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

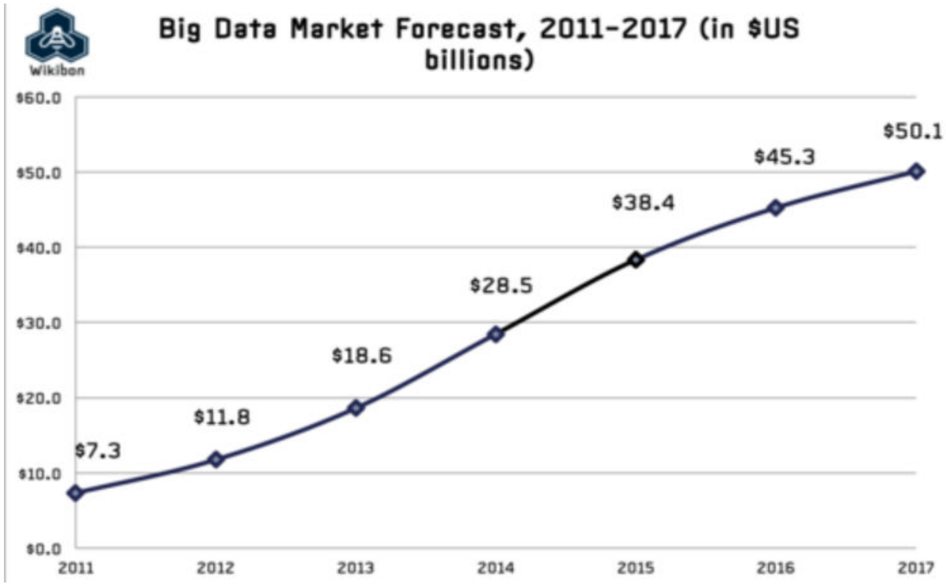
Graph Database and Visualization

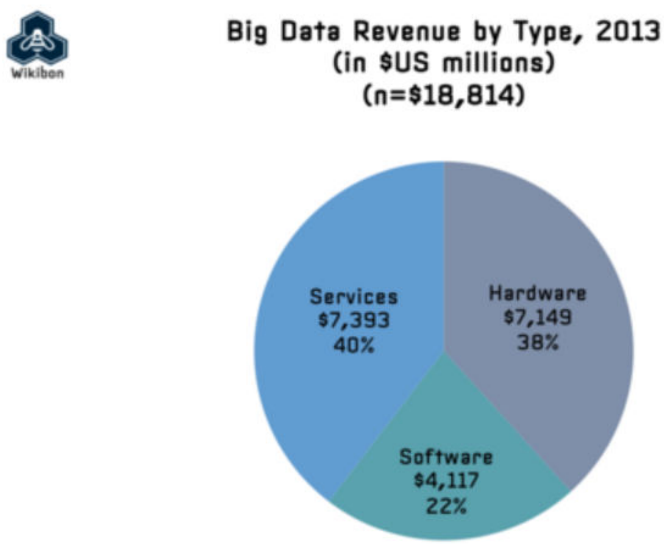
Social Media Solution and Insider Threat Solution

Other Use Cases and Ongoing Researches

Questions and Open Discussion

Big Data Market



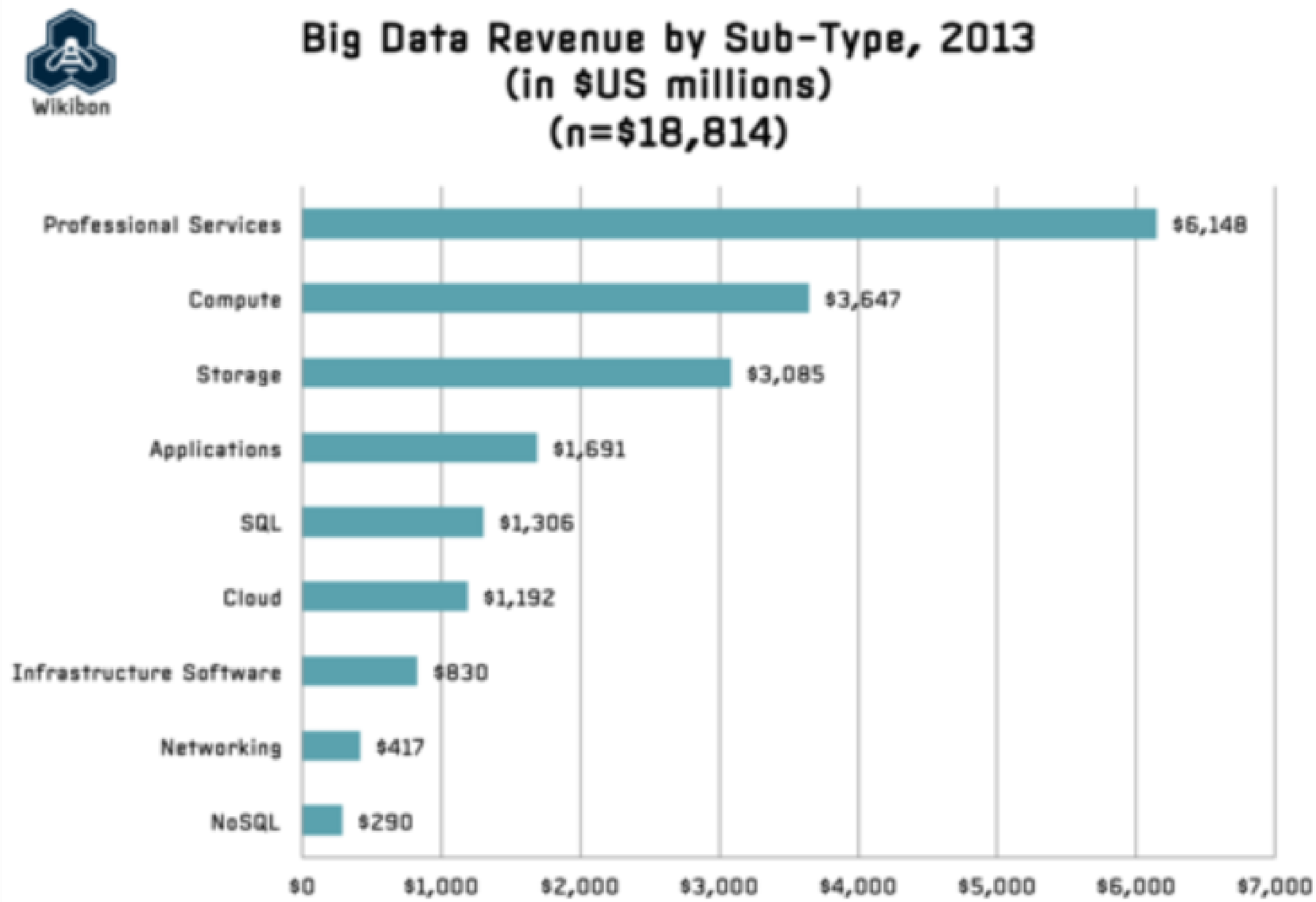


2013 Worldwide Big Data Revenue by Vendor ($US millions)

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

http://wikibon.org/wiki/v/Big_Data_Vendor_Revenue_and_Market_Forecast_2013-2017

Big Data Revenue by Sub-Type, 2013



System G Team

5 Key Big Data Use Case Categories

Big Data Exploration

Find, visualize, understand all big data to improve decision making

Enhanced 360º 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

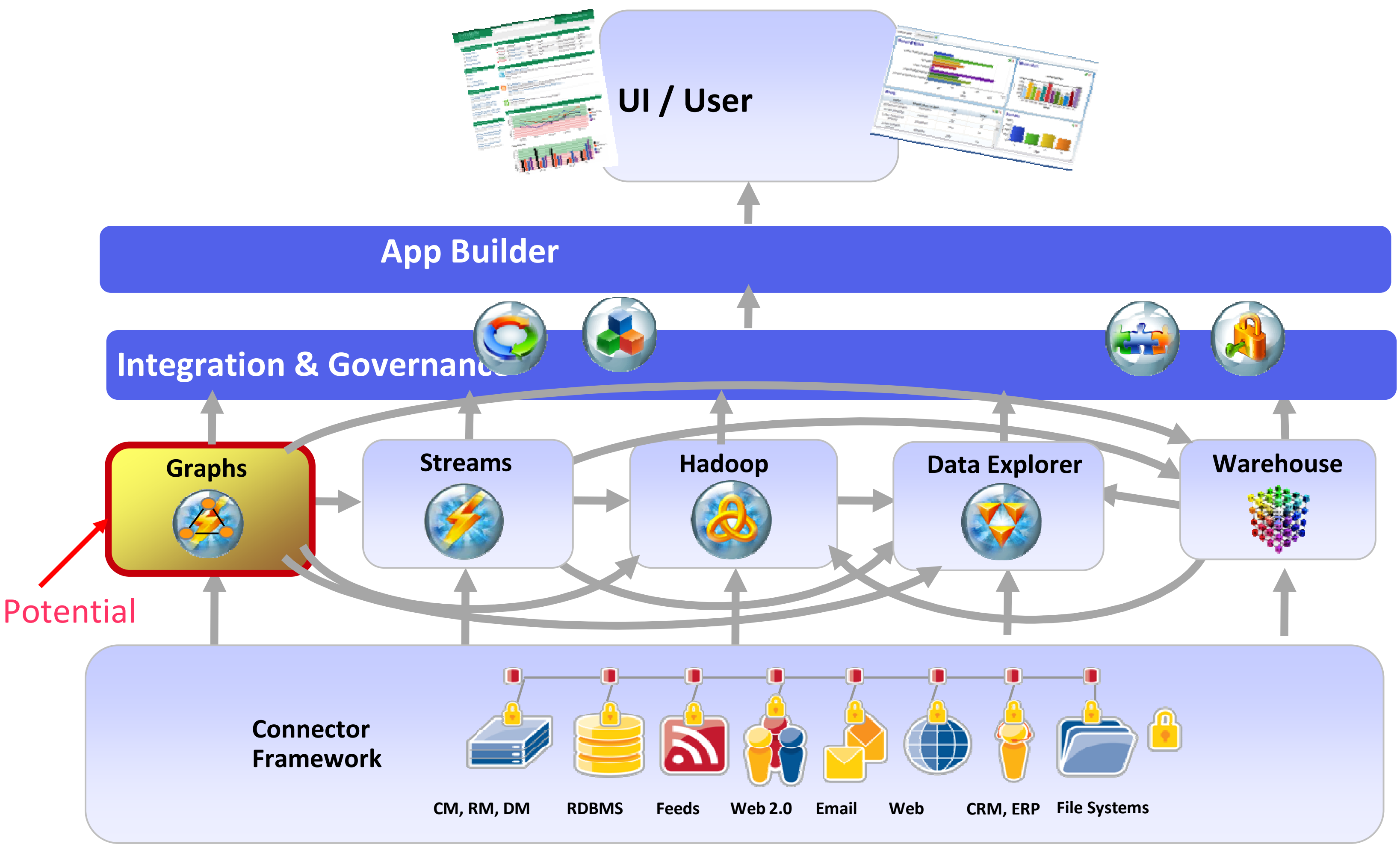
Operations Analysis

Analyze a variety of machine data for improved business results

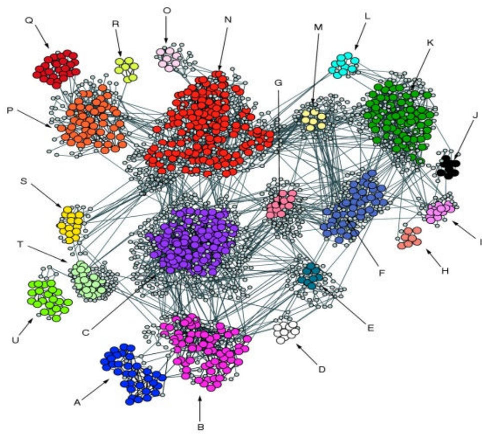
Data Warehouse Augmentation

Integrate big data and data warehouse capabilities to increase operational efficiency

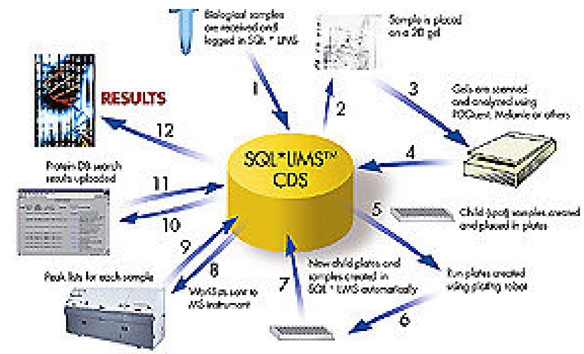
A missing pillar for Big Data



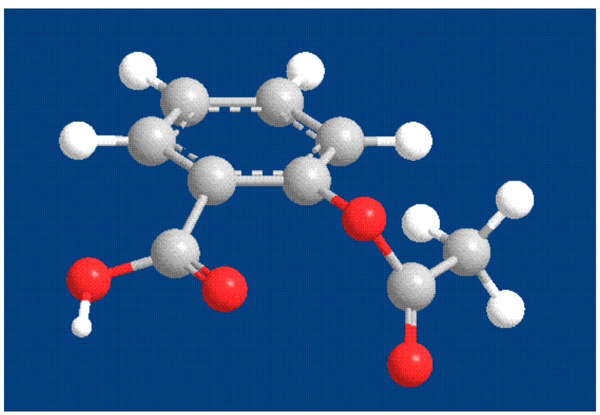
Big Data includes all sorts of Networks



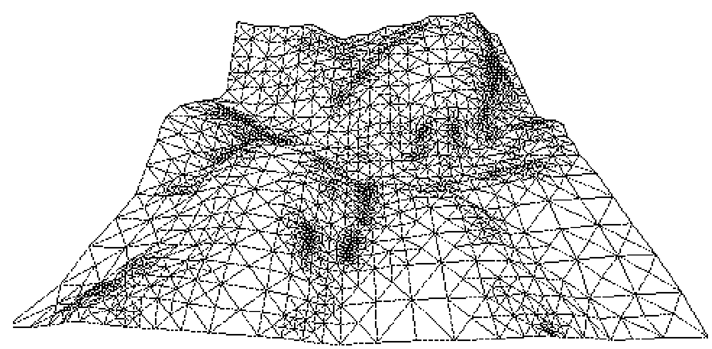
Social/Economic/Political Network



Information/Knowledge Network



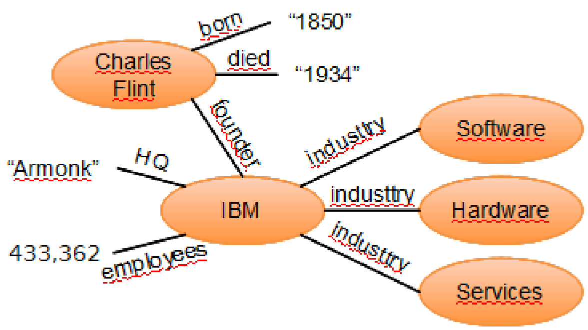
Nature/Bio/Cognitive Network



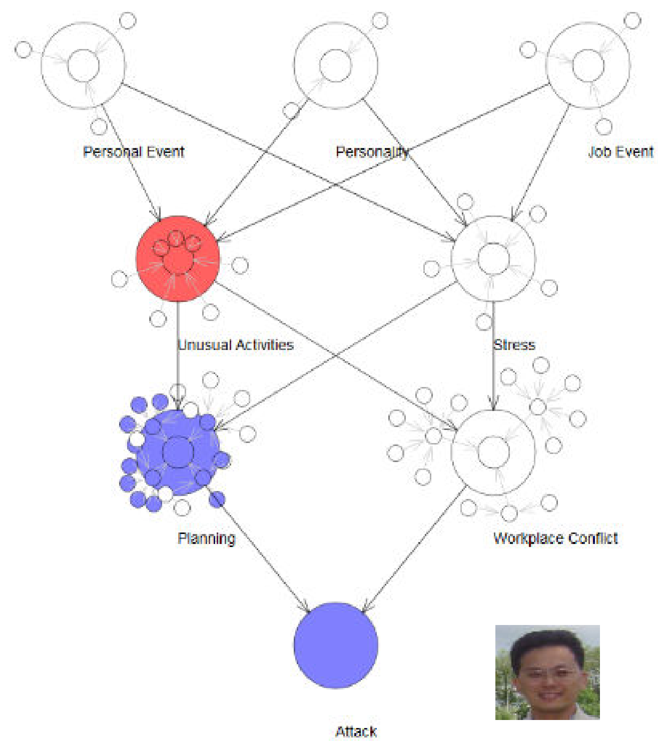
Man-Made Technology Network

Graphs

Graph Database	Topological Analytics	Graphical Models
RDF / Property Graph	Collective Graph	Activity Graph
Attributes	*Macro*	*Micro & Reasoning*



subject	predicate	object
Charles Flint	born	"1850"
Charles Flint	died	"1934"
Charles Flint	founder	IBM
IBM	HQ	"Armonk"
IBM	employees	433,362
IBM	industry	Software
IBM	industry	Hardware
IBM	industry	Services



Contextual Analysis

Collective Analysis

Cognitive Understanding

What is the fundamental challenge for RDB on Linked Data?

User

UserID	User	Address	Phone	Email	Alternate
1	Alice	123 Foo St.	12345678	alice@example.org	alice@neo4j.org
2	Bob	456 Bar Ave.		bob@example.org	
...	...	...	...	...	...
99	Zach	99 South St.		zach@example.org	

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

Order

OrderID	UserID
1234	1
5678	1
...	...
5588	99

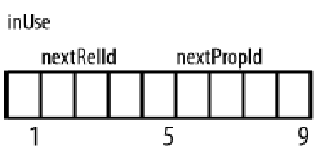
LineItem

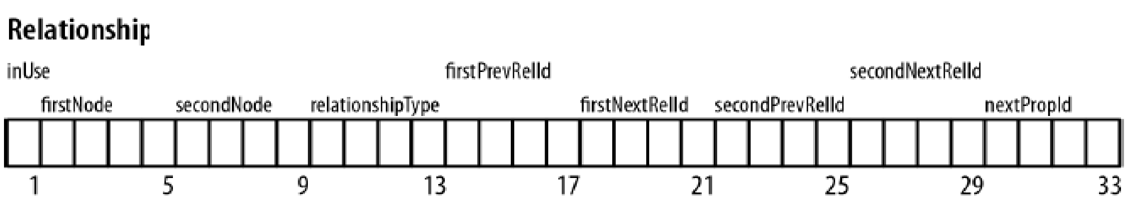
OrderID	ProductID	Quantity
1234	765	2
1234	987	1
...	...	...
5588	765	1

Product

ProductID	Description	Handling
321	strawberry ice cream	freezer
765	potatoes	
...	...	
987	dried spaghetti	

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



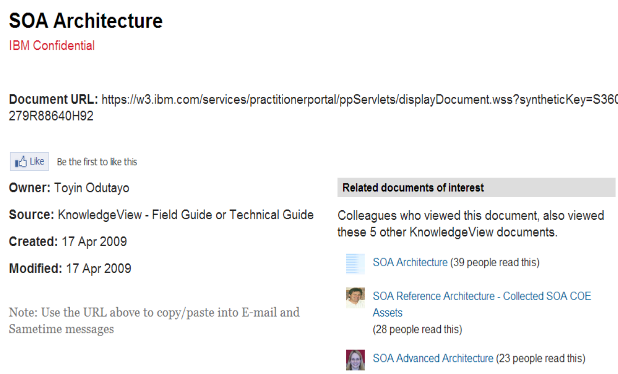


*Retrieving multi-step relationships is a '**graph traversal**' problem*

Cited “Graph Database” O’liey 2013

Preliminary datastore comparison for Recommendation & Visualization

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



SOA Architecture
IBM Confidential

Document URL: https://w3.ibm.com/services/practitionerportal/ppServlets/displayDocument.wss?syntheticKey=S36C
279R88640H92

Like Be the first to like this

Owner: Toyin Odutayo
Source: KnowledgeView - Field Guide or Technical Guide
Created: 17 Apr 2009
Modified: 17 Apr 2009

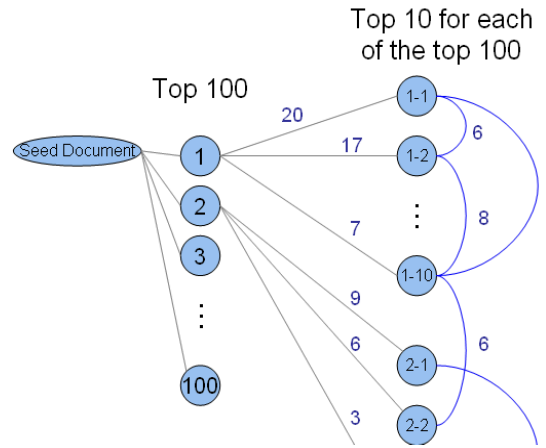
Note: Use the URL above to copy/paste into E-mail and Sametime messages

Related documents of interest

Colleagues who viewed this document, also viewed these 5 other KnowledgeView documents.

SOA Architecture (39 people read this)
SOA Reference Architecture - Collected SOA COE Assets (28 people read this)
SOA Advanced Architecture (23 people read this)

Recommendation ==> 2-hop traversal & ranking



For Visualization ==> 4-hop traversal & rankings

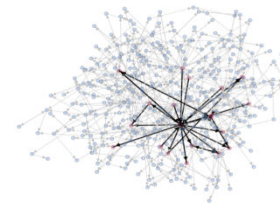
Query Time (sec) / App. Type	DB2 via SQL	Oracle via SQL	DB2RDF via SPARQL	Neo4j	Titan (Berk.DB)	Titan (HBase)	System G GBase	System G Native Store
Recommendation	0.24	0.35	TBD	0.068	0.281	0.414	0.201	0.015
Visualization	52.0 (cold) 50.6 (cache)	201.0 (cold) 42.0 (cache)	TBD	4.8 (cold) 1.2 (cache)	17.3 (cold) 6.8 (cache)	24.2 (cold) 5.7 (cache)	27.0 (cold) 2.4 (cache)	4.2 (cold) 0.07 (cache)
	Products			*Startup*	*Open Sources*		*System G*	

*All performance numbers are preliminary

What is IBM System G?

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.

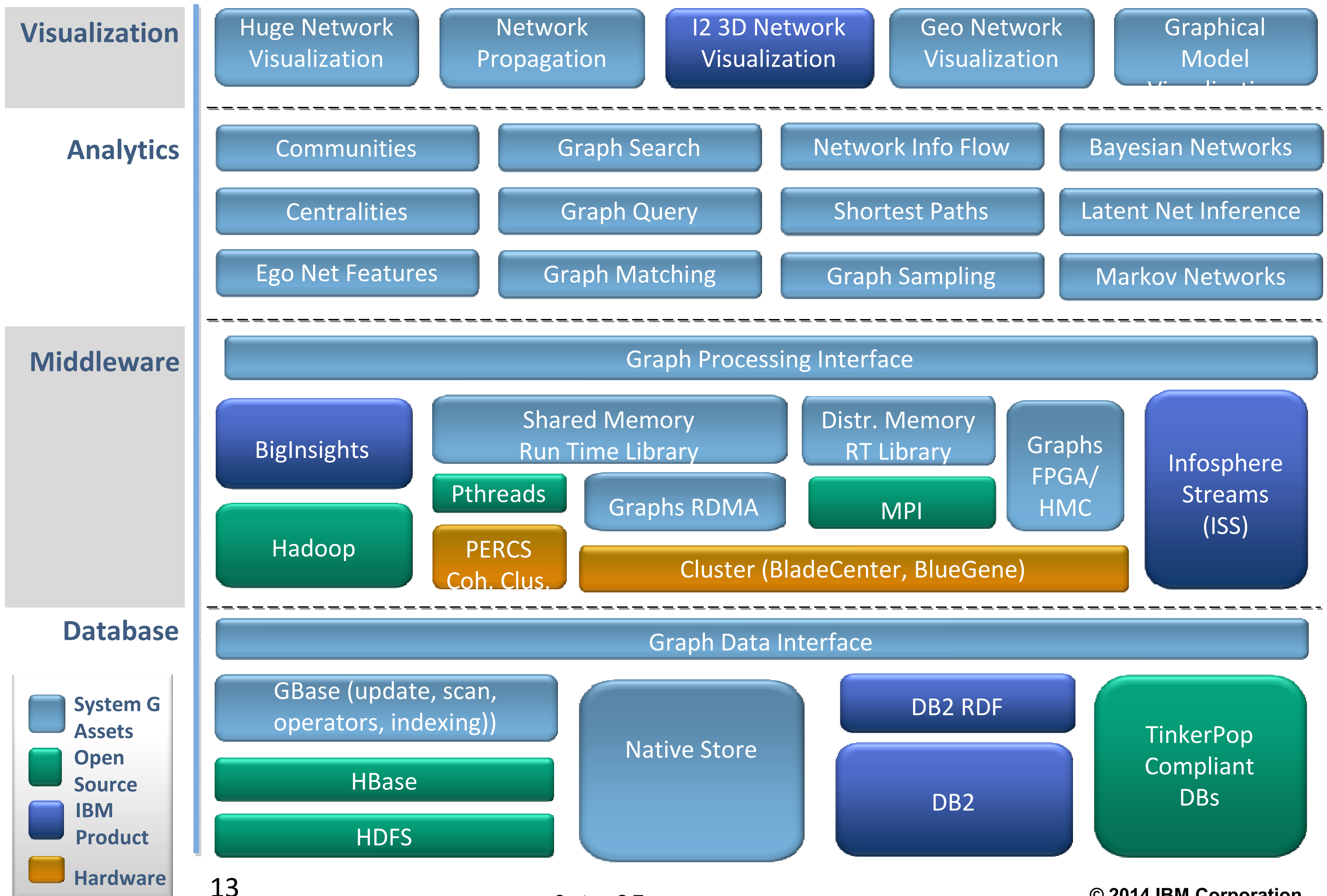
And More:

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

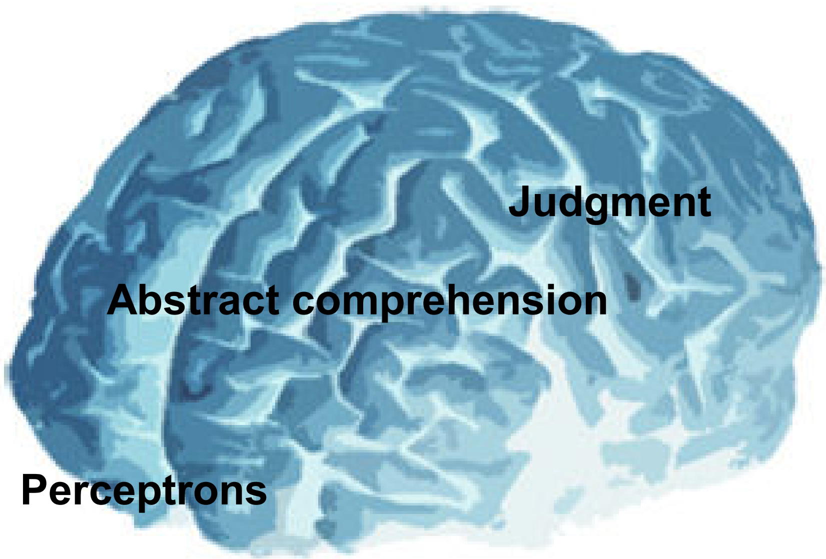
100+ research innovations/papers including 7 best paper awards

Including ACM CIKM 2012 Best Paper Award; IEEE BigData 2013 Best Paper Award

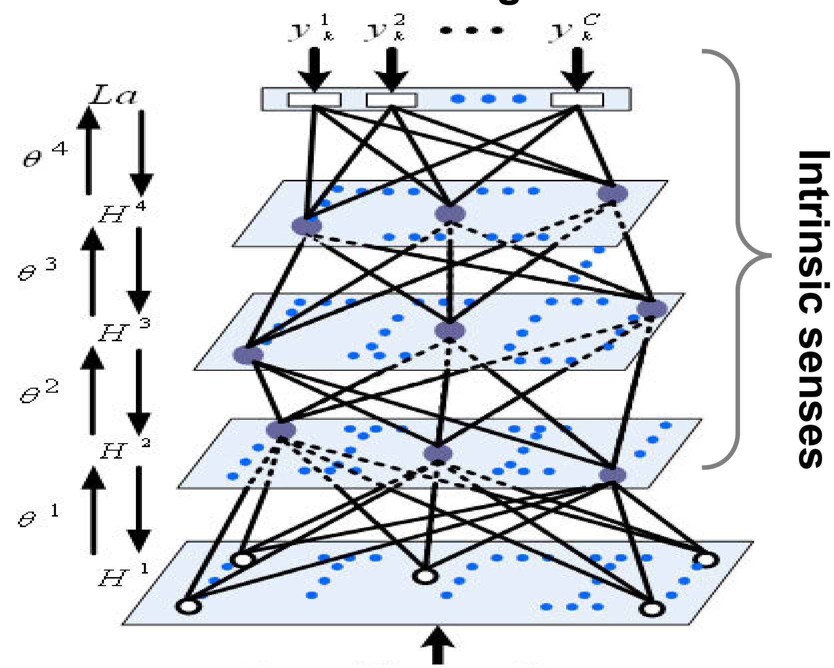
System G Graph Computing Tools



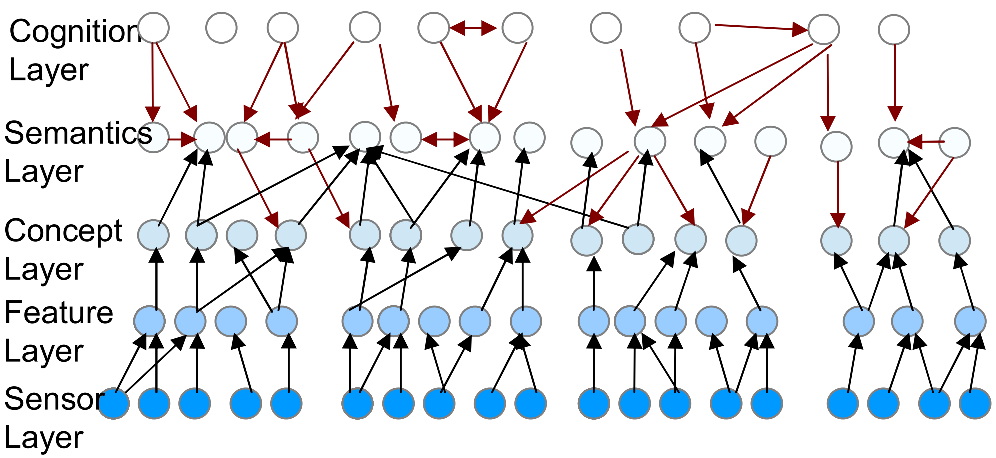
System G Cognitive Network and Cognitive Analytics



Reasoning



Observations



Multi-Modality Multi-Layer Understanding

Text/Visual Sentiments, Feeing and Emotions

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 Celluar Network Monitoring
15. System G for Cloud Monitoring
16. System G for Code Life Cycle Management
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)

http://wikibon.org/wiki/v/Big_Data_Database_Revenue_and_Market_Forecast_2012-2017

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

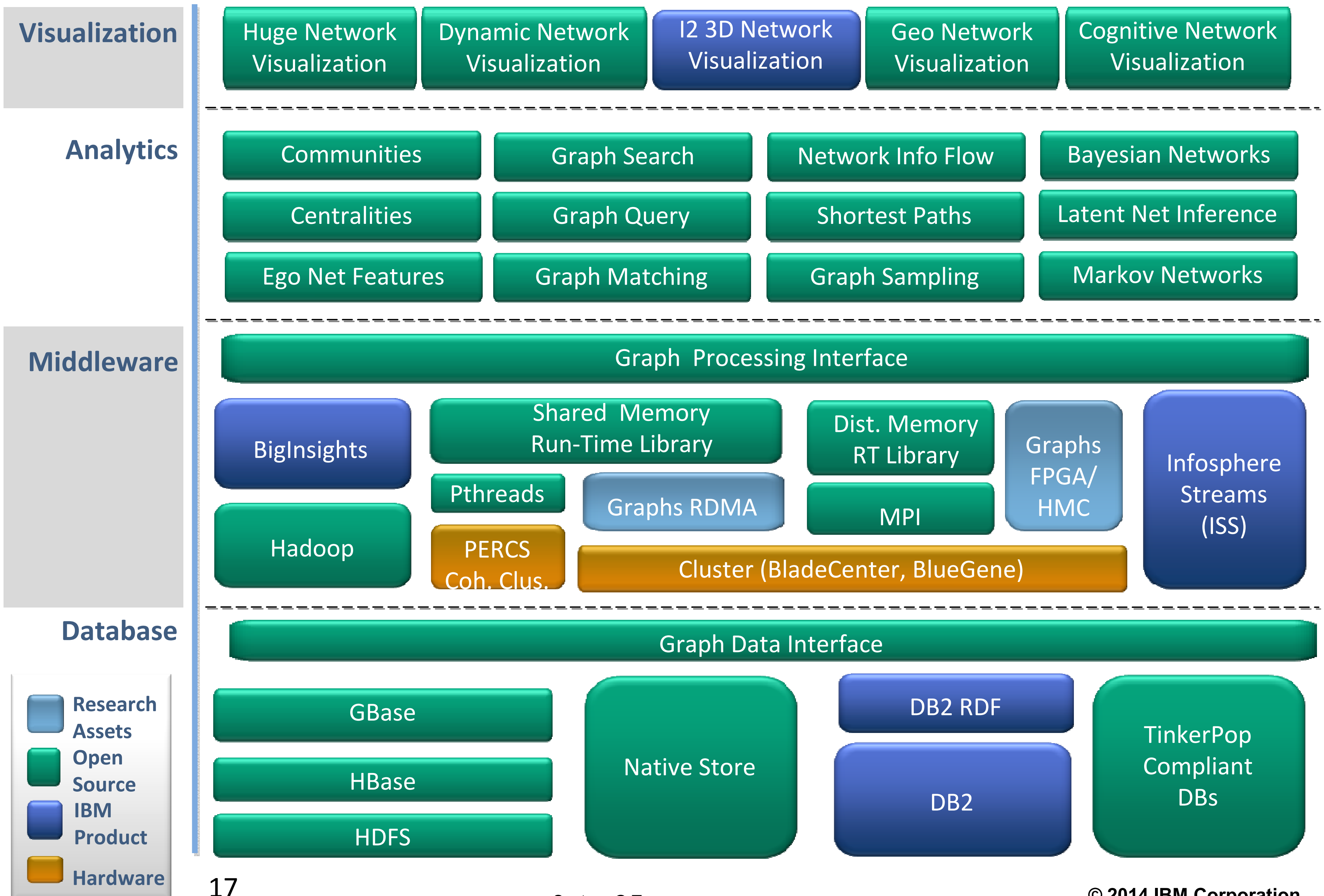
Wikibon Reports:

- "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-Orinted DB, Graph NoSQL DB, and In-Memory NoSQL DB.
- Graphs = Graph DB + Applications + Services…...

(pending approval) Open System G?



Big Data, Graphs, and System G

Graph Database and Visualization

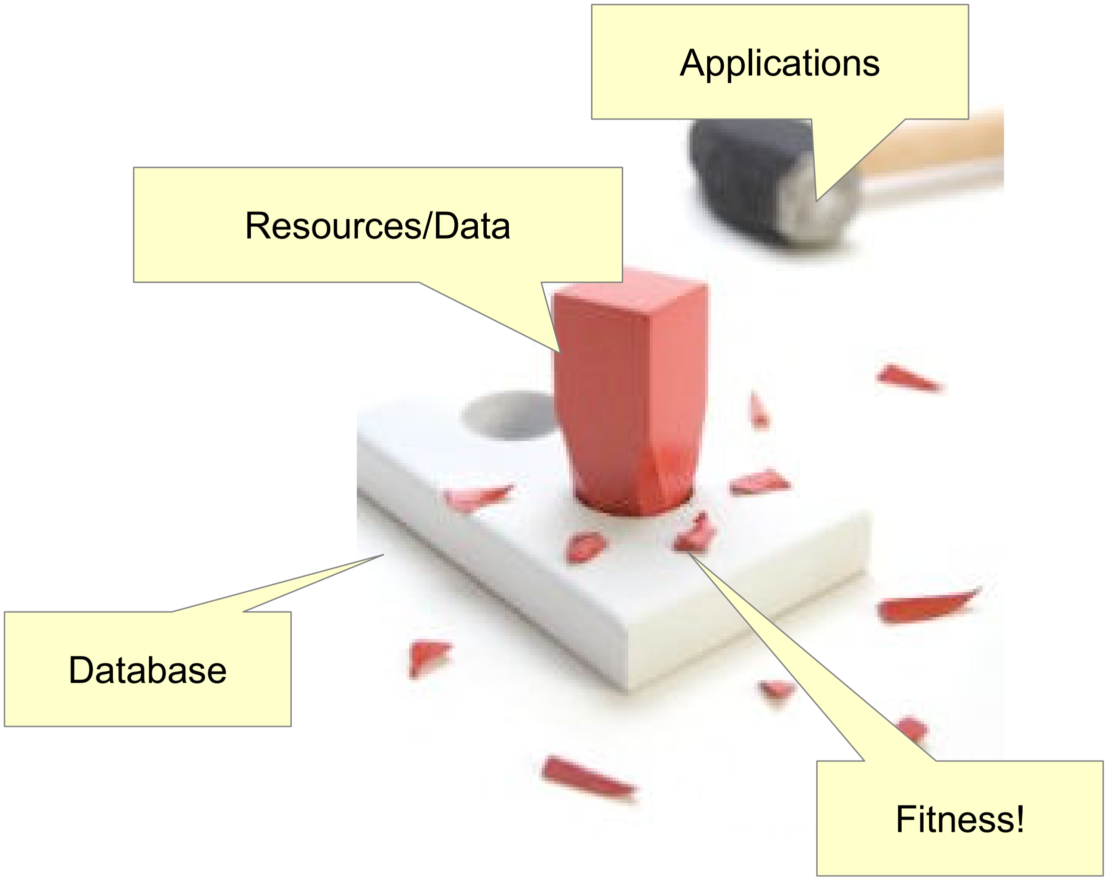
Social Media Solution and Insider Threat Solution

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 Database – Where is the graph?

- Graph at **<u>front end</u>** 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 **<u>back end</u>** allows real graph representation in memory/disk, and **optimization** of graph operations

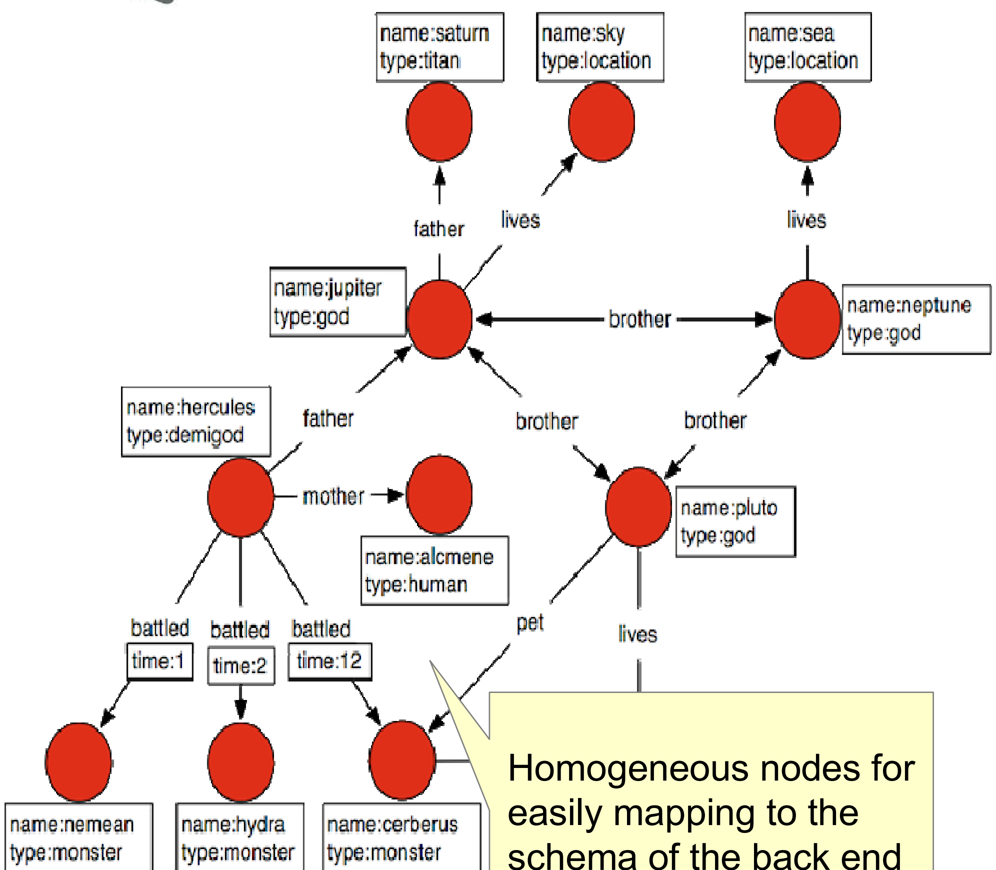
Appearance and Essence: even a relational database can be wrapped as a graph, but it can not help the efficiency for graph computing and data management, as it is not a real graph store

Which Graph is Used?

The two underlying graph models are not identical

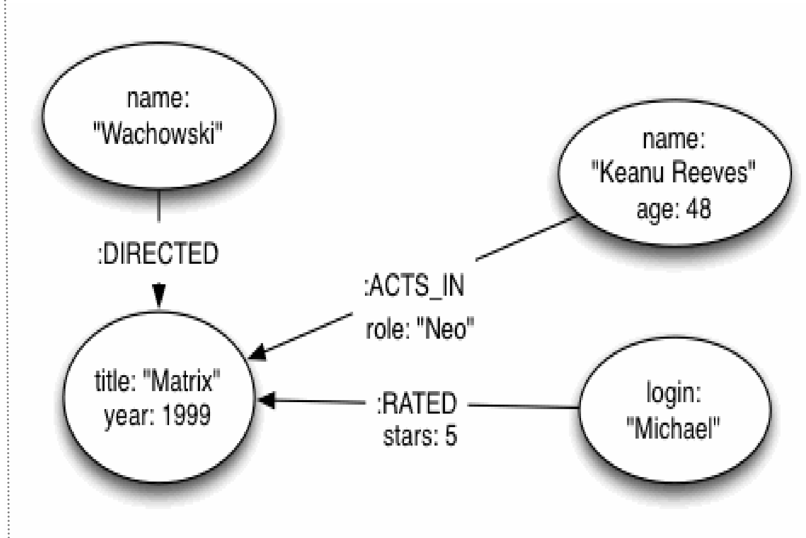
≠

The example in Apache Titan tutorial



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

The example in Neo4j tutorial



Schemaless allows more flexibility and handles various linked data scenarios

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

Highlighted index in a segment of Titan source code

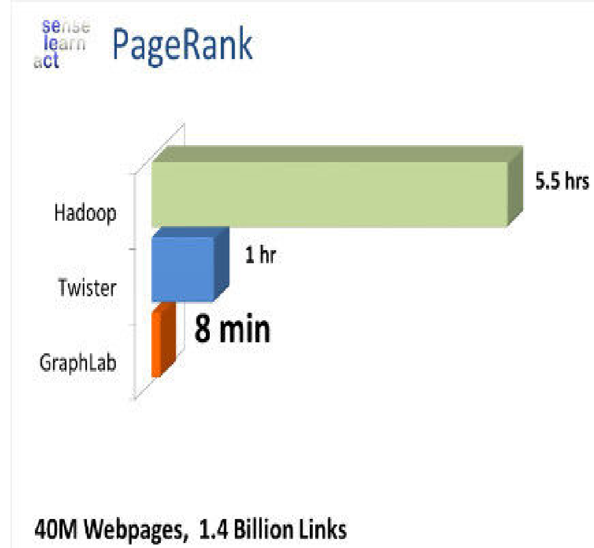
```
public static TitanGraph create(final String directory) {
    BaseConfiguration config = new BaseConfiguration();
    Configuration storage = config.subset(GraphDatabaseConfiguration.STORAGE_NAMESPACE);
    // configuring local backend
    storage.setProperty(GraphDatabaseConfiguration.STORAGE_BACKEND_KEY, "local");
    storage.setProperty(GraphDatabaseConfiguration.STORAGE_DIRECTORY_KEY, directory);
    // configuring elastic search index
    Configuration index = storage.subset(GraphDatabaseConfiguration.INDEX_NAMESPACE).subset(INDEX_
    index.setProperty(INDEX_BACKEND_KEY, "elasticsearch");
    index.setProperty("local-mode", true);
    index.setProperty("client-only", false);
    index.setProperty(STORAGE_DIRECTORY_KEY, directory + File.separator + "es");

    TitanGraph graph = TitanFactory.open(config);
    GraphOfTheGodsFactory.load(graph);
    return graph;
}

public static void load(final TitanGraph graph) {

    graph.makeKey("name").dataType(String.class).indexed(Vertex.class).unique().make();
    graph.makeKey("age").dataType(Integer.class).indexed(INDEX_NAME, Vertex.class).make();
    graph.makeKey("type").dataType(String.class).make();

    final TitanKey time = graph.makeKey("time").dataType(Integer.class).make();
    final TitanKey reason = graph.makeKey("reason").dataType(String.class).indexed(INDEX_NAME, Ed
    graph.makeKey("place").dataType(Geoshape.class).indexed(INDEX_NAME, Edge.class).make();
```

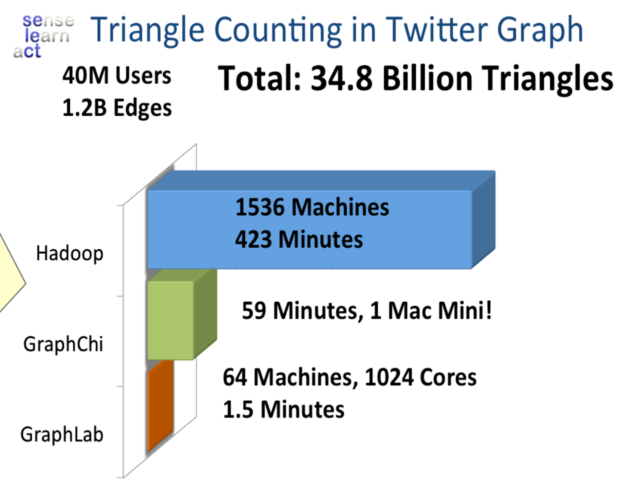
Scalability – Do not be cheated! (Graph Analytics Examples)

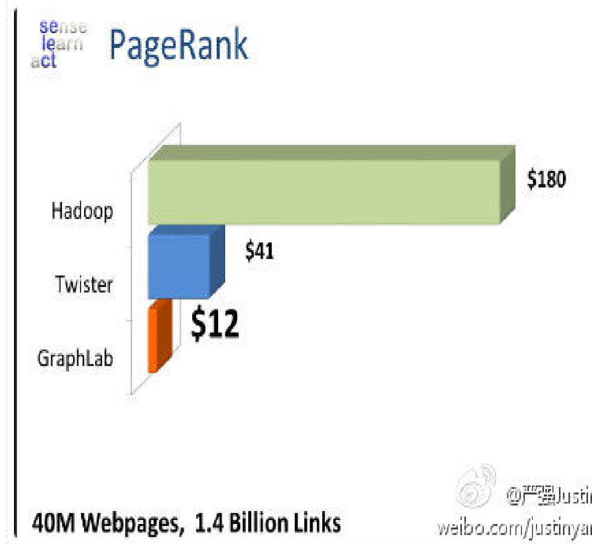
XXX is a scalable YYY optimized for bla bla bla...

<u>Graph size v.s. Machine size:</u>
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):

storage_size = (index_size+1) + edgeList_size
1 TB = ((#v+1) + #v*25) * 8bytes
#v ~= 5 Billion





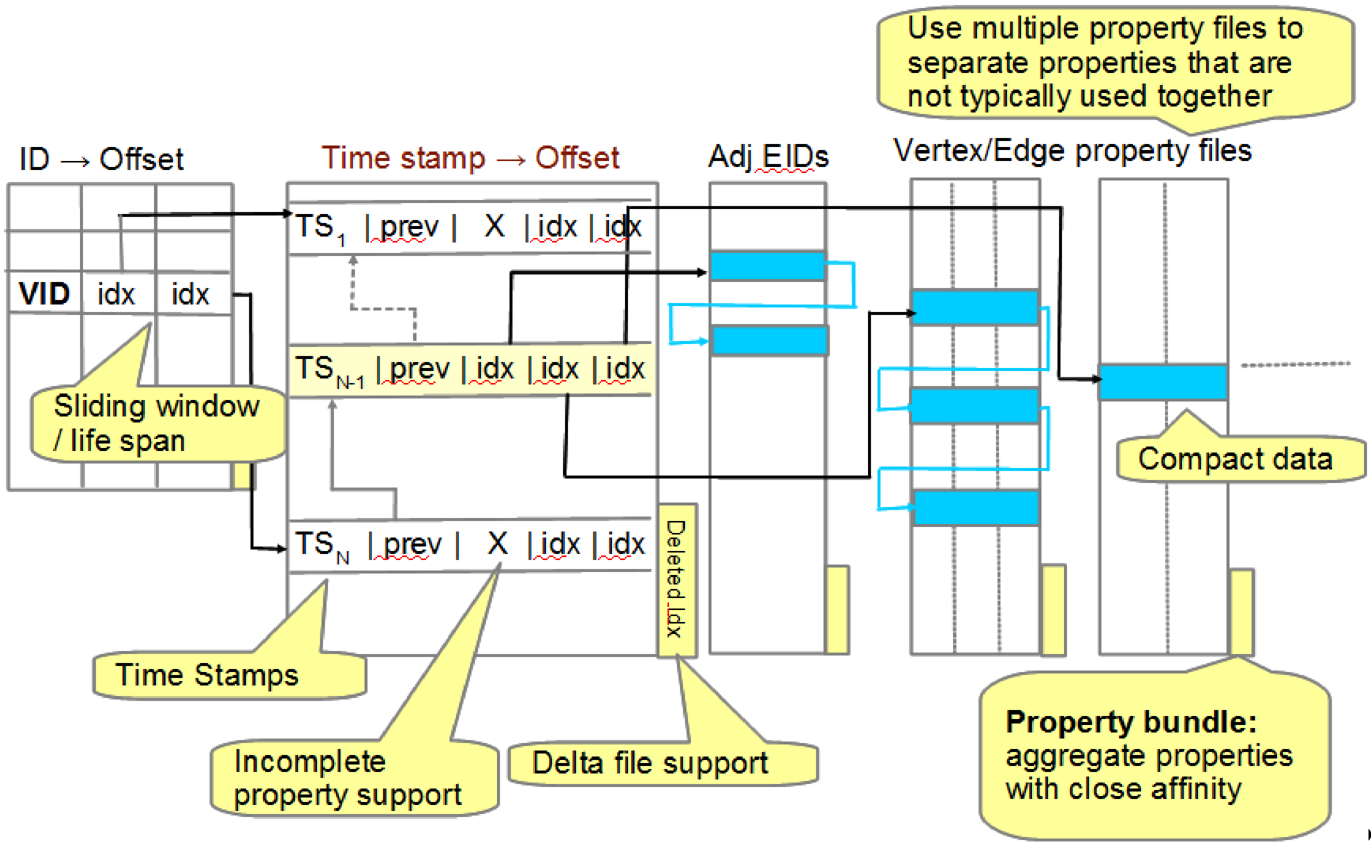


<u>Scale up & out:</u>
If the Hadoop based solution scales *linearly* to **18 million m/c**, it is just equivalent to the GraphLab in terms of performance, but the cost....

System G Native Store

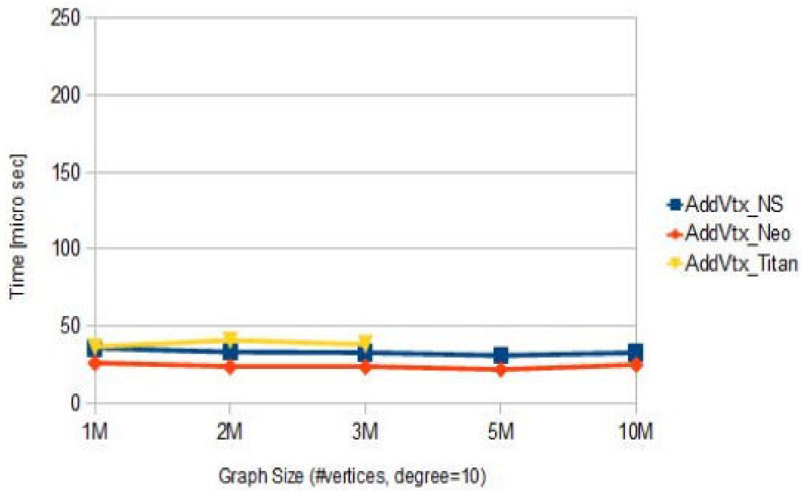
- **System G Native store represents graphs in-memory and on-disk**
  - Organizing graph data for representing a graph that stores both graph structure and vertex properties and edge properties
  - Caching graph data in memory in either batch-mode or on-demand from the on-disk streaming graph data
  - 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

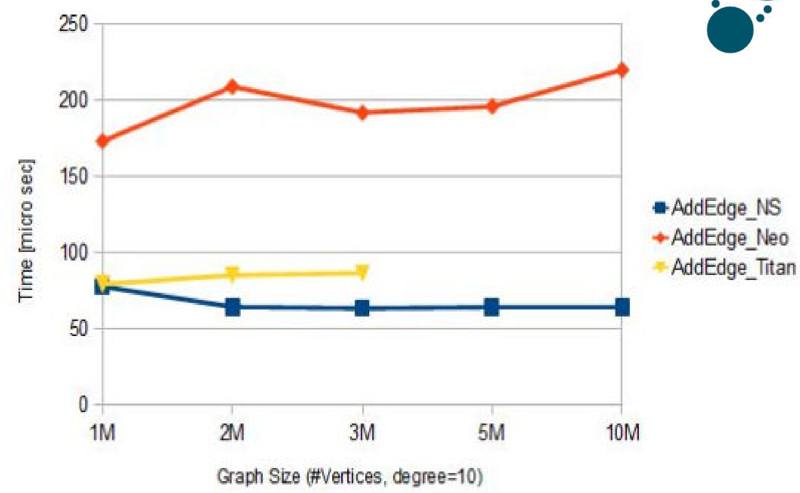
Example: System G On-disk persistent graph:

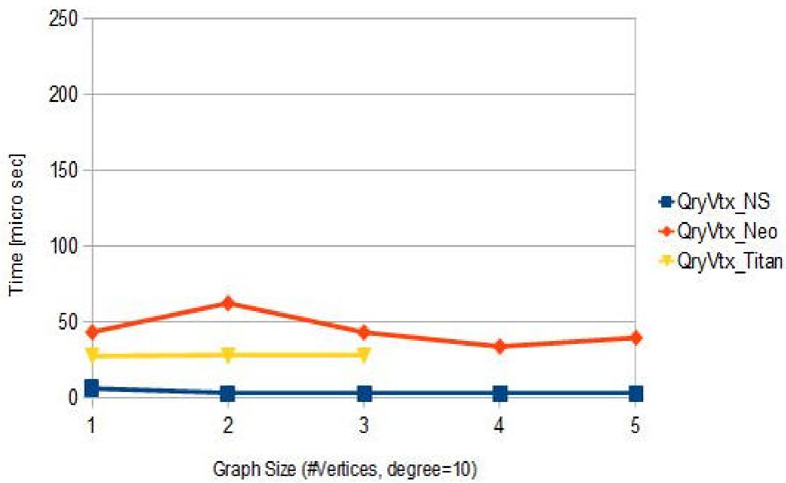


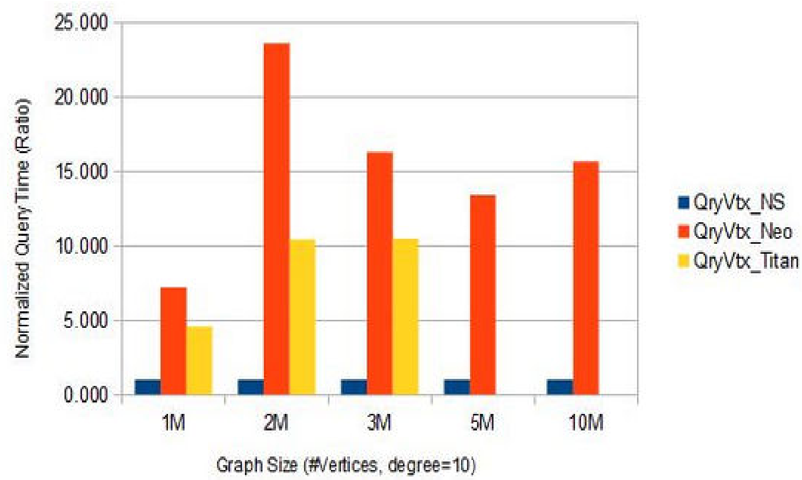
Performance of Graph Primitives -- I

- Graph Primitives: AddVertex, AddEdge, QueryVertex
- Consistent results as what we observed
- Titan over Berkeley DB, not finished in a day.
- On Intel Hashwell (Xeon E5-2697 v2) at 2.7 GHz, 256 GB memory, RedHat Linux



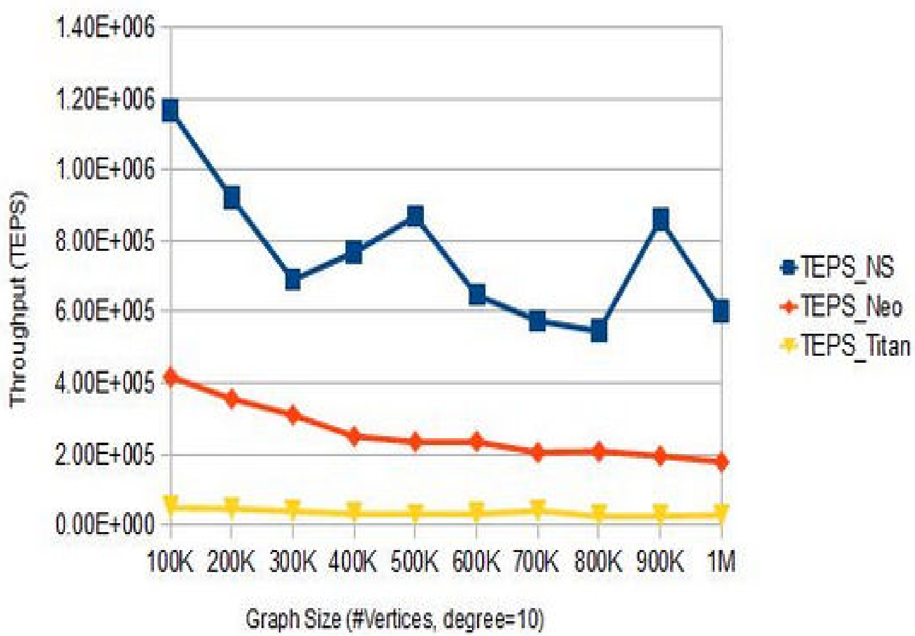


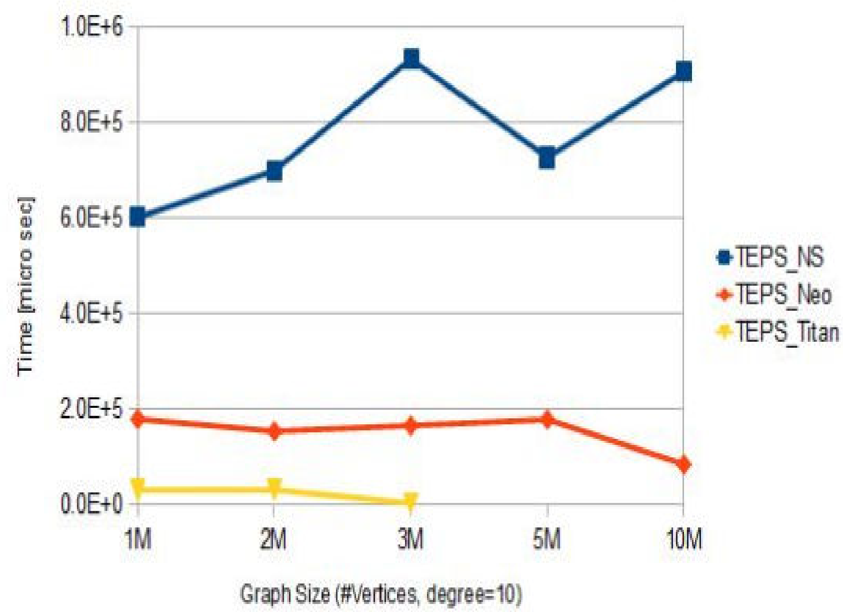




Performance of Graph Primitives - 2

- Many graph analytics requires traversal of subgraphs → critically important
- Experiment setup
  - Start from a vertex and travse 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

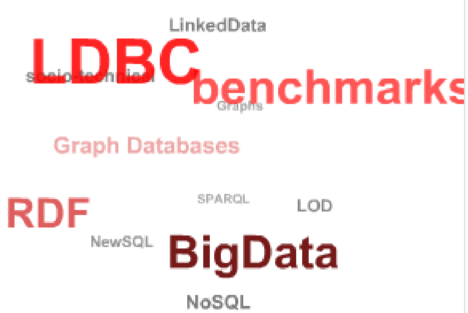




An Emerging Benchmark

LDBC Linked Data Benchmark Council

LDBC is funded by the European Community's Seventh Framework Programme FP7/2007-2013



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

76425 species

Tree of Life by Dr. Yifan Hu

14.8 million tweets

The information diffusion graph of the death of Osama bin Laden by Gilad Lotan

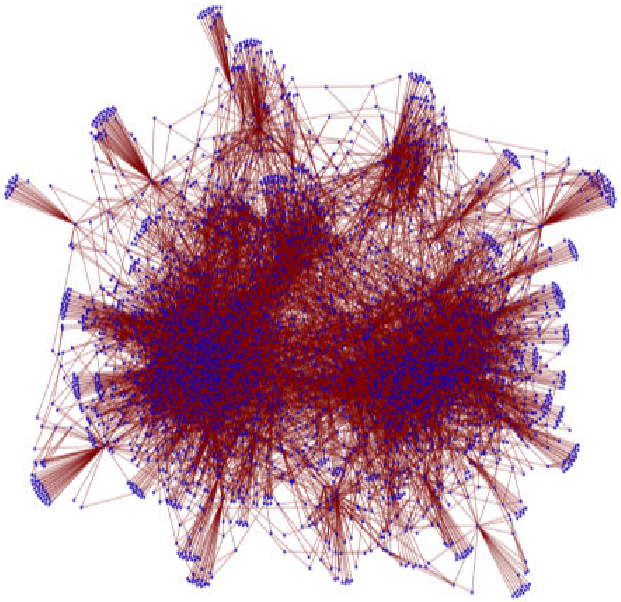
500 million users

Facebook friendship graph by Paul Butler

Challenging Task :

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

Visualization Key Challenges



Visual clutter

How can we encode the information intuitively?

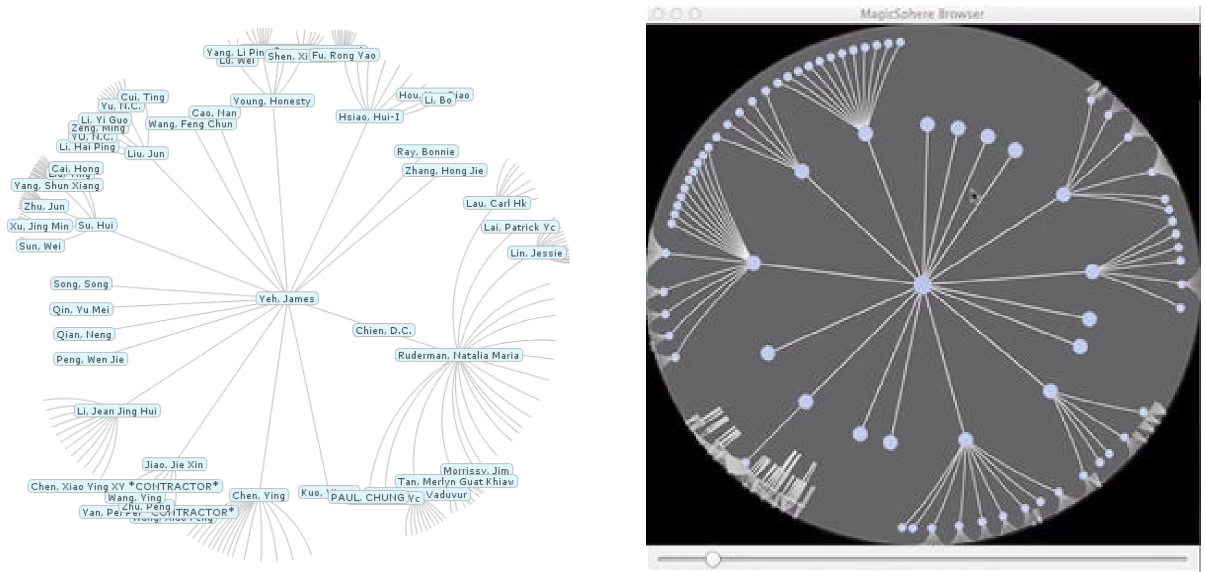
Performance issues

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



Visualization of an organization tree with more than 10,000 nodes

Challenge:

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

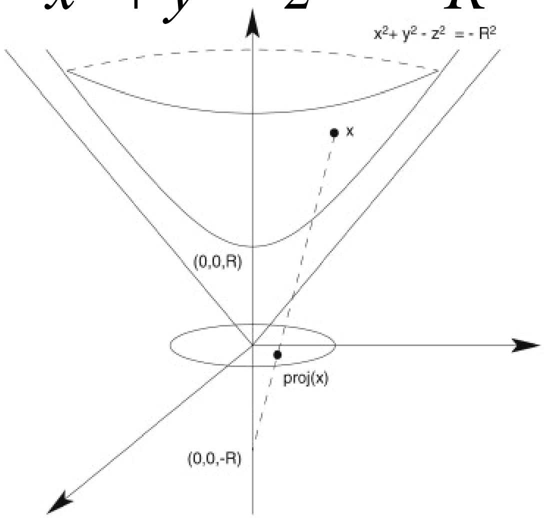
Idea:

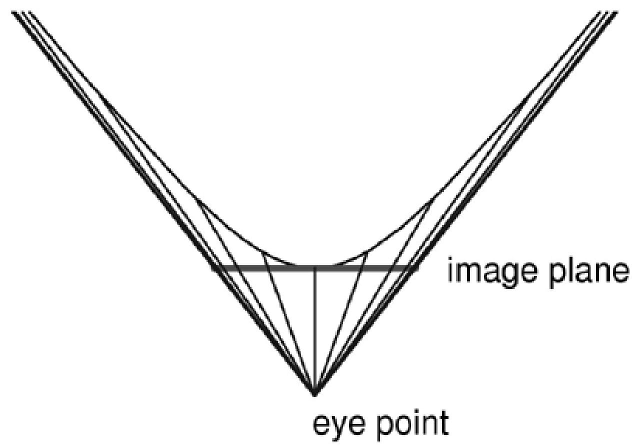
Layout graph onto an infinity plain

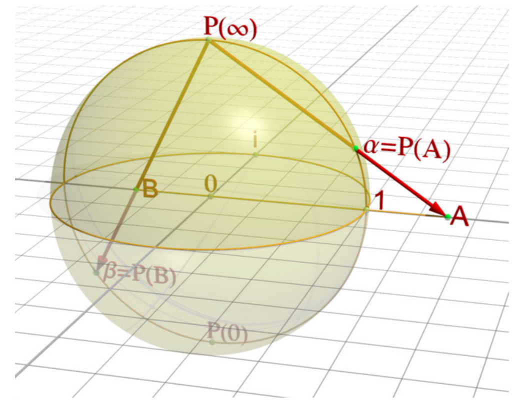
Project the infinity plain into a screen window with finite size

Hyperbolic Projection and Riemann Sphere

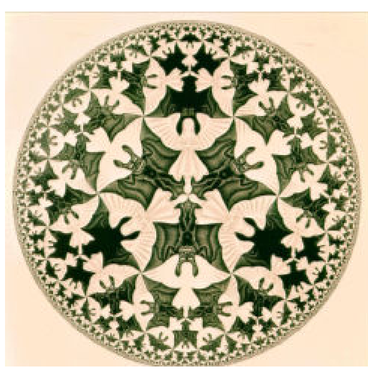
$$x^2 + y^2 - z^2 = -R^2$$

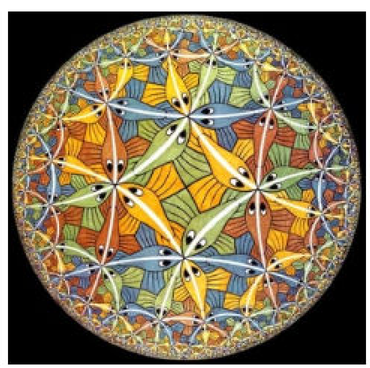




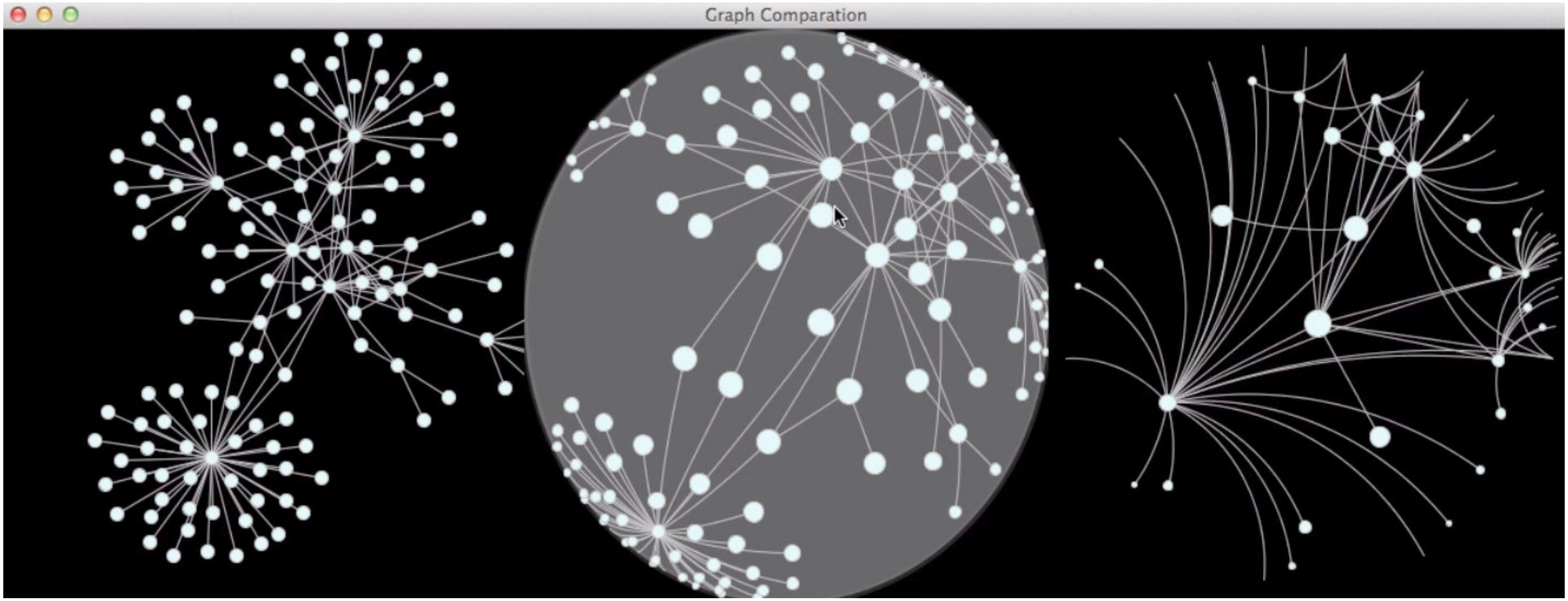


Project the infinity plain onto a finite circular disc called Poincaré disk



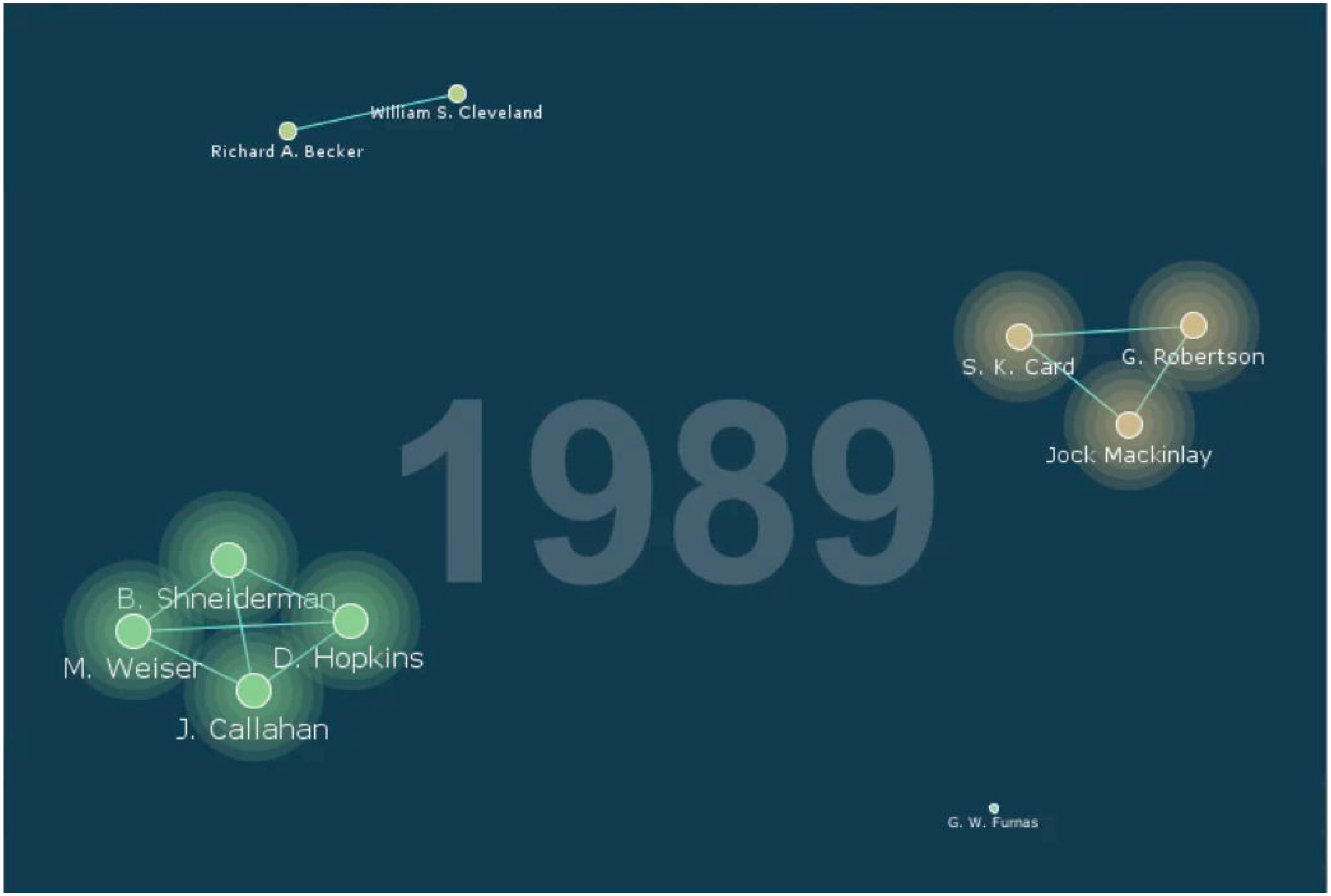


Static Visualization Demo



Conformal: The angle between any two lines on the sphere must be the same between their projected counterparts on the map

Dynamic Visualization Demo



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

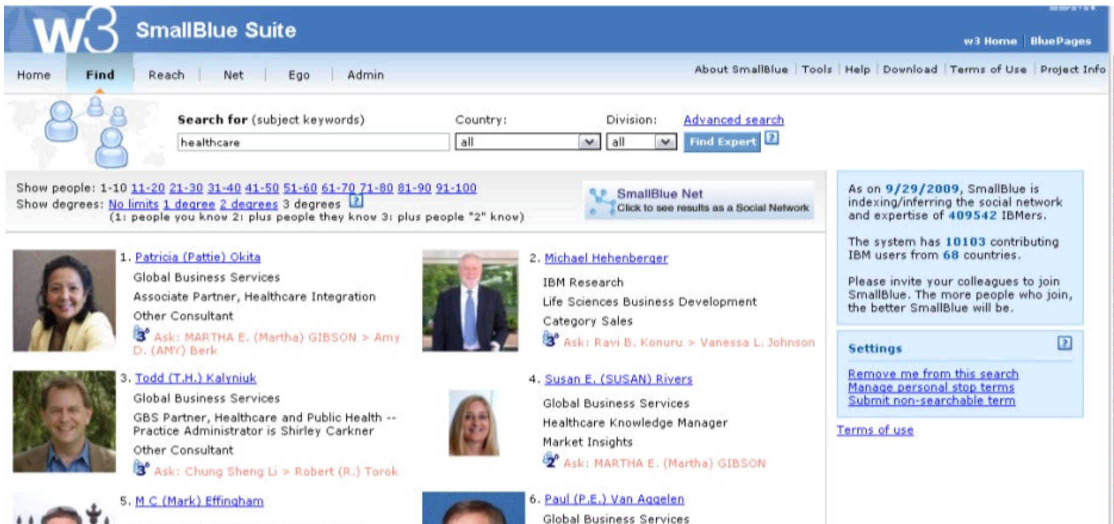
1,000,000+ Lotus Connections (blogs, file sharing, bookmark) data

200,000 people's consulting project & earning data

Shortest Paths

Centralities

Graph Search



– 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 benefi
– APQC (WW leader in Knowledge Practice) April 2013:
"The Industry Leader and Best Practice in Expertise Location"

Dynamic networks of 400,000+ IBMers:

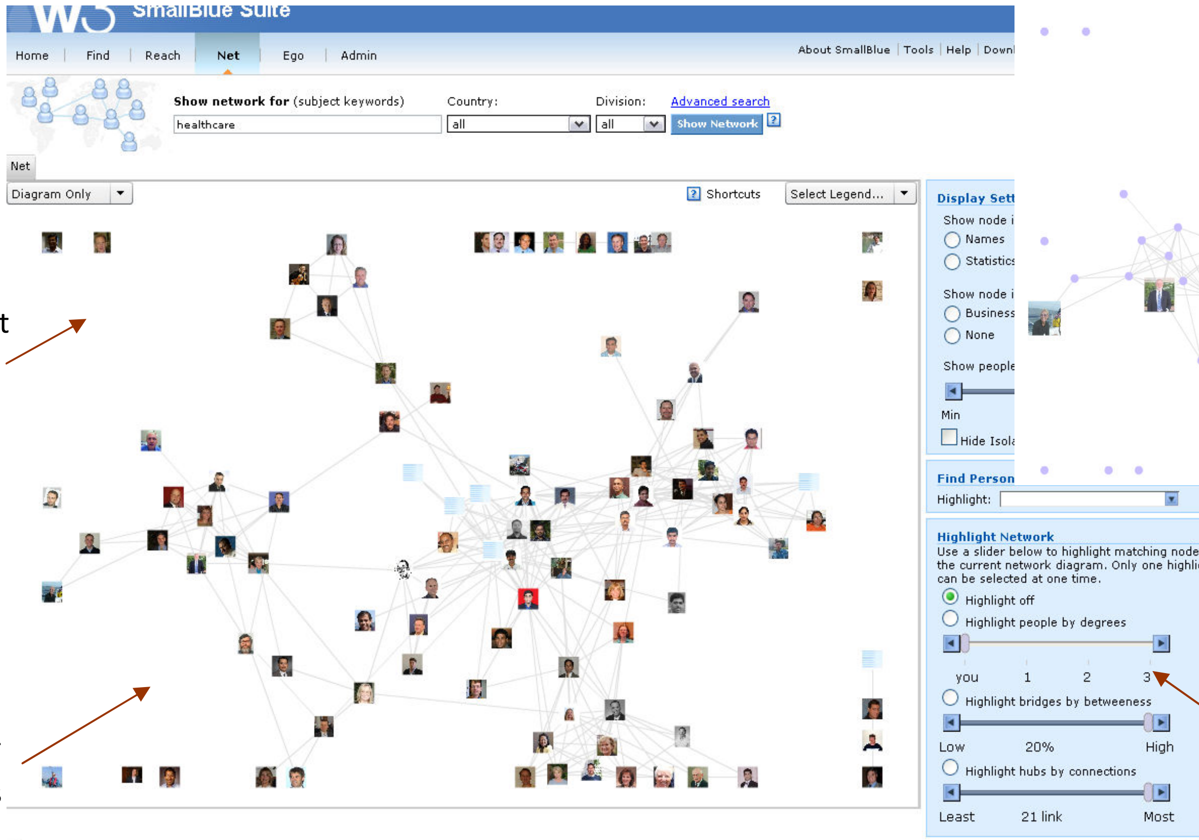
Shortest Paths
Social Capital
Bridges
Hubs
Expertise Search
Graph Search
Graph Recomm.

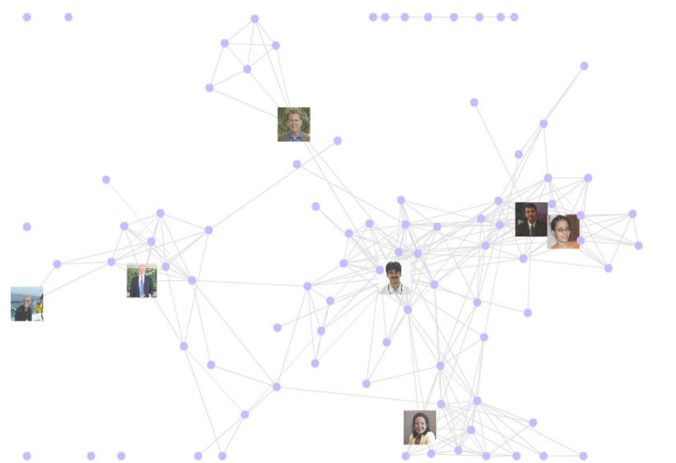
Finding and Ranking Expertise – Social Network Analysis

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

Independent experts on healthcare

A cluster of XYZ experts





Influencers are the one with high 'Betweeness' and 'Degree' value

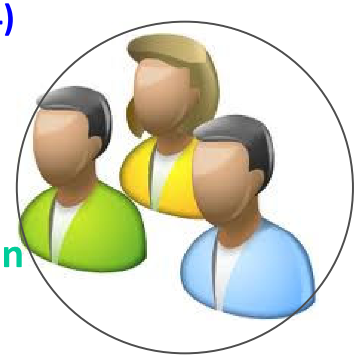
UI to highlight experts based on my social proximity, the number of experts she connects, or the 'social bridges' importance

SmallBlue analyzes underlining dynamic network structure in enterprise

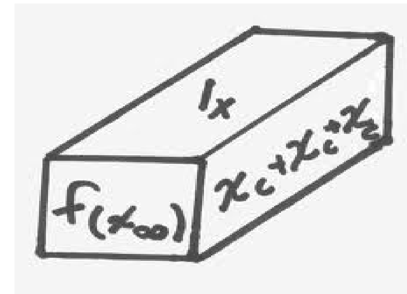
IBM System G Social Media Solution Overview

Actionable Applications (April 2014)

- **Live Monitoring**
- **Anomaly Detection**
- **Impact Trend Analysis**
- **Flow Analytics & Visualization**
- **Person Analytics**
- **Multimedia Analytics**
- *Auto-Counter Messaging*

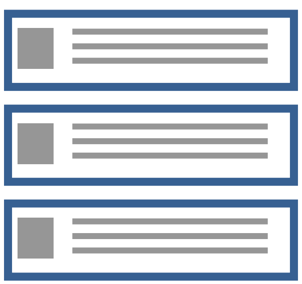


Analytics & Predictive Models

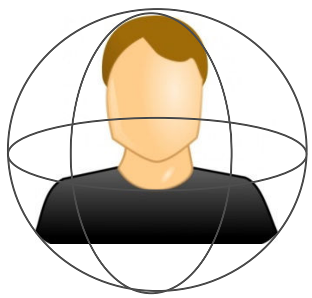


Inferred Social Network Traits

- **Roles**
- **Dynamic Analysis**
- **Topological Analysis**
- **Location Analysis**



Social Media Posts



Inferred Cognitive Traits

(Human Essential)

- **Personality**
- **Needs**
- **Value**
- **Trustworthiness**

(Human Dynamic)

- **Contextual Behavior**
- **Emotional State**

(Information Dynamic)

- **Info Reasoning & Morphing**
- **Visual Sentiment**

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)

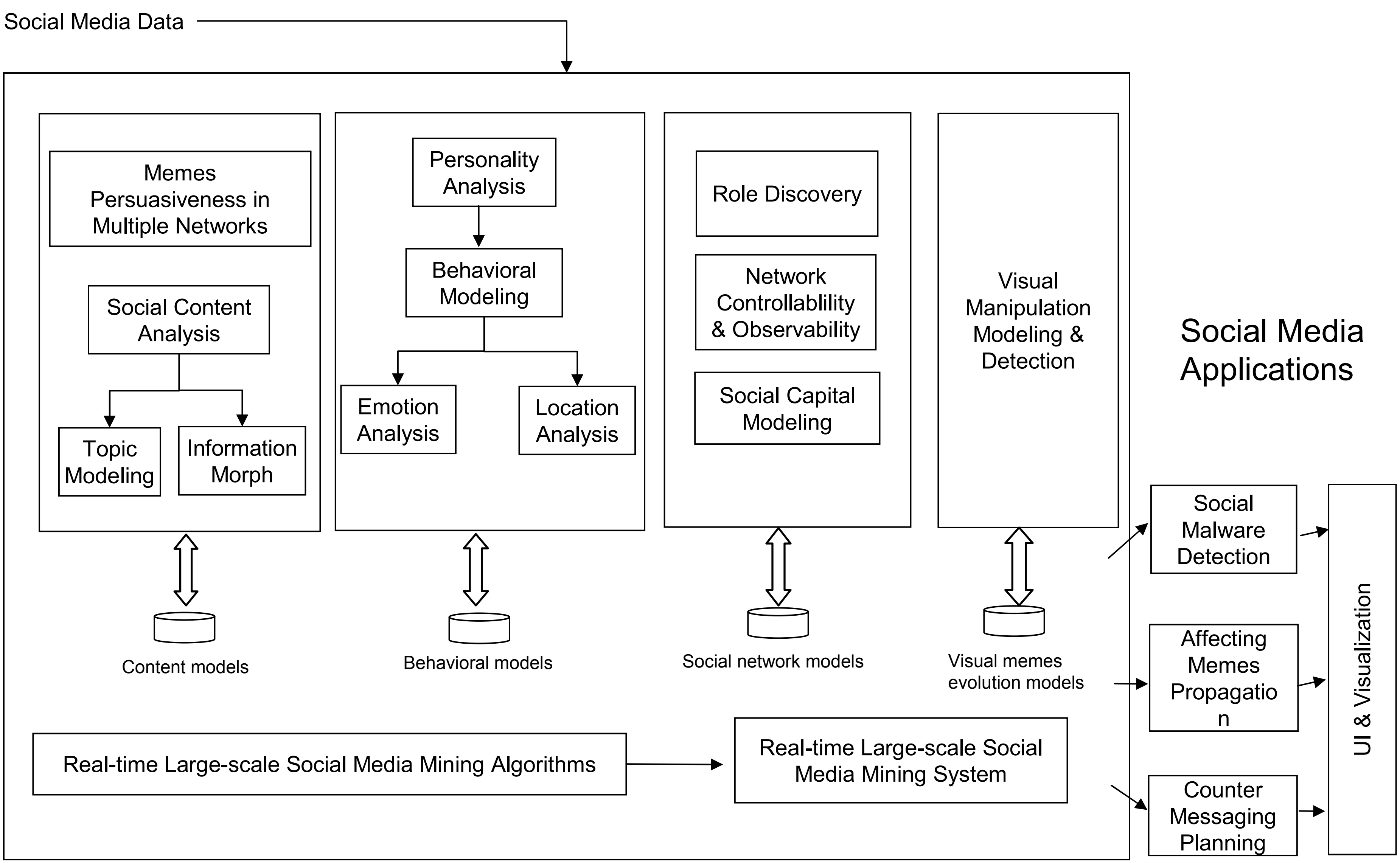
IBM System G Social Media Solution Research Tasks

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 Mining Architecture

System G Team

X-Bank Use Case

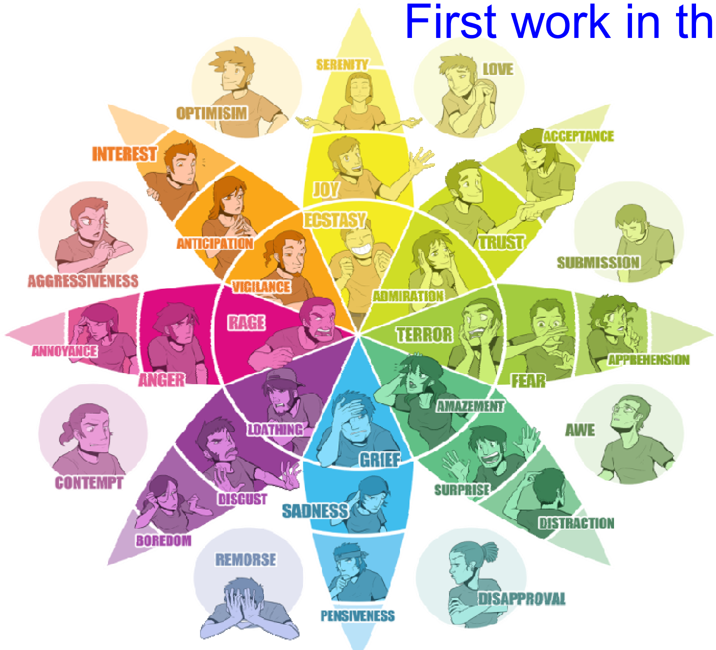
Objective: Detect unexpected social media movements that may impact a major bank's business

- **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

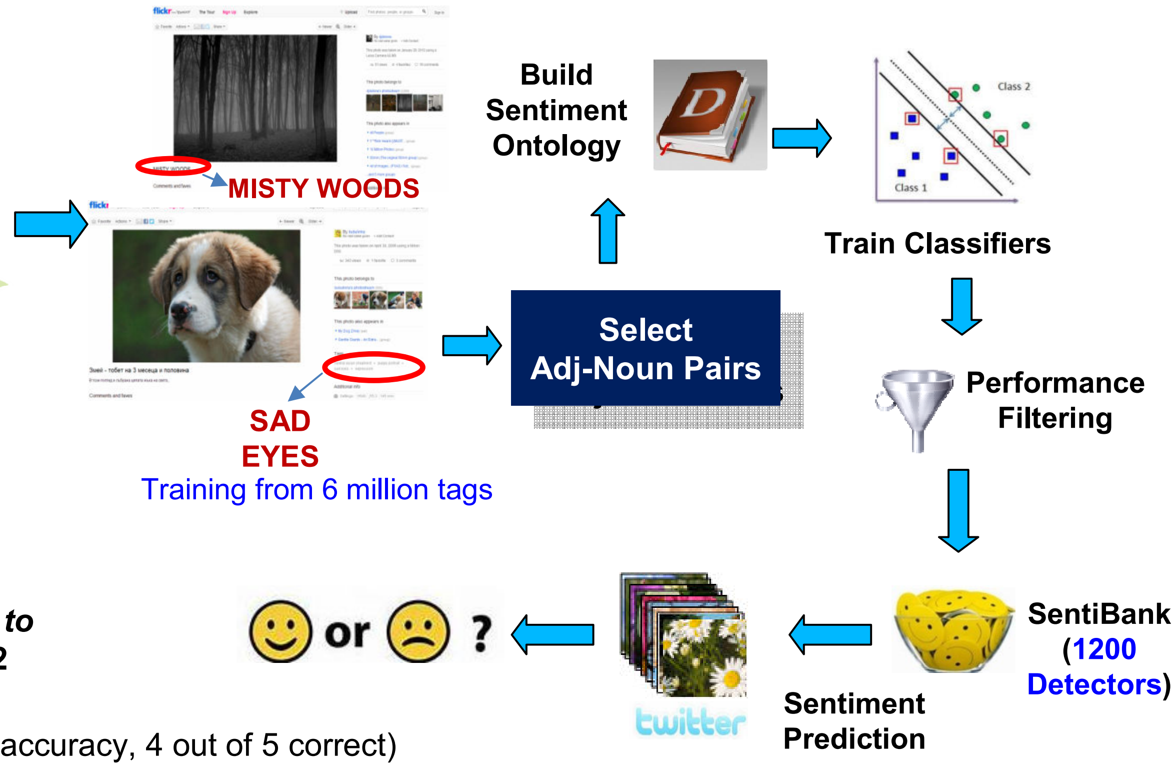
An image tweet (without mentioning "*the bank name*") sparks a lot critiques of their unfair practice

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**



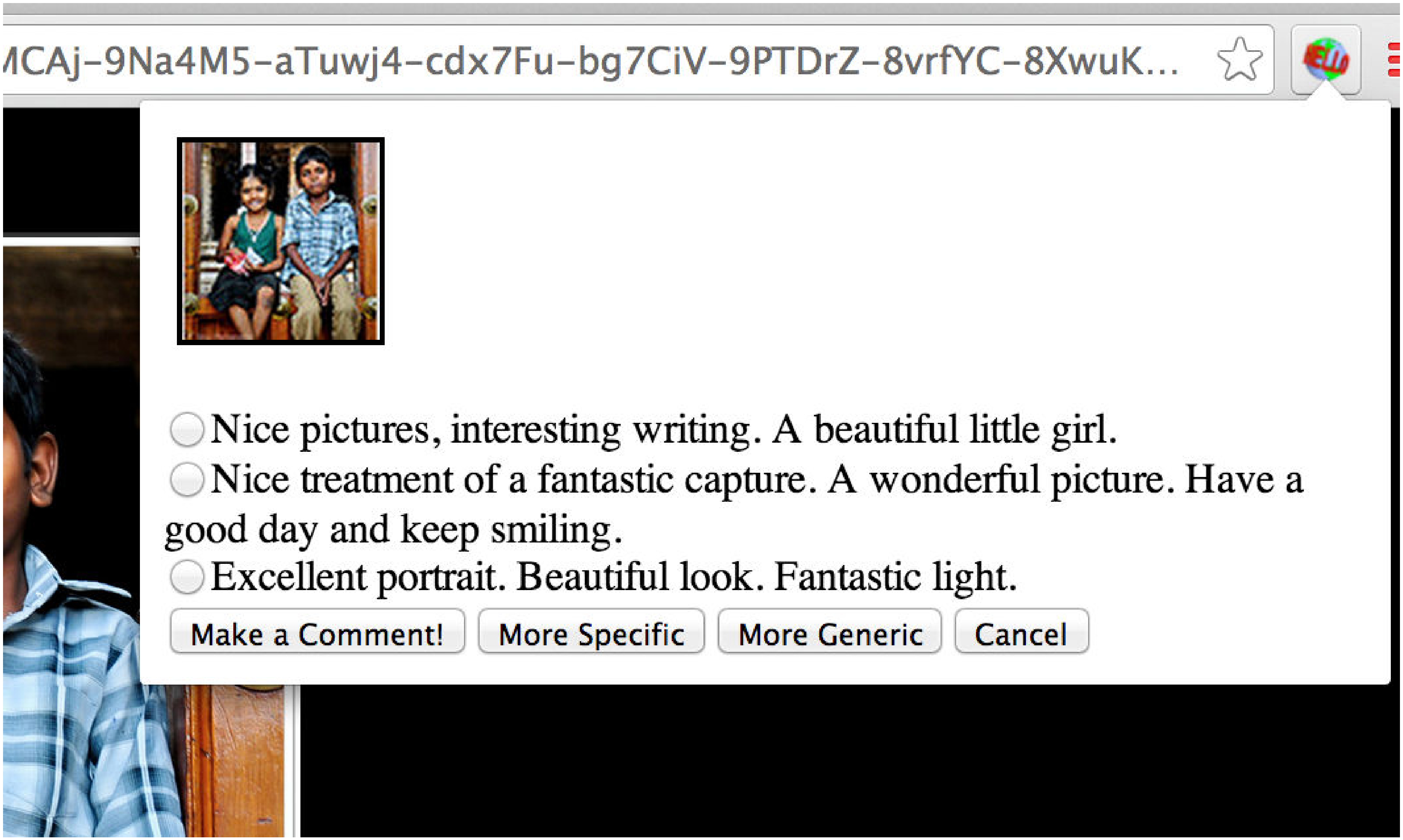
Detection results of "lonely dog" (80% accuracy, 4 out of 5 correct)

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

Experiment on Sentiment Detection Accuracy on Twitter

Text	0.43
Visual	0.70
T+V	0.72

Automatic Comments on Images



Measuring Human Essential Traits in Social Media

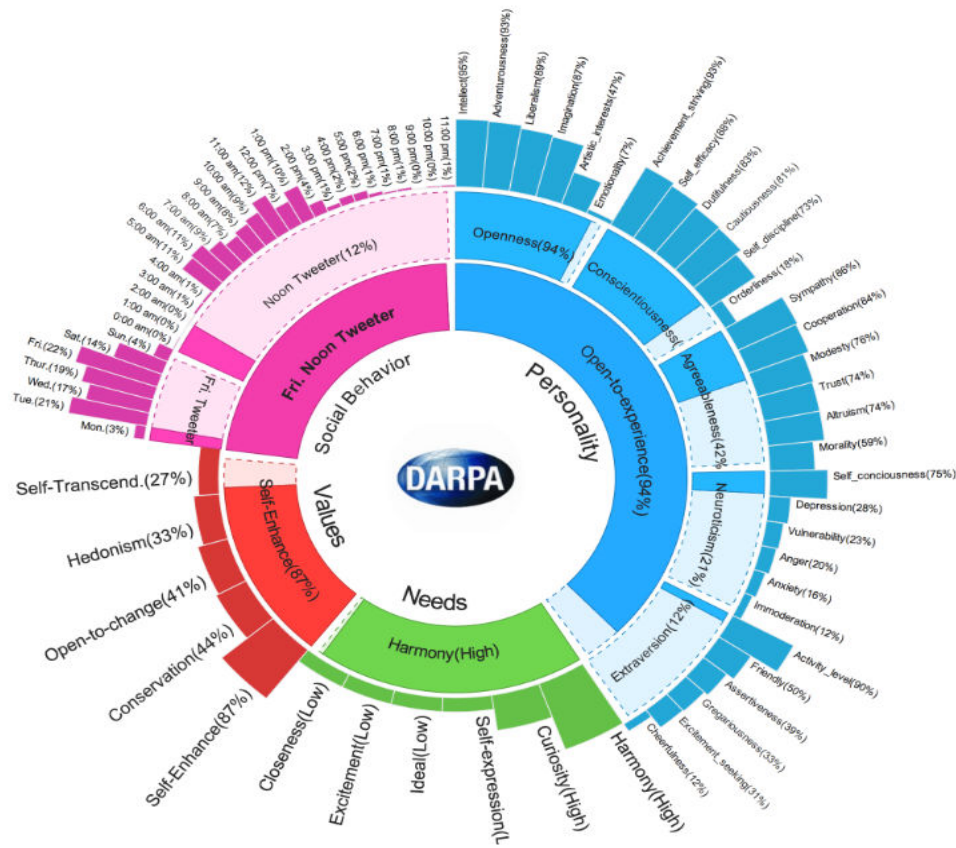
– **Personality:** Mapping personal/organizational social media postings to scores of BIG 5 Personality (*Openness, Conscientiousness, Extraversion, Agreeableness, and Neurocism*)

– **Needs:** Mapping personal/organizational social media postings to scores of *Harmony, Curiousity, 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*.

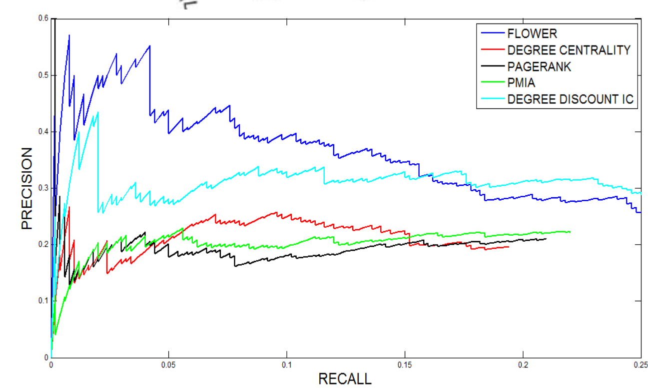
– **Trustingness and Trustworthiness:** Deriving from *interaction and propagation* history between the user and his followers and the people he follows.

– **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

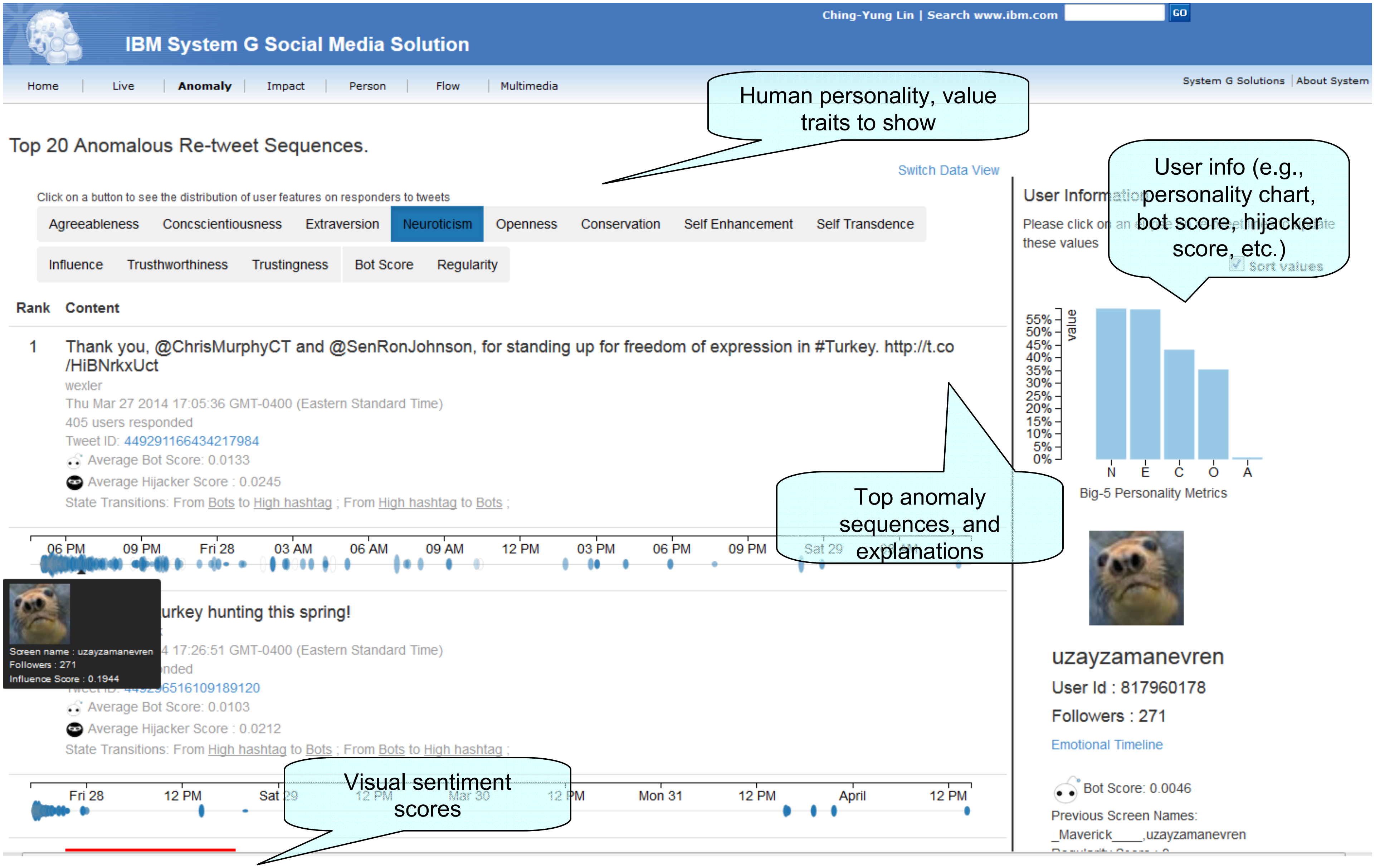
IBM

Live Tweets, Sentiment, Keywords

Growing Influential Between Graphs

Real-Time Translation, Locations, Top Retweets

Anomaly Detection



Impact Analysis

IBM

MERCAN TOPKARA | Search www.ibm.com GO

IBM System G Social Media Solution

Home | Live | Anomaly | **Impact** | Person | Flow | Multimedia

System G Solutions | About System G | News

Real-time hashtag monitoring

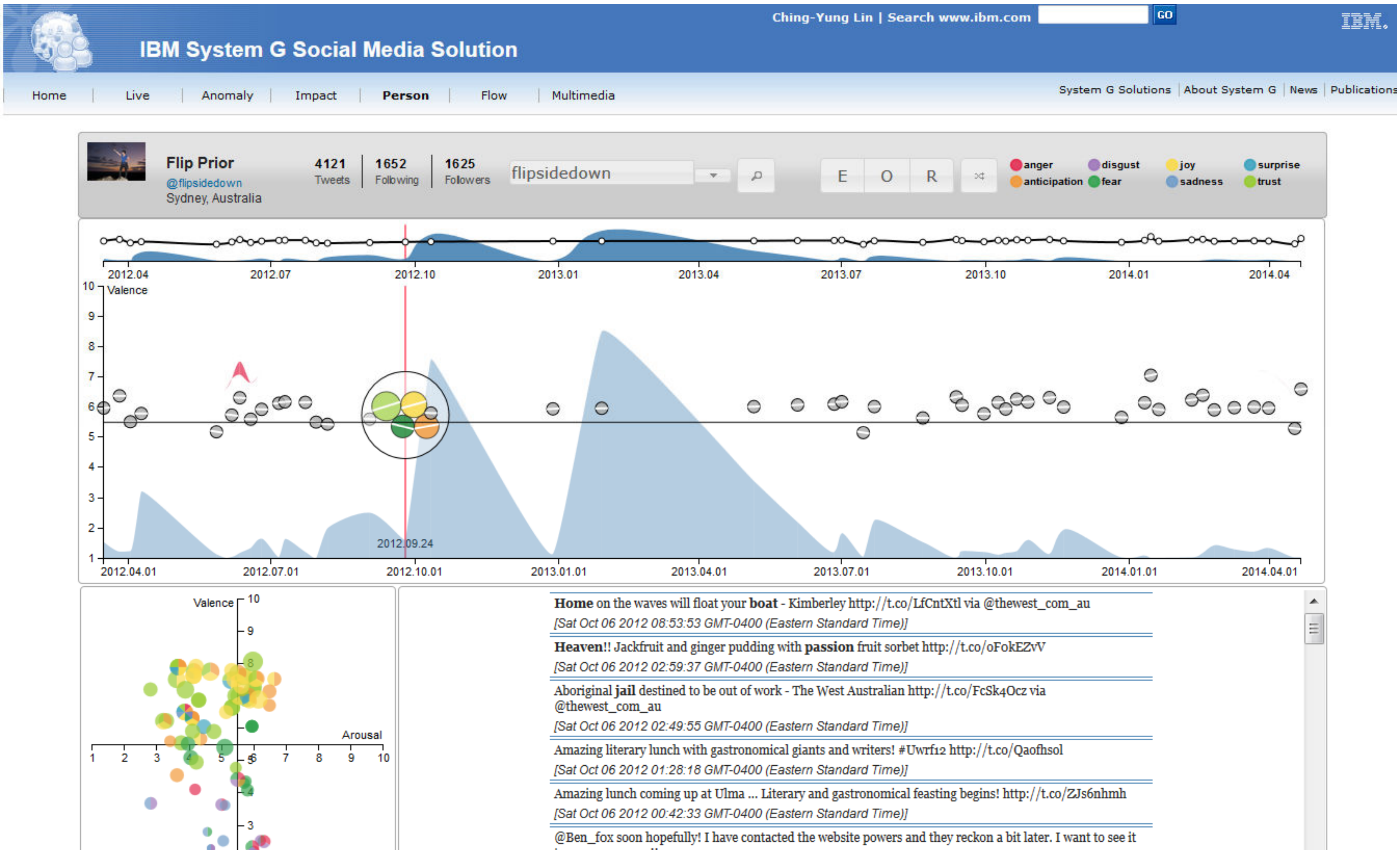
Predicting the business impact of tweet messages grouped by hashtags.

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

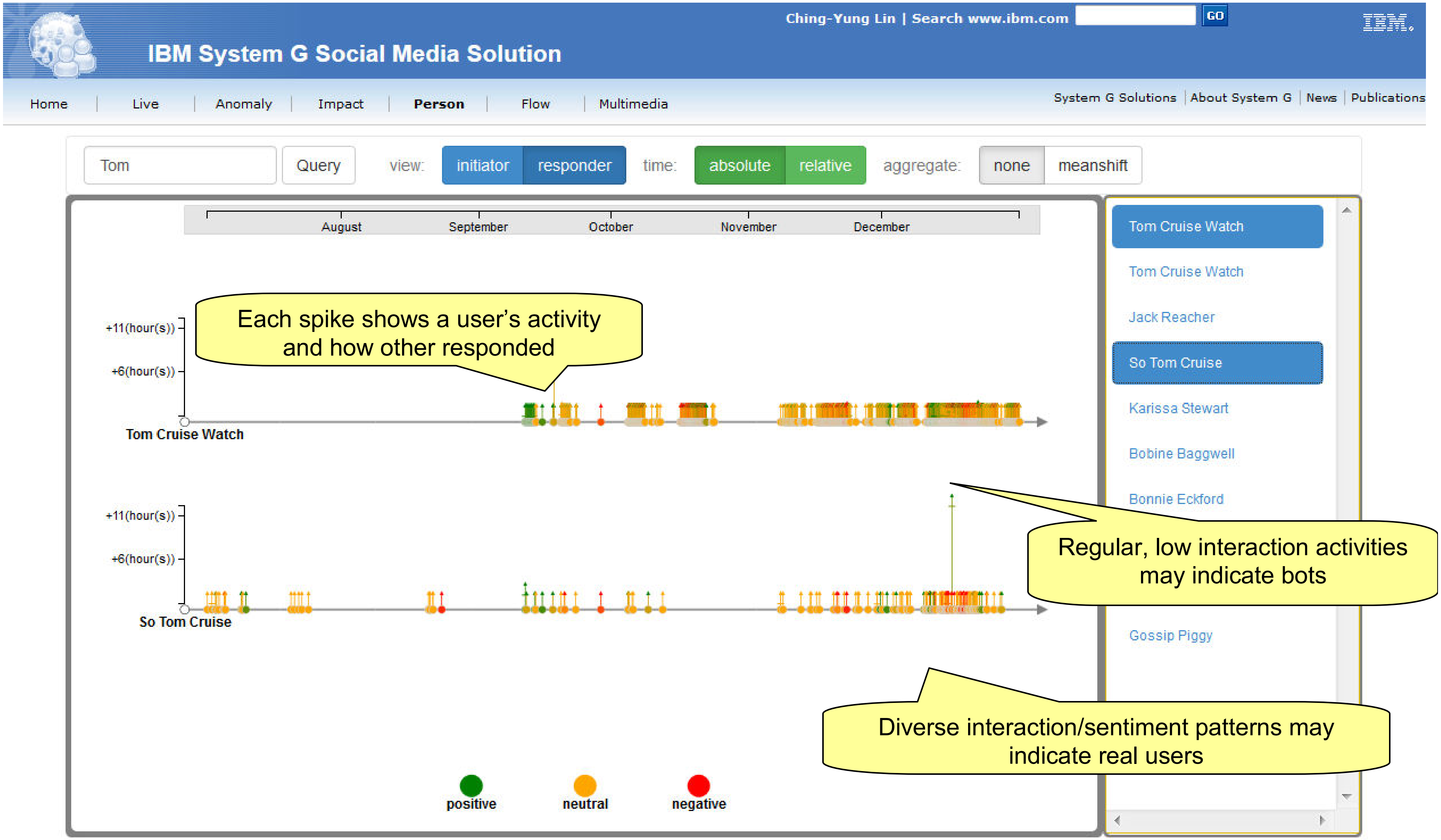
Sort By Date

#	Conversations	Impact	Impact Score	Prediction Detail	First Tweet Time	Last Tweet Time	Duration
10	gulen, turkey	HIGH	35.2	URL	2014-04-21 19:52:21	2014-05-01 13:48:05	233 hours
11	erdogan	HIGH	34.36	URL	2014-04-30 09:14:40	2014-05-01 14:08:12	28 hours
12	erdogan, turkey	HIGH	34.3	URL	2014-04-30 08:45:57	2014-05-01 14:08:12	29 hours
13	istanbul, taksim, turkey	HIGH	34.24	URL	2014-05-01 04:37:22	2014-05-01 14:07:28	9 hours
14	istanbul, mayday, taksim, turkey	HIGH	34.23	URL	2014-05-01 03:28:43	2014-05-01 13:49:07	10 hours
15	israel, news, opinion, turkey, world	HIGH	31.63	URL	2014-04-26 06:01:12	2014-04-26 09:33:51	3 hours
16	news, opinion, turkey, world	HIGH	31.63	URL	2014-04-26 06:01:12	2014-04-26 09:33:51	3 hours

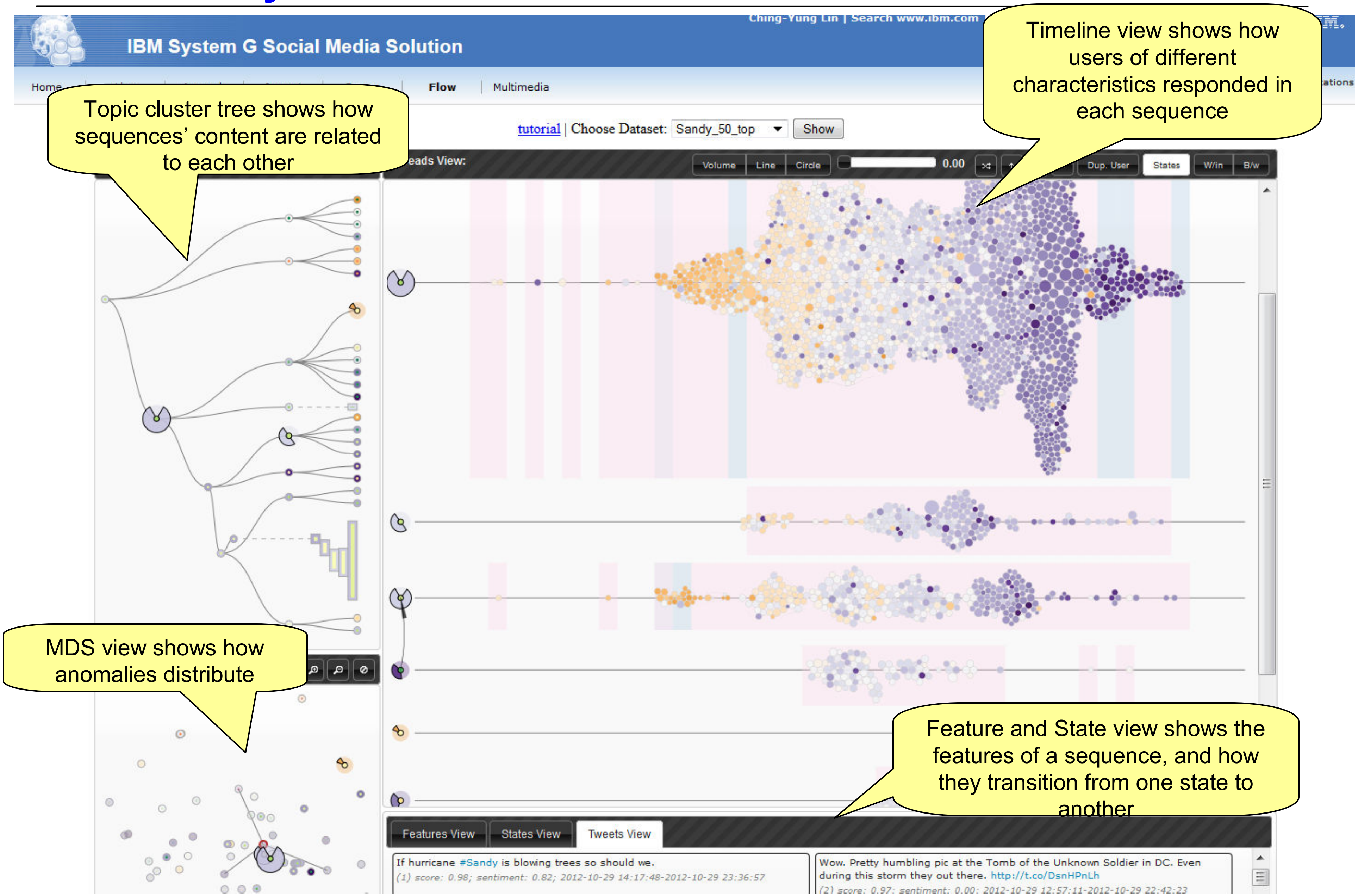
Person Analytics



Person Analytics - II

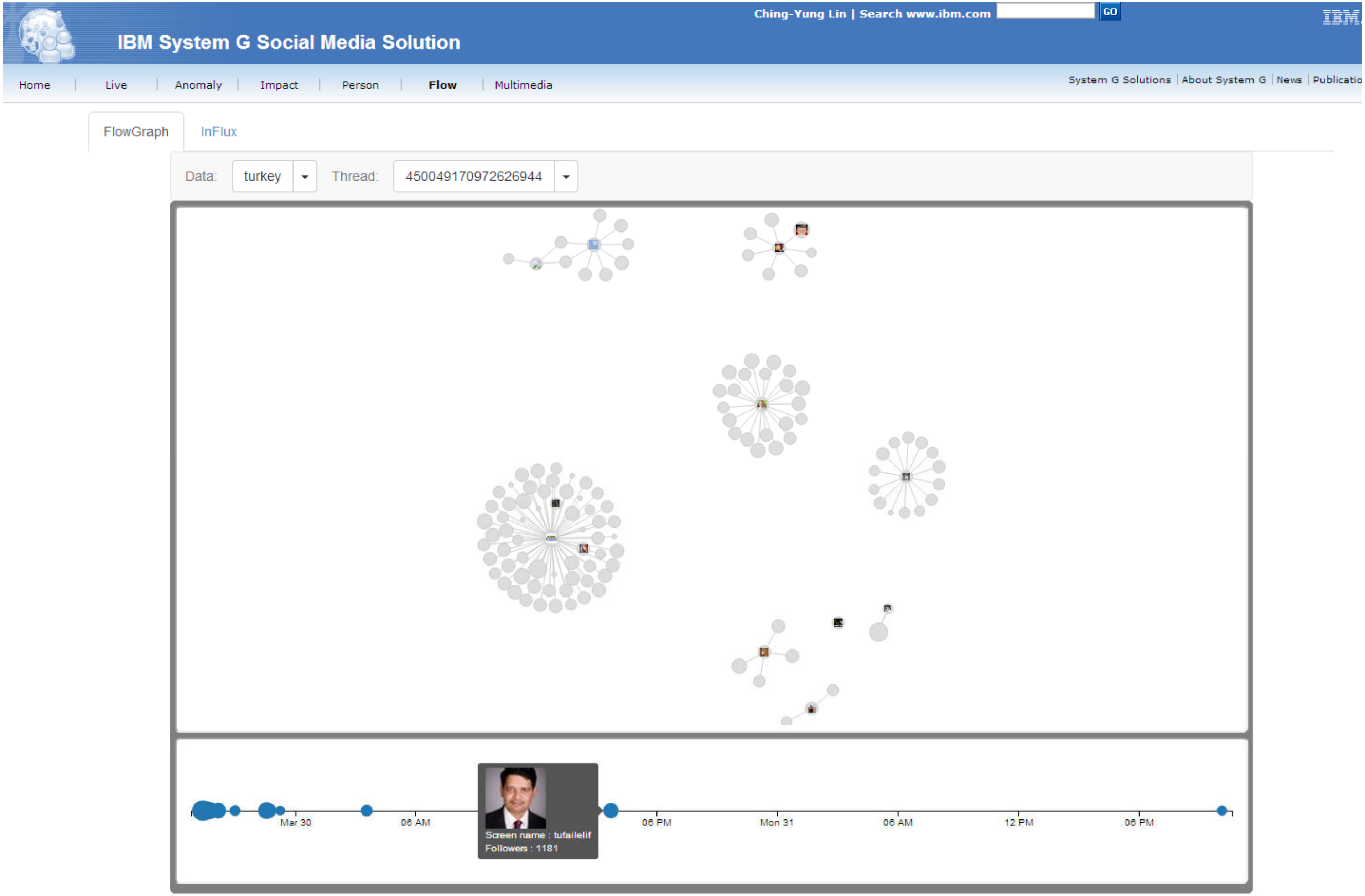


Flow Analytics - I

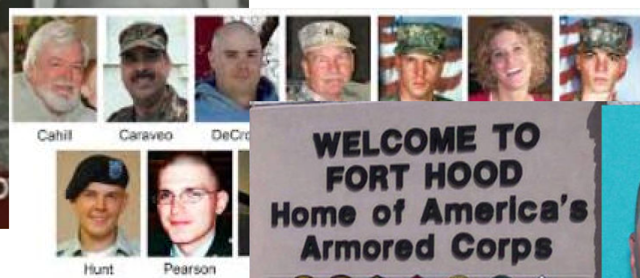


System G Team

Flow Analytics - II



Insider Threats

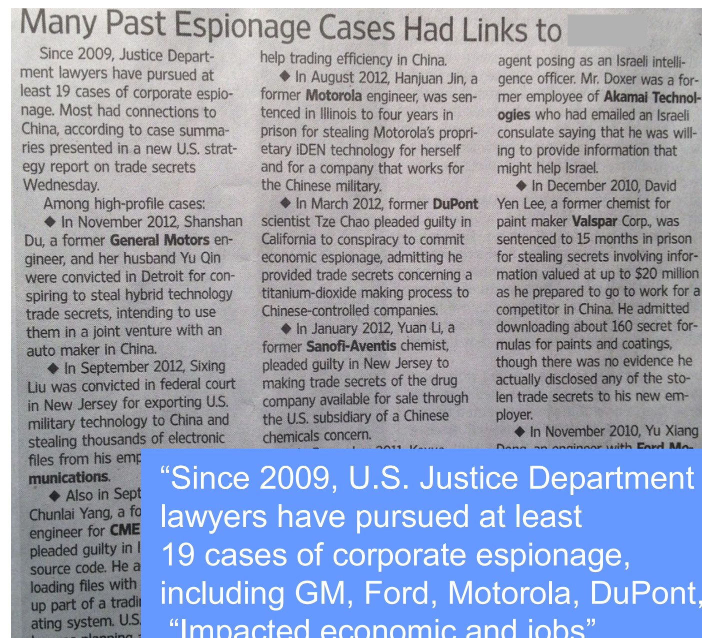


2009

2014

2010

2013



Many Past Espionage Cases Had Links to

Since 2009, Justice Department lawyers have pursued at least 19 cases of corporate espionage. Most had connections to China, according to case summaries presented in a new U.S. strategy report on trade secrets Wednesday.

Among high-profile cases:

◆ In November 2012, Shanshan 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, Sixing Liu was convicted in federal court in New Jersey for exporting U.S. military technology to China and stealing thousands of electronic files from his emp... **munications**.

◆ Also in Sept... Chunlai Yang, a fo... engineer for **CME**... pleaded guilty in I... source code. He a... loading files with... up part of a tradi... ating system. U.S... he was planning a...

...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 Tze Chao pleaded guilty in California to conspiracy to commit economic espionage, admitting he provided trade secrets concerning a titanium-dioxide making process to Chinese-controlled companies.

◆ In January 2012, Yuan Li, a former **Sanofi-Aventis** chemist, pleaded guilty in New Jersey to making trade secrets of the drug company available for sale through the U.S. subsidiary of a Chinese chemicals concern.

...agent posing as an Israeli intelligence officer. Mr. Doxer was a former employee of **Akamai Technologies** who had emailed an Israeli consulate saying that he was willing to provide information that might help Israel.

◆ In December 2010, David Yen Lee, a former chemist for paint maker **Valspar** Corp., was sentenced to 15 months in prison for stealing secrets involving information valued at up to $20 million as he prepared to go to work for a competitor in China. He admitted downloading about 160 secret formulas for paints and coatings, though there was no evidence he actually disclosed any of the stolen trade secrets to his new employer.

◆ In November 2010, Yu Xiang...

"Since 2009, U.S. Justice Department lawyers have pursued at least 19 cases of corporate espionage, including GM, Ford, Motorola, DuPont, …
"Impacted economic and jobs"
– WSJ Feb 21, 2013

Insider Threat comes with sequence of ‘weak signals’

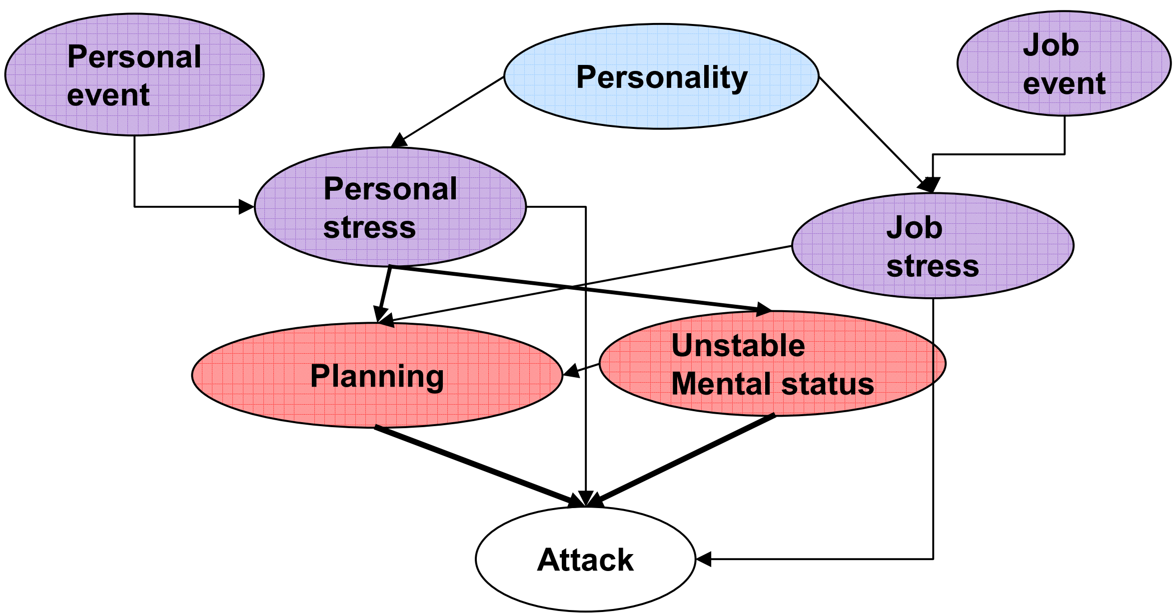
- Personal stress:
  - Gender identity confusion
  - Family change (termination of a stable relationship)
- Job stress:

– 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

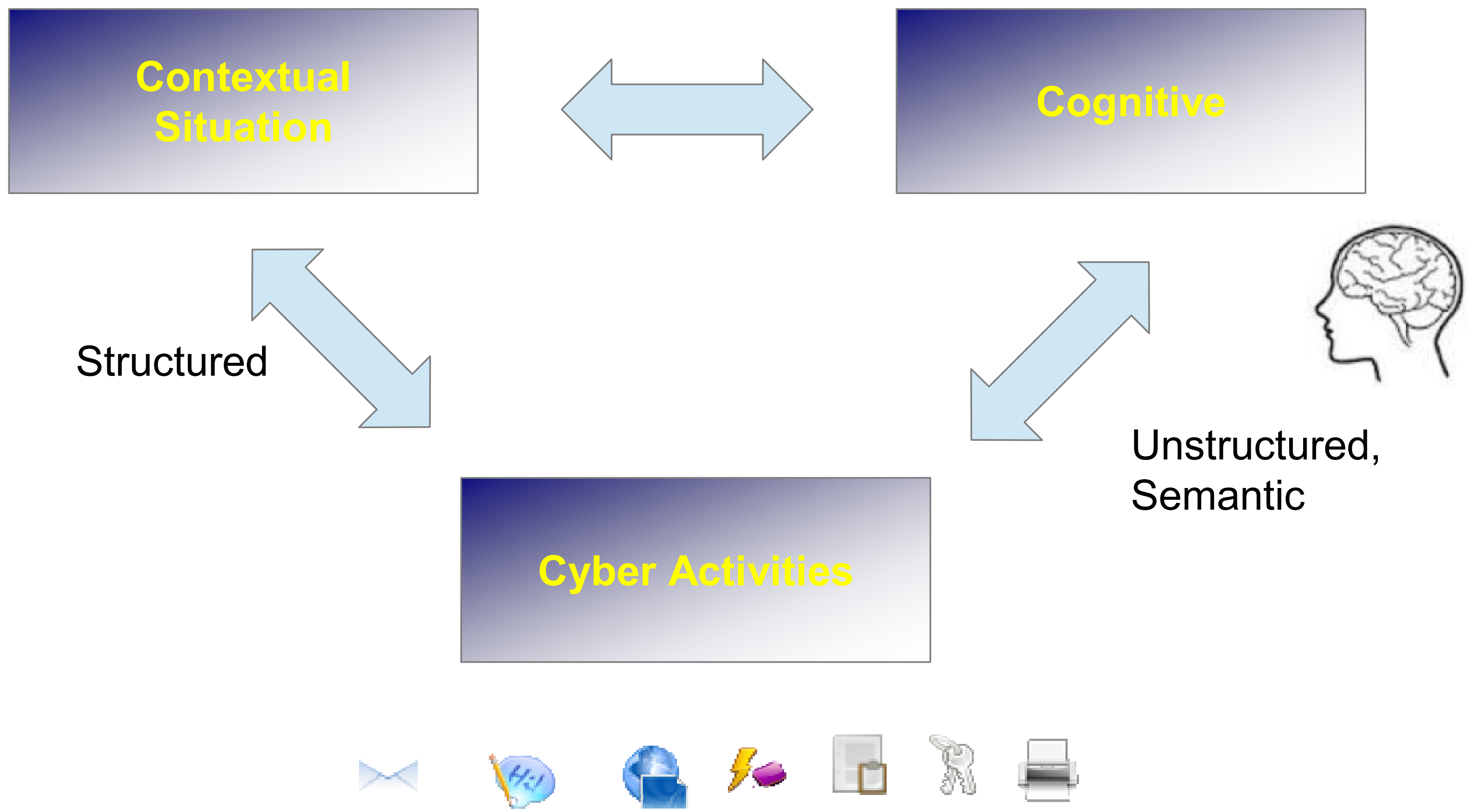


- Attack:

– Brought music CD to work and downloaded/ copied documents onto it with his own account

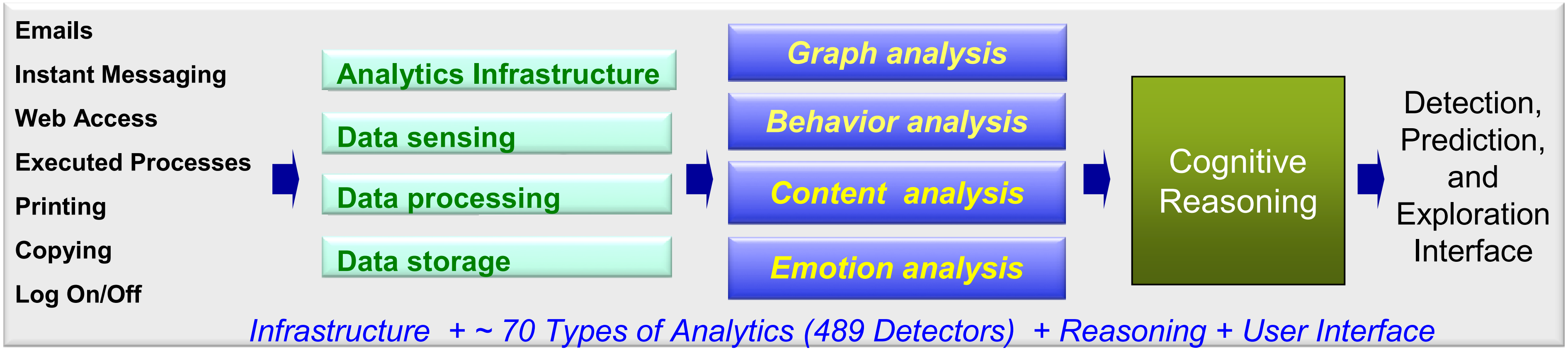
Insider Threat Signals

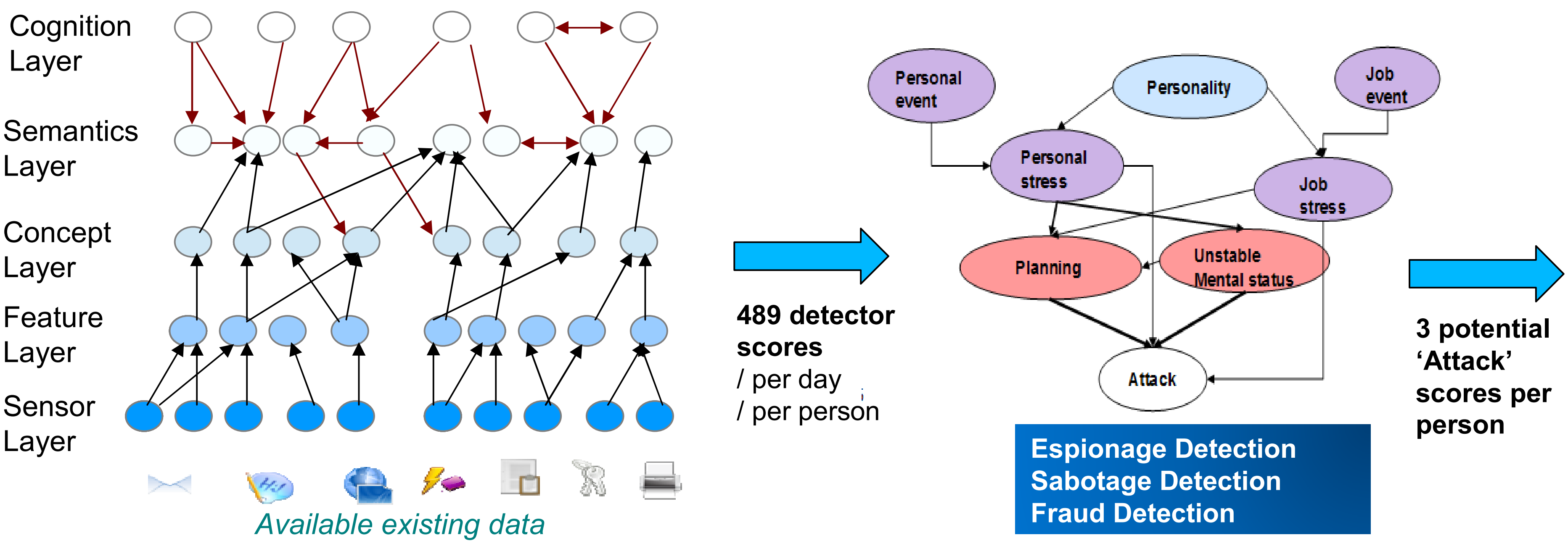
Emerging 'Cognitive Security'



IBM System G Insider Threat Solution (ADAMS) Summary

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



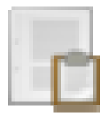


: observations : hidden states

 System G Team

Performance Evaluation and Test Scenarios

Emails



Copy

Log On/Off

Instant Messaging

Process

Printing

Http

- 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)

		Sabotage	Espionage	Fraud	All
Dec	Sabotage (Scenario 12)	4	241	1667	9
	Espionage (Scenario 8)	981	1	120	1
	Fraud (Scenario 13)	1526	454	1	2
Jan	Fraud (Scenario 13)	4230	3367	1	2
	Espionage (Scenario 14)	11	44	574	30
	Fraud (Scenario 5)	4230	1462	3	8
Feb	Espionage (Scenario 14)	1936	73	232	203
	Espionage (Scenario 4)	4101	9	803	26
	Sabotage (Scenario 15)	65	4101	654	181
Mar	Sabotage (Scenario 16)	1	1690	294	1
	Fraud (Scenario 5))	1544	9	5	10
	Espionage (Scenario 4)	4325	11	46	27

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.

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.

24 more Benchmarks Reported in the April 2014 meeting

IBM

- 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.

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

Big Data, Graphs, and System G

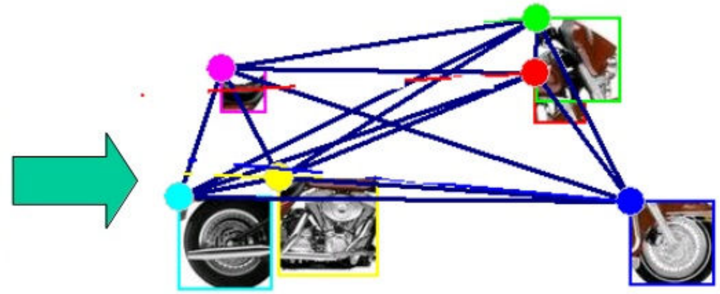
Graph Database and Visualization

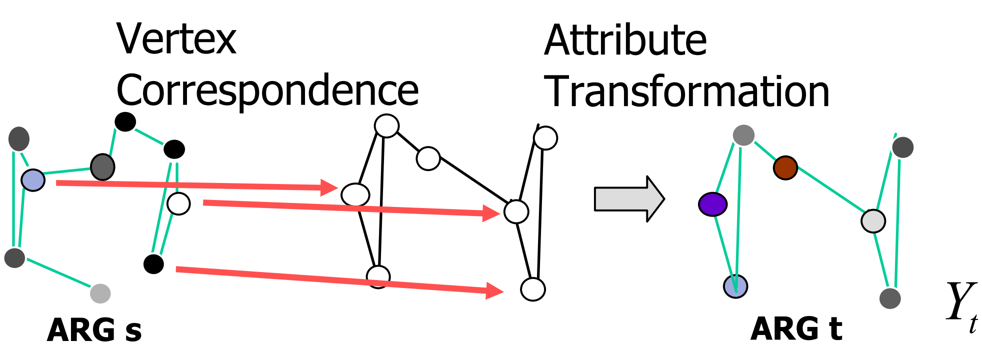
Social Media Solution and Insider Threat Solution

Other Use Cases and Ongoing Researches

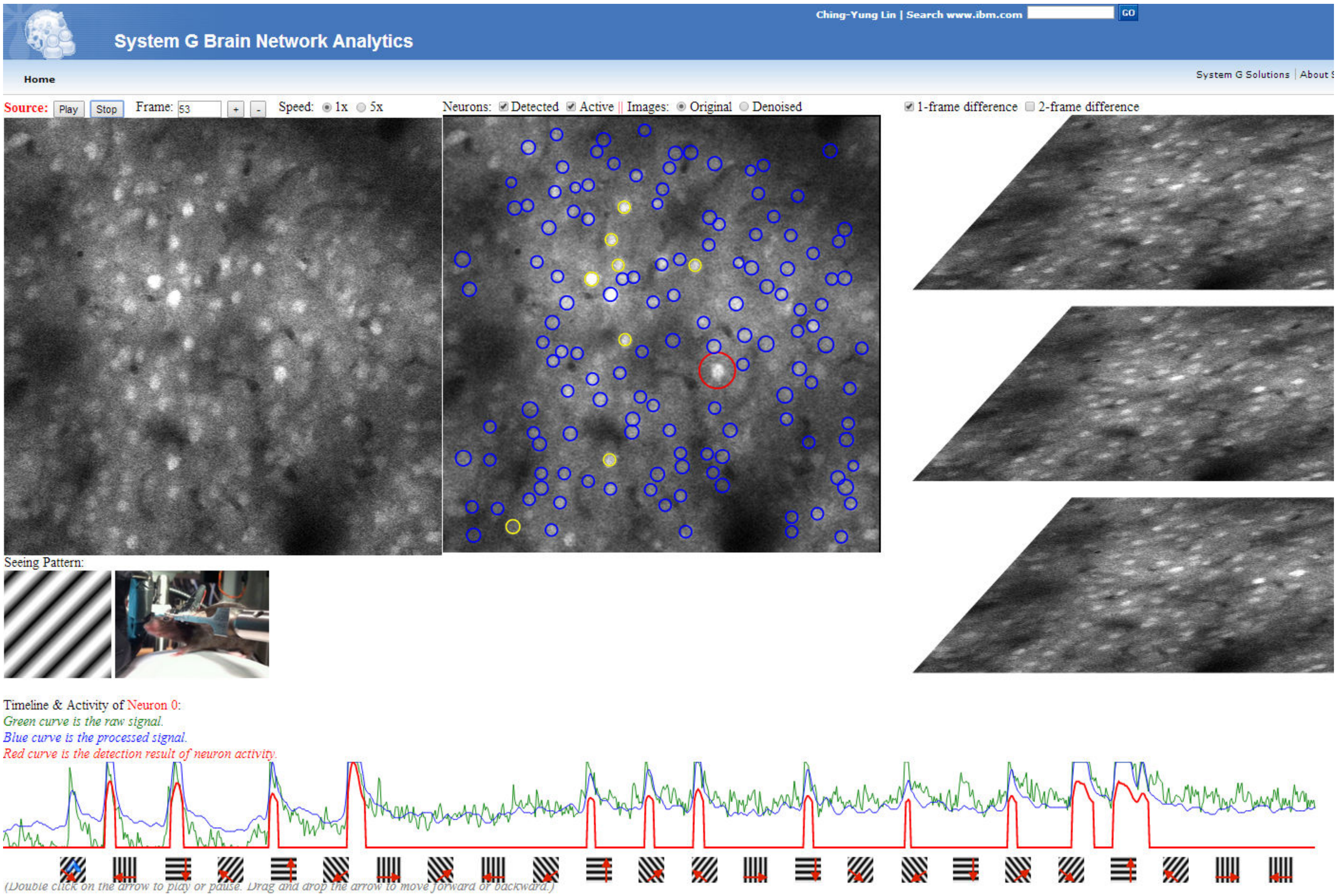
Questions and Discussions

Use Case 18: Graph Analysis for Computer Vision

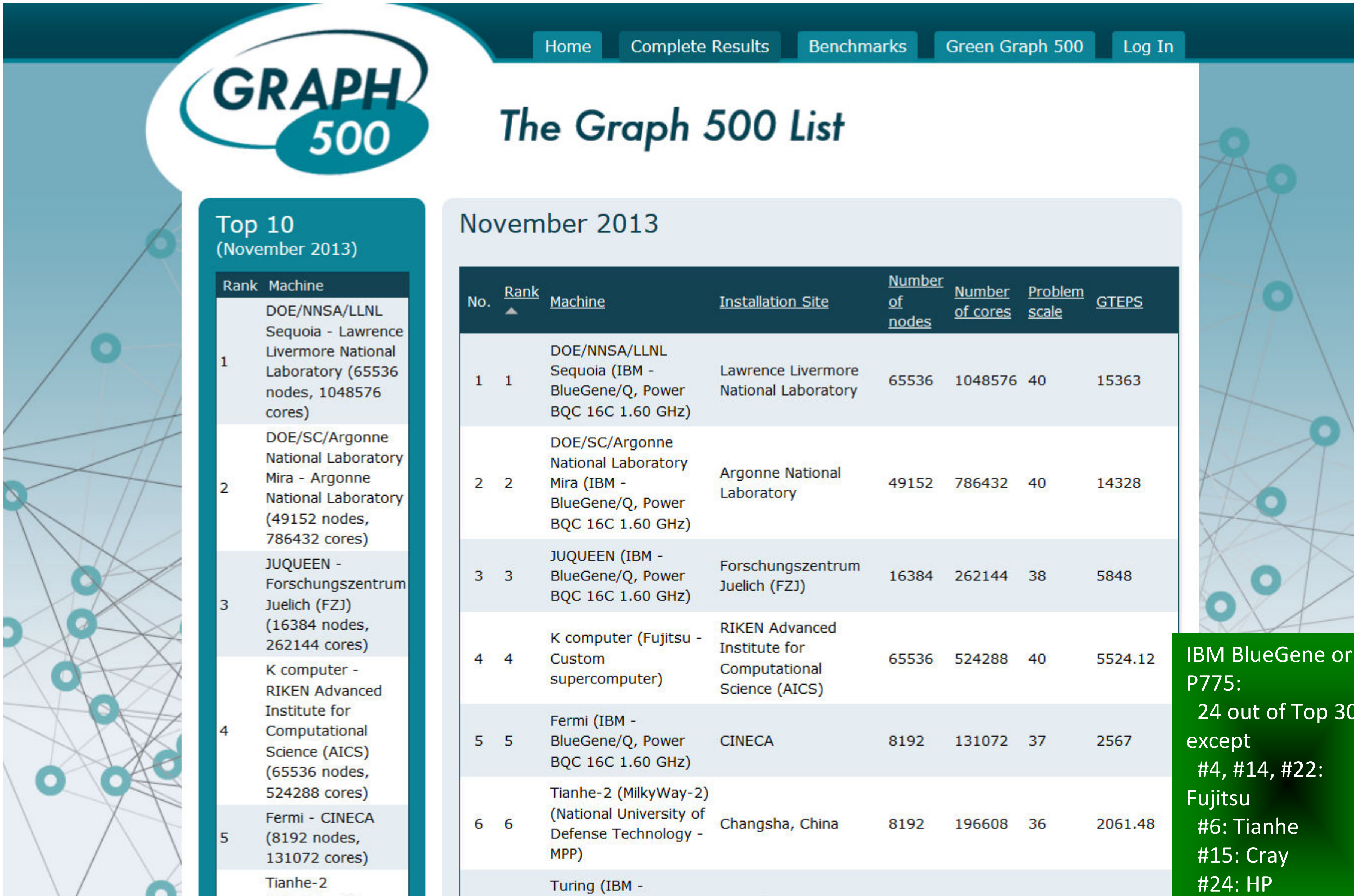




Use Case 21: Understanding Brain Network



Large-scale graph benchmark – Graph 500



Home | Complete Results | Benchmarks | Green Graph 500 | Log In

GRAPH 500

The Graph 500 List

Top 10
(November 2013)

Rank	Machine
1	DOE/NNSA/LLNL Sequoia - Lawrence Livermore National Laboratory (65536 nodes, 1048576 cores)
2	DOE/SC/Argonne National Laboratory Mira - Argonne National Laboratory (49152 nodes, 786432 cores)
3	JUQUEEN - Forschungszentrum Juelich (FZJ) (16384 nodes, 262144 cores)
4	K computer - RIKEN Advanced Institute for Computational Science (AICS) (65536 nodes, 524288 cores)
5	Fermi - CINECA (8192 nodes, 131072 cores)
	Tianhe-2

November 2013

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

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

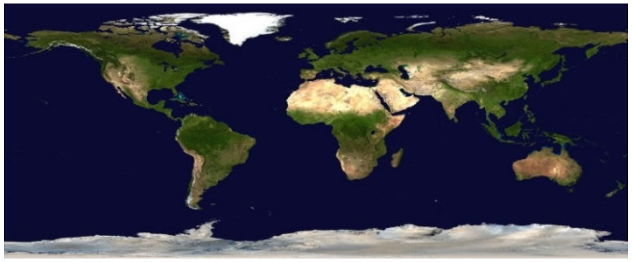
Graph Database and Visualization

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, Minnessota; 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)
  - Middleware & Hardware (*Deep Computing Systems*, Doug Joseph, Fabrizio Petrini, Fabio Checconi)
- **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