I2: Scalable, Human-Centric Information Network System (IPP)

Information Network Academic Research Center

Center Government Lead: Lance Kaplan (ARL)
Center Academic Lead: Jiawei Han (UIUC)

Project Lead: Xifeng Yan (UCSB)

General Members:
- IBM Research (C. Aggarwal)
- UC-Santa Barbara (X. Yan, A. Singh, Tobias Höllerer)
- UIUC (J. Han)

Sub-awardees
- Palo Alto Research Center (P. Pirolli)

Collaborators
- SCNARC (W. Gray, C. Lin), IRC (M. Endsley, J. Hendler), CNARC (G. Cao)
I2: Scalable, Human-Centric Information Network System

Key Research Questions:
1. How to organize, manage, and query information networks in a scalable manner? and
2. How to analyze and visualize information networks to provide situation awareness to end-users?

Provision / Understanding
(Höllerer, Pirolli, Yan)

Analytical Processing
(Yan, Han)

Search / Query
(Singh, Yan, Aggarwal)

Human Centric Design

Online Analytics

Info. Net. Management

Comm. Net. Devices

Scale
Key Research Questions:

- How to organize provenance metadata to enable robust, multi-resolution computation of QoI?
- How to develop a network language that enables flexible network information access?
- How to index large graphs in order to support scalable access methods?
I2.1: Information Network Management

Key Technical Innovations

- Graph query language to composite networks with temporal dimension
- Graph indexing for processing graph queries in large-scale social, information, and communication networks
- Stochastic information dissemination models and Bayes probabilistic models for information dissemination

Highlights:

1. Scalable Discovery of Best Clusters on Large Networks (Macropol and Singh, VLDB 2010)
2. Differential Graph Index (Yan et al. ICDE’2010)
3. Graph Density Index (Nan and Yan, submitted to ICDE 2011)
Scalable Discovery of Best Clusters on Large Graphs

Time vs. Graph Size on Generated Graphs.

Top 20 Cluster Scores on Yeast Dataset.

Table 3: Summary of Experiments on Large Real World Networks

<table>
<thead>
<tr>
<th>Graph</th>
<th>Nodes</th>
<th>Edges</th>
<th>Time (sec)</th>
<th>Avg. Score</th>
<th>Avg. Size</th>
<th>Time (sec)</th>
<th>Avg. Score</th>
<th>Avg. Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>LiveJournal</td>
<td>4,847,571</td>
<td>68,993,773</td>
<td>199</td>
<td>3.1</td>
<td>13.9</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>2,394,385</td>
<td>5,021,410</td>
<td>37.3</td>
<td>1.4</td>
<td>6.6</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Webgraph</td>
<td>685,230</td>
<td>7,600,595</td>
<td>28.6</td>
<td>2.6</td>
<td>12</td>
<td>102</td>
<td>1.5</td>
<td>9</td>
</tr>
<tr>
<td>Amazon</td>
<td>410,236</td>
<td>3,356,824</td>
<td>22.2</td>
<td>2.7</td>
<td>9.1</td>
<td>71.8</td>
<td>2.8</td>
<td>13</td>
</tr>
<tr>
<td>Slashdot</td>
<td>82,168</td>
<td>948,464</td>
<td>11.0</td>
<td>2.0</td>
<td>7.7</td>
<td>19.9</td>
<td>0.78</td>
<td>6</td>
</tr>
</tbody>
</table>

Singh's Group
Distributed Graph Computing

- Searching and mining large graphs is critical to scientific discovery in complex networks.
- Graph operations tend to crawl across many links, resulting in extremely large memory footprints.
- Scalable query processing on massive graphs requires careful partitioning and distributing graphs across clusters.

Recent Progress

We are building an experimental prototype and are testing various kinds of partitioning algorithms.
Key Research Questions:

- How to perform multi-dimensional analytics that allows a non-expert to explore information networks in real time?
- What are new fundamental operators that could facilitate network-wise graph modeling and mining?
Key Research Questions:

- How to perform multi-dimensional analytics that allows a non-expert to explore information networks in real time?
- What are new fundamental operators that could facilitate network-wise graph modeling and mining?
I2.2: Info. Network Online Analytical Processing (Highlights)

Key Technical Innovations
- OLAP framework in network space: informational OLAP and topological OLAP
- Top-k graph aggregation over large networks, facilitating local structure search
- Path schema-based similarity search in heterogeneous information networks

Highlights:
1. Shortest Path Decomposition of Vertices (Zhao and Han, VLDB 2010)
2. Top-K Aggregation Over Large Networks (Yan et al. ICDE’2010)
3. Node similarity ranking model (Sun et al. 2010, submitted to ICDE 2011)
4. New information network OLAP operators: graph alignment (Khan and Yan 2010, submitted to ICDE 2011)
Neighborhood signatures of vertices are built to maintain indexing features: Effective search space pruning ability.

Processing (Query Decomposition): Decompose the query graph into a set of indexed shortest paths in Spath (VLDB’10)
Semantics-Based Similarity Search in InfoNet

Search top-k similar objects of the same type for a query

- Applications: *Which terror group or event is similar to this one?*

Two critical concepts to define a similarity

- Feature space
  - Traditional data: attributes denoted as numerical value/vector, set, etc.
  - Network data: a relation sequence called "**meta-path**"
    - Existing homogeneous network-based similarity does not deal with this problem

- Measure defined on the feature space
  - Cosine, Euclidean distance, Jaccard coefficient, etc.
  - **PathSim**

- INARC
Meta-Path for DBLP Queries

Meta-path: A path of InfoNet schema, e.g., APC, APA

Who are most similar to Christos Faloutsos?

(a) Path: APA

<table>
<thead>
<tr>
<th>Rank</th>
<th>Author</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Christos Faloutsos</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Spiros Papadimitriou</td>
<td>0.127</td>
</tr>
<tr>
<td>3</td>
<td>Jimeng Sun</td>
<td>0.12</td>
</tr>
<tr>
<td>4</td>
<td>Jia-Yu Pan</td>
<td>0.114</td>
</tr>
<tr>
<td>5</td>
<td>Agra J. M. Trina</td>
<td>0.110</td>
</tr>
<tr>
<td>6</td>
<td>Jure Leskovec</td>
<td>0.096</td>
</tr>
<tr>
<td>7</td>
<td>Caetano Traina Jr.</td>
<td>0.096</td>
</tr>
<tr>
<td>8</td>
<td>Hanghang Tong</td>
<td>0.091</td>
</tr>
<tr>
<td>9</td>
<td>Deepayan Chakrabarti</td>
<td>0.083</td>
</tr>
<tr>
<td>10</td>
<td>Flip Korn</td>
<td>0.053</td>
</tr>
</tbody>
</table>

(b) Path: APCPA

<table>
<thead>
<tr>
<th>Rank</th>
<th>Author</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Christos Faloutsos</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Jiawei Han</td>
<td>0.842</td>
</tr>
<tr>
<td>3</td>
<td>Rakesh Agrawal</td>
<td>0.838</td>
</tr>
<tr>
<td>4</td>
<td>Jian Pei</td>
<td>0.8</td>
</tr>
<tr>
<td>5</td>
<td>Charu C. Aggarwal</td>
<td>0.739</td>
</tr>
<tr>
<td>6</td>
<td>H. V. Jagadish</td>
<td>0.705</td>
</tr>
<tr>
<td>7</td>
<td>Raghu Ramakrishnan</td>
<td>0.697</td>
</tr>
<tr>
<td>8</td>
<td>Nick Koudas</td>
<td>0.689</td>
</tr>
<tr>
<td>9</td>
<td>Surajit Chaudhuri</td>
<td>0.677</td>
</tr>
<tr>
<td>10</td>
<td>Divesh Srivastava</td>
<td>0.661</td>
</tr>
</tbody>
</table>

(c) Path: APTPA

<table>
<thead>
<tr>
<th>Rank</th>
<th>Author</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Christos Faloutsos</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Jian Pei</td>
<td>0.661</td>
</tr>
<tr>
<td>3</td>
<td>Srinivasan Parthasarathy</td>
<td>0.600</td>
</tr>
<tr>
<td>4</td>
<td>Jeffrey Xu Yu</td>
<td>0.587</td>
</tr>
<tr>
<td>5</td>
<td>Ming-Syan Chen</td>
<td>0.579</td>
</tr>
<tr>
<td>6</td>
<td>Jian Pei</td>
<td>0.576</td>
</tr>
<tr>
<td>7</td>
<td>Mohammed Javeed Zaki</td>
<td>0.571</td>
</tr>
<tr>
<td>8</td>
<td>Hans-Peter Krieger</td>
<td>0.563</td>
</tr>
<tr>
<td>9</td>
<td>Yannis Manolopoulos</td>
<td>0.548</td>
</tr>
<tr>
<td>10</td>
<td>Rakesh Agrawal</td>
<td>0.545</td>
</tr>
</tbody>
</table>
Flickr: Which Pictures Are Most Similar?

Some path schema leads to similarity closer to human intuition
But some others are not

Figure 5: Top-6 images in Flickr network under path schema ITI

Figure 6: Top-6 images in Flickr network under path schema ITIGITI
I2.3: Information Network Visualization

Key Research Questions:

• How to best visualize (potentially large) information networks for Army personnel in different user situations?

• What are the principles behind effective information network dissemination on various interaction platforms?

• How is situation awareness for information networks achieved in different user contexts?
I2.3: Information Network Visualization

Key Technical Innovations

- Scalable visualization techniques of large node-link graphs covering information, social and communication networks in a web browser
- Novel network visualization methods that improve situational awareness
- Interactive topic visualization for information networks
- Wikipedia Dashboards: Supporting Credibility and Semantic Feedback

Highlights:

1. **WiGis**, Graph Vis framework (Gretarsson et al. 2010, GraphDrawing ‘09)
3. **TopicNets**, visual analysis of large text corpora using information network (Gretarsson et al. 2010, under review)

INARC
Platform Evaluations

Mobile Platforms
Immersive Situation Room
Desktop / Networked Collaboration
Platform Evaluations

Mobile Platforms
Immersive Situation Room
Desktop / Networked Collaboration
Platform Evaluations

Mobile Platforms
Immersive Situation Room
Desktop / Networked Collaboration
“Mold” a graph (here: Facebook profiles) through very simple interactions into meaningful layouts

Visual Editing of Networks (here: Wikipedia Data)
Exploring Network Data through Interaction

- **Wikipedia**: structured data (infoboxes, categories, etc.)
- **DBpedia**: structured semantic data (RDF) allows SQL-like querying
- **WiGipedia**: visual interface allows single-click Wiki edits

Höllerer's Group
Exposure to the editing/authoring histories of Wikipedia increases credibility judgments

Information Network Academic Research Center

Center Government Lead: Lance Kaplan (ARL)
Center Academic Lead: Jiawei Han (UIUC)

Project Lead: Xifeng Yan (UCSB)

General Members:

IBM Research (A. Iyengar, M. Srivatsa, C. Aggarwal)
UC-Santa Barbara (X. Yan, A. Singh)
UIUC (T. Abdelzaher)

Sub-awardees

Carnegie Mellon Univ. (C. Faloutsos)

Collaborators

CNARC (G. Cao, S. Zhu, P. Mahopatra), IRC (V. Kawadia, J. Hendler, P. Basu), SCNARC (Z. Wen)
APP I2: Storage and Processing of Information Network Systems

Key Research Questions:
How to achieve the right balance between availability and efficiency when replicating information over tactical composite networks?

Key Technical Innovations
- Quality-aware dissemination and caching algorithms that understand the mobility models of users and the data models to satisfy user’s information needs.

- Novel cache replacement policies based on maximizing the diversity of information stored in a network cache.
**APP I2.1: In-Network Storage**

**Key Objectives:**
Develop in-network storage techniques to enhance the availability and efficiency of information retrieval in tactical networks.

<table>
<thead>
<tr>
<th>Role</th>
<th>Researchers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lead</td>
<td>M. Srivatsa, IBM (INARC)</td>
</tr>
<tr>
<td>Primary</td>
<td>T. Abdelzaher, UIUC (INARC)</td>
</tr>
<tr>
<td></td>
<td>A. Iyengar, IBM (INARC)</td>
</tr>
<tr>
<td></td>
<td>X. Yan, UCSB (INARC)</td>
</tr>
<tr>
<td>Collab</td>
<td>G. Cao, PSU (CNARC) (joint with C2.1)</td>
</tr>
<tr>
<td></td>
<td>P. Mohapatra, UCDavis (CNARC) (linked w. T1.3)</td>
</tr>
<tr>
<td></td>
<td>V. Kawadia, BBN (IRC) (linked with T2.3)</td>
</tr>
<tr>
<td></td>
<td>S. Zhu, PSU (CNARC) (linked with T2.3)</td>
</tr>
</tbody>
</table>
Overview

- Information must be replicated, cached for both performance and high availability
- Disconnections are common
  - Caching becomes essential for data dissemination
    Achieving information dissemination in presence of disconnections
- Caching is critically important for:
  - Improving performance
  - Achieving data dissemination
- Data Dissemination
  - Model, analyze and quantify the impact of rich data
  - Query Performance Improvement
  - Performance bound of query response time with in-network caching
  - models on data dissemination schemes
Key Research Topics

- Efficient cache replacement algorithms
- Query processing and caching
  - Queries for provenance information
  - More general queries
- Cache Diversity – Maximizing total information stored in the network
  - Determine degree of redundancy between different data to reduce redundancy of cached content
  - Reducing redundancy among cached objects results in better cache space utilization, reduction in data transmissions for updates
Disruption Tolerant Networks (DTNs)

Consist of handheld mobile devices:
- Laptops, Smartphones, PDAs
- Communicate in an ad-hoc manner

Opportunistic and intermittent network connectivity
- Low node density
- Unpredictable node mobility
Node mobility is exploited
Nodes temporarily hold the data if no network connectivity is available
  - Nodes are used as “relays”
Node make data forwarding decision when contact with each other
Carry-and-Forward in DTNs

An example
APP I2.2: Large-Scale Information Network Processing

Key Technical Innovations

- Dynamic graph indexing structures for time-varying information networks
- Graph partitioning algorithms (overlapping, caching, and re-balancing) to (1) improve locality of data, and (2) accommodate network data updates and query workload changes

<table>
<thead>
<tr>
<th>Role</th>
<th>Researchers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lead</td>
<td>Xifeng Yan, UCSB (INARC)</td>
</tr>
<tr>
<td>Primary</td>
<td>Ambuj Singh, UCSB (INARC) Charu Aggarwal, IBM (INARC) Christos Faloutsos, CMU (INARC)</td>
</tr>
<tr>
<td>Collab</td>
<td>Zhen Wen, IBM (SCNARC) (linked with S1.1) Jim Hendler, RPI (IRC) (linked with R2.3) Prithwish Basu, BBN (IRC) (linked with E1.2)</td>
</tr>
</tbody>
</table>
APP I2.2: Large-Scale Information Network Processing

Key Research Questions:

1. What is the performance limit of using graph index for time-varying information networks?

2. Can distributed graph computing significantly improve the processing of information networks, in which situation?