Graph Computing and Linked Big Data

Ching-Yung Lin
Manager, Dept. of Network Science, IBM T. J. Watson Research Center
Adjunct Professor, Columbia University and New York University

Keynote Speech @ International Conference on Semantic Computing 2014
June 17th, 2014
Outline

Big Data, Graphs, and System G
Graph Database and Visualization
Social Media Solution and Insider Threat Solution
Other Use Cases and Ongoing Researches
Questions and Open Discussion
Big Data, Graphs, and System G

Graph Database and Visualization

Social Media Solution and Insider Threat Solution

Other Use Cases and Ongoing Researches

Questions and Open Discussion
Big Data Market


![Big Data Revenue by Type, 2013 (in $US millions)](image)

<table>
<thead>
<tr>
<th>Vendor</th>
<th>Big Data Revenue</th>
<th>Total Revenue</th>
<th>Big Data Revenue as % of Total Revenue</th>
<th>% Big Data Hardware Revenue</th>
<th>% Big Data Software Revenue</th>
<th>% Big Data Services Revenue</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM</td>
<td>$1,368</td>
<td>$99,751</td>
<td>1%</td>
<td>31%</td>
<td>27%</td>
<td>42%</td>
</tr>
<tr>
<td>HP</td>
<td>$869</td>
<td>$114,100</td>
<td>1%</td>
<td>42%</td>
<td>14%</td>
<td>44%</td>
</tr>
<tr>
<td>Dell</td>
<td>$652</td>
<td>$54,550</td>
<td>1%</td>
<td>85%</td>
<td>0%</td>
<td>15%</td>
</tr>
<tr>
<td>SAP</td>
<td>$545</td>
<td>$22,900</td>
<td>2%</td>
<td>0%</td>
<td>76%</td>
<td>24%</td>
</tr>
<tr>
<td>Teradata</td>
<td>$518</td>
<td>$2,665</td>
<td>19%</td>
<td>36%</td>
<td>30%</td>
<td>34%</td>
</tr>
<tr>
<td>Oracle</td>
<td>$491</td>
<td>$37,552</td>
<td>1%</td>
<td>28%</td>
<td>37%</td>
<td>36%</td>
</tr>
<tr>
<td>SAS Institute</td>
<td>$480</td>
<td>$3,020</td>
<td>16%</td>
<td>0%</td>
<td>68%</td>
<td>32%</td>
</tr>
<tr>
<td>Palantir</td>
<td>$418</td>
<td>$418</td>
<td>100%</td>
<td>0%</td>
<td>50%</td>
<td>50%</td>
</tr>
<tr>
<td>Accenture</td>
<td>$415</td>
<td>$30,606</td>
<td>1%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>PWC</td>
<td>$312</td>
<td>$32,580</td>
<td>1%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Deloitte</td>
<td>$305</td>
<td>$33,050</td>
<td>1%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Pivotal</td>
<td>$300</td>
<td>$300</td>
<td>100%</td>
<td>15%</td>
<td>50%</td>
<td>35%</td>
</tr>
<tr>
<td>Cisco Systems</td>
<td>$295</td>
<td>$50,200</td>
<td>1%</td>
<td>72%</td>
<td>12%</td>
<td>16%</td>
</tr>
</tbody>
</table>

Big Data Revenue by Sub-Type, 2013

Big Data Revenue by Sub-Type, 2013
(in $US millions)
(n=18,814)

- Professional Services: $6,148
- Compute: $3,647
- Storage: $3,085
- Applications: $1,691
- SQL: $1,306
- Cloud: $1,192
- Infrastructure Software: $830
- Networking: $417
- NoSQL: $290
5 Key Big Data Use Case Categories

Big Data Exploration
Find, visualize, understand all big data to improve decision making

Enhanced 360° View of the Customer
Extend existing customer views (MDM, CRM, etc) by incorporating additional internal and external information sources

Security/Intelligence Extension
Lower risk, detect fraud and monitor cyber security in real-time

Operations Analysis
Analyze a variety of machine data for improved business results

Data Warehouse Augmentation
Integrate big data and data warehouse capabilities to increase operational efficiency
A missing pillar for Big Data
Big Data includes all sorts of Networks

Social/Economic/Political Network

Information/Knowledge Network

Nature/Bio/Cognitive Network

Man-Made Technology Network
Graphs

Graph Database

RDF / Property Graph

Attributes

Charles Flint
- born: "1850"
- died: "1934"
- HQ: "Armonk"
- employees: 433,362
- industry: Software, Hardware, Services

Topological Analytics

Macro

Collective Graph

Contextual Analysis

Collective Analysis

Graphical Models

Micro & Reasoning

Activity Graph

Cognitive Understanding

System G Team

© 2014 IBM Corporation
What is the fundamental challenge for RDB on Linked Data?

In Relational DB, relationships are **distributed**. It takes a long time to **JOIN** to retrieve a graph from data.

Native Graph DB stores nodes and relationships directly, it makes retrieval efficient.

Retrieving multi-step relationships is a **'graph traversal'** problem.

Cited “Graph Database” O’ley 2013
Preliminary datastore comparison for Recommendation & Visualization

IBM KnowledgeView 1-year Access Log: 72.3K users, 82.1K docs, and 1.74 million downloads

Recommendation ==> 2-hop traversal & ranking

For Visualization ==> 4-hop traversal & rankings

<table>
<thead>
<tr>
<th>Query</th>
<th>Time (sec) / App. Type</th>
<th>DB2 via SQL</th>
<th>Oracle via SQL</th>
<th>DB2RDF via SPARQL</th>
<th>Neo4j</th>
<th>Titan (Berk. DB)</th>
<th>Titan (HBase)</th>
<th>System G GBase</th>
<th>System G Native Store</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recomendation</td>
<td></td>
<td>0.24</td>
<td>0.35</td>
<td>TBD</td>
<td>0.068</td>
<td>0.281</td>
<td>0.414</td>
<td>0.201</td>
<td>0.015</td>
</tr>
<tr>
<td>Visualization</td>
<td></td>
<td>52.0 (cold) 50.6 (cache)</td>
<td>201.0 (cold) 42.0 (cache)</td>
<td>TBD</td>
<td>4.8 (cold) 1.2 (cache)</td>
<td>17.3 (cold) 6.8 (cache)</td>
<td>24.2 (cold) 5.7 (cache)</td>
<td>27.0 (cold) 2.4 (cache)</td>
<td>4.2 (cold) 0.07 (cache)</td>
</tr>
</tbody>
</table>

*All performance numbers are preliminary*
What is IBM System G?

A Complete Set of Graph Data Store, Visualizations, Algorithms, and Middleware to Support Big Data Analytics Applications

http://systemG.research.ibm.com (Internet) or http://systemG.ibm.com (IBM internal site)

Rich Graph Algorithm/Functions Primitives

- Centralities
- Communities
- Graph Sampling
- Network Info Flow
- Shortest Paths
- Ego Net Features
- Graph Matching
- Graph Query
- Graph Search
- Bayesian Networks
- Latent Net Inference
- Markov Networks
- Spatio-Temporal Ana.

Multi Graph Type Support

- Few, very **large graphs** (e.g. social, Internet of things)
- Many, many **small graphs** (e.g. protein, healthcare)
- Large **semantic graph** (Semantic web, RDF, Graph search, Graph recommendation)
- Large **Probabilistic graphical models**: Bayesian networks, Markovian networks, HMMs, etc.

And More:

- Graph Visualizations
- Graph Databases
- Graph Middleware for Hardware Platform Optimization
- Cognitive Networks and Cognitive Analytics
- Graph-Embedded Industry Solutions

100+ research innovations/papers including 7 best paper awards

*Including ACM CIKM 2012 Best Paper Award; IEEE BigData 2013 Best Paper Award*
System G Graph Computing Tools

Visualization
- Huge Network Visualization
- Network Propagation
- I2 3D Network Visualization
- Geo Network Visualization
- Graphical Model Visualization

Analytics
- Communities
- Graph Search
- Network Info Flow
- Bayesian Networks
- Centralities
- Graph Query
- Shortest Paths
- Latent Net Inference
- Ego Net Features
- Graph Matching
- Graph Sampling
- Markov Networks

Middleware
- Graph Processing Interface
- BigInsights
- Shared Memory Run Time Library
- Distr. Memory RT Library
- Graphs FPGA/ HMC
- Infosphere Streams (ISS)
- Pthreads
- Graphs RDMA
- MPI
- Cluster (BladeCenter, BlueGene)

Database
- Graph Data Interface
- GBase (update, scan, operators, indexing))
- Native Store
- DB2 RDF
- TinkerPop Compliant DBs
- HBase
- DB2
- HDFS

Icons:
- System G Assets
- Open Source
- IBM Product
- Hardware
System G Cognitive Network and Cognitive Analytics

Perceptrons

Judgment

Abstract comprehension

Multi-Modality Multi-Layer Understanding

Observations

Text/Visual Sentiments, Feeling and Emotions

Reasoning

Intrinsic senses

Sensor Layer

Feature Layer

Concept Layer

Semantics Layer

Cognition Layer

© 2014 IBM Corporation
IBM System G Application Use Cases

1. System G for Expertise Location
2. System G for Recommendation
3. System G for Commerce
4. System G for Financial Analysis
5. System G for Social Media Monitoring
6. System G for Telco Customer Analysis
7. System G for Watson
8. System G for Data Exploration and Visualization
9. System G for Personalized Search
10. System G for Anomaly Detection (Espionage, Sabotage, etc.)
11. System G for Fraud Detection
12. System G for Cybersecurity
13. System G for Sensor Monitoring (Smarter another Planet)
14. System G for Cellular Network Monitoring
15. System G for Cloud Monitoring
17. System G for Traffic Navigation
18. System G for Image and Video Semantic Understanding
19. System G for Genomic Medicine
20. System G for Brain Network Analysis
21. System G for Data Curation
22. System G for Near Earth Object Analysis
Graph Market Analysis (in Big Data Market)


<table>
<thead>
<tr>
<th>USD: billions</th>
<th>2014</th>
<th>2015</th>
<th>2016</th>
<th>2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>Big Data XaaS Revenue</td>
<td>$1.71</td>
<td>$2.43</td>
<td>$2.87</td>
<td>$3.19</td>
</tr>
<tr>
<td>Big Data Professional Services Revenue</td>
<td>$9.24</td>
<td>$12.31</td>
<td>$14.06</td>
<td>$15.30</td>
</tr>
<tr>
<td>Big Data Application (Analytic and Transactional) Revenue</td>
<td>$3.24</td>
<td>$4.94</td>
<td>$6.05</td>
<td>$6.89</td>
</tr>
<tr>
<td>Big Data NoSQL Database Revenue</td>
<td>$0.73</td>
<td>$1.14</td>
<td>$1.41</td>
<td>$1.62</td>
</tr>
<tr>
<td>Big Data SQL Database Revenue</td>
<td>$2.00</td>
<td>$2.48</td>
<td>$2.74</td>
<td>$2.91</td>
</tr>
<tr>
<td>Big Data Infrastructure Revenue</td>
<td>$0.67</td>
<td>$0.93</td>
<td>$1.08</td>
<td>$1.19</td>
</tr>
<tr>
<td>Big Data Networking Revenue</td>
<td>$0.67</td>
<td>$0.89</td>
<td>$1.02</td>
<td>$1.11</td>
</tr>
<tr>
<td>Big Data Storage Revenue</td>
<td>$4.39</td>
<td>$5.85</td>
<td>$6.68</td>
<td>$7.27</td>
</tr>
<tr>
<td>Big Data Compute Revenue</td>
<td>$5.23</td>
<td>$6.70</td>
<td>$7.50</td>
<td>$8.06</td>
</tr>
<tr>
<td>Total Big Data Revenue</td>
<td>$27.9</td>
<td>$37.7</td>
<td>$43.4</td>
<td>$47.5</td>
</tr>
</tbody>
</table>

Wikibon Reports:

- “Hadoop-related software and services matured rapidly in 2012. the NoSQL market is largely up for grabs.” [Oct 2013]
- “It is not uncommon for an enterprise IT organization to support multiple NoSQL DBs alongside legacy RDBMSs. Indeed, there are single applications that often deploy two or more NoSQL solutions, e.g., pairing a document-oriented DB with a graph DB for an analytics solution.” [Dec 2013]

Observations:

- **Service revenue is bigger than the DB** ($15.3B vs $4.53B in 2017).
- NoSQL’s market in 2017 will be $1.62B vs. SQL’s market of $2.91B in the Big Data space.
- Graph DB is one of the 4 categories in NoSQL DB ==> Distributed DB, Document-Oriented DB, Graph NoSQL DB, and In-Memory NoSQL DB.
- Graphs = Graph DB + Applications + Services……
(pending approval) Open System G?

Visualization
- Huge Network Visualization
- Dynamic Network Visualization
- I2 3D Network Visualization
- Geo Network Visualization
- Cognitive Network Visualization

Analytics
- Communities
- Graph Search
- Network Info Flow
- Bayesian Networks
- Centralities
- Graph Query
- Shortest Paths
- Latent Net Inference
- Ego Net Features
- Graph Matching
- Graph Sampling
- Markov Networks

Middleware
- Graph Processing Interface
- BigInsights
- Shared Memory Run-Time Library
- Dist. Memory RT Library
- Graphs FPGA/HMC
- Infosphere Streams (ISS)
- Pthreads
- Graphs RDMA
- MPI
- Cluster (BladeCenter, BlueGene)
- PERCS Coh. Clus.

Database
- Graph Data Interface
- GBase
- DB2 RDF
- TinkerPop Compliant DBs
- HBase
- Native Store
- DB2
- HDFS
- Research Assets
- Open Source
- IBM Product
- Hardware

© 2014 IBM Corporation
Big Data, Graphs, and System G

Graph Database and Visualization

Social Media Solution and Insider Threat Solution

Other Use Cases and Ongoing Researches

Questions and Open Discussion
Graph Database – Why use graph?

Overcome the **inefficiency** of:
1) representing linked data using relational tables
2) accessing graph data by searching in relational database records

Think in terms of **entities (nodes)** and their **relationships (edges)**, not in terms of **entities** and **actions**

➔ **Speed, Scalability, and Schemaless**
• Graph at **front end** allows users to easily model a problem using entities and relationships, but it may **not** be able to offer efficient implementation of **graph operations** due to the different underlying data structure.

• Graph at **back end** allows real graph representation in memory/disk, and **optimization** of graph operations.

**Appearance and Essence:**
even a relational database can be wrapped as a graph, but it can not help the efficiency for graph computing and data management, as it is not a real graph store.
Which Graph is Used?

The example in Apache Titan tutorial

- Homogeneous nodes for easily mapping to the schema of the back end database

The example in Neo4j tutorial

- Schemless allows more flexibility and handles various linked data scenarios

The two underlying graph models are not identical.
What is the Issue for Non-Graph Backend

Strongly relies on **indexing**, which can adversely impact the database performance when the graph is **large** and **dynamic**;

The underlying data structure is **not** a linked data, thus we can not directly follow the links when **traversing** a graph. So, overhead is introduced.
Scalability – Do not be cheated! (Graph Analytics Examples)

**Graph size v.s. Machine size:**
Let's consider storing the topology (in CRS-like format) of a graph in a server with 1TB memory (assuming average vertex degree is 25):

\[
\text{storage\_size} = (\text{index\_size} + 1) + \text{edgeList\_size} \\
1 \text{ TB} = ((\#v+1) + \#v\times25) \times 8\text{bytes} \\
\#v \approx 5 \text{ Billion}
\]

**Scale up & out:**
If the Hadoop based solution scales *linearly* to 18 million m/c, it is just equivalent to the GraphLab in terms of performance, but the cost....
System G Native store represents graphs in-memory and on-disk
- Organizing graph data for representing a graph that stores both graph structure and vertex properties and edge properties
- Caching graph data in memory in either batch-mode or on-demand from the on-disk streaming graph data
- Persisting graph updates along with the time stamps from in-memory graph to on-disk graph
- Performing graph queries by loading graph structure and/or property data
Performance of Graph Primitives -- I

- Graph Primitives: AddVertex, AddEdge, QueryVertex
- Consistent results as what we observed
- Titan over Berkeley DB, not finished in a day.
- On Intel Haswell (Xeon E5-2697 v2) at 2.7 GHz, 256 GB memory, RedHat Linux
Many graph analytics require traversal of subgraphs → critically important

Experiment setup
- Start from a vertex and traverse an ego network subgraph in BFS manner for three hops
- Sort the vertices at three hops away from the starting vertex

Measured the number of edges traversed in each query (TEPS)
Native Store exhibits the best performance, and Neo4j is the second
An Emerging Benchmark

Test Set: data generator of full social media activity simulation of any number of users

We are participating this EU effort. We're also anticipating a potential benchmarking with an important customer.
Visualizing Huge Static Graph

Challenging Task:

Squeezing millions and even billions of records into million pixels (1600 X 1200 ≈ 2 million pixels)

- 76425 species
- 14.8 million tweets
- 500 million users
Visualization Key Challenges

**Visual clutter**
How can we encode the information intuitively?

**Performance issues**
How can we render the huge datasets in real time with rich interactions?

**Cognition**
How can users understand the visual representation when the information is overwhelming?
Geometry Based Techniques

**Challenge:**
How to squeezing millions and even billions of records into million pixels

**Idea:**
Layout graph onto an infinity plain
Project the infinity plain into a screen window with finite size

*Visualization of an organization tree with more than 10,000 nodes*
Project the infinity plain onto a finite circular disc called Poincaré disk
Conformal: The angle between any two lines on the sphere must be the same between their projected counterparts on the map
Big Data, Graphs, and System G

Graph Database and Visualization

Social Media Solution and Insider Threat Solution

Other Use Cases and Ongoing Researches

Questions and Open Discussion
IBM System G Enterprise Expertise Solution

Production Live System used by IBM GBS since 2009 – verified ~$100M contribution

15,000 contributors in 76 countries; 92,000 annual unique IBM users
25,000,000+ emails & SameTime messages (incl. Content features)
1,000,000+ Learning clicks; 14M KnowledgeView, SalesOne, ..., access data
1,000,000+ Lotus Connections (blogs, file sharing, bookmark) data
200,000 people’s consulting project & earning data

– On BusinessWeek four times, including being the Top Story of Week, April 2009
– Help IBM earned the 2012 Most Admired Knowledge Enterprise Award
– Wharton School study: $7,010 gain per user per year using the tool
– In 2012, contributing about 1/3 of GBS Practitioner Portal $228.5 million savings and benefits
– APQC (WW leader in Knowledge Practice) April 2013:
  “The Industry Leader and Best Practice in Expertise Location”

Dynamic networks of 400,000+ IBMers:

Shortest Paths
Social Capital
Bridges
Hubs
Expertise Search
Graph Search
Graph Recomm.
Finding and Ranking Expertise – Social Network Analysis

- Who are the key bridges? Who have the most connections? How do these experts cluster?

SmallBlue analyzes underlining dynamic network structure in enterprise

Independent experts on healthcare

A cluster of XYZ experts

Influencers are the one with high ‘Betweeness’ and ‘Degree’ value

UI to highlight experts based on my social proximity, the number of experts she connects, or the ‘social bridges’ importance
IBM System G Social Media Solution Overview

Actionable Applications (April 2014)
- Live Monitoring
- Anomaly Detection
- Impact Trend Analysis
- Flow Analytics & Visualization
- Person Analytics
- Multimedia Analytics
- Auto-Counter Messaging

Inferred Cognitive Traits
(Human Essential)
- Personality
- Needs
- Value
- Trustworthiness
(Human Dynamic)
- Contextual Behavior
- Emotional State
(Information Dynamic)
- Info Reasoning & Morphing
- Visual Sentiment

Social Media Posts

74 papers published & submitted; 12+ patents filed
ACM CIKM 2012 Best Paper Award
IEEE BigData 2013 Best Paper Award
PNAS Cover Article Jan 2013; Science (1); Nature (2)

Inferred Social Network Traits
- Roles
- Dynamic Analysis
- Topological Analysis
- Location Analysis

Analytics & Predictive Models
IBM System G Social Media Solution Research Tasks

**Thrust 1. Modeling Information Dissemination:**

Task 1.1. Computational Modeling of User Dynamic Behavior
Task 1.2. Computational Models of Trust and Social Capital
Task 1.3. Information Morphing Modeling
Task 1.4. Persuasiveness of Memes
Task 1.5. The Observability of Social Systems
Task 1.6. Culture-Dependent Social Media Modeling
Task 1.7. Dynamics of Influence in Social Networks
Task 1.8. Understanding the Optimal Immunization Policy
Task 1.9. Modeling and Identification of Campaign Target Audience
Task 1.10. Modeling and Predicting Competing Memes

**Thrust 2. Detecting and Tracking Information Dissemination:**

Task 2.1. Real-Time and Large-Scale Social Media Mining
Task 2.2. Role and Function Discovery
Task 2.3. Detecting Malicious Users and Malware Propagation
Task 2.4. Emergent Topic Detection and Tracking
Task 2.5. Detecting Evolution History and Authenticity of Multimedia Memes
Task 2.6. Synchronistic Social Media Information and Social Proof Opinion Mining
Task 2.7. Community Detection and Tracking
Task 2.8. Interplay Across Multiple-Networks
Task 2.9: Assessing Affective Impact of Multi-Modal Social Media

**Thrust 3. Affecting Information Dissemination:**

Task 3.1. Crowd-sourcing Evidence Gathering to Formulate Counter-messaging Objectives
Task 3.2. Delivery and Evaluation of a Counter-messaging Campaign
Task 3.3. Optimal Target People Selection
Task 3.4. Automated Generation of Counter Messaging
Task 3.5. User Interfaces for Semi-Automatic Counter Messaging
Task 3.6. Controlling the Dynamics of Influence in Social Networks
Task 3.7. Influencing the Outcome of Competing Memes and Counter Messaging
Solution Architecture

Social Media Data

Memes
Persuasiveness in
Multiple Networks

Social Content
Analysis

Topic
Modeling

Information
Morph

Content models

Behavioral models

Real-time Large-scale Social Media Mining Algorithms

Personality
Analysis

Behavioral
Modeling

Emotion
Analysis

Location
Analysis

Behavioral models

Real-time Large-scale Social Media Mining System

Role Discovery

Network
Controllability
& Observability

Social Capital
Modeling

Social network models

Visual memes
evolution models

Visual Manipulation
Modeling &
Detection

Social Media
Applications

Social Malware
Detection

Affecting
Memes
Propagatio

Counter
Messaging
Planning

UI & Visualization

Social Mining Architecture
System G Team

© 2014 IBM Corporation
X-Bank Use Case

Objective: Detect unexpected social media movements that may impact a major bank’s business

- X-Bank:
  - Major bank in Spain

- Client needs:
  - Monitor Catalan independence movement: independence may bring bankruptcy since X-Bank needs ECB support
  - Detect potential PR crisis by analyzing the formation and spreading of grassroots opinion on their employees and services

- Challenges:
  - Existing social media monitoring tools miss important tweets that don’t contain specified keywords and are not from specified users
  - Existing tools lack of predictive capability of tweets’ potential influence

An image tweet (without mentioning “the bank name”) sparks a lot critiques of their unfair practice
Visual Sentiment and Semantic Analysis

First work in the literature on automatic visual sentiment analysis

“For content to go viral, it needs to be emotional,” Dan Jones, 2012

Detection results of “lonely dog” (80% accuracy, 4 out of 5 correct)

Detection results of “crazy car” (100% accuracy, 5 out of 5 correct)

Experiment on Sentiment Detection Accuracy on Twitter

<table>
<thead>
<tr>
<th></th>
<th>Text</th>
<th>Visual</th>
<th>T+V</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.43</td>
<td>0.70</td>
<td>0.72</td>
</tr>
</tbody>
</table>

© 2014 IBM Corporation
Automatic Comments on Images

○ Nice pictures, interesting writing. A beautiful little girl.
○ Nice treatment of a fantastic capture. A wonderful picture. Have a good day and keep smiling.

Make a Comment!  More Specific  More Generic  Cancel
– **Personality:** Mapping personal/organizational social media postings to scores of BIG 5 Personality (*Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism*).

– **Needs:** Mapping personal/organizational social media postings to scores of *Harmony, Curiosity, Self-expression, Ideal, Excitement, and Closeness*.

– **Values:** Mapping personal/organizational social media postings to scores of *Self-Enhance, Conservation, Open-to-Change, Hedonism, and Self-Transcend*.

– **Trustiness and Trustworthiness:** Deriving from interaction and propagation history between the user and his followers and the people he follows.

– **Influence:** Total *attention* received by user as leader across all discovered flows.

Precision-Recall performance of predicting info propagation by different features (Our proposed influence index: FLOWER)
Live Monitoring

System G SMISC Social Media Monitoring

Monitoring categories

Monitoring filter

Live Tweets, Sentiment, Keywords

Growing Influential Between Graphs

Real-Time Translation, Locations, Top Retweets
Anomaly Detection

- Human personality, value traits to show
- User info (e.g., personality chart, bot score, hijacker score, etc.)
- Top anomaly sequences, and explanations
- Visual sentiment scores
Impact Analysis

Real-time hashtag monitoring
Predicting the business impact of tweet messages grouped by hashtags.

Please click on the "hashtags" to learn more about each conversations content.
Last updated at 2014-05-01 18:10:02 GMT

<table>
<thead>
<tr>
<th></th>
<th>Conversations</th>
<th>Impact</th>
<th>Impact Score</th>
<th>Prediction</th>
<th>First Tweet Time</th>
<th>Last Tweet Time</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>[erdogan, turkey]</td>
<td>HIGH</td>
<td>34.3</td>
<td>URL</td>
<td>2014-04-30 05:45:07</td>
<td>2014-05-01 14:06:12</td>
<td>29 hours</td>
</tr>
</tbody>
</table>
Person Analytics - II

Each spike shows a user’s activity and how other responded.

Regular, low interaction activities may indicate bots.

Diverse interaction/sentiment patterns may indicate real users.
Flow Analytics - I

- Topic cluster tree shows how sequences' content are related to each other.
- Timeline view shows how users of different characteristics responded in each sequence.
- MDS view shows how anomalies distribute.
- Feature and State view shows the features of a sequence, and how they transition from one state to another.
Since 2009, U.S. Justice Department lawyers have pursued at least 19 cases of corporate espionage. Most had connections to China, according to cases summarized in a new U.S. strategy report on trade secrets.

Among high-profile cases:
- In November 2012, Shenhuan Du, a former General Motors engineer, and her husband Yu Qin were convicted in Detroit for conspiring to steal hybrid technology trade secrets, intending to use them in a joint venture with an auto maker in China.
- In September 2012, Shuang Xue was convicted in federal court in New Jersey for exporting United States military technology to China and stealing thousands of electronic files from his employer.
- Also in September, Churlia Yang, a former CME Group engineer, pleaded guilty in Chicago for downloading files with up part of a trading system. He was planning to help trading efficiency in China.
- In August 2012, Hanjuan Jin, a former Motorola engineer, was sentenced in Illinois to four years in prison for stealing Motorola's proprietary IDEN technology for herself and for a company that works for the Chinese military.
- In March 2012, former DuPont scientist Tai Chou pleaded guilty in California to conspiracy to commit economic espionage, admitting he provided trade secrets regarding a titanium-dioxide making process to China-controlled companies.
- In January 2012, Yuan Li, a former Sanofi-Aventis chemist, pleaded guilty in New Jersey for making trade secrets of the drug company available for sale through a Chinese subsidiary of a Chinese chemicals company.

“Since 2009, U.S. Justice Department lawyers have pursued at least 19 cases of corporate espionage, including GM, Ford, Motorola, DuPont, … impact economic and jobs”

– WSJ Feb 21, 2013
Insider Threat comes with sequence of ‘weak signals’

- **Personal stress:**
  - Gender identity confusion
  - Family change (termination of a stable relationship)

- **Job stress:**
  - Dissatisfaction with work
    - Job roles and location (sent to Iraq)
    - Long work hours (14/7)

- **Unstable Mental Status:**
  - Fight with colleagues, write complaining emails to colleagues
  - Emotional collapse in workspace (crying, violence against objects)
  - Large number of unhappy Facebook posts (work-related and emotional)

- **Planning:**
  - Online chat with a hacker confiding his first attempt of leaking the information

- **Attack:**
  - Brought music CD to work and downloaded/copied documents onto it with his own account
Emerging 'Cognitive Security'

- Contextual Situation
- Cognitive
- Cyber Activities

Structured

Unstructured, Semantic
A novel **Cognitive Security System** to Detect and Predict Abnormal Behaviors in Organization.

![Diagram of Cognitive Security System]

- **Infrastructure** + ~70 Types of Analytics (489 Detectors) + Reasoning + User Interface
- **Sensor Layer**: Available existing data
  - : observations
  - : hidden states
- **Feature Layer**: 489 detector scores / per day / per person
- **Concept Layer**
- **Semantics Layer**
- **Cognition Layer**
- **Behavior analysis**
- **Content analysis**
- **Data processing**
- **Data storage**
- **Data sensing**
- **Analytics Infrastructure**
- **Graph analysis**

**Detection, Prediction, and Exploration Interface**

- Espionage Detection
- Sabotage Detection
- Fraud Detection

© 2014 IBM Corporation
Real-world data were collected from ~ 5,500 de-identified people since July 2012 (a.k.a. Vegas Data) ~ 600,000 records per person from July 2012 to present

Red team (CERT) inserted 3 test scenarios per month

Each scenario has one (or sometimes more) abnormal person (by modifying data of real people)
Evaluations on the Real-World Data in Vegas Lab (Oct 2013)

- Each month, DARPA ADAMS program inserted 3 cases (1 abnormal person per case) in the “Vegas Lab”.
- Each performer system retrieved top abnormal people out of the 5,500 people per month.
- Below chart showed where the 3 IBM ADAMS systems (Sabotage, Espionage, and Fraud) ranked the abnormal person in each case. “All” is a combined rank list of the 3 systems. (Oct 2013 review on 12/12 ~ 03/13 data)

<table>
<thead>
<tr>
<th>Month</th>
<th>Sabotage (Scenario)</th>
<th>Espionage (Scenario)</th>
<th>Fraud (Scenario)</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dec</td>
<td>4</td>
<td>241</td>
<td>1667</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>981</td>
<td>1</td>
<td>120</td>
<td>1</td>
</tr>
<tr>
<td>Jan</td>
<td>1526</td>
<td>454</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>4230</td>
<td>3367</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Feb</td>
<td>11</td>
<td>44</td>
<td>574</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>4230</td>
<td>1462</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>Mar</td>
<td>1936</td>
<td>73</td>
<td>232</td>
<td>203</td>
</tr>
<tr>
<td></td>
<td>4101</td>
<td>9</td>
<td>803</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>65</td>
<td>4101</td>
<td>654</td>
<td>181</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1650</td>
<td>294</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1544</td>
<td>9</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>4325</td>
<td>11</td>
<td>46</td>
<td>27</td>
</tr>
</tbody>
</table>

Promising results. IBM’s system successfully caught the bad guys of the 12 cases: 4 as Top #1, 3 in Top #2-#5, 2 in Top #6-#20, 1 in Top #21-#50, and 2 in Top #51-#100.

12. Layoff Logic Bomb: An engineer is worried about rumors of impending layoffs feels that he needs some kind of an “insurance policy”, in case he gets laid-off or fired. He creates a “logic bomb” which will delete all files from a number of company Linux systems in five days, unless he resets the timer before then.

13. Outsourcer’s Apprentice: (http://www.bbc.co.uk/news/technology-21043693) A software developer outsources his job to China and spends his workdays surfing the web. Most surfing occurs on a second laptop. He pays just a small fraction of his salary to a Chinese company to do his job. The developer provides his VPN credentials to the company and enabling Terminal Services on his workstation. The Chinese consulting firm sends the developer PayPal invoices.

8. Anomalous Encryption: A Subject wishes to pass sensitive information to a foreign government in exchange for that government setting him up with his own business. Subject researches NSA monitoring capabilities, generates a long random passphrase and then tests encrypting and mails data to personal account. The subject encrypts documents and emails the key.
Abnormal-user ranking of ~5,500 users over a month. The number shows where the red team abnormal person is in the ranked list.

Pink background represents the same type of the scenario and the detector.

Rank-All is based on the highest anomaly scores of the 3 detectors.

Out of the 24 new cases, our same-type detectors ranked:
- 9 in Top #1-#5,
- 4 in Top #6-#10,
- 3 in Top #11-#20,
- 5 in Top #21-#40,
- 1 in Top #51-#100, and
- 2 in Top #101-#160

We ran our system only once, without any knowledge of the red team scenarios and how they were inserted. This was again the best performance and showed the promising potential use of system.

### 24 more Benchmarks Reported in the April 2014 meeting

<table>
<thead>
<tr>
<th>Month</th>
<th>Scenario</th>
<th>Espionage</th>
<th>Fraud</th>
<th>Sabotage</th>
<th>Rank_All</th>
</tr>
</thead>
<tbody>
<tr>
<td>April</td>
<td>14 (Espionage)</td>
<td>6</td>
<td>2218</td>
<td>1833</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>10 (Sabotage)</td>
<td>3341</td>
<td>1817</td>
<td>158</td>
<td>444</td>
</tr>
<tr>
<td></td>
<td>17 (Fraud)</td>
<td>3284</td>
<td>5</td>
<td>436</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>18 (Fraud)</td>
<td>2371</td>
<td>9</td>
<td>1733</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>19 (Fraud)</td>
<td>1131</td>
<td>4</td>
<td>750</td>
<td>11</td>
</tr>
<tr>
<td>May</td>
<td>6 (Espionage)</td>
<td>3</td>
<td>4140</td>
<td>4140</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>6 (Espionage)</td>
<td>34</td>
<td>2645</td>
<td>4140</td>
<td>102</td>
</tr>
<tr>
<td></td>
<td>6 (Espionage)</td>
<td>6</td>
<td>807</td>
<td>91</td>
<td>17</td>
</tr>
<tr>
<td>June</td>
<td>6 (Espionage)</td>
<td>107</td>
<td>3939</td>
<td>971</td>
<td>246</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>715</td>
<td>394</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>July</td>
<td>20 (Fraud)</td>
<td>586</td>
<td>1</td>
<td>126</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>21 (Fraud)</td>
<td>980</td>
<td>4</td>
<td>793</td>
<td>11</td>
</tr>
<tr>
<td>Oct</td>
<td>22 (Espionage)</td>
<td>2</td>
<td>2578</td>
<td>79</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>22 (Espionage)</td>
<td>3</td>
<td>672</td>
<td>1034</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>22 (Espionage)</td>
<td>5</td>
<td>1061</td>
<td>1512</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>96</td>
<td>1262</td>
<td>858</td>
<td>176</td>
<td></td>
</tr>
<tr>
<td>Nov</td>
<td>21 (Fraud)</td>
<td>1898</td>
<td>7</td>
<td>511</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>23 (Sabotage)</td>
<td>3833</td>
<td>4384</td>
<td>27</td>
<td>68</td>
</tr>
<tr>
<td></td>
<td>2721</td>
<td>910</td>
<td>39</td>
<td>113</td>
<td></td>
</tr>
<tr>
<td>Dec</td>
<td>24 (Fraud)</td>
<td>12</td>
<td>33</td>
<td>341</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>24 (Fraud)</td>
<td>2132</td>
<td>18</td>
<td>1271</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td>25 (Fraud)</td>
<td>50</td>
<td>13</td>
<td>768</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td>26 (Fraud)</td>
<td>1207</td>
<td>21</td>
<td>3976</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td>27 (Sabotage)</td>
<td>579</td>
<td>1517</td>
<td>33</td>
<td>85</td>
</tr>
</tbody>
</table>
Big Data, Graphs, and System G

Graph Database and Visualization

Social Media Solution and Insider Threat Solution

Other Use Cases and Ongoing Researches

Questions and Discussions
Use Case 18: Graph Analysis for Computer Vision

Vertex Correspondence

Attribute Transformation

ARG s

ARG t

Y_t
Use Case 21: Understanding Brain Network

Possible Network Patterns:

Timeline & Activity of Neuron 0:
- Green curve is the raw signal.
- Blue curve is the processed signal.
- Red curve is the detection result of neuron activity.

(Double click on the arrow to play or pause. Drag and drop the arrow to move forward or backward.)
### Top 10 (November 2013)

<table>
<thead>
<tr>
<th>Rank</th>
<th>Machine</th>
<th>Installation Site</th>
<th>No. of nodes</th>
<th>No. of cores</th>
<th>Problem scale</th>
<th>GTEPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DOE/NNSA/LLNL Sequoia (IBM - BlueGene/Q, Power QBC 16C 1.60 GHz)</td>
<td>Lawrence Livermore National Laboratory</td>
<td>65536</td>
<td>1048576</td>
<td>40</td>
<td>15363</td>
</tr>
<tr>
<td>2</td>
<td>DOE/SC/Argonne National Laboratory Mira - Argonne National Laboratory (49152 nodes, 786432 cores)</td>
<td>Argonne National Laboratory</td>
<td>49152</td>
<td>786432</td>
<td>40</td>
<td>14328</td>
</tr>
<tr>
<td>3</td>
<td>JUQUEEN (IBM - BlueGene/Q, Power QBC 16C 1.60 GHz)</td>
<td>Forschungszentrum Juelich (FZJ)</td>
<td>16384</td>
<td>262144</td>
<td>38</td>
<td>5848</td>
</tr>
<tr>
<td>4</td>
<td>K computer - RIKEN Advanced Institute for Computational Science (AICS) (65536 nodes, 524288 cores)</td>
<td>RIKEN Advanced Institute for Computational Science (AICS)</td>
<td>65536</td>
<td>524288</td>
<td>40</td>
<td>5524.12</td>
</tr>
<tr>
<td>5</td>
<td>Fermi (IBM - BlueGene/Q, Power QBC 16C 1.60 GHz)</td>
<td>CINECA</td>
<td>8192</td>
<td>131072</td>
<td>37</td>
<td>2567</td>
</tr>
<tr>
<td>6</td>
<td>Tianhe-2 (MilkyWay-2) (National University of Defense Technology - MPI)</td>
<td>Changsha, China</td>
<td>8192</td>
<td>196608</td>
<td>36</td>
<td>2061.48</td>
</tr>
<tr>
<td>7</td>
<td>Turing (IBM - Power QBC 16C 1.60 GHz)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**IBM BlueGene or P775:**
- 24 out of Top 30, except #4, #14, #22: Fujitsu
- #6: Tianhe
- #15: Cray
- #24: HP
Big Data, Graphs, and System G

Graph Database and Visualization

Social Media Solution and Insider Threat Solution

Other Use Cases and Ongoing Researches

Questions and Open Discussion
Acknowledgement: IBM System G Team and SMISC & ADAMS Team

---

**SMISC (Social Media Solution):** Shih-Fu, Chang, Columbia; Laszlo Barabasi, Northeastern; Brian Uzzi, Norwestern; Jaideep Srivastava, Minnesota; Tina Eliassi-Ra, Rutgers; Michalas Faloutsos, UNM; Christos Faloutsos, CMU; Trevor Darrell, Berkeley; Ajay Divakaran, SRI.

**ADAMS (Insider Threat Solution):** Yan Liu and Ram Nevatia, USC

---

- **Watson Research Center teams:**
  - System, Analysis, Middleware, Database, and Visualization (*Network Science Group:* Ching-Yung Lin *(Lead)*, Zhen Wen ³, Hanghang Tong, Danny Yeh, Jason Crawford ¹, Yinglong Xia ², Sabrina Lin, Keith Houck, Julie Macnaught, Jie Liu, Larry Lai, Lifeng Nai, Nan Cao ⁴)
  - Database and Analysis (*Database Research Group:* Yuan-Chi Chang, Mustafa Canim, Bishwaranjan Bhattacharjee)
  - Database (*DB2RDF Group:* Kavitha Srinivas, Anastasios Kementsietsidis, Achille Fokoue, Julian Dolby, Mihaela Bornea)
  - Analysis (*Streams System and Analytics Group:* Kun-Lung Wu, Gabriela Jacques Da Silva, Kanat Tangwongsan)
  - Analysis (*Mobile Network Analytics:* Kang-Won Lee, Ting He, Ramya Raghavendra, Murtaza Zafer)
  - Analysis (*Machine Learning Group:* Rick Lawrence)
  - Middleware (*Middleware Research Group:* Liana Fong, Wei Tan, Xavier Guerin, Yanbin Liu)
  - Middleware & Hardware (*Scalable Systems Group:* Gabriel Tanase, Peng Wu, Mauricio Serrano)

- **India Research Lab:** Analysis (*SNAzzy Group:* Amit Nanavati, Natwar Modani)
- **China Research Lab:** Analysis (*X-RIME, Information Management Group:* Chen Wang, Ju Wei Shi)
- **Australia Research Lab:** Suraj Pandey and Wanita Sherchan
- **Austin Research Lab:** Peter Hofstee, Jian Li
- **Brazil Research Lab:** Ana Appel; **Ireland Research Lab:** Shoukat Ali

---

© 2014 IBM Corporation