An Introduction to INARC Research Program

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

Prime Member: UIUC (Jiawei Han [I3, I3.1, I3.2], Tarek Abdelzaher [I1, I1.1], Thomas Huang [I1.2], Dan Roth [I3.3])

General Members:

- IBM Research (Dakshi Agrawal [T1], Charu Aggarwal [I1, I1.3], Anastasios Kementsietsidis [T1.2], Spyridon Papadimitrio [I3.2], Mudhakar Srivatsa)
- UC-Santa Barbara (Ambuj Singh [E3, I2.1], Xifeng Yan [I2, I2.1, I2.2], Tobias Hollerer [I2.3], B. S. Manjunath)
- CUNY (City Univ. of New York): (Ted Brown [Educ], Amotz Bar-Noy, Heng Ji [I1.3])

Sub-awardees

- Carnegie Mellon Univ.: Christos Faloutsos
- Univ. of Michigan: Lada Adamic
- Northwestern Univ.: Noshir Contractor
- Palo Alto Research Center (PARC): Peter Pirolli [T1.2]
Modern warfare and other military missions demand superior mastery of information.

Information as well as the underlying data is massive, interrelated, and forms gigantic information networks.

Military networks are also heterogeneous, resource-constrained, interacting and rapidly evolving, thereby posing great challenges for effective analysis, operation, and adaptation of such networks.

The challenges must not be addressed in isolation; they demand a multi-disciplinary approach that breaks new grounds and builds upon the research in communication, social, and cognitive networks.

Key tasks for information network-oriented research:
- dynamics/evolution of information networks
- Trustworthiness and quality in information networks
- Data integration/information fusion
- Construction and access of information networks
- Knowledge discovery in information networks
Two CCRI Projects and three INARC dedicated projects
Collaborations with Other Centers and ARL Researchers

Objectives

- Collaborations with the other centers and ARL researchers for better understanding and exploration of information networks
- Education and training across multiple centers and learn application demands, scenarios, and advanced network technology

Plan for collaborations

- Develop concrete collaboration plan with other centers and ARL for each project — Cross-center linkage shown in the research plans
- Develop education and training plan with other centers
- Actively contribute to NS-CTA website, information exchange, data sharing, and collaborative publications
- Regular mutual visits and research collaborations with other centers and ARL
- IBM summer interns for students working with cross-center research teams
Project I1: Distributed & Real-time Data Integration and Information Fusion

**Project Leads:**

<table>
<thead>
<tr>
<th>IBM</th>
<th>Charu Aggarwal (IBM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Tarek Abdelzaher (UIUC)</td>
</tr>
</tbody>
</table>

**Primary Research Staff:**

- Abdelzaher, Tarek (UIUC)
- Aggarwal, Charu (IBM)
- Bar-Noy, Amotz (CUNY)
- Höllerer, Tobias (UCSB)
- Huang, Thomas (UIUC)
- Ji, Heng (CUNY)
- Manjunath, B. S. (UCSB)
- Papadimitriou, Spyridon (IBM)
- Roth, Dan (UIUC)
- Singh, Ambuj (UCSB)

**Collaborators:**

Govindan, Ramesh (USC)
Lin, Ching-Yung (IBM)
Pentland, Alex (MIT)
Wen, Zhen (IBM)
Project I1: Goals and Vision

- **Two general themes**
  - Optimizing the quality of information
  - Quantifying uncertainty as a first class attribute
Project I1: Project Goals and Vision

Goal: Integrate, Extract, and Fuse Data, and Estimate Uncertainties in Fusion

- Create a semantic information network through data integration and fusion from a wide variety of sources (CNARC links)
- Exploit the links in the network during the fusion process and use the fusion process to derive linkages (including social linkages => SCNARC links)
- Develop models for information quality and uncertainty to drive data collection and fusion (CNARC links)

Tasks

- Task 1: QoI-aware Signal Data Fusion
  - T. Abdelzaher (lead), C. Aggarwal, A. Barnoy, S. Papadimitriou
- Task 2: Human and Visual Data Fusion
  - T. Huang (lead), T. Hollerer, H. Ji, B. S. Manjunath
- Task 3: Modeling Uncertainty in Data Fusion
  - H. Ji (co-lead), C. Aggarwal (co-lead), T. Huang, D. Roth, A. Singh
Project I1: Key Research Questions and Hypothesis

- **Overall Research Questions**
  - How to infer information objects and linkages from widely distributed and diverse data sources in order to build an information network?
  - How to quantify uncertainty regarding state of such objects and linkages?
  - How to utilize object linkages to optimize data collection/fusion and reduce the uncertainty?
  - How to address the above problems in a way that optimizes the quality of information?

- **Hypothesis**

  ![Diagram](Diagram.png)

  - Data collection → Data fusion → Information Network
  - Significantly enhanced

  Higher QoI, Less uncertainty

  Exploit **linkages**

  Information Network

INARC
Task I1.1: QoI-Aware Signal Data Fusion

- **Challenges and Goals**
  - How to fuse the sensor feeds from physical sources in a manner cognizant of quality of *information needs, resource constraints, and data uncertainty*?
  - How to utilize information network feedback?
  - How to analyze timeliness of real-time information fusion?

- **Recent Progress**
  - Modeling/Prediction from Sparse, High-dimensional Data (UIUC)
    - Identify *categories* of cases, determine predictors, quantify uncertainty
  - Feedback Mechanisms to Sensor Networks (UIUC, CNARC)
    - Uncertainty-aware congestion control & transmission reliability
  - Data Linkage Exploitation (IBM, CUNY)
    - Exploit data linkages to minimize sensing cost

---

INARC
Task 11.1: QoI-Aware Signal Data Fusion

<table>
<thead>
<tr>
<th></th>
<th>Census Wages</th>
<th>Social Security</th>
<th>Property Values</th>
<th>Residential Energy</th>
<th>Vehicular Mileage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Model</td>
<td>6.73%</td>
<td>17.01%</td>
<td>22.13%</td>
<td>10.02%</td>
<td>12.27%</td>
</tr>
<tr>
<td>Reg. Cube</td>
<td>6.59%</td>
<td>21.12%</td>
<td>23.76%</td>
<td>8.01%</td>
<td>24.56%</td>
</tr>
<tr>
<td>Samp. Cube</td>
<td>77.48%</td>
<td>183.2%</td>
<td>81.42%</td>
<td>34.17%</td>
<td>43.11%</td>
</tr>
<tr>
<td>Our Method</td>
<td>3.4%</td>
<td>12.45%</td>
<td>11.15%</td>
<td>0.002%</td>
<td>2.22%</td>
</tr>
</tbody>
</table>

Better prediction: (UIUC, 2010)
Sampling regression cube analysis

Physical Models
- Lower energy
- Higher accuracy
(UUIC, ICCPS 2010)

Feedback on QoI/Uncertainty

Unattended ground sensors
Data collection and fusion

Physical Models
Task I1.2: Human and Visual Data Fusion

☑ Challenges and Goals

☑ Large body of research is devoted to analysis and interpretation of homogeneous data (e.g., documents only, images only)

☑ In an information network, the data are collected from multiple sources and are naturally heterogeneous

☑ Monitoring and fusing large amounts of heterogeneous data is critical for maintaining situational awareness on and off the battlefield

☑ A need exists to filter and summarize heterogeneous data

☑ A robust information fusion architecture can be realized with the use of explicit (e.g., hyperlinks) and virtual (e.g., image similarity) linkages

☑ “Information overload is an important problem” Captain Scott Shaffer, NS CTA Seminar, February 2010
Task I1.2: Human and Visual Data Fusion

- **Recent Progress:** Interpreting surveillance video data in an information network setting (UIUC, UCSB)
  - Video cameras make observations on locations, people and events
  - Human observers generate reports on events

**Inference On Network Parameters**

<table>
<thead>
<tr>
<th>Traffic Flow and Motion Patterns</th>
<th>Social and Hierarchical Organization</th>
<th>Normal and Abnormal Events</th>
</tr>
</thead>
</table>

- **Dataset:** UCSB Campus Camera Network
  - 27 indoor cameras deployed in Harold Frank Hall
  - Currently there are 40 fixed outdoor cameras deployed along the bike paths on the campus
Task I1.3: Modeling Uncertainty in Data Fusion

- **Challenges**
  - Data can be noisy, incorrect, or misleading
  - Unstructured data is difficult to interpret
  - Difficult to assure accuracy or even coherence among the data sources: The fusion process may itself cause data uncertainties

- **Goals**
  - Quantify uncertainty regarding the state of objects and linkages
  - Utilize object linkages to reduce the uncertainty in inference
  - Model the uncertainty resulting from heterogeneous data fusion
  - Resolve linkage challenges
    - Co-reference
    - Noisy, incorrect, or misleading data
    - Unstructured data
  - Understand how the fusion process may itself cause data uncertainties

- **Recent Progress:** New models and inference methods, ...
Motivation

- Media processing is typically organized as a pipeline; errors accumulate
- New: methods that make multiple decisions jointly using global information

Method

- Modeling Joint inference over the entire information network with feedback from other centers
I2: Scalable, Human-Centric Information Network System

Project Lead: Xifeng Yan (UCSB)

Primary Research Staff:
- Aggarwal Charu (IBM)
- Han Jiawei (UIUC)
- Höllerer Tobias (UCSB)
- Singh Ambuj (UCSB)
- Pirolli Peter (PARC)
- Yan Xifeng (UCSB)

Collaborators:
- Wayne Gray, RPI/SCNARC
- M. Endsley, SA Technologies/IRC
- Ching-Yung Lin, IBM/SCNARC
- Guohong Cao, PSU/CNARC
- James Hendler, RPI/IRC
**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?

**Visualization / Understanding**  
(Höllerer, Pirolli, Yan)

**Analytical Processing**  
(Yan, Han)

**Operational Information Network**  
(Aggarwal, Singh, Yan)

**Communication Networks, Devices**  
Scale

**Human-Centric Design**

**Online Analytical Processing**

**Information Network Management**
Task I2.1: Information Network Organization and Management

**Motivation**
- How to develop a network language that enables flexible network information access?
- How to index large graphs in order to support scalable access methods?
- How to organize information for fast dissemination?

**Goals: Operational Information Network**
- Graph Query Language; Graph Index; and Dissemination Strategy

**Recent Research Progress**
- Novel Network Queries
  - Top-k Graph Clusters in Large Networks, Proximity Graph Search
- Differential Graph Index (ICDE’2010, Collaborate with UIUC)
- Distributed Graph Computing Environment
- Bayes Probabilistic Model for Information Dissemination
Top-k Graph Clusters in Large Networks

- **Answering Top-k clusters queries in graphs**
  - Current graph clustering algorithms scale poorly
    - MCL can take $O(n^2)$ running time
    - Graclus and MLR-MCL take memory of $O(n)$
  - Idea: Prune search space by finding *best* clusters only.
    - Still relevant to many applications
    - Allows for greater speed and less memory usage

- **Top Graph Clusters (TopGC)**
  - Modify Locality Sensitive Hashing (LSH) to find groups of nodes with similar graph neighborhoods in linear time
    - Nodes with similar neighbors should belong to same cluster
    - Edge weights may be added by interpreting weight as probability of node being included in neighborhood
  - Split algorithm into two separate phases, pruning and clustering, to reduce memory usage
Scalable Discovery of Best Clusters on Large Graphs

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>TopGC Avg. Score</th>
<th>TopGC Avg. Size</th>
<th>MLR-MCL Avg. Score</th>
<th>MLR-MCL 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>
</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>
</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>
</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>
</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>
</tr>
</tbody>
</table>
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.
Task I2.2: Information Network Online Analytical Processing

Motivation

- 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?

Goals:

- Multiple Online Analytical Operators for Information Networks

Recent Research Progress

- OLAP Framework: Informational OLAP and Topological OLAP
- Top-K Aggregation Over Large Networks (ICDE’2010)
- Path Schema-Based Similarity Search in Information Networks (in submission)
**Top-K Aggregation Over Large Networks**

**Node Relevance, \( f(u) \):** For a given query, \( f(u) \) could be as simple as 0/1, e.g., if a user recommends a movie or not, or it can be a classification function, e.g., how likely a user is a data mining expert.

**Aggregation Function, \( F(u) \):** Evaluate the collective strength of a node \( u \)'s \( h \)-hop neighbors—\( F(u) \) can be as simple as SUM and AVG. It can also be as complicated as a non-linear function, e.g., learned by a collective classification method.

**Top-k Mining:** Find top-k nodes with the highest aggregate scores.
Path Schema-Based Similarity Search in Information Networks

- **Motivation:** Similarity search in heterogeneous information networks
  - Applications: \textit{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

---

**Path: ITI**

(a) top-1  (b) top-2  (c) top-3

(d) top-4  (e) top-5  (f) top-6

**Path: ITIGITI**

(a) top-1  (b) top-2  (c) top-3

(d) top-4  (e) top-5  (f) top-6
Motivation

- 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?

Goals:

- Situation-Aware, Adaptive Information Network Visualization

Recent Research Progress

- Visualization and Interactive Manipulation of Heterogeneous Information Networks
- Exploring Network Data through Interaction
Abundance of different types of network data:

Intelligence Data, Social Web, Biological Networks, Financial Networks, Interlinked Publication Databases, Tele Communications, ...

Real-time visual interaction and dynamic probing are powerful tools for data analysis

Our approach

☐ Make interaction feasible for *large-scale data*
☐ Use interaction to predictably explore the data
☐ Make the tool easily accessible & embeddable
Exploring Network Data through Interaction

“Mold” a graph (here: Facebook profiles) through very simple interactions into meaningful layouts

Visual Editing of Networks (here: Wikipedia Data)
Project I3: Knowledge Discovery in Information Networks

Project Lead: Jiawei Han (UIUC)

Primary Research Staff:
- Faloutsos, Christos (CMU)
- Han, Jiawei (UIUC)
- Ji, Heng (CUNY)
- Papadimitriou, Spyridon (IBM)
- Roth, Dan (UIUC)
- Yan, Xifeng (UCSB)

Collaborators:
- M. Faloutsos (UCR, IRC)
- J. Hendler (RPI, IRC)
- B. Szymanski (RPI, SCNARC)
- S. Adali (RPI, SCNARC)
- T. LaPorta (PSU, CNARC)
Project I3: Knowledge Discovery in Information Networks

- **Goal**
  - Develop efficient and effective knowledge discovery mechanisms that automatically uncover patterns, knowledge, and other critical information in information networks for military applications

- **Organization of Project Tasks**
  - Task I1: Methods for scalable mining of dynamic, heterogeneous information networks (*J. Han*, X. Yan, and C. Faloutsos) [Collaborators: M. Faloutsos (UCR, IRC), J. Hendler (RPI, IRC) and B. Szymanski (RPI, SCNARC)]
  - Task I2: Real-time methods for mining spatiotemporal information-related cyber-physical networks (*S. Papadimitriou*, *J. Han*, and X. Yan) [Collaborators: S. Adali (RPI, SCNARC), T. LaPorta (PSU, CNARC)]
  - Task I3: Text and unstructured data mining for information network analysis (*D. Roth*, J. Han and H. Ji) [Collaborators: M. Faloutsos (UCR, IRC) and B. Szymanski (RPI, SCNARC)]

INARC  

Project Lead: Jiawei Han (UIUC)
Task I3.1: Mining Dynamic, Heterogeneous Information Networks

- **IPP Major Research Tasks and New Research Progress**
  - Classification of heterogeneous information networks
    - Recent progress: RankClus, NetClus, iNextCube
    - New work: GRaphClass (KDD’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 sub)
    - Evolution of heterogeneous information networks: KDD’10 sub
    - Graph association patterns in information networks (SIGMOD’10, UCSB & IBM)
  - Outlier detection for dynamic heterogeneous information networks
    - Anomaly discovery in large information networks (PAKDD’10, ICDM’09 award paper: Christos Faloutsos)
    - New outlier detection method in large databases

INARC

Han (UIUC), Yan (UCSB), Faloutsos (CMU)
Integrating Clustering with Ranking for Heterogeneous Information Network Analysis

Global Ranking? vs.

Clustering in heterogeneous network?

What feature can I use?

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

Flickr: Query “Raleigh”

DBLP

<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>FOCS</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>EUROCRYPT</td>
</tr>
<tr>
<td>10</td>
<td>SDM</td>
<td>SenSys</td>
<td>CP</td>
<td>CIVR</td>
</tr>
</tbody>
</table>

Han’s Group (UIUC)
OddBall: Spotting Anomalies in Weighted Graphs

Leman Akoglu, Mary McGlohon, Christos Faloutsos
To appear in PAKDD 2010, Hyderabad, India

Near-Clique/Star

ENRON

1.3581x + (0.10897) = y
1.0438x + (-0.10446) = y
2.0438x + (-0.40549) = y

Example: GIM-V At Work
Connected Components

Generalized Iterated Matrix Vector Multiplication (GIMV)

U Kang, Charalampos E. Tsourakakis, and Christos Faloutsos
ICDM 2009 (Dec.'09), Miami, Florida, USA.
Best Application Paper (runner-up)
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 (sub)
- 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

Han’s Group (UIUC)
Motivation and Goals

- Most military applications need to handle text data, including documents, emails, 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:
- Text information extraction: Wikification (Linking entities to Wikipedia)
- Google News analyzer: Integrate text mining and information network analysis

INARC

Roth (UIUC), Han (UIUC), Heng Ji (CUNY)
I3.3: Recent Progress on Dynamic Language Modeling & Text Mining

- Knowledge Discovery from Web-Scale N-Grams for Information Extraction (Ji and Lin, Journal of LRE 2010; Lin et al., LREC 2010; Ji and Lin, PACLIC 2009)
- Query-time dynamic language model adaptation (Ji et al., MMIES book chapter)
- Domain Adaptation for Mining Scientific Literatures (Ji et al., LREC 2010)
- Cross-lingual Predicate Cluster Acquisition to Improve Bilingual Event Extraction by Inductive Learning (Ji, Machine Learning Symposium 2009)
- Graph-clustering methods for Information Extraction and Class Acquisition (Zheng and Ji, ACL2010 submission)
- Multi-document Abstractive Summarization based on Information Extraction (Ji et al., COLING 2010 submission)
- Cross-document Cross-media Knowledge Base Population: Entity linking and Slot Filling in a joint inference framework (Ji coordinating the international competition)
Consider a networked knowledge base of text documents (e.g., Wikipedia).

Given a new text document (e.g., a blog) Wikification is the process of identifying important mentions in the blog, context sensitive disambiguation of them and linking them to the corresponding Knowledge Base entry (e.g., Wikipedia page).

An American official announced that American President Bill Clinton met his Russian counterpart, Vladimir Putin, today. The president said that Russia was a great country.
Google News Analyzer: Proposed Goals and Research Outline

- **Motivation:** Military missions need to analyze a lot of textual and unstructured data

- **Google News Analyzer**
  - Integration of (1) information retrieval, (2) natural language processing, (3) multidimensional text cube, (4) information network analysis, and (5) text information integration and information quality

- **Recent progress**
  - Text/topic cube construction and usage
  - TopCell: Query-based ranking of text cube (ICDE’10)
  - MiTexCube: multidimensional text clustering

- **Research Plan:**
  - Construct a Google News Analyzer for automated online multidimensional news analysis

---

INARC

Han’s Group (UIUC)
CCRI EDIN Projects: Evolving Dynamic Integrated Networks

- Project E1: Ontology and Shared Metrics for Composite Networks
- Project E2: Mathematical Modeling of Composite Networks
- Project E3: Structure and Co-evolution of Composite Networks
- Project E4: Modeling Mobility and Its Impact on Composite Networks

Progress
- Ongoing teleconferences
- Research collaborations
- San Diego meeting on March 16

INARC Participants
- Singh, Adamic, Aggarwal, Contractor, Faloutsos, Han, Yan
How to Find Most Trusted Colleagues in Composite Networks?

Derive lower & upper bounds on affinity
Expand network selectively using bounds
Experimental results show that less than 0.4% of network explored during searches
Can define the composite network dynamically
Extending to richer class of mappings between component networks (future)

EDIN E2.1: Singh (UCSB)
Synopsis Construction of Massive Networks

- **Motivation**
  - Massive networks need to be queried, mined, and analyzed for a variety of purposes.
  - In many cases, dynamic changes may happen to the network, which may evolve over time.
  - Need synopsis which requires less space, and can evolve with the changes in the underlying network.

- **Core Approach**
  - Design a synopsis which can be used for approximate mining and querying purposes.
  - Use a hash-based compression in conjunction with query-processing on the compressed structure.
  - Tested approach on a number of real social networks and bibliographic networks.

---

**INARC**

**Task E2.1: Aggarwal (IBM) & Yan (UCSB)**
Co-evolution of Communication, Social, and Economic Networks

- **Findings**
  - Sellers utilize groups, social ties, and chat to maintain and acquire customer relationships

- **Next steps**
  - Time series analysis to trace co-evolution of trading and chat networks

Evolution of sellers’ customer networks

- Eytan Bakshy, Matthew Simmons, David Huffaker, Chun-Yuen Teng, Lada Adamic, "The Social Dynamics of Economic Activity in a Virtual World", ICWSM 2010

---

INARC

Task E3.3: Adamic (Michigan)
CCRI TRUST Projects

- Project T1: Trust Models and Metrics: Computation, Inference and Aggregation
  - T1.1: Unified Models and Metrics of Trust
  - T1.2a: Information Credibility: Definitions, Assessments, and Fundamental Trade-offs
  - T1.2b: Models and Maintenance of Provenance Metadata for Enhancing Trust in Information
  - T1.3: Cognitive Trust Models

- Project T2: Understanding the Interactions between Network Characteristics and Trust
  - INARC participation deferred until next year

- Project T3: Fundamental Paradigms for Enhancing Trust
  - T3.1: Trust Establishment via Distributed Oracles

INARC Participants
- PIs: Abdelzaher, Agrawal, Brown, Han, Hollerer, Iyengar, Ji, Kementsietsidis, Pirolli, Roth, Srivatsa, and Yan
Subtask T1.2a: Information Credibility

Motivation:
In a dynamic and volatile battlefield, sensors or personnel may not be 100% credible – information fusion may further introduce uncertainty: for trust-based decision-making, an accurate assessment of the credibility of information is a pre-requisite.

Methods proposed and being studied:
- Exploit network structure for:
  - Cross-checking & mutually consolidating info from multiple sources
  - Resolving ambiguities and conflicts automatically and efficiently in a dynamic situation
- Notion of credibility in a multidimensional space: Associated with multiple conditions, e.g., time, location, ...., and need cross-environment analysis
  - Credibility of a sensor or information source will be highly linked with the context in a heterogeneous information network
- Incremental updates of current information network and subsequent revision of credibility assessments
- Use background knowledge and guard against malicious information sources
- Collaboration with CNARC for taking communication network information
- Collaboration with SCNARC for assessing credibility of human sources
Subtask T1.2b: Information Provenance

Motivation:
Provenance = “To Come From"
Identifying information’s source and lineage is vital for trust-based decision making

Methods proposed and being studied:
- Provenance tracking algorithms (IBM, UIUC)
  - Use annotation propagation and query rewriting to track info provenance
  - Use annotations to encode security metadata (e.g., roles, clearance, etc.)
- Disseminating provenance metadata on an unreliable communication network (w/CNARC)
  - Exploit correlations (based on provenance graph) between accesses to provenance metadata to build novel data placement (caching) and dissemination protocols
- Empirically evaluate the impact of provenance data on trust-based decision making (w/SCNARC)
  - Use enterprise social network data to quantify the impact of provenance data on individual effectiveness
Subtask T1.3: Cognitive Trust Models

Goal:
Development of a cognitive model of credibility judgments as a result of human-information and human-human interaction in information networks and social networks that are similar to those that occur in military operations.

Key Highlights:
- Our hypothesis is that cognitive models of trust can be developed on a foundation of Bayesian cognitive models of categorization, judgment, and reasoning
- Psychological experiments are being designed to test the model using novel visualizations and interaction techniques
- Experimental tasks will be designed to be reflective of the structure and dynamics of real information and social networks (collaboration with SCNARC)
- Validation of Twitter dataset – short messages regarding evolving situations – we are building and studying large-scale Twitter data sets to develop the visualization technologies and the experimental tasks.
- Close collaboration with RPI/SCNARC for complementary study of cognitive models
Subtask T3.1: Trust Establishment via Distributed Oracles

Motivation:
In a dynamic environment, distributed and decentralized solutions to create, propagate and revoke trust are required.

Key Points:
- Scalable dissemination and propagation of trust through a network
  - Explore tradeoffs between scalability and expressiveness in propagating trust metadata on heterogeneous networks
- Bootstrapping trust (for new comers) in a network
  - Bootstrap trust using cross-network metrics (e.g., use credentials from a social network to bootstrap trust in information sources)
- Revoking trust (terminate or ignore misbehaving entities) on a network
  - Explore tradeoffs between revocation latency and accuracy using distributed and hierarchical TDOs (Trusted Distributed Oracles)
- Task wide F2F meeting at IBM Research on March 8th, 2010
Key Highlights

- Collaboration
  - IPP written with close collaboration of CNARC/SCNARC/IRC PIs
  - Teams within INARC and across other centers are working together on concrete research problems in each of the INARC subtasks

- Staff Rotations
  - Summer intern from PSU/CNARC to IBM/INARC to work on data placement issues in trust establishment via distributed oracles
  - Summer intern from UIUC/INARC to IBM/INARC to work on provenance models

- Key publication submitted

- Provenance demo being prepared as a possible candidate for the ribbon-cutting ceremony – input from multiple PIs.

- Weekly telecons underway to work out details of various research tasks
Conclusions

- INARC has been performing active research and collaborations on IPP research on information network and cross-center collaborations
- INARC has so far got 16 research papers submitted to, published by, or accepted by major conferences and journals
- INARC is performing and seeking more close collaborations with other centers and ARL researchers
- INARC will actively contribute to the June 23 NS-CTA opening ceremony on poster and system demonstrations
- Education: Actively participating NS-CTA seminar and setting up INARC seminar at 4th Wed. of the month (at the same hour as NS-CTA seminar)
  - Christos Faloutsos (CMU) will give the 1st seminar on March 24
  - Noshir Contractor (Northwestern) will give the 2nd seminar on April 28
- IBM summer interns for summer active research on NS-CTA projects and collaborations