlotta Domeniconi**  
George Mason University, USA

**Christos Faloutsos**  
Carnegie Mellon University, USA

**Wei Fan**  
Huawei, China

**Jiawei Han**  
University of Illinois at Urbana-Champaign,  
USA

### Rong Jin

Michigan State University, USA

### George Karypis

University of Minnesota, USA

### Emonn Keogh

University of California, Riverside, USA

### Tamara Kolda

Sandia National Laboratories, USA

### Balaji Krishnapuram

Siemens, USA

### Ee-Peng Lim

Singapore Management University,  
Singapore

### Huan Liu

Arizona State University, USA

### Sofus Macskassy

Facebook, USA

### Claire Monteleoni

George Washington University, USA

### Jennifer Neville

Purdue University, USA

### Nikunj Oza

NASA Ames Research Center, USA

### Jian Pei

Simon Fraser University, Canada

### Jimeng Sun

IBM T.J. Watson Research Center, USA

### Jianyong Wang

Tsinghua University, China

### Takashi Washio

Osaka University, Japan

### Jeffrey Yu

Chinese University of Hong Kong, Hong  
Kong

### Philip Yu

University of Illinois at Chicago, USA

### Xiao Jin Zhu

University of Wisconsin-Madison, USA

### Program Committee

#### Evrin Acar

University of Copenhagen, Denmark

#### Nitin Agarwal

University of Arkansas at Little Rock, USA

#### Leman Akoglu

SUNY Stony Brook, USA

#### Anthony Bagnall

University of East Anglia, United Kingdom

**James Bailey**

University of Melbourne, Australia

**Justin Basilico**

Netflix, USA

**Sugato Basu**

Google, USA

**Gustavo E. A. P. A. Batista**

University of Sao Paulo, Brazil

**Michele Benzi**

Emory University, USA

**Jonathan Berry**

Sandia National Labs, USA

**Kanishka Bhaduri**

Netflix, USA

**Sam Blasiak**

George Mason University, USA

**Daniel Boley**

University of Minnesota, USA

**Shyam Boriah**

University of Minnesota, USA

**Surafna Budalakoti**

University of Texas Austin, USA

**Polo Chau**

Georgia Tech, USA

**Ming-Syan Chen**

Academia Sinica, Taiwan

**Phoebe Chen**

La Trobe University, Australia

**Yixin Chen**

Washington University in St Louis, USA

**Hong Cheng**

Chinese University of Hong Kong, Hong Kong

**William K. Cheung**

Hong Kong Baptist University, Hong Kong

**Eric C. Chi**

Rice University, USA

**Gao Cong**

Nanyang Technological University, Singapore

**John Conroy**

IDA Center for Computing Sciences, USA

**Diane Cook**

Washington State University, USA

**Alfredo Cuzzocrea**

Italian National Research Council, Italy

**Kamalika Das**

NASA/University Affiliated Research Center, USA

**Santanu Das**

NASA/University Affiliated Research Center, USA

**Christian Desrosiers**

Ecole de Technologie Supérieure, Canada

**Wei Ding**

University of Massachusetts Boston, USA

**Murat Dundar**

Indiana University-Purdue University  
Indianapolis, USA

**Danny Dunlavy**

Sandia National Laboratories in Albuquerque, USA

**Haimonti Dutta**

Columbia University, USA

**Saso Dzeroski**

Jozef Stefan Institute, Slovenia

**Ya J. Fan**

Lawrence Livermore National Laboratory, USA

**Brian Gallagher**

Lawrence Livermore National Labs, USA

**Yang Gao**

Nanjing University, China

**Eric Gaussier**

Université Joseph Fourier, France

**Yong Ge**

University of North Carolina, Charlotte, USA

**Rayid Ghani**

University of Chicago, USA

**Shalini Ghosh**

SRI, USA

**Fosca Giannotti**

ISTI-CNR Pisa, Italy

**Aristides Gionis**

Aalto University, Finland

**Aris Gkoulalas-Divanis**

IBM Research, Ireland

**David Gleich**

Purdue University, USA

**Bart Goethals**

University of Antwerp, Belgium

**Francesco Gullo**

Yahoo! Labs, Spain

**Jingrui He**

Stevens Institute of Technology, USA

**Qi He**

LinkedIn, USA

**Steven Hoi**

Nanyang Technological University, Singapore

**Bao Qing Hu**

Wuhan University, China

**Jun (Luke) Huan**

University of Kansas, USA

**Joshua Huang**

Shenzhen Institute of Advanced Technology, China

**Seung-won Hwang**

Postech, Korea

**Akihiro Inokuchi**

Osaka University, Japan

**Nathalie Japkowicz**

University of Ottawa, Canada

**Madhav Jha**

Sandia National Laboratories, USA

**Shuiwang Ji**

Old Dominion University, USA

**Ruoming Jin**

Kent State University, USA

**Vladimir Jojic**

UNC Chapel Hill, USA

**Yoshinobu Kawahara**

Osaka University, Japan

**Jaya Kawale**

Adobe Research, USA

**Yiping Ke**

A\*Star, Singapore

**Philip Kegelmeyer**

Sandia National Laboratories, USA

**Kristian Kersting**

University of Bonn, Germany

**Irwin King**

The Chinese University of Hong Kong, Hong Kong

**James Kwok**

The Hong Kong University of Science and Technology, Hong Kong

**Hady Wirawan Lauw**

Singapore Management University, Singapore

**Jiuyong Li**

University of South Australia, Australia

**Lei Li**

UC Berkeley, USA

**Ming Li**

Nanjing University, China

**Tao Li**

Florida International University, USA

**Wujun Li**

Shanghai Jiao Tong University, China

**Zhenhui Li**

Pennsylvania State University, USA

**Jessica Lin**

George Mason University, USA

**Shou-de Lin**

National Taiwan University, Taiwan

**Jun Liu**

SAS, USA

**Yan Liu**

University of Southern California, USA

**David Lo**

Singapore Management University, Singapore

**Aurelie Lozano**

IBM Research, USA

**Chang-Tien Lu**

Virginia Tech, USA

**Ping Luo**

HP Labs, USA

**Nikos Mamoulis**

University of Hong Kong, Hong Kong

**Prakash MandayamComar**

Dow Chemical, USA

**Ole Mengshoel**

CMU Silicon Valley, USA

**Abdullah Mueen**

Microsoft, USA

**Emmanuel Mueller**Karlsruhe Institute of Technology (KIT),  
Germany**Vijay Narayanan**

Microsoft Research, USA

**Xia Ning**

NEC Labs America, USA

**Jia-Yu Pan**

Google, USA

**Spiros Papadimitriou**

Rutgers University, USA

**Ali Pinar**

Sandia National Labs, USA

**B. Aditya Prakash**

Virginia Tech, USA

**YanJun Qi**

NEC Labs America, USA

**Yuan Qi**

Purdue University, USA

**Buyue Qian**

IBM T. J. Watson, USA

**Chedy Raissi**

INRIA, France

**Thanawin (Art) Rakthanmanon**

Kasetsart University, Thailand

**Huzefa Rangwala**

George Mason University, USA

**Chandan Reddy**

Wayne State University, USA

**Marie-Christine Rousset**

University of Grenoble, France

**Ansaf Salleb-Aoussi**

Columbia University, USA

**Hanhuai Shan**

Microsoft, USA

**Shashi Shekhar**

University of Minnesota, USA

**Dou Shen**

Baidu, China

**Myra Spiliopoulou**

Magdeburg University, Germany

**Ashok Srivastava**

Verizon, USA

**Jaideep Srivastava**

University of Minnesota, USA

**Michael Steinbach**

University of Minnesota, USA

**Karsten Steinhäuser**

University of Minnesota, USA

**Fabio Stella**

University of Milano-Bicocca, Italy

**Masashi Sugiyama**

Tokyo Institute of Technology, Japan

**Yizhou Sun**

Northeastern University, USA

**Einoshin Suzuki**

Kyushu University, Japan

**Andrea Tagarelli**

University of Calabria, Italy

**Ichigaku Takigawa**

Hokkaido University, Japan

**Ah Hwee Tan**Nanyang Technological University,  
Singapore**Jie Tang**

Tsinghua University, China

**Lei Tang**

WalmartLabs, USA

**Dacheng Tao**

University of Technology Sydney, Australia

**Alexandre Termier**

Universite Joseph Fourier, France

**Evimaria Terzi**

Boston University, USA

**Hanghang Tong**

IBM Research, USA

**Shusaku Tsumoto**

Shimane University, Japan

**Giorgio Valentini**

University of Milan, Italy

**Hamed Valizadegan**

NASA Ames Research Center, USA

**Ranga Raju Vatsavai**

Oak Ridge National Laboratory, USA

**Michail Vlachos**IBM Zurich Research Laboratory,  
Switzerland**Ke Wang**

Simon Fraser University, Canada

**Lei Wang**

Australia Wollongong University, Australia

**Xin Wang**

University of Calgary, Canada

**Xiang Wang**

IBM T. J. Watson, USA

**Tim Weninger**

University of Notre Dame, USA

**Raymond Chi-Wing Wong**Hong Kong University of Science and  
Technology, Hong Kong**Weng-Keen Wong**

Oregon State University, USA

**Junjie Wu**

Beihang University, China

**Xintao Wu**

University of North Carolina at Charlotte, USA

**Jierui Xie**

Samsung R&amp;D Center, USA

**Jianliang Xu**

Hong Kong Baptist University, Hong Kong

**Guirong Xue**

Shanghai Jiao Tong University, China

**Rong Yan**

Facebook, USA

**De-Nian Yang**

Academia Sinica, Taiwan

**Mi-Yen Yeh**

Academia Sinica, Taiwan

**Guoxian Yu**

Southwest University, China

**Lei Yu**

Binghamton University, USA

**Shipeng Yu**

Siemens Medical Solutions, USA

**De-Chuan Zhan**

Nanjing University, China

**Chengqi Zhang**

The University of Technology Sydney,  
Australia

**Dell Zhang**

Birkbeck University of London, United  
Kingdom

**Min Zhang**

Tsinghua University, China

**Ping Zhang**

IBM TJ Watson, USA

**Ya (Anne) Zhang**

Shanghai Jiao Tong University, China

**Zhongfei (Mark) Zhang**

SUNY Binghamton, USA

**Ying Zhao**

Tsinghua University, China

**Zheng Alan Zhao**

SAS, USA

**Yu Zheng**

Microsoft Research Asia, China

**Bin Zhou**

University of Maryland Baltimore County,  
USA

**Jiayu Zhou**

Arizona State University, USA

**Feida Zhu**

Singapore Management University, Singapore

**Xingquan Zhu**

Florida Atlantic University, USA

**Child Care**

The Sheraton Society Hill Hotel highly recommends Your Other Hand. They can be reached at +1-215-790-0990.

**Hotel Address**

Sheraton Society Hill Hotel  
1 Dock Street  
Philadelphia, PA 19106  
Direct Telephone: +1-215-238-6000  
Toll Free (Central Reservations):  
+1-800-325-3535  
Hotel Fax: +1-215-238-6652  
[www.sheratonphiladelphiasocietyhill.com/](http://www.sheratonphiladelphiasocietyhill.com/)

**SIAM Registration Desk**

The SIAM registration desk is located in the Bromley/Claypoole Room. It is open during the following hours:

Wednesday, April 23  
5:00 PM - 7:00 PM

Thursday, April 24  
7:00 AM - 7:30 PM

Friday, April 25  
7:30 AM - 3:30 PM

Saturday, April 26  
7:30 AM - 4:00 PM

**Hotel Check-in  
and Check-out Times**

Check-in time is 3:00 PM and check-out time is 12:00 PM.

**Corporate Members  
and Affiliates**

SIAM corporate members provide their employees with knowledge about, access to, and contacts in the applied mathematics and computational sciences community through their membership benefits. Corporate membership is more than just a bundle of tangible products and services; it is an expression of support for SIAM and its programs. SIAM is pleased to acknowledge its corporate members and sponsors. In recognition of their support, non-member attendees who are employed by the following organizations are entitled to the SIAM member registration rate.

**Corporate Institutional  
Members**

The Aerospace Corporation  
Air Force Office of Scientific Research  
AT&T Laboratories - Research  
Bechtel Marine Propulsion Laboratory  
The Boeing Company  
CEA/DAM  
Department of National Defence (DND/CSEC)  
DSTO- Defence Science and Technology Organisation  
Hewlett-Packard  
IBM Corporation  
IDA Center for Communications Research, La Jolla  
IDA Center for Communications Research, Princeton  
Institute for Computational and Experimental Research in Mathematics (ICERM)  
Institute for Defense Analyses, Center for Computing Sciences  
Lawrence Berkeley National Laboratory  
Lockheed Martin  
Los Alamos National Laboratory  
Mathematical Sciences Research Institute  
Max-Planck-Institute for Dynamics of Complex Technical Systems  
Mentor Graphics  
National Institute of Standards and Technology (NIST)  
National Security Agency (DIRNSA)  
Oak Ridge National Laboratory, managed by UT-Battelle for the Department of Energy  
Sandia National Laboratories  
Schlumberger-Doll Research  
Tech X Corporation  
U.S. Army Corps of Engineers, Engineer Research and Development Center  
United States Department of Energy

*List current March 2014.*



## Funding Agencies

SIAM and the Conference Organizing Committee wish to extend their thanks and appreciation to the U.S. National Science Foundation for its support of this conference.



## Leading the applied mathematics community... Join SIAM and save!

SIAM members save up to \$130 on full registration for the 2014 SIAM International Conference on Data Mining! Join your peers in supporting the premier professional society for applied mathematicians and computational scientists. SIAM members receive subscriptions to *SIAM Review*, *SIAM News*, and *Unwrapped*, and enjoy substantial discounts on SIAM books, journal subscriptions, and conference registrations.

If you are not a SIAM member and paid the Non-Member rate to attend the conference, you can apply the difference between what you paid and what a member would have paid (\$130 for a Non-Member) towards a SIAM membership. Contact SIAM Customer Service for details or join at the conference registration desk.

If you are a SIAM member, it only costs \$10 to join the SIAM Activity Group on Data Mining and Analytics (SIAG/DMA). As a SIAG/DMA member, you are eligible for an additional \$10 discount on this conference, so if you paid the SIAM member rate to attend the conference, you might be eligible for a free SIAG/DMA membership. Check at the registration desk.

Free Student Memberships are available to students who attend an institution that is an Academic Member of SIAM, are members of Student Chapters of SIAM, or are nominated by a Regular Member of SIAM.

Join onsite at the registration desk, go to [www.siam.org/joinsiam](http://www.siam.org/joinsiam) to join online or download an application form, or contact SIAM Customer Service:

Telephone: +1-215-382-9800 (worldwide); or 800-447-7426 (U.S. and Canada only)

Fax: +1-215-386-7999

E-mail: [membership@siam.org](mailto:membership@siam.org)

Postal mail: Society for Industrial and Applied Mathematics, 3600 Market Street, 6th floor, Philadelphia, PA 19104-2688 USA

## Standard Audio/Visual Set-Up in Meeting Rooms

SIAM does not provide computers for any speaker. When giving an electronic presentation, speakers must provide their own computers. SIAM is not responsible for the safety and security of speakers' computers.

The Plenary Session Room will have two (2) screens, one (1) data projector and one (1) overhead projector. Cables or adaptors for Apple computers are not supplied, as they vary for each model. Please bring your own cable/adaptor if using an Apple computer.

All other concurrent/breakout rooms will have one (1) screen and one (1) data projector. Cables or adaptors for Apple computers are not supplied, as they vary for each model. Please bring your own cable/adaptor if using an Apple computer. Overhead projectors will be provided only if requested.

If you have questions regarding availability of equipment in the meeting room of your presentation, or to request an overhead projector for your session, please see a SIAM staff member at the registration desk.

## E-mail Access

The Sheraton Society Hill offers wireless Internet access in guest rooms and in the lobby area of the hotel. Complimentary wireless Internet access in the meeting space is also available.

SIAM will also provide a limited number of email stations for attendees during registration hours.

## Registration Fee Includes

- Admission to all technical sessions
- Admission to tutorial sessions
- Admission to a workshop
- Business Meeting (open to SIAG/DMA members)
- USB of conference proceedings, workshop and tutorial notes
- Coffee breaks daily
- Continental breakfast daily
- Room set-ups and audio/visual equipment
- Welcome Reception and Poster Session

## Job Postings

Please check with the SIAM registration desk regarding the availability of job postings or visit <http://jobs.siam.org>.

## Important Notice to Poster Presenters

The poster session is scheduled for Thursday, April 24 at 6:45 PM. Poster presenters are requested to set up their poster material on the provided 4' x 6' poster boards in the Hamilton Room after 10:00 AM on Thursday, April 24. All materials must be posted by 6:45 PM on Thursday, the official start time of the session. Posters will remain on display through 10:00 AM on Friday, April 25. Poster displays must be removed by this time. Posters remaining after this time will be discarded. SIAM is not responsible for discarded posters.

## SIAM Books and Journals

Display copies of books and complimentary copies of journals are available on site. SIAM books are available at a discounted price during the conference. If a SIAM books representative is not available, completed order forms and payment (credit cards are preferred) may be taken to the SIAM registration desk. The books table will close at 1:30 PM on Saturday, April 26.

## Tabletop Displays

SIAM  
Springer

## Conference Sponsors



## Name Badges

A space for emergency contact information is provided on the back of your name badge. Help us help you in the event of an emergency!

## Comments?

Comments about SIAM meetings are encouraged! Please send to:

Cynthia Phillips, SIAM Vice President for Programs (vpp@siam.org).

## Get-togethers

- Welcome Reception and Poster Session

Thursday, April 24

6:45 PM – 9:00 PM



- Business Meeting  
(open to SIAG/DMA members)

Friday, April 25

6:30 PM – 7:00 PM



*Complimentary beer and wine will be served.*

## Please Note

SIAM is not responsible for the safety and security of attendees' computers. Do not leave your laptop computers unattended. Please remember to turn off your cell phones, pagers, etc. during sessions.

## Recording of Presentations

Audio and video recording of presentations at SIAM meetings is prohibited without the written permission of the presenter and SIAM.

## Social Media

SIAM is promoting the use of social media, such as Facebook and Twitter, in order to enhance scientific discussion at its meetings and enable attendees to connect with each other prior to, during and after conferences. If you are tweeting about a conference, please use the designated hashtag to enable other attendees to keep up with the Twitter conversation and to allow better archiving of our conference discussions. The hashtag for this meeting is #SIAMSDM14.

# *Exciting Overseas Opportunities in Vienna, Austria*

## *with the International Atomic Energy Agency*

Be part of the  
**NEXT GENERATION SAFEGUARDS INITIATIVE**



### **Cost Free Experts**

The U.S. Support Program to IAEA Safeguards funds supplemental technical positions, called Cost Free Experts, to assist the IAEA with short-term, technical projects for which the IAEA lacks expertise or manpower. Available CFE positions are advertised at <http://www.bnl.gov/ISPO/CFE/cfe.asp>. Send applications or requests for information to [tcollins@bnl.gov](mailto:tcollins@bnl.gov).

### **Regular Staff**

The IAEA Department of Safeguards seeks outstanding, mid-career engineers and scientists for three to seven year appointments. Vacancies are posted at [www.iaea.org](http://www.iaea.org). Interested parties should apply directly to the IAEA, but can also contact the International Safeguards Project Office (ISPO) for information and assistance with application status.

### **Junior Professional Officers**

Safeguards positions are available for recent graduates in the area of software development, engineering, technical writing, information collection and analysis, information security and instrumentation support. Positions provide valuable professional experience and can lead to future opportunities in the national laboratories and U.S. government. Visit [www.bnl.gov/ispo/jpo/jpo.asp](http://www.bnl.gov/ispo/jpo/jpo.asp) for additional information and application materials.

ISPO Recruitment Program

Sponsored by the International Safeguards Project Office • Brookhaven National Laboratory



## Invited Plenary Speakers

*\*\* All Invited Plenary Presentations will take place in Ballroom B/C/D\*\**

**Thursday, April 24**

**8:15 AM - 9:30 AM**

**IP1** Tensor Factorizations for Multi-way Data: Past, Present, and Future

**Tamara G. Kolda**, *Sandia National Laboratories, USA*

**1:30 PM - 2:45 PM**

**IP2** The Battle for the Future of Data Mining

**Oren Etzioni**, *Allen Institute for Artificial Intelligence, USA*

**Friday, April 25**

**8:15 AM - 9:30 AM**

**IP3** Beyond Big Data: Combining Cognitive Computing and Analytics

**Brenda L. Dietrich**, *IBM Research, USA*

**1:30 PM - 2:45 PM**

**IP4** Why Most Published Studies Are Wrong But Can Be Fixed

**David Madigan**, *Columbia University, USA*

# SIAM Activity Group on Data Mining (SIAG/DMA)

[www.siam.org/activity/dma](http://www.siam.org/activity/dma)



## A GREAT WAY TO GET INVOLVED!

Collaborate and interact with mathematicians and applied scientists whose work involves data mining.

### ACTIVITIES INCLUDE:

- Special sessions at SIAM Annual Meetings
- Annual conference
- Website

### BENEFITS OF SIAG/DMA MEMBERSHIP:

- Listing in the SIAG's online membership directory
- Additional \$10 discount on registration at SIAM International Conference on Data Mining (excludes student)
- Electronic communications from your peers about recent developments in your specialty
- Eligibility for candidacy for SIAG/DMA office
- Participation in the selection of SIAG/DMA officers

### ELIGIBILITY:

- Be a current SIAM member.

### COST:

- \$10 per year
- Student members can join two activity groups for free!

### 2014-15 SIAG/DMA OFFICERS

- Chair: Zoran Obradovic, Temple University
- Vice Chair: Jeremy Kepner, Massachusetts Institute of Technology
- Program Director: Takashi Washio, Osaka University
- Secretary: Kirk Borne, George Mason University

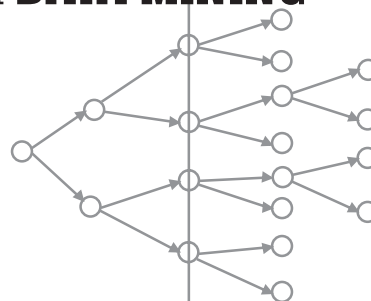
### TO JOIN:

SIAG/DMA: [my.siam.org/forms/join\\_siag.htm](http://my.siam.org/forms/join_siag.htm)

SIAM: [www.siam.org/joinsiam](http://www.siam.org/joinsiam)

## 2014 SIAM International Conference on DATA MINING

April 24-26, 2014



Sheraton Society Hill Hotel  
Philadelphia, Pennsylvania, USA

## Tutorials

**Thursday, April 24**

**10:00 AM - 12:05 PM**

**TS1:** Leveraging Social Media and Web of Data  
to Assist Crisis Response Coordination

*Ballroom E2*

**3:00 PM - 5:05 PM**

**TS2:** Safer Data Mining: A Tutorial on Algorithmic Techniques  
in Differential Privacy

*Ballroom E2*

---

**Friday, April 25**

**10:00 AM - 12:05 PM**

**TS3:** Data Mining in Drug Discovery and Development

*Ballroom E2*

**3:00 PM - 5:05 PM**

**TS4:** Node Similarity, Graph Similarity and Matching:  
Theory and Applications

*Ballroom E2*

---

**Saturday, April 26**

**8:30 AM - 10:35 AM**

**TS5:** Stochastic Optimization for Analyzing  
and Mining Big Data

*Cook*

## Workshops and Minisymposium

**Saturday, April 26**

### **Workshops**

**8:30 AM - 5:00 PM**

**Workshop 1:** Data Mining for Medicine and Healthcare

*Ballroom C*

**Workshop 2:** Exploratory Data Analysis

*Ballroom D*

**Workshop 3:** Heterogeneous Machine Learning

*Ballroom E*

**Workshop 4:** Mining Networks and Graphs: A BigData Analytics Challenge

*Ballroom B*

**1:30 PM - 5:00 PM**

**Workshop 5:** Optimization Methods for Anomaly Detection

*Cook*

For detailed schedules of each workshop, visit the conference workshop page:

<http://www.siam.org/meetings/sdm14/workshops.php>

### **Minisymposium**

**8:30 AM - 12:00 PM**

**MS1:** MultiClust: Multiple Clusterings, Multi-view Data,  
and Multi-source Knowledge-driven Clustering

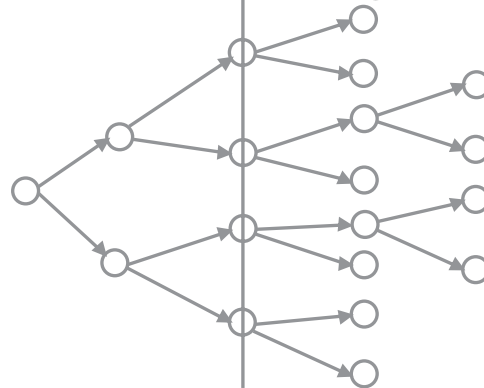
*Flower*



## SDM14 Program

### 2014 SIAM International Conference on **DATA MINING**

April 24-26, 2014



Sheraton Society Hill Hotel  
Philadelphia, Pennsylvania, USA

## Wednesday, April 23

### Registration

5:00 PM-7:00 PM

Room: Bromley/Claypoole

## Thursday, April 24

### Registration

7:00 AM-7:30 PM

Room: Bromley/Claypoole

### Continental Breakfast

7:30 AM

Room: Ballroom Foyer



### Announcements

8:00 AM-8:15 AM

Room: Ballroom B/C/D

Thursday, April 24

## IP1

### Tensor Factorizations for Multi-way Data: Past, Present, and Future

8:15 AM-9:30 AM

Room: Ballroom B/C/D

Chair: Zoran Obradovic, Temple University, USA

In the past decade, tensor factorizations have exploded onto the data mining and graph analysis scene. In this talk, I'll provide some historical context about the use of multi-way analysis, survey some recent advances in the field, and close with some challenging questions for moving forward. I'll focus mainly on factorizations that result in sums of rank-1 components, known variously as canonical decomposition, CANDECOMP, or PARAFAC. In any data or graph task, we are seeking to model real-world observations. Assumptions about the data and the model have a major impact on the effectiveness of the approach, especially in the case of sparse data. I'll talk the importance of understanding the underlying data distribution, incorporating constraints for the tensor models (symmetries, nonnegativity, sparsity, etc.), ways to validate algorithms for fitting these models, and challenges for the future.

Tamara G. Kolda

Sandia National Laboratories, USA

### Coffee Break

9:30 AM-10:00 AM

Room: Ballroom Foyer



Thursday, April 24

## CP1

### Classification 1

10:00 AM-12:05 PM

Room: Ballroom A1

Chair: Varun Chandola, State University of New York, Buffalo, USA

#### 10:00-10:20 Class Augmented Active Learning

Hong Cao, Chunyu Bao, and Xiaoli Li, Institute for Infocomm Research, Singapore; Yew-Kwong Wong, EADS Innovation Works South Asia

#### 10:25-10:45 Batch Mode Active Learning with Hierarchical-Structured Embedded Variance

Yu Cheng and ZhengZhang Chen, Northwestern University, USA; Hongliang Fei and Fei Wang, IBM T.J. Watson Research Center, USA; Alok Choudhary, Northwestern University, USA

#### 10:50-11:10 Learning from Multi-User Multi-Attribute Annotations

Ou Wu and Shuxiao Li, Chinese Academy of Sciences, China; Honghui Dong, Beijing Jiaotong University, China; Ying Chen and Weiming Hu, Chinese Academy of Sciences, China

#### 11:15-11:35 Selective Sampling on Probabilistic Data

Peng Peng and RaymondChi-Wing Wong, Hong Kong University of Science and Technology, Hong Kong

#### 11:40-12:00 Disambiguation-Free Partial Label Learning

Min-Ling Zhang, Southeast University, China

Thursday, April 24

## CP2

### Networks, Graphs 1

10:00 AM-12:05 PM

Room: Ballroom A2

Chair: Leman Akoglu, Stony Brook University, USA

#### 10:00-10:20 Dava: Distributing Vaccines over Networks under Prior Information

Yao Zhang and B. Aditya Prakash, Virginia Tech, USA

#### 10:25-10:45 Influence Maximization with Viral Product Design

Nicola Barbieri and Francesco Bonchi, Yahoo! Research Barcelona, Spain

#### 10:50-11:10 A Generative Model with Network Regularization for Semi-Supervised Collective Classification

Ruichao Shi, Harbin Institute of Technology, China; Qingyao Wu, Nanyang Technological University, Singapore; Yunming Ye, Harbin Institute of Technology, China; Shen-Shyang Ho, Nanyang Technological University, Singapore

#### 11:15-11:35 Local Learning for Mining Outlier Subgraphs from Network Datasets

Manish Gupta, Microsoft Research, India; Arun Mallya, Subhro Roy, Jason Cho, and Jiawei Han, University of Illinois at Urbana-Champaign, USA

#### 11:40-12:00 A Probabilistic Approach to Uncovering Attributed Graph Anomalies

Nan Li, University of California, Santa Barbara, USA; Huan Sun and Kyle Chipman, University of California, Santa Barbara, USA; Jemin George, U.S. Army Research Laboratory, USA; Xifeng Yan, University of California, Santa Barbara, USA

Thursday, April 24

## CP3

### Large Scale Matrix/Tensor

10:00 AM-12:05 PM

Room: Ballroom E1

Chair: Jieping Ye, Arizona State University, USA

#### 10:00-10:20 Vog: Summarizing and Understanding Large Graphs

Danai Koutra, Carnegie Mellon University, USA; U Kang, KAIST, Korea; Jilles Vreeken, Saarland University, Germany; Christos Faloutsos, Carnegie Mellon University, USA

#### 10:25-10:45 Context-Preserving Hashing for Fast Text Classification

Bin Li and Lianhua Chi, University of Technology, Australia; Xingquan Zhu, Florida Atlantic University, USA

#### 10:50-11:10 FlexiFaCT: Scalable Flexible Factorization of Coupled Tensors on Hadoop

Alex Beutel, Abhimanu Kumar, Evangelos Papalexakis, Partha Talukdar, Christos Faloutsos, and Eric Xing, Carnegie Mellon University, USA

#### 11:15-11:35 Turbo-Smt: Accelerating Coupled Sparse Matrix-Tensor Factorizations by 200x

Evangelos Papalexakis and Tom Mitchell, Carnegie Mellon University, USA; Nicholas Sidiropoulos, University of Minnesota, USA; Christos Faloutsos and Partha Talukdar, Carnegie Mellon University, USA; Brian M. Murphy, Queen's University, Belfast, United Kingdom

#### 11:40-12:00 Dusk: A Dual Structure-Preserving Kernel for Supervised Tensor Learning with Applications to Neuroimages

Lifang He, South China University of Technology, China; Xiangnan Kong and Philip Yu, University of Illinois, Chicago, USA; Ann Ragin, Northwestern University, USA; Zhifeng Hao, Guangdong University of Technology, China; Xiaowei Yang, South China University of Technology, China

Thursday, April 24

**TS1****Tutorial Session: Leveraging Social Media and Web of Data to Assist Crisis Response Coordination**

10:00 AM-12:05 PM

Room: Ballroom E2

Chair: Suresh Venkatasubramanian,  
University of Utah, USA

There is an ever increasing number of users in social media (1B+ Facebook users, 500M+ Twitter users) and ubiquitous mobile access (6B+ mobile phone subscribers) who share their observations and opinions. In addition, the Web of Data and existing knowledge bases keep on growing at a rapid pace. In this scenario, we have unprecedented opportunities to improve crisis response by extracting social signals, creating spatio-temporal mappings, performing analytics on social and Web of Data, and supporting a variety of applications. Such applications can help provide situational awareness during an emergency, improve preparedness, and assist during the rebuilding/recovery phase of a disaster. Data mining can provide valuable insights to support emergency responders and other stakeholders during crisis. However, there are a number of challenges and existing computing technology may not work in all cases. Therefore, our objective here is to present the characterization of such data mining tasks, and challenges that need further research attention.

**Carlos Castillo**

Qatar Computing Research Institute, Qatar

**Fernando Diaz**

Microsoft Research, USA

**Hemant Purohit**

Kno.e.sis, USA

**Lunch Break**

12:05 PM-1:30 PM

Attendees on their own

Thursday, April 24

**IP2****The Battle for the Future of Data Mining**

1:30 PM-2:45 PM

Room: Ballroom B/C/D

Chair: Pang-Ning Tan, Michigan State  
University, USA

Deep learning has catapulted to the front page of the New York Times, formed the core of the so-called “Google brain”, and achieved impressive results in vision, speech recognition, and elsewhere. Yet researchers have offered simple conundrums that deep learning doesn’t address. For example: The large ball crashed right through the table because it was made of styrofoam. What was made of styrofoam?

- the large ball
- the table

The answer is obviously “the table”, but if we change the word “Styrofoam” to “steel”, the answer is clearly “the large ball”. To automatically answer this type of question, our computers require an extensive body of common-sense knowledge. We believe that text mining can provide the requisite body of knowledge. My talk will describe work at the new Allen Institute for AI towards building the next-generation of text-mining systems.

**Oren Etzioni**Allen Institute for Artificial Intelligence,  
USA**Coffee Break**

2:45 PM-3:00 PM

Room: Ballroom Foyer



Thursday, April 24

**CP4****Feature Selection/extraction**

3:00 PM-5:05 PM

Room: Ballroom A1

Chair: Huan Liu, Arizona State University,  
USA
**3:00-3:20 How Can I Index My  
Thousands of Photos Effectively and  
Automatically? An Unsupervised  
Feature Selection Approach**

Juhua Hu and Jian Pei, Simon Fraser  
University, Canada; Jie Tang, Tsinghua  
University, P. R. China

**3:25-3:45 Unveiling Variables in  
Systems of Linear Equations**

Behzad Golshan and Evimaria Terzi, Boston  
University, USA

**3:50-4:10 Transductive Hsic Lasso**

Dan He, Irina Rish, and Laxmi Parida, IBM  
T.J. Watson Research Center, USA

**4:15-4:35 Robust Subspace Discovery  
Through Supervised Low-Rank  
Constraints**

Sheng Li and Yun Fu, Northeastern  
University, USA

**4:40-5:00 Adaptive Quantization  
for Hashing: An Information-Based  
Approach to Learning Binary Codes**

Caiming Xiong, Wei Chen, Gang Chen,  
David Johnson, and Jason Corso, State  
University of New York at Buffalo, USA



Thursday, April 24

## CP5

### Multi-task/Transfer Learning

3:00 PM-5:05 PM

Room: Ballroom A2

Chair: Jing Gao, State University of New York at Buffalo, USA

#### 3:00-3:20 Linking Heterogeneous Input Spaces with Pivots for Multi-Task Learning

Jingrui He, Stevens Institute of Technology, USA; Yan Liu, University of Southern California, USA; Qiang Yang, Hong Kong University of Science and Technology, Hong Kong

#### 3:25-3:45 Active Multitask Learning Using Both Latent and Supervised Shared Topics

Ayan Acharya, University of Texas at Austin, USA

#### 3:50-4:10 Multi-Task Feature Selection on Multiple Networks via Maximum Flows

Mahito Sugiyama, Chloé-Agathe Azencott, and Dominik Grimm, Max Plank Institute, Germany; Yoshinobu Kawahara, Osaka University, Japan; Karsten Borgwardt, Max Planck Institute for Intelligent Systems, Germany

#### 4:15-4:35 Mixed-Transfer: Transfer Learning over Mixed Graphs

Ben Tan, Hong Kong University of Science and Technology, Hong Kong; Erheng Zhong, Sun Yat-Sen University, China; Michael K. Ng, Hong Kong Baptist University, Hong Kong; Qiang Yang, Hong Kong University of Science and Technology, Hong Kong

#### 4:40-5:00 Multi-Graph Learning with Positive and Unlabeled Bags

Jia Wu, Zhibin Hong, and Shirui Pan, University of Technology, Australia; Xingquan Zhu, Florida Atlantic University, USA; Chengqi Zhang, University of Technology, Australia; Zhihua Cai, China University of Geoscience, P.R. China

Thursday, April 24

## CP6

### Applications

3:00 PM-5:05 PM

Room: Ballroom E1

Chair: Ramesh Natarajan, IBM T.J. Watson Research Center, USA

#### 3:00-3:20 Consumer Segmentation and Knowledge Extraction from Smart Meter and Survey Data

Tri Kurniawan Wijaya, EPFL, Switzerland; Tanuja Ganu and Dipanjan Chakraborty, IBM Research, India; Karl Aberer, EPFL, Switzerland; Deva Seetharam, IBM Research, India

#### 3:25-3:45 Keeping Up with Innovation: A Predictive Framework for Modeling Healthcare Data with Evolving Clinical Interventions

Sunil K. Gupta, Santu Rana, Dinh Phung, and Svetha Venkatesh, Deakin University, Australia

#### 3:50-4:10 Turning the Tide: Curbing Deceptive Yelp Behaviors

Bogdan Carbutar and Mahmudur Rahman, Florida International University, USA; Jaime Ballesteros, Nokia, USA; George Burri, Jive Software, USA; Duen Horng Chau, Georgia Institute of Technology, USA

#### 4:15-4:35 Predictive Learning in the Presence of Heterogeneity and Limited Training Data

Anuj Karpatne, University of Minnesota, USA; Ankush Khandelwal, University of Minnesota, Twin Cities, USA; Shyam Boriah and Vipin Kumar, University of Minnesota, USA

#### 4:40-5:00 Forecasting a Moving Target: Ensemble Models for Ili Case Count Predictions

Prithwish Chakraborty, Pejman Khadivi, Bryan Lewis, Aravindan Mahendiran, Jiangzhou Chen, Patrick Bulter, and Elaine O. Nsoesie, Virginia Tech, USA; Sumiko Mekaru and John Brownstein, Children's Hospital of Boston, USA; Madhav Marathe and Naren Ramakrishnan, Virginia Tech, USA

Thursday, April 24

## TS2

### Tutorial Session: Safer Data Mining: A Tutorial on Algorithmic Techniques in Differential Privacy

3:00 PM-5:05 PM

Room: Ballroom E2

Chair: Suresh Venkatasubramanian, University of Utah, USA

We present recent algorithmic developments in privacy-preserving data analysis and data mining. The goal of privacy-preserving data analysis is to reveal useful statistics about a population, while still preserving the privacy of the individuals. Sensitive data sets are increasingly common: examples include surveys, census data, social networks, medical records, financial statements, and streams of search queries. The tutorial focuses on Differential Privacy as the formal privacy guarantee. Differential privacy is a strong privacy guarantee that, intuitively speaking, hides the presence or absence of any individual in the data set. Differential Privacy has seen rapid developments in the last seven years drawing interest from several different communities. The research frontier remains active with many important open problems. This tutorial will present the basics of Differential Privacy and state of the art algorithms in three application areas: Supervised Machine Learning (with a focus on regularized regression), Unsupervised Machine Learning (with a focus on spectral methods), and Streaming Computation (with a focus on streaming data structures). A particular emphasis of this tutorial is on practical methods that will aid the practitioner in applying Differential Privacy in academic and industrial applications.

**Moritz Hardt**

IBM Research, USA

**Alexander Nikolov**

Rutgers University, USA

### Organizational Break

5:05 PM-5:15 PM

Thursday, April 24

## Poster Spotlights

5:15 PM-6:45 PM

Room: Ballroom B/C/D

### PP1

#### Welcome Reception and Poster Session

6:45 PM-9:00 PM



Room: Hamilton

#### Discovering Groups of Time Series with Similar Behavior in Multiple Small Intervals of Time

Gowtham Atluri, Michael Steinbach, Kelvin Lim, Angus MacDonald Iii, and Vipin Kumar, University of Minnesota, USA

#### Large Building Associations Between Markers of Environmental Stressors and Adverse Human Health Impacts Using Frequent Itemset Mining

Shannon M. Bell and Stephen Edwards, U.S. Environmental Protection Agency, USA

#### Characterising Seismic Data

Roel Bertens and Arno Siebes, Universiteit Utrecht, The Netherlands

#### Density-Based Clustering Validation

Davoud Moulavi and Pablo Jaskowiak, University of Alberta, Canada; Ricardo J. Campello, University of São Paulo at São Carlos, Brazil; Arthur Zimek, Ludwig-Maximilians-Universität München, Germany; Joerg Sander, University of Alberta, Canada

#### Online Discovery of Group Level Events in Time Series

XI Chen, University of Minnesota, USA; Abdullah Mueen, University of New Mexico, USA; Vijay Narayanan, Nikos Karampatziakis, and Gagan Bansal, Microsoft Corporation, USA; Vipin Kumar, University of Minnesota, USA

#### Frugal Traffic Monitoring with Autonomous Participatory Sensing

Vladimir Coric, Nemanja Djuric, and Slobodan Vucetic, Temple University, USA

#### Improving Credit Card Fraud Detection with Calibrated Probabilities

Alejandro Correa Bahnsen, Aleksandar Stokanovic, Djamila Aouada, and Bjorn Ottersten, Luxembourg University, Luxembourg

#### Iterative Framework Radiation Hybrid Mapping

Raed Seetan and Anne M. Denton, North Dakota State University, USA; Omar Al-Azzam, University of Minnesota, USA; Ajay Kumar, North Dakota State University, USA; Jazarai Sturdivant, Virginia State University, USA; Shahryar Kianian, University of Minnesota, USA

#### Mining Interpretable and Predictive Diagnosis Codes from Multi-Source Electronic Health Records

Sanjoy Dey, Gyorgy Simon, Bonnie Westra, Michael Steinbach, and Vipin Kumar, University of Minnesota, USA

#### Online Matrix Completion Through Nuclear Norm Regularisation

Charanpal Dhanjal, Télécom ParisTech, France; Romaric Gaudel, University of Lille, France; S Cléménçon, Télécom ParisTech, France

#### Classifying Imbalanced Data Streams Via Dynamic Feature Group Weighting with Importance Sampling

Ke Wu and Andrea Edwards, Xavier University, USA; Wei Fan, Huawei Noah's Ark Lab, Hong Kong; Jing Gao, State University of New York at Buffalo, USA; Kun Zhang, Xavier University, USA

#### Exploiting Monotonicity Constraints in Active Learning for Ordinal Classification

Ad Feelders and Pieter Soons, Universiteit Utrecht, The Netherlands

#### Classifier-Adjusted Density Estimation for Anomaly Detection and One-Class Classification

Lisa Friedland, Amanda Gentzel, and David Jensen, University of Massachusetts, Amherst, USA

#### Online Anomaly Detection by Improved Grammar Compression of Log Sequences

Yun Gao, Wei Zhou, Zhang Zhang, Jizhong Han, and Dan Meng, Chinese Academy of Sciences, China; Zhiyong Xu, Suffolk University Boston, USA

#### Submodularity in Team Formation Problem

Dinesh Garg, IBM India Research Lab, India; Avradeep Bhowmik, University of Texas, Austin, USA; Vivek Borkar, IIT Bombay, India; Madhavan Pallan, IBM India Research Lab, India

#### Quantifying Trust Dynamics in Signed Graphs, the S-Cores Approach

Christos Giatsidis and Michalis Vazirgiannis, Ecole Polytechnique, France; Dimitrios Thilikos, CNRS, France; Silviu Maniu, University of Hong Kong, China; Bogdan Cautis, Université Paris-Sud, France

#### On Finding the Point Where There Is No Return: Turning Point Mining on Game Data

Wei Gong, Ee-Peng Lim, Feida Zhu, Palakorn Achananuparp, and David Lo, Singapore Management University, Singapore

#### Memory-Efficient Query-Driven Community Detection with Application to Complex Disease Associations

Steve Harenberg, Ramona Seay, Stephen Ranshous, Kanchana Padmanabhan, Jitendra Harlalka, Eric Schendel, Michael O'Brien, and Rada Chirkova, North Carolina State University, USA; William Hendrix and Alok Choudhary, Northwestern University, USA; Vipin Kumar, University of Minnesota, USA; Murali Doraiswamy, Duke University, USA; Nagiza Samatova, North Carolina State University and Oak Ridge National Laboratory, USA

#### Unsupervised Feature Learning by Deep Sparse Coding

Yunlong He, Georgia Institute of Technology, USA; Koray Kavukcuoglu, DeepMind Technologies, United Kingdom; Yun Wang, Princeton University, USA; Arthur Szlam, City College of New York, USA; Yanjun Qi, University of Virginia, USA

#### Less Is More: Similarity of Time Series under Linear Transformations

Frank Höppner, Ostfalia University of Applied Sciences, Germany

#### Recovering Missing Labels of Crowdsourcing Workers

Qingyang Hu, Zhejiang University, China; Kevin Chiew, Provident Technology Pte. Ltd., Singapore; Hao Huang, National University of Singapore, Singapore; Qinming He, Zhejiang University, China

#### Detecting Influence Relationships from Graphs

Chuan Hu, Huiping Cao, and Chaomin Ke, New Mexico State University, USA

#### Meta: Multi-Resolution Framework for Event Summarization

Yexi Jiang, Florida International University, USA; Chang-Shing Perng, IBM Research, China; Tao Li, Florida International University, USA

continued on next page

### Beating Human Analysts in Nowcasting Corporate Earnings by Using Publicly Available Stock Price and Correlation Features

Michael Kamp, Mario Boley, and Thomas Gärtner, Fraunhofer Institute for Autonomous Intelligent Systems, Germany

### Adversarial Learning with Bayesian Hierarchical Mixtures of Experts

Yan Zhou and Murat Kantarcioglu, University of Texas at Dallas, USA

### Large-Scale Multi-Label Learning with Incomplete Label Assignments

Xiangnan Kong, University of Illinois, Chicago, USA; Zhaoming Wu, Tsinghua University, P. R. China; Li-Jia Li, Yahoo! Research, USA; Ruofei Zhang, Microsoft Corporation, USA; Philip Yu, University of Illinois, Chicago, USA; Hang Wu, Tsinghua University, P. R. China; Wei Fan, IBM T.J. Watson Research Center, USA

### Directed Interpretable Discovery in Tensors with Sparse Projection

Chia-Tung Kuo and Ian Davidson, University of California, Davis, USA

### Large-Scale Kernel Ranksvm

Tzu-Ming Kuo, Ching-Pei Lee, and Chih-Jen Lin, National Taiwan University, Taiwan

### Efficient Matching of Substrings in Uncertain Sequences

Yuxuan Li, James Bailey, and Lars Kulik, University of Melbourne, Australia; Jian Pei, Simon Fraser University, Canada

### Result Integrity Verification of Outsourced Bayesian Network Structure Learning

Wendy Hui Wang and Ruilin Liu, Stevens Institute of Technology, USA; Changhe Yuan, Queens College, CUNY, USA

### The Benefits of Personalized Smartphone-Based Activity Recognition Models

Jeff Lockhart, Fordham University, USA

### A New Framework for Traffic Anomaly Detection

Jinsong Lan, Chinese Academy of Sciences, China; Cheng Long and Raymond C.-W. Wong, Hong Kong University of Science and Technology, Hong Kong; Youyang Chen, Beijing Institute of Technology, China; Yanjie Fu, Rutgers University, USA; Danhuai Guo and Shuguang Liu, Chinese Academy of Sciences, China; Yong Ge, University of North Carolina, Charlotte, USA; Yuanchun Zhou and Jianhui Li, Chinese Academy of Sciences, China

### A Graphical Model Approach to Atlas-Free Mining of Mri Images

Chris Magnano and Ameet Soni, Swarthmore College, USA; Sriram Natarajan, Indiana University, USA; Gautam Kunapuli, UtopiaCompression Corporation, USA

### ROCsearch --- An ROC-guided Search Strategy for Subgroup Discovery

Marvin Meeng, Leiden University, Netherlands; Wouter Duivesteijn, Technische Universität Dortmund, Germany; Arno Knobbe, Leiden University, Netherlands

### Community Discovery in Social Networks Via Heterogeneous Link Association and Fusion

Lei Meng and Ah Hwee Tan, Nanyang Technological University, Singapore

### Discriminative Density-Ratio Estimation

Yun-Qian Miao, Ahmed Farahat, and Mohamed Kamel, University of Waterloo, Canada

### Accelerating Graph Adjacency Matrix Multiplications with Adjacency Forest

Masaaki Nishino, NTT Communication Science Laboratories, Japan; Norihito Yasuda, Japan Science and Technology Agency, Japan; Shin-Ichi Minato, Hokkaido University, Japan; Masaaki Nagata, NTT Communication Science Laboratories, Japan

### Graphical Models for Identifying Fraud and Waste in Healthcare Claims

Peder A. Olsen, Ramesh Natarajan, and Sholom Weiss, IBM T.J. Watson Research Center, USA

### An Optimization-Based Framework to Learn Conditional Random Fields for Multi-Label Classification

Mahdi Pakdaman Naeni, University of Pittsburgh, USA; Iyad Batal, General Electric Global Research Center, USA; Zitao Liu, CharmGil Hong, and Milos Hauskrecht, University of Pittsburgh, USA

### Extracting Researcher Metadata with Labeled Features

Sujatha Das Gollapalli, Pennsylvania State University, USA; Yanjun Qi, University of Virginia, USA; Prasenjit Mitra and Lee Giles, Pennsylvania State University, USA

### Multi-Task Clustering Using Constrained Symmetric Non-Negative Matrix Factorization

Samir Al-Stouhi and Chandan Reddy, Wayne State University, USA

### A Weighted Adaptive Mean Shift Clustering Algorithm

Yazhou Ren and Carlotta Domeniconi, George Mason University, USA; Guoji Zhang, South China University of Technology, China; Guoxian Yu, Southwest University, China

### Generalized Outlier Detection with Flexible Kernel Density Estimates

Arthur Zimek, Erich Schubert, and Hans-Peter Kriegel, Ludwig-Maximilians-Universität München, Germany

### Membership Detection Using Cooperative Data Mining Algorithms

Lisa Singh, Calvin Newport, and Yiqing Ren, Georgetown University, USA

### A Guide to Selecting a Network Similarity Method

Sucheta Soundarajan and Tina Eliassi-Rad, Rutgers University, USA; Brian Gallagher, Lawrence Livermore National Laboratory, USA

### Discriminant Analysis for Unsupervised Feature Selection

Jiliang Tang, Xia Hu, Huiji Gao, and Huan Liu, Arizona State University, USA

### Factor Matrix Trace Norm Minimization for Low-Rank Tensor Completion

Yuanyuan Liu, Fanhua Shang, Hong Cheng, and James Cheng, Chinese University of Hong Kong, Hong Kong; Hanghang Tong, City College of CUNY, USA

### Auto-Play: A Data Mining Approach to Odi Cricket Simulation and Prediction

Vignesh Veppur Sankaranarayanan, Junaed Sattar, and Laks V.S. Lakshmanan, University of British Columbia, Canada

### Relational Regularization and Feature Ranking

Mathias Verbeke, K.U. Leuven, Belgium; Fabrizio Costa, Albert-Ludwigs University, Germany; Luc De Raedt, Katholieke Universiteit Leuven, Belgium

### Future Influence Ranking of Scientific Literature

Senzhang Wang and Sihong Xie, University of Illinois, Chicago, USA; Xiaoming Zhang and Zhoujun Li, BeiHang University, China; Philip S Yu, University of Illinois, Chicago, USA; Xinyu Shu, China Agricultural University, China

### Kaleido: Network Traffic Attribution Using Multifaceted Footprinting

Ting Wang, IBM Research, USA; Fei Wang, IBM T.J. Watson Research Center, USA; Reiner Sailer and Douglas Schales, IBM Research, USA

### Modeling Asymmetry and Tail Dependence among Multiple Variables by Using Partial Regular Vine

Wei Wei, Junfu Yin, Jinyan Li, and Longbing Cao, University of Technology, Australia



Thursday, April 24

## PP1

### Welcome Reception and Poster Session

6:45 PM-9:00 PM

continued

#### WCAMiner: A Novel Knowledge Discovery System for Mining Concept Associations Using Wikipedia

Peng Yan and Wei Jin, North Dakota State  
University, USA

#### Subgraph Search in Large Graphs with Result Diversification

Huiwen Yu and Dayu Yuan, Pennsylvania  
State University, USA

#### To Sample Or To Smash? Estimating Reachability in Large Time-Varying Graphs

Feng Yu, City University of New York, USA;  
Prithwish Basu, Raytheon Systems, USA;  
Amotz Bar-Noy and Dror Rawitz, City  
University of New York, USA

#### Finding Friends on a New Site Using Minimum Information

Reza Zafarani and Huan Liu, Arizona State  
University, USA

#### Unsupervised Selection of Robust Audio Feature Subsets

Maia Zaharieva, Technische Universitaet  
Wien, Austria; Gerhard Sageder, University  
of Vienna, Austria; Matthias Zeppelzauer,  
Vienna University of Technology, Austria

#### Towards Breaking the Curse of Dimensionality for High-Dimensional Privacy

Hessam Zakerzadeh, University of Calgary,  
Canada; Charu C. Aggarwal, IBM T.J.  
Watson Research Center, USA; Ken Barker,  
University of Calgary, Canada

#### Towards Accurate Histogram Publication under Differential Privacy

Xiaojuan Zhang, Renmin University of China,  
China; Rui Chen and Jianliang Xu, Hong  
Kong Baptist University, Hong Kong;  
Xiaofeng Meng, Renmin University of  
China, China; Yingtao Xie, China West  
Normal University, China

#### Addressing Human Subjectivity Via Transfer Learning: An Application to Predicting Disease Outcome in Multiple Sclerosis Patients

Yijun Zhao and Carla Brodley, Tufts  
University, USA; Tanuja Chitnis and Brian  
Healy, Harvard Medical School, USA

## Friday, April 25

### Continental Breakfast

7:30 AM

Room: Ballroom Foyer



### Registration

7:30 AM-3:30 PM

Room: Bromley/Claypoole

### Announcements

8:00 AM-8:15 AM

Room: Ballroom B/C/D

Friday, April 25

## IP3

### Beyond Big Data: Combining Cognitive Computing and Analytics

8:15 AM-9:30 AM

Room: Ballroom B/C/D

Chair: Mohammed Zaki, Rensselaer  
Polytechnic Institute, USA

In 2011 a computer named Watson demonstrated super human question answering ability by defeating two Jeopardy! grand champions. This marked the beginning of a new era of computing, the cognitive computing era. Watson was built upon a strong foundation in natural language understanding and machine learning, and excels at question answering. In parallel, significant advances have been made in data analytics, both in the scale of data that can be analyzed, and in the methods that can be applied to extract signal from data. We have an opportunity to combine the technologies to deal with large volumes of both structured and unstructured data, to answer questions using both written facts and signals extracted from data. This talk will explore examples and identify research opportunities.

Brenda L. Dietrich  
IBM Research, USA

### Coffee Break

9:30 AM-10:00 AM



Room: Ballroom Foyer



Friday, April 25

**CP7****Classification 2**

10:00 AM-12:05 PM

*Room: Ballroom A1**Chair: Jingrui He, Stevens Institute of Technology, USA***10:00-10:20 AUC Dominant Unsupervised Ensemble of Binary Classifiers***Priyanka Agrawal, Vikas C. Raykar, and Amrita Saha, IBM Research, India***10:25-10:45 Convex Optimization for Binary Classifier Aggregation in Multiclass Problems***Sunho Park and Tae Hyun Hwang, University of Texas Southwestern Medical Center at Dallas, USA; Seungjin Choi, Pohang University of Science and Technology, South Korea***10:50-11:10 A Deep Learning Approach to Link Prediction in Dynamic Networks***Xiaoyi Li, Nan Du, Hui Li, Kang Li, Jing Gao, and Aidong Zhang, State University of New York at Buffalo, USA***11:15-11:35 A Marginalized Denoising Method for Link Prediction in Relational Data***Zheng Chen and Weixiong Zhang, Washington University in St. Louis, USA***11:40-12:00 Learning on Probabilistic Labels***Peng Peng and Raymond ChiWing Wong, Hong Kong University of Science and Technology, Hong Kong; Philip S. Yu, University of Illinois, Chicago, USA*

Friday, April 25

**CP8****Networks, Graphs 2**

10:00 AM-12:05 PM

*Room: Ballroom A2**Chair: B. Aditya Prakash, Virginia Tech, USA***10:00-10:20 Adaptive User Distance Modeling in Social Media***Erheng Zhong, Sun Yat-Sen University, China; Wei Fan, Huawei Noah's Ark Lab, Hong Kong; Qiang Yang, Hong Kong University of Science and Technology, Hong Kong***10:25-10:45 Make It Or Break It: Manipulating Robustness in Large Networks***Hau Chan and Leman Akoglu, Stony Brook University, USA; Hanghang Tong, City College of CUNY, USA***10:50-11:10 It Takes Two to Tango: Exploring Social Tie Development with Both Online and Offline Interactions***Peifeng Yin, Pennsylvania State University, USA; Qi He, LinkedIn, USA; Xingjie Liu, Square Enix Co. Ltd., Japan; Wang-Chien Lee, Pennsylvania State University, USA***11:15-11:35 Laplacian Spectral Properties of Graphs from Random Local Samples***Zhengwei Wu and Victor Preciado, University of Pennsylvania, USA***11:40-12:00 Triangle Counting in Streamed Graphs Via Small Vertex Covers***David Garcia-Soriano, Yahoo! Research Barcelona, Spain*

Friday, April 25

**CP9****Probabilistic Topic Modeling**

10:00 AM-12:05 PM

*Room: Ballroom E1**Chair: Carlotta Domeniconi, George Mason University, USA***10:00-10:20 A Constrained Hidden Markov Model Approach for Non-Explicit Citation Context Extraction***Parikshit Sondhi and ChengXiang Zhai, University of Illinois at Urbana-Champaign, USA***10:25-10:45 Joint Author Sentiment Topic Model***Subhabrata Mukherjee, Max-Planck-Institut fuer Informatik, Germany; Gaurab Basu and Sachindra Joshi, IBM Research, India***10:50-11:10 A Dynamic Nonparametric Model for Characterizing the Topical Communities in Social Streams***Ziqi Liu and Qinghua Zheng, Xi'an Jiaotong University, P.R. China; Fei Wang, IBM T.J. Watson Research Center, USA; Zhenhua Tian and Bo Li, Xi'an Jiaotong University, P.R. China***11:15-11:35 Recurrent Chinese Restaurant Process with a Duration-Based Discount for Event Identification from Twitter***Qiming Diao and Jing Jiang, Singapore Management University, Singapore***11:40-12:00 Automatic Construction and Ranking of Topical Keyphrases on Collections of Short Documents***Marina Danilevsky, Chi Wang, Nihit Desai, and Xiang Ren, University of Illinois at Urbana-Champaign, USA; Jingyi Guo, University of Massachusetts, Amherst, USA; Jiawei Han, University of Illinois at Urbana-Champaign, USA*

Friday, April 25

**TS3****Tutorial Session: Data Mining in Drug Discovery and Development**

10:00 AM-12:05 PM

Room: Ballroom E2

Chair: Suresh Venkatasubramanian,  
University of Utah, USA

Increasingly, effective drug discovery and development involve the searching and data mining of large amounts of heterogeneous information from many sources covering the domains of chemistry, biology and pharmacology amongst others. In this tutorial, we provide a review of the publicly-available large-scale databases relevant to drug discovery, describe the kinds of data mining approaches that can be applied to them, and identify directions for future research. Many of those insights come from drug discovery community, which is highly related to data mining but focuses on bioinformatics and/or cheminformatics specifics. We survey various related articles from data mining venues as well as from bioinformatics/cheminformatics venues to share with the audience key problems and trends in drug discovery research, with different applications such as drug combinations, drug repositioning, personalized medicine, and payer evidence.

**Ping Zhang**

IBM T.J. Watson Research Center, USA

**Lun Yang**

GlaxoSmithKline, USA

**Lunch Break**

12:05 PM-1:30 PM

Attendees on their own

Friday, April 25

**IP4****Why Most Published Studies Are Wrong But Can Be Fixed**

1:30 PM-2:45 PM

Room: Ballroom B/C/D

Chair: Arindam Banerjee, University of  
Minnesota, USA

Observational healthcare data, such as administrative claims and electronic health records, play an increasingly prominent role in healthcare. Pharmacoepidemiologic studies in particular routinely estimate temporal associations between medical product exposure and subsequent health outcomes of interest and such studies influence prescribing patterns and healthcare policy more generally. Some authors have questioned the reliability and accuracy of such studies, but few previous efforts have attempted to measure their performance. The Observational Medical Outcomes Partnership (OMOP, <http://omop.org>) has conducted a series of experiments to empirically measure the performance of various observational study designs with regard to predictive accuracy for discriminating between true drug effects and negative controls. In this talk, I describe the past work of the Partnership, explore opportunities to expand the use of observational data to further our understanding of medical products, and highlight areas for future research and development.

David Madigan

Columbia University, USA

**Coffee Break**

2:45 PM-3:00 PM

Room: Ballroom Foyer



Friday, April 25

**CP10****Clustering and Density Estimation**

3:00 PM-5:05 PM

Room: Ballroom A1

Chair: Suresh Venkatasubramanian,  
University of Utah, USA**3:00-3:20 On Randomly Projected Hierarchical Clustering with Guarantees**Johannes Schneider, Zurich Insurances,  
Switzerland**3:25-3:45 Self-Taught Spectral Clustering Via Constraint Augmentation**

Xiang Wang, IBM Research, USA; Jun Wang, IBM T.J. Watson Research Center, USA; Buyue Qian, IBM Research, USA; Fei Wang, IBM T.J. Watson Research Center, USA; Ian Davidson, University of California, Davis, USA

**3:50-4:10 Embed and Conquer: Scalable Embeddings for Kernel K-Means on MapReduce**

Ahmed Elgohary, Ahmed Farahat, Mohamed Kamel, and Fakhri Karray, University of Waterloo, Canada

**4:15-4:35 A Constructive Setting for the Problem of Density Ratio Estimation**

Vladimir Vapnik, NEC Laboratories America, USA; Igor Braga, University of Sao Paulo, Brazil; Rauf Izmailov, Applied Communication Sciences, USA

**4:40-5:00 Density Estimation with Adaptive Sparse Grids for Large Data Sets**

Benjamin Peherstorfer, Massachusetts Institute of Technology, USA; Dirk Pflüger, Universität Stuttgart, Germany; Hans-Joachim Bungartz, Technische Universität München, Germany

Friday, April 25

## CP11

### Recommendation and Social Media Analysis

3:00 PM-5:05 PM

Room: Ballroom A2

Chair: Hui Xiong, Rutgers University, USA

#### 3:00-3:20 Latent Factor Transition for Dynamic Collaborative Filtering

Chenyi Zhang, Zhejiang University, China;  
Ke Wang and Hongkun Yu, Simon Fraser University, Canada; Jianling Sun, Zhejiang University, China; Ee-Peng Lim, Singapore Management University, Singapore

#### 3:25-3:45 Contextual Combinatorial Bandit and Its Application on Diversified Online Recommendation

Lijing Qin, Tsinghua University, P. R. China; Shouyuan Chen, Chinese University of Hong Kong, Hong Kong; Xiaoyan Zhu, Tsinghua University, P. R. China

#### 3:50-4:10 User Preference Learning with Multiple Information Fusion for Restaurant Recommendation

Hui Xiong, Yanjie Fu, and Bin Liu, Rutgers University, USA; Yong Ge, University of North Carolina, Charlotte, USA; Zijun Yao, Rutgers University, USA

#### 4:15-4:35 On Modeling Community Behaviors and Sentiments in Microblogging

Tuan-Anh Hoang, Singapore Management University, Singapore; William W. Cohen, Carnegie Mellon University, USA; Ee-Peng Lim, Singapore Management University, Singapore

#### 4:40-5:00 Selecting a Representative Set of Diverse Quality Reviews Automatically

Nana Xu, Beijing Institute of Technology, China; Hongyan Liu, Tsinghua University, P. R. China; Jiawei Chen and Jun He, Renmin University of China, China

Friday, April 25

## CP12

### Pattern Mining and Time Series

3:00 PM-5:05 PM

Room: Ballroom E1

Chair: Jilles Vreeken, Saarland University, Germany

#### 3:00-3:20 Finding the True Frequent Itemsets

Matteo Riondato, Brown University, USA; Fabio Vandin, University of Southern Denmark, Denmark

#### 3:25-3:45 A Statistical Learning Theory Framework for Supervised Pattern Discovery

Jonathan H. Huggins and Cynthia Rudin, Massachusetts Institute of Technology, USA

#### 3:50-4:10 When and Where: Predicting Human Movements Based on Social Spatial-Temporal Events

Ning Yang, Sichuan University, China; Xiangnan Kong, Fengjiao Wang, and Philip Yu, University of Illinois, Chicago, USA

#### 4:15-4:35 Ensembles of Elastic Distance Measures for Time Series Classification

Jason Lines and Anthony Bagnall, University of East Anglia, United Kingdom

#### 4:40-5:00 Discovery of Rare Sequential Topic Patterns in Document Stream

Zhongyi Hu, Hongan Wang, and Jiaqi Zhu, Chinese Academy of Sciences, China; Maozhen Li, Brunel University, United Kingdom; Ying Qiao and Changzhi Deng, Chinese Academy of Sciences, China

Friday, April 25

## TS4

### Tutorial Session: Node Similarity, Graph Similarity and Matching: Theory and Applications

3:00 PM-5:05 PM

Room: Ballroom E2

Chair: Suresh Venkatasubramanian, University of Utah, USA

Given a graph how can we find similarities among its nodes? Given a sequence of graphs, can we spot similarities or differences between them and detect temporal anomalies?

How can we match the nodes in different graphs in order to reveal similarities between them? In a microscopic level, the nodes can be similar in terms of their proximity or their structural features and roles. In a macroscopic level, two graphs are similar if the structures within them are similar, or equivalently their nodes are connected in similar ways. How can we find the similarity between two graphs when the node correspondence is known or unknown? When the correspondence is unknown, how can we efficiently match the two graphs, i.e., infer the node mapping? The objective of this tutorial is to provide a concise and intuitive overview of the most important concepts and tools, which can detect similarities between nodes of a static graph and different snapshots of dynamic graphs or distinct networks. We review the state of the art in three related fields: (a) node similarity and role discovery, (b) graph similarity, and (c) graph matching. The emphasis of this tutorial is to give the intuition behind these powerful mathematical concepts and tools, as well as case studies that illustrate their practical use in data mining.

#### Danai Koutra

Carnegie Mellon University, USA

#### Tina Eliassi-Rad

Rutgers University, USA

#### Christos Faloutsos

Carnegie Mellon University, USA

Friday, April 25

## Organizational Break

5:05 PM-5:15 PM

## NSF Panel

5:15 PM-6:30 PM

Room: Ballroom B/C/D

## SIAG/DMA Business Meeting

6:30 PM-7:00 PM

Room: Ballroom B/C/D



Complimentary beer and wine will be served.

## Doctoral Forum and Student Posters

7:00 PM-9:00 PM

Room: Hamilton



## Saturday, April 26

### Registration

7:30 AM-4:00 PM

Room: Bromley/Claypoole

### Continental Breakfast

8:00 AM

Room: Ballroom Foyer



Saturday, April 26

## MS1

### MultiClust: Multiple Clusterings, Multi-view Data, and Multi-source Knowledge-driven Clustering

8:30 AM-12:00 PM

Room: Flower

The aim of the MultiClust minisymposium, to be held in conjunction with the 2014 SIAM Data Mining Conference, is to establish a venue for the growing community interested in multiple clustering solutions and in the different research topics related to this field. The minisymposium will increase the visibility of the topic itself, but also bridge the closely related research areas such as ensemble clustering, co-clustering, clustering with constraints, and frequent pattern mining. In particular, we solicit discussions on approaches for solving emerging problems such as clustering ensembles, semi-supervised clustering, subspace/projected clustering, co-clustering, and multi-view clustering.

Organizer: Ira Assent

Aarhus University, Denmark

Organizer: Carlotta Domeniconi

George Mason University, USA

Organizer: Francesco Gullo

Yahoo! Research, USA

Organizer: Andrea Tagarelli

University of Calabria, Italy

Organizer: Arthur Zimek

Ludwig-Maximilians-Universität München, Germany

### Clustering with Linear Programming

Arindam Banerjee, University of Minnesota, USA

### Clustering Evaluation and Validation: Some Results, Challenges, and Research Questions

Ricardo J. Campello, University of São Paulo at São Carlos, Brazil

### Utilizing Multiple Clusterings: Beyond Consensus Clustering

Xiaoli Z. Fern, Oregon State University, USA

### Title Not Available

George Karypis, University of Minnesota and Army HPC Research Center, USA

Saturday, April 26

**Workshop 1 (full day):  
Data Mining for Medicine  
and Healthcare**

8:30 AM-5:00 PM

Room: Ballroom C

**Workshop 2 (full day):  
Exploratory Data Analysis**

8:30 AM-5:00 PM

Room: Ballroom D

**Workshop 3 (full day):  
Heterogeneous Machine  
Learning**

8:30 AM-5:00 PM

Room: Ballroom E

**Workshop 4 (full day):  
Mining Networks and  
Graphs: A BigData Analytics  
Challenge**

8:30 AM-5:00 PM

Room: Ballroom B

Saturday, April 26

**TS5**

**Tutorial Session: Stochastic  
Optimization for Analyzing  
and Mining Big Data**

8:30 AM-10:35 AM

Room: Cook

Chair: Suresh Venkatasubramanian,  
University of Utah, USA

In recent years of machine learning applications, the size of data has been observed with an unprecedented growth. To tackle millions of billions of data points, stochastic optimization has emerged as a dominant approach for big data analytics. In this talk, we shall present recent advances in stochastic optimization for solving fundamental problems arising in big data analytics. In particular, we will present efficient stochastic optimization algorithms for solving classification (e.g., support vector machines, logistic regression) and regression (e.g. ridge regression and lasso) problems. This talk consists of four parts. The first part gives an introduction to machine learning and stochastic optimization that motivates stochastic optimization for big data analytics. The second part presents several start-of-the-art stochastic optimization algorithms for solving big data classification and regression problems. The third part presents general strategies of stochastic optimization, including stochastic gradient descent for a variety of objective functions, accelerated stochastic gradient descent for composite optimization, variance reduced stochastic optimization algorithms, parallel and distributed optimization algorithms. In the fourth part, we discuss some implementation issues and introduce a practical library of distributed stochastic optimization for solving big data classification and regression problems.

**Tianbao Yang**

NEC Laboratories America, USA

**Rong Jin**

Michigan State University, USA

**Shenghuo Zhu**

NEC Laboratories America, USA

**Coffee Break**

10:00 AM-10:30 AM

Room: Ballroom Foyer



Saturday, April 26

**Lunch Break**

12:00 PM-1:30 PM

Attendees on their own

**Workshop 5 (half day):  
Optimization Methods for  
Anomaly Detection**

1:30 PM-5:00 PM

Room: Cook

**Coffee Break**

3:00 PM-3:30 PM

Room: Ballroom Foyer



For detailed schedules of each workshop, visit the conference workshop page:

<http://www.siam.org/meetings/sdm14/workshops.php>

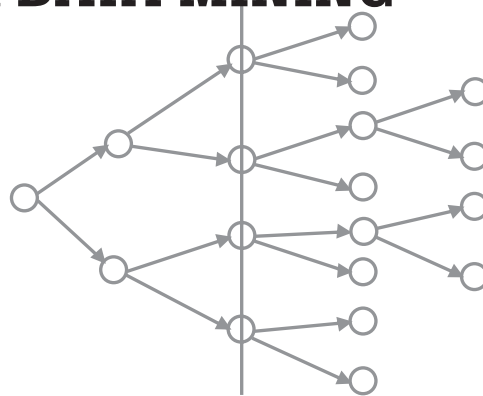


## Notes

## SDM14 Abstracts

2014 SIAM  
International Conference  
on **DATA MINING**

April 24-26, 2014



**Sheraton Society Hill Hotel**  
**Philadelphia, Pennsylvania, USA**

Abstracts are printed as submitted by the authors.

## IP1

**Tensor Factorizations for Multi-way Data: Past, Present, and Future**

In the past decade, tensor factorizations have exploded onto the data mining and graph analysis scene. In this talk, I'll provide some historical context about the use of multi-way analysis, survey some recent advances in the field, and close with some challenging questions for moving forward. I'll focus mainly on factorizations that result in sums of rank-1 components, known variously as canonical decomposition, CANDECOMP, or PARAFAC. In any data or graph task, we are seeking to model real-world observations. Assumptions about the data and the model have a major impact on the effectiveness of the approach, especially in the case of sparse data. I'll talk the importance of understanding the underlying data distribution, incorporating constraints for the tensor models (symmetries, non-negativity, sparsity, etc.), ways to validate algorithms for fitting these models, and challenges for the future.

Tamara G. Kolda

Sandia National Laboratories  
tgkolda@sandia.gov

## IP2

**The Battle for the Future of Data Mining**

Deep learning has catapulted to the front page of the New York Times, formed the core of the so-called Google brain, and achieved impressive results in vision, speech recognition, and elsewhere. Yet researchers have offered simple conundrums that deep learning doesn't address. For example: The large ball crashed right through the table because it was made of styrofoam. What was made of styrofoam? - the large ball - the table The answer is obviously the table, but if we change the word Styrofoam to steel, the answer is clearly the large ball. To automatically answer this type of question, our computers require an extensive body of common-sense knowledge. We believe that text mining can provide the requisite body of knowledge. My talk will describe work at the new Allen Institute for AI towards building the next-generation of text-mining systems.

Oren Etzioni

Allen Institute for Artificial Intelligence, USA  
OrenE@vulcan.com

## IP3

**Beyond Big Data: Combining Cognitive Computing and Analytics**

In 2011 a computer named Watson demonstrated super human question answering ability by defeating two Jeopardy! grand champions. This marked the beginning of a new era of computing, the cognitive computing era. Watson was built upon a strong foundation in natural language understanding and machine learning, and excels at question answering. In parallel, significant advances have been made in data analytics, both in the scale of data that can be analyzed, and in the methods that can be applied to extract signal from data. We have an opportunity to combine the technologies to deal with large volumes of both structured and unstructured data, to answer questions using both written facts and signals extracted from data. This talk will explore examples and identify research opportu-

nities.

Brenda L. Dietrich

IBM Research  
dietric@us.ibm.com

## IP4

**Why Most Published Studies Are Wrong But Can Be Fixed**

Observational healthcare data, such as administrative claims and electronic health records, play an increasingly prominent role in healthcare. Pharmacoepidemiologic studies in particular routinely estimate temporal associations between medical product exposure and subsequent health outcomes of interest and such studies influence prescribing patterns and healthcare policy more generally. Some authors have questioned the reliability and accuracy of such studies, but few previous efforts have attempted to measure their performance. The Observational Medical Outcomes Partnership (OMOP, <http://omop.org>) has conducted a series of experiments to empirically measure the performance of various observational study designs with regard to predictive accuracy for discriminating between true drug effects and negative controls. In this talk, I describe the past work of the Partnership, explore opportunities to expand the use of observational data to further our understanding of medical products, and highlight areas for future research and development.

David Madigan

Columbia University  
david.madigan@columbia.edu

## CP1

**Class Augmented Active Learning**

Traditional active learning encounters a cold start issue when few labelled examples are present for learning a decent initial classifier. Its poor quality subsequently affects selection of the next query and stability of the iterative learning, resulting in more human annotation effort. This paper presents a novel class augmentation technique, which enhances each class's representation which initially consists of only limited set of labelled examples. Our augmentation employs a connectivity-based influence computation algorithm with a decaying mechanism for the unlabelled samples. Besides augmentation, our method also introduces structure preserving oversampling to correct class imbalance. Extensive experiments on ten publicly available data sets demonstrate the effectiveness of our proposed method over existing state-of-the-art methods. Moreover, our proposed modules perform at the fundamental data level without any requirement to modify the standard machine learning tools.

Hong Cao, Chunyu Bao

Institute for Infocomm Research, A\*STAR  
hcao@i2r.a-star.edu.sg, cbao@i2r.a-star.edu.sg

Xiaoli Li

Institute for Infocomm Research  
xlli@i2r.a-star.edu.sg

Yew-Kwong Wong

EADS Innovation Works South Asia

david.woon@eads.net

ppeng@ust.hk, raywong@cse.ust.hk

**CP1****Batch Mode Active Learning with Hierarchical-Structured Embedded Variance**

We consider active learning when the categories are represented as a tree. Recent work has improved the traditional techniques by moving beyond “flat” structure through incorporation of the label hierarchy into the uncertainty measure. However, these methods have two major limitations when used. First, these methods roughly use the information in the label structure but do not take into account the training samples. Second, none of these methods can work in a batch mode to reduce the computational time of training. We propose a batch mode active learning scheme that exploits both the hierarchical structure of the labels and the characteristics of the training data. In addition, the selection criterion is designed to construct batches and incorporate a diversity measure. Experimental results indicate that our technique achieves a notable improvement in performance over the state-of-the-art approaches.

Yu Cheng  
Northwestern University  
IBM T.J. Watson Research Center  
chengyu05@gmail.com

ZhengZhang Chen  
Northwestern University  
zhengzhang.chen@gmail.com

Hongliang Fei, Fei Wang  
IBM T.J. Watson Research Center  
hfei@us.ibm.com, fwang@us.ibm.com

Alok Choudhary  
Dept. of Electrical Engineering and Computer Science  
Northwestern University, Evanston, USA  
choudhar@eecs.northwestern.edu

**CP1****Selective Sampling on Probabilistic Data**

In the literature of supervised learning, most existing studies assume that the labels provided by the labelers are *deterministic*, which may introduce noise easily in many real-world applications. In many applications like crowdsourcing, however, many labelers may simultaneously label the same group of instances and thus the label of each instance is associated with a probability. Motivated by this observation, we propose a new framework where each label is enriched with a probability. In this paper, we study an interactive sampling strategy, namely, *selective sampling*, in which each selected instance is labeled with a probability. Specifically, we flip a coin every time when we read a new instance and decide whether it should be labeled according to the flipping result. We prove that in our setting the label complexity can be reduced dramatically. Finally, we conducted comprehensive experiments in order to verify the effectiveness of our proposed labeling framework.

Peng Peng, RaymondChi-Wing Wong  
Department of Computer Science and Engineering  
HKUST

**CP1****Learning from Multi-User Multi-Attribute Annotations**

There are cases in numerous annotation tasks wherein despite of the presence of multiple users, each user should classify or rate multiple attributes for each sample. This situation is referred to as multi-user multi-attribute annotations in this paper. This work deals with the learning problem under multi-user multi-attribute annotations. A generative model is introduced to describe the human labeling for multi-user multi-attribute annotations. A maximum likelihood approach is leveraged to infer the parameters in the generative model, namely, ground-truth labels, user expertise, and annotation difficulties. The classifiers for each attribute are learned simultaneously. Furthermore, the correlations among attributes are taken into account during inference and learning using conditional random field. The experimental results reveal that our approach can obtain better estimation of the ground truth labels, user experts, annotation difficulties as well as attribute classifiers.

Ou Wu, Shuxiao Li  
NLPR, Institute of Automation, Chinese Academy of Sciences  
wuou@nlpr.ia.ac.cn, sxli@nlpr.ia.ac.cn

Honghui Dong  
Beijing Jiaotong University  
hhdong@bjtu.edu.cn

Ying Chen, Weiming Hu  
NLPR, Institute of Automation, Chinese Academy of Sciences  
ychen@nlpr.ia.ac.cn, wmhu@nlpr.ia.ac.cn

**CP1****Disambiguation-Free Partial Label Learning**

Partial label learning deals with the problem where each training example is associated with a set of *candidate* labels, among which only one is correct. The common strategy is to try to disambiguate their candidate labels, such as by identifying the ground-truth label iteratively or by treating each candidate label equally. Nevertheless, the above disambiguation strategy is prone to be misled by the false positive label(s) within candidate label set. In this paper, a new disambiguation-free approach to partial label learning is proposed by employing the well-known error-correcting output codes (ECOC) techniques. Specifically, to build the binary classifier with respect to each column coding, any partially labeled example will be regarded as a positive or negative training example only if its candidate label set *entirely* falls into the coding dichotomy. Experiments on controlled and real-world data sets clearly validate the effectiveness of the proposed approach.

Min-Ling Zhang  
Southeast University  
zhangml@seu.edu.cn

**CP2****Influence Maximization with Viral Product Design**

*Product design and viral marketing* are two popular con-

cepts in the marketing literature that aim at the same goal: maximizing the adoption of a new product. While the effect of the social network is nowadays kept in great consideration in any marketing-related activity, the interplay between product design and social influence is surprisingly still largely unexplored. In this paper we move a first step in this direction and study the problem of designing the features of a novel product such that its adoption, fueled by peer influence, is maximized. Our experimental evaluation on real-world data from the domain of social music consumption and social movie consumption confirms the effectiveness of the proposed framework in integrating product design in viral marketing.

Nicola Barbieri  
Yahoo Labs  
barbieri@yahoo-inc.com

Francesco Bonchi  
Yahoo! Research  
bonchi@yahoo-inc.com

## CP2

### Local Learning for Mining Outlier Subgraphs from Network Datasets

Various systems can be modeled using entity-relationship graphs. Given such a graph, one may be interested in identifying suspicious or anomalous subgraphs. A user may want to identify suspicious subgraphs matching a query template. A subgraph can be defined as anomalous based on the connectivity structure within itself as well as with its neighborhood. Existence of low-probability links and absence of high-probability links can be a good indicator of subgraph outlierness. Probability of an edge can in turn be modeled based on the weighted similarity between the attribute values of the nodes linked by the edge. We claim that the attribute weights must be learned locally for accurate link existence probability computations. We design a system that finds subgraph outliers given a graph and a query by modeling the problem as a linear optimization. Experimental results on several synthetic and real datasets show the effectiveness of the proposed approach in computing interesting outliers.

Manish Gupta, Arun Mallya, Subhro Roy, Jason Cho  
Univ of Illinois at Urbana-Champaign  
manishg.iitb@gmail.com, amallya2@illinois.edu,  
sroy9@illinois.edu, hcho33@illinois.edu

Jiawei Han  
UIUC  
hanj@illinois.edu

## CP2

### A Probabilistic Approach to Uncovering Attributed Graph Anomalies

We introduce a probabilistic model to identify anomalous subgraphs containing a significantly different percentage of a certain vertex attribute. Our framework, gAnomaly, models generative processes of vertex attributes and divides the graph into regions that are governed by such processes. Two types of regularizers are employed to smoothen the regions and facilitate vertex assignment. An iterative procedure is further proposed to find fine-grained anomalies. Experiments show gAnomaly outperforms a state-of-

the-art algorithm at uncovering anomalous subgraphs.

Nan Li, Huan Sun  
Computer Science Department  
University of California, Santa Barbara  
nanli@cs.ucsb.edu, huansun@cs.ucsb.edu

Kyle Chipman  
Neural Science Research Institute  
University of California, Santa Barbara  
kchipman@gmail.com

Jemin George  
U.S. Army Research Laboratory  
jemin.george.civ@mail.mil

Xifeng Yan  
Department of Computer Science  
University of California at Santa Barbara  
xyan@cs.ucsb.edu

## CP2

### A Generative Model with Network Regularization for Semi-Supervised Collective Classification

In recent years much effort has been devoted to Collective Classification (CC) techniques to predict labels of linked instances given a large number of labeled data. However, in many real-world applications labeled data are limited and very expensive to obtain. Recently, Semi-Supervised Collective Classification (SSCC) has been examined to leverage unlabeled data for enhancing the classification performance of CC. In this paper we propose a probabilistic generative model with network regularization (GMNR) for SSCC. Our main idea is to compute label probability distributions for unlabeled instances by maximizing log-likelihood in the generative model and the label smoothness on the network topology of data. We develop an effective EM algorithm to compute label probability distributions for label prediction. Experimental results on three real sparsely-labeled network datasets show that GMNR outperforms state-of-the-art CC algorithms and other SSCC algorithms at 0.10 significance level.

Ruichao Shi  
Shenzhen Key Laboratory of Internet Information  
Collaboratio  
Shenzhen Graduate School, Harbin Institute of  
Technology  
shrek.black@gmail.com

Qingyao Wu  
School of Computer Engineering  
Nanyang Technological University  
qywu@ntu.edu.sg

Yunming Ye  
Harbin Institute of Technology, China  
yeyunming@hit.edu.cn

Shen-Shyang Ho  
School of Computer Engineering  
Nanyang Technological University  
ssho@ntu.edu.sg

## CP2

### Dava: Distributing Vaccines over Networks under



### Prior Information

In this paper, we study how to immunize healthy nodes, in presence of already infected nodes. Efficient algorithms for such a problem can help public-health experts make more informed choices. First we formulate the Data-Aware Vaccination problem, and prove it is NP-hard and also that it is hard to approximate. Secondly, we propose two effective polynomial-time heuristics DAVA and DAVA-fast. Finally, we also demonstrate the scalability and effectiveness of our algorithms through extensive experiments on multiple real networks including epidemiology datasets, which show substantial gains of up to 10 times more healthy nodes at the end.

Yao Zhang  
Virginia Tech  
yaozhang@cs.vt.edu

B. Aditya Prakash  
CS, VT  
badityap@cs.vt.edu

### CP3

#### FlexiFaCT: Scalable Flexible Factorization of Coupled Tensors on Hadoop

Given multiple datasets of relational data, how can we efficiently decompose our data into latent factors? Factorization of a single matrix or tensor has attracted much attention, e.g. in the Netflix challenge, with users rating movies. However, we often have additional side information, e.g. demographic data. Incorporating this information leads to the coupled factorization problem. So far, it has been solved for small datasets. We provide a distributed, scalable method for matrix, tensor, and coupled factorization through stochastic gradient descent. We offer the following contributions: (1) Versatility: Our algorithm can perform factorizations with flexible objective functions, e.g. the Frobenius norm, sparse factorization, and non-negative factorization. (2) Scalability: FlexiFaCT scales to unprecedented sizes in both the data and model, with up to billions of parameters, and runs on standard Hadoop. (3) Convergence proofs showing that FlexiFaCT converges, even with projections.

Alex Beutel, Abhimanu Kumar  
Carnegie Mellon University  
abeutel@cs.cmu.edu, abhimank@cs.cmu.edu

Evangelos Papalexakis  
CMU  
epapalex@cs.cmu.edu

Partha Talukdar, Christos Faloutsos  
Carnegie Mellon University  
ppt@cs.cmu.edu, christos@cs.cmu.edu

Eric Xing  
School of Computer Science  
CMU  
epxing@cs.cmu.edu

### CP3

#### Context-Preserving Hashing for Fast Text Classification

There have been a number of approximate algorithms for text similarity computation, such as min-wise hashing, ran-

dom projection, and feature hashing, which are based on the bag-of-words representation. A limitation of their ‘flat-set’ representation is that context information and semantic hierarchy cannot be preserved. In this paper, we aim to fast compute similarities between texts while also preserving context information. To take into account semantic hierarchy, we consider a notion of ‘multi-level exchangeability’ which can be applied at word-level, sentence-level, paragraph-level, etc. We employ a nested-set to represent a multi-level exchangeable object. To fingerprint nested-sets for fast comparison, we propose a Recursive Min-wise Hashing (RMH) algorithm at the same computational cost of the standard min-wise hashing algorithm. The empirical studies show that the proposed context-preserving hashing method can significantly outperform min-wise hashing.

Bin Li, Lianhua Chi  
University of Technology, Sydney  
libin82cn@gmail.com, lianhua1221@gmail.com

Xingquan Zhu  
Florida Atlantic University  
hill.zhu@gmail.com

### CP3

#### Dusk: A Dual Structure-Preserving Kernel for Supervised Tensor Learning with Applications to Neuroimages

We introduce a new scheme to design structure-preserving kernels for supervised tensor learning. Specifically, we demonstrate how to leverage the naturally available structure within the tensorial representation to encode prior knowledge in the kernel. Our approach is an extension of the conventional kernels in the vector space to tensor space. We applied our novel kernel in conjunction with SVM to real-world tensor classification problems including brain fMRI classification for three different diseases.

Lifang He  
South China University of Technology  
lifanghescut@gmail.com

Xiangnan Kong, Philip Yu  
University of Illinois at Chicago  
xkong4@uic.edu, psyu@cs.uic.edu

Ann Ragin  
Northwestern University  
ann-ragin@northwestern.edu

Zhifeng Hao  
Faculty of Computer, Guangdong University of Technology  
mazfhao@scut.edu.cn

Xiaowei Yang  
South China University of Technology, China  
xwyang@scut.edu.cn

### CP3

#### Vog: Summarizing and Understanding Large Graphs

How can we succinctly describe a million-node graph with a few simple sentences? How can we measure the importance of a set of discovered subgraphs in a large graph? These are exactly the problems we focus on. Our main ideas are

to construct a 'vocabulary' of subgraph-types that often occur in real graphs (e.g., stars, cliques, chains), and from a set of subgraphs, find the most succinct description of a graph in terms of this vocabulary. We measure success in a well-founded way by means of the Minimum Description Length (MDL) principle: a subgraph is included in the summary if it decreases the total description length of the graph. Our contributions are three-fold: (a) formulation: we provide a principled encoding scheme to choose vocabulary subgraphs; (b) algorithm: we develop VOG, an efficient method to minimize the description cost, and (c) applicability: we report experimental results on multi-million-edge real graphs, including Flickr and the Notre Dame web graph.

Danai Koutra  
Carnegie Mellon University  
danai@cs.cmu.edu

U Kang  
KAIST  
ukang@cs.kaist.ac.kr

Jilles Vreeken  
Max Planck Institute for Informatics  
Saarland University  
jilles@mpi-inf.mpg.de

Christos Faloutsos  
Carnegie Mellon University  
christos@cs.cmu.edu

### CP3

#### **Turbo-Smt: Accelerating Coupled Sparse Matrix-Tensor Factorizations by 200x**

How can we correlate the neural activity in the human brain as it responds to typed words, with properties of these terms (like edible, fits in hand)? This is one of many settings of the Coupled Matrix-Tensor Factorization (CMTF) problem. We introduce Turbo-SMT, a meta-method capable of boosting the performance of any CMTF algorithm, by up to 200x, along with an up to 65 fold increase in sparsity, with comparable accuracy to the baseline. We apply Turbo-SMT to BrainQ, a dataset consisting of a (nouns, brain voxels, human subjects) tensor and a (nouns, properties) matrix, with coupling along the nouns dimension. Turbo-SMT is able to find meaningful latent variables, as well as to predict brain activity with competitive accuracy.

Evangelos Papalexakis  
CMU  
epapalex@cs.cmu.edu

Tom Mitchell  
Carnegie Mellon University  
tom.mitchell@cmu.edu

Nicholas Sidiropoulos  
University of Minnesota  
nikos@ece.umn.edu

Christos Faloutsos  
Carnegie Mellon University  
christos@cs.cmu.edu

Partha Talukdar  
CMU

partha.talukdar@cs.cmu.edu

Brian M. Murphy  
Queens University of Belfast

### CP4

#### **Unveiling Variables in Systems of Linear Equations**

Motivated by emerging applications including vehicular traffic networks and crowdsourcing systems, we introduce the UNVEIL problem defined as follows: given a system of linear equations with less equations than unknowns identify a set of  $k$  variables to query their values so that the number of variables that can be uniquely deduced in the new system is maximized. We will discuss the complexity of this problem and algorithms for solving it. Finally, we will present experimental evidence of the utility of these algorithms in real datasets.

Behzad Golshan, Evimaria Terzi  
Boston University  
behzad@cs.bu.edu, evimaria@cs.bu.edu

### CP4

#### **Transductive Hsic Lasso**

Sparse regression methods such as Lasso, are commonly used for analysis of high-dimensional, small-sample datasets, due to their good generalization and feature-selection properties. In this work, we develop a novel method, called Transductive HSIC Lasso, that incorporates transduction into a nonlinear sparse regression approach known as HSIC Lasso. Our method exploits the structure of the HSIC Lasso, which maximizes the relevance between the selected features and the label, while minimizing the redundancy between the selected features; transduction is achieved by including unlabeled samples into the redundancy computation. Our experiments demonstrate advantages of the proposed method over the state-of-the-art approaches, both on simulated and real-life data.

Dan He, Irina Rish, Laxmi Parida  
IBM T.J. Watson Research  
dhe@us.ibm.com, rish@us.ibm.com, parida@us.ibm.com

### CP4

#### **How Can I Index My Thousands of Photos Effectively and Automatically? An Unsupervised Feature Selection Approach**

It is not easy for a person to organize a large collection of photos. To automatically index a large photo collection, we first propose a simple strategy to extract meaningful semantics from original images as candidate dimensions, and then propose an efficient unsupervised feature selection method to select a dimension subset that is small and sufficient to index every image uniquely. Our empirical study demonstrates both the efficiency and the effectiveness of our method.

Juhua Hu  
Simon Fraser University  
juhuah@sfu.ca

Jian Pei  
School of Computing Science  
Simon Fraser University  
jpei@cs.sfu.ca

Jie Tang  
Tsinghua University  
jietang@tsinghua.edu.cn

#### CP4

##### Robust Subspace Discovery Through Supervised Low-Rank Constraints

In this paper, we propose a Supervised Regularization based Robust Subspace (SRRS) approach via low-rank learning. Unlike existing subspace methods, our approach jointly learns low-rank representations and a robust subspace from noisy observations. To improve the classification performance, class label information is incorporated as supervised regularization. The problem can be formulated as a constrained rank minimization objective function. Experimental results on four datasets demonstrate that our approach outperforms the state-of-the-art subspace and low-rank learning methods.

Sheng Li, Yun Fu  
Northeastern University  
shengli@ece.neu.edu, yunfu@ece.neu.edu

#### CP4

##### Adaptive Quantization for Hashing: An Information-Based Approach to Learning Binary Codes

Large-scale data mining and retrieval applications have increasingly turned to compact binary data representations as a way to achieve both fast queries and efficient data storage. Most of existing algorithms focus on learning a set of projection hyperplanes for the data and simply binarizing the result from each hyperplane, but this neglects the fact that informativeness may not be uniformly distributed across the projections. In this paper, we address this issue by proposing a novel adaptive quantization (AQ) strategy that adaptively assigns varying numbers of bits to different hyperplanes based on their information content. Our method provides an information-based schema that preserves the neighborhood structure of data points, and we jointly find the globally optimal bit-allocation for all hyperplanes.

Caiming Xiong, Wei Chen, Gang Chen, David Johnson, Jason Corso  
SUNY UB  
cxiong@buffalo.edu, wchen23@buffalo.edu,  
gangchen@buffalo.edu, davidjoh@buffalo.edu,  
jcorso@buffalo.edu

#### CP5

##### Active Multitask Learning Using Both Latent and Supervised Shared Topics

Multitask learning (MTL) via a shared representation has been adopted to alleviate problems with sparsity of labeled data across different learning tasks. Active learning, on the other hand, reduces the cost of labeling examples by making informative queries over an unlabeled pool of data. A unification of both of these approaches can potentially be useful in settings where labeled information is expensive to obtain but the learning tasks have some common characteristics. This paper introduces two such models – Active Doubly Supervised Latent Dirichlet Allocation and its non-parametric variation that integrate MTL and active learning in the same framework. These models make use

of both latent and supervised shared topics to accomplish multitask learning. Experimental results on both document and image classification show that integrating MTL and active learning along with shared latent and supervised topics is superior to other methods which do not employ all of these components.

Ayan Acharya  
University of Texas at Austin  
Department of ECE  
aacharya@utexas.edu

#### CP5

##### Linking Heterogeneous Input Spaces with Pivots for Multi-Task Learning

Most existing works on multi-task learning (MTL) assume the same input space for different tasks. In this paper, we address a general setting where different tasks have heterogeneous input spaces. Our key observation is that in many real applications, there might exist some correspondence among the inputs of different tasks, which is referred to as pivots. For such applications, we first propose a learning scheme for multiple tasks and analyze its generalization performance. Then we focus on the problems where only a limited number of the pivots are available, and propose a general framework to leverage the pivot information. We further propose an effective optimization algorithm to find both the mappings and the prediction model. Experimental results demonstrate its effectiveness, especially with very limited number of pivots.

Jingrui He  
IBM T.J. Watson Research Center  
jingrui.he@gmail.com

Yan Liu  
University of Southern California  
yanliu.cs@usc.edu

Qiang Yang  
Hong Kong University of Science and Technology  
Department of Computer Science and Engineering  
qyang@cse.ust.hk

#### CP5

##### Multi-Task Feature Selection on Multiple Networks via Maximum Flows

We propose a new formulation of multi-task feature selection coupled with multiple network regularizers, and show that the problem can be exactly and efficiently solved by maximum flow algorithms. This method contributes to one of the central topics in data mining: How to exploit structural information in multivariate data analysis, which has numerous applications, such as gene regulatory and social network analysis. On simulated data, we show that the proposed method leads to higher accuracy in discovering causal features by solving multiple tasks simultaneously using networks over features. Moreover, we apply the method to multi-locus association mapping with *Arabidopsis thaliana* genotypes and flowering time phenotypes, and demonstrate its ability to recover more known phenotype-related genes than other state-of-the-art methods.

Mahito Sugiyama, Chloé-Agathe Azencott  
Max Planck Institutes Tübingen  
mahito.sugiyama@tuebingen.mpg.de,  
chloe-agathe.azencott@mines-paristech.fr

Dominik Grimm  
Max Planck Institutes Tübingen  
Eberhard Karls Universität Tübingen  
dominik.grimm@tuebingen.mpg.de

Yoshinobu Kawahara  
ISIR, Osaka University  
kawahara@ar.sanken.osaka-u.ac.jp

Karsten Borgwardt  
Max Planck Institute for Intelligent Systems  
Eberhard Karls Universität Tübingen  
karsten.borgwardt@tuebingen.mpg.de

## CP5

### Mixed-Transfer: Transfer Learning over Mixed Graphs

Heterogeneous transfer learning has been proposed as a new learning strategy to improve performance in a target domain by leveraging data from other heterogeneous source domains where feature spaces can be different across different domains. In this paper, we propose a novel algorithm named Mixed-Transfer to solve heterogeneous transfer learning with noise co-occurrence data from the Web. It is composed of three components, that is, a cross domain harmonic function to avoid personal biases, a joint transition probability graph of mixed instances and features to model the heterogeneous transfer learning problem, a random walk process to simulate the label propagation on the graph and avoid the data sparsity problem.

Ben Tan  
Hong Kong University of Science and Technology,  
Clear Water Bay, Hong Kong  
btan@cse.ust.hk

Erheng Zhong  
Hong Kong University of Science and Technology  
Department of Computer Science and Engineering  
ezhong@cse.ust.hk

Michael K. Ng  
Department of Mathematics, Hong Kong Baptist  
University  
mng@math.hkbu.edu.hk

Qiang Yang  
Hong Kong University of Science and Technology  
Department of Computer Science and Engineering  
qyang@cse.ust.hk

## CP5

### Multi-Graph Learning with Positive and Unlabeled Bags

In this paper, we formulate a new multi-graph learning task with only positive and unlabeled bags, where labels are only available for bags but not for individual graphs inside the bag. This problem setting raises significant challenges because bag-of-graph setting does not have features to directly represent graph data, and no negative bags exists for deriving discriminative classification models. To solve the challenge, we propose a puMGL learning framework which relies on two iteratively combined processes for multi-graph learning: (1) deriving features to represent graphs for learning; and (2) deriving discriminative models with only positive and unlabeled graph bags. Experiments

and comparisons on real-world multi-graph data demonstrate the algorithm performance.

Jia Wu  
Centre for Quantum Computation and Intelligent  
Systems, FEIT  
University of Technology Sydney, Australia  
jia.wu@student.uts.edu.au

Zhibin Hong, Shirui Pan  
University of Technology Sydney, Australia  
zhibin.hong@student.uts.edu.au,  
shirui.pan@student.uts.edu.au

Xingquan Zhu  
Florida Atlantic University, USA  
xzhu3@fau.edu

Chengqi Zhang  
University of Technology, Sydney  
chengqi.zhang@uts.edu.au

Zhihua Cai  
China University of Geosciences, Wuhan, China  
zhcai@cug.edu.cn

## CP6

### Forecasting a Moving Target: Ensemble Models for Ili Case Count Predictions

Modern epidemiological forecasts of common illnesses, such as the flu, rely on both traditional surveillance sources as well as digital surveillance data. However, most published studies have been retrospective. The reports about flu are also lagged by several weeks and revised over several weeks. We posit that effectively handling this uncertainty is one of the key challenges for a real-time prediction system in this sphere. In this paper, we present a detailed prospective analysis on the generation of robust quantitative predictions about temporal trends of flu activity, using several *surrogate* data sources for 15 Latin American countries. We present our findings about the effects of correcting the uncertainty associated with official flu estimates and compare accuracy at model and data level fusion. In the process we present a novel matrix factorization approach using neighborhood embedding and compare the method to several baseline methods.

Prithwish Chakraborty  
Dept. of Computer Science, Virginia Tech, Blacksburg,  
VA, USA  
PhD  
prithwi@vt.edu

Pejman Khadivi  
Dept. of Computer Science, Virginia Tech, Blacksburg,  
VA, US  
pejman@vt.edu

Bryan Lewis  
Virginia Bioinformatics Institute, VA Tech  
blewis@vbi.vt.edu

Aravindan Mahendiran  
Dept. of Computer Science, Virginia Tech, Blacksburg,  
VA, US  
aravind@vt.edu

Jiangzhou Chen  
Virginia Tech  
chenj@vbi.vt.edu

Patrick Bulter  
Dept. of Computer Science, Virginia Tech, Blacksburg,  
VA, US  
pabutler@vbi.vt.edu

Elaine O. Nsoesie  
Virginia Tech  
Boston Children's Hospital  
onelaine@vbi.vt.edu

Sumiko Mekaru  
Children's Hospital Boston, MA, USA  
sumiko.mekaru@childrens.harvard.edu

John Brownstein  
Children's Hospital Boston, MA, USA  
john.brownstein@childrens.harvard.edu

Madhav Marathe  
Virginia Bioinformatics Institute, VA Tech  
mmarathe@vbi.vt.edu

Naren Ramakrishnan  
Computer Science  
Virginia Tech  
naren@cs.vt.edu

## CP6

### Keeping Up with Innovation: A Predictive Framework for Modeling Healthcare Data with Evolving Clinical Interventions

Medical outcomes are inexorably linked to patient illness and clinical interventions. Interventions change the course of disease, crucially determining outcome. Traditional outcome prediction models build a single classifier by augmenting interventions with disease information. Interventions, however, differentially affect prognosis, thus a single prediction rule may not suffice to capture variations. Interventions also evolve over time as more advanced interventions replace older ones. To this end, we propose a Bayesian nonparametric, supervised framework that models a set of intervention groups through a mixture distribution building a separate prediction rule for each group, and allows the mixture distribution to change with time. Experiments on synthetic and medical cohorts for 30-day readmission prediction demonstrate the superiority of the proposed model over clinical and data mining baselines.

Sunil K. Gupta, Santu Rana, Dinh Phung, Svetha Venkatesh  
Deakin University, Geelong Waurn Ponds Campus  
Victoria, Australia  
sunil.gupta@deakin.edu.au, santu.rana@deakin.edu.au,  
dinh.phung@deakin.edu.au, svetha.venkatesh@deakin.edu.au

## CP6

### Predictive Learning in the Presence of Heterogeneity and Limited Training Data

Many real-world applications possess heterogeneity in their data, and further lack sufficient training data, making predictive learning difficult. However, there often exists a

structure among the data instances, which can be leveraged in predictive learning. We present an approach that reduces the model complexity while addressing heterogeneity, and demonstrate its application for forest cover estimation. We show that our framework captures meaningful information about heterogeneity in data, improves prediction performance, and is robust to over-fitting.

Anuj Karpatne  
University of Minnesota  
anuj@cs.umn.edu

Ankush Khandelwal  
University of Minnesota- Twin Cities  
ankush@cs.umn.edu

Shyam Boriah  
Department of Computer Science  
University of Minnesota  
sboriah@cs.umn.edu

Vipin Kumar  
University of Minnesota  
kumar@cs.umn.edu

## CP6

### Turning the Tide: Curbing Deceptive Yelp Behaviors

The popularity and influence of reviews, make sites like Yelp ideal targets for malicious behaviors. We present Marco, a novel system that exploits the unique combination of social, spatial and temporal signals gleaned from Yelp, to detect venues whose ratings are impacted by fraudulent reviews. Marco increases the cost and complexity of attacks, by imposing a tradeoff on fraudsters, between their ability to impact venue ratings and their ability to remain undetected. We contribute a new dataset to the community, which consists of both ground truth and gold standard data. We show that Marco significantly outperforms state-of-the-art approaches, by achieving 94% accuracy in classifying reviews as fraudulent or genuine, and 95.8% accuracy in classifying venues as deceptive or legitimate. Marco successfully flagged 244 deceptive venues from our large dataset with 7,435 venues, 270,121 reviews and 195,417 users.

Bogdan Carbutar, Mahmudur Rahman  
FIU  
carbutar@gmail.com, mrahman.fiu@gmail.com

Jaime Ballesteros  
Nokia  
jaime.ballesteros@here.com

George Burri  
Jive Software  
george@burrillmail.com

Duen Horng Chau  
Georgia Tech  
polo@gatech.edu

## CP6

### Consumer Segmentation and Knowledge Extraction from Smart Meter and Survey Data

Many electricity suppliers around the world are deploying



smart meters to gather fine-grained spatiotemporal consumption data and to effectively manage the collective demand of their consumer base. In this paper, we introduce a structured framework and a discriminative index that can be used to segment the consumption data along multiple contextual dimensions such as locations, communities, seasons, weather patterns, holidays, etc. The generated segments can enable various higher-level applications such as usage-specific tariff structures, theft detection, consumer-specific demand response programs, etc. Our framework is also able to track consumers' behavioral changes, evaluate different temporal aggregations, and identify main characteristics which define a cluster.

Tri Kurniawan Wijaya

EPFL

tri-kurniawan.wijaya@epfl.ch

Tanuja Ganu

IBM Research, India

tanuja.ganu@in.ibm.com

Dipanjan Chakraborty

IBM Research India

cdipanjan@in.ibm.com

Karl Aberer

EPFL

karl.aberer@epfl.ch

Deva Seetharam

IBM Research India

dseetharam@in.ibm.com

## CP7

### A Marginalized Denoising Method for Link Prediction in Relational Data

We propose a novel link prediction algorithm, where the problem of predicting missing links is cast as a problem of matrix denoising. We train a mapping function by recovering the originally observed matrix from conceptually “infinite” corrupted matrices where some links are randomly masked from the observed matrix. By re-applying the learned function to the observed relational matrix, we aim to “denoise” the observed matrix and thus to recover the unobserved links.

Zheng Chen, Weixiong Zhang

Washington University in St. Louis

zheng.chen@wustl.edu, weixiong.zhang@wustl.edu

## CP7

### A Deep Learning Approach to Link Prediction in Dynamic Networks

Time varying problems usually have complex underlying structures represented as dynamic networks where entities and relationships appear and disappear over time. In this paper, we propose a novel deep learning framework, i.e., Conditional Temporal Restricted Boltzmann Machine (ctrBM), which predicts links based on individual transition variance as well as influence introduced by local neighbors.

Xiaoyi Li, Nan Du, Hui Li

State University of New York at Buffalo

xiaoyili@buffalo.edu, nandu@buffalo.edu,

hli24@buffalo.edu

Kang Li

The State University of New York at Buffalo

kli22@buffalo.edu

Jing Gao

University at Buffalo

jing@buffalo.edu

Aidong Zhang

Department of Computer Science

State University of New York at Buffalo

azhang@buffalo.edu

## CP7

### Convex Optimization for Binary Classifier Aggregation in Multiclass Problems

We present a convex optimization method for an optimal aggregation of binary classifiers to estimate class membership probabilities in multiclass problems.

Sunho Park, Tae Hyun Hwang

Department of Clinical Sciences, UTSW Medical Center

Sunho.Park@UTSouthwestern.edu,

taehyun.hwang@utsouthwestern.edu

Seungjin Choi

Department of Computer Science and Engineering,

Pohang University of Science and Technology

seungjin@postech.ac.kr

## CP7

### Learning on Probabilistic Labels

Probabilistic information becomes more prevalent nowadays and can be found easily in many applications like crowdsourcing and pattern recognition. In this paper, we focus on a dataset which contains probabilistic information for classification. Based on this probabilistic dataset, we propose a classifier and give a theoretical bound linking the error rate of our classifier and the number of instances needed for training. Interestingly, we find that our theoretical bound is asymptotically at least no worse than the previously best-known bounds developed based on the traditional dataset. Furthermore, our classifier guarantees a fast rate of convergence compared with traditional classifiers. Experimental results show that our proposed algorithm has a higher accuracy than the traditional algorithm.

Peng Peng, Raymond ChiWing Wong

Department of Computer Science and Engineering

HKUST

ppeng@ust.hk, raywong@cse.ust.hk

Philip S. Yu

Department of Computer Science

University of Illinois at Chicago

psyu@cs.uic.edu

## CP7

### AUC Dominant Unsupervised Ensemble of Binary Classifiers

Ensemble methods are widely used in practice with the hope of obtaining better predictive performance than could be obtained from any of the constituent classifiers in the ensemble. Most of the existing literature is concerned with learning ensembles in a supervised setting. In this paper

we propose an unsupervised iterative algorithm to combine the discriminant scores from different binary classifiers. We prove that (under certain assumptions) the Area Under the ROC Curve (AUC) of the resulting ensemble is greater than or equal to the AUC of the best classifier (with maximum AUC). We also experimentally validate this claim on a number of datasets and also show that the performance is better than the supervised ensembles.

Priyanka Agrawal, Vikas C. Raykar, Amrita Saha  
IBM Research, Bangalore  
priyanka.svnit@gmail.com, vikasraykar@gmail.com, amr-saha4@in.ibm.com

## CP8

### Make It Or Break It: Manipulating Robustness in Large Networks

The function and performance of networks rely on their robustness, defined as their ability to continue functioning in the face of damage (targeted attacks or random failures) to parts of the network. Prior research has proposed a variety of measures to quantify robustness and various manipulation strategies to alter it. In this paper, our contributions are two-fold. First, we critically analyze various robustness measures and identify their strengths and weaknesses. Our analysis suggests natural connectivity, based on the weighted count of loops in a network, to be a reliable measure. Second, we propose the first principled manipulation algorithms that directly optimize this robustness measure, which lead to significant performance improvement over existing, ad-hoc heuristic solutions. Extensive experiments on real-world datasets demonstrate the effectiveness and scalability of our methods against a long list of competitor strategies.

Hau Chan  
Stony Brook University  
hauchan@cs.stonybrook.edu

Leman Akoglu  
Stonybrook University  
leman@cs.stonybrook.edu

Hanghang Tong  
City Collge, CUNY  
tong@cs.cuny.cuny.edu

## CP8

### Triangle Counting in Streamed Graphs Via Small Vertex Covers

We present a new randomized algorithm for estimating the number of triangles in massive graphs revealed as a stream of edges in arbitrary order. It exploits the fact that real-world graphs often have small vertex covers, which allows to reduce the space and sample complexity of triangle counting algorithms. Experiments on real-world graphs validate our theoretical analysis and show that our algorithm yields more accurate estimates than state-of-the-art approaches using the same amount of space.

David Garcia-Soriano  
Yahoo! Research Barcelona  
davidgs@yahoo-inc.com

Konstantin Kutzkov  
IT University of Copenhagen, Denmark

kutzkov@gmail.com

## CP8

### Laplacian Spectral Properties of Graphs from Random Local Samples

In this paper, we study the relationship between the eigenvalue spectrum of normalized Laplacian matrix and the structure of ‘local’ subgraphs. We propose techniques to estimate the spectral properties from a random collection of local subgraphs. Particularly, an algorithm is provided to estimate the spectral moments of the normalized Laplacian matrix. Moreover, we propose to compute bounds on the spectral radius based on convex optimization. Numerical results on a large-scale e-mail network are illustrated.

Zhengwei Wu, Victor Preciado  
University of Pennsylvania  
zhengwei@seas.upenn.edu, preciado@seas.upenn.edu

## CP8

### It Takes Two to Tango: Exploring Social Tie Development with Both Online and Offline Interactions

Understanding social tie development is crucial for user engagement in social networking services. In this paper, we analyze the social interactions, both online and offline, and investigate the development of their social ties. In this study, we aim to answer three key questions: 1) is there a correlation between online and offline interactions? 2) how is the social tie developed via heterogeneous interaction channels? 3) would the development of social tie between two users be affected by their common friends? To achieve our goal, we develop a Social-aware Hidden Markov Model (SaHMM) that explicitly takes into account the factor of common friends in measure of the social tie development. Our experiments show that the social tie development captured by our SaHMM is significantly more consistent to lifetime profiles of users.

Peifeng Yin  
Pennsylvania State University  
pzy102@cse.psu.edu

Qi He  
LinkedIn  
qhe@linkedin.com

Xingjie Liu  
Square  
liuxingjie03@gmail.com

Wang-Chien Lee  
Pennsylvania State University  
wlee@cse.psu.edu

## CP8

### Adaptive User Distance Modeling in Social Media

One important challenge in social network analysis is how to model users’ distance as a single measure. We propose to model this distance by simultaneously exploring users’ profile attributes and local network structures. Due to the sparsity of data, where each user may interact with just a few people and only a few users provide their profile information, it is typically difficult to learn effective distance measures for any individual network. One important observation is that, people nowadays engage in multiple social

networks, such as Facebook, Twitter, etc., where auxiliary knowledge from related networks can help alleviate the data sparsity problem. Nonetheless, due to the network differences, borrowing knowledge directly does not work well. Instead, we propose an adaptive metric learning framework. The basic idea is to exploit knowledge from related networks collectively through embedding and employ boosting-based techniques to eliminate irrelevant attributes.

Erheng Zhong  
Hong Kong University of Science and Technology  
Department of Computer Science and Engineering  
ezhong@cse.ust.hk

Wei Fan  
Huawei Noah's Ark Lab  
david.fanwei@huawei.com

Qiang Yang  
Hong Kong University of Science and Technology  
Department of Computer Science and Engineering  
qyang@cse.ust.hk

## CP9

### Automatic Construction and Ranking of Topical Keyphrases on Collections of Short Documents

We introduce a framework for topical keyphrase generation and ranking, based on the output of a topic model run on a collection of short documents. By shifting from the unigram-centric traditional methods of keyphrase extraction and ranking to a phrase-centric approach, we are able to directly compare and rank phrases of different lengths. Our method defines a function to rank topical keyphrases so that more highly ranked keyphrases are considered to be more representative phrases for that topic. We study the performance of our framework on multiple real world document collections, and also show that it is more scalable than comparable phrase-generating models.

Marina Danilevsky  
University of Illinois Urbana-Champaign  
danilev1@illinois.edu

Chi Wang  
University of Illinois at Urbana-Champaign  
chiwang1@illinois.edu

Nihit Desai, Xiang Ren  
University of Illinois Urbana-Champaign  
nhdesai2@illinois.edu, xren7@illinois.edu

Jingyi Guo  
University of Massachusetts Amherst  
jingyi@cs.umass.edu

Jiawei Han  
UIUC  
hanj@illinois.edu

## CP9

### Recurrent Chinese Restaurant Process with a Duration-Based Discount for Event Identification from Twitter

Event identification and analysis on Twitter has become an important task. Recently the Recurrent Chinese Restaurant

Process (RCRP) has been successfully used for event identification from news streams and news-centric social media streams. However, these models cannot be directly applied to Twitter for two reasons: (1) Events emerge and die out fast on Twitter, (2) Most Twitter posts are personal interest oriented while only a small fraction is event related. Motivated by these challenges, we propose a new nonparametric model which considers burstiness. We further combine this model with traditional topic models to identify both events and topics simultaneously.

Qiming Diao, Jing Jiang  
Singapore Management University  
qiming.ustc@gmail.com, jingjiang@smu.edu.sg

## CP9

### A Dynamic Nonparametric Model for Characterizing the Topical Communities in Social Streams

The data in social networks often come as *streams*, i.e., both text content (e.g., emails, user postings) and network structure (e.g., user friendship) evolve over time. To capture the time-evolving latent structures in such social streams, we propose a fully *nonparametric Dynamic Topical Community Model* (nDTCM), where infinite latent community variables coupled with infinite latent topic variables in each epoch, and the temporal dependencies between variables across epochs are modeled via the rich-get-richer scheme. nDTCM can effectively characterize three dynamic aspects in social streams: the number of communities or topics changes (e.g., new communities or topics are born and old ones die out); the popularity of communities or topics evolves; the semantics such as community topic distribution, community participant distribution and topic word distribution drift.

Ziqi Liu  
MOEKLINNS Lab, Department of Computer Science  
Xi'an Jiaotong University, China  
zqilau@gmail.com

Qinghua Zheng  
Department of Computer Science  
Xi'an Jiaotong University, China  
qhzheng@mail.xjtu.edu.cn

Fei Wang  
IBM T J Watson Research Center  
feiwang03@gmail.com

Zhenhua Tian, Bo Li  
Department of Computer Science  
Xi'an Jiaotong University, China  
zhhtian@gmail.com, boli.cs8@gmail.com

## CP9

### Joint Author Sentiment Topic Model

Traditional works in sentiment analysis and aspect rating prediction do not take *author preferences* and *writing style* into account during rating prediction of reviews. In this work, we introduce Joint Author Sentiment Topic Model (JAST), a generative process of writing a review by an author. Authors have different topic preferences, 'emotional' attachment to topics, writing style based on the distribution of *semantic* (topic) and *syntactic* (background) words and tendency to *switch topics*. JAST uses Latent Dirichlet Allocation to learn the distribution of author-specific topic preferences and emotional attachment to topics. It

uses a Hidden Markov Model to capture short range syntactic and long range semantic dependencies in reviews to capture *coherence* in author writing style. JAST jointly discovers topics in a review, author preferences for topics, topic ratings as well as overall review rating from the point of view of an author.

Subhabrata Mukherjee  
Max-Planck-Institut für Informatik  
subhabrata.mukherjee.ju@gmail.com

Gaurab Basu, Sachindra Joshi  
IBM Research Lab  
gaurabas@in.ibm.com, jsachind@in.ibm.com

## CP9

### A Constrained Hidden Markov Model Approach for Non-Explicit Citation Context Extraction

In this paper we present a constrained hidden markov model based approach for extracting non-explicit citing sentences in research articles. Our method involves first independently training a separate HMM for each citation in the article being processed and then performing a constrained joint inference to label non-explicit citing sentences. Results on a standard test collection show that our method significantly outperforms the baselines and is comparable to the state of the art approaches.

Parikshit Sondhi  
University of Illinois Urbana Champaign  
Walmart Labs  
sondhi1.uiuc@gmail.com

ChengXiang Zhai  
University of Illinois at Urbana Champaign  
czhai@illinois.edu

## CP10

### A Constructive Setting for the Problem of Density Ratio Estimation

We introduce a constructive setting for the problem of density ratio estimation through the solution of a multi-dimensional integral equation. In this equation, not only its right hand side is approximately known, but also the integral operator is approximately defined. We show that this ill-posed problem has a rigorous solution and obtain the solution in a closed form. The key element of this solution is the novel  $V$ -matrix, which captures the geometry of the observed samples. We compare our method with previously proposed ones, using both synthetic and real data. Our experimental results demonstrate the good potential of the new approach.

Vladimir Vapnik  
NEC Laboratories America  
vlad@nec-labs.com

Igor Braga  
University of Sao Paulo  
igorab@icmc.usp.br

Rauf Izmailov  
Applied Communication Sciences

rizmailov@appcomsci.com

## CP10

### Embed and Conquer: Scalable Embeddings for Kernel $k$ -Means on MapReduce

The kernel  $k$ -means is an effective method for data clustering which extends the commonly-used  $k$ -means algorithm to work on a similarity matrix over complex data structures. It is, however, computationally very complex as it requires the complete kernel matrix to be calculated and stored. Further, its kernelized nature hinders the parallelization of its computations on modern scalable infrastructures for distributed computing. In this paper, we are defining a family of low-dimensional embeddings that allows for scaling kernel  $k$ -means on MapReduce via an efficient and unified parallelization strategy. Afterwards, we propose two practical methods for low-dimensional embedding that adhere to our definition of the embeddings family. Exploiting the proposed parallelization strategy, we present two scalable MapReduce algorithms for kernel  $k$ -means. We demonstrate the effectiveness and efficiency of the proposed algorithms through an empirical evaluation on benchmark datasets.

Ahmed Elgohary, Ahmed Farahat, Mohamed Kamel, Fakhri Karray  
University of Waterloo  
aalgohary@uwaterloo.ca, afarahat@uwaterloo.ca, mkamel@uwaterloo.ca, karray@uwaterloo.ca

## CP10

### Density Estimation with Adaptive Sparse Grids for Large Data Sets

Even though kernel density estimation is widely used, its performance highly depends on the kernel bandwidth, and it can become computationally expensive for large data sets. We present a sparse-grid-based density estimation method which discretizes the density on basis functions centered at grid points rather than on kernels centered at the data points. Thus, the costs of evaluating the estimated density function are independent from the number of data points. Numerical results confirm that our method is competitive to current kernel-based approaches with respect to accuracy and runtime.

Benjamin Peherstorfer  
SCCS, Department of Informatics  
Technische Universität München  
pehersto@mit.edu

Dirk Pflüger  
Universität Stuttgart, SimTech-IPVS  
Simulation of Large Systems  
Dirk.Pflueger@ipvs.uni-stuttgart.de

Hans-Joachim Bungartz  
Technische Universität München, Department of Informatics  
Chair of Scientific Computing in Computer Science  
bungartz@in.tum.de

## CP10

### On Randomly Projected Hierarchical Clustering with Guarantees

Hierarchical clustering algorithms are very popular owing

to their simplicity of implementation and ease of interpretation of the results. However, they impose a large runtime cost, thus limiting their applicability to typically only small data instances. Here we mitigate this shortcoming and explore fast hierarchical clustering algorithms based on random projections. We present adaptations for well-known HC variants, such as single (SLC) and average (ALC) linkage clustering. The algorithms maintain, with arbitrary high probability, the outcome of hierarchical clustering as well as the worst-case running-time guarantees.

Johannes Schneider  
Zurich insurances  
vollkoff@gmail.com

## CP10

### Self-Taught Spectral Clustering Via Constraint Augmentation

Although constrained spectral clustering has been used extensively for the past few years, all work assumes the guidance (constraints) are given by humans. Original formulations of the problem assumed the constraints are given passively whilst later work allowed actively polling an Oracle (human experts). In this paper, for the first time to our knowledge, we explore the problem of augmenting the given constraint set for constrained spectral clustering algorithms. This moves spectral clustering towards the direction of self-teaching as has occurred in the supervised learning literature. We present a formulation for self-taught spectral clustering and show that the self-teaching process can drastically improve performance without further human guidance.

Xiang Wang  
IBM Research  
wangxi@us.ibm.com

Jun Wang  
IBM Thomas J. Watson Research Center  
Business Analytics and Mathematical Sciences  
Department  
wangjun@us.ibm.com

Buyue Qian  
IBM Research  
bqian@us.ibm.com

Fei Wang  
IBM T.J. Watson Research Center  
fwang@us.ibm.com

Ian Davidson  
University of California, Davis  
davidson@cs.ucdavis.edu

## CP11

### User Preference Learning with Multiple Information Fusion for Restaurant Recommendation

In this paper, we develop a generative probabilistic model to exploit the multi-aspect ratings of restaurants for restaurant recommendation. Also, the geographic proximity is integrated into the probabilistic model to capture the geographic influence. Moreover, the profile information, which contains customer/restaurant-independent features and the shared features, is also integrated into the model. Finally, we conduct a comprehensive experimental study

on a real-world data set. The experimental results clearly demonstrate the benefit of exploiting multi-aspect ratings and the improvement of the developed generative probabilistic model.

Hui Xiong  
Rutgers, the State University of New Jersey  
hxiong@rutgers.edu

Yanjie Fu, Bin Liu  
Rutgers University  
yanjie.fu@rutgers.edu, binben.liu@rutgers.edu

Yong Ge  
UNC Charlotte  
yong.ge@uncc.edu

Zijun Yao  
Rutgers University  
zijun.yao@rutgers.edu

## CP11

### On Modeling Community Behaviors and Sentiments in Microblogging

In this paper, we propose the *CBS* topic model, a probabilistic graphical model, to derive the user communities in microblogging networks based on the sentiments they express on their generated content and behaviors they adopt. As a topic model, *CBS* can uncover hidden topics and derive user topic distribution. In addition, our model associates topic-specific sentiments and behaviors with each user community. Notably, *CBS* has a general framework that accommodates multiple types of behaviors simultaneously. Our experiments on two Twitter datasets show that the *CBS* model can effectively mine the representative behaviors and emotional topics for each community. We also demonstrate that *CBS* model perform as well as other state-of-the-art models in modeling topics, but outperforms the rest in mining user communities.

Tuan-Anh Hoang  
Living Analytics Research Centre  
Singapore Management University  
tahoang.2011@phdis.smu.edu.sg

William W. Cohen  
Carnegie Mellon University  
wcohen@cs.cmu.edu

Ee-Peng Lim  
Singapore Management University  
eplim@smu.edu.sg

## CP11

### Contextual Combinatorial Bandit and Its Application on Diversified Online Recommendation

we propose a principled framework called *contextual combinatorial bandit* and introduce its application on diversified online recommendation task. Specifically, we formulate the recommendation task as a diversity promoting exploration/exploitation problem. On each of  $n$  rounds, the learning algorithm sequentially selects a set of items according to the item-selection strategy that balances *exploration* and *exploitation*, and collects the user feedback on these selected items. For the above bandit problem, we further provide a algorithm that achieves  $\tilde{O}(\sqrt{n})$  regret after



playing  $n$  rounds.

Lijing Qin  
Tsinghua University  
qinlj09@mail.tsinghua.edu.cn

Shouyuan Chen  
The Chinese University of Hong Kong  
sychen@cse.cuhk.edu.hk

Xiaoyan Zhu  
Tsinghua University  
zxy-dcs@tsinghua.edu.cn

## CP11

### Selecting a Representative Set of Diverse Quality Reviews Automatically

In this paper, we study how to find a representative set of high quality reviews to cover diversified aspects of user opinions. Existing work cannot solve this problem well. To overcome the drawbacks of existing methods, we define a new problem of finding the minimum set of reviews to cover all of features with different sentiment polarity and high quality without user-defined parameters. To solve the problem efficiently, we define potential objective function and develop greedy algorithm to find the solution in polynomial-time with approximation guarantee. We also propose two strategies to further reduce the number of reviews and to prune the search space respectively. Comprehensive experiments conducted on real review sets show that the proposed methods are effective and outperform existing methods.

Nana Xu  
Beijing  
100872  
nanaxu23@gmail.com

Hongyan Liu  
Tsinghua University  
liuhy@sem.tsinghua.edu.cn

Jiawei Chen, Jun He  
Renmin University of China  
orangeruc@ruc.edu.cn, hej@ruc.edu.cn

## CP11

### Latent Factor Transition for Dynamic Collaborative Filtering

User preferences change over time and capturing such changes is essential for developing accurate recommender systems. Despite its importance, only a few works in collaborative filtering have addressed this issue. In this paper, we consider evolving preferences and we model user dynamics by introducing and learning a transition matrix for each user's latent vectors between consecutive time windows. Intuitively, the transition matrix for a user summarizes the time-invariant pattern of the evolution for the user. We first extend the conventional probabilistic matrix factorization and then improve upon this solution through its fully Bayesian model. These solutions take advantage of the model complexity and scalability of conventional Bayesian matrix factorization, yet adapt dynamically to user's evolving preferences. We evaluate the effectiveness of these solutions through empirical studies on six large-

scale real life data sets.

Chenyi Zhang  
Zhejiang University  
zhangchenyi.zju@gmail.com

Ke Wang  
Simon Fraser University, Canada  
wangk@cs.sfu.ca

Hongkun Yu  
Simon Fraser University  
hongkuny@sfu.ca

Jianling Sun  
Zhejiang University  
sunjl@zju.edu.cn

Ee-Peng Lim  
Singapore Management University  
eplim@smu.edu.sg

## CP12

### A Statistical Learning Theory Framework for Supervised Pattern Discovery

This paper formalizes a latent variable inference problem we call *supervised pattern discovery*, the goal of which is to find sets of observations that belong to a single ‘pattern.’ We discuss two versions of the problem and prove uniform risk bounds for both. In the first version, collections of patterns can be generated in an arbitrary manner and the data consist of multiple labeled collections. In the second version, the patterns are assumed to be generated independently by identically distributed processes. These processes are allowed to take an arbitrary form, so observations within a pattern are not in general independent of each other. The bounds for the second version of the problem are stated in terms of a new complexity measure, the quasi-Rademacher complexity.

Jonathan H. Huggins  
MIT  
jhuggins@mit.edu

Cynthia Rudin  
Massachusetts Institute of Technology  
rudin@mit.edu

## CP12

### Ensembles of Elastic Distance Measures for Time Series Classification

Many alternative distance measures for comparing time series have been proposed for Time Series Classification (TSC) problems. These include variants of Dynamic Time Warping (DTW), such as weighted and derivative DTW, and edit distance-based measures, including Longest Common Subsequence and Time Warp Edit Distance. Our aim is to test two hypotheses related to these measures. Firstly, we test whether there is any significant difference when training nearest neighbour classifiers with these measures for TSC problems. Secondly, we test whether combining these measures through simple ensemble schemes gives significantly better accuracy. We have carried out one of the largest studies ever conducted into TSC. Our first finding is that there is no significant difference in accuracy between the distance measures on our data sets. Our second find-

ing, and the major contribution of this work, is to define an ensemble classifier that significantly outperforms the individual classifiers.

Jason Lines, Anthony Bagnall  
University of East Anglia  
j.lines@uea.ac.uk, Anthony.Bagnall@uea.ac.uk

## CP12

### Finding the True Frequent Itemsets

Frequent Itemsets (FIs) mining from a dataset  $D$  is a fundamental primitive in data mining. In many applications  $D$  is a collection of samples obtained from an unknown probability distribution  $\pi$  on transactions. By extracting the FIs in  $D$  one attempts to infer itemsets that are generated by  $\pi$  with probability at least  $\theta$ , which we call the True Frequent Itemsets (TFIs). The set of FIs often contains a huge number of *false positives* w.r.t. the TFIs. We design and analyze an algorithm to identify a threshold  $\hat{\theta}$  such that the collection of itemsets with frequency at least  $\hat{\theta}$  in  $D$  contains only TFIs with probability at least  $1 - \delta$ , for some user-specified  $\delta$ . Our method uses the (empirical) VC-dimension of the problem at hand. This allows us to identify almost all the TFIs without including any false positive.

Matteo Riondato  
Department of Computer Science  
Brown University  
matteo@cs.brown.edu

Fabio Vandin  
Department of Mathematics and Computer Science  
University of Southern Denmark  
vandinf@imada.sdu.dk

## CP12

### When and Where: Predicting Human Movements Based on Social Spatial-Temporal Events

Predicting both the time and the location of human movements is valuable but challenging for a variety of applications. To address this problem, we propose an approach considering both the periodicity and the sociality of human movements. We first define a new concept, Social Spatial-Temporal Event (SSTE), to represent social interactions among people. For the time prediction, we characterise the temporal dynamics of SSTE with an ARMA (AutoRegressive Moving Average) model. To dynamically capture the SSTE kinetics, we propose a Kalman Filter based learning algorithm to learn and incrementally update the ARMA model as a new observation becomes available. For the location prediction, we propose a ranking model the periodicity and the sociality of human movements are simultaneously taken into consideration for improving the prediction accuracy. Extensive experiments conducted on real data sets validate our proposed approach.

Ning Yang  
College of Computer Science  
Sichuan University  
yangning@scu.edu.cn

Xiangnan Kong, Fengjiao Wang, Philip Yu  
University of Illinois at Chicago

xkong4@uic.edu, fwang27@uic.edu, psyu@uic.edu

## CP12

### Discovery of Rare Sequential Topic Patterns in Document Stream

In this paper, we propose the novel mining problem of rare Sequential Topic Patterns (STPs) from Internet document streams, which are rare on the whole but relatively often for specific users. It can be applied in many practical fields, such as personalized context-aware recommendation and real-time monitoring on abnormal user behaviors. We design a group of effective algorithms to discover these user-related rare STPs based on the temporal and probabilistic information of extracted topics.

Zhongyi Hu, Hongan Wang  
Institute of Software,  
Chinese Academy of Sciences  
sunnyddhzy@126.com, hongan@iscas.ac.cn

Jiaqi Zhu  
Institute of Software, Chinese Academy of Sciences  
zhujq@ios.ac.cn

Maozhen Li  
School of Engineering and Design,  
Brunel University  
maozhen.li@brunel.ac.uk

Ying Qiao, Changzhi Deng  
Institute of Software,  
Chinese Academy of Sciences  
qiaoying@iscas.ac.cn, changzhi@iscas.ac.cn

## MS1

### Clustering with Linear Programming

Arindam Banerjee  
University of Minnesota  
banerjee@cs.umn.edu

## MS1

### Clustering Evaluation and Validation: Some Results, Challenges, and Research Questions

Ricardo J. Campello  
Department of Computer Sciences  
University of So Paulo at So Carlos  
campello@icmc.usp.br

## MS1

### Utilizing Multiple Clusterings: Beyond Consensus Clustering

Xiaoli Z. Fern  
Oregon State University  
xfern@eecs.oregonstate.edu

## MS1

### Title Not Available - Karypis

George Karypis

University of Minnesota / AHPARC  
karypis@cs.umn.edu

## PP1

### Discovering Groups of Time Series with Similar Behavior in Multiple Small Intervals of Time

The focus of this paper is to address the problem of discovering groups of time series that share similar behavior in multiple small intervals of time. This problem has two characteristics: i) There are exponentially many combinations of time series that needs to be explored to find these groups, ii) The groups of time series of interest need to have similar behavior only in some subsets of the time dimension. We present an Apriori based approach to address this problem. We evaluate it on a synthetic dataset and demonstrate that our approach can directly find all groups of intermittently correlated time series without finding spurious groups unlike other alternative approaches that find many spurious groups. We also demonstrate, using a neuroimaging dataset, that groups of intermittently coherent time series discovered by our approach are reproducible on independent sets of time series data. In addition, we demonstrate the utility of our approach on an S&P 500 stocks data set.

Gowtham Atluri  
Dept. of Computer Science  
University of Minnesota  
gowtham@cs.umn.edu

Michael Steinbach  
University of Minnesota  
steinbac@cs.umn.edu

Kelvin Lim  
Dept. of Psychiatry  
University of Minnesota  
kolim@umn.edu

Angus MacDonald Iii  
Dept. of Psychology  
University of Minnesota  
angus@umn.edu

Vipin Kumar  
University of Minnesota  
kumar@cs.umn.edu

## PP1

### Large Building Associations Between Markers of Environmental Stressors and Adverse Human Health Impacts Using Frequent Itemset Mining

Health effects are unknown for most of the >83,000 chemicals in commerce, creating challenges in balancing industrial needs with health protection. Here we describe the use of frequent itemset mining to identify exposure  $\Rightarrow$  health associations from the NHANES, a large-scale survey measuring biomarkers of environmental exposure and human health. Our approach is designed to enable more effective knowledge discovery of potential health impacts of environmental chemicals by facilitating the comprehensive data mining and meta-analysis of the NHANES dataset.

Shannon M. Bell  
Oak Ridge Institute for Science and Education  
U.S. Environmental Protection Agency

bell.shannon@epa.gov

Stephen Edwards  
US Environmental Agency  
edwards.stephen@epa.gov

## PP1

### Characterising Seismic Data

When a seismologist analyses a new seismogram it is often useful to have access to a set of similar seismograms. For example if she tries to determine the event, if any, that caused the particular readings on her seismogram. So, the question is: when are two seismograms similar? We propose a framework based on wavelets and MDL for this. Using MULTi-KRIMP we are able to extract succinct sets of characteristics from seismograms on which the similarity is based.

Roel Bertens, Arno Siebes  
Universiteit Utrecht  
Department of Information and Computing Sciences  
R.Bertens@uu.nl, a.p.j.m.siebes@uu.nl

## PP1

### Density-Based Clustering Validation

One of the most challenging aspects of clustering is validation, which is the objective and quantitative assessment of clustering results. A number of different relative validity criteria have been proposed for the validation of *globular* clusters. Not all data, however, are composed of globular clusters. Density-based clustering algorithms seek partitions with high density areas of points (clusters, not necessarily globular) separated by low density areas, possibly containing noise objects. In these cases relative validity indices proposed for globular cluster validation may fail. In this paper we propose a relative validation index for density-based, arbitrarily shaped clusters. The index assesses clustering quality based on the relative density connection between pairs of objects. Our index is formulated on the basis of a new kernel density function, which is used to compute the density of objects and to evaluate the within- and between-cluster density connectedness of clustering results. Experiments on synthetic and real world data show the effectiveness of our approach for the evaluation and selection of clustering algorithms and their respective appropriate parameters.

Davoud Moulavi  
Department of Computing Science, University of Alberta  
moulavi@ualberta.ca

Pablo Jaskowiak  
Department of Computing Sciences, University of Alberta  
jaskowia@ualberta.ca

Ricardo J. Campello  
Department of Computer Sciences  
University of So Paulo at So Carlos  
campello@icmc.usp.br

Arthur Zimek  
LMU Munich  
zimek@dbis.lmu.de

Joerg Sander  
Department of Computing Science, University of Alberta

jsander@ualberta.ca

**PP1****Online Discovery of Group Level Events in Time Series**

Recent advances in high throughput data collection and storage technologies have led to a dramatic increase in the availability of high-resolution time series datasets in various domains. These time series reflect the dynamics of the underlying physical processes. Detecting changes in a time series over time or changes in the relationships among the time series in a dataset containing multiple contemporaneous time series can be useful to detect events in these physical processes. Contextual events detection algorithms detect changes in the relationships between multiple related time series. In this work, we introduce a new type of contextual events, called group level contextual change events. In contrast to individual contextual change events that reflect the change in behavior of one target time series against a context, group level events reflect the change in behavior of a target group of time series relative to a context group of time series.

XI Chen

University of Minnesota  
Department of Computer Science and Engineering  
chen@cs.umn.edu

Abdullah Mueen  
University of New Mexico.  
mueen@cs.unm.edu

Vijay Narayanan, Nikos Karampatziakis, Gagan Bansal  
Microsoft Corporation  
vkn@microsoft.com, nikosk@microsoft.com,  
gagan.bansal@microsoft.com

Vipin Kumar  
University of Minnesota  
kumar@cs.umn.edu

**PP1****Frugal Traffic Monitoring with Autonomous Participatory Sensing**

In this paper we propose a strategy for frugal sensing in which the participants send only a fraction of the observed traffic information to reduce costs while achieving high accuracy. The strategy is based on autonomous sensing, in which participants make decisions to send traffic information without guidance from the central server, thus reducing the communication overhead and improving privacy. We propose to use traffic flow theory in deciding whether or not to send an observation to the server. To provide accurate and computationally efficient estimation of the current traffic, we propose to use a budgeted version of the Gaussian Process model on the server side. The experiments on real-life traffic data set indicate that the proposed approach can use up to two orders of magnitude less samples than a baseline approach when estimating traffic speed on a highway network, with only a negligible loss in accuracy.

Vladimir Coric, Nemanja Djuric, Slobodan Vucetic  
Temple University  
vladimir.coric@temple.edu, nemanja.djuric@temple.edu,

vucetic@temple.edu

**PP1****Improving Credit Card Fraud Detection with Calibrated Probabilities**

In this paper, two different methods for calibrating probabilities are evaluated and analyzed in the context of credit card fraud detection, with the objective of finding the model that minimizes the real losses due to fraud. It is shown that by calibrating the probabilities and then using Bayes minimum Risk the losses due to fraud are reduced. Furthermore, because of the good overall results, the aforementioned card processing company is currently incorporating the methodology proposed in this paper into their fraud detection system. Finally, the methodology has been tested on a different application, namely, direct marketing.

Alejandro Correa Bahnsen, Aleksandar Stokanovic,  
Djamila Aouada, Bjorn Ottersten  
Luxembourg University  
al.bahnsen@gmail.com, aleksandar.stokanovic@uni.lu,  
djamila.aouada@uni.lu, bjorn.ottersten@uni.lu

**PP1****Iterative Framework Radiation Hybrid Mapping**

Building comprehensive radiation hybrid maps for large sets of markers is a computationally expensive process, since the basic mapping problem is equivalent to the traveling salesman problem. The mapping problem is also susceptible to noise, and as a result, it is often beneficial to remove markers that are not trustworthy. The resulting framework maps are typically more reliable but don't provide information about as many markers. We present an approach to mapping most markers by first creating a framework map and then incrementally adding the remaining markers. We consider chromosomes of the human genome, for which the correct ordering is known, and compare the performance of our two-stage algorithm with the Carthage radiation hybrid mapping software. We show that our approach is not only much faster than mapping the complete genome in one step, but that the quality of the resulting maps is also much higher.

Raed Seetan  
North Dakota State University  
raed.seetan@ndsu.edu

Anne M. Denton  
Department of Computer Science  
North Dakota State University  
anne.denton@ndsu.edu

Omar Al-Azzam  
University of Minnesota Crookston  
oalazzam@crk.umn.edu

Ajay Kumar  
North Dakota State University  
ajay.kumar.2@ndsu.edu

Jazarai Sturdivant  
Virginia State University  
jstu0608@students.vsu.edu

Shahryar Kianian

University of Minnesota  
shahryar.kianian@ars.usda.gov

## PP1

### Mining Interpretable and Predictive Diagnosis Codes from Multi-Source Electronic Health Records

Mining patterns from multi-source electronic health-care records (EHR) can potentially lead to better and more cost-effective treatments. We aim to find the groups of ICD-9 diagnosis codes from EHRs that can predict the improvement of urinary incontinence of patients and are also interpretable to domain experts. In particular, we incorporate two additional information from EHRs: the clinical information such as demographic, behavioral, physiological, and psycho-social variables available for same set of samples and the prior information available from clinical domain knowledge such as the clinical classification system (CCS) while grouping the ICD-9 codes to enhance their interpretability. Our results obtained from a large-scale EHR data set show that the proposed integrative framework enhances clinical interpretability substantially as compared to the baseline model obtained from ICD-9 codes only, while achieving almost the same predictive capability.

Sanjoy Dey  
University of Minnesota  
sanjoy@cs.umn.edu

Gyorgy Simon  
Institute of Health Informatics,  
University of Minnesota  
simo0342@umn.edu

Bonnie Westra  
School of Nursing,  
University of Minnesota  
westr006@umn.edu

Michael Steinbach, Vipin Kumar  
University of Minnesota  
steinbac@cs.umn.edu, kumar@cs.umn.edu

## PP1

### Online Matrix Completion Through Nuclear Norm Regularisation

It is the main goal of this paper to propose a novel method to perform matrix completion *on-line*. Motivated by a wide variety of applications, ranging from the design of recommender systems to sensor network localization through seismic data reconstruction, we consider the matrix completion problem when entries of the matrix of interest are observed *gradually*. The algorithm promoted in this article builds upon the SOFT IMPUTE approach introduced in Mazumder et al. (2010). The major novelty essentially arises from the use of a randomised technique for both computing and updating the *Singular Value Decomposition (SVD)* involved in the algorithm. Though of disarming simplicity, the method proposed turns out to be very efficient, while requiring reduced computations. Several numerical experiments based on real datasets illustrating its performance are displayed, together with preliminary results giving it a theoretical basis.

Charanpal Dhanjal  
Telecom ParisTech

charanpal.dhanjal@telecom-paristech.fr

Romarc Gaudel  
University of Lille  
romarc.gaudel@univ-lille3.fr

S Cl  men  on  
Telecom ParisTech  
stephan.clemencon@telecom-paristech.fr

## PP1

### Classifying Imbalanced Data Streams Via Dynamic Feature Group Weighting with Importance Sampling

Data stream classification and imbalanced data learning are two important areas of data mining research. Each has been well studied to date with many interesting algorithms developed. However, only a few approaches reported in literature address the intersection of these two fields due to their complex interplay. In this work, we proposed an importance sampling driven, dynamic feature group weighting framework for classifying data streams of imbalanced distribution. We derived the theoretical upper bound for the generalization error of the proposed algorithm. We also studied the empirical performance of our method on a set of benchmark synthetic and real world data, and significant improvement has been achieved over the competing algorithms in terms of standard evaluation metrics and parallel running time.

Ke Wu, Andrea Edwards  
Xavier Univ. Of Louisiana  
kwu@xula.edu, aedwards@xula.edu

Wei Fan  
Huawei Noah's Ark Lab  
david.fanwei@huawei.com

Jing Gao  
University at Buffalo  
jing@buffalo.edu

Kun Zhang  
Xavier University of Louisiana  
kzhang@xula.edu

## PP1

### Exploiting Monotonicity Constraints in Active Learning for Ordinal Classification

We consider ordinal classification and instance ranking problems where each attribute is known to have an increasing or decreasing relation with the class label. We aim to exploit such monotonicity constraints by using labeled attribute vectors to draw conclusions about the class labels of order related unlabeled ones. Assuming we have a pool of unlabeled attribute vectors, and an oracle that can be queried for class labels, the central problem is to choose a query point whose label is expected to provide the most information. We evaluate several query strategies on artificial and real data.

Ad Feelders, Pieter Soons  
Universiteit Utrecht

A.J.Feelders@uu.nl, p.s@live.nl

## PP1

### Classifier-Adjusted Density Estimation for Anomaly Detection and One-Class Classification

Density estimation methods are often regarded as unsuitable for anomaly detection in high-dimensional data due to the difficulty of estimating multivariate probability distributions. Instead, the scores from popular distance- and local-density-based methods, such as local outlier factor (LOF), are used as surrogates for probability densities. We question this infeasibility assumption and explore a family of simple statistically-based density estimates constructed by combining a probabilistic classifier with a naive density estimate. Across a number of semi-supervised and unsupervised problems formed from real-world data sets, we show that these methods are competitive with LOF and that even simple density estimates that assume attribute independence can perform strongly. We show that these density estimation methods scale well to data with high dimensionality and that they are robust to the problem of irrelevant attributes that plagues methods based on local estimates.

Lisa Friedland, Amanda Gentzel  
University of Massachusetts Amherst  
lfriedl@cs.umass.edu, agentzel@cs.umass.edu

David Jensen  
UMass Amherst  
jensen@cs.umass.edu

## PP1

### Online Anomaly Detection by Improved Grammar Compression of Log Sequences

Log sequences mining is widely used in detecting anomalies. The state-of-the-art detection methods either need significant computation, or require specific assumptions on the logs distribution. Our proposed CADM needs no assumptions and can generate alerts online with only  $O(n)$  computation. It exploits the relative entropy between test and normal logs for anomalous levels discovery and an improved grammar-based compression method for relative entropy estimation. Experiments prove CADM can achieve higher detection accuracy with minimal computation.

Yun Gao, Wei Zhou, Zhang Zhang, Jizhong Han, Dan Meng  
Institute of Information Engineering  
Chinese Academy of Sciences  
gaoyun@iie.ac.cn, zhouwei@iie.ac.cn,  
zhangzhang@iie.ac.cn, hanjizhong@iie.ac.cn,  
mengdan@iie.ac.cn

Zhiyong Xu  
Department of Math and Computer Science  
Suffolk University Boston  
zxu@mcs.suffolk.edu

## PP1

### Submodularity in Team Formation Problem

The team formation problem is about finding a team of experts for a project. Several formulations have been proposed for this problem but each of them focuses only on

a subset of design criteria such as skill coverage, social compatibility, economy, skill redundancy, etc. In this paper, for the first time, we show that most of these criteria can be encapsulated within one single formulation, namely unconstrained submodular function maximization. In our formulation, the submodular function turns out to be non-negative and non-monotone. The maximization of this function is much less explored than its monotone constrained counterpart. Recently, for this problem, a simulated annealing based randomized scheme having 0.41 approximation ratio has been proposed. We customize this algorithm to our setting and conduct an extensive set of experiments to show its efficacy. Our formulation offers many advantages such as skill cover softening, better team communication, and connectivity relaxation.

Dinesh Garg  
IBM India Research Lab  
New Delhi  
garg.dinesh@in.ibm.com

Avradeep Bhowmik  
University of Texas, Austin  
avradeep.l@utexas.edu

Vivek Borkar  
IIT Bombay  
borkar@ee.iitb.ac.in

Madhavan Pallan  
IBM India Research Lab,  
New Delhi  
mapallan@in.ibm.com

## PP1

### Quantifying Trust Dynamics in Signed Graphs, the S-Cores Approach

This paper focuses on the issue of defining models and metrics for reciprocity in signed graphs. In unsigned networks, reciprocity quantifies the predisposition of network members in creating mutual connections. On the other hand, this concept has not yet been investigated in the case of signed graphs. We capitalize on the graph degeneracy concept to identify subgraphs of the signed network in which reciprocity is more likely to occur. This enables us to assess reciprocity at a global level, rather than at a local one as in existing approaches. The large scale experiments we perform on real world trust networks lead to both interesting and intuitive results. These reciprocity measures can be used in various social applications such as trust management, community detection and evaluation of individuals. The global reciprocity we define in this paper is closely correlated to the clustering structure of the graph, more than the local reciprocity, indicated by our experimental evaluation.

Christos Giatsidis, Michalis Vazirgiannis  
Ecole Polytechnique  
xristosakamad@gmail.com, mvazirg@lix.polytechnique.fr

Dimitrios Thilikos  
CNRS, LIRMM, Montpellier  
sedthilk@thilikos.info

Silviu Maniu  
University of Hong Kong  
smaniu@cs.hku.hk



Bogdan Cautis  
Universite Paris-Sud  
bogdan.cautis@u-psud.fr

## PP1

### On Finding the Point Where There Is No Return: Turning Point Mining on Game Data

Gaming expertise is usually accumulated through playing or watching many game instances, and identifying critical moments in these game instances called turning points. Turning Point Rules (TPRs) are game patterns that almost always lead to some irreversible outcomes. In this paper, we formulate the notion of irreversible outcome property which can be combined with pattern mining so as to automatically extract TPRs from any given game datasets. To show the usefulness of TPRs, we apply them to Tetris, a popular game. We mine TPRs from Tetris games and generate challenging game sequences so as to help training an intelligent Tetris algorithm.

Wei Gong  
School of Information Systems  
Singapore Management University  
wei.gong.2011@smu.edu.sg

Ee-Peng Lim, Feida Zhu  
Singapore Management University  
eplim@smu.edu.sg, fdzhu@smu.edu.sg

Palakorn Achananuparp, David Lo  
School of Information Systems  
Singapore Management University  
palakorna@smu.edu.sg, davidlo@smu.edu.sg

## PP1

### Memory-Efficient Query-Driven Community Detection with Application to Complex Disease Associations

Rather than enumerating all communities in a network, mining for communities containing user-specified query nodes produces output more relevant to the users preference. In this paper, we propose and systematically compare two memory efficient approaches: out-of-core and index-based. The achieved scalability of our methods enables the discovery of diseases that are known to be or likely associated with Alzheimer's, when a genome-scale network is mined with Alzheimer's biomarker genes as query nodes.

Steve Harenberg, Ramona Seay, Stephen Ranshous, Kanchana Padmanabhan, Jitendra Harlalka, Eric Schendel, Michael OBrien, Rada Chirkova  
North Carolina State University  
harenberg@ncsu.edu, rgseay@ncsu.edu, smransho@ncsu.edu, kpadman@ncsu.edu, jkharlal@ncsu.edu, erschend@ncsu.edu, mpobrie@ncsu.edu, rychirko@ncsu.edu

William Hendrix  
Northwestern University  
whendrix@northwestern.edu

Alok Choudhary  
Dept. of Electrical Engineering and Computer Science  
Northwestern University, Evanston, USA  
choudhar@eecs.northwestern.edu

Vipin Kumar  
University of Minnesota  
kumar@cs.umn.edu

Murali Doraiswamy  
Duke University  
murali.doraiswamy@duke.edu

Nagiza Samatova  
North Carolina State University  
Oak Ridge National Laboratory  
samatova@csc.ncsu.edu

## PP1

### Unsupervised Feature Learning by Deep Sparse Coding

In this paper, we propose a new unsupervised feature learning framework, namely Deep Sparse Coding (DeepSC), that extends sparse coding to a multi-layer architecture for visual object recognition tasks. The main innovation of the framework is that it connects the sparse-encoders from different layers by a sparse-to-dense module. The sparse-to-dense module is a composition of a local spatial pooling step and a low-dimensional embedding process, which takes advantage of the spatial smoothness information in the image. As a result, the new method is able to learn multiple layers of sparse representations of the image which capture features at a variety of abstraction levels and simultaneously preserve the spatial smoothness between the neighboring image patches. Combining the feature representations from multiple layers, DeepSC achieves the state-of-the-art performance on multiple object recognition tasks.

Yunlong He  
Georgia Institute of Technology  
he.yunlong@gmail.com

Koray Kavukcuoglu  
DeepMind Technologies  
koray@deepmind.com

Yun Wang  
Princeton University  
yunwang@princeton.edu

Arthur Szlam  
The City College of New York  
aszlam@ccny.cuny.edu

YanJun Qi  
University of Virginia  
yanjun@virginia.edu

## PP1

### Recovering Missing Labels of Crowdsourcing Workers

Labels collected from crowdsourcing platforms may be incorrect or missing. While most existing work focuses on modeling the labeling errors, this paper proposes an algorithm to predict the missing labels of crowd workers. We adopt thoughts from semi-supervised learning and utilize the particular consistency between crowd workers. Experiments show that our algorithm can recover such missing labels properly, which are useful in predicting the ground

truth and discovering properties of crowd workers.

Qingyang Hu

College of Computer Science and Technology, Zhejiang  
Univers  
huqingyang@zju.edu.cn

Kevin Chiew

Provident Technology Pte. Ltd., Singapore  
kev.chiew@gmail.com

Hao Huang

School of Computing,  
National University of Singapore,  
huanghao@comp.nus.edu.sg

Qinming He

College of Computer Science and Technology,  
Zhejiang University  
hqm@zju.edu.cn

## PP1

### Detecting Influence Relationships from Graphs

Graphs have been used to represent objects and object connections in many applications. Mining influence relationships from graphs has gained increasing interests because providing influence information about object connections can facilitate graph exploration, connection recommendations, etc. In this paper, we study the problem of detecting influence aspects, on which objects are connected, and influence degrees, with which one graph node influences other nodes on given aspects. Existing techniques focus on inferring either the influence degrees or influence types from graphs. We propose two generative Aspect Influence Models, OAIM and LAIM, to detect both influence aspects and influence degrees at the aspect level. We compare these models with one baseline approach which considers only the text content of objects. The empirical studies on citation graphs and networks of users from Twitter show that our models can discover more effective results than the baseline approach.

Chuan Hu

New Mexico State University  
chuanhu@nmsu.edu

Huiping Cao

Computer Science  
New Mexico State University  
hcao@cs.nmsu.edu

Chaomin Ke

New Mexico State University  
raykcm@nmsu.edu

## PP1

### Less Is More: Similarity of Time Series under Linear Transformations

When comparing time series, z-normalization preprocessing and dynamic time warping (DTW) distance became almost standard procedure. This paper makes a point against carelessly using this setup by discussing implications and alternatives. A (conceptually) simpler distance measure is proposed that allows for a linear transformation of amplitude and time only, but is also open for other normalizations (unachievable by z-normalization preprocess-

ing). Lower bounding techniques are presented for this measure that apply directly to raw series.

Frank Höppner

Ostfalia University of Applied Sciences  
f.hoeppner@ostfalia.de

## PP1

### Meta: Multi-Resolution Framework for Event Summarization

Event summarization is an effective process that mines and organizes event patterns to represent the original events. It allows the analysts to quickly gain the general idea of the events. In recent years, several event summarization algorithms have been proposed, but they all focus on how to find out the optimal summarization results, and are designed for one-time analysis. As event summarization is a comprehensive analysis work, merely handling this problem with a single optimal algorithm is not enough. In the absence of an integrated summarization solution, we propose an extensible framework META to enable analysts to easily and selectively extract and summarize events from different views with different resolutions. In this framework, we store the original events in a carefully-designed data structure that enables an efficient storage and multi-resolution analysis.

Yexi Jiang

Florida International University  
yjian004@cs.fiu.edu

Chang-Shing Perng

IBM Research  
perng@us.ibm.com

Tao Li

Florida International University  
taoli@cs.fiu.edu

## PP1

### Beating Human Analysts in Nowcasting Corporate Earnings by Using Publicly Available Stock Price and Correlation Features

Corporate earnings are a crucial indicator for investment and business valuation. Despite their importance and the fact that classic econometric approaches fail to match analyst forecasts by orders of magnitude, the automatic prediction of corporate earnings from public data is not in the focus of current machine learning research. In this paper, we present for the first time a fully automatized machine learning method for earnings prediction that at the same time a) only relies on publicly available data and b) can outperform human analysts. The latter is shown empirically in an experiment involving all S&P 100 companies in a test period from 2008 to 2012. The approach employs a simple linear regression model based on a novel feature space of stock market prices and their pairwise correlations. With this work we follow the recent trend of nowcasting, i.e., of creating accurate contemporary forecasts of undisclosed target values based on publicly observable proxy variables.

Michael Kamp, Mario Boley, Thomas Gärtner

Fraunhofer IAIS  
michael.kamp@iais.fraunhofer.de,  
mario.bolei@iais.fraunhofer.de,

thomas.gaertner@iais.fraunhofer.de

## PP1

### Adversarial Learning with Bayesian Hierarchical Mixtures of Experts

Many data mining applications operate in adversarial environment, for example, webpage ranking in the presence of web spam. A growing number of adversarial data mining techniques are recently developed, providing robust solutions under specific defense-attack models. Existing techniques are tied to distributional assumptions geared towards minimizing the undesirable impact of given attack models. However, the large variety of attack strategies renders the adversarial learning problem multimodal. Therefore, it calls for a more flexible modeling ideology for equivocal input. In this paper we present a Bayesian hierarchical mixtures of experts for adversarial learning. Optimal attacks minimizing the likelihood of malicious data are modeled interactively at both expert and gating levels in the learning hierarchy. We demonstrate that our adversarial hierarchical-mixtures-of-experts learning model is robust against adversarial attacks on both artificial and real data.

Yan Zhou, Murat Kantarcioglu  
University of Texas at Dallas  
yzhou07@gmail.com, muratk@utdallas.edu

## PP1

### Large-Scale Multi-Label Learning with Incomplete Label Assignments

Multi-label learning deals with the classification problems where each instance can be assigned with multiple labels simultaneously. Conventional multi-label learning approaches mainly focus on exploiting label correlations. The label sets for training instances are usually assumed to be fully labeled without any missing labels. However, in many real-world cases, the label assignments for training instances can be incomplete. This problem is typical when the number instances is very large, and the labeling cost is very high, which makes it almost impossible to get a fully labeled training set. In this paper, we study the problem of large-scale multi-label learning with incomplete label assignments. We propose an approach based upon positive and unlabeled stochastic gradient descent and stacked models, which can effectively and efficiently consider missing labels and label correlations simultaneously, and is very scalable, that has linear time complexities over the size of the data.

Xiangnan Kong  
University of Illinois at Chicago  
xkong4@uic.edu

Zhaoming Wu  
Tsinghua University, China  
arieswuthu@gmail.com

Li-Jia Li  
Yahoo! Research, USA.  
lijiali@yahoo-inc.com

Ruofei Zhang  
Microsoft, USA.  
rfzhang@gmail.com

Philip Yu

University of Illinois at Chicago  
psyu@cs.uic.edu

Hang Wu  
Tsinghua University, China  
wuhang56@gmail.com

Wei Fan  
IBM T.J.Watson Research  
wei.fan@gmail.com

## PP1

### Large-Scale Kernel Ranksvm

Learning to rank is an important task for recommendation systems, online advertisement and web search. Among those learning to rank methods, rankSVM is a widely used model. Both linear and nonlinear (kernel) rankSVM have been extensively studied, but the lengthy training time of kernel rankSVM remains a challenging issue. In this paper, after discussing difficulties of training kernel rankSVM, we propose an efficient method to handle these problems. The idea is to reduce the number of variables from quadratic to linear with respect to the number of training instances, and efficiently evaluate the pairwise losses. Our setting is applicable to a variety of loss functions. Further, general optimization methods can be easily applied to solve the reformulated problem. Implementation issues are also carefully considered. Experiments show that our method is faster than state-of-the-art methods for training kernel rankSVM.

Tzu-Ming Kuo, Ching-Pei Lee, Chih-Jen Lin  
National Taiwan University  
b99902073@csie.ntu.edu.tw, r00922098@csie.ntu.edu.tw,  
cjlin@csie.ntu.edu.tw

## PP1

### Directed Interpretable Discovery in Tensors with Sparse Projection

Tensors or multiway arrays are useful constructs capable of representing complex graphs and multi-source data, etc. Analyzing such complex data requires enforcing simplicity so as to give meaningful insights. Sparse tensor decomposition typically requires appending penalty term(s) to the optimization, which provides a number of challenges. Not only is tuning the penalty weights time consuming, but the resulting decomposition often has a non-uniform distribution of sparsity. This is undesirable for some data sets where we want each factor to have similar sparsity for easier interpretation. We propose an alternative method that allows the user to specify the exact amount of sparsity and where the sparsity should be. This is achieved by augmenting the alternating non-negative least squares algorithm with a projection step. We demonstrate our works usefulness in finding interpretable features for real world problems in fMRI scan data and face image analysis without any pre-processing.

Chia-Tung Kuo  
Department of Computer Science  
University of California, Davis  
tomkuo@ucdavis.edu

Ian Davidson  
University of California, Davis

davidson@cs.ucdavis.edu

## PP1

### Efficient Matching of Substrings in Uncertain Sequences

Substring matching is fundamental to data mining methods for sequential data. It involves checking the existence of a short subsequence within a longer sequence, ensuring no gaps within a match. Whilst a large amount of existing work has focused on substring matching and mining techniques for certain sequences, there are only a few results for uncertain sequences. Uncertain sequences provide powerful representations for modelling sequence behavioural characteristics in emerging domains, such as bioinformatics, sensor streams and trajectory analysis. In this paper, we focus on the core problem of computing substring matching probability in uncertain sequences and propose an efficient dynamic programming algorithm for this task. We demonstrate our approach is both competitive theoretically, as well as effective and scalable experimentally. Our results contribute towards a foundation for adapting classic sequence mining methods to deal with uncertain data.

Yuxuan Li

University of Melbourne  
yuxli@student.unimelb.edu.au

James Bailey  
The University of Melbourne  
baileyj@unimelb.edu.au

Lars Kulik  
University of Melbourne  
lkulik@unimelb.edu.au

Jian Pei  
School of Computing Science  
Simon Fraser University  
jpei@cs.sfu.ca

## PP1

### Result Integrity Verification of Outsourced Bayesian Network Structure Learning

There has been considerable recent interest in the data-mining-as-a-service paradigm: the client that lacks computational resources outsources his/her data and data mining needs to a third-party service provider. One of the security issues of this outsourcing paradigm is how the client can verify that the service provider indeed has returned correct data mining results. In this paper, we focus on the problem of result verification of outsourced Bayesian network (BN) structure learning. We consider the untrusted service provider that intends to return wrong BN structures. We develop three efficient probabilistic verification approaches to catch the incorrect BN structure with high probability and cheap overhead. Our experimental results demonstrate that our verification methods can capture wrong BN structure effectively and efficiently.

Wendy Hui Wang  
Stevens Institute of Technology  
hwang4@stevens.edu

Ruilin Liu

Department of Computer Science  
Stevens Institute of Technology

rliu3@stevens.edu

Changhe Yuan  
Queens College, CUNY  
changhe.yuan@qc.cuny.edu

## PP1

### The Benefits of Personalized Smartphone-Based Activity Recognition Models

Abstract not available at time of publication.

Jeff Lockhart

Fordham University  
lockhart@cis.fordham.edu

## PP1

### A New Framework for Traffic Anomaly Detection

Trajectory data is becoming more and more popular nowadays and extensive studies have been conducted on trajectory data. One important research direction about trajectory data is the anomaly detection which is to find all anomalies based on trajectory patterns in a road network. In this paper, we introduce a *road segment-based* anomaly detection problem, which is to detect the abnormal road segments each of which has its “real” traffic deviating from its “expected” traffic and to infer the *major causes* of anomalies on the road network. First, a *deviation-based* method is proposed to quantify the anomaly of reach road segment. Second, based on the observation that one anomaly from a road segment can trigger other anomalies from the road segments nearby, a *diffusion-based* method based on a *heat diffusion model* is proposed to infer the major causes of anomalies on the whole road network. To validate our methods, we conduct intensive experiments on a large real-world GPS dataset of about 23,000 taxis in Shenzhen, China to demonstrate the performance of our algorithms.

Jinsong Lan  
Computer Network Information Center,  
Chinese Academy of Sciences  
lanjinsong@cnic.cn

Cheng Long

Department of Computer Science and Engineering,  
The Hong Kong University of Science and Technology  
clong@cse.ust.hk

Raymond C.-W. Wong  
Department of Computer Science and Engineering  
The Hong Kong University of Science and Technology  
raywong@cse.ust.hk

Youyang Chen  
School of Computer Science,  
Beijing Institute of Technology  
chen.youyang@qq.com

Yanjie Fu  
Rutgers University  
yanjie.fu@rutgers.edu

Danhuai Guo  
Computer Network Information Center,  
Chinese Academy of Sciences  
guodanhuai@cnic.cn

Shuguang Liu  
Chongqing Institutes of Green and Intelligent Technology  
Chinese Academy of Sciences  
liushuguang@cigit.ac.cn

Yong Ge  
UNC Charlotte  
yong.ge@uncc.edu

Yuanchun Zhou, Jianhui Li  
Computer Network Information Center,  
Chinese Academy of Sciences  
zyc@cnic.cn, lijh@cnic.cn

## PP1

### A Graphical Model Approach to Atlas-Free Mining of Mri Images

We propose a graphical model framework based on conditional random fields (CRFs) to mine MRI brain images. As a proof-of-concept, we apply CRFs to the problem of brain tissue segmentation. Experimental results show robust and accurate performance on tissue segmentation comparable to other state-of-the-art segmentation methods. In addition, results show that our algorithm generalizes well across data sets and is less susceptible to outliers. Our method relies on minimal prior knowledge unlike atlas-based techniques, which assume images map to a normal template. Our results show that CRFs are a promising model for tissue segmentation, as well as other MRI data mining problems such as anatomical segmentation and disease diagnosis where atlas assumptions are unreliable in abnormal brain images.

Chris Magnano, Ameet Soni  
Swarthmore College  
Computer Science Department  
cmagnan1@swarthmore.edu, soni@cs.swarthmore.edu

Sriraam Natarajan  
Indiana University  
School of Informatics and Computing  
natarasr@indiana.edu

Gautam Kunapuli  
UtopiaCompression Corporation  
gautam@utopiacompression.com

## PP1

### ROCsearch — An ROC-guided Search Strategy for Subgroup Discovery

ROCsearch, an ROC-based beam search variant, automatically adapts its search behavior to the properties and resulting search landscape of a given dataset. A sensible search width for each search level is automatically determined by analyzing previous level results in ROC space. While producing equivalent results, ROCsearch is an order of magnitude more efficient than traditional beam search, and domain experts can now eschew the hard task of ascertaining a proper beam width parameter.

Marvin Meeng  
LIACS, Leiden University  
m.meeng@liacs.leidenuniv.nl

Wouter Duivesteijn  
Fakultät für Informatik , Technische Universität

Dortmund  
wouter.duivesteijn@tu-dortmund.de

Arno Knobbe  
LIACS, Leiden University  
a.j.knobbe@liacs.leidenuniv.nl

## PP1

### Community Discovery in Social Networks Via Heterogeneous Link Association and Fusion

In this paper, we adapt a heterogeneous data clustering algorithm, called Generalized Heterogeneous Fusion Adaptive Resonance Theory (GHF-ART), for discovering user communities in heterogeneous social networks. Comparing with existing algorithms, GHF-ART has several advantages, including 1) performing real-time matching of patterns and one-pass learning which guarantee low computational cost; 2) Do not need the number of clusters a priori and 3) employing a weighting function to incrementally assess the importance of all the feature channels.

Lei Meng, Ah Hwee Tan  
Nanyang Technological University  
menglei.thunder@gmail.com, asahtan@ntu.edu.sg

## PP1

### Discriminative Density-Ratio Estimation

In order to deal with covariate shift in classification problems, we propose a novel method to discriminatively reweight the training samples through estimating the density ratio between the joint distributions of the training and test data for each class. The proposed algorithm is an iterative procedure that alternates between estimating the class information for the test data and estimating the new class-wise density ratios between joint distributions. Experiments on synthetic and benchmark datasets demonstrate the superiority of the proposed method.

Yun-Qian Miao, Ahmed Farahat, Mohamed Kamel  
University of Waterloo  
yqmiao@uwaterloo.ca, afarahat@uwaterloo.ca, mkamel@uwaterloo.ca

## PP1

### Accelerating Graph Adjacency Matrix Multiplications with Adjacency Forest

We propose a method for accelerating matrix multiplications that are iteratively performed with a sparse adjacency matrix. We exploit the fact that the intermediate computational results for the matrix multiplication of equivalent partial row vectors of a matrix are the same. Our new data structure, the adjacency forest, uses this property and represents an adjacency matrix as a rooted forest that is made by sharing the common suffixes of the row vectors of the matrix. We confirmed experimentally that our approach can speed up the computation of Personalized PageRank and Non-negative matrix factorization up to 300%.

Masaaki Nishino  
NTT Communication Science Laboratories  
NTT Corporation  
nishino.masaaki@lab.ntt.co.jp

Norihito Yasuda  
Japan Science and Technology Agency

yasuda@erato.ist.hokudai.ac.jp

Shin-Ichi Minato  
Hokkaido University  
minato@ist.hokudai.ac.jp

Masaaki Nagata  
NTT Communication Science Laboratories  
nagata.masaaki@lab.ntt.co.jp

## PP1

### Graphical Models for Identifying Fraud and Waste in Healthcare Claims

We describe graphical model based methods for analyzing prescription and medical claims data in order to identify fraud and waste. Our approach draws on ideas from speech recognition and language modeling to identify patients, doctors and pharmacies whose prescription encounters show significant departure from normative behavior. We have analyzed claims data from a large healthcare provider, consisting of over 53 million individual prescription claims in the calendar year 2011.

Peder A. Olsen

IBM, TJ Watson Research Center  
pederao@us.ibm.com

Ramesh Natarajan  
IBM Thomas J. Watson Research Center  
rnamesh@us.ibm.com

Sholom Weiss  
IBM, TJ Watson Research Cent  
sholom@us.ibm.com

## PP1

### An Optimization-Based Framework to Learn Conditional Random Fields for Multi-Label Classification

This paper studies multi-label classification problem in which data instances are associated with multiple, possibly high-dimensional, label vectors. This problem is especially challenging when labels are dependent and one cannot decompose the problem into a set of independent classification problems. To address the problem and properly represent label dependencies we propose and study a pairwise conditional random field (CRF) model. We develop a new approach for learning the structure and parameters of the CRF from data. The approach maximizes the pseudo likelihood of observed labels and relies on the fast proximal gradient descent for learning the structure and limited memory BFGS for learning the parameters of the model. Empirical results on several datasets show that our approach outperforms several multi-label classification baselines, including recently published state-of-the-art methods.

Mahdi Pakdaman Naeini

Intelligent Systems Program  
University of Pittsburgh  
pakdaman@cs.pitt.edu

Iyad Batal  
General Electric Global Research Center  
iyad@cs.pitt.edu

Zitao Liu

University of Pittsburgh  
ztliu@cs.pitt.edu

CharmGil Hong, Milos Hauskrecht  
Computer Science Department  
University of Pittsburgh  
charmgil@cs.pitt.edu, milos@pitt.edu

## PP1

### Extracting Researcher Metadata with Labeled Features

Due to inherent diversity in values for certain metadata fields on researcher homepages (e.g., affiliation) supervised algorithms require a large number of labeled examples for accurately identifying values for these fields. We address this issue with *feature labeling*, a recent semi-supervised machine learning technique. We apply feature labeling to researcher metadata extraction from homepages by combining a small set of expert-provided feature distributions with few fully-labeled examples. We study “dictionary” and “proximity” labeled features for our task in two stages. We experimentally show that this two-stage approach provides significant improvements in the tagging performance. In one experiment with only ten labeled homepages and 22 expert-specified labeled features, we obtained a 45% relative increase in the F1 value for the affiliation field, while the overall F1 improves by 9%.

Sujatha Das Gollapalli  
The Pennsylvania State University  
gsdas@cse.psu.edu

YanJun Qi  
University of Virginia  
yanjun@virginia.edu

Prasenjit Mitra, Lee Giles  
College of Information Sciences and Technology  
The Pennsylvania State University  
pmitra@ist.psu.edu, giles@ist.psu.edu

## PP1

### Multi-Task Clustering Using Constrained Symmetric Non-Negative Matrix Factorization

A promising new approach to improve clustering quality is to combine data from multiple related datasets (tasks) and apply multi-task clustering. We present a novel framework that can simultaneously cluster multiple tasks through balanced Intra-Task (within-task) and Inter-Task (between-task) knowledge sharing. We propose an effective and flexible geometric affine transformation (contraction or expansion) of the distances between Inter-Task and Intra-Task instances for an improved Intra-Task clustering without overwhelming the individual tasks with the bias accumulated from other tasks. We impose an Intra-Task soft orthogonality constraint to a Symmetric Non-Negative Matrix Factorization (NMF) based formulation to generate basis vectors that are near orthogonal within each task. Inducing orthogonal basis vectors within each task imposes the prior knowledge that a task should have orthogonal (independent) clusters.

Samir Al-Stouhi, Chandan Reddy  
Wayne State University

s.alstouhi@wayne.edu, reddy@cs.wayne.edu

## PP1

### A Weighted Adaptive Mean Shift Clustering Algorithm

The performance of the mean shift algorithm significantly deteriorates with high dimensional data due to the sparsity of the input space. Noisy features can also disguise the mean shift procedure. In this paper we extend the mean shift algorithm to overcome these limitations, while maintaining its desirable properties. To achieve this goal, we first estimate the relevant subspace for each data point, and then embed such information within the mean shift algorithm, thus avoiding computing distances in the full dimensional input space. The resulting approach achieves the best-of-two-worlds: effective management of high dimensional data and noisy features, while preserving a non-parametric nature. Our approach can also be combined with random sampling to speedup the clustering process with large scale data, without sacrificing accuracy. Extensive experimental results on both synthetic and real-world data demonstrate the effectiveness of the proposed method.

Yazhou Ren, Carlotta Domeniconi  
George Mason University  
ryzasia@gmail.com, carlotta@cs.gmu.edu

Guoji Zhang  
South China University of Technology  
magjzh@scut.edu.cn

Guoxian Yu  
Southwest University  
gxyu@swu.edu.cn

## PP1

### Generalized Outlier Detection with Flexible Kernel Density Estimates

We analyse the interplay of density estimation and outlier detection in density-based outlier detection. By clear and principled decoupling of both steps, we formulate a generalization of density-based outlier detection methods based on kernel density estimation. Embedded in a broader framework for outlier detection, the resulting method can be easily adapted to detect novel types of outliers: while common outlier detection methods are designed for detecting objects in sparse areas of the data set, our method can be modified to also detect unusual local concentrations or trends in the data set if desired. It allows for the integration of domain knowledge and specific requirements. We demonstrate the flexible applicability and scalability of the method on large real world data sets.

Arthur Zimek, Erich Schubert  
LMU Munich  
zimek@dbis.lmu.de, schube@dbis.lmu.de

Hans-Peter Kriegel  
Ludwig-Maximilians University Munich  
kriegel@dbis.lmu.de

## PP1

### Membership Detection Using Cooperative Data Mining Algorithms

More and more companies are providing data mining and

analytics solutions to customers using social media data. Unfortunately, given the exponential increase in the volume of social media data, building local database snapshots and running computationally expensive algorithms is not always plausible. As an alternative to the centralized approach, we study the feasibility of cooperative algorithms where data never leaves the mined social media network, and instead the network users themselves work together, using only the communication primitives provided by the social media site, to solve data mining problems. We focus on the task of group membership detection. After validating the potential of cooperative solutions on Twitter, we empirically evaluate a collection of cooperative strategies on a snapshot of the Twitter network containing over 50 million users. Our best solution, brokered token passing, can reliably and efficiently detect group membership.

Lisa Singh, Calvin Newport, Yiqing Ren  
Georgetown University  
singh@cs.georgetown.edu, cnewport@cs.georgetown.edu, yiqingr@cs.georgetown.edu

## PP1

### A Guide to Selecting a Network Similarity Method

We consider the problem of determining how similar two networks are. Many network-similarity methods exist; and it is unclear how one can select a method. We provide the first empirical study on the relationships between different network-similarity methods. We demonstrate that (1) different network-similarity methods are well correlated, (2) some complex network-similarity methods can be closely approximated by a much simpler method, and (3) a few network-similarity methods produce rankings that are very close to the consensus ranking.

Sucheta Soundarajan, Tina Eliassi-Rad  
Department of Computer Science  
Rutgers University  
ssoundarajan@gmail.com, eliasi@cs.rutgers.edu

Brian Gallagher  
Lawrence Livermore National Laboratory  
bgallagher@llnl.gov

## PP1

### Discriminant Analysis for Unsupervised Feature Selection

Feature selection has been proven to be efficient in handling high-dimensional data. As most data is unlabeled, unsupervised feature selection has attracted more and more attention in recent years. Discriminant analysis has been proven to be a powerful technique to select discriminative features. To apply discriminant analysis, we usually need label information which is absent for unlabeled data. This gap makes it challenging to apply discriminant analysis for unsupervised feature selection. In this paper, we investigate how to exploit discriminant analysis in unsupervised scenarios to select discriminative features. We introduce the concept of pseudo labels, which enable discriminant analysis on unlabeled data, propose a novel unsupervised feature selection framework DisUFS which incorporates learning discriminative features with generating pseudo labels, and develop an effective algorithm for DisUFS. Experimental results demonstrate the effectiveness of DisUFS.

Jiliang Tang



Arizona State University  
ARIZONA STATE UNIVERISTY  
Jiliang.Tang@asu.edu

Xia Hu, Huiji Gao, Huan Liu  
Arizona State University  
xia.hu@asu.edu, huiji.gao@asu.edu, huan.liu@asu.edu

## PP1

### Factor Matrix Trace Norm Minimization for Low-Rank Tensor Completion

Most existing low-n-rank minimization algorithms for tensor completion suffer from high computational cost. To address this issue, we propose a novel factor matrix trace norm minimization method. We introduce a tractable relaxation of our rank function, which leads to a convex combination problem of much smaller scale matrix nuclear norm minimization. Then we develop an efficient ADMM scheme to solve the proposed problem. Experimental results on both synthetic and real-world data validate the effectiveness of our approach.

Yuanyuan Liu  
Department of Systems Engineering and Engineering Management  
The Chinese University of Hong Kong  
yyliu@se.cuhk.edu.hk

Fanhua Shang  
Department of Computer Science and Engineering  
The Chinese University of Hong Kong  
fhshang@cse.cuhk.edu.hk

Hong Cheng  
Department of Systems Engineering and Engineering Management  
The Chinese University of Hong Kong  
hcheng@se.cuhk.edu.hk

James Cheng  
The Chinese University of Hong Kong  
jcheng@cse.cuhk.edu.hk

Hanghang Tong  
City College, CUNY  
tong@cs.cuny.cuny.edu

## PP1

### Auto-Play: A Data Mining Approach to Odi Cricket Simulation and Prediction

Cricket is a popular sport played by 16 countries, is the second most watched sport in the world after soccer, and enjoys a multi-million dollar industry. There is tremendous interest in simulating games and predicting their outcome. In this paper, we build a prediction system that takes in historical match data as well as the instantaneous state of a match, and predicts future match events culminating in a victory or loss.

Vignesh Veppur Sankaranarayanan, Junaed Sattar, Laks V.S. Lakshmanan  
University of British Columbia

vsvicky@cs.ubc.ca, junaed@cs.ubc.ca, laks@cs.ubc.ca

## PP1

### Relational Regularization and Feature Ranking

We propose a regularization and feature selection technique for logical and relational learning. To this end, we introduce a notion of locality that ties together features according to their proximity in a transformed representation of the relational learning problem obtained via a procedure that we call “graphicalization”. We present a wrapper and an embedded approach to identify the most relevant sets of predicates which yields more readily interpretable results than selecting low-level propositionalized features.

Mathias Verbeke  
Department of Computer Science  
KU Leuven  
mathias.verbeke@cs.kuleuven.be

Fabrizio Costa  
Institut für Informatik  
Albert-Ludwigs-Universität  
costa@informatik.uni-freiburg.de

Luc De Raedt  
Katholieke Universiteit Leuven  
luc.deraedt@cs.kuleuven.be

## PP1

### Future Influence Ranking of Scientific Literature

A new model called MRFRank is proposed to rank the future popularity of new publications and young researchers. First, words and words co-occurrence are extracted to characterize innovative papers and authors. Then, time-aware weighted graphs are constructed to distinguish the various importance of links. Finally, by leveraging above two features, MRFRank ranks the future importance of papers and authors simultaneously. Experimental results on the ArnetMiner dataset demonstrate the effectiveness of MRFRank.

Senzhang Wang  
Beihang University  
University of Illinois at Chicago  
wangsenzhang@126.com

Sihong Xie  
University of Illinois at Chicago  
sxie6@uic.edu

Xiaoming Zhang  
Beihang University  
yolixs@buaa.edu.cn

Zhoujun Li  
Beihang University  
lizj@buaa.edu.cn

Philip. S Yu  
University of Illinois at Chicago  
psyu@uic.edu

Xinyu Shu  
China Agricultural University

shu921@sina.com

longbing.cao@uts.edu.au

**PP1****Kaleido: Network Traffic Attribution Using Multifaceted Footprinting**

Network traffic attribution, namely, inferring users responsible for activities observed on network interfaces, is one fundamental yet challenging task in security forensics. Compared with other user-system interaction records, network traces are inherently coarse-grained, context-sensitive, and detached from user ends. This paper presents KALEIDO, a new network traffic attribution tool with a series of key features: a) it adopts a new class of inductive discriminant models to capture user- and context-specific patterns (“footprints”) from different aspects of network traffic; b) it applies efficient learning methods to extracting and aggregating such footprints from noisy historical traces; c) with the help of novel indexing structures, it performs efficient, runtime traffic attribution over high-volume network traces. The efficacy of KALEIDO is evaluated using the real network traces collected over three months in a large enterprise network.

Ting WangIBM T.J. Watson Research Center  
tingwang@us.ibm.com

Fei Wang

IBM T J Watson Research Center  
feiwang03@gmail.com

Reiner Sailer, Douglas Schales

IBM Research  
sailer@us.ibm.com, schales@us.ibm.com**PP1****Modeling Asymmetry and Tail Dependence among Multiple Variables by Using Partial Regular Vine**

Modelling high-dimensional dependence is widely studied to explore deep relations in multiple variables particularly useful for financial risk assessment. Very often, strong restrictions are applied on a dependence structure by existing high-dimensional dependence models. These restrictions disabled the detection of sophisticated structures such as asymmetry between multiple variables. The paper proposes a partial regular vine copula model to relax these restrictions. The new model employs partial correlation to construct the regular vine structure, which is algebraically independent. Our method is tested on a cross-country stock market data set to construct and analyse the asymmetry and tail dependence.

Wei WeiAdvanced Analytic Institution, FEIT,  
University of Technology, Sydney  
timajia@gmail.com

Junfu Yin, Jinyan Li

Advanced Analytic Institution, FEIT  
University of Technology, Sydney  
junfu.yin@gmail.com, jinyan.li@uts.edu.au

Longbing Cao

University of Technology, Sydney

**PP1****WCAMiner: A Novel Knowledge Discovery System for Mining Concept Associations Using Wikipedia**

This paper presents WCAMiner, a system focusing on detecting how concepts are associated by incorporating Wikipedia knowledge. We propose to combine content analysis and link analysis techniques over Wikipedia resources, and define various association mining models to interpret such queries. Specifically, our algorithm can automatically build a Concept Association Graph (CAG) from Wikipedia for two given topics of interest, and generate a ranked list of concept chains as potential associations between the two given topics. In comparison to traditional cross-document mining models where documents are usually domain-specific, the system proposed here is capable of handling different query scenarios across domains without being limited to the given documents. We highlight the importance of this problem in various domains, present experiments on different datasets and compare the mining results with two competitive baseline models to demonstrate the improved performance of our system.

Peng Yan, Wei JinNorth Dakota State University  
pengyancdl@gmail.com, wei.jin@ndsu.edu**PP1****Subgraph Search in Large Graphs with Result Diversification**

The problem of subgraph search in large graphs has wide applications in both nature and social science. The subgraph search results are typically ordered based on graph similarity score. In this paper, we study the problem of ranking the subgraph search results based on diversification. We design two ranking measures based on both similarity and diversity, and formalize the problem as an optimization problem. We give two efficient algorithms, the greedy selection and the swapping selection with provable performance guarantee. We also propose a novel local search heuristic with at least 100 times speedup and a similar solution quality. We demonstrate the efficiency and effectiveness of our approaches via extensive experiments.

Huiwen Yu, Dayu YuanDepartment of Computer Science and Engineering  
The Pennsylvania State University  
hwyu@cse.psu.edu, duy113@cse.psu.edu**PP1****To Sample Or To Smash? Estimating Reachability in Large Time-Varying Graphs**

Time-varying graphs (T-graph) consist of a time-evolving set of graph snapshots (or graphlets). A T-graph property with potential applications in both computer and social network forensics is *T-reachability*, which identifies the nodes reachable from a source node using the T-graph edges over time period T. In this paper, we consider the problem of estimating the T-reachable set of a source node in two different settings – when a time-evolution of a T-graph is specified by a probabilistic model, and when the actual T-graph snapshots are known and given to us offline

("data aware" setting).

Feng Yu  
Graduate Center of CUNY  
fyu@gc.cuny.edu

Prithwish Basu  
Raytheon BBN Technologies  
pbasu@bbn.com

Amotz Bar-Noy, Dror Rawitz  
Graduate Center of CUNY  
amotz@sci.brooklyn.cuny.edu, rawitz@gmail.com

## PP1

### Finding Friends on a New Site Using Minimum Information

With the emergence of numerous social media sites, individuals, with their limited time, often face a dilemma of choosing a few sites over others. Users prefer more engaging sites, where they can find familiar faces such as friends, relatives, or colleagues. Link prediction methods help find friends using link or content information. Unfortunately, whenever users join any site, they have no friends or any content generated. In this case, sites have no chance other than recommending random influential users to individuals hoping that users by befriending them create sufficient information for link prediction techniques to recommend meaningful friends. In this study, by considering social forces that form friendships, namely, influence, homophily, and confounding, and by employing minimum information available for users, we demonstrate how one can significantly improve random predictions without link or content information.

Reza Zafarani, Huan Liu  
Arizona State University  
reza@asu.edu, huanliu@asu.edu

## PP1

### Unsupervised Selection of Robust Audio Feature Subsets

We propose an unsupervised, filter-based feature selection approach that preserves the natural assignment of feature components to semantically meaningful features. Experiments on different tasks in the audio domain show that the proposed approach outperforms well-established feature selection methods in terms of retrieval performance and runtime. Results achieved on different audio datasets for the same retrieval task indicate that the proposed method is more robust in selecting consistent feature sets across different datasets than compared approaches.

Maia Zaharieva  
TU Wien, Austria  
zaharieva@ims.tuwien.ac.at

Gerhard Sageder  
University of Vienna, Austria  
gerhard.sageder@univie.ac.at

Matthias Zeppelzauer  
Vienna University of Technology, Austria

mzz@ims.tuwien.ac.at

## PP1

### Towards Breaking the Curse of Dimensionality for High-Dimensional Privacy

We present a general approach to cope the problem of curse of dimensionality in privacy models  $k$ -anonymity,  $\ell$ -diversity, and  $t$ -closeness. In the presence of inter-attribute correlations, such an approach continues to be much more robust in higher dimensionality, without losing accuracy. We present experimental results illustrating the effectiveness of the approach. This approach is resilient enough to prevent identity, attribute, and membership disclosure attack.

Hessam Zakerzadeh  
University of Calgary  
hzakerza@ucalgary.ca

Charu C. Aggarwal  
IBM T. J. Watson Research Center  
charu@us.ibm.com

Ken Barker  
University of Calgary  
kbarker@ucalgary.ca

## PP1

### Towards Accurate Histogram Publication under Differential Privacy

This paper considers the problem of publishing histograms under differential privacy, one of the strongest privacy models. Existing differentially private histogram publication schemes have shown that clustering is a promising idea to improve the accuracy of sanitized histograms. However, none of them fully exploits the benefit of clustering. We introduce a new clustering framework, which features the trade-off between the approximation error due to clustering and the Laplace error due to Laplace noise injected.

Xiaojuan Zhang  
School of Information, Renmin University of China  
xjzhang82@gmail.com

Rui Chen, Jianliang Xu  
Department of Computer Science, Hong Kong Baptist University  
ruichen@comp.hkbu.edu.hk, xujl@comp.hkbu.edu.hk

Xiaofeng Meng  
School of Information, Renmin University of China  
xfmeng@ruc.edu.cn

Yingtao Xie  
Experiment Center, China West Normal University  
yingtaoxie@outlook.com

## PP1

### Addressing Human Subjectivity Via Transfer Learning: An Application to Predicting Disease Outcome in Multiple Sclerosis Patients

Predicting disease course in chronic progressive diseases such as multiple sclerosis (MS) is complicated by the im-

pact of physician subjectivity in the data and because of patients biases in their choice of physician. We introduce a new transfer learning approach to address both issues. For a longitudinal MS dataset, our transfer learning approach performs significantly better than either forming a classifier for the entire dataset or from a single physician's dataset alone.

Yijun Zhao  
Tufts University  
yzhao@cs.tufts.edu

Carla Brodley  
Department of Computer Science, Tufts University  
brodley@cs.tufts.edu

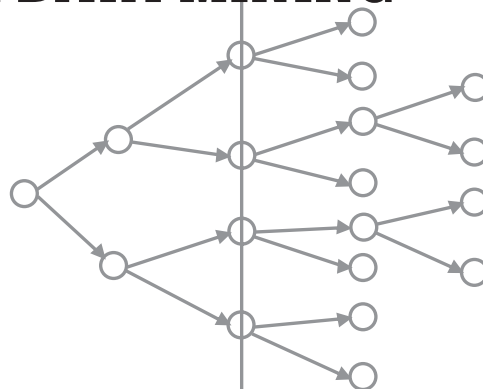
Tanuja Chitnis, Brian Healy  
Harvard Medical School  
Partners MS Center, Brigham and Womens Hospital  
tchitnis@rics.bwh.harvard.edu, bchealy@mgh.harvard.edu

## Notes

## Organizer and Speaker Index

2014 SIAM  
International Conference  
on **DATA MINING**

April 24-26, 2014



Sheraton Society Hill Hotel  
Philadelphia, Pennsylvania, USA

**A**

Acharya, Ayan, CP5, 3:25 Thu  
*Assent, Ira, MS1, 8:30 Sat*  
 Atluri, Gowtham, PP1, 6:45 Thu

**B**

Banerjee, Arindam, MS1, 8:30 Sat  
 Barbieri, Nicola, CP2, 10:25 Thu  
 Bell, Shannon M., PP1, 6:45 Thu  
 Bertens, Roel, PP1, 6:45 Thu  
 Beutel, Alex, CP3, 10:50 Thu  
 Braga, Igor, CP10, 4:15 Fri

**C**

Campello, Ricardo J., MS1, 8:30 Sat  
 Campello, Ricardo J., PP1, 6:45 Thu  
 Cao, Hong, CP1, 10:00 Thu  
 Castillo, Carlos, TS1, 10:00 Thu  
 Chakraborty, Prithwish, CP6, 4:40 Thu  
 Chan, Hau, CP8, 10:25 Fri  
 Chen, XI, PP1, 6:45 Thu  
 Chen, Zheng, CP7, 11:15 Fri  
 Cheng, Yu, CP1, 10:25 Thu  
 Chi, Lianhua, CP3, 10:25 Thu  
 Coric, Vladimir, PP1, 6:45 Thu  
 Correa Bahnsen, Alejandro, PP1, 6:45 Thu

**D**

Danilevsky, Marina, CP9, 11:40 Fri  
 Denton, Anne M., PP1, 6:45 Thu  
 Dey, Sanjoy, PP1, 6:45 Thu  
 Dhanjal, Charanpal, PP1, 6:45 Thu  
 Diao, Qiming, CP9, 11:15 Fri  
 Diaz, Fernando, TS1, 10:00 Thu  
 Dietrich, Brenda L., IP3, 8:15 Fri  
*Domeniconi, Carlotta, MS1, 8:30 Sat*

**E**

Eliassi-Rad, Tina, TS4, 3:00 Fri  
 Etzioni, Oren, IP2, 1:30 Thu

**F**

Faloutsos, Christos, TS4, 3:00 Fri  
 Fan, Wei, PP1, 6:45 Thu  
 Farahat, Ahmed, CP10, 3:50 Fri  
 Feelders, Ad, PP1, 6:45 Thu  
 Fern, Xiaoli Z., MS1, 8:30 Sat  
 Friedland, Lisa, PP1, 6:45 Thu  
 Fu, Yanjie, CP11, 3:50 Fri

**G**

Gao, Yun, PP1, 6:45 Thu  
 Garcia-Soriano, David, CP8, 11:40 Fri  
 Garg, Dinesh, PP1, 6:45 Thu  
 Giatsidis, Christos, PP1, 6:45 Thu  
 Golshan, Behzad, CP4, 3:25 Thu  
 Gong, Wei, PP1, 6:45 Thu  
*Gullo, Francesco, MS1, 8:30 Sat*  
 Gupta, Sunil K., CP6, 3:25 Thu

**H**

Han, Jiawei, CP2, 11:15 Thu  
 Hardt, Moritz, TS2, 3:00 Thu  
 Harenberg, Steve, PP1, 6:45 Thu  
 He, Dan, CP4, 3:50 Thu  
 He, Jingrui, CP5, 3:00 Thu  
 He, Lifang, CP3, 11:40 Thu  
 He, Yunlong, PP1, 6:45 Thu  
 Hoang, Tuan-Anh, CP11, 4:15 Fri  
 Höppner, Frank, PP1, 6:45 Thu  
 Hu, Chuan, PP1, 6:45 Thu  
 Hu, Juhua, CP4, 3:00 Thu  
 Hu, Qingyang, PP1, 6:45 Thu  
 Huggins, Jonathan H., CP12, 3:25 Fri

**J**

Jiang, Yexi, PP1, 6:45 Thu  
 Jin, Rong, TS5, 8:30 Sat

**K**

Kamp, Michael, PP1, 6:45 Thu  
 Kantarcioglu, Murat, PP1, 6:45 Thu  
 Karpadne, Anuj, CP6, 4:15 Thu

Karypis, George, MS1, 8:30 Sat  
 Kolda, Tamara G., IP1, 8:15 Thu  
 Kong, Xiangnan, PP1, 6:45 Thu  
 Koutra, Danai, CP3, 10:00 Thu  
 Koutra, Danai, TS4, 3:00 Fri  
 Kuo, Chia-Tung, PP1, 6:45 Thu  
 Kuo, Tzu-Ming, PP1, 6:45 Thu

**L**

Li, Nan, CP2, 11:40 Thu  
 Li, Sheng, CP4, 4:15 Thu  
 Li, Xiaoyi, CP7, 10:50 Fri  
 Li, Yuxuan, PP1, 6:45 Thu  
 Lines, Jason, CP12, 4:15 Fri  
 Liu, Ruilin, PP1, 6:45 Thu  
 Liu, Ziqi, CP9, 10:50 Fri  
 Lockhart, Jeff, PP1, 6:45 Thu  
 Long, Cheng, PP1, 6:45 Thu

**M**

Madigan, David, IP4, 1:30 Fri  
 Magnano, Chris, PP1, 6:45 Thu  
 Meeng, Marvin, PP1, 6:45 Thu  
 Meng, Lei, PP1, 6:45 Thu  
 Miao, Yun-Qian, PP1, 6:45 Thu  
 Mukherjee, Subhabrata, CP9, 10:25 Fri

**N**

Nikolov, Alexander, TS2, 3:00 Thu  
 Nishino, Masaaki, PP1, 6:45 Thu

**O**

Olsen, Peder A., PP1, 6:45 Thu

**P**

Pakdaman Naeini, Mahdi, PP1, 6:45 Thu  
 Papalexakis, Evangelos, CP3, 11:15 Thu  
 Park, Sunho, CP7, 10:25 Fri  
 Peherstorfer, Benjamin, CP10, 4:40 Fri  
 Peng, Peng, CP1, 11:15 Thu  
 Peng, Peng, CP7, 11:40 Fri  
 Purohit, Hemant, TS1, 10:00 Thu



**Q**

Qi, Yanjun, PP1, 6:45 Thu  
 Qin, Lijing, CP11, 3:25 Fri

**R**

Rahman, Mahmudur, CP6, 3:50 Thu  
 Raykar, Vikas C., CP7, 10:00 Fri  
 Reddy, Chandan, PP1, 6:45 Thu  
 Ren, Yazhou, PP1, 6:45 Thu  
 Riondato, Matteo, CP12, 3:00 Fri

**S**

Schneider, Johannes, CP10, 3:00 Fri  
 Schubert, Erich, PP1, 6:45 Thu  
 Singh, Lisa, PP1, 6:45 Thu  
 Sondhi, Parikshit, CP9, 10:00 Fri  
 Soundarajan, Sucheta, PP1, 6:45 Thu  
 Sugiyama, Mahito, CP5, 3:50 Thu

**T**

*Tagarelli, Andrea, MS1, 8:30 Sat*  
 Tan, Ben, CP5, 4:15 Thu  
 Tang, Jiliang, PP1, 6:45 Thu  
 Tong, Hanghang, PP1, 6:45 Thu

**V**

*Venkatasubramanian, Suresh, TS 1, 10:00 Thu*  
*Venkatasubramanian, Suresh, TS 2, 3:00 Thu*  
*Venkatasubramanian, Suresh, TS 3, 10:00 Fri*  
*Venkatasubramanian, Suresh, TS 4, 3:00 Fri*  
*Venkatasubramanian, Suresh, TS 5, 8:30 Sat*  
 Veppur Sankaranarayanan, Vignesh, PP1, 6:45 Thu  
 Verbeke, Mathias, PP1, 6:45 Thu

**W**

Wang, Senzhang, PP1, 6:45 Thu  
 Wang, Ting, PP1, 6:45 Thu  
 Wang, Xiang, CP10, 3:25 Fri  
 Wei, Wei, PP1, 6:45 Thu  
 Wijaya, Tri Kurniawan, CP6, 3:00 Thu

Wu, Jia, CP5, 4:40 Thu  
 Wu, Ou, CP1, 10:50 Thu  
 Wu, Qingyao, CP2, 10:50 Thu  
 Wu, Zhengwei, CP8, 11:15 Fri

**X**

Xiong, Caiming, CP4, 4:40 Thu  
 Xu, Nana, CP11, 4:40 Fri

**Y**

Yan, Peng, PP1, 6:45 Thu  
 Yang, Lun, TS3, 10:00 Fri  
 Yang, Ning, CP12, 3:50 Fri  
 Yang, Tianbao, TS5, 8:30 Sat  
 Yin, Peifeng, CP8, 10:50 Fri  
 Yu, Feng, PP1, 6:45 Thu  
 Yu, Huiwen, PP1, 6:45 Thu

**Z**

Zafarani, Reza, PP1, 6:45 Thu  
 Zaharieva, Maia, PP1, 6:45 Thu  
 Zakerzadeh, Hessam, PP1, 6:45 Thu  
 Zhang, Chenyi, CP11, 3:00 Fri  
 Zhang, Min-Ling, CP1, 11:40 Thu  
 Zhang, Ping, TS3, 10:00 Fri  
 Zhang, Xiaojian, PP1, 6:45 Thu  
 Zhang, Yao, CP2, 10:00 Thu  
 Zhao, Yijun, PP1, 6:45 Thu  
 Zhong, Erheng, CP8, 10:00 Fri  
 Zhu, Jiaqi, CP12, 4:40 Fri  
 Zhu, Shenghuo, TS5, 8:30 Sat  
*Zimek, Arthur, MS1, 8:30 Sat*

## Notes

## SDM14 Budget

### Conference Budget SIAM Conference on Data Mining April 24-26, 2014 Philadelphia, PA

**Expected Paid Attendance**

235

**Revenue**

Registration Income

	Total	<u>\$96,570</u> \$96,570
--	-------	-----------------------------

**Expenses**

Printing		\$1,200
Organizing Committee		\$2,500
Invited Speakers		\$9,600
Food and Beverage		\$21,900
AV Equipment and Telecommunication		\$12,900
Advertising		\$4,488
Proceedings		\$9,400
Conference Labor (including benefits)		\$35,220
Other (supplies, staff travel, freight, misc.)		\$1,800
Administrative		\$10,931
Accounting/Distribution & Shipping		\$5,367
Information Systems		\$9,598
Customer Service		\$3,555
Marketing		\$5,520
Office Space (Building)		\$3,014
Other SIAM Services		<u>\$3,417</u>
	Total	\$140,410

Net Conference Expense		(\$43,840)
------------------------	--	------------

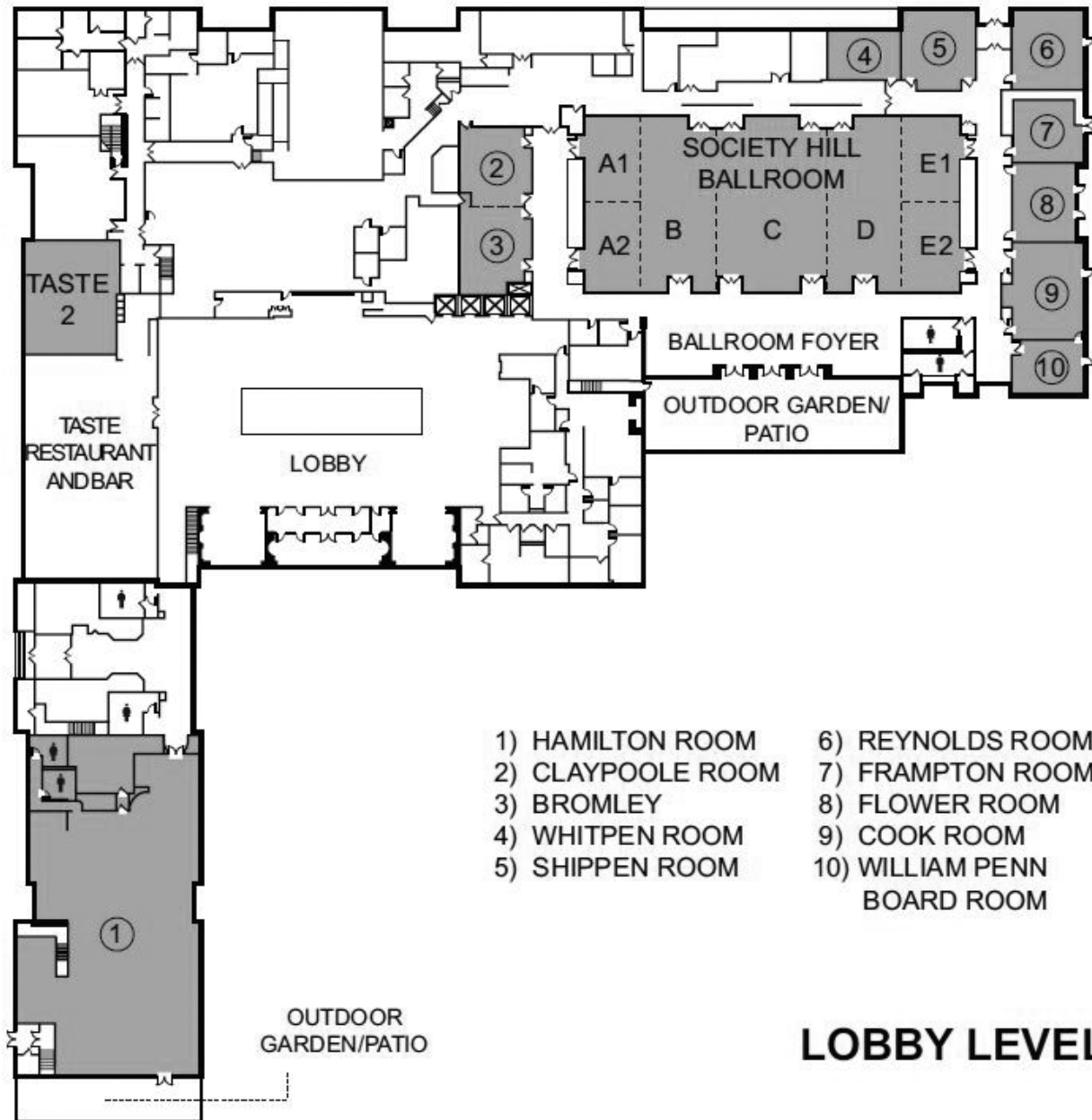
Support Provided by SIAM		<u>\$43,840</u> <u>\$0</u>
--------------------------	--	-------------------------------

**Estimated Support for Travel Awards not included above:**

Early Career	3	\$2,400
--------------	---	---------

# Sheraton Society Hill Hotel Floor Plan

## Philadelphia, Pennsylvania



FSC logo text box indicating size & layout of logo. Conlins to insert logo.