Madness Session
(26th)

CHAIRS: ARISTIDES GIONIS AND JIE TANG
Madness Session (26th)
Chairs: Aristides Gionis and Jie Tang

10:30am - 12:00am
Industry & Govt 4 Education, Privacy and Best Paper Presentations
Chair: Prem Melville
Research 13 Diffusion in Social and Information Networks
Chair: Le Song
Research 14 Recommender Systems
Chair: Shou-De Lin
Research 15 Security and Privacy
Chair: Shengrui Wang
Research 16 Active Learning
Chair: Heng Huang

1:30pm - 3:00pm
Industry & Govt 5 Monitoring & Maintenance
Chair: Balaji Krishnapuram
Research 17 Graph Mining and Modeling
Chair: B. Aditya Prakash
Research 18 Clustering
Chair: Ian Davidson
Research 19 Trend, Anomaly, and Novelty Detection
Chair: Naoki Abe

3:30pm - 5:00pm
Industry & Govt 6 Workforce Analytics & Personalization
Chair: Mohak Shah
Research 20 Social Network Analysis
Chair: Tina Eliassi-Rad
Research 21 Large-scale Optimization and Learning
Chair: Christos Boutsidis
Research 22 Topic Modeling
Chair: Jonathan Chang
Industry & Govt 4
Education, Privacy and Best paper Presentations

CHAIR: PREM MELVILLE

1. Targeting Direct Cash Transfers to the Extremely Poor

2. Style in the Long Tail: Discovering Unique Interests with Latent Variable Models in Large Scale Social E-commerce

3. Predicting Student Risks Through Longitudinal Analysis

Targeting Direct Cash Transfers to the Extremely Poor

Brian Abelson, Kush R. Varshney, and Joy Sun
Style in the Long Tail

Discovering Unique Interests with Latent Variable Models in Large Scale Social E-commerce

Diane Hu, Rob Hall, Josh Attenberg
Etsy
Predicting Student Risks Through Longitudinal Analysis
Ashay Tamhane, Shajith Ikbal, Bikram Sengupta, Mayuri Duggirala, James Appleton

- What ➔ Predicting academic performance risks for K-12 students using historical student's data.
- Why ➔ Early prediction can help in planning better personalized interventions.
- Highlight ➔ The scale of our study – on a large collection of student's data spanning over several years, several schools, students/features from a major US school district (Gwinnett County).

Key findings through 8th grade prediction experiments:

- Possible to predict risks accurately

<table>
<thead>
<tr>
<th>Test</th>
<th>ROC-AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRCT 8th Maths</td>
<td>0.924</td>
</tr>
<tr>
<td>CRCT 8th Science</td>
<td>0.907</td>
</tr>
<tr>
<td>ITBS 8th Maths</td>
<td>0.896</td>
</tr>
</tbody>
</table>

- Possible to predict risks well ahead in time with reasonably good accuracy

- Past scores are important for prediction

<table>
<thead>
<tr>
<th>Features</th>
<th>ROC-AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>all</td>
<td>0.924</td>
</tr>
<tr>
<td>scores</td>
<td>0.902</td>
</tr>
<tr>
<td>demographics</td>
<td>0.866</td>
</tr>
<tr>
<td>behavior</td>
<td>0.576</td>
</tr>
</tbody>
</table>
A Case Study: Privacy Preserving Release of Spatio-temporal Density in Paris

Gergely Acs (INRIA)
Claude Castelluccia (INRIA)

- **Story:** A national project (http://xdata.fr) combines large datasets provided by different service providers (telecom, electricity, postal service, water management, etc.)

- **European Data Protection laws** (Directive 95/46/EC): all datasets have to be anonymized, prior cross-processing, such that data subjects are no longer identifiable.

- **This paper:** How many individuals visited a given area (out of 989) of Paris in any hour?
  - Source: Call-data-records (CDR) of 2 million Orange mobile subscribers in Paris
  - “European” Privacy is addressed within the Differential Privacy model.
Madness Session (26th)
Chairs: Aristides Gionis and Jie Tang

10:30am - 12:00am
- Industry & Govt 4 Education, Privacy and Best Paper Presentations
  Chair: Prem Melville
- Research 13 Diffusion in Social and Information Networks
  Chair: Le Song
- Research 14 Recommender Systems
  Chair: Shou-De Lin
- Research 15 Security and Privacy
  Chair: Shengrui Wang
- Research 16 Active Learning
  Chair: Heng Huang

1:30pm - 3:00pm
- Industry & Govt 5 Monitoring & Maintenance
  Chair: Balaji Krishnapuram
- Research 17 Graph Mining and Modeling
  Chair: B. Aditya Prakash
- Research 18 Clustering
  Chair: Ian Davidson
- Research 19 Trend, Anomaly, and Novelty Detection
  Chair: Naoki Abe

3:30pm - 5:00pm
- Industry & Govt 6 Workforce Analytics & Personalization
  Chair: Mohak Shah
- Research 20 Social Network Analysis
  Chair: Tina Eliassi-Rad
- Research 21 Large-scale Optimization and Learning
  Chair: Christos Boutsidis
- Research 22 Topic Modeling
  Chair: Jonathan Chang
Research 13
Diffusion in Social and Information Networks

CHAIR: LE SONG

1. A Bayesian Framework for Estimating Properties of Network Diffusions
2. Scalable Diffusion-Aware Optimization of Network Topology
3. Probabilistic Latent Network Visualization: Inferring and Embedding Diffusion Networks
4. MMRate: Inferring Multi-aspect Diffusion Networks with Multi-pattern Cascades
5. Stability of Influence Maximization
A Bayesian Framework for Estimating Properties of Network Diffusions
V Embar, RK Pasumarthi, I Bhattacharya (IBM Research India)

Problem: Properties of networks, diffusions & cascades
Example: Users who are well-connected and frequently retweeted

Challenge: Network connections & transmission paths unobserved
Naïve Solution: Plug-in frequentist point estimate of network & path

Our Solution: Compute expectation under posterior distr. of nw & path
Analysis: Characterization of nice, partly nice and not-nice properties
Computation: Analytical when nice, Gibbs Sampling when not nice

Actual | Naïve | Proposed
How to strategically modify networks to optimize susceptibility to cascades?

**New Theoretical Framework for LT**
- New Properties
- Supermodular Objective Functions

**Scalable Approximation Algorithms**
- Scale to Million-node Networks
- Linear Time & Space Complexities

**State-of-the-art Performance**
- Edge Deletion: +10-20%
- Edge Addition: +100%

Which k pairs do I connect to maximize future product adoption?

Which k pairs do I disconnect to minimize future infections?

DELETE EDGE  ADD EDGE
Probabilistic Latent Network Visualization: Inferring and Embedding Diffusion Networks

Inferring the underlying diffusion network by embedding it into a low-dimensional Euclidean space

Cascade data

{node 1, event A, timestamp}
{node 2, event A, timestamp}

... e.g. information propagation, disease infection

Latent coordinates of nodes that best explain the observed cascades

Advantages of the low-dimensional embedding

- Suggesting network layouts most suitable for browsing
- High accuracy in inferring network when analyzing the diffusion process of new or rare information, rumors, and disease

Takeshi Kurashima, Tomoharu Iwata, Noriko Takaya, and Hiroshi Sawada  NTT Corporation
MMRate: Inferring Multi-aspect Diffusion Networks with Multi-pattern Cascades

Senzhang Wang*, Xia Hu+, Philip S. Yu++, Zhoujun Li*
*Beihang University, +Arizona State University, ++University of Illinois at Chicago

In social media:

Cascades: multi-pattern and Relationships: multi-aspect

generate

Cascades: $\mathcal{C}$

$\mathbf{t}^1$ : (user$_1$: t$_{11}$); (user$_2$: t$_{12}$); ... (user$_5$: t$_{15}$)

$\mathbf{t}^2$ : (user$_1$: t$_{21}$); (user$_2$: t$_{22}$); ... (user$_5$: t$_{25}$)

. . .

$\mathbf{t}^{|\mathcal{C}|}$ : (user$_1$: t$_{c1}$); (user$_2$: t$_{c2}$); ... (user$_5$: t$_{c5}$)

infer

Aspect distribution

\[ \Pi = [\pi_1 = 0.7, \pi_2 = 0.3] \]

\[ \alpha_{i,j} \]

$\alpha_{i,5}^1 = 0.4$  $\alpha_{i,5}^2 = 0.8$  $\alpha_{i,4}^1 = 0.6$  $\alpha_{i,4}^2 = 0.3$

$\alpha_{i,3}^1 = 0.2$  $\alpha_{i,3}^2 = 0.4$  $\alpha_{i,4} = 0.4$

$\alpha_{i,5} = 0.3$  $\alpha_{i,5} = 0.5$

$\alpha_{i,3} = 0.4$

$\alpha_{i,4} = 0.1$

$\alpha_{i,4} = 0.4$

$\alpha_{i,4} = 0.4$

$\alpha_{i,4} = 0.5$
Stability of Influence Maximization

Xinran He and David Kempe
University of Southern California

Research 13, Tuesday 10:30AM-12PM, NY Ballroom West

You're not sure which of the three is your input. Will this be a problem?
Madness Session (26th)
Chairs: Aristides Gionis and Jie Tang

10:30am - 12:00am
Industry & Govt 4 Education, Privacy and Best Paper Presentations
Chair: Prem Melville
Research 13 Diffusion in Social and Information Networks
Chair: Le Song
Research 14 Recommender Systems
Chair: Shou-De Lin
Research 15 Security and Privacy
Chair: Shengrui Wang
Research 16 Active Learning
Chair: Heng Huang

1:30pm - 3:00pm
Industry & Govt 5 Monitoring & Maintenance
Chair: Balaji Krishnapuram
Research 17 Graph Mining and Modeling
Chair: B. Aditya Prakash
Research 18 Clustering
Chair: Ian Davidson
Research 19 Trend, Anomaly, and Novelty Detection
Chair: Naoki Abe

3:30pm - 5:00pm
Industry & Govt 6 Workforce Analytics & Personalization
Chair: Mohak Shah
Research 20 Social Network Analysis
Chair: Tina Eliassi-Rad
Research 21 Large-scale Optimization and Learning
Chair: Christos Boutsidis
Research 22 Topic Modeling
Chair: Jonathan Chang
Research 14
Recommender Systems

CHAIR: SHOU-DE LIN

1. COM: a Generative Model for Group Recommendation
2. Leveraging User Libraries to Bootstrap Collaborative Filtering
3. Topic-Factorized Ideal Point Estimation Model for Legislative Voting Network
4. Jointly Modeling Aspects, Ratings and Sentiments for Movie Recommendation (JMARS)
5. User Effort Minimization Through Adaptive Diversification
We propose an LDA-based recommender system “COM” to suggest items for a GROUP of users.

COM outperforms the state-of-the-art methods by more than 24% in recommendation accuracy.

See you in Research Session 14 (Empire East) at 10:30 AM.
Leveraging User Libraries to Bootstrap Collaborative Filtering
Laurent Charlin, Richard Zemel, Hugo Larochelle

Columbia University, Toronto University, Université de Sherbrooke

**Problem:** Can we keep up with hundreds of new papers every year?

**Experiments:** On three real-life datasets yields good performance in all data regimes

**Graphical Model Combines:**

- User library (weak preferences)
- Document Content
- User preferences

**Employed in:** NIPS’13 recommendation system
Topic-Factorized Ideal Point Estimation Model for Legislative Voting Network

Yupeng Gu\(^1\), Yizhou Sun\(^1\), Ning Jiang\(^2\), Bingyu Wang\(^1\), Ting Chen\(^1\)

\(^1\)Northeastern University, \(^2\)UIUC

Input: Legislative Network and Bill Texts

Output: Politicians’ Ideology on Different Topics

Maximizing:

\[
(1 - \lambda) \cdot avelogL(text) + \lambda \cdot avelogL(voting)
\]
Jointly Modeling Aspects, Ratings and Sentiments for Movie Recommendation (JMARS)

Qiming Diao¹, Minghui Qiu¹, Chao-Yuan Wu², Alexander J. Smola²,³, Jing Jiang¹, Chong Wang²

¹SMU, ²CMU, ³Google

Traditional

\[
\begin{align*}
\text{Rating Prediction} & \quad \rightarrow \quad \text{JMARS} \\
\end{align*}
\]

JMARS

\[
\begin{align*}
\text{Better Recommendation} & \quad + \quad \text{User Interests} & \quad + \quad \text{Movie Properties} \\
\end{align*}
\]
Result Diversification has been advocated in the literature as a way of improving the user experience during search.

- Top-k results selected based on their *relevance* to the query and the *diversity* between them

- Previous works balance *relevance* and *diversity* mostly by a predefined fixed way.

We propose a novel framework for *Adaptive Diversification*

**Balances relevance and diversity dynamically**

**Minimizes user navigation effort**
Madness Session (26th)
Chairs: Aristides Gionis and Jie Tang

10:30am - 12:00am
- Industry & Govt 4 Education, Privacy and Best Paper Presentations
  Chair: Prem Melville
- Research 13 Diffusion in Social and Information Networks
  Chair: Le Song
- Research 14 Recommender Systems
  Chair: Shou-De Lin

1:30pm - 3:00pm
- Industry & Govt 5 Monitoring & Maintenance
  Chair: Balaji Krishnapuram
- Research 17 Graph Mining and Modeling
  Chair: B. Aditya Prakash
- Research 18 Clustering
  Chair: Ian Davidson
- Research 19 Trend, Anomaly, and Novelty Detection
  Chair: Naoki Abe

3:30pm - 5:00pm
- Industry & Govt 6 Workforce Analytics & Personalization
  Chair: Mohak Shah
- Research 20 Social Network Analysis
  Chair: Tina Eliassi-Rad
- Research 21 Large-scale Optimization and Learning
  Chair: Christos Boutsidis
- Research 22 Topic Modeling
  Chair: Jonathan Chang
Research 15
Security and Privacy

CHAIR: SHENGRUI WANG

1. Differentially Private Network Data Release via Structural Inference
2. Exponential Random Graph Estimation under Differential Privacy
3. Top-k Frequent Itemsets via Differentially Private FP-trees
4. CatchSync: Catching Synchronized Behavior in Large Directed Graphs
5. Mobile App Recommendations with Security and Privacy Awareness
Differentially Private Network Data Release via Structural Inference

Qian Xiao, Rui Chen, Kian-Lee Tan

Input:
Private network

Output:
Perturbed network to release

- Random walk with MCMC in the entire HRG space
- Sample a good-fitting HRG
- Generate a random graph using the noisy connection probabilities

An HRG sample
Calibrate the sampling distribution to satisfy Differential Privacy
Exponential Random Graph Estimation under Differential Privacy

Network data + Exponential random graph modeling

\[ p(x|\theta) = \frac{\exp(\theta \cdot f(x))}{Z_\theta} \]

Differential privacy requirement

Estimated parameters

Wentian Lu and Gerome Miklau  \{wen,miklau\}@cs.umass.edu
1. Frequent Itemset Discovery

\[
\text{if } \sigma(X) + \text{noise} > \tau + \text{noise} \text{ then } \\
X \text{ is frequent} \\
\text{else} \\
X \text{ is infrequent}
\]

2. Noisy Support Derivation

- Given a set of large itemsets $\mathcal{L}$, find MFIs
- For each MFI $M$, build an FP-tree
Why spikes?

Twitter’09 41M nodes

3.17M

0.41M

d=20

Synchronized!

Follow 20 from

3 million followers

@Buy_AB22
@Buy_BT27
@Buy_BT68
@imc215
@VoteSink
@xAsherzx

1,500

followees

musicians
actors
politicians
financial groups
shopping sites
sport news
e-magazines
...

Smooth & Win!

Weibo’11

Before CatchSync

After CatchSync

Before CatchSync

After CatchSync

Before CatchSync

After CatchSync

CatchSync+SPOT

CatchSync

SPOT

OutRank

Precision

Recall

0.25

0.43

0.62

0.83

0.98

0.5

0.55

0.6

0.65

0.7

0.75

0.8

0.85

0.9

1

1

7

20

878

51

2664

1E+5

7E+6

frequency

out-degree

Security and Privacy Session: Fraud Detection

CatchSync: Catch Synchronized Behavior in Large Directed Graph

Meng Jiang, Peng Cui, Alex Beutel, Christos Faloutsos, Shiqiang Yang
Mobile App Recommendations with Security and Privacy Awareness

Hengshu Zhu, Hui Xiong, Yong Ge, Enhong Chen

Popularity or Security?—That is the question...
Madness Session (26th)

Chairs: Aristides Gionis and Jie Tang

10:30am - 12:00am
- Industry & Govt 4: Education, Privacy and Best Paper Presentations
  Chair: Prem Melville
- Research 13: Diffusion in Social and Information Networks
  Chair: Le Song
- Research 14: Recommender Systems
  Chair: Shou-De Lin
- Research 15: Security and Privacy
  Chair: Shengrui Wang

1:30pm - 3:00pm
- Industry & Govt 5: Monitoring & Maintenance
  Chair: Balaji Krishnapuram
- Research 17: Graph Mining and Modeling
  Chair: B. Aditya Prakash
- Research 18: Clustering
  Chair: Ian Davidson
- Research 19: Trend, Anomaly, and Novelty Detection
  Chair: Naoki Abe

3:30pm - 5:00pm
- Industry & Govt 6: Workforce Analytics & Personalization
  Chair: Mohak Shah
- Research 20: Social Network Analysis
  Chair: Tina Eliassi-Rad
- Research 21: Large-scale Optimization and Learning
  Chair: Christos Boutsidis
- Research 22: Topic Modeling
  Chair: Jonathan Chang
Research 16

Active Learning

CHAIR: HENG HUANG

1. Active-Transductive Learning with Label-Adapted Kernels
2. Active Learning For Sparse Bayesian Multilabel Classification
3. Large-Scale Adaptive Semi-Supervised Learning via Unified Inductive and Transductive Model
4. Active Semi-Supervised Learning Using Sampling Theory for Graph Signals
5. Active Collaborative Permutation Learning
Active-Transductive Learning with Label-Adapted Kernels

Dan Kushnir
Bell Laboratories, Alcatel-Lucent,
dan.kushnir@alcatel-lucent.com

• How to choose the best training set for classification?

• How to do that in linear running time with graphs?

• Get some cool ideas about future work and challenges in active and transductive learning.
Active Learning for Sparse Bayesian Multilabel Classification

- New technique for active sparse multilabel learning based on Mutual Information
- Achieves higher accuracy than the state-of-the-art in few AL rounds
- Is non-myopic, provably near optimal and computationally efficient

Deepak Vasisht (MIT), Andreas Damianou (Sheffield), Manik Varma (MSR), Ashish Kapoor (MSR)
New objective:
\[
\min_{W,b,Y} \left\| X^T W + 1_n b^T - Y_l \right\|_F^2 + \sum_{i=1}^{n} \sum_{k=1}^{c} y_{ik} \left\| x_i^T W + b^T - t_k \right\|_F^2
\]
\[
s.t. \ \forall i, y_{ik} \in [0, 1], \ \sum_{k=1}^{c} y_{ik} = 1
\]

- Computational efficient
- Adaptive and robust to boundary points
- Adaptive optimization procedure
- Only one parameter which serves for multiple purposes

(a) Original toy data  
(b) After classification  
(c) Sample weights
Active Semi-supervised Learning Using Sampling Theory for Graph Signals

**Problem**
- Data points in feature space
- Construct similarity graph
- Choose points to label
- Predict labels for the rest

**Methodology**
Extending Nyquist Shannon sampling theory to signals on graphs

- Sampling rate of $B$ allows perfect reconstruction of signals with bandwidth $B/2$

For given budget, choose a sampling pattern that can represent signals of maximum bandwidth.

**Result**
- $\Psi$-max
- LLR
- LLGC bound
- Proposed

Akshay Gadde, Aamir Anis and Antonio Ortega
University of Southern California
Active Collaborative Permutation Learning

Jialei Wang, Nathan Srebro and James Evans

- Given $m$ objects, $n$ users gave their pairwise preferences, infer $n$ ranking lists based on observations.
- Learn independently: $\mathcal{O}(nm^2)$ observations required for random queries; $\mathcal{O}(nm \log m)$ observations required using active learning strategies.
- Learn collaboratively: only $\mathcal{O}(n + m)$ observations required (based on max-norm regularized matrix completion), active learning further reduces the label cost (based on margin sampling).
- Extensive empirical studies on simulated and real data verify the effectiveness of the proposed approach.
Madness Session (26th)
Chairs: Aristides Gionis and Jie Tang

10:30am - 12:00am
Industry & Govt 4 Education, Privacy and Best Paper Presentations
Chair: Prem Melville
Research 13 Diffusion in Social and Information Networks
Chair: Le Song
Research 14 Recommender Systems
Chair: Shou-De Lin
Research 15 Security and Privacy
Chair: Shengrui Wang
Research 16 Active Learning
Chair: Heng Huang

1:30pm - 3:00pm
Industry & Govt 5 Monitoring & Maintenance
Chair: Balaji Krishnapuram
Research 17 Graph Mining and Modeling
Chair: B. Aditya Prakash
Research 18 Clustering
Chair: Ian Davidson
Research 19 Trend, Anomaly, and Novelty Detection
Chair: Naoki Abe

3:30pm - 5:00pm
Industry & Govt 6 Workforce Analytics & Personalization
Chair: Mohak Shah
Research 20 Social Network Analysis
Chair: Tina Eliassi-Rad
Research 21 Large-scale Optimization and Learning
Chair: Christos Boutsidis
Research 22 Topic Modeling
Chair: Jonathan Chang
Log-based Predictive Maintenance
Applying Data Mining Techniques to Address Critical Process Optimization Needs in Advanced Manufacturing
Unveiling Clusters of Events for Alert and Incident Management in Large-Scale Enterprise IT
Scalable Near Real-Time Failure Localization of Data Center Networks
Correlating Events with Time Series for Incident Diagnosis
Large Scale Predictive Modeling for Micro-Simulation of 3G Air Interface Load
Log-based Predictive Maintenance

Ruben Sipos @ Cornell U., Dmitriy Fradkin @ Siemens, Fabian Moerchen @ Amazon, Zhuang (John) Wang @ Skytree

Come to see the success story of a DEPLOYED log-based predictive maintenance platform. (Tuesday afternoon)

<table>
<thead>
<tr>
<th>Time Stamp</th>
<th>EventCode</th>
<th>Message</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010-09-08 20:10:51</td>
<td>ABC 59</td>
<td>ReconDone(IsvStatusSuccess) after 0 received images for ...</td>
</tr>
<tr>
<td>2010-09-08 20:10:51</td>
<td>ABC 49</td>
<td>Timer was started waiting for ...</td>
</tr>
<tr>
<td>2010-09-08 20:10:52</td>
<td>ABD 46</td>
<td>Receiving IsvMsgReconRequestDone message of size 80 Bytes from IRS.</td>
</tr>
<tr>
<td>2010-09-08 20:10:52</td>
<td>CDE 46</td>
<td>Control info CDE (E c0 03 25 20 a2 00 00)</td>
</tr>
<tr>
<td>2010-09-08 20:10:52</td>
<td>XYZ 10</td>
<td>SsqCtrl: Mode loading initiated: @PatientID@=...</td>
</tr>
<tr>
<td>2010-09-08 20:10:52</td>
<td>FGH 17</td>
<td>hidden</td>
</tr>
<tr>
<td>2010-09-08 20:10:53</td>
<td>CDE 45</td>
<td>IS Imaging System Finished received.</td>
</tr>
</tbody>
</table>
Applying Data Mining Techniques to Address Critical Process Optimization Needs in Advanced Manufacturing

Li Zheng, Chunqiu Zeng, Lei Li, Yexi Jiang, Wei Xue, Jingxuan Li, Chao Shen, Wubai Zhou, Hongtai Li, Liang Tang, Tao Li, Bing Duan, Ming Lei and Pengnian Wang

- Implement and deploy PDP-Miner, an Integrated Solution for Data Analysis in Plasma Display Panel Manufacturing.

- Overall PDP yield rate has increased from 91% to 94%.
- Monthly production is boosted by 10,000 panels, which brings more than 117 million RMB of revenue improvement per year.
Unveiling Clusters of Events for Alert and Incident Management in Large-Scale Enterprise IT

Derek Lin, Rashmi Raghu, Vivek Ramamurthy, Jin Yu, Regunathan Radhakrishnan, Joseph Fernandez

<table>
<thead>
<tr>
<th>time</th>
<th>type</th>
<th>message</th>
</tr>
</thead>
<tbody>
<tr>
<td>8/19/13 14:50</td>
<td>I</td>
<td>Playspan,S IAD,08-19-2013,10:46:55,PROBLEM,Critical,P-DBM-MSQ</td>
</tr>
<tr>
<td>8/19/13 14:54</td>
<td>I</td>
<td>sw551wxy01 :I - REDLINKMAIL SFTPWRAPPER 8/19/2013 2:53:55</td>
</tr>
<tr>
<td>8/19/13 14:54</td>
<td>I</td>
<td>sw551wxy01 :I - REDLINKPAGE SFTPWRAPPER 8/19/2013 2:53:58</td>
</tr>
<tr>
<td>8/19/13 15:31</td>
<td>U</td>
<td>NC - Node not responding to ping or unknown ICMP error</td>
</tr>
<tr>
<td>8/19/13 15:36</td>
<td>U</td>
<td>Error replicat P4APCT11 is 1461 sec(s) behind on CW531GGXY214N3</td>
</tr>
<tr>
<td>8/19/13 15:36</td>
<td>U</td>
<td>Error replicat P4APCT1G is 2346 sec(s) behind on CW750GGSPA04N2</td>
</tr>
<tr>
<td>8/19/13 15:36</td>
<td>U</td>
<td>SCOM: Service Windows Image Acquisition (WIA) has RESTARTED</td>
</tr>
<tr>
<td>8/19/13 15:36</td>
<td>U</td>
<td>Error replicat P4APCT1F is 1835 sec(s) behind on CW531GGXY214N3</td>
</tr>
<tr>
<td>8/19/13 15:36</td>
<td>U</td>
<td>Error replicat P4APCT1B is 1834 sec(s) behind on CW531GGXY214N3</td>
</tr>
<tr>
<td>8/19/13 15:36</td>
<td>U</td>
<td>SCOM: Machine: SW879AA8BCXI04 Severity:ERROR User: pat44e.mr</td>
</tr>
<tr>
<td>8/19/13 15:36</td>
<td>U</td>
<td>Service Tripwire Enterprise Agent is not RUNNING</td>
</tr>
<tr>
<td>8/19/13 15:37</td>
<td>U</td>
<td>Error replicat P4APCT1C is 2164 sec(s) behind on CW531GGXY214N3</td>
</tr>
<tr>
<td>8/19/13 15:37</td>
<td>U</td>
<td>Error replicat P4APCT1A is 251734 sec(s) behind on CW531GGXY214N3</td>
</tr>
<tr>
<td>8/19/13 15:37</td>
<td>U</td>
<td>Splunk Alert: JVM PID file /JPG:tcServer-6.0/tomcat2/logs/tcservlet.p</td>
</tr>
<tr>
<td>8/19/13 15:37</td>
<td>U</td>
<td>SCOM: Service Windows Image Acquisition (WIA) has RESTARTED</td>
</tr>
<tr>
<td>8/19/13 15:37</td>
<td>U</td>
<td>Error replicat P4APCT1F is 1840 sec(s) behind on CW531GGXY214N3</td>
</tr>
<tr>
<td>8/19/13 15:37</td>
<td>U</td>
<td>rs551mbmnn45: I - This is to notify you that an error was reported for</td>
</tr>
<tr>
<td>8/19/13 15:37</td>
<td>U</td>
<td>I - Alert Threshold exceeded - MMG === E:\Visa.IMG\WebServer\Log</td>
</tr>
<tr>
<td>8/19/13 15:37</td>
<td>U</td>
<td>AIX HARDWARE ERROR: ATTACHED SCSI TARGET DEVICE ERROR, ADA</td>
</tr>
<tr>
<td>8/19/13 15:38</td>
<td>I</td>
<td>The chassis power supply for st93zbbyhx1 is in a failure condition</td>
</tr>
<tr>
<td>8/19/13 15:38</td>
<td>I</td>
<td>A chassis fan for st93zbbyhx1 is in a failure condition</td>
</tr>
</tbody>
</table>
Scalable Near Real-Time Failure Localization of Data Center Networks

Herodotos Herodotou\textsuperscript{1}, Bolin Ding\textsuperscript{1}, Shobana Balakrishnan\textsuperscript{1}, Geoff Outhred\textsuperscript{2}, Percy Fitter\textsuperscript{2}

\textsuperscript{1}Microsoft Research and \textsuperscript{2}Windows Azure, Microsoft

**Goal:** Localize \textit{user-perceived} network failures in the \textit{entire} Windows Azure network in near \textit{real time}

### Input Data

**Network Topology**

```plaintext
N1 -- N2
N2 -- N3
N3 -- N4
```

**Canary Ping Data**

- Periodical pings
  - Source Vertex | Destination Vertex | Ping Successes | Ping Failures |
  - v1           | e10               | 45             | 5            |
  - v1           | e10               | 38             | 2            |
  - v2           | e10               | 46             | 0            |
  - v2           | e10               | 37             | 3            |

**Network Alerts**

- Generated from devices or links
- Hundreds to thousands per minute
- Often noisy

### Step 1: Probability Routing

**Challenge**

- Unknown ping routes

**Approach**

- Model using shortest paths
- Compute routes & probabilities

### Step 2: Failure Modeling

**Challenge**

- Partial network failures

**Approach**

- Model failure scores as probabilities
- Formulate as a data fitting problem

### Step 3: Deviation Detection

**Challenge**

- Noise in input data & failure scores

**Approach**

- Use statistical hypothesis testing to detect significant deviations

### Step 4: Correlation Analysis

**Challenge**

- Indistinguishable links/devices
  - Both aggregation switches will have equal failure scores

**Approach**

- Intersect links/devices with top failure scores with network alerts

### Failure Localization Result

- Generate ranked list of suspect links/devices
- Overlay on top of topology in web report
- Mean time to localize: 12 minutes

### Experimental Evaluation

#### Localization Accuracy

- Based on 73 real network incidents

#### Localization Precision

- Precision = \( \frac{\text{#Links/Devices After Localization}}{\text{#Links/Devices Before Localization}} \)

- Ruling from Network Alerts
- Ruling from Localization

#### Number of Network Issues

- TP = True Positive
- TN = True Negative
- FP = False Positive
- FN = False Negative
- LP = Likely Positive
- LN = Likely Negative
- NA = Non Applicable

#### Mean Time to Localize

- 12 minutes
Correlating Events with Time Series for Incident Diagnosis
Chen Luo\textsuperscript{1}, Jian-Guang Lou\textsuperscript{2}, Qingwei Lin\textsuperscript{2}, Qiang Fu\textsuperscript{2}, Rui Ding\textsuperscript{2}, Dongmei Zhang\textsuperscript{2}, and Zhe Wang\textsuperscript{1}
\textsuperscript{1}Jilin University, China \textsuperscript{2}Microsoft Research

• Correlation is a very important tool for incident diagnosis.

• Three questions in incident diagnosis
  • Are they correlated?
  • Temporal relationship?
  • Monotonic effect?

• A two-sample based method is proposed to answer the questions

• Existing tools cannot work well
Large Scale Predictive Modeling for Micro-Simulation of 3G Air Interface Load

AKA

How to Make Mobile Network Analysts Who (Want to) Know Nothing About Data Mining Use It To Run Their Own Simulations of Mobile Network Service Quality Across Thousands of Cells And Trust The Results

Dejan Radosavljevik – T-Mobile NL & Leiden University
Peter van der Putten – Leiden University
# Madness Session (26th)

**Chairs:** Aristides Gionis and Jie Tang

<table>
<thead>
<tr>
<th>Time</th>
<th>Session Title</th>
<th>Chair</th>
</tr>
</thead>
<tbody>
<tr>
<td>10:30am - 12:00am</td>
<td>Industry &amp; Govt 4 Education, Privacy and Best Paper Presentations</td>
<td>Prem Melville</td>
</tr>
<tr>
<td></td>
<td>Research 13 Diffusion in Social and Information Networks</td>
<td>Le Song</td>
</tr>
<tr>
<td></td>
<td>Research 14 Recommender Systems</td>
<td>Shou-De Lin</td>
</tr>
<tr>
<td></td>
<td>Research 15 Security and Privacy</td>
<td>Shengrui Wang</td>
</tr>
<tr>
<td></td>
<td>Research 16 Active Learning</td>
<td>Heng Huang</td>
</tr>
<tr>
<td>1:30pm - 3:00pm</td>
<td>Industry &amp; Govt 5 Monitoring &amp; Maintenance</td>
<td>Balaji Krishnapuram</td>
</tr>
<tr>
<td></td>
<td>Research 17 Graph Mining and Modeling</td>
<td>B. Aditya Prakash</td>
</tr>
<tr>
<td></td>
<td>Research 18 Clustering</td>
<td>Ian Davidson</td>
</tr>
<tr>
<td></td>
<td>Research 19 Trend, Anomaly, and Novelty Detection</td>
<td>Naoki Abe</td>
</tr>
<tr>
<td>3:30pm - 5:00pm</td>
<td>Industry &amp; Govt 6 Workforce Analytics &amp; Personalization</td>
<td>Mohak Shah</td>
</tr>
<tr>
<td></td>
<td>Research 20 Social Network Analysis</td>
<td>Tina Eliassi-Rad</td>
</tr>
<tr>
<td></td>
<td>Research 21 Large-scale Optimization and Learning</td>
<td>Christos Boutsidis</td>
</tr>
<tr>
<td></td>
<td>Research 22 Topic Modeling</td>
<td>Jonathan Chang</td>
</tr>
</tbody>
</table>
Research 17
Graph Mining and Modeling

**Chair:** B. Aditya Prakash

1. Core Decomposition of Uncertain Graphs
2. Learning Multifractal Structure in Large Networks
3. Temporal Skeletonization on Sequential Data: Patterns, Categorization, and Visualization
4. Focused Clustering and Outlier Detection in Large Attributed Graphs
5. Inside the Atoms: Ranking on a Network of Networks
Core Decomposition of Uncertain Graphs
Francesco Bonchi\textsuperscript{1}, Francesco Gullo\textsuperscript{1}, Andreas Kaltenbrunner\textsuperscript{2} and Yana Volkovich\textsuperscript{2}
\textsuperscript{(1)Yahoo Labs, \textsuperscript{2}Barcelona Media, Spain}

We extend the graph tool of core decomposition to the context of uncertain graphs

- **uncertain graph** is a graph whose edges are assigned a probability of existence
- **k-core** of a graph is defined as a maximal subgraph in which every vertex is connected to at least $k$ other vertices within that subgraph
- **$\eta$-degree** of a vertex is the maximum degree such that the probability for this vertex to have that degree is no less than $\eta$

We define the $(k, \eta)$-core concept and devise efficient algorithms

**Applications:**
- influence-maximization problem
- task-driven team formation problem
Learning multifractal structure in large networks

Austin Benson, Carlos Riquelme, Sven Schmit  Stanford University
Graph Mining and Modeling Tuesday 1:30PM-3:00PM NY Ballroom West

- Analyze and improve the multifractal network generators (MFNG) introduced by Palla et al. and apply them to social networks.
- Show how to quickly count graph features in MFNG → Use this theory to model large networks.
- Effectively simulate important graph properties of social and information networks.
Temporal Skeletonization on Sequential Data
1:30PM-3:00PM | NY Ballroom West
Chuanren Liu, Kai Zhang, Hui Xiong, Geoff Jiang, Qiang Yang
Focused Clustering and Outlier Detection in Large Attributed Graphs

Bryan Perozzi, Leman Akoglu
Stony Brook University

Patricia Iglesias Sánchez*, Emmanuel Müller†
*Karlsruhe Institute of Technology
†University of Antwerp

Our Contributions:
- new user-centric problem
- search steered by user preference
- local scalable algorithm
- analysis on synthetic & real graphs

The Problem:
Given user input & graph
Infer attribute weights
Extract focused clusters
Detect focused outliers
Inside the Atoms: Ranking on a Network of Networks

Jingchao Ni¹, Hanghang Tong², Wei Fan³, Xiang Zhang¹

¹Department of Electrical Engineering and Computer Science, Case Western Reserve University
²School of Computing, Informatics, Decision Systems Engineering, Arizona State University
³Huawei Noahs Ark Lab

---

**New data Model**
Network of Networks (NoN)

**CrossRank**
A new ranking algorithm on NoN

**CrossQuery**
A fast query algorithm on NoN

---

Location network of social networks

Disease network of protein interaction networks

---

**Research 17: Graph Mining and Modeling**
Tuesday 1:30PM-3:00PM, NY Ballroom West
Madness Session (26th)
Chairs: Aristides Gionis and Jie Tang

10:30am - 12:00am
- Industry & Govt 4
  Education, Privacy and Best Paper Presentations
  Chair: Prem Melville
- Research 13
  Diffusion in Social and Information Networks
  Chair: Le Song
- Research 14
  Recommender Systems
  Chair: Shou-De Lin
- Research 15
  Security and Privacy
  Chair: Shengrui Wang
- Research 16
  Active Learning
  Chair: Heng Huang

1:30pm - 3:00pm
- Industry & Govt 5
  Monitoring & Maintenance
  Chair: Balaji Krishnapuram
- Research 17
  Graph Mining and Modeling
  Chair: B. Aditya Prakash
- Research 18
  Clustering
  Chair: Ian Davidson
- Research 19
  Trend, Anomaly, and Novelty Detection
  Chair: Naoki Abe

3:30pm - 5:00pm
- Industry & Govt 6
  Workforce Analytics & Personalization
  Chair: Mohak Shah
- Research 20
  Social Network Analysis
  Chair: Tina Eliassi-Rad
- Research 21
  Large-scale Optimization and Learning
  Chair: Christos Boutsidis
- Research 22
  Topic Modeling
  Chair: Jonathan Chang
Research 18
Clustering

1. Relevant Overlapping Subspace Clusters on Categorical Data
2. Batch Discovery of Recurring Rare Classes toward Identifying Anomalous Samples
3. A Dirichlet Multinomial Mixture Model-based Approach for Short Text Clustering
4. Representative Clustering of Uncertain Data
5. SMVC: Semi-Supervised Multi-View Clustering in Subspace Projections
Relevant Overlapping Subspace Clusters on Categorical Data

1. Overlapping
2. Non-Redundant
3. No Enumeration
4. Parameter-free
5. Outliers

Xiao He, Jing Feng, Bettina Konte, Son T.Mai, Claudia Plant
We present a clustering algorithm for discovering rare classes across a batch of samples in the presence of random effects.

A normal sample is a composition of data points each originating from a known class. An anomalous sample contain data points originating from classes not known beforehand.

Our model uses a hierarchical Dirichlet process prior to jointly perform clustering and cluster matching across samples. Each sample data is modeled by an infinite mixture of infinite Gaussian mixture models, which allows modeling of skewed and multi-modal clusters.

We demonstrate the utility of the proposed algorithm, processing a flow cytometry data set containing two extremely rare cell populations, and report results that significantly outperform competing techniques.
In this paper, we proposed an approach for short text clustering with the following nice properties:

- It can infer the number of clusters automatically;
- It has a clear way to balance the completeness and homogeneity of the clustering results;
- Its time complexity is linear to the number of documents, and it is fast to converge;
- It can obtain the representative words of each cluster.
Uncertain Spatial Data

- How to perform clustering?
- What should result look like?

Our Approach:

- Copes with ambiguity
- Provides confidence
- In accordance with Possible World Semantics

Tuesday 26.8.2014
Research 18, Clustering (Empire East)
SMVC: Semi-Supervised Multi-View Clustering in Subspace Projections

S. Günnemann, I. Färber, M. Rüdiger, T. Seidl

Challenge: Learn association of constraints to views

Learned through Bayesian inference

Challenge: Learn association of dimensions to views

instance-level user constraints

multiple clustering views in subspace projections
Madness Session (26th)

Chairs: Aristides Gionis and Jie Tang

10:30am - 12:00am

- Industry & Govt 4 Education, Privacy and Best Paper Presentations
  Chair: Prem Melville

- Research 13 Diffusion in Social and Information Networks
  Chair: Le Song

- Research 14 Recommender Systems
  Chair: Shou-De Lin

- Research 15 Security and Privacy
  Chair: Shengrui Wang

- Research 16 Active Learning
  Chair: Heng Huang

1:30pm - 3:00pm

- Industry & Govt 5 Monitoring & Maintenance
  Chair: Balaji Krishnapuram

- Research 17 Graph Mining and Modeling
  Chair: B. Aditya Prakash

- Research 18 Clustering
  Chair: Ian Davidson

- Research 19 Trend, Anomaly, and Novelty Detection
  Chair: Naoki Abe

3:30pm - 5:00pm

- Industry & Govt 6 Workforce Analytics & Personalization
  Chair: Mohak Shah

- Research 20 Social Network Analysis
  Chair: Tina Eliassi-Rad

- Research 21 Large-scale Optimization and Learning
  Chair: Christos Boutsidis

- Research 22 Topic Modeling
  Chair: Jonathan Chang
Research 19

Trend, Anomaly, and Novelty Detection

**Chair:** Naoki Abe

1. Sleep Analytics and Online Selective Anomaly Detection

2. GLAD: Group Anomaly Detection in Social Media Analysis

3. FBLG: A Simple and Effective Approach for Temporal Dependence Discovery from Time Series Data

4. Learning Time-Series Shapelets

5. Utilizing Temporal Patterns for Estimating Uncertainty in Interpretable Early Decision Making
We spend nearly one-third of our lives sleeping!

GOOD SLEEP $\rightarrow$ ALERTNESS $\rightarrow$ PRODUCTIVITY

EEG can be used to study sleep patterns

Data Mining (DM) $\rightarrow$ Control Theory

See you @ 1:30 PM, Empire West!
GLAD: Group Anomaly Detection in Social Media Analysis

Rose Yu, Xinran He and Yan Liu
University of Southern California

Research 19, Tuesday 1:30PM-3:00PM, Empire West

- Anomalies in social media data may not only appear as individual points, but also as groups. E.g. Group review spam, malicious social collateral.
- GLAD takes both pair-wise and point-wise data as input, infers the groups and detects group anomalies simultaneously.
- Experiments: detects “anomalous” paper groups from the scientific publication corpus and uncovers the party affiliation changes in the Senate.

\[
\pi_p \rightarrow G_p \rightarrow R_p \\
\beta \rightarrow X_p
\]

\[
\theta \rightarrow M
\]

\[
\alpha
\]

\[
\alpha, \beta, \theta, B, X_p, N^2, N, N^2, \beta
\]

Diagram:

- Ben Nelson
- James Jeffords
FBLG: A Simple and Effective Approach for Temporal Dependence Discovery from Time Series Data

Dehua Cheng, Mohammad Taha Bahadori, and Yan Liu
University of Southern California

Research 19, Tuesday 1:30PM-3:00PM, Empire West
“Simple, efficient and effective: a new algorithm for temporal dependence recovery"

- We create a “new” time series by reversing the time stamps of original time series and combine the results from both time series to improve the performance.
- Theoretical explanation and empirical evaluation on real SNS time series are provided.
Learning Time-Series **Shapelets**
Josif Grabocka, Nicolas Schilling, Martin Wistuba, Lars Schmidt-Thieme (josif@ismll.de)

- **Shapelets** are discriminative sub-sequences of time series

- This paper **learns** the discriminative shapelets
Utilizing Temporal Patterns for Estimating Uncertainty in Interpretable Early Decision Making
Mohamed Ghalwash, Vladan Radosavljevic, and Zoran Obradovic
Tuesday 1:30PM-3:00PM | Empire West

**PROBLEM:** The black and red time series are very similar at the beginning.

![Time Series Comparison](image)

How to provide early, interpretable and uncertainty estimate for temporal classification?

**RESULTS:**
1) The red example is confidently classified correctly and early.

2) Our method (blue) outperforms a non-interpretable alternative method (red) in more than 15 out of 20 public time series datasets.
Madness Session (26th)
Chairs: Aristides Gionis and Jie Tang

10:30am - 12:00am
- Industry & Govt 4 Education, Privacy and Best Paper Presentations
  Chair: Prem Melville
- Research 13 Diffusion in Social and Information Networks
  Chair: Le Song
- Research 14 Recommender Systems
  Chair: Shou-De Lin
- Research 15 Security and Privacy
  Chair: Shengrui Wang
- Research 16 Active Learning
  Chair: Heng Huang

1:30pm - 3:00pm
- Industry & Govt 5 Monitoring & Maintenance
  Chair: Balaji Krishnapuram
- Research 17 Graph Mining and Modeling
  Chair: B. Aditya Prakash
- Research 18 Clustering
  Chair: Ian Davidson
- Research 19 Trend, Anomaly, and Novelty Detection
  Chair: Naoki Abe

3:30pm - 5:00pm
- Industry & Govt 6 Workforce Analytics & Personalization
  Chair: Mohak Shah
- Research 20 Social Network Analysis
  Chair: Tina Eliassi-Rad
- Research 21 Large-scale Optimization and Learning
  Chair: Christos Boutsidis
- Research 22 Topic Modeling
  Chair: Jonathan Chang
Industry & Govt 6

Workforce Analytics & Personalization

Chair: Mohak Shah

1. Predicting Employee Expertise for Talent Management in the Enterprise
2. Modeling Professional Similarity by Mining Professional Career Trajectories
3. A System to Grade Computer Programming Skills using Machine Learning
4. Large Scale Visual Recommendations From Street Fashion Images
5. We Know What You Want to Buy: A Demographic-based System for Product Recommendation On Microblogs
6. Modeling Impression Discounting in Large-scale Recommender Systems
Predicting Employee Expertise for Talent Management in the Enterprise

The Waves of Business Analytics

<table>
<thead>
<tr>
<th>The Industrial Economy</th>
<th>The Financial Economy</th>
<th>The Customer Economy and Web</th>
<th>The Talent Economy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steel, Oil, Railroads</td>
<td>Conglomerates</td>
<td>Customer Segmentation</td>
<td>Globalization, Demographics</td>
</tr>
<tr>
<td></td>
<td>Financial Engineering</td>
<td>Personalized Products</td>
<td>Skills and Leadership</td>
</tr>
<tr>
<td>Early 1900s</td>
<td>1950s-60s</td>
<td>1970s-80s</td>
<td>Shortages</td>
</tr>
</tbody>
</table>

Kush R. Varshney, Vijil Chenthamarakshan, Scott W. Fancher, Jun Wang, Dongping Fang, and Aleksandra Mojsilović
Modeling Professional Similarity by Mining Professional Career Trajectories
Ye Xu, Zang Li, Abhishek Gupta, Ahmet Bugdayci, Anmol Bhasin

Session:
- Tuesday
- 3:30PM-5:00PM
- NY Ballroom East

Discover LinkedIn profiles
By measuring career trajectory similarity
A System to Grade Computer Programming Skills using Machine Learning

What’s wrong with the current approach?

- Test cases only to evaluate programs? → Does not mimic experts ❌

What are good features?

- Only bag of words to predict correctness? → Is not accurate enough ❌
- Use the control and data dependency graphs to extract features which capture intent and logic

Supervised learning using these features gives great results!

Shashank Srikant
Varun Aggarwal

3:45 PM
NY Ballroom East
Large Scale Visual Recommendations from Street Fashion Images

Vignesh Jagadeesh, Robinson Piramuthu, Anurag Bhardwaj, Wei Di, Neel Sundaresan

- Clothing recommender
- Mobile Use Case

Query Skirt

Query Skirt

Coordinating Tops

- Recommend clothing that co-ordinate with query using image information alone
- Predictive models learn visual patterns that co-ordinate well from large scale data
- Crowdsourcing to validate retrievals

Contextual recommendations delivered in real time using data driven models learnt from the web

eBay Inc.
We Know What You Want to Buy: A Demographic-based System for Product Recommendation On Microblogs

Xin Wayne Zhao, Renmin University of China;
Yanwei Guo, Peking University;
Yulan He, Aston University;
Han Jiang, Peking University;
Yuexin Wu, wuyuexin@gmail.com;
Xiaoming Li, Peking University;
Modeling Impression Discounting In Large-scale Recommender Systems

Pei Lee, Laks V.S. Lakshmanan
University of British Columbia

Mitul Tiwari, Sam Shah
LinkedIn

- Impressions: recommendations shown to user
- Impression Discounting Model
  - Learn a discounting function to capture ignored impressions
  - Linear or multiplicative aggregation model
  - Anti-noise regression model
- Offline and Online evaluation (Bucket testing)
Madness Session (26th)

Chairs: Aristides Gionis and Jie Tang

10:30am - 12:00am
- Industry & Govt 4: Education, Privacy, and Best Paper Presentations
  Chair: Prem Melville
- Research 13: Diffusion in Social and Information Networks
  Chair: Le Song
- Research 14: Recommender Systems
  Chair: Shou-De Lin
- Research 15: Security and Privacy
  Chair: Shengrui Wang
- Research 16: Active Learning
  Chair: Heng Huang

1:30pm - 3:00pm
- Industry & Govt 5: Monitoring & Maintenance
  Chair: Balaji Krishnapuram
- Research 17: Graph Mining and Modeling
  Chair: B. Aditya Prakash
- Research 18: Clustering
  Chair: Ian Davidson
- Research 19: Trend, Anomaly, and Novelty Detection
  Chair: Naoki Abe

3:30pm - 5:00pm
- Industry & Govt 6: Workforce Analytics & Personalization
  Chair: Mohak Shah
- Research 20: Social Network Analysis
  Chair: Tina Eliassi-Rad
- Research 21: Large-scale Optimization and Learning
  Chair: Christos Boutsidis
- Research 22: Topic Modeling
  Chair: Jonathan Chang
Research 20
Social Network Analysis

**CHAIR: TINA ELIASSI-RAD**

1. Using Strong Triadic Closure to Characterize Ties in Social Networks
2. Network Structural Analysis via Core-Tree-Decomposition
3. Analyzing Expert Behaviors in Collaborative Networks
4. Predicting Long-Term Impact of CQA Posts: A Comprehensive Viewpoint
5. Who Are Experts Specializing in Landscape Photography?
Who are true friends in Online Social Networks?

Use the graph structure: Label social ties as Strong or Weak by enforcing the Strong Triadic Closure property
Network Structural Analysis via Core-Tree-Decomposition

@Research 20: Social Network Analysis (15:30-)

Takuya Akiba, Takanori Maehara, Ken-ichi Kawarabayashi

- Core/Fringe analysis with Tree Decomposition
- 10%-20% vertices of real-world networks are in core
- relation of algorithm performance and core-size
Analyzing Expert Behaviors in Collaborative Networks

Collaborative network - information routing

- How do experts make routing decisions?
- Who have made inefficient routing decisions?
- How to optimize the routing performance through targeted training?
- Can the completion time of a task be predicted so that one can act early for difficult tasks?

e.g.,
1. ticket processing in customer service
2. bug solving in software development

Huan Sun, Mudhakar Srivatsa, Shulong Tan, Yang Li, Lance Kaplan, Shu Tao, Xifeng Yan

IBM

UCSB
What is `&&&` operation in C

```c
#include <stdio.h>
int main(void)
{
    int i, c;
    for (i = 0; i < 3; i++) {
        c = i &&& i;
        printf("%d\n", c);
    }
    return 0;
}
```

The output of the above program compiled using `gcc` is

```
0
1
1
```

Q: How many users will find it beneficial?

Predicting Long-Term Impact of CQA Posts: A Comprehensive Viewpoint.
Authors: Yuan Yao, Hanghang Tong, Feng Xu, Jian Lu
Who Are Experts Specializing in Landscape Photography?

Analyzing Topic-specific Authority on Content Sharing Services

- Identify topic-specific authorities who generated high-quality content

Leverage data of Like clicks

Challenge: Users didn’t specify topical causes under Like clicks

Bin Bi, Ben Kao, Chang Wan, Junghoo Cho
Madness Session (26th)
Chairs: Aristides Gionis and Jie Tang

10:30am - 12:00am
Industry & Govt 4 Education, Privacy and Best Paper Presentations
Chair: Prem Melville
Research 13 Diffusion in Social and Information Networks
Chair: Le Song
Research 14 Recommender Systems
Chair: Shou-De Lin
Research 15 Security and Privacy
Chair: Shengrui Wang
Research 16 Active Learning
Chair: Heng Huang

1:30pm - 3:00pm
Industry & Govt 5 Monitoring & Maintenance
Chair: Balaji Krishnapuram
Research 17 Graph Mining and Modeling
Chair: B. Aditya Prakash
Research 18 Clustering
Chair: Ian Davidson
Research 19 Trend, Anomaly, and Novelty Detection
Chair: Naoki Abe

3:30pm - 5:00pm
Industry & Govt 6 Workforce Analytics & Personalization
Chair: Mohak Shah
Research 20 Social Network Analysis
Chair: Tina Eliassi-Rad
Research 21 Large-scale Optimization and Learning
Chair: Christos Boutsidis
Research 22 Topic Modeling
Chair: Jonathan Chang
Research 21
Large-scale Optimization and Learning

Chair: Christos Boutsidis

1. Efficient Mini-batch Training for Stochastic Optimization
2. Streaming Submodular Maximization: Massive Data Summarization on the Fly
3. Distance Queries from Sampled Data: Accurate and Efficient
4. Improved Testing of Low Rank Matrices
5. DeepWalk: Online Learning of Social Representations
Efficient Mini-batch Training for Stochastic Optimization

Mu Li, Tong Zhang, Yuqiang Chen, and Alex Smola

Motivation
- Mini-batch SGD is widely used for large scale optimization
- However

Our Solution
- For each mini-batch, solve a better optimization problem

Experiment
- 12 machines, logistic regression
- fix run time:

![Graph showing the relationship between minibatch size and system performance, convergence rate, and objective.](image-url)
Streaming Submodular Maximization: Massive Data Summarization on the Fly!

Data Stream

\[ F \left\{ \begin{array}{c}
\text{\includegraphics[width=0.1\textwidth]{image1}} \\
\text{\includegraphics[width=0.1\textwidth]{image2}} \\
\text{\includegraphics[width=0.1\textwidth]{image3}} \\
\text{\includegraphics[width=0.1\textwidth]{image4}} \\
\end{array} \right\} \rightarrow \text{max} \]

Ashwinkumar Badanidiyuru, Baharan Mirzasoleiman, Amin Karbasi & Andreas Krause
Distance Queries from Sampled Data: Accurate and Efficient

<table>
<thead>
<tr>
<th></th>
<th>Su</th>
<th>Mo</th>
<th>Tu</th>
<th>We</th>
<th>Th</th>
<th>Fr</th>
<th>Sa</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>40</td>
<td>30</td>
<td>10</td>
<td>43</td>
<td>55</td>
<td>30</td>
<td>20</td>
</tr>
<tr>
<td>z</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>h</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>60</td>
<td>3</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>b</td>
<td>20</td>
<td>10</td>
<td>5</td>
<td>24</td>
<td>15</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>f</td>
<td>0</td>
<td>7</td>
<td>6</td>
<td>3</td>
<td>8</td>
<td>5</td>
<td>20</td>
</tr>
<tr>
<td>g</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td>84</td>
<td>70</td>
<td>50</td>
</tr>
<tr>
<td>d</td>
<td>13</td>
<td>10</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>

Query(DATA) → Estimation

\[ \sum_{h \in D} |v_{hi} - v_{hj}|^p \]

Microsoft Research

Edith Cohen
Property Testing:
- A matrix has rank (or stable rank) \( \leq d \), or
- Need to modify an enough fraction of the matrix to make it have rank (stable rank) \( \leq d \)

- Entry access model / Row access model
- Randomized algorithm
- Near-optimal sample complexity
DeepWalk: Online Learning of Social Representations

Our contributions:
- Generalizing Deep Learning for NLP to graphs
- Evaluation on Multi-label classification tasks
- Parallelizable Method

DeepWalk at-a-glance

1. Input: Graph
2. Deep Learning for NLP
3. Output: Representation
Madness Session (26th)
Chairs: Aristides Gionis and Jie Tang

10:30am - 12:00am
Industry & Govt 4 Education, Privacy and Best Paper Presentations
Chair: Prem Melville
Research 13 Diffusion in Social and Information Networks
Chair: Le Song
Research 14 Recommender Systems
Chair: Shou-De Lin
Research 15 Security and Privacy
Chair: Shengrui Wang
Research 16 Active Learning
Chair: Heng Huang

1:30pm - 3:00pm
Industry & Govt 5 Monitoring & Maintenance
Chair: Balaji Krishnapuram
Research 17 Graph Mining and Modeling
Chair: B. Aditya Prakash
Research 18 Clustering
Chair: Ian Davidson
Research 19 Trend, Anomaly, and Novelty Detection
Chair: Naoki Abe

3:30pm - 5:00pm
Industry & Govt 6 Workforce Analytics & Personalization
Chair: Mohak Shah
Research 20 Social Network Analysis
Chair: Tina Eliassi-Rad
Research 21 Large-scale Optimization and Learning
Chair: Christos Boutsidis
Research 22 Topic Modeling
Chair: Jonathan Chang
Research 22
Topic Modeling

1. TCS: Efficient Topic Discovery over Crowd-oriented Service Data
2. SigniTrend: Scalable Detection of Emerging Topics in Textual Streams by Hashed Significance Thresholds
3. Experiments with Non-parametric Topic Models
4. Reducing the Sampling Complexity of Topic Models
5. Dynamics of News Events and Social Media Reaction
Crowd-Oriented Services

- Crowd-Oriented Service Systems: Yahoo! Answers, Quora, Stack Overflow, etc.
- Unique Structure of Crowd-Oriented Service Data: Task-Response Pairs
- Fundamental Tasks: Discovering Topics and Tracking Topic Evolution over Time

Our Approach: Topic Crowd Service (TCS) Model

- **TCS Model**: Capturing the Structure of Task-Response Pairs
- **Pairwise Sketch (pSketch)**: Only Storing Significant Word Pairs
- **BPE Method**: Fast Parameter Estimation using Belief Propagation and pSketch
**SigniTrend**: Scalable Detection of Emerging Topics in Textual Streams by Hashed Significance Thresholds

**Popular words** occur at a **high frequencies**

- “Facebook” 460/h
- “WhatsApp” 250/h

**Pairs are less frequent → significant bursts**

- “Facebook” AND “WhatsApp” 190/h 2/h

Facebook bought WhatsApp

How to track all pairs **efficiently**?

**Tuesday 26.8.2014, Research 22 @ Empire West**

Erich Schubert, Michael Weiler, Hans-Peter Kriegel
Experiments with Non-parametric Topic Models by Wray Buntine and Swapnil Mishra

1. Builds high-fidelity topics: improves performance dramatically
2. Burstiness handling can be added to all kinds of models
3. Works with multi-core too!!!
4. Gibbs sampling comparable with variational methods
Reducing the Sampling Complexity of Topic Models
Aaron Q. Li, Amr Ahmed, Sujith Ravi, Alex J. Smola - Google and CMU

Problem
• Collapsed sampling for topic models is $O(k)$
• This happens for fancy models and big data

Solution
• Approximate changing distribution using Metropolis Hastings
• Amortized $O(1)$ draws using the Alias Method for proposal

Results
• Highest throughput samplers for LDA (>1M/s)
• Application to Pitman Yor Process and HDP
We take social- and news media as *dynamic system*, reacting to events by outputting news publications.

Our model reconstructs correct event importance, longitude and timing from volume of publications.

We correlate important events to sentiment shifts from Twitter, revealing their dependency.