Madness Session
(25th)

CHAIRS: ARISTIDES GIONIS AND JIE TANG
Madness Session (25th)
Chairs: Aristides Gionis and Jie Tang

10:30am - 12:00am
Industry & Govt 1
Health & Safety
Chair: Ted Senator
Research 1
Applications to Healthcare and Medicine I
Chair: Nitesh Chawla
Research 2
Location-based Services
Chair: Vincent S. Tseng
Research 3
Statistical Techniques for Big Data
Chair: Vikas Sindhwani
Research 4
Feature Selection
Chair: Katharina Morik

1:30pm - 3:00pm
Industry & Govt 2
Social Media & Listening
Chair: James Shanahan
Research 5
Dynamic Graph Analysis
Chair: Tanya Berger-Wolf
Research 6
Supervised Learning I
Chair: Xiaodan Zhang
Research 7
Scaling-up Methods for Big Data
Chair: Chih-Jen Lin
Research 8
Data Streams
Chair: Yan Liu

3:30pm - 5:00pm
Industry & Govt 3
Fraud/Threat Detection & Environment
Chair: Rob Cooley
Research 9
Scaling-up Graph Algorithms
Chair: Polo Chau
Research 10
Text Mining
Chair: Jiawei Han
Research 11
Supervised Learning II
Chair: Evimaria Terzi
Research 12
Applications to Healthcare and Medicine II
Chair: Xiang Zhang
Industry & Govt 1
Health & Safety

**CHAIR: TED SENATOR**

1. How to Create a $1B Model in 20 Days
2. Proactive Workflow Modeling by Stochastic Processes with Application to Healthcare Operation and Management
3. ISIS: A Networked-Epidemiology Based Pervasive Web App for Infectious Disease Pandemic Planning and Response
4. FoodSIS: A Text Mining System to Improve the State of Food Safety in Singapore
5. Reducing Gang Violence Through Network Influence Based Targeting of Social Programs
6. Spatially Embedded Co-offence Prediction Using Supervised Learning
How to Create a $1B Model in 20 Days

Predictive Modeling in the Real World
- A Sprint Case Study

Jeremy Howard: Enlitic, CEO
Tracey De Poalo: Sprint, Manager of Predictive Modeling
Proactive Workflow Modeling for Healthcare Management
Monday 10:30AM-12PM | NY Ballroom East
Chuanren Liu, Yong Ge, Hui Xiong, Keli Xiao, Wei Geng, Matt Perkins
isis: a networked-epidemiology based web app for infectious disease pandemic planning/response

richard beckman, keith bisset, jiangzhuo chen, bryan lewis, madhav marathe, paula stretz

use – inform decision makers for pandemic planning and response

- diseases include
  - influenza
  - ebola
  - pertussis

- interventions
  - pharmaceuticals
  - school closure
  - sequestration
  - k-neighbors
  - dynamic trigger

- hpc models
  - individual based
  - highly parallel

- up to national scale

use - design experiment, model, analyze, visualize, forecast, planning, decision cycle
FoodSIS: A Text Mining System to Improve the State of Food Safety in Singapore

Kiran Kate, IBM Research;
Sneha Chaudhari, Carnegie Mellon University, Pittsburgh, USA;
Andy Prapanca, IBM Research;
Jayant Kalagnanam, IBM Research;
Reducing Gang Violence through Network Influence Based Targeting of Social Programs

Police approach or “call-in” gang members to encourage dis-enrollment from the gang.
- Current techniques based on ad-hoc selection of gang members
- Anecdotal evidence of dis-enrollment spreading like a contagion

Can we leverage influence maximization on a social network to encourage gang-member dis-enrollment?

In this talk we provide:
- A diffusion model based on police observations
- An influence problem as a non-monotonic submodular maximization problem
- An algorithm with approximation guarantee
- A model geographic aspects of influence
- An examination of performance on a real-world police dataset
- An overview of our ongoing work with the Chicago Police

Paulo Shakarian (Arizona State University), Joseph Salmento (West Point), William Pulleyblank (West Point), John Bertetto (Chicago Police)
Spatially Embedded Co-offence Prediction Using Supervised Learning

Mohammad Tayebi, Martin Ester
Uwe Glässer, Patricia Brantingham

Police Arrest Data

Co-offending Network Extraction

Co-offence Prediction

- **Predictive policing**: Tackling crime before it happens
- **Research question**: Can we predict criminal collaborations?
- **Approach**: Feature-based co-offence prediction framework

**Conclusion**: Our approach can predict 90% of the co-offences correctly.
Madness Session (25th)

Chairs: Aristides Gionis and Jie Tang

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
<th>Chair</th>
</tr>
</thead>
<tbody>
<tr>
<td>10:30am - 12:00am</td>
<td>Industry &amp; Govt 1 Health &amp; Safety</td>
<td>Ted Senator</td>
</tr>
<tr>
<td></td>
<td>Research 1 Applications to Healthcare and Medicine I</td>
<td>Nitesh Chawla</td>
</tr>
<tr>
<td></td>
<td>Research 2 Location-based Services</td>
<td>Vincent S. Tseng</td>
</tr>
<tr>
<td></td>
<td>Research 3 Statistical Techniques for Big Data</td>
<td>Vikas Sindhwani</td>
</tr>
<tr>
<td></td>
<td>Research 4 Feature Selection</td>
<td>Katharina Morik</td>
</tr>
<tr>
<td>1:30pm - 3:00pm</td>
<td>Industry &amp; Govt 2 Social Media &amp; Listening</td>
<td>James Shanahan</td>
</tr>
<tr>
<td></td>
<td>Research 5 Dynamic Graph Analysis</td>
<td>Tanya Berger-Wolf</td>
</tr>
<tr>
<td></td>
<td>Research 6 Supervised Learning I</td>
<td>Xiaodan Zhang</td>
</tr>
<tr>
<td></td>
<td>Research 7 Scaling-up Methods for Big Data</td>
<td>Chih-Jen Lin</td>
</tr>
<tr>
<td></td>
<td>Research 8 Data Streams</td>
<td>Yan Liu</td>
</tr>
<tr>
<td>3:30pm - 5:00pm</td>
<td>Industry &amp; Govt 3 Fraud/Threat Detection &amp; Environment</td>
<td>Rob Cooley</td>
</tr>
<tr>
<td></td>
<td>Research 9 Scaling-up Graph Algorithms</td>
<td>Polo Chau</td>
</tr>
<tr>
<td></td>
<td>Research 10 Text Mining</td>
<td>Jiawei Han</td>
</tr>
<tr>
<td></td>
<td>Research 11 Supervised Learning II</td>
<td>Evimaria Terzi</td>
</tr>
<tr>
<td></td>
<td>Research 12 Applications to Healthcare and Medicine II</td>
<td>Xiang Zhang</td>
</tr>
</tbody>
</table>
Research 1
Applications to Healthcare and Medicine I

CHAIR: NITESH CHAWLA

1. LUDIA: An Aggregate-Constrained Low-Rank Reconstruction Algorithm to Leverage Publicly Released Health Data
2. People on Drugs: Credibility of User Statements in Health Communities
3. Unfolding Physiological State: Mortality Modelling in Intensive Care Units
4. Unsupervised Learning of Disease Progression Models
5. Good-Enough Brain Model: Challenges, Algorithms and Discoveries in Multi-Subject Experiments
LUDIA: An Aggregate-Constrained Low-Rank Reconstruction Algorithm to Leverage Publicly Released Health Data

- Yubin Park and Joydeep Ghosh from UT Austin

What can we say about the relationship between physical inactivity and diabetes at individual-level?

Ecological Fallacy!
People on Drugs: Credibility of User Statements in Health Communities

Subhabrata Mukherjee, Gerhard Weikum and Cristian Danescu-Niculescu-Mizil

I took a cocktail of meds. Xanax gave me hallucinations and a demonic feel.

Xanax and Prozac are known to cause drowsiness.

Xanax made me dizzy and sleepless.

User Trustworthiness

Language Objectivity

Statement Credibility

"Applications to Healthcare and Medicine 1", NY Ballroom West, 10:45am
Unfolding Physiological State Mortality Modeling in Intensive Care Units

R1-3

Marzyeh Ghassemi, Tristan Naumann Finale Doshi-Velez, Nicole Brimmer Rohit Joshi, Anna Rumshisky Peter Szolovits

NY Ballroom West @ 10:30 am!
Unsupervised Learning of Disease Progression Models
Xiang Wang (IBM Research), David Sontag (NYU), Fei Wang (IBM Research)

- **Goal**: Learn the progression trajectory of diseases from EHR data
- **Challenge**: Irregular visits, limited supervision, censored records
- **Approach**: Markov Jump Process + Noisy-Or Bayesian Networks

**Input**: Patient records
- 06/2008 – Hypertension
- 06/2009 – Chronic Bronchitis
- 11/2009 – Back Pain
- 01/2010 – Pneumonia
- 06/2010 – Chest Pain, Anxiety
- 03/2011 – Heart Failure

**Output**: Progression stages and semantics
- **Mild**: Presence of cardiovascular and respiratory disorders
- **Moderate**: Onset of musculoskeletal disorders; acute respiratory infection leads to exacerbation
- **Severe**: Psychological disorders, heart failure, kidney failure.
“apple” “Is it edible?” (y/n)
“knife” “Can it hurt you?” (y/n)

Frontal lobe (attention)

Parietal lobe (movement)

Occipital lobe (vision)

Temporal lobe (language)

**Good-Enough Brain Model: Challenges, Algorithms and Discoveries in Multi-Subject Experiments**

Evangelos Papalexakis¹, Alona Fyshe¹, Nicholas Sidiropoulos², Partha Talukdar¹, Tom Mitchell¹, Christos Faloutsos¹

¹Carnegie Mellon University, ²University of Minnesota
Madness Session (25th)
Chairs: Aristides Gionis and Jie Tang

10:30am - 12:00am
- Industry & Govt 1: Health & Safety
  Chair: Ted Senator
- Research 1: Applications to Healthcare and Medicine I
  Chair: Nitesh Chawla
- Research 2: Location-based Services
  Chair: Vincent S. Tseng
- Research 3: Statistical Techniques for Big Data
  Chair: Vikas Sindhwani
- Research 4: Feature Selection
  Chair: Katharina Morik

1:30pm - 3:00pm
- Industry & Govt 2: Social Media & Listening
  Chair: James Shanahan
- Research 5: Dynamic Graph Analysis
  Chair: Tanya Berger-Wolf
- Research 6: Supervised Learning I
  Chair: Xiaodan Zhang
- Research 7: Scaling-up Methods for Big Data
  Chair: Chih-Jen Lin
- Research 8: Data Streams
  Chair: Yan Liu

3:30pm - 5:00pm
- Industry & Govt 3: Fraud/Threat Detection & Environment
  Chair: Rob Cooley
- Research 9: Scaling-up Graph Algorithms
  Chair: Polo Chau
- Research 10: Text Mining
  Chair: Jiawei Han
- Research 11: Supervised Learning II
  Chair: Evimaria Terzi
- Research 12: Applications to Healthcare and Medicine II
  Chair: Xiang Zhang
Research 2
Location-based Services

CHAIR: VINCENT S. TSENG

1. Prediction of Human Emergency Behavior and their Mobility following Large-scale Disaster
2. Inferring User Demographics and Social Strategies in Mobile Social Networks
3. Travel Time Estimation of a Path using Sparse Trajectories
4. Modeling Human Location Data with Mixtures of Kernel Densities
5. A Cost-Effective Recommender System for Taxi Drivers
Can we predict human emergency behavior and its movements by modeling its past movements during disaster?

If some future disaster occur, given person’s current observed movements:
- which place will it go next time period?
- How about its traveling routes?

Session: Location-based Services, Monday 10:30 am, Empire East
Did you know: As of 2014, there are **7.3** billion mobile users.

Users average **22** calls, **23** messages, and **110** status checks per day.

**Inferring User Demographics and Social Strategies in Mobile Social Networks**

Yuxiao Dong, Yang Yang, Jie Tang, Yang Yang, Nitesh V. Chawla

Monday at **10:30am**, Room **Empire East**, Research Session 2: Location-based Services
Travel Time Estimation of a Path using Sparse Trajectories

For any path in a road network

Instantly Using a sample of vehicles

To be presented at 11:00am Today in Research 2 session

Data Released!
Modeling Human Location Data with Mixtures of Kernel Densities

Moshe Lichman and Padhraic Smyth
Department of Computer Science
University of California, Irvine
A Cost-Effective Recommender System for Taxi Drivers
Meng Qu, Hengshu Zhu, Junming Liu, Guannan Liu, Hui Xiong

- **Cost-Effective Mobile Recommender System**
  Focusing on recommending an entire driving route for taxi drivers which yields maximum potential net profit per unit searching time.

- **Recursive Recommendation Strategy**

- **The net profit for each route $R$ start at $r_1$:**

  $G(R, r_1, M) = g(r_1) + \sum_{i=2}^{M} g(r_i) \prod_{j=1}^{i-1} (1 - P(r_j))$

- **Empirical Studies on Recommendations**
## Madness Session (25th)

**Chairs:** Aristides Gionis and Jie Tang

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
</table>
| 10:30am - 12:00am | Industry & Govt 1 Health & Safety  
Chair: Ted Senator  
Research 1 Applications to Healthcare and Medicine I  
Chair: Nitesh Chawla |
|               | Research 2 Location-based Services  
Chair: Vincent S. Tseng  
Research 3 Statistical Techniques for Big Data  
Chair: Vikas Sindhwani |
|               | Research 4 Feature Selection  
Chair: Katharina Morik |
| 1:30pm - 3:00pm | Industry & Govt 2 Social Media & Listening  
Chair: James Shanahan  
Research 5 Dynamic Graph Analysis  
Chair: Tanya Berger-Wolf  
Research 6 Supervised Learning I  
Chair: Xiaodan Zhang |
|               | Research 7 Scaling-up Methods for Big Data  
Chair: Chih-Jen Lin  
Research 8 Data Streams  
Chair: Yan Liu |
| 3:30pm - 5:00pm | Industry & Govt 3 Fraud/Threat Detection & Environment  
Chair: Rob Cooley  
Research 9 Scaling-up Graph Algorithms  
Chair: Polo Chau  
Research 10 Text Mining  
Chair: Jiawei Han |
|               | Research 11 Supervised Learning II  
Chair: Evimaria Terzi  
Research 12 Applications to Healthcare and Medicine II  
Chair: Xiang Zhang |
Research 3
Statistical Techniques for Big Data

Chair: Vikas Sindhwani

1. Parallel Gibbs Sampling for Hierarchical Dirichlet Processes via Gamma Processes Equivalence
2. Empirical Glitch Explanations
3. Learning with Dual Heterogeneity: A Nonparametric Bayes Model
4. Online Chinese Restaurant Process
5. Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion
Parallel Gibbs Sampling for Hierarchical Dirichlet Processes via Gamma Processes Equivalence

Dehua Cheng and Yan Liu
University of Southern California

Research 3, Monday 10:30AM-12PM, Empire West

“An approximate parallel sampling algorithm for HDP-equivalent model by replacing Dirichlet-multinomial hierarchy with Gamma-Poisson hierarchy”

• Experiments show that our algorithm can achieve considerable speedup as well as better inference accuracy for HDP compared with existing parallel sampling algorithms.
**EMPIRICAL GLITCH EXPLANATIONS**: Dasu, Loh & Srivastava

- **Convention**: Identify data glitches & repair e.g. “Phone # is unique” → De-duplicate

- **Our thesis**: Not all data glitches are bad, just misunderstood ...

- **Our proposal**: Associate explanations with groups of data glitches, automatically e.g. “New Hire with Supervisor’s phone number” (blue)

- **How to use**:
  1. Unexplained glitches (red block) → Repair;
  2. Explanations → Domain experts → Refine domain knowledge e.g. “Phone # is unique except when (a) new hire has supervisor’s phone number”

- **Outcome**
  - Only 1/3 of “bad” data repaired, reducing statistical distortion induced by cleaning, as well as cost and cycle times.

<table>
<thead>
<tr>
<th>Empl.</th>
<th>Status</th>
<th>Phone</th>
<th>Dept.</th>
<th>Room</th>
<th>Super.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID_1</td>
<td>Active</td>
<td>1AAA36000000</td>
<td>D4000</td>
<td></td>
<td>ID_4</td>
</tr>
<tr>
<td>ID_2</td>
<td></td>
<td>1AAA36000000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ID_3</td>
<td>Retired</td>
<td>1AAA36000000</td>
<td>D2200</td>
<td>E260</td>
<td>ID_6</td>
</tr>
<tr>
<td>ID_5</td>
<td>Active</td>
<td>1AAA3608776</td>
<td>D2300</td>
<td>A115</td>
<td>ID_9</td>
</tr>
<tr>
<td>ID_7</td>
<td>New Hire</td>
<td>1AAA3608776</td>
<td>D2300</td>
<td>D284</td>
<td>ID_5</td>
</tr>
<tr>
<td>ID_8</td>
<td>New Hire</td>
<td>1AAA3608776</td>
<td>D2300</td>
<td></td>
<td>ID_5</td>
</tr>
<tr>
<td>ID_10</td>
<td>Active</td>
<td>1AAA3605519</td>
<td>D8000</td>
<td>A132</td>
<td>ID_13</td>
</tr>
<tr>
<td>ID_11</td>
<td>Active</td>
<td>1AAA3605519</td>
<td>D8000</td>
<td>A132</td>
<td>ID_13</td>
</tr>
<tr>
<td>ID_12</td>
<td>Active</td>
<td>1AAA3605519</td>
<td>D8000</td>
<td>A132</td>
<td>ID_13</td>
</tr>
</tbody>
</table>
LEARNING WITH DUAL HETEROGENEITY: A NONPARAMETRIC BAYES MODEL

Hongxia Yang and Jingrui He

Motivation: Many real-world problems exhibit dual-heterogeneity, e.g., social media, manufacturing, traffic analytics.

Solution: Graphical representation embedded in a Nonparametric Bayes framework

Application

Better Sharing and Learning
The probability of table assignment for the next customer

\[ P(z_i=j|z_{-i}, y_{-i}, x_i, \theta, G_0, \alpha) \propto \begin{cases} 
(1 + \gamma_1)^j (1 - \gamma_2)^{e_j} \frac{m_j}{i-1+\alpha} H(x_i, \theta_j), & \text{if } j \leq k \\
\frac{\alpha}{i-1+\alpha} \int H(x_i, \theta_j) dG_0(\theta_j), & \text{if } j = k + 1 
\end{cases} \]

The probability of table assignment for the next customer

\[ (1 + \gamma_1) \frac{2}{3+\alpha} H(x_i, \theta_1') \quad (1 - \gamma_2) \frac{1}{3+\alpha} H(x_i, \theta_2') \quad \frac{\alpha}{3+\alpha} \int H(x_i, \theta_3) dG_0(\theta_3) \]
Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion

Xin Luna Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, Wei Zhang

- Data from web
  - Unstructured text
  - Semi-structured DOM trees
  - Structured WebTables
- "Prior" data from KG
# Madness Session (25th)

**Chairs:** Aristides Gionis and Jie Tang

<table>
<thead>
<tr>
<th>Time</th>
<th>Session 1</th>
<th>Session 2</th>
<th>Session 3</th>
<th>Session 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>10:30am-12:00am</td>
<td>Industry &amp; Govt 1 Health &amp; Safety Chair: Ted Senator</td>
<td>Research 1 Applications to Healthcare and Medicine I Chair: Nitesh Chawla</td>
<td>Research 2 Location-based Services Chair: Vincent S. Tseng</td>
<td>Research 4 Feature Selection Chair: Katharina Morik</td>
</tr>
<tr>
<td>1:30pm-3:00pm</td>
<td>Industry &amp; Govt 2 Social Media &amp; Listening Chair: James Shanahan</td>
<td>Research 5 Dynamic Graph Analysis Chair: Tanya Berger-Wolf</td>
<td>Research 6 Supervised Learning I Chair: Xiaodan Zhang</td>
<td>Research 7 Scaling-up Methods for Big Data Chair: Chih-Jen Lin</td>
</tr>
<tr>
<td>3:30pm-5:00pm</td>
<td>Industry &amp; Govt 3 Fraud/Threat Detection &amp; Environment Chair: Rob Cooley</td>
<td>Research 9 Scaling-up Graph Algorithms Chair: Polo Chau</td>
<td>Research 10 Text Mining Chair: Jiawei Han</td>
<td>Research 11 Supervised Learning II Chair: Evimaria Terzi</td>
</tr>
</tbody>
</table>
Research 4

Feature Selection

CHAIR: KATHARINA MORIK

1. Effective Global Approaches for Mutual Information Based Feature Selection
2. Gradient Boosted Feature Selection
3. Simultaneous Feature and Feature Group Selection through Hard Thresholding
4. Safe and Efficient Screening For Sparse Support Vector Machine
5. Factorized Sparse Learning Models with Interpretable High Order Feature Interactions
Effective **Global Approaches** for Mutual Information Based Feature Selection

**Monday 25th Aug 10:30AM-12PM Riverside Ballroom**

- **Existing incremental selection methods**
  - Select 1 feature at a time
  - Minimum Redundancy Maximum Relevance
  - No coming back: cannot deselect features

- **Our global approach:**
  - Consider all features **concurrently**: assign optimal weights to all features
  - Simple global solution: dominant eigenvector

- **Code available online**

**Vinh Nguyen, Jeffrey Chan, Simone Romano and James Bailey**

*Department of Computing and Information Systems, The University of Melbourne, Victoria, Australia*
Gradient Boosted Feature Selection
Zhixiang (Eddie) Xu, Gao Huang, Kilian Q. Weinberger, Alice X. Zheng

**Introduction**

- **Combines** feature selection and classification.
- Scales to very large data sets, \( n \gg p \)
- Discovers non-linear feature dependency,
- Incorporates known feature sparsity structure.

**Method**

**Intuitive** optimization problem

\[
\ell(H) + \lambda \|H\|_1 + \mu q(H)
\]

loss regularization feature selection

Optimized using Gradient Boosted Regression Trees (GBRT)

A modification to the GBRT impurity function

\[
\sum_{i=1}^{n} \left( g_i - h(x_i) \right)^2 + \mu f(\cdot)
\]

Encourage feature re-using

**Results**

![Graph showing feature selection results](image)
Simultaneously Feature and Feature Group Selection through Hard Thresholding

Variable selection
Feature group selection
Flexible control

Shuo Xiang, Tao Yang and Jieping Ye, Arizona State University
Model selection is crucial for $L_1$-regularized SVM. But, it is a very expensive process… Can we make it faster?

Safe Screening

- Makes use of the variational inequality
- Closed-form solution for screening inactive features
- Every feature that is removed is guaranteed to be inactive in the optimal solution
- Most inactive features are removed in each iteration
- Speedup is from tens to hundreds times
Factorized Sparse Learning Models with Interpretable High Order Feature Interactions
Sanjay Purushotham, Martin Renqiang Min, C.-C. Jay Kuo, Rachel Ostroff

Problem: Identify interpretable discriminative high order feature interactions
Example: Pairwise feature interactions

Our approach: Factorized High order Interactions Model (FHIM)
- Capture High Order interactions using tensor product
- Solved by Greedy Alternating Optimization algorithm

Biomarker Discovery for Blood-based cancer diagnosis
Madness Session (25th)
Chairs: Aristides Gionis and Jie Tang

10:30am - 12:00am
Industry & Govt 1
Health & Safety
Chair: Ted Senator
Research 1
Applications to Healthcare and Medicine I
Chair: Nitesh Chawla
Research 2
Location-based Services
Chair: Vincent S. Tseng
Research 3
Statistical Techniques for Big Data
Chair: Vikas Sindhwani
Research 4
Feature Selection
Chair: Katharina Morik

1:30pm - 3:00pm
Industry & Govt 2
Social Media & Listening
Chair: James Shanahan
Research 5
Dynamic Graph Analysis
Chair: Tanya Berger-Wolf
Research 6
Supervised Learning I
Chair: Xiaodan Zhang
Research 7
Scaling-up Methods for Big Data
Chair: Chih-Jen Lin
Research 8
Data Streams
Chair: Yan Liu

3:30pm - 5:00pm
Industry & Govt 3
Fraud/Threat Detection & Environment
Chair: Rob Cooley
Research 9
Scaling-up Graph Algorithms
Chair: Polo Chau
Research 10
Text Mining
Chair: Jiawei Han
Research 11
Supervised Learning II
Chair: Evimaria Terzi
Research 12
Applications to Healthcare and Medicine II
Chair: Xiang Zhang
Industry & Govt 2
Social Media & Listening

CHAIR: JAMES SHANAHAN

1. Activity Ranking in LinkedIn Feed
2. Large-Scale High-Precision Topic Modeling on Twitter
3. LASTA: Large Scale Topic Assignment on Multiple Social Networks
5. 'Beating the News' with EMBERS: Forecasting Civil Unrest using Open Source Indicators
6. Modeling Mass Protest Adoption in Social Network Communities using Geometric Brownian Motion
Activity Ranking in LinkedIn Feed

- LinkedIn feed data
  - Hierarchy of activities
  - Diversity and freshness
  - Connection relationship
  - Platform dependence

- Relevance system
  - Large scale distributed training platform
  - Offline replay evaluation

- Online bucket tests
  - Freshness
  - Impression discounting
  - Personalization
Large scale high-precision topic modeling on Twitter
Shuang, Alek, Andy and Pankaj

Infer topics of tweets

At full Twitter scale
- 270M+ MAUs, 500M+ tweets/day, 400B aggregate …

In real-time
- 150K+ request per second …

Over an ontology of hundreds of topics
- 350+ atomic topics with hierarchical relations …

With super-high precision
- 90+% precision, 40% overall coverage
LASTA: Large Scale Topic Assignment on Multiple Social Networks

Nemanja Spasojevic, Jinyun Yan, Adithya Rao, Prantik Bhattacharyya
{nemanja, jinyun, adithya, prantik}@klout.com

- Topical interest mining production pipeline.
- ~10,000 topics assigned to hundreds of millions of users daily.
- Features generated from posts, profiles and social graph connections.
- Cross-network analysis for data from Twitter, Facebook, LinkedIn, Google+, Wikipedia.

<table>
<thead>
<tr>
<th>User</th>
<th>Top 10 Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marissa Mayer</td>
<td>yahoo, google, technology, business, twitter, social-media, flickr, design, marketing, seo, gmail</td>
</tr>
<tr>
<td>Lady Gaga</td>
<td>music, lady-gaga, celebrities, art, fashion, born-this-way, venus, entertainment, radio</td>
</tr>
<tr>
<td>Barack Obama</td>
<td>politics, affordable-care-act, healthcare, new-york-times, congress, chicago, twitter, washington, illinois</td>
</tr>
</tbody>
</table>

LASTA topic assignment examples
EARS (Earthquake Alert and Report System): a Real Time Decision Support System for Earthquake Crisis Management

Stefano Cresci, Maurizio Tesconi, Marco Avvenuti, Andrea Marchetti and Carlo Meletti
‘Beating the News’ with EMBERS: Forecasting Civil Unrest using Open Source Indicators


---

Planned Protest

Dynamic Query Expansion

Volume-Based

Cascades Model

Baseline Model

Fusion and Suppression

Date

# of Warnings

Warning ID: W1793
When: 01/04/2014
Where: Ecuador, Pinchincha, Quito
Who: Ethnic
Why: Energy and Resources
Probability: 0.87
Forecast Date: 12/27/13

GSR Event ID: E1859
When: 01/02/14
Where: Ecuador, Pinchincha, Quito
Who: Ethnic
Why: Energy and Resources
Reported Date: 01/05/14

Evaluated by MITRE
Modeling Mass Protest Adoption in Social Network Communities using Geometric Brownian Motion

Fang Jin¹,², Rupinder Paul Khandpur¹,², Nathan Self¹,², Edward Dougherty², Sheng Guo³, Feng Chen⁴, B. Aditya Prakash¹,², Naren Ramakrishnan¹,²

¹ Discovery Analytics Center, ² Virginia Tech, ³ LinkedIn Inc., ⁴ University at Albany

What types of mobilization patterns underlie protests?

Can we employ geometric Brownian motion to capture adoption in social networks?
Madness Session (25th)

Chairs: Aristides Gionis and Jie Tang

10:30am - 12:00am

- Industry & Govt 1: Health & Safety
  Chair: Ted Senator
- Research 1: Applications to Healthcare and Medicine I
  Chair: Nitesh Chawla
- Research 2: Location-based Services
  Chair: Vincent S. Tseng
- Research 3: Statistical Techniques for Big Data
  Chair: Vikas Sindhwani
- Research 4: Feature Selection
  Chair: Katharina Morik

1:30pm - 3:00pm

- Industry & Govt 2: Social Media & Listening
  Chair: James Shanahan
- Research 5: Dynamic Graph Analysis
  Chair: Tanya Berger-Wolf
- Research 6: Supervised Learning I
  Chair: Xiaodan Zhang
- Research 7: Scaling-up Methods for Big Data
  Chair: Chih-Jen Lin
- Research 8: Data Streams
  Chair: Yan Liu

3:30pm - 5:00pm

- Industry & Govt 3: Fraud/Threat Detection & Environment
  Chair: Rob Cooley
- Research 9: Scaling-up Graph Algorithms
  Chair: Polo Chau
- Research 10: Text Mining
  Chair: Jiawei Han
- Research 11: Supervised Learning II
  Chair: Evimaria Terzi
- Research 12: Applications to Healthcare and Medicine II
  Chair: Xiang Zhang
Research 5
Dynamic Graph Analysis

**Chair: Tanya Berger-Wolf**

1. Non-Parametric Scan Statistics for Event Detection and Forecasting in Heterogeneous Social Media Graphs
2. Event Detection in Activity Networks
3. FEMA: Flexible Evolutionary Multi-faceted Analysis for Dynamic Behavioral Pattern Discovery
4. Profit-maximizing Cluster Hires
5. On Social Event Organization
Non-Parametric Scan Statistics for Event Detection and Forecasting in Heterogeneous Social Media Graphs

Feng Chen¹, Daniel B. Neill²
SUNY Albany¹, Carnegie Mellon University²

\[
\arg\max_{S \subseteq \mathcal{V}} F(S), \quad \text{s.t. } S \text{ is connected}
\]

Non-parametric scan statistics function
Event detection in activity networks

Polina Rozenshtein, Aris Gionis, Nikolaj Tatti
Aris Anagnostopoulos

given a graph with vertex weights and edge distances
find a heavy subgraph with vertices in short distances

application: find events in cities

(a) 01.06.12 Primavera sound music festival
(b) 18.09.12 festival of the Poblenou neighborhood
(c) 31.10.12 Halloween
**FEMA**: Flexible Evolutionary Multi-faceted Analysis for Dynamic Behavior Pattern Discovery

Meng Jiang, Peng Cui, Fei Wang, Xinran Xu, Wenwu Zhu, Shiqiang Yang

**Dynamic Graph Analysis Session**: Behavior Modeling

**Challenges**
- Sparsity: high-order
- Complexity: long seq.

- Multi-faceted
- Dynamic
- Human behavior

- Problem
  - Behavior modeling
- Pattern discovery
- Behavior prediction

- Tensor sequence
- Decomposition
- Completion

- Time
  - User
  - Item
  - t₁, t₂, t₃
Profit-maximizing Cluster Hires
Behzad Golshan, Theodoros Lappas, Evimaria Terzi

You are the Manager.
Which workers to hire and which jobs to do?

Contributions:
• Formalizing the problem
• Proposing an efficient method
• Evaluating real-world datasets
On Social Event Organization

Keqian Li, Wei Lu, Laks V. S. Lakshmanan
Smriti Bhagat
Cong Yu

You want to organize the annual corporate day, and plan activities/events that the employees will enjoy with one another.

Participants have preferences for events, however, want to be with friends. Events too have participation constraints.

Come at 1:30pm, NY Ballroom West to hear about a novel social network based, event organization problem!
Madness Session (25th)
Chairs: Aristides Gionis and Jie Tang

10:30am - 12:00am
- Industry & Govt 1 Health & Safety
  Chair: Ted Senator
- Research 1 Applications to Healthcare and Medicine I
  Chair: Nitesh Chawla
- Research 2 Location-based Services
  Chair: Vincent S. Tseng
- Research 3 Statistical Techniques for Big Data
  Chair: Vikas Sindhwani
- Research 4 Feature Selection
  Chair: Katharina Morik

1:30pm - 3:00pm
- Industry & Govt 2 Social Media & Listening
  Chair: James Shanahan
- Research 5 Dynamic Graph Analysis
  Chair: Tanya Berger-Wolf
- Research 6 Supervised Learning I
  Chair: Xiaodan Zhang
- Research 7 Scaling-up Methods for Big Data
  Chair: Chih-Jen Lin
- Research 8 Data Streams
  Chair: Yan Liu

3:30pm - 5:00pm
- Industry & Govt 3 Fraud/Threat Detection & Environment
  Chair: Rob Cooley
- Research 9 Scaling-up Graph Algorithms
  Chair: Polo Chau
- Research 10 Text Mining
  Chair: Jiawei Han
- Research 11 Supervised Learning II
  Chair: Evimaria Terzi
- Research 12 Applications to Healthcare and Medicine II
  Chair: Xiang Zhang
Research 6
Supervised Learning I

CHAIR: XIAODAN ZHANG

1. FastXML: A Fast, Accurate and Stable Tree-classifier for eXtreme Multi-label Learning
2. A Multi-class Boosting Method with Direct Optimization
3. An Efficient Algorithm For Weak Hierarchical Lasso
4. Online Multiple Kernel Regression
5. Class-Distribution Regularized Consensus Maximization for Alleviating Overfitting in Model Combination
FastXML: A Fast & Stable Tree-classifier for eXtreme Multi-label Learning (ID 532)

Yash Prabhu (IIT Delhi) & Manik Varma (MSR)

• Extreme Classification
  • Learning with millions of categories
  • New paradigm for ranking and recommendation

• Limitations of the state-of-the-art
  • Require a very large cluster for training

• Our Contributions
  • Significantly improve classification accuracy
  • Train on a single box (code publicly available)
A Multi-class Boosting Method with Direct Optimization

Shaodan Zhai, Tian Xia, and Shaojun Wang
Kno.e.sis Center, Wright State University

- **Direct Multi-Class Boosting** (DMCBoost)
- Direct solves the multi-class classification tasks
- Only requires very weak base classifiers
- Direct optimizes the non-convex performance measures, including classification error and margin
- Robust to noisy data
An Efficient Algorithm For Weak Hierarchical Lasso

Yashu Liu, Jie Wang, Jieping Ye
Arizona State University

**Linear Regression with pairwise interactions:**

\[ y = w_0 + \sum_{i=1}^{d} x_i w_i + \frac{1}{2} \sum_{i=1}^{d} \sum_{j=1}^{d} x_i x_j Q_{ij} + \epsilon \]

**Weak hierarchy constraint:**

\[ Q_{ij} \neq 0 \text{ only if } w_i \neq 0 \text{ OR } w_j \neq 0 \]

- **Show the non-convex sub-problem admits a closed-form solution**
- **Further accelerate the computation: from quadratic to lineararithmic**
Online Multiple Kernel Regression

Doyen Sahoo¹, Steven C.H. Hoi¹, Bin Li²
1. Singapore Management University, 2. Wuhan University

- From Batch to Online Learning
- From Linear to Kernel Methods
- From Single to Multiple Kernels

\[ y_t \leftarrow f(x_t) \]
\[ \ell(y_t, f(x_t)) \]
\[ x_t \]

OMKR

\[ x_t \]
\[ K_1 \quad \text{Kernel Regressor 1} \quad \theta_1 \]
\[ K_2 \quad \text{Kernel Regressor 2} \quad \theta_2 \]
\[ K_m \quad \text{Kernel Regressor m} \quad \theta_m \]

Update

\[ \min_{\theta \in \Delta} \min_{f \in \mathcal{H}_{K(\theta)}} \left\{ \frac{1}{2} |f|^2_{\mathcal{H}_{K(\theta)}} + C \sum_{i=1}^{n} \ell(f(x_i), y_i) \right\} \]
\[ \Delta = \{ \theta \in \mathbb{R}^m_+ | \theta^T 1_m = 1 \} \]
\[ K(\theta)(\cdot, \cdot) = \sum_{i=1}^{m} \theta_i K_i(\cdot, \cdot) \]
Class-Distribution Regularized Consensus Maximization for Alleviating Overfitting in Model Combination

Sihong Xie, Jing Gao, Wei Fan, Deepak Turaga and Philip S. Yu
# Madness Session (25th)

**Chairs:** Aristides Gionis and Jie Tang

## Timetable

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
</table>
| 10:30am - 12:00am | Industry & Govt 1 Health & Safety Chair: Ted Senator  
                               Research 1 Applications to Healthcare and Medicine I Chair: Nitesh Chawla  
                               Research 2 Location-based Services Chair: Vincent S. Tseng  
                               Research 3 Statistical Techniques for Big Data Chair: Vikas Sindhwani  
                               Research 4 Feature Selection Chair: Katharina Morik |
| 1:30pm - 3:00pm | Industry & Govt 2 Social Media & Listening Chair: James Shanahan  
                               Research 5 Dynamic Graph Analysis Chair: Tanya Berger-Wolf  
                               Research 6 Supervised Learning I Chair: Xiaodan Zhang  
                               Research 7 Scaling-up Methods for Big Data Chair: Chih-Jen Lin  
                               Research 8 Data Streams Chair: Yan Liu |
| 3:30pm - 5:00pm | Industry & Govt 3 Fraud/Threat Detection & Environment Chair: Rob Cooley  
                               Research 9 Scaling-up Graph Algorithms Chair: Polo Chau  
                               Research 10 Text Mining Chair: Jiawei Han  
                               Research 11 Supervised Learning II Chair: Evimaria Terzi  
                               Research 12 Applications to Healthcare and Medicine II Chair: Xiang Zhang |
Research 7
Scaling-up Methods for Big Data

CHAIR: CHIH-JEN LIN

1. Improving the Modified Nyström Method Using Spectral Shifting

2. Fast Flux Discriminant for Large-Scale Sparse Nonlinear Classification

3. Scalable Histograms on Large Probabilistic Data

4. Correlation Clustering in MapReduce

5. Scaling Out Big Data Missing Value Imputations
Improving the Modified Nystrom Method Using Spectral Shifting

Shusen Wang, Chao Zhang, Hui Qian, Zhihua Zhang
Monday 1:30PM-3:00PM | Empire West

- Standard /Modified Nyström: \( K \approx C U C^T \)
  - \( C = K[:,:,:] \), \( U = (K[:,:,\text{idx}])^\dagger \) (Standard) or \( C^\dagger K(C^\dagger)^T \) (Modified)
  - Performs badly when the eigenvalue spectrum of the original matrix decays slowly, such as Gauss kernel matrix with big \( \alpha \)

- Spectral shifting Nyström: \( K - \lambda I_m \approx C U C^T \)
  - Give a shift to the spectrum first and then do the approximation
  - Better upper bound than other Nyström method
Fast Flux Discriminant for Large-Scale Sparse Nonlinear Classification

Wenlin Chen, Yixin Chen, Kilian Q. Weinberger
## Scalable Histograms on Large Probabilistic Data

Mingwang Tang and Feifei Li

### Histogram on Probabilistic Data
- Possible world semantic
- Histogram on probabilistic data based on expectation \([ICDE09]\)
- Optimal histogram using SSE: \(O(Bn^2)\) time

### Scalable histogram based on partition and merge principle
- Recursive merge method with constant approximate ratio
- Distributed and parallel methods on Hadoop
- Tradeoff between communication and computation
Correlation Clustering in Map-Reduce

Flavio Chierichetti
Sapienza

Nilesh Dalvi
Trooly

Ravi Kumar
Google

CC: important clustering tool
MR: large-scale data computation

Cluster 100s of millions of elements
Scaling out Big Data Missing Value Imputations

Pythia vs. Godzilla

Christos Anagnostopoulos & Peter Triantafillou
Madness Session (25th)
Chairs: Aristides Gionis and Jie Tang

10:30am - 12:00am
Industry & Govt 1 Health & Safety
Chair: Ted Senator
Research 1 Applications to Healthcare and Medicine I
Chair: Nitesh Chawla
Research 2 Location-based Services
Chair: Vincent S. Tseng
Research 3 Statistical Techniques for Big Data
Chair: Vikas Sindhwani
Research 4 Feature Selection
Chair: Katharina Morik

1:30pm - 3:00pm
Industry & Govt 2 Social Media & Listening
Chair: James Shanahan
Research 5 Dynamic Graph Analysis
Chair: Tanya Berger-Wolf
Research 6 Supervised Learning I
Chair: Xiaodan Zhang
Research 7 Scaling-up Methods for Big Data
Chair: Chih-Jen Lin
Research 8 Data Streams
Chair: Yan Liu

3:30pm - 5:00pm
Industry & Govt 3 Fraud/Threat Detection & Environment
Chair: Rob Cooley
Research 9 Scaling-up Graph Algorithms
Chair: Polo Chau
Research 10 Text Mining
Chair: Jiawei Han
Research 11 Supervised Learning II
Chair: Evimaria Terzi
Research 12 Applications to Healthcare and Medicine II
Chair: Xiang Zhang
Research 8
Data Streams

Prototype-based Learning on Concept-drifting Data Streams
Detecting Moving Object Outliers In Massive-Scale Trajectory Streams
The Setwise Stream Classification Problem
Streamed Approximate Counting of Distinct Elements
Time-Varying Learning and Content Analytics via Sparse Factor Analysis
Prototype-based Learning on Concept-drifting Data Streams

Junming Shao, Zahra Ahmadi and Stefan Kramer
Detecting Moving Object Outliers In Massive-Scale Trajectory Streams

Yanwei Yu @Yantai U., Lei Cao @WPI, Elke A. Rundensteiner @ WPI, Qin Wang @USTB

Research 8: Data Streams  Monday 1:30PM-3:00PM | Riverside Ballroom
The problem of classification is typically formulated in the context of individual instances of records.

In many applications, it is desirable to classify entire sets of records as indivisible entities.

- The problem is equivalent to that of classifying sets of data sets.

- Each of the data sets may have a set of records of different cardinality.
Streamed Approximate Counting of Distinct Elements.  
(Daniel Ting)

**Problem:** Estimate the cardinality of a set

- **Our work:** Unify/understand *all* sketching methods for cardinality estimation.
  - Drive practical improvements
  - Provably beat existing “optimal” methods

- 8 MB memory per set (Cardinality=10M)
  - 134 KB LPCA
  - 40 KB Min-count
  - 6.1 KB HyperLogLog
SPARFA-Trace: Time-varying Learning and Content Analytics via Sparse Factor Analysis

Andrew Lan, Christoph Studer, Richard Baraniuk

- Analyze learner responses and activity logs in classrooms/MOOCs
- Trace learner knowledge state evolution over time
- Estimate learning resource content organization, quality and difficulty
Madness Session (25th)
Chairs: Aristides Gionis and Jie Tang

10:30am - 12:00am
Industry & Govt 1
Health & Safety
Chair: Ted Senator
Research 1
Applications to Healthcare and Medicine I
Chair: Nitesh Chawla
Research 2
Location-based Services
Chair: Vincent S. Tseng
Research 3
Statistical Techniques for Big Data
Chair: Vikas Sindhwani
Research 4
Feature Selection
Chair: Katharina Morik

1:30pm - 3:00pm
Industry & Govt 2
Social Media & Listening
Chair: James Shanahan
Research 5
Dynamic Graph Analysis
Chair: Tanya Berger-Wolf
Research 6
Supervised Learning I
Chair: Xiaodan Zhang
Research 7
Scaling-up Methods for Big Data
Chair: Chih-Jen Lin
Research 8
Data Streams
Chair: Yan Liu

3:30pm - 5:00pm
Industry & Govt 3
Fraud/Threat Detection & Environment
Chair: Rob Cooley
Research 9
Scaling-up Graph Algorithms
Chair: Polo Chau
Research 10
Text Mining
Chair: Jiawei Han
Research 11
Supervised Learning II
Chair: Evimaria Terzi
Research 12
Applications to Healthcare and Medicine II
Chair: Xiang Zhang
Industry & Govt 3
Fraud/Threat Detection & Environment

Chair: Rob Cooley

1. Guilt by Association: Large Scale Malware Detection by Mining File-relation Graphs
2. Knock It Off: Profiling the Online Storefronts of Counterfeit Merchandise
3. Corporate Residence Fraud Detection
4. Improving Management of Aquatic Invasions by Integrating Shipping Network, Ecological, and Environmental Data: Data Mining for Social Good
5. Novel Geospatial Interpolation Analytics for General Meteorological Measurements
Guilt by Association:
Large Scale Malware Detection by Mining File-relation Graphs

Acar Tamersoy, Kevin Roundy, Polo Chau

- How to detect malware sooner at higher accuracy?
- AESOP achieves 99.6% recall with 0.01% false alarms
  - Locality-sensitive hashing to compute file co-occurrence
  - Belief propagation to spread labels to unknown files
- Being deployed by Symantec for billions of files

% of files AESOP detects at least 1 week ahead of current technology

**True Positive Rate (TP)**

**False Positive Rate (FP)**
Knock It Off: Profiling the Online Storefronts of Counterfeit Merchandise

Matthew F. Der, Lawrence K. Saul, Stefan Savage, Geoffrey M. Voelker
Industry & Govt 3:
Fraud/Threat Detection & Environment

3:30PM-5PM | NY Ballroom East

Bart Minnaert  David Martens  Enric Junqué de Fortuny
Marija Stankova  Julie Moeyersoms  Foster Provost
Improving Management of Aquatic Invasions by Integrating Shipping Network, Ecological, and Environmental Data: Data Mining for Social Good

Jian Xu, Thanuka Wickramarathne, Nitesh Chawla, Erin Grey, Karsten Steinhaeuser, Reuben Keller, John Drake, David Lodge

Annual losses by ship-borne aquatic invasion: $800 million in the Great Lakes, more globally. An invisible tax that everyone has to pay.
Novel Geospatial Interpolation Analytics for General Meteorological Measurements

Bingsheng Wang and Jinjun Xiong
# Madness Session (25th)

**Chairs:** Aristides Gionis and Jie Tang

## Time Schedule

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
</table>
| 10:30am - 12:00am | Industry & Govt 1
  - Health & Safety  
  Chair: Ted Senator |
| 10:30am - 12:00am | Research 1
  - Applications to Healthcare and Medicine I  
  Chair: Nitesh Chawla |
| 10:30am - 12:00am | Research 2
  - Location-based Services  
  Chair: Vincent S. Tseng |
| 10:30am - 12:00am | Research 3
  - Statistical Techniques for Big Data  
  Chair: Vikas Sindhwani |
| 10:30am - 12:00am | Research 4
  - Feature Selection  
  Chair: Katharina Morik |
| 1:30pm - 3:00pm  | Industry & Govt 2
  - Social Media & Listening  
  Chair: James Shanahan |
| 1:30pm - 3:00pm  | Research 5
  - Dynamic Graph Analysis  
  Chair: Tanya Berger-Wolf |
| 1:30pm - 3:00pm  | Research 6
  - Supervised Learning I  
  Chair: Xiaodan Zhang |
| 1:30pm - 3:00pm  | Research 7
  - Scaling-up Methods for Big Data  
  Chair: Chih-Jen Lin |
| 1:30pm - 3:00pm  | Research 8
  - Data Streams  
  Chair: Yan Liu |
| 3:30pm - 5:00pm  | Industry & Govt 3
  - Fraud/Threat Detection & Environment  
  Chair: Rob Cooley |
| 3:30pm - 5:00pm  | Research 9
  - Scaling-up Graph Algorithms  
  Chair: Polo Chau |
| 3:30pm - 5:00pm  | Research 10
  - Text Mining  
  Chair: Jiawei Han |
| 3:30pm - 5:00pm  | Research 11
  - Supervised Learning II  
  Chair: Evimaria Terzi |
| 3:30pm - 5:00pm  | Research 12
  - Applications to Healthcare and Medicine II  
  Chair: Xiang Zhang |
Research 9
Scaling-up Graph Algorithms

CHAIR: POLO CHAU

1. Almost Linear-Time Algorithms for Adaptive Betweenness Centrality using Hypergraph Sketches
2. Efficient SimRank Computation via Linearization
3. FAST-PPR: Scaling Personalized PageRank Estimation for Large Graphs
4. Graph Sample and Hold: A Framework for Big-Graph Analytics
5. Balanced Graph Edge Partition
Almost Linear-Time Algorithms for Adaptive Betweenness Centrality Using Hypergraph Sketches

by Yuichi Yoshida (NII)

Ordering by adaptive betweenness centrality
for $i = 1$ to $n$:

$v_i \leftarrow$ node with the highest BC.
Remove (the effect of) $v_i$.

Applications
- Distance oracle
- Immunization
- Community detection

Contributions
- Almost linear-time approximation algorithm
- Provable guarantee
- 1000x times faster by experiment
Efficient **SimRank** Computation via **Linearization**

Takanori Maehara, Mituru Kusumoto, Ken-ichi Kawarabayashi

- SimRank computation is **EXPENSIVE** because it is a solution of **NONLINEAR** recursion.
- We find a **LINEAR** formulation of SimRank
  1. Linearize SimRank
  2. Solve **linear equation** to compute SimRank

→ Scales up to $|V|=100M$, $|E|=4G$ graphs
FAST-PPR: Making Search Personal
Peter Lofgren, Sid Banerjee, Ashish Goel (Stanford), C. Seshadhri (Sandia)

Scaling Personalized PageRank to Large Graphs

Bi-Directional Method

20x Faster
Given a large graph $G$ represented as a stream of edges $e_1, e_2, e_3, \ldots$

We show how to efficiently sample from $G$ using limited memory space to calculate unbiased estimates of various graph properties.

**Graph Sample and Hold:**

A Framework for Big-Graph Analytics

---

Nesreen Ahmed$^1$, Nick Duffield$^2$, Jennifer Neville$^1$, Ramana Kompella$^1$

$^1$Purdue University, $^2$Rutgers University
Balanced Graph Edge Partition

Florian Bourse, ENS;
Marc Lelarge, INRIA-ENS;
Milan Vojnovic, Microsoft Research;
Madness Session (25th)
Chairs: Aristides Gionis and Jie Tang

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
</table>
| 10:30am - 12:00am | Industry & Govt 1 Health & Safety  
Chair: Ted Senator  
Research 1 Applications to Healthcare and Medicine I  
Chair: Nitesh Chawla  
Research 2 Location-based Services  
Chair: Vincent S. Tseng  
Research 3 Statistical Techniques for Big Data  
Chair: Vikas Sindhwani  
Research 4 Feature Selection  
Chair: Katharina Morik |
| 1:30pm - 3:00pm | Industry & Govt 2 Social Media & Listening  
Chair: James Shanahan  
Research 5 Dynamic Graph Analysis  
Chair: Tanya Berger-Wolf  
Research 6 Supervised Learning I  
Chair: Xiaodan Zhang  
Research 7 Scaling-up Methods for Big Data  
Chair: Chih-Jen Lin  
Research 8 Data Streams  
Chair: Yan Liu |
| 3:30pm - 5:00pm | Industry & Govt 3 Fraud/Threat Detection & Environment  
Chair: Rob Cooley  
Research 9 Scaling-up Graph Algorithms  
Chair: Polo Chau  
Research 10 Text Mining  
Chair: Jiawei Han  
Research 11 Supervised Learning II  
Chair: Evimaria Terzi  
Research 12 Applications to Healthcare and Medicine II  
Chair: Xiang Zhang |
Research 10
Text Mining

Chair: Jiawei Han

1. Mining Topics in Documents: Standing on the Shoulders of Big Data
2. Integrating Spreadsheet Data via Accurate and Low-Effort Extraction
3. Sentiment Expression Conditioned by Affective Transitions and Social Forces
4. Entity Profiling with Varying Source Reliabilities
5. Open Question Answering Over Curated and Extracted Knowledge Bases
Mining Topics in Documents: Standing on the Shoulders of Big Data

Zhiyuan (Brett) Chen     Bing Liu

Session: Text Mining
Time: 3:30pm – 3:48pm
Integrating Spreadsheet Data via Accurate and Low-Effort Extraction

Shirley Zhe Chen and Michael Cafarella
{chenzhe, michjc}@umich.edu

What is the strength of the connection between smoking and lung cancer for each of the 50 U.S. states?

Smoking Spreadsheet
From census.gov

Lung Cancer Spreadsheet
From census.gov
Sentiment Expression Conditioned by Affective Transitions and Social Forces

Moritz Sudhof, Andrés Gómez Emilsson, Andrew L. Maas, Christopher Potts
{sudhof, nc07agom, amaas, cgpotts}@stanford.edu
Entity Profiling with Varying Source Reliabilities

Furong Li, Mong Li Lee, Wynne Hsu (NUS)

- Various name representations
- Erroneous attribute values
- Incomplete information
- Ambiguous references

Data from multiple sources

- Model source reliabilities
- Lower impact of erroneous values

Complete & accurate profiles

<table>
<thead>
<tr>
<th>Name</th>
<th>Address</th>
<th>Phone</th>
<th>Cuisine</th>
<th>Recommend</th>
<th>Price</th>
<th>Weekday Hours</th>
<th>Weekend Hours</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frank Restaurant</td>
<td>407 Colorado St</td>
<td>512-494-6916</td>
<td>American</td>
<td>Hot dog</td>
<td>$</td>
<td>Normal</td>
<td>Extend</td>
<td>8.2</td>
</tr>
<tr>
<td>Frank &amp; Angie’s Pizzeria</td>
<td>508 West Ave</td>
<td>512-472-3524</td>
<td>Italian</td>
<td>Pizza</td>
<td>$$</td>
<td>Normal</td>
<td>Night</td>
<td>8.0</td>
</tr>
</tbody>
</table>
Why can't your computer...

...answer your questions?!

1. Your computer is ignorant!
2. Your computer doesn't understand you!

Open Question Answering Over Curated and Extracted Knowledge Bases

Fader, Zettlemoyer, & Etzioni
Text Mining Session
Madness Session (25th)
Chairs: Aristides Gionis and Jie Tang

10:30am - 12:00am
Industry & Govt 1
Health & Safety
Chair: Ted Senator
Research 1
Applications to Healthcare and Medicine I
Chair: Nitesh Chawla
Research 2
Location-based Services
Chair: Vincent S. Tseng
Research 3
Statistical Techniques for Big Data
Chair: Vikas Sindhwani
Research 4
Feature Selection
Chair: Katharina Morik

1:30pm - 3:00pm
Industry & Govt 2
Social Media & Listening
Chair: James Shanahan
Research 5
Dynamic Graph Analysis
Chair: Tanya Berger-Wolf
Research 6
Supervised Learning I
Chair: Xiaodan Zhang
Research 7
Scaling-up Methods for Big Data
Chair: Chih-Jen Lin
Research 8
Data Streams
Chair: Yan Liu

3:30pm - 5:00pm
Industry & Govt 3
Fraud/Threat Detection & Environment
Chair: Rob Cooley
Research 9
Scaling-up Graph Algorithms
Chair: Polo Chau
Research 10
Text Mining
Chair: Jiawei Han
Research 11
Supervised Learning II
Chair: Evi Maria Terzi
Research 12
Applications to Healthcare and Medicine II
Chair: Xiang Zhang
Research 11
Supervised Learning II

CHAIR: EVIMARIA TERZI

1. Large Margin Distribution Machine
2. Distance Metric Learning Using Dropout: A Structured Regularization Approach
3. Box Drawings for Learning with Imbalanced Data
4. Incremental and Decremental Training for Linear Classification
5. Supervised Deep Learning with Auxiliary Networks
The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining

Large Margin Distribution Machine

Teng Zhang and Zhi-Hua Zhou

LAMDA Group
National Key Laboratory for Novel Software Technology
Nanjing University, China
{zhangt, zhouzh}@lamda.nju.edu.cn
Distance Metric Learning Using Dropout
A Structured Regularization Approach
Qi Qian (MSU), Juhua Hu (SFU), Rong Jin (MSU), Jian Pei (SFU), Shenghuo Zhu (NEC)

Too many may hurt!
Please drop some!
Too many parameters in DML to learn can easily lead to overfitting!

How about Dropout in DML?

Dropout on Metric

Dropout on Data

$L_p$ Norm & Structured Norm

Trace Norm

Better Generalization Performance
Models are made of Axis Aligned Rectangles for Interpretability.

Two algorithms at polar extremes:
- **Exact Boxes**: single mixed-integer linear program to perfectly place all the boxes.
- **Fast Boxes**: clusters the minority class, then uses 1-dimensional ML for each decision boundary.

Some thought provoking questions
1. Can you really encode Exact Boxes as a single mixed-integer linear program?
2. Can Fast Boxes perform almost as well as Exact Boxes?
3. What is the main reason many other methods fail? (it’s surprising!)
4. Do you need to sacrifice accuracy to gain interpretability?

Find out at the poster and the talk!
Incremental and Decremental Training for Linear Classification

Cheng-Hao Tsai, Chieh-Yen Lin, Chih-Jen Lin
National Taiwan University

- Incremental/decremental learning are useful in updating models when data are not changed much.

- But they are not guaranteed to be faster than re-training. Also implementation may be complicated.

- We propose and analyze a warm start strategy to effectively reduce the training time.

- Our main conclusion: warm start for a primal-based high-order optimization method (e.g., Newton) is preferred.
Supervised Deep Learning with Auxiliary Networks

Junbo Zhang\textsuperscript{1,2}, Guangjian Tian\textsuperscript{2}, Yadong Mu\textsuperscript{2}, Wei Fan\textsuperscript{2}

\textsuperscript{1}School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, China
\textsuperscript{2}Huawei Noah's Ark Lab, Hong Kong

jbzhang@my.swjtu.edu.cn, \{tian.guangjian, mu.yadong, david.fanwei\}@huawei.com

Motivation: Huge data, but few labeled

1. Labeling Data is Very Expensive
   - Sample-specific annotations
   - Side information (similarity/dissimilarity constraints)
     - More flexible
     - Greatly mitigates the workload of annotators

2. Existing Deep Learning Schemes
   - Unsupervised Pre-training + Supervised Fine-turning
     - DBN, Stacked Autoencoders
   - Semi-supervised or Guided Autoencoder

3. Problems and Shortcoming
   - Ineffectively handle sparse side information
   - Sample-specific annotations are always required

Solution: SUGAR

- Effectively Handle Side Information
- More Robust, Flexible, Easily Extendible
- General Model for \textbf{Feature Learning} from both unlabeled and labeled data

Potential Application Areas

1. Handwriting Recognition
2. Domain Adaptation
3. Telecommunication Data Mining
4. Others
   - Multi-source data
   - Few Labeled data

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure.png}
\caption{Experiments}
\end{figure}
Madness Session (25th)

Chairs: Aristides Gionis and Jie Tang

10:30am - 12:00am
- Industry & Govt 1: Health & Safety
  Chair: Ted Senator
- Research 1: Applications to Healthcare and Medicine I
  Chair: Nitesh Chawla
- Research 2: Location-based Services
  Chair: Vincent S. Tseng
- Research 3: Statistical Techniques for Big Data
  Chair: Vikas Sinderswani
- Research 4: Feature Selection
  Chair: Katharina Morik

1:30pm - 3:00pm
- Industry & Govt 2: Social Media & Listening
  Chair: James Shanahan
- Research 5: Dynamic Graph Analysis
  Chair: Tanya Berger-Wolf
- Research 6: Supervised Learning I
  Chair: Xiaodan Zhang
- Research 7: Scaling-up Methods for Big Data
  Chair: Chih-Jen Lin
- Research 8: Data Streams
  Chair: Yan Liu

3:30pm - 5:00pm
- Industry & Govt 3: Fraud/Threat Detection & Environment
  Chair: Rob Cooley
- Research 9: Scaling-up Graph Algorithms
  Chair: Polo Chau
- Research 10: Text Mining
  Chair: Jiawei Han
- Research 11: Supervised Learning II
  Chair: Evimaria Terzi
- Research 12: Applications to Healthcare and Medicine II
  Chair: Xiang Zhang
Research 12
Applications to Healthcare and Medicine II

CHAIR: XIANG ZHANG

1. Scalable Noise Mining in Long-Term Electrocardiographic Time-Series to Predict Death Following Heart Attacks
3. FUNNEL: Automatic Mining of Spatially Coevolving Epidemics
4. From Micro to Macro: Data Driven Phenotyping by Densification of Longitudinal Electronic Medical Records
5. Clinical Risk Prediction with Multilinear Sparse Logistic Regression
6. Dual Beta Process Priors for Latent Cluster Discovery in Chronic Obstructive Pulmonary Disease
Scalable Noise Mining in Long-Term Electrocardiographic Time-Series to Predict Death Following Heart Attacks

Chih-Chun Chia, University of Michigan, Ann Arbor; Zeeshan Syed, University of Michigan, Ann Arbor;
Can we derive interpretable medical concepts with minimal supervision?

MARBLE: **HIGH-THROUGHPUT PHENOTYPING FROM ELECTRONIC HEALTH RECORDS VIA SPARSE NONNEGATIVE TENSOR FACTORIZATION**

Joyce C. Ho\(^1\), Joydeep Ghosh\(^1\), Jimeng Sun\(^2\)

\(^1\) University of Texas at Austin, \(^2\) Georgia Institute of Technology
FUNNEL: Automatic Mining of Spatially Coevolving Epidemics

**P1 Seasonality**
- April (4)
- May (5)
- June (6)
- July (7)
- August (8)
- September (9)
- October (10)
- November (11)
- December (12)
- January (1)
- February (2)
- March (3)
- Local patterns
- Vaccination

**P2 Vaccination**
- Original
- S(t)
- I(t)
- V(t)

**P3 Local patterns**
- External shocks

**P4 External shocks**
- Mistakes, errors

Yasuko Matsubara, Yasushi Sakurai, Willem G. van Panhuis, Christos Faloutsos
From Micro to Macro: **Data Driven Phenotyping** by Densification of Longitudinal Electronic Medical Records

Jiayu Zhou, Fei Wang, Jianying Hu, Jieping Ye

---

**Extremely Sparse!**

Patient Longitudinal Electronic Medical Records

Patient ID (100153)

<table>
<thead>
<tr>
<th>DMII Ophth Uncrtnd</th>
<th>Aortic Atherosclerosis</th>
<th>Hx Antineoplastic Chemotherapy</th>
<th>Pure Hypercholesterolemia</th>
<th>Acq Absence Cert/Unirec</th>
</tr>
</thead>
<tbody>
<tr>
<td>(250.52)</td>
<td>(440.0)</td>
<td>(877.41)</td>
<td>(272.0)</td>
<td>(186.01)</td>
</tr>
<tr>
<td>Bmi 32.0 – 32.9 (V85.32)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bmi 38.0 – 38.9 (V85.36)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cont Exp Haz Aromat NEC (V87.19)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bmi 31.0 – 31.9 (V85.31)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bmi 30.0 – 30.9 (V85.30)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bmi 46 And Over (V85.4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hy Kid Noc W Cr Kid I – Iv (V85.90)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chronic Kidney Dis NOS (V85.5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dmi Ren T St Uncrtnd (V85.42)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chr Kidney Dis Stage III (V85.53)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comp Oth Vasc Dev/v Wtr (V85.74)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bmi 35.0 – 35.9 (V85.35)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Estrogen Recep Neg Stat (V85.1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bmi 34.0 – 34.9 (V85.34)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bmi 36.0 – 36.9 (V85.36)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bmi 39.0 – 39.9 (V85.39)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bmi 37.0 – 37.9 (V85.37)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Edema (V82.3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bmi 32.0 – 32.9 (V85.32)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dmi Wo Crp St St Uncrtnd (V85.50)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anemia In Chr Kidney Dis (V85.21)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

Concept Evolution

Multi-Task Matrix Completion Problem

MALSAR

Multi-Task Learning via structural regularization
Clinical Risk Prediction with Multilinear Sparse Logistic Regression

Fei Wang, Ping Zhang, Buyue Qian, Xiang Wang, Ian Davidson

Healthcare Analytics Research Group, IBM T. J. Watson Research Department of Computer Science, UC Davis

Multi-dimensional mode medical data

\[
\min_{\mathcal{W}} J(\mathcal{W}, b) = \ell(\mathcal{W}, b) + R(\mathcal{W})
\]

\[
f_{\mathcal{W}, b}(x^i) = x^i \times_1 w_1 \times_2 w_2 \cdots \times_K w^K + b
\]

\[
\ell_l(\mathcal{W}, b) = \log[1 + \exp(-y_i f_{\mathcal{W}, b}(x^i))]
\]

\[
R(\mathcal{W}) = R_1(\mathcal{W}) + R_2(\mathcal{W}) = \sum_{k=1}^{K} \lambda_k \| w^k \|_1 + \frac{1}{2} \sum_{k=1}^{K} \mu_k \| w^k \|_2^2
\]
Dual Beta Process Priors for Latent Cluster Discovery in Chronic Obstructive Pulmonary Disease
J. C. Ross, P. J. Castaldi, M. H. Cho, J. G. Dy

Problem
COPD is a major health concern, not completely understood, and HETEROGENEOUS

Approach
• Group people based on smoke exposure response (DISEASE TRAJECTORIES)
• Use GAUSSIAN PROCESSES to model trajectories
• Use BETA PROCESSES to identify latent clusters and salient features

Results
Four clusters identified with significant association to genetic markers