ARL On-Site Visit Meeting at UIUC

- INARC ARL government lead and visitors
  - Lance Kaplan (ARL, INARC Government Lead)
  - Elizabeth Bowman (ARL)
  - John Kosinski (CERDEC)
  - Robert Winkler (ARL)

- INARC UIUC Principal Investigators
  - Tarek Abdelzaher
  - Jiawei Han
  - Thomas Huang
  - Dan Roth

- Students and Researchers Representing the Data Mining Group
  - Dr. Hongbo Deng (INARC Research Scientist)
  - Manish Gupta, Ming Ji, Zhenhui Li, Yizhou Sun, Chi Wang, Bo Zhao
Information Network Research at UIUC and at the Data Mining Group

Jiawei Han
Department of Computer Science
University of Illinois at Urbana-Champaign
Outline

- Information Network Research at UIUC
  - Tarek Abdelzaher
  - Jiawei Han
  - Thomas Huang
  - Dan Roth
- Information Network Research at the Data Mining Group@CS.UIUC
  - Research Team
  - Research Projects
Information Network Research at UIUC

- **Prime Member of INARC:**
  - Four researchers, each supervising a team of junior researchers

- **PI and research tasks**
  - Jiawei Han [I3, I3.1, I3.2, Edin, Trust]
  - Tarek Abdelzaher [I1, I1.1, Edin]
  - Thomas Huang [I1.2]
  - Dan Roth [I3.3, Trust]

- Participating all NS-CTA and INARC activities

- Bi-weekly research meetings & weekly PI-based research meetings

- Actively collaborating and contributing to INARC research frontiers
Information Network Research at the Data Mining Group

- **Research Team and Research Tasks Covered**
  - Yizhou Sun, Ming Ji, Chi Wang, Jing Gao: Mining Information Networks (I3.1)
  - Zhenhui Li, Lu An Tang, Xiao Yu: Cyber-Physical Networks (I3.2)
  - Hongbo Deng, Bo Zhao, Cindy Xide Lin, Yintao Yu: Google News and Text Mining (I3.3)
  - Yizhou Sun, Peixiang Zhao: Similarity Search in Information Networks (I2.2)
  - Manish Gupta, Yizhou Sun: Evolution and Dynamics in Information Networks (E3.3)
  - Manish Gupta, Zhijun Yin: Trust in Information Networks (T1.2)

- **Collaborative contributions**
  - Xin Jin, Zhijun Yin: Multimedia mining in information networks (working with Thomas Huang’s group)
  - Hyung Sul Kim, Zhijun Yin, Xin Jin: Sensor network mining and information networks (working with Tarek Abdelzaher’s group)
Task I3.1: Mining Dynamic, Heterogeneous Information Networks

- **IPP Major Research Tasks and New Research Progress**
  - Classification of heterogeneous information networks
  - Starting background: RankClus, NetClus, iNextCube
  - New work: GNetClass (PKDD’10 sub)
  - Methods for pattern discovery in evolutionary heterogeneous information networks
  - Discovery of hidden semantic relationships (e.g., identifying leaders and their followers by mining records of group activities: KDD’10)
  - Evolution of heterogeneous information networks: PKDD’10 sub
  - Anomaly and outlier discovery in dynamic heterogeneous information networks
  - Detection of community outliers in information networks (KDD’10)
Integrating Clustering with Ranking for Heterogeneous Information Network Analysis (RankClus, NetClus and iNextCube)

Global Ranking? vs. Clustering in heterogeneous network?

What feature can I use?

Cluster and rank people or events
Find highly suspicious groups/events

Table 5: Top-10 Conferences in 5 Clusters Using RankClus

<table>
<thead>
<tr>
<th>DB</th>
<th>Network</th>
<th>AI</th>
<th>Theory</th>
<th>IR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>VLDB</td>
<td>INFOCOM</td>
<td>AAMAS</td>
<td>SODA</td>
</tr>
<tr>
<td>2</td>
<td>ICDE</td>
<td>SIGMETRICS</td>
<td>IJCAI</td>
<td>STOC</td>
</tr>
<tr>
<td>3</td>
<td>SIGMOD</td>
<td>ICNP</td>
<td>AAAI</td>
<td>FOCOS</td>
</tr>
<tr>
<td>4</td>
<td>KDD</td>
<td>SIGCOMM</td>
<td>Agents</td>
<td>ICALP</td>
</tr>
<tr>
<td>5</td>
<td>ICDM</td>
<td>MOBICOM</td>
<td>AAAI/IAAI</td>
<td>CCC</td>
</tr>
<tr>
<td>6</td>
<td>EDBT</td>
<td>ICDCS</td>
<td>ECAI</td>
<td>SPAA</td>
</tr>
<tr>
<td>7</td>
<td>DASFAA</td>
<td>NETWORKING</td>
<td>RoboCup</td>
<td>PODC</td>
</tr>
<tr>
<td>8</td>
<td>PODS</td>
<td>MobiHoc</td>
<td>IAT</td>
<td>CRYPTO</td>
</tr>
<tr>
<td>9</td>
<td>SSDBM</td>
<td>ISCC</td>
<td>ICMAS</td>
<td>APPROX-RANDOM</td>
</tr>
<tr>
<td>10</td>
<td>SDM</td>
<td>SenSys</td>
<td>CP</td>
<td>EUROCRYPT</td>
</tr>
</tbody>
</table>

Flickr: Query “Raleigh”
Role Discovery and Community Outlier Analysis

- Role discovery in information networks (KDD’10)
  - Latent knowledge in information network:
    - Roles: who is the commander/manager/supervisor?
    - Relationships: friends/relatives/colleagues/enemies?
    - Influence: who influences/is influenced by me most? Who is the opinion leader?
  - Can we discover them by exploring the links?
  - If we can, it will benefit our study in
    - Community structure $\leftarrow$ clustering & classification
    - Authoritative sources $\leftarrow$ search & ranking
    - Evolution patterns $\leftarrow$ prediction & recommendation

- Community outlier detection (KDD’10)
  - Outlier if in the community but have properties or behaviors different from it
Role Discovery: Advisor-Advisee

- **Input:** Research publication network
- **Output:** potential advising relationship & their ranking: \((r, [st, ed])\)
**Task I3.2: Real-Time Methods for Spatiotemporal Information-Related Cyber-Physical Network Analysis**

**Objectives and Collaborations**

- Knowledge discovery in cyber-physical networks ⇒ Interact with CNARC
- Spatiotemporal information-centered clustering ⇒ Interact with CNARC and SCNARC when modeling needs spatial-based reasoning
- Evolution analysis of cyber-physical networks ⇒ EDIN and IRC

**Recent Progress and On-going work**

- User-guided clustering of cyber-physical networks
  - Put physical networks and info. networks under a common analytic foundation and study their interdependencies
- Online clustering of dynamic and evolving cyber-physical networks
  - Cluster analysis in cyber-physical networks
  - Classification in cyber-physical networks
  - Military applications in cyber-physical networks
Discovery of Periodic Patterns of Moving Object Clusters

- A system that mines moving object patterns: Z. Li, et al., “MoveMine: Mining Moving Object Databases”, SIGMOD’10 (system demo)
- Z. Li, B. Ding, J. Han, and R. Kays, “Mining Hidden Periodic Behaviors for Moving Objects”, KDD’10

- Z. Li, B. Ding, J. Han, and R. Kays, “Swarm: Mining Relaxed Temporal Moving Object Clusters”. VLDB’10 (sub)

Bird flying paths shown on Google Earth

Mined periodic patterns by our new method

Swarm discovers more patterns

Convoy discovers only restricted patterns
Zhenhui Li, Ming Ji, Jae-Gil Lee, LuAn Tang, Yintao Yu, Jiawei Han, and Roland Kays, “MoveMine: Mining Moving Object Databases” (system demo), Proc. 2010 ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'10), Indianapolis, Indiana, June 2010.
Cyber-Physical Networks

- From moving objects to cyber-physical network ...
Task I3.3: Mining Text and Unstructured Data for InfoNet Analysis

Motivation and Goals

- Most military applications need to handle text data, including documents, e-mails, telecommunication messages, micro blogs and conversations
- Text and unstructured data form a critical part of information networks
- Multidimensional analysis of information networks associated with text data
- Combine text mining and network analysis and leverage the power of statistical topic modeling and discrete regularization

Recent Research Progress

- Multi-dimensional text information network analysis
- **Text Cube, Topic Cube, iTopicModel**: Information network enhanced topic modeling
- Dynamic language modeling
- **Google News analyzer**: Integrate text mining and information network analysis
Effective OLAP Exploration

- TEXplorer (submitted to VLDB’10): integrating keyword-based ranking and OLAP exploration.
I3.3. Google News Project: Object Identification and Network Analysis

Top 20 nodes

- Washington, D.C.
- Sasha
- American
- George Washington
- United States of America
- John F. Kennedy, Kennedy
- Denzel Washington
- Michelle
- Bruce Springsteen
- Aretha Franklin
- U.S.
- Texas
- Martin Luther King Jr.
- Dr. Martin Luther King, King
- Dr. King, King
- Senator John McCain, McCain
- Iraq
- Middle East
- California
- [Mr. Brown, Brown]
- [Bill Clinton, Clinton, Hillary Rodham Clinton]
- [Obama, Mr. Obama, Barack Obama]
- [Joe Biden, Biden]
- Afghanistan
- Israel
- Britain
- [George W. Bush, George W. Bush]
- [Bush, Mr. Bush, Mr. Bush]
- George Bush
I2.2. Path Schema-Based Similarity Search in Information Networks

- **Motivation:** Similarity search in heterogeneous information networks
  - Applications: *Which terror group or event is similar to this one?*

- **Key ideas**
  - New similarity definition for heterogeneous information networks
    - **Feature space**
      - Traditional data: attributes denoted as numerical value/vector, set
      - Networked data: a relation/link sequence—“path schema”
    - **Measure defined on the feature space**
      - PathSim, not cosine, Euclidean distance, Jaccard coefficient, etc.
  - Y. Sun, J. Han, X. Yan, et al. “PathSim: Path Schema-Based Top-K Similarity Search in Heterogeneous Information Networks” (in submission)

- **Experiments:** Flickr pictures using image-tag-group paths: (1) ITI vs. (2) ITIGITI
Search in Image-Rich Information Networks

Images

- FLOWERS
- FLOWER-POWER
- Sunsets

Tags
- red
- flor
- flower
- blosum
- clouds
- sunset
iRIN: Image Retrieval in Image-Rich Information Networks (WWW’10 demo)

Flickr Image Search (query="navy")

Flickr Image Retrieval Demo

Example queries rabbit cat rose moon flower face bird ride night

Results 1 - 10 from 10 for navy

Which algorithm gets the best result for this query? ◁ 1 (Content) ◁ 2 (SimRank) ◁ 3 (SimLearn) Submit
E3.3. Evolutionary Clustering and Analysis of Heterogeneous Bibliographic Networks

Clusters are more consistent.

Table 1: *Consistency versus prior weight*

<table>
<thead>
<tr>
<th>Prior wt</th>
<th>0.0</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author</td>
<td>0.108</td>
<td>2.432</td>
<td>1.199</td>
<td>1.342</td>
<td>2.818</td>
<td>5.273</td>
</tr>
<tr>
<td>Term</td>
<td>0.470</td>
<td>3.105</td>
<td>2.168</td>
<td>2.222</td>
<td>3.646</td>
<td>6.024</td>
</tr>
<tr>
<td>Conf</td>
<td>0.567</td>
<td>2.730</td>
<td>1.800</td>
<td>1.326</td>
<td>3.293</td>
<td>6.709</td>
</tr>
</tbody>
</table>

Table 2: *Quality Variation with Prior Weight*

<table>
<thead>
<tr>
<th>Prior wt Compacted</th>
<th>0.0</th>
<th>0.2</th>
<th>0.4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4594.17</td>
<td>2166.86</td>
<td>1978.39</td>
</tr>
<tr>
<td>Prior wt Compacted</td>
<td>0.6</td>
<td>0.8</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>2932.19</td>
<td>4267.58</td>
<td>3972.62</td>
</tr>
</tbody>
</table>

Dataset

DBLP (1993 to 2008, 654K papers, 484K authors, 107K title terms and 3900 conferences)

Four area (DM, DB, IR, ML papers; 1993 to 2008, 24K papers, 26K authors, 12K title terms, 20 conferences)
Recent Research Publications


(Li, et al., 2010b) Li, Z.; Lee, J.-G.; Li, X.; Han, J., “Incremental Clustering for Trajectories”, *Proc. 2010 Int. Conf. on Database Systems for Advanced Applications (DASFAA’10)*, Tsukuba, Japan, April 2010.


(Gao, et al., 2010) Gao, J. (UIUC); Liang, F. (UIUC); Fan, W. (IBM); Wang, C. (UIUC); Sun, Y. (UIUC); Han, J. (UIUC), “On Community Outliers and their Efficient Detection in Information Networks”, accepted by *KDD’10*

(Wang, et al., 2010) Wang, C. (UIUC); Han, J. (UIUC); Jia, Y. (UIUC); Zhang, D. (UIUC); Yu, Y. (UIUC); Tang, J. (Tsnghua U); Guo, J. (Tsnghua U.), “Mining Advisor-Advisee Relationships from Research Publication Networks”, accepted by *KDD’10*

(Li, et al., 2010d) Li, Z. (UIUC); Ding, B. (UIUC); Han, J. (UIUC); Kays, R. (New York State Museum), “Mining Hidden Periodic Behaviors for Moving Objects”, accepted by *KDD’10*

(Gupta, et al, 2010) Gupta, M.(UIUC), Aggarwal, C. (IBM), Han, J. (UIUC), Sun, Y. (UIUC), "Evolutionary Clustering and Analysis of Heterogeneous Information Networks", submitted.

(Ji, M. et al., 2010) Ji, M. (UIUC); Sun, Y. (UIUC); Danilevsky, M. (UIUC); Gao, J. (UIUC); Han, J. (UIUC), “Graph-based Classification on Heterogeneous Information Networks”, submitted.

(Li, et al., 2010c) Li, Z. (UIUC); Ding, B. (UIUC); Han, J. (UIUC); Kays, R. (New York State Museum), “Swarm: Mining Relaxed Temporal Moving Object Clusters”, submitted for publication, 2010.

(Sun, et al., 2010) Sun, Y. (UIUC); Han, J. (UIUC); Wu, T. (UIUC); Yan, X. (UCSB); Yu, P. S. (UIC); Liu, L. (UIUC), “PathSim: Path SchemaBased TopK Similarity Search in Heterogeneous Information Networks”, submitted.
Research Awards, Tutorials, and Invited Talks

- **Conference tutorial**: Jiawei Han, Zhenhui Li, and Lu An Tang, “Mining Moving Object, Trajectory and Traffic Data”, Conference Tutorial of 2010 Int. Conf. on Database Systems for Advanced Applications (DASFAA’10), Japan, April 2010.

- **Conference tutorial**: Jing Gao, Wei Fan, and Jiawei Han, “On the Power of Ensemble: Supervised and Unsupervised Methods Reconciled”, Conference Tutorial of 2010 SIAM Int. Conf. on Data Mining (SDM’10), Columbus, OH, April 2010.

- **Conference tutorial**: Jiawei Han, Yizhou Sun, Xifeng Yan, and Philip S. Yu, “Mining Knowledge from Databases: An Information Network Analysis Approach” (tutorial), Proc. 2010 ACM SIGMOD Int. Conf. on Management of Data (SIGMOD’10), Indianapolis, Indiana, June 2010.

- **Distinguished seminar**: Jiawei Han, “Mining Heterogeneous Information Networks By Exploring the Power of Links”, Department of Computer Science, Univ. of California at San Diego, Feb. 2010.

- **Distinguished seminar**: Jiawei Han, “RankClus: Integrated Clustering and Ranking at Mining Heterogeneous Information Networks”, Department of Computer Science, Univ. of North Carolina at Charlotte, Feb. 2010.

- **NSF IGERT invited talk**: Jiawei Han, “Mining Moving Object and Traffic Data”, Department of Computer Science, Univ. of Illinois at Chicago, March 2010.

- **Conference keynote speech**: Jiawei Han, ECML/PKDD 2010 (Barcelona, Spain)

- **IBM Ph.D. Fellowship**: Jing Gao (2010-2012)

- **DOE and NSF Ph.D. Scholarship/Fellowship**: Tim Weninger (2009-2011, …)
Summary and Looking Forward

- Data Mining Group has been making good progress and fruitful results on Information Network Research
- We are working on many exciting research frontiers
- Forming APP brainstorming and proposal discussion groups
- Looking forward to collaborations with other groups at UIUC, other sites at INARC, other NS-CTA networks, and ARL
Thank you

May, 2010, Urbana-Champaign