

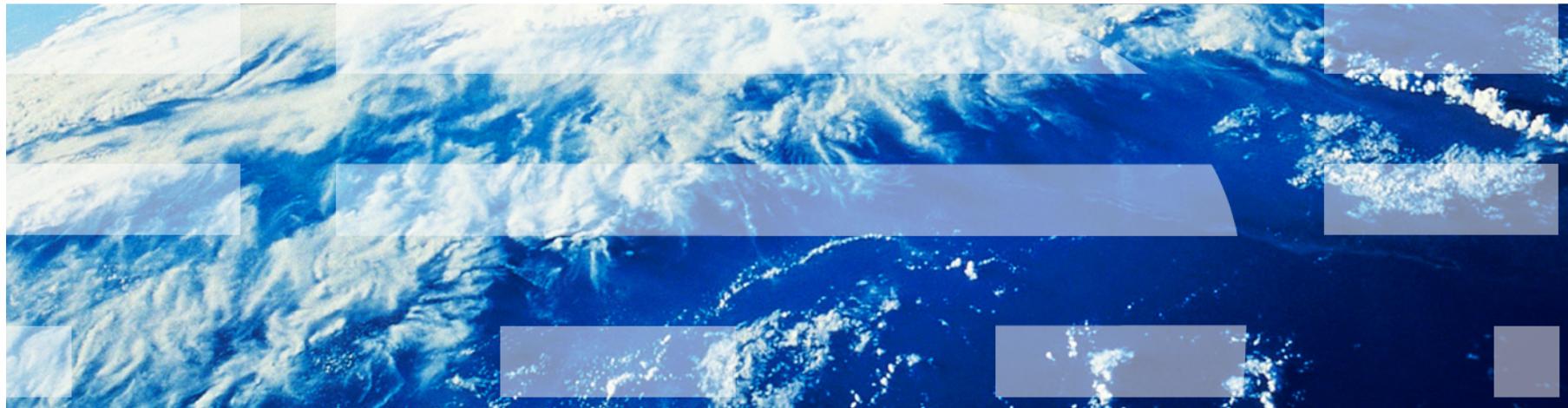
E6893 Big Data Analytics Lecture 1:

Overview of Big Data Analytics

Ching-Yung Lin, Ph.D.

Adjunct Professor, Dept. of Electrical Engineering and Computer Science

IBM Chief Scientist, Graph Computing and Distinguished Researcher



September 10th, 2015

Big Data



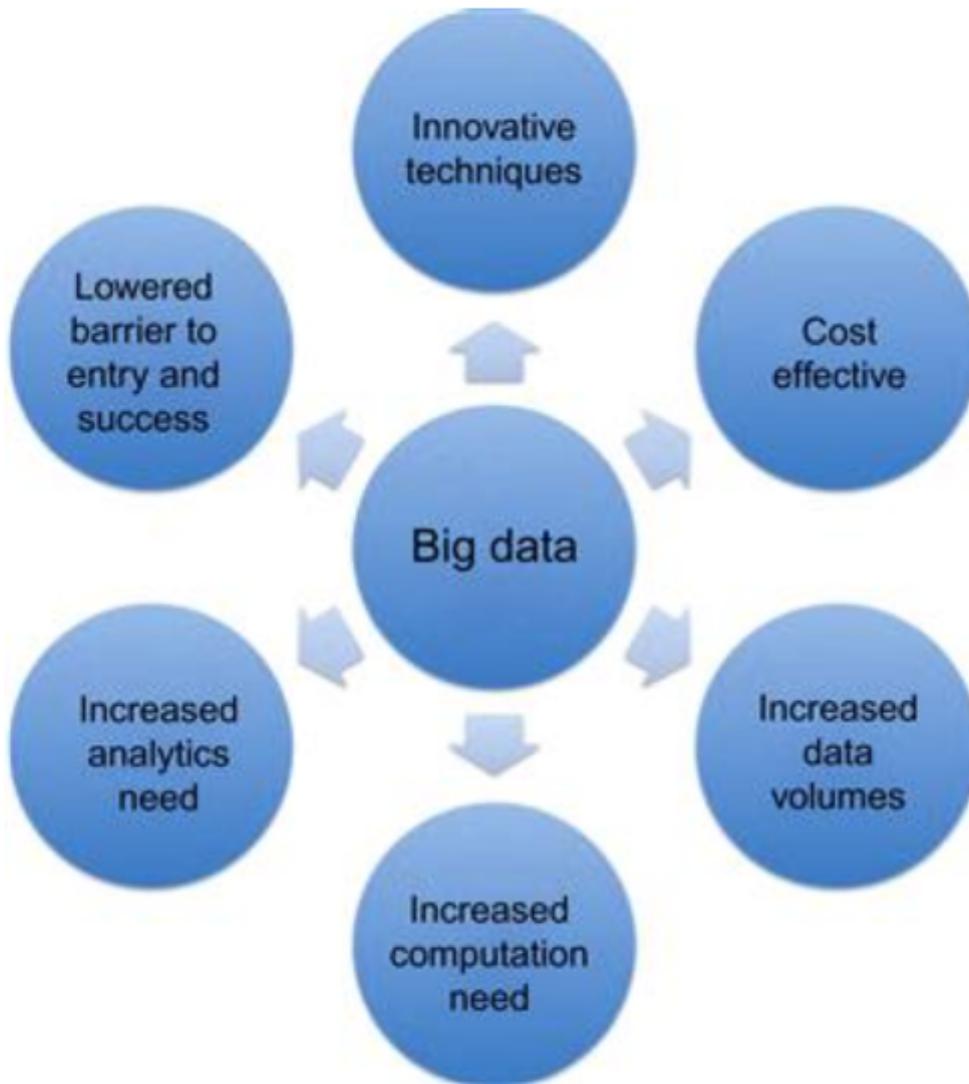
Definition and Characteristics of Big Data

*“Big data is **high-volume**, **high-velocity** and **high-variety** information assets that demand **cost-effective**, **innovative** forms of information processing for **enhanced insight and decision making**.”* -- Gartner

which was derived from:

*“While enterprises struggle to consolidate systems and collapse redundant databases to enable greater operational, analytical, and collaborative consistencies, changing economic conditions have made this job more difficult. E-commerce, in particular, has exploded data management challenges along three dimensions: **volumes, velocity and variety**. In 2001/02, IT organizations much compile a variety of approaches to have at their disposal for dealing each.”* – Doug Laney

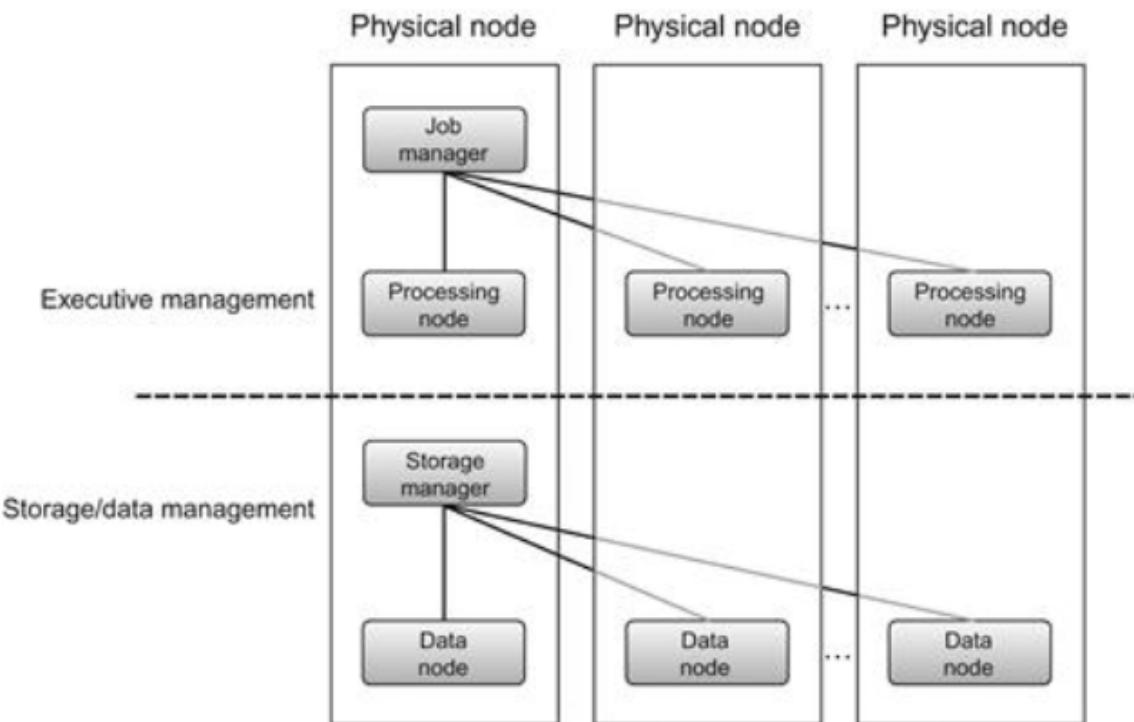
What made Big Data needed?



"Big Data Analytics", David Loshin, 2013

Key Computing Resources for Big Data

- Processing capability: CPU, processor, or node.
- Memory
- Storage
- Network



“Big Data Analytics”, David Loshin, 2013

Techniques towards Big Data

- Massive Parallelism
- Huge Data Volumes Storage
- Data Distribution
- High-Speed Networks
- High-Performance Computing
- Task and Thread Management
- Data Mining and Analytics
- Data Retrieval
- Machine Learning
- Data Visualization

→ Techniques exist for years to decades. Why did Big Data become **hot** now?

Why Big Data now?

- More data are being collected and stored
- Open source code
- Commodity hardware

Aspect	Typical Scenario	Big Data
Application development	Applications that take advantage of massive parallelism developed by specialized developers skilled in high-performance computing, performance optimization, and code tuning	A simplified application execution model encompassing a distributed file system, application programming model, distributed database, and program scheduling is packaged within Hadoop, an open source framework for reliable, scalable, distributed, and parallel computing
Platform	Uses high-cost massively parallel processing (MPP) computers, utilizing high-bandwidth networks, and massive I/O devices	Innovative methods of creating scalable and yet elastic virtualized platforms take advantage of clusters of commodity hardware components (either cycle harvesting from local resources or through cloud-based utility computing services) coupled with open source tools and technology
Data management	Limited to file-based or relational database management systems (RDBMS) using standard row-oriented data layouts	Alternate models for data management (often referred to as NoSQL or “Not Only SQL”) provide a variety of methods for managing information to best suit specific business process needs, such as in-memory data management (for rapid access), columnar layouts to speed query response, and graph databases (for social network analytics)
Resources	Requires large capital investment in purchasing high-end hardware to be installed and managed in-house	The ability to deploy systems like Hadoop on virtualized platforms allows small and medium businesses to utilize cloud-based environments that, from both a cost accounting and a practical perspective, are much friendlier to the bottom line

“Big Data Analytics”, David Loshin, 2013

Big Data Analytics

From Strategic Planning to
Enterprise Integration with Tools,
Techniques, NoSQL, and Graph



David Loshin

- Chapter 1: Market and Business Drivers for Big Data Analysis
- Chapter 2: Business Problems Suited to Big Data Analytics
- Chapter 3: Achieving Organizational Alignment for Big Data Analytics
- Chapter 4: Developing a Strategy for Integrating Big Data Analytics into the Enterprise
- Chapter 5: Data Governance for Big Data Analytics: Considerations for Data Policies and Processes
- Chapter 6: Introduction to High-Performance Appliances for Big Data Management
- Chapter 7: Big Data Tools and Techniques
- Chapter 8: Developing Big Data Applications
- Chapter 9: NoSQL Data Management for Big Data
- Chapter 10: Using Graph Analytics for Big Data
- Chapter 11: Developing the Big Data Roadmap

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

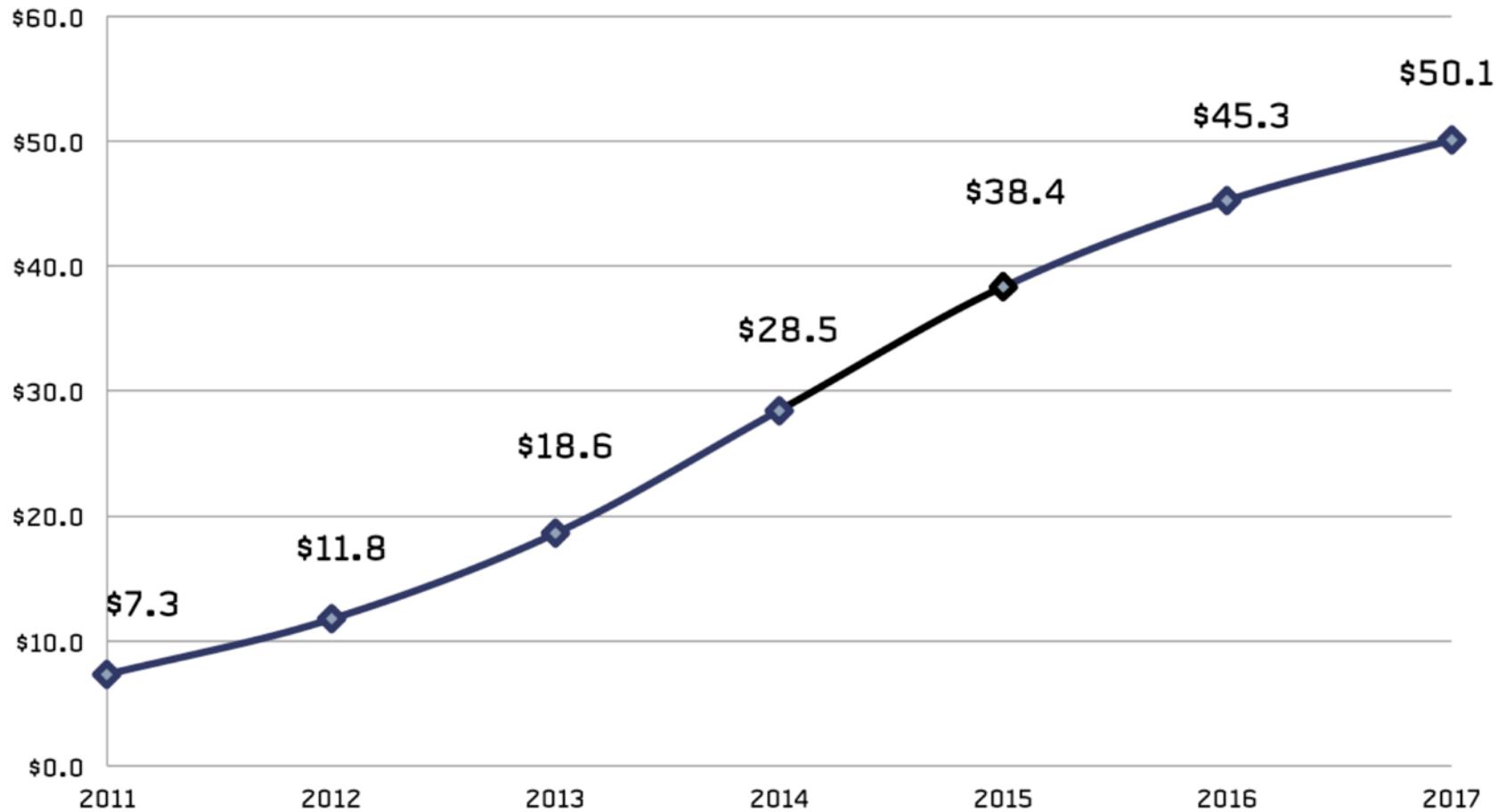


Data Warehouse Augmentation

Integrate big data and data warehouse capabilities to increase operational efficiency



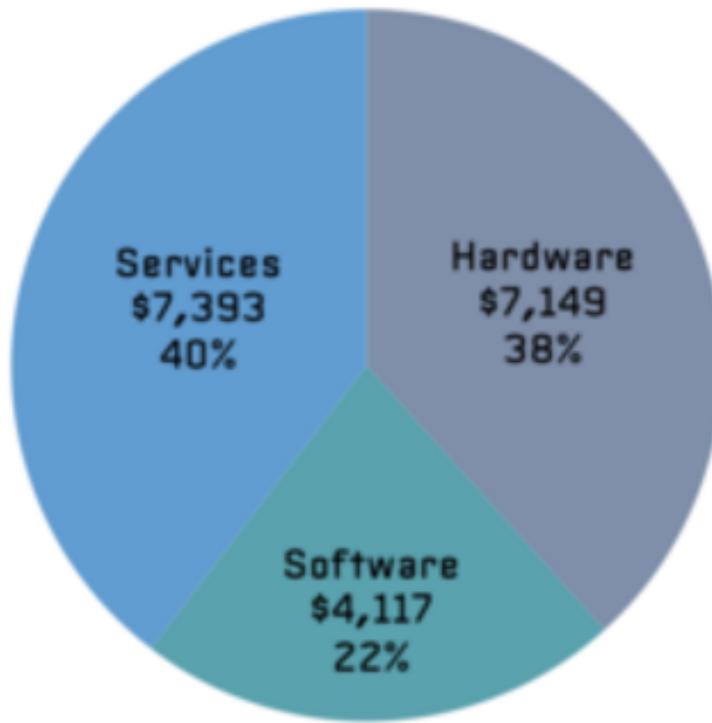
Big Data Market Forecast, 2011-2017 (in \$US billions)



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



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



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

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

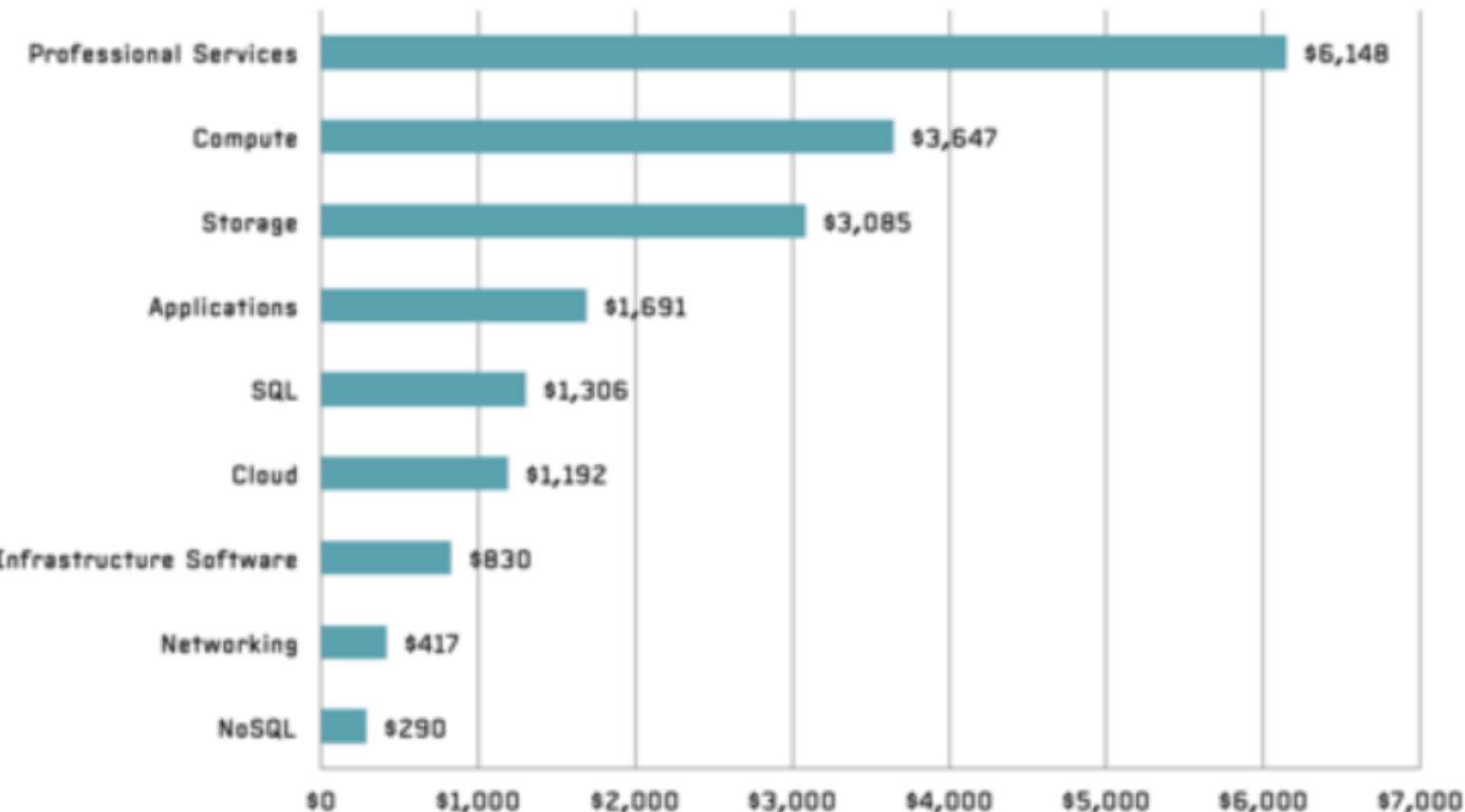
Vendor	Big Data Revenue	Total Revenue	Big Data Revenue as % of Total Revenue	% Big Data Hardware Revenue	% Big Data Software Revenue	% Big Data Services Revenue
IBM	\$1,368	\$99,751	1%	31%	27%	42%
HP	\$869	\$114,100	1%	42%	14%	44%
Dell	\$652	\$54,550	1%	85%	0%	15%
SAP	\$545	\$22,900	2%	0%	76%	24%
Teradata	\$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



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



Big Data Market further breakdown

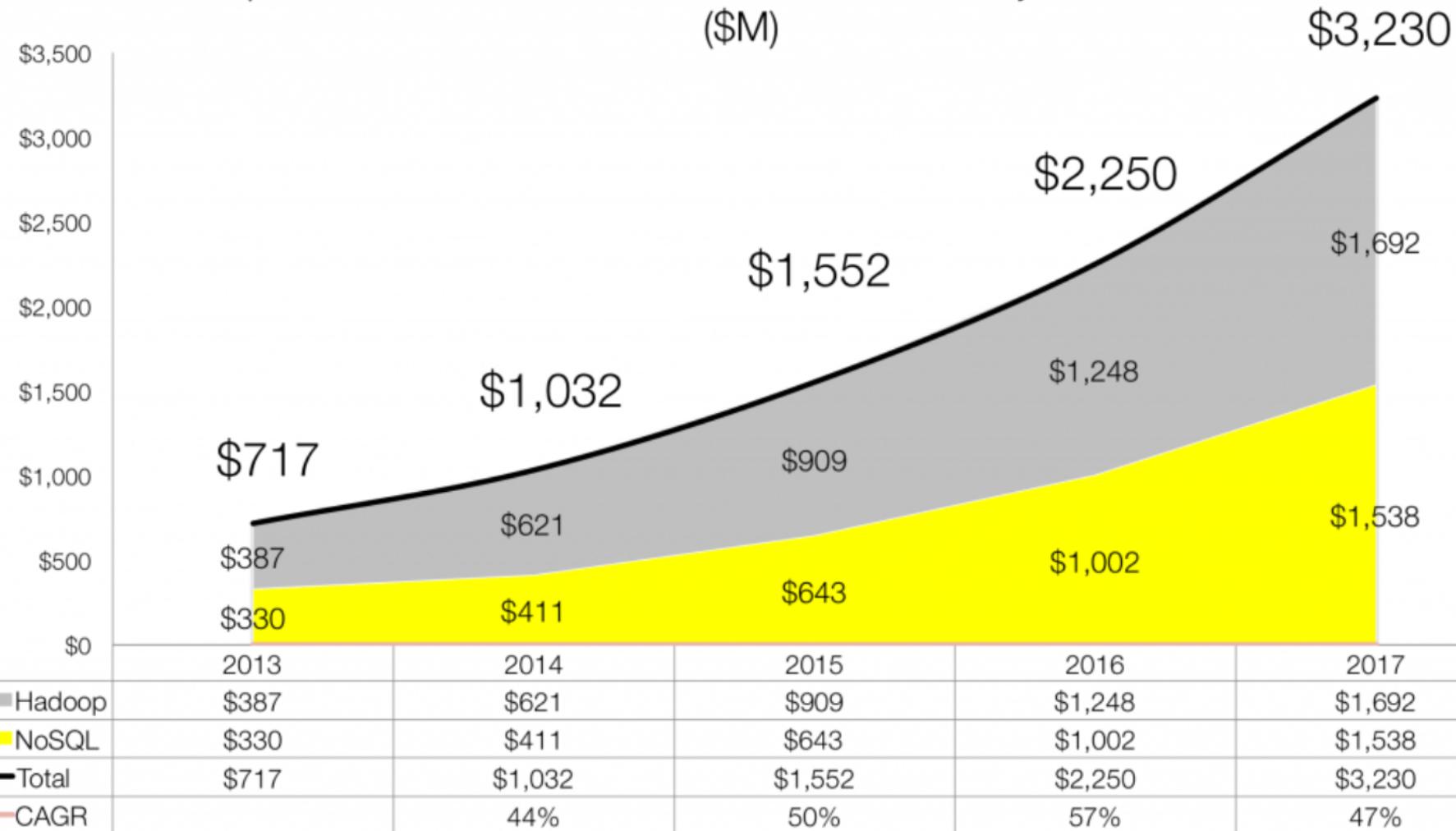
http://wikibon.org/wiki/v/Big_Data_Database_Revenue_and_Market_Forecast_2012-2017

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

- NoSQL DB ==> Distributed DB, Document-Oriented DB, Graph NoSQL DB, and In-Memory NoSQL DB.
- “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]

Hadoop & NoSQL Software/Services Revenue Projection, 2013-2017

($\$M$)



<http://wikibon.com/hadoop-nosql-software-and-services-market-forecast-2013-2017/>

Course Structure

Class Data	Number	Topics Covered
09/10/15	1	Introduction to Big Data Analytics
09/17/15	2	Big Data Platforms
09/24/15	3	Big Data Storage and Processing
10/01/15	4	Big Data Analytics Algorithms — I (recommender)
10/08/15	5	Big Data Analytics Algorithms — II (clustering)
10/15/15	6	Big Data Analytics Algorithms — III (classification)
10/22/15	7	Spark and Data Analytics
10/29/15	8	Linked Big Data — I (graph DB)
11/05/15	9	Linked Big Data — II (graph analytics)
11/12/15	10	Big Data Application (Guest Speaker)
11/19/15	11	Final Project Proposal Presentation
11/26/15		<i>Thanksgiving Holiday</i>
12/03/15	12	Big Data Application (Guest Speaker)
12/10/15	13	Big Data Application (Guest Speaker)
12/17/15 & 12/18/15	14-15	Two-Day Big Data Analytics Workshop – Final Project Presentations

Course Grading

- 3 Homeworks: 50%
 - Individual work; Language Requirement: Java, JavaScript, Python, C/C++, Perl
 - Report and source code
- Data Store & Processing
- Analytics
- In-Memory and Graph Computing

- Final Project: 50%
 - Teamwork: 2 - 3 students per team (on campus); 1+ per team for CVN
- Proposal (slides)
- Final Report (paper, up to 12 pages)
- Workshop Presentation (Oral and Demo)
- Open Source

Course Information

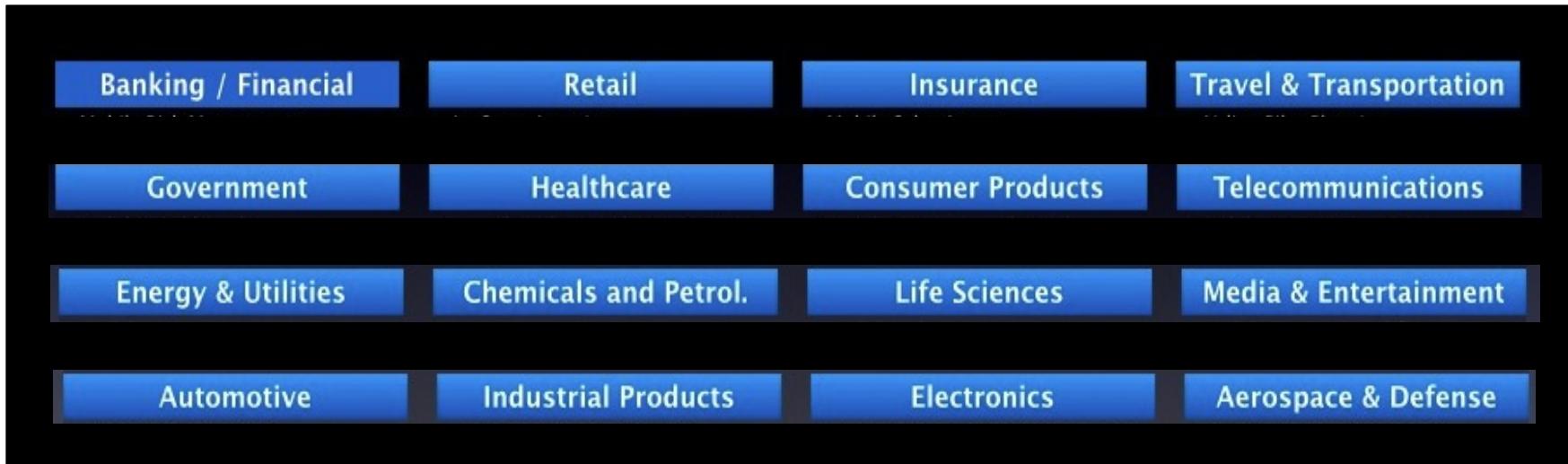
- Website:

<http://www.ee.columbia.edu/~cylin/course/bigdata/>

- Textbook:

-- None, but reference book(s) and/or articles/papers will be provided each lecture.

- **Goal:** Create a Big Data open source toolsets for various industries (and disciplines)



- **Dataset and Use Cases:** Welcome!!

Crowdsourcing of our collective effort!!

Other Issues

- Professor Lin:
 - Office Hours:
Thursday 9:30pm – 10:00pm (SIPA 415, lecture room) (every week)
 - Contact: c {dot} lin {at} columbia {dot} edu (the same as <cl300>)
 - Telephone: 914-945-1897
- TAs:
Ghazel Fazelnia <gf2293>, EE; Office Hours and Location: TBD
We are recruiting more TAs..



The Apache™ Hadoop® project develops open-source software for reliable, scalable, distributed computing.

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

The project includes these modules:

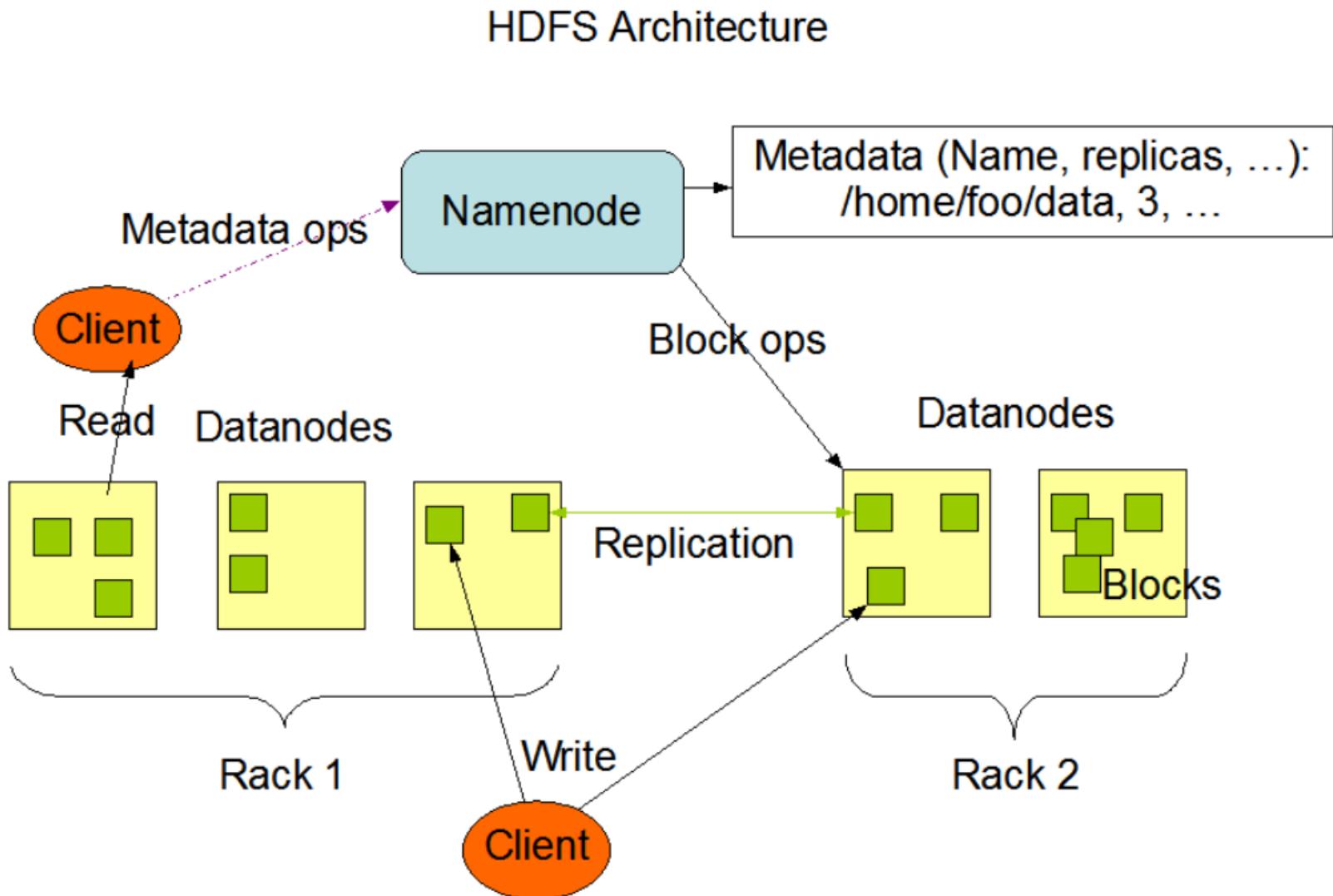
- **Hadoop Common:** The common utilities that support the other Hadoop modules.
- **Hadoop Distributed File System (HDFS™):** A distributed file system that provides high-throughput access to application data.
- **Hadoop YARN:** A framework for job scheduling and cluster resource management.
- **Hadoop MapReduce:** A YARN-based system for parallel processing of large data sets.

<http://hadoop.apache.org>

Hadoop-related Apache Projects

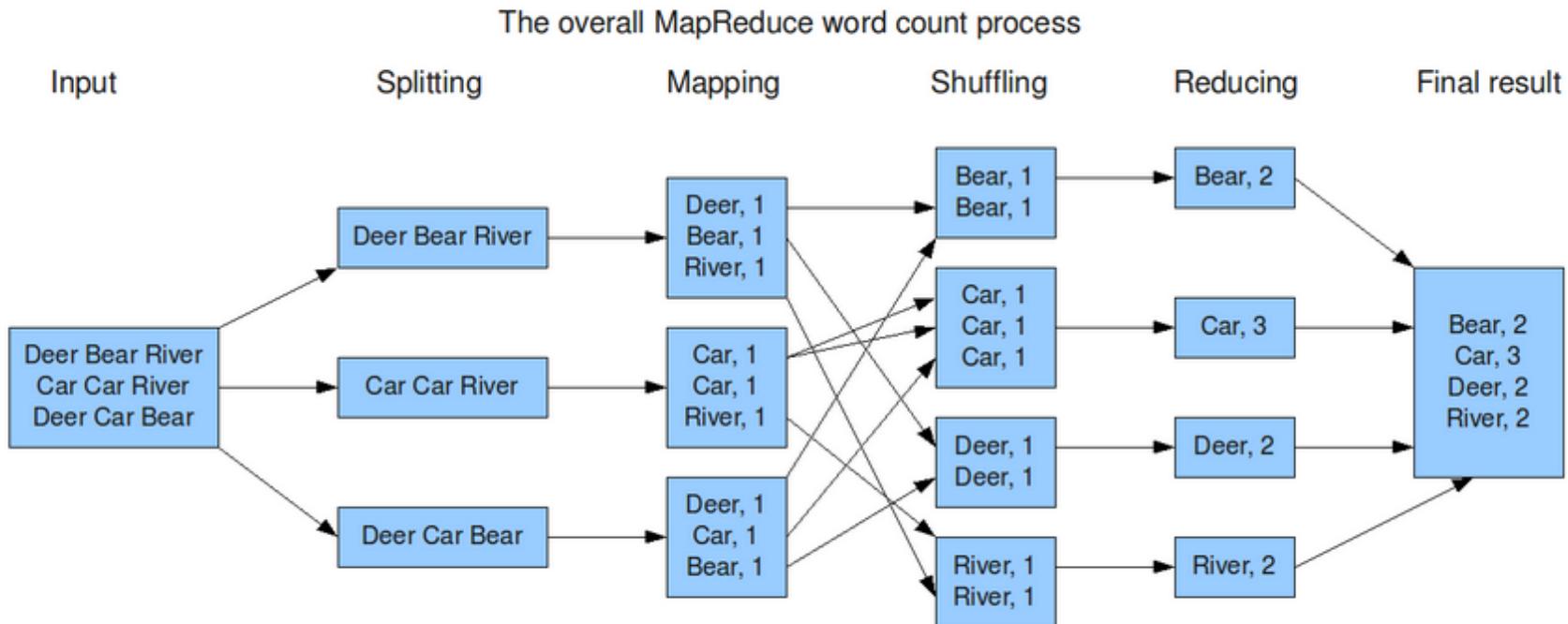
- [Ambari™](#): A web-based tool for provisioning, managing, and monitoring Hadoop clusters. It also provides a dashboard for viewing cluster health and ability to view MapReduce, Pig and Hive applications visually.
- [Avro™](#): A data serialization system.
- [Cassandra™](#): A scalable multi-master database with no single points of failure.
- [Chukwa™](#): A data collection system for managing large distributed systems.
- [HBase™](#): A scalable, distributed database that supports structured data storage for large tables.
- [Hive™](#): A data warehouse infrastructure that provides data summarization and ad hoc querying.
- [Mahout™](#): A Scalable machine learning and data mining library.
- [Pig™](#): A high-level data-flow language and execution framework for parallel computation.
- [Spark™](#): A fast and general compute engine for Hadoop data. Spark provides a simple and expressive programming model that supports a wide range of applications, including ETL, machine learning, stream processing, and graph computation.
- [Tez™](#): A generalized data-flow programming framework, built on Hadoop YARN, which provides a powerful and flexible engine to execute an arbitrary DAG of tasks to process data for both batch and interactive use-cases.
- [ZooKeeper™](#): A high-performance coordination service for distributed applications.

Hadoop Distributed File System (HDFS)



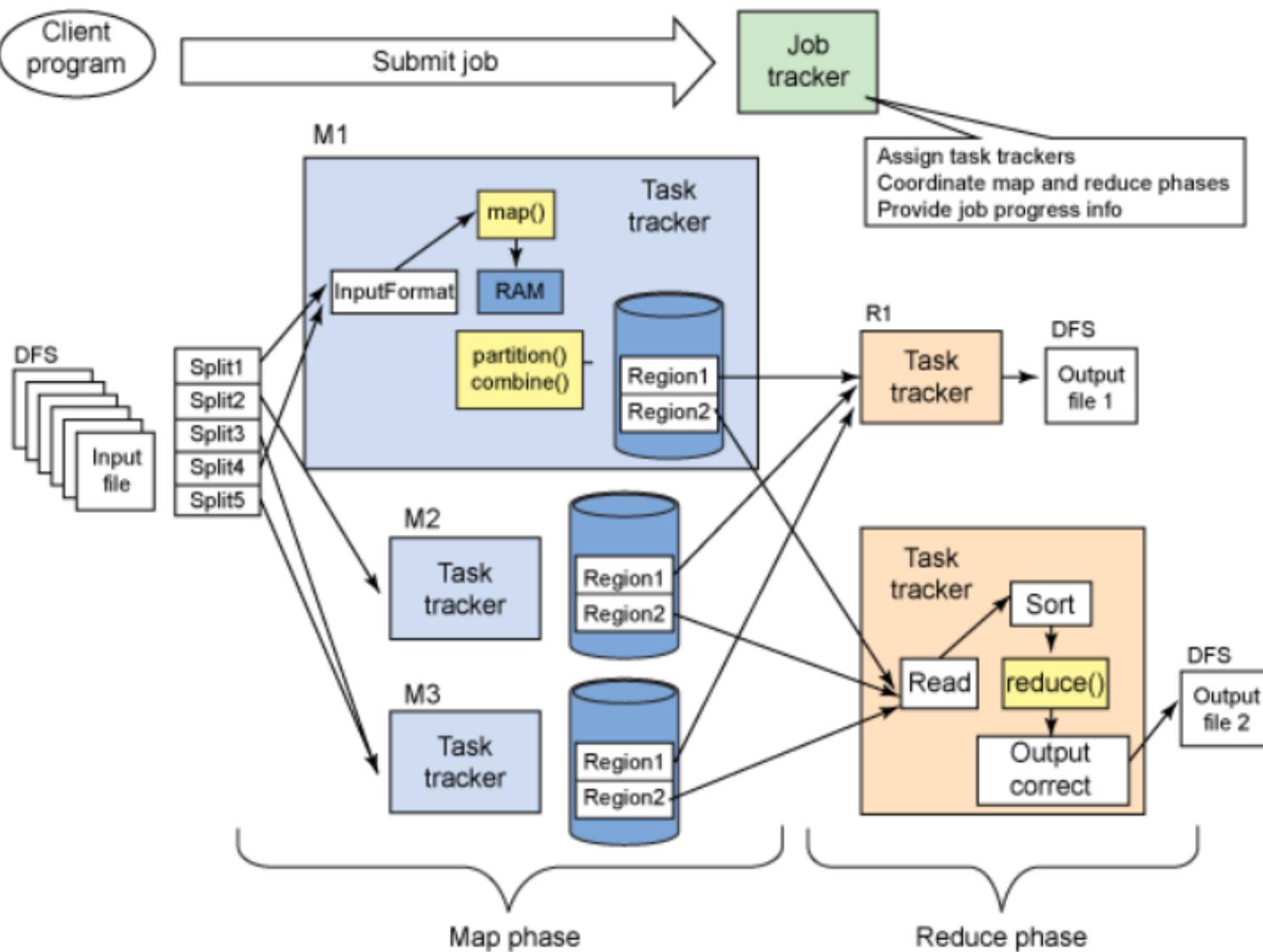
<http://hortonworks.com/hadoop/hdfs/>

MapReduce example



<http://www.alex-hanna.com>

MapReduce Data Flow



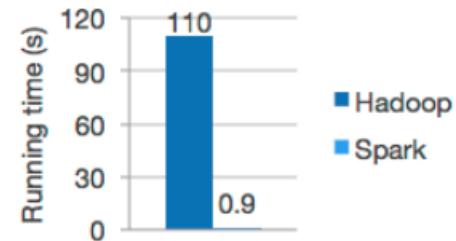
<http://www.ibm.com/developerworks/cloud/library/cl-openstack-deployhadoop/>

Building on top of HDFS

Speed

Run programs up to 100x faster than Hadoop MapReduce in memory, or 10x faster on disk.

Spark has an advanced DAG execution engine that supports cyclic data flow and in-memory computing.



Logistic regression in Hadoop and Spark



Ease of Use

Write applications quickly in Java, Scala or Python.

Spark offers over 80 high-level operators that make it easy to build parallel apps. And you can use it *interactively* from the Scala and Python shells.

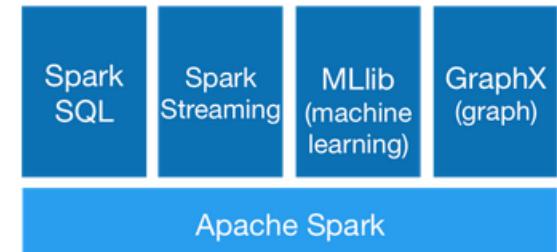
```
file = spark.textFile("hdfs://...")  
  
file.flatMap(lambda line: line.split())  
 .map(lambda word: (word, 1))  
 .reduceByKey(lambda a, b: a+b)
```

Word count in Spark's Python API

Generality

Combine SQL, streaming, and complex analytics.

Spark powers a stack of high-level tools including [Spark SQL](#), [MLlib](#) for machine learning, [GraphX](#), and [Spark Streaming](#). You can combine these frameworks seamlessly in the same application.



IBM System G

- Home
- Overview
- Toolkits
- Solutions
- Cloud
- Documents
- Resource

Graph Analytics

Linked data analysis for intelligence

The Graph 500 List					
November 2013					
No.	Rank	Machine	Installation Site	Number of nodes	Number of cores
1	1	DOE/NNSA/LNL Sequoia (IBM - BlueGene/Q, Power BQC 16C 1.60 GHz)	Lawrence Livermore National Laboratory	65536	1048576
2	2	DOE/SC/Argonne National Laboratory Mira (IBM - BlueGene/Q, Power BQC 16C 1.60 GHz)	Argonne National Laboratory	49152	786432
3	3	JUQUEEN (Forschungszentrum Jülich (FZJ))	Forschungszentrum Jülich (FZJ)	16384	262144
4	4	K computer (Fujitsu - Custom supercomputer)	RIKEN Advanced Institute for Computational Science (AICS)	65536	524288

- Key-Value Store
- Document Store
- Tabular Store
- Object Database
- Graph Database (property graphs, RDF graphs)

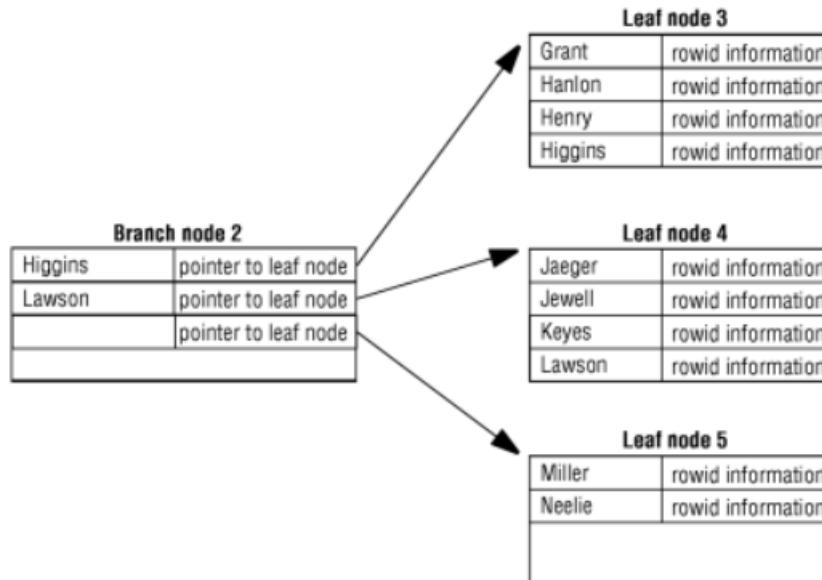
Key Value Store

Example Data Represented in a Key–Value Store

Key	Value
...	
“BMW”	{“1-Series”, “3-Series”, “5-Series”, “5-Series GT”, “7-Series”, “X3”, “X5”, “X6”, “Z4”}
“Buick”	{“Enclave”, “LaCrosse”, “Lucerne”, “Regal”}
“Cadillac”	{“CTS”, “DTS”, “Escalade”, “Escalade ESV”, “Escalade EXT”, “SRX”, “STS”}
...	

- Get(*key*), which returns the value associated with the provided *key*.
- Put(*key, value*), which associates the *value* with the *key*.
- Multi-get(*key₁, key₂,.., key_N*), which returns the list of values associated with the list of *keys*.
- Delete(*key*), which removes the entry for the *key* from the data store.

“Big Data Analytics”, David Loshin, 2013

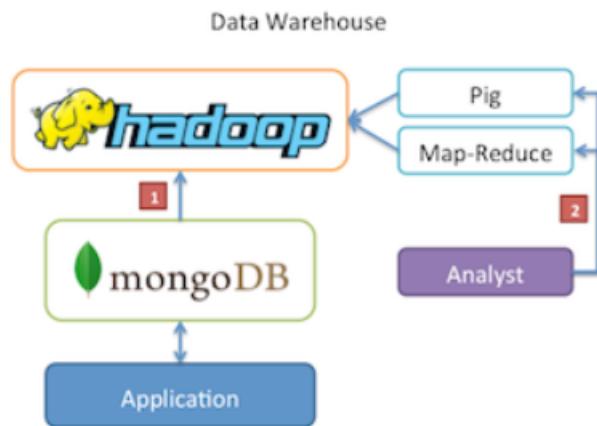


IBM Informix K-V store



Example Application: Spatio-Temporal Analysis

Document Store



The following diagram highlights the components of a MongoDB insert operation:

```
db.users.insert ( ← collection
  {
    name: "sue", ← field: value
    age: 26, ← field: value
    status: "A" ← field: value
  }
)
```

The components of a MongoDB insert operations.

The following diagram shows the same query in SQL:

```
INSERT INTO users ← table
  ( name, age, status ) ← columns
VALUES      ( "sue", 26, "A" ) ← values/row
```

The components of a SQL INSERT statement.



Relational data model

Highly-structured table organization with rigidly-defined data formats and record structure.

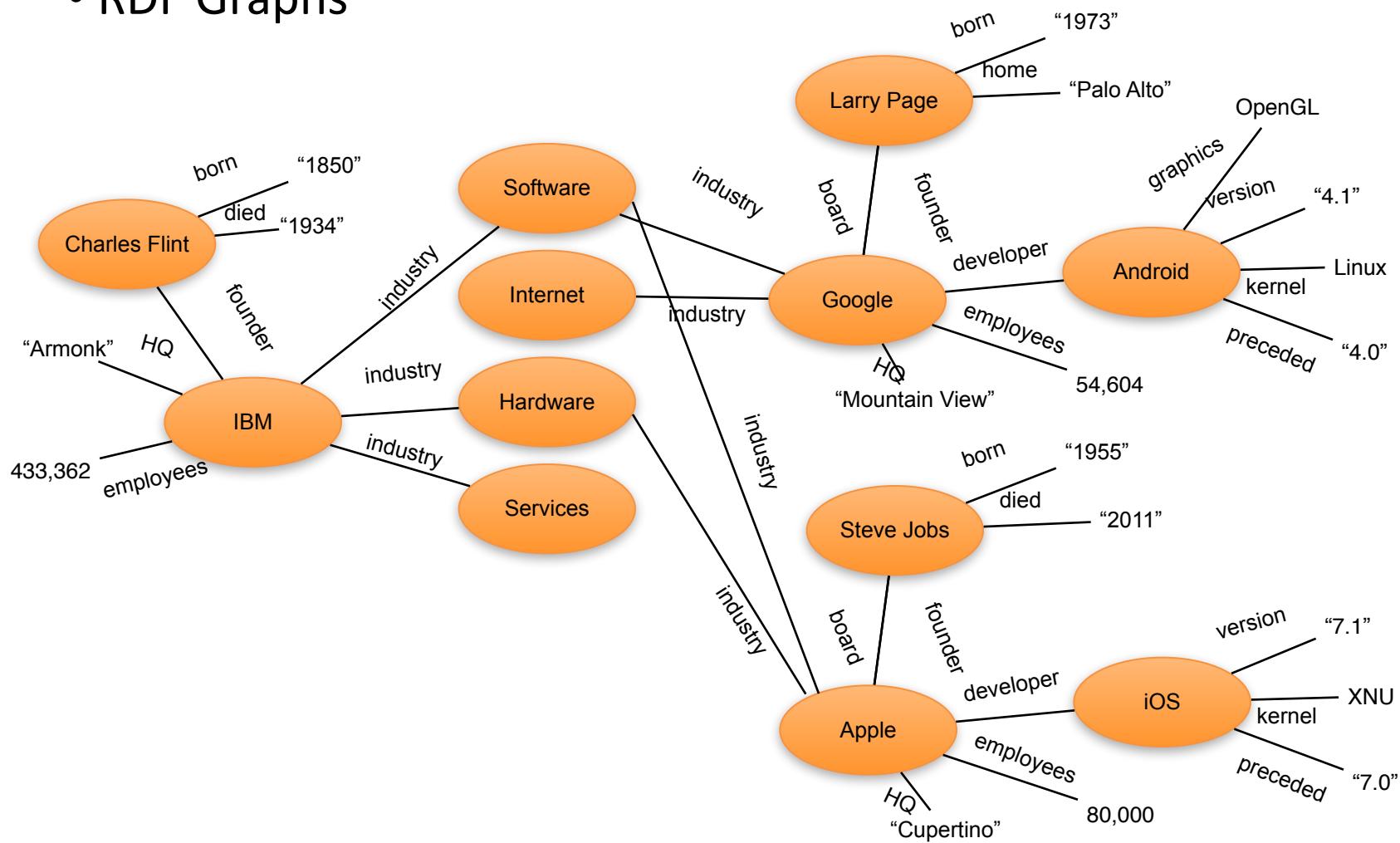


Document data model

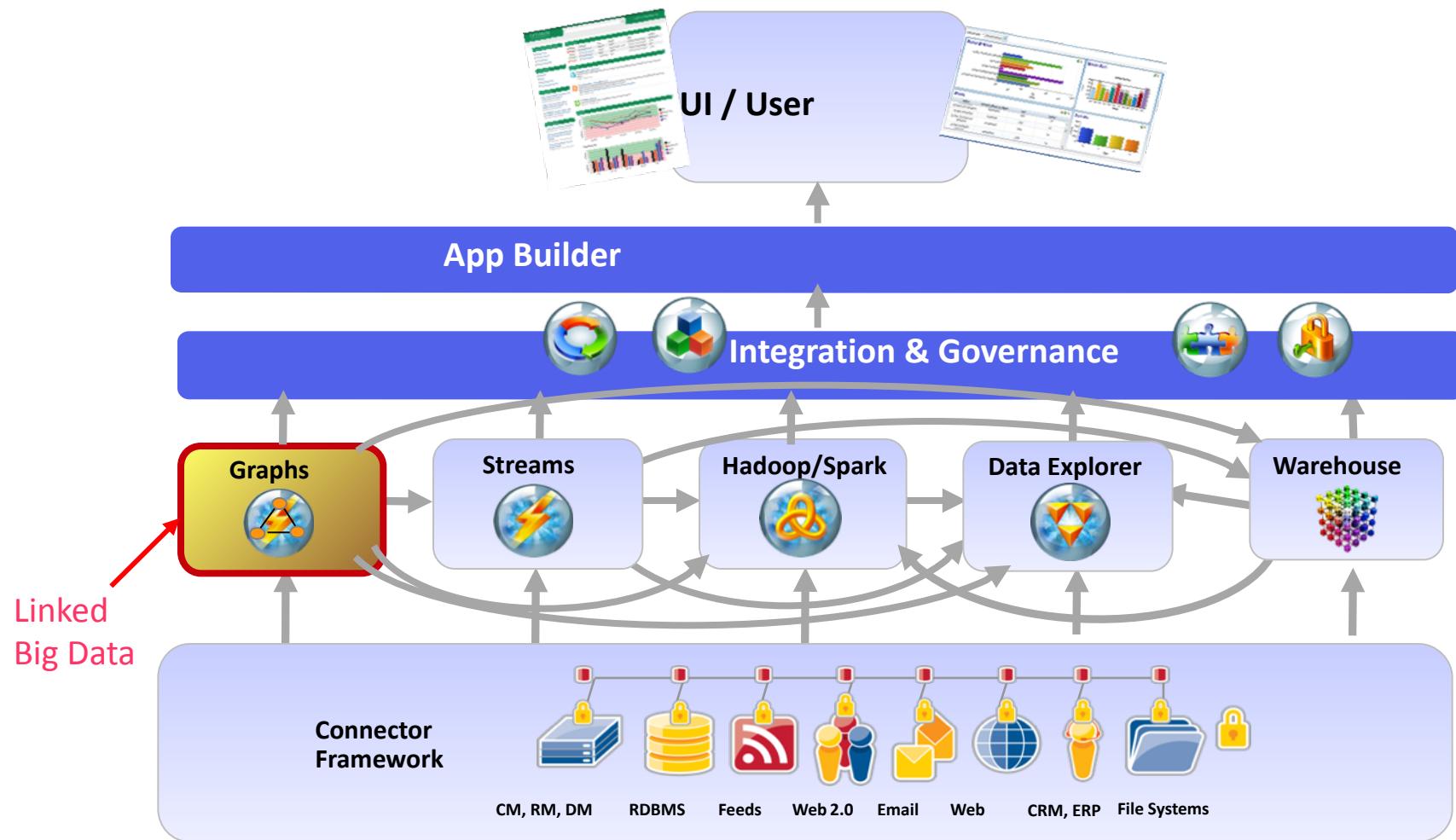
Collection of complex documents with arbitrary, nested data formats and varying "record" format.

Graph Data

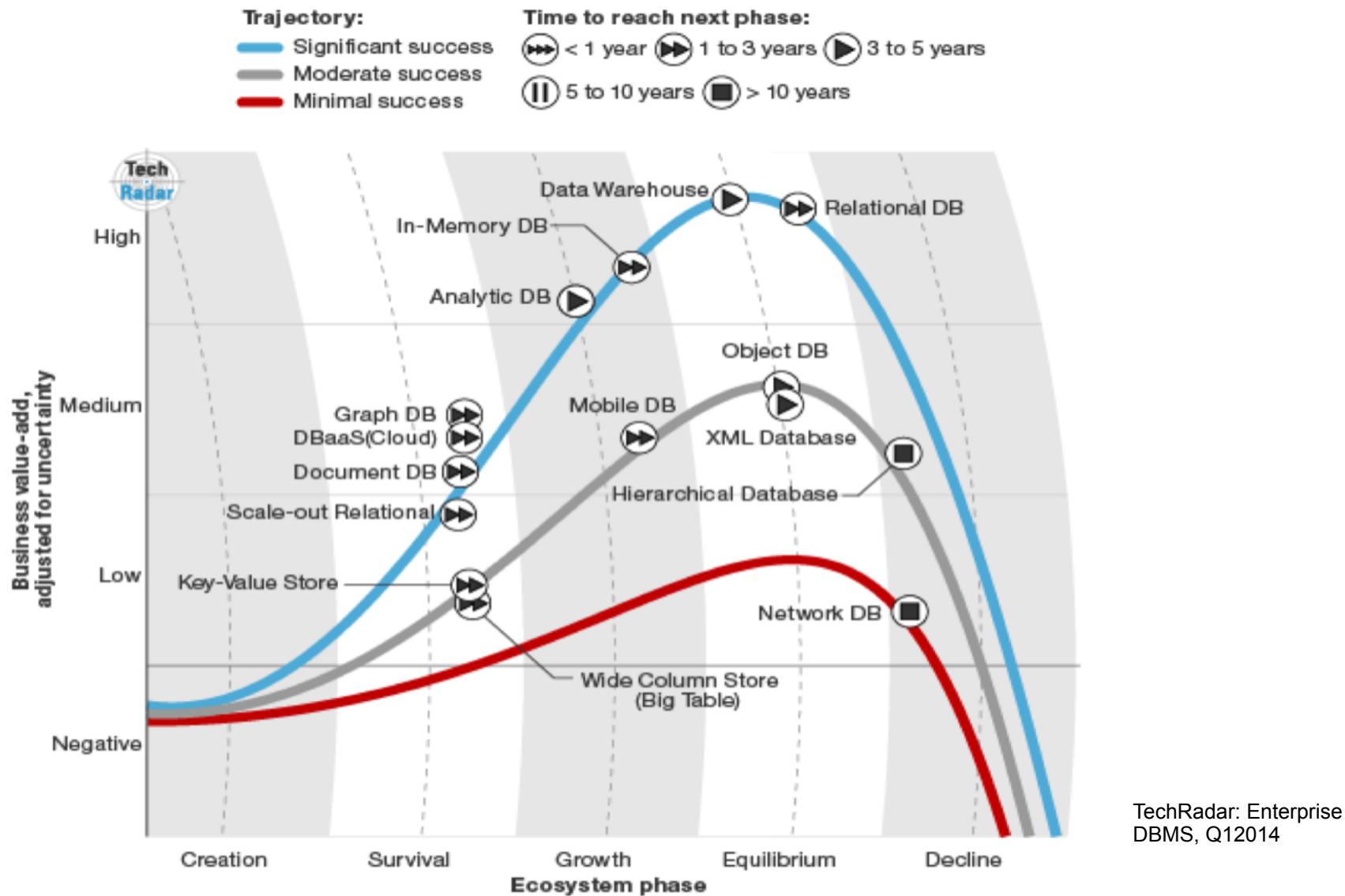
- Property Graphs
- RDF Graphs



Graph is a missing pillar in the existing Big Data foundation

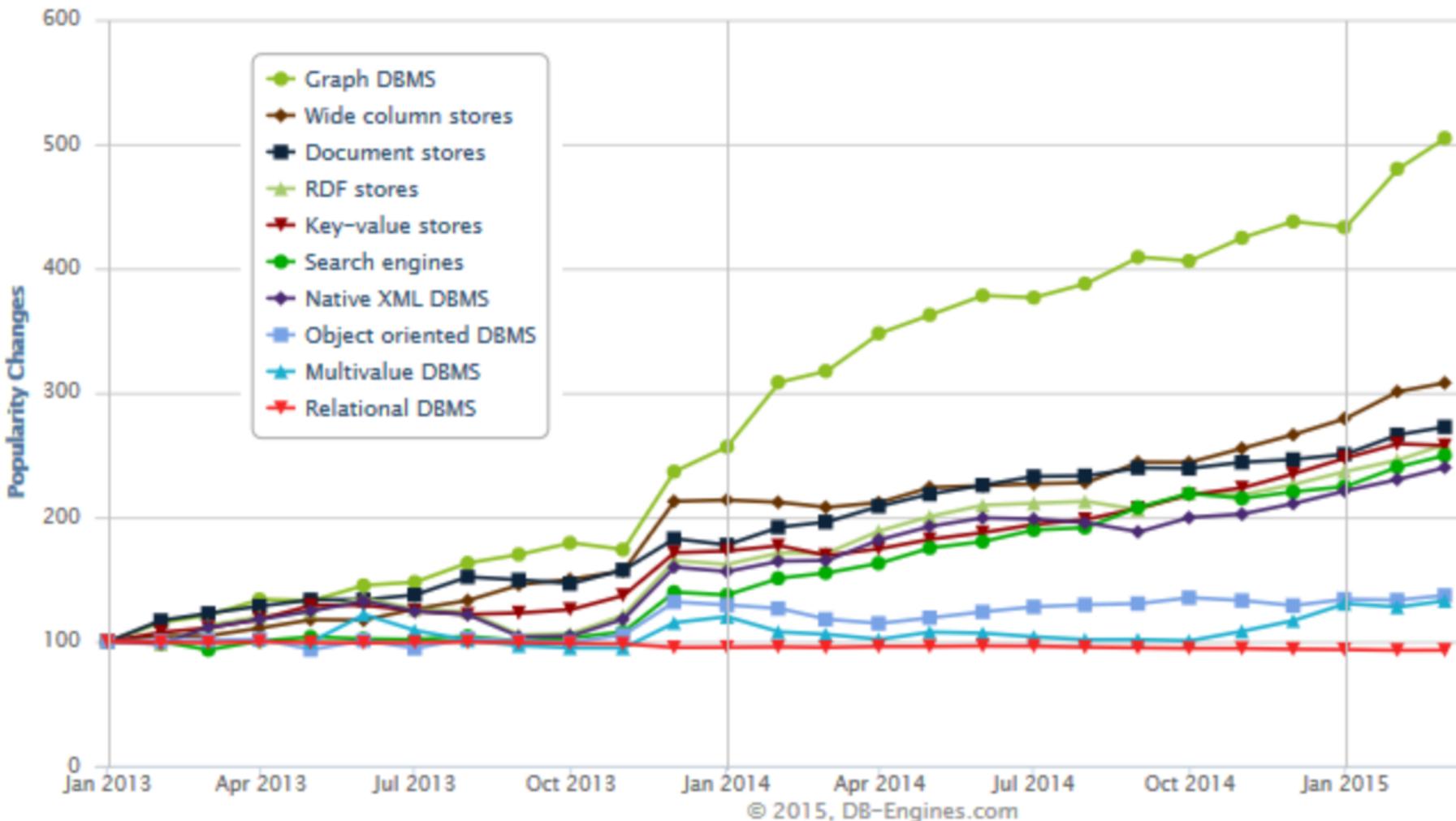


Volume ==> Hadoop / Spark; Velocity ==> Streams; Variety ==> Graphs



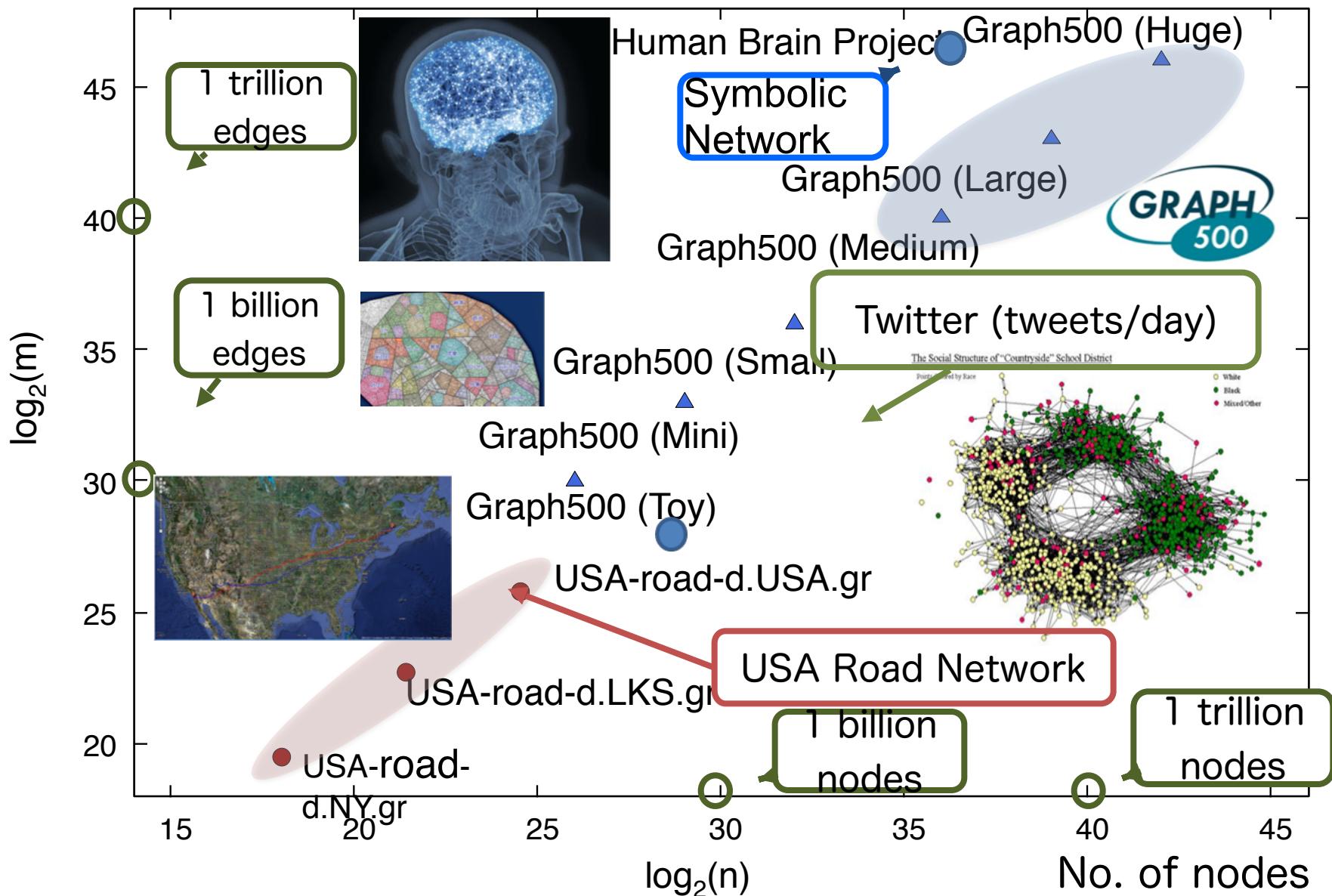
Graph DB is in the significant success trajectory, and has the highest business value among the upcoming DBs.

GraphDB has the largest Popularity Change among DBMS lately



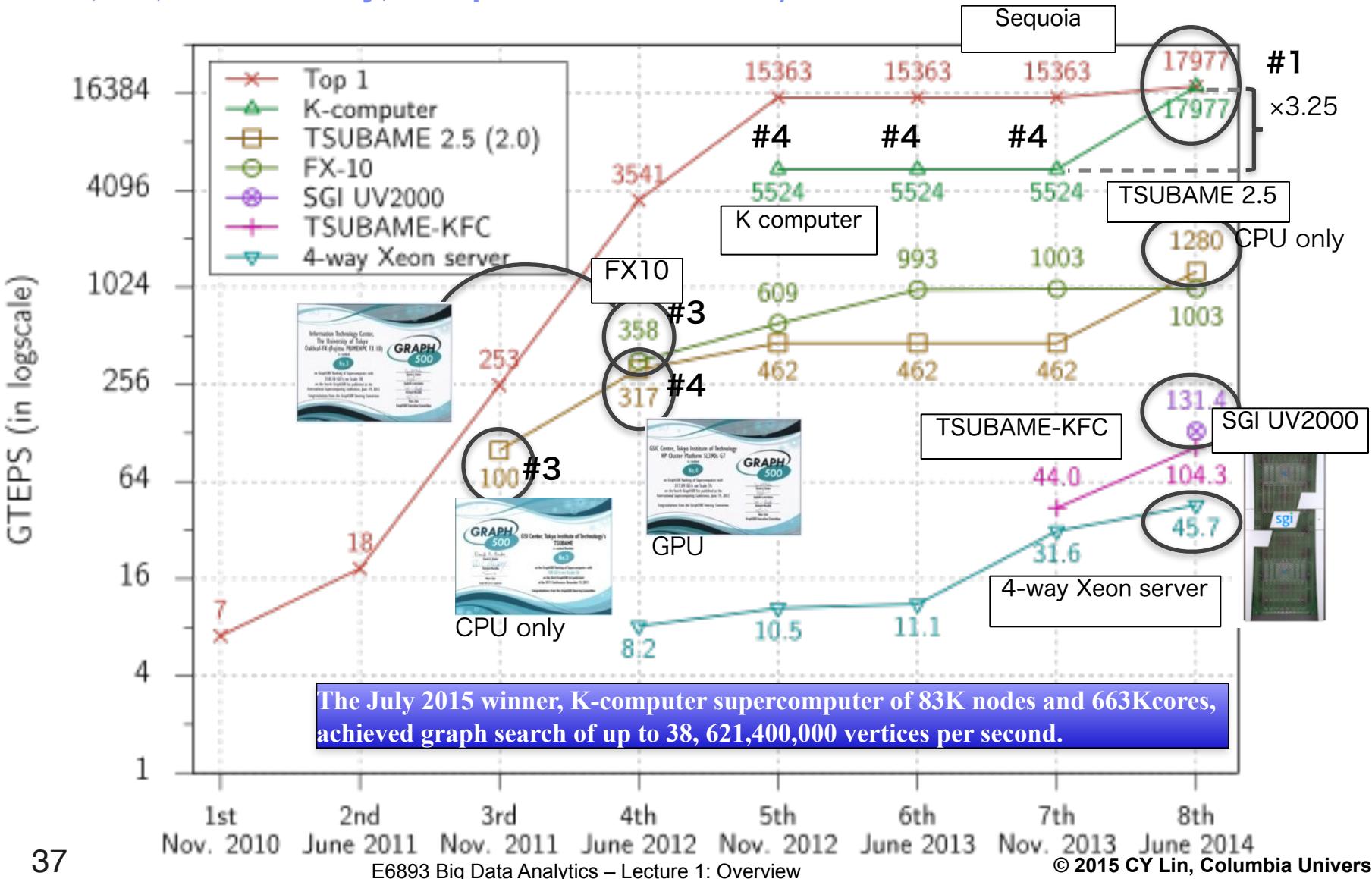
Comparison of linked data size

No. of edges





July 2015: IBM Research's Software powered all Top 3 winners of Graph 500 benchmark and 9 out of the Top 10 winners (supercomputers in US, Japan, France, UK, and Germany; except Tianhe 2 in China).



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



Category 1: 360° View

Recommendation

amazon.com Ching's Store See All 32 Product Categories Your Account | Cart | Your Lists | Help | Find Gifts

Hello, Ching Yung Lin. We have recommendations for you. (If you're not Ching Yung Lin, click here.) Make this

BROWSE

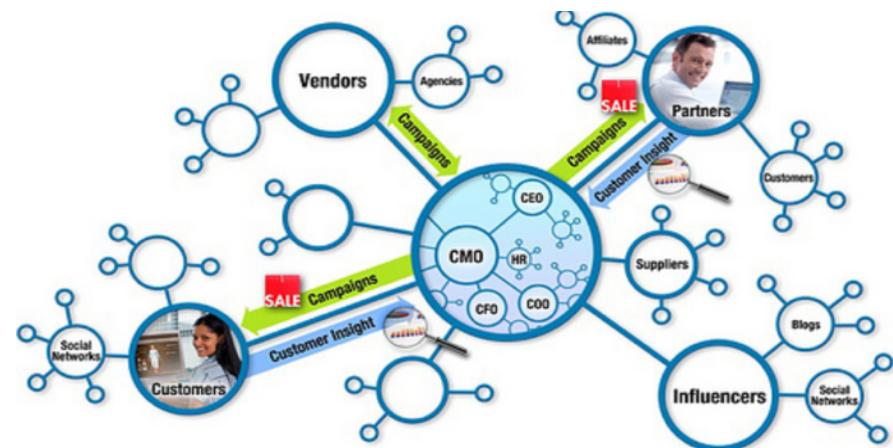
- Your Favorites
 - Books
 - Software
- Featured Stores
 - Apparel & Accessories
 - Beauty
 - DVD's TV Central

Recommended for you

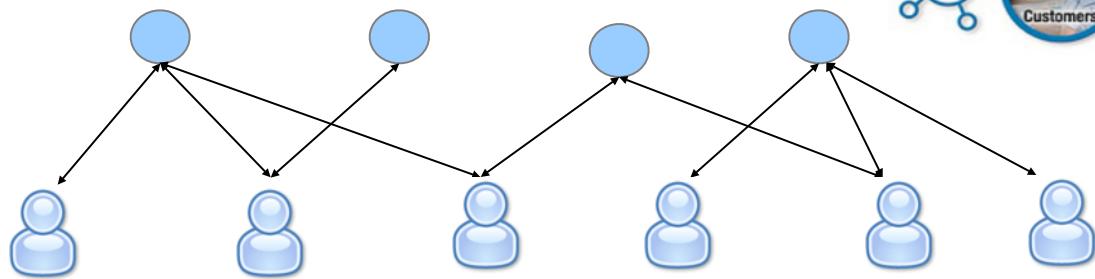
- Spikes [Reprint] Paperback by Fred Rieke
- Spiking Neuron Models Paperback by Wulfram Gerstner
- Methods in Neuronal Modeling - 2nd Edition Hardcover by Christof Koch

(Why is this recommended to me?)

See more Recommendations



Enhancing:



Graph Visualizations

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 and Database

Use Case 1: Social Network Analysis in Enterprise for Productivity

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

15,000 contributors in 76 countries; 92,000 annual unique IBM users

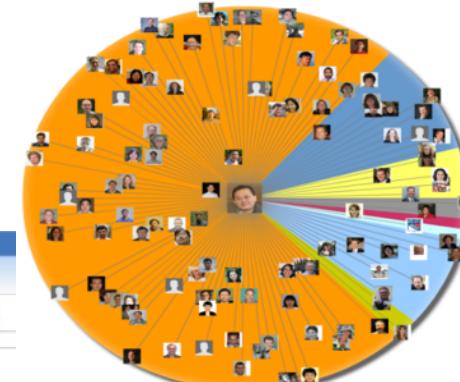
25,000,000+ emails & SameTime messages (incl. Content features)

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

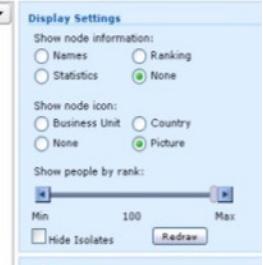
The screenshot shows the SmallBlue Suite interface with a search bar for 'subject keywords' set to 'healthcare'. Below the search bar, it says 'Show people: 1-10 11-20 21-30 31-40 41-50 51-60 61-70 71-80 81-90 91-100'. It also shows 'Show degrees: No limits 1 degree 2 degrees 3 degrees' and '(1: people you know 2: plus people they know 3: plus people "2" know)'. On the left, there are profiles for 1. Patricia (Patti) Okita, 2. Michael Hehenberger, 3. Todd (T.H.) Kalynuk, 4. Susan E. (SUSAN) Rivers, and 5. M.C. (Mark) Effingham. On the right, there is a network visualization titled 'SmallBlue Net' with the instruction 'Click to see results as a Social Network'.



Shortest
Paths

Centralities

Graph
Search



Dynamic networks of
400,000+ IBMers:

Shortest Paths

Social Capital

Bridges

Hubs

Expertise Search

Graph Search

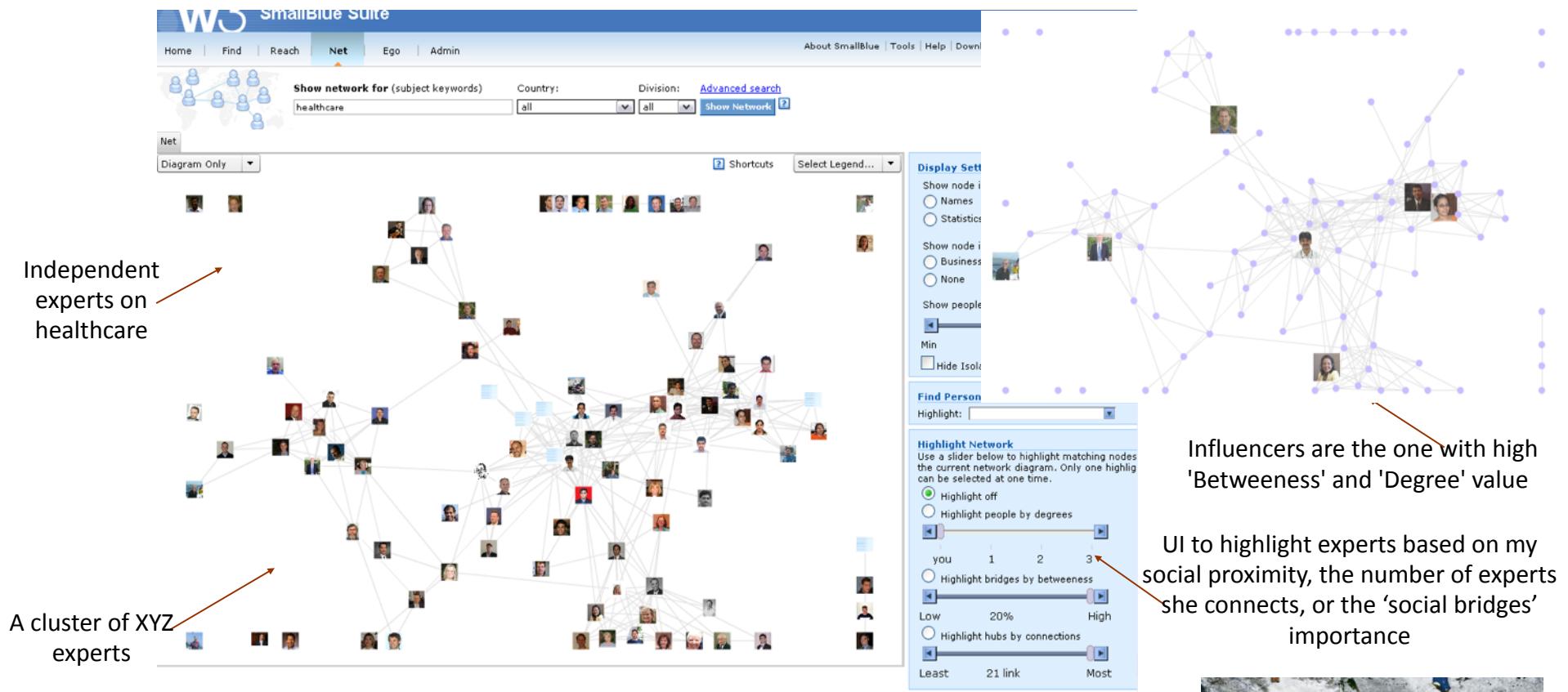
Graph Recomm.

- 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"

Finding and Ranking Expertise – Social Network Analysis

- Decades of Social Science studies demonstrates that (social) network structure is the key indicator determining a person's influence, organizational operation efficiency, social capital to get help, potential to be successful, etc.
- Who are the key bridges? Who have the most connections? How do these experts cluster?
- Analogy – Google founders utilized the concept of network analysis on webpages to create ranking.



SmallBlue analyzes underlining dynamic network structure in enterprise



User Interface of finding knowledgeable and influential colleagues

- Search for the most knowledgeable colleagues within organization or my 3-degree network for who knows topic XYZ (or within a country, a division, a job role, or any group/community)
- Based on IBM HR requirements, adding the 'sponsored search' for business department needs
- IBM HR gives a list of about 10,000 IBMers whose name should not be listed in the search result – mostly high level managers, lawyers, people involving acquisition, etc.
- A list of 2,000+ words that are inappropriate to search in enterprise.

W3 SmallBlue Suite

Home | **Find** | Reach | Net | Ego | Admin About SmallBlue | Tools | Help | Download | Terms of Use | Project Info

Search for (subject keywords): healthcare Country: all Division: Advanced search Find Expert

Show people: 1-10 11-20 21-30 31-40 41-50 51-60 61-70 71-80 81-90 91-100
 Show degrees: No limits 1 degree 2 degrees 3 degrees (1: people you know 2: plus people they know 3: plus people "2" know)

SmallBlue Net Click to see results as a Social Network

As on **9/29/2009**, SmallBlue is indexing/inferred the social network and expertise of **409542** IBMers.
 The system has **10103** contributing IBM users from **68** countries.
 Please invite your colleagues to join SmallBlue. The more people who join, the better SmallBlue will be.

Settings Remove me from this search Manage personal stop terms Submit non-searchable term

Terms of use

My shortest path to Susan

As a user, you can only see their public information. Private info is used internally to rank expertise but private data can never be exposed.

Click a name to see their profile (SmallBlue Reach)

	1. Patricia (Pattie) Okita Global Business Services Associate Partner, Healthcare Integration Other Consultant Ask: MARTHA E. (Martha) GIBSON > Amy D. (AMY) Berk		2. Michael Hohenberger IBM Research Life Sciences Business Development Category Sales Ask: Ravi B. Konuru > Vanessa L. Johnson
	3. Todd (T.H.) Kalyniuk Global Business Services GBS Partner, Healthcare and Public Health -- Practice Administrator is Shirley Carkner Other Consultant Ask: Chung Sheng Li > Robert (R.) Torok		4. Susan E. (SUSAN) Rivers Global Business Services Healthcare Knowledge Manager Market Insights Ask: MARTHA E. (Martha) GIBSON
	5. M. C. (Mark) Effingham IBM Sales & Distribution, Public Sector Client Technical Advisor Ask: Ari Fishkind > Julie A. Reid		6. Paul (P.E.) Van Aqqelen Global Business Services Pacific Development Center, Business Development Manager Other Consultant Ask: Michael W. Ticknor > Kinson (K.W.) Lee
	7. Eric S. (ERIC) Minkoff Global Business Services US GBS Learning & Knowledge Learning Deployment Lead - Public Sector Ask: James (JAMES) Stupak > Andrea R.		8. Thomas (Tom) Cocozza Global Business Services Healthcare Transformation Services Ask: MARTHA E. (Martha) GIBSON > Alan J. (ALAN) Lauder

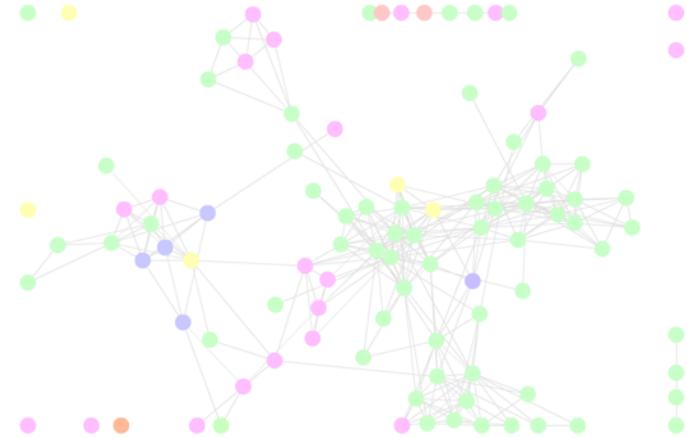
Visualize social roles of individuals in company



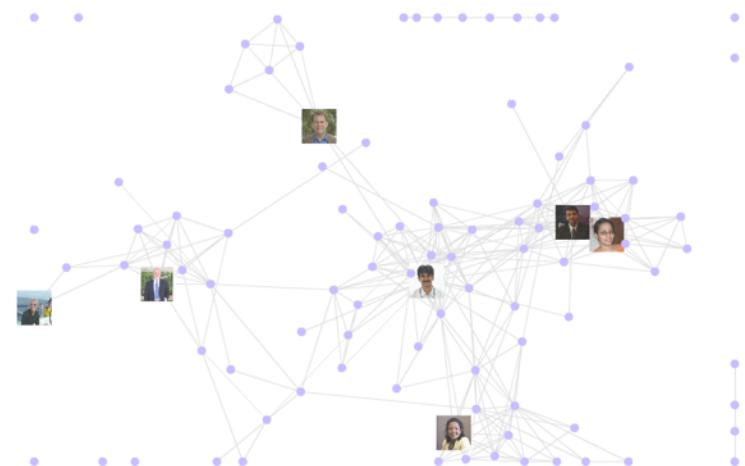
Example: Healthcare experts in the world



Example: Healthcare experts in the U.S.



Connections between different divisions



Key social bridges

Shortest Paths between two people in enterprise

- Example: Is Tom a right person to me?

His official job role, title, contact info

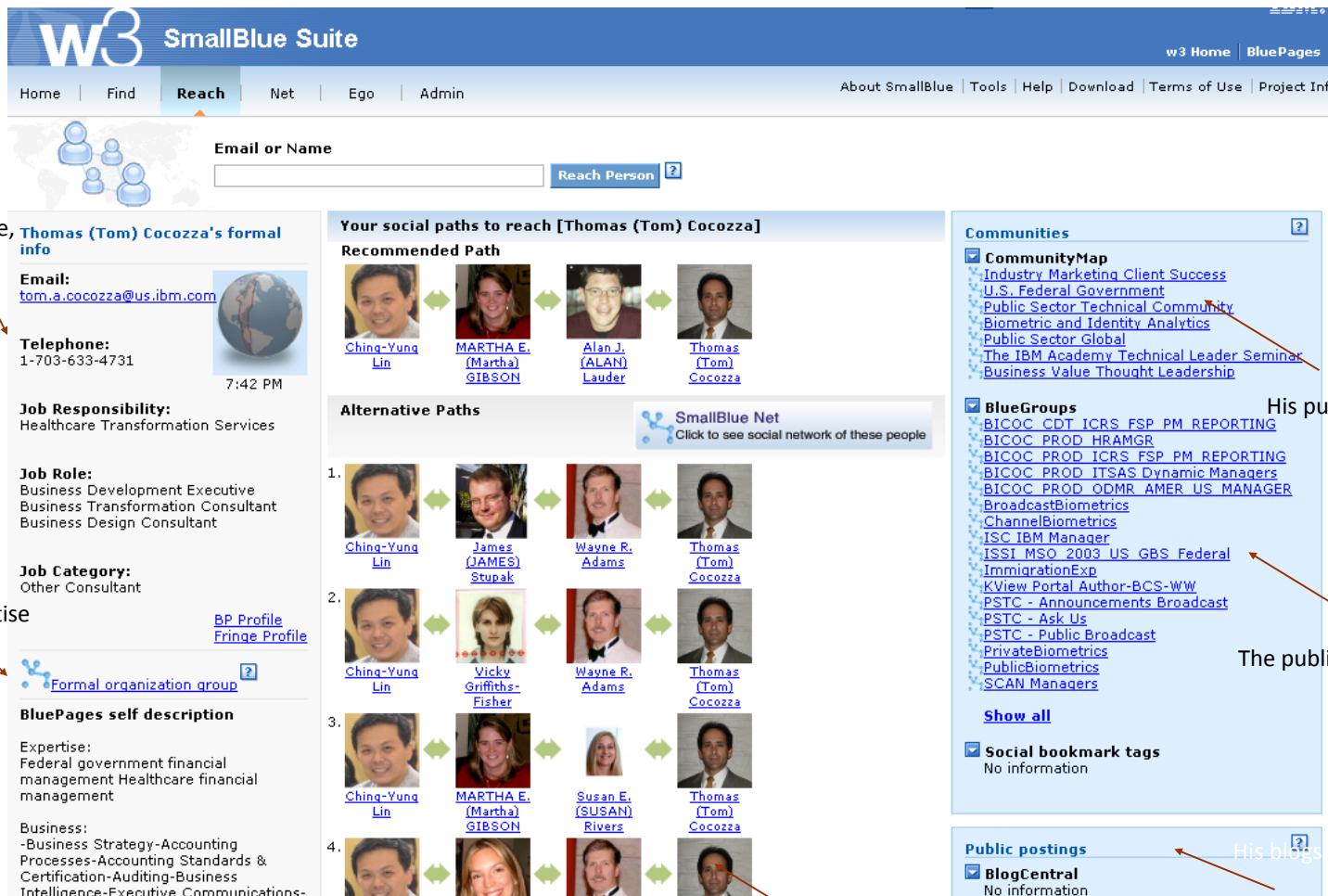
His self-described expertise

His public communities

The public interest groups he is in

His blogs

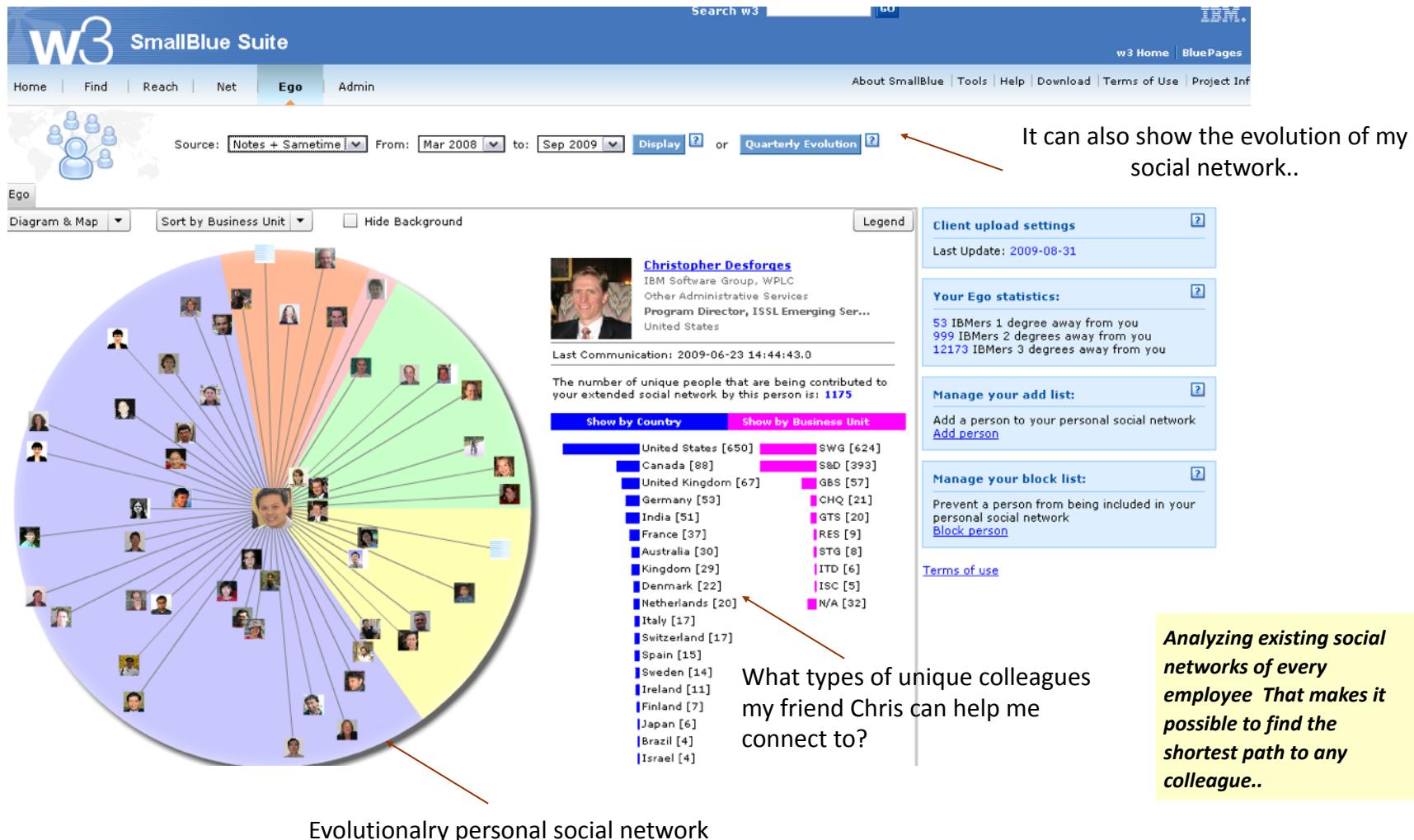
My various paths to Tom. SmallBlue can show the paths to any colleagues up to 6-degree away



The screenshot shows the SmallBlue Suite interface with the 'Reach' tab selected. The main search bar contains 'Thomas (Tom) Cocozza'. Below it, under 'Your social paths to reach [Thomas (Tom) Cocozza]', there are four sections: 'Recommended Path' (4 steps), 'Alternative Paths' (4 steps), 'Public postings' (1 step), and 'Social bookmark tags' (No information). Each path section shows a sequence of profile pictures and names. To the left, a sidebar displays Thomas's formal info (Email: tom.a.cocozza@us.ibm.com, Telephone: 1-703-633-4731, Job Responsibility: Healthcare Transformation Services, Job Role: Business Development Executive, Business Transformation Consultant, Business Design Consultant, Job Category: Other Consultant) and his self-described expertise (Formal organization group, BluePages self description: Expertise: Federal government financial management, Healthcare financial management; Business: Business Strategy-Accounting, Processes-Accounting Standards & Certification-Auditing-Business Intelligence-Executive Communications-). To the right, a sidebar lists 'Communities' (CommunityMap, Industry Marketing Client Success, U.S. Federal Government, Public Sector Technical Community, Biometric and Identity Analytics, Public Sector Global, The IBM Academy Technical Leader Seminar, Business Value Thought Leadership), 'BlueGroups' (BICOC_CDT_ICRS_FSP_PM_REPORTING, BICOC_PROD_HRAMGR, BICOC_PROD_ICRS_FSP_PM_REPORTING, BICOC_PROD_ITSA_Dynamic_Managers, BICOC_PROD_ODMR_AMER_US_MANAGER, BroadcastBiometrics, ChannelBiometrics, ISC_IBM_Manager, ISSI_MSO_2003_US_GBS_Federal, ImmigrationExp, KView_Portal_Author-BCS-WW, PSTC - Announcements Broadcast, PSTC - Ask Us, PSTC - Public Broadcast, Private_Biometrics, Public_Biometrics, SCAN_Managers), 'Social bookmark tags' (No information), and 'Public postings' (BlogCentral, No information).

Personal social network capital management

- What is a friend's social capital to me? Am I losing an 'important' friend?



Network Value Analysis – First Large-Scale Economical Social Network Study



BusinessWeek

TOP NEWS | BW MAGAZINE | INVESTING | ASIA | EUROPE | TECHNOLOGY | AUTOS | INNOVATION | SMALL BIZ | B-SCHOOLS | CAREERS

SEARCH SITE GO Advanced Search

APRIL 10, 2009

Insider Newsletter

A weekly summary of the best In BusinessWeek and BusinessWeek.com

NEWS THIS WEEK'S TOP STORY

Putting a Price on Social Connections

Researchers at IBM and Massachusetts Institute of Technology say that indulging in certain types of electronic communications makes for higher productivity at work. Our Insider Top Story this week, "Putting a Price on Social Connections," is part of our Special Report, The Value of Virtual Friends. Find out what sort of networking increases your value as an employee.

Researchers at IBM and MIT have found that certain e-mail patterns at work correlate with higher revenue production.

You've seen the advertisements on TV and in the paper—apparently there has never been a better time to sell Grandma's lara. But hold on. Click through our slide show, "Selling Your Jewelry," and read the accompanying story first to see what you can get for what you've got. Have you ever had the thing appraised?

Beer sales are down in the mature markets, but one brewer is hoping an untapped market will make up for it. BWV correspondent Kerry Capell follows giant SABMiller into Africa. The idea is to make commercial beer that's more affordable, to compete with local brews.

And we have stories on buying shares in distressed sectors and forming nonprofits in this troubled economy.

—Katherine Davis

Productivity effect from network variables

- An additional person in network size ~ \$986 revenue per year
- Each person that can be reached in 3 steps ~ \$0.163 in revenue per month
- A link to manager ~ \$1074 in revenue per month
- 1 standard deviation of network diversity (1 - constraint) ~ \$758
- 1 standard deviation of btw ~ -\$300K
- 1 strong link ~ \$-7.9 per month

- Structural Diverse networks with abundance of structural holes are associated with higher performance.
 - *Having diverse friends helps.*
- Betweenness is negatively correlated to people but highly positive correlated to projects.
 - *Being a bridge between a lot of people is bottleneck.*
 - *Being a bridge of a lot of projects is good.*
- Network reach are highly corrected.
 - *The number of people reachable in 3 steps is positively correlated with higher performance.*
- Having too many strong links — the same set of people one communicates frequently is negatively correlated with performance.
 - *Perhaps frequent communication to the same person may imply redundant information exchange.*

Use Case 2: Recommendation

w3 Search Pages(w3)

Practitioner Portal Translate this page: English Tell a friend How-to videos Portal help Site map Feedback

People in your network

Network for: [Lin, China-Yung](#)

81 colleagues are 1 degree from you
1615 colleagues are 2 degrees from you
18270 colleagues are 3 degrees from you

Your 1st degree network diagram ([Show list](#))

View networks: [Lotus Connections & SmallBlue](#) ▾

Sort by: Division | Country | Social proximity



[Edit SmallBlue](#)

[View all tags](#) | [Tags by person](#)

▶ Portlet social rating information

Buzz in your network

Share your status with your network:

[Post status](#)

Network buzz for networks:

IBM Connections & SmallBlue ▾

Sources:

Profiles Blogs ➔

1 of 1 items Network: All Sources: All Sort by: Most recent | Person

 [Jeffrey Nichols](#) Re: Thoughts (and Questions) on Answers [July 09 10:50 AM](#) [Comment](#)

[RSS Feed](#)

▶ Portlet social rating information

Popular in the Practitioner Portal

Here's what is currently popular in the Practitioner Portal with your colleagues.

▶ Top 5 document searches

SAP, cloud pattern, bao, signature, solutions, bob, sc, KM and KS case studies

▶ Top accessed content

▶ Top Bookmarks

▶ Portlet social rating information

Recently shared content in your network

See what content people in your network have been sharing to others. Select the network and sources you are interested in and click go.

Networks:

Direct (1st degree) ▾

Sources:

IC Bookmarks IC Files IC Wikis
 Practitioner Portal Media Library ILX [GO](#)

5 of top 18 Sort by: Social Proximity | Date | Source

Network: direct Sources: All

 [Welcome to Graph Technologies](#) [09 Jul 2013](#)

 [Mobile security Workshop \(Bharti Airtel\)](#) [15 Jul 2013](#)

 [hci-and-smartphone-data-at-scale-ibm-Jul2013.pptx](#) [30 Jul 2013](#)

Popular learning

See what education is popular with the people in your network. Select the sources you are interested in and click go.

Sources:

L@IBM Media Library ILX [GO](#)

5 of top 30 Sort by: Popularity | Source

Sources: All

[Leadership in a Project Team Environment](#) 

PMKN eShareNet June 13, 2013 - Worldwide Project Management Method (WWPMM) 3.0 Release Preview: Improving PM Method Adaptability. Presented by Stacey Lopez and Todd Fredrickson - IBM Rational Asset Manager 

New2Blue - Mid-Year Review - Personal Business Commitments (Session Replay) [New Employee Experience 2013 Events] 

Junos Pulse for Android Smartphone 

Project Management Orientation 

[Show more](#)

- Integrated Practitioner Portal, KnowledgeView, Media Library, Lotus Connections, and Learning@IBM and for a personalized ranking



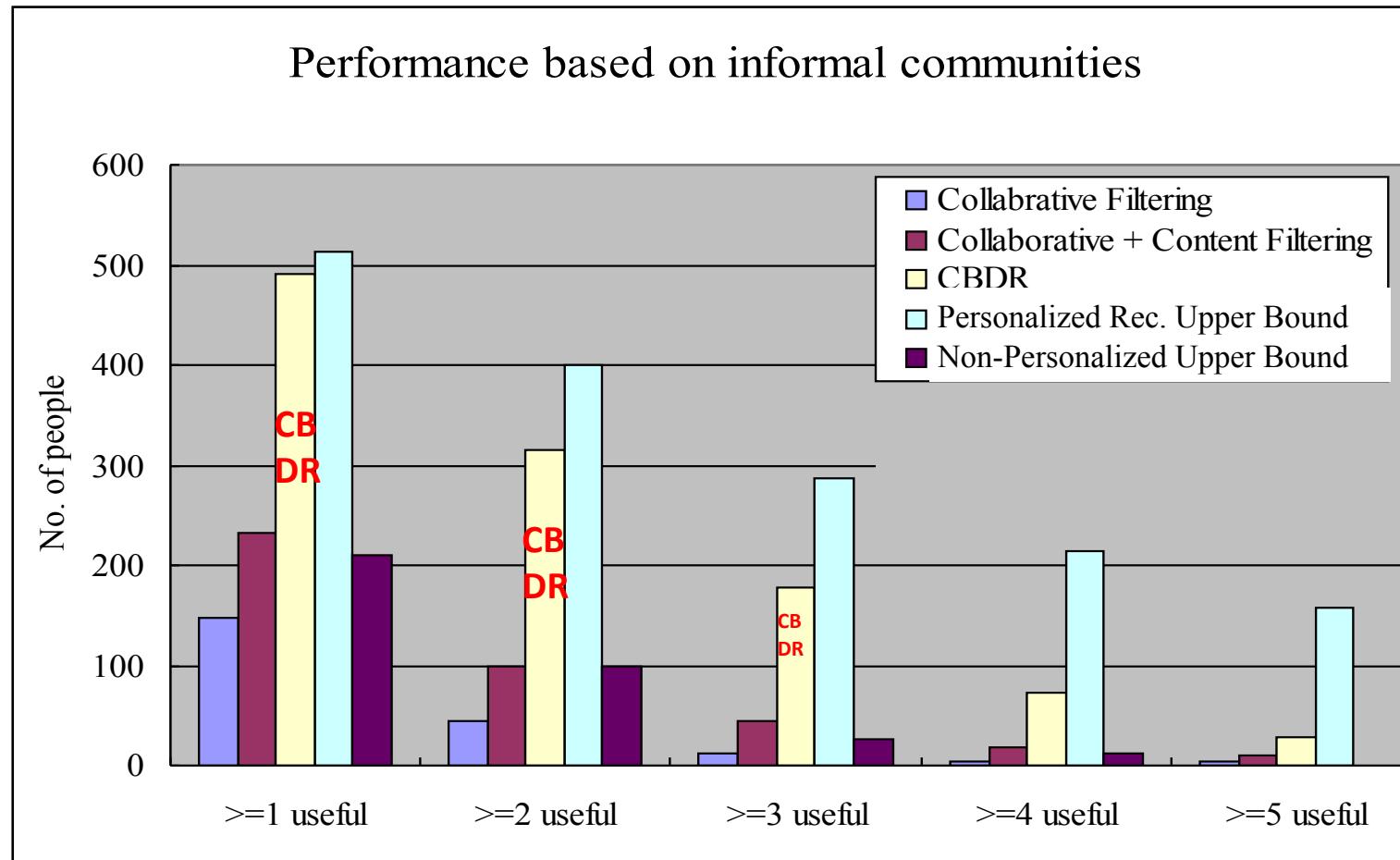
Graph
Communities

Improving Recommendation Quality by Graph Community Analytics

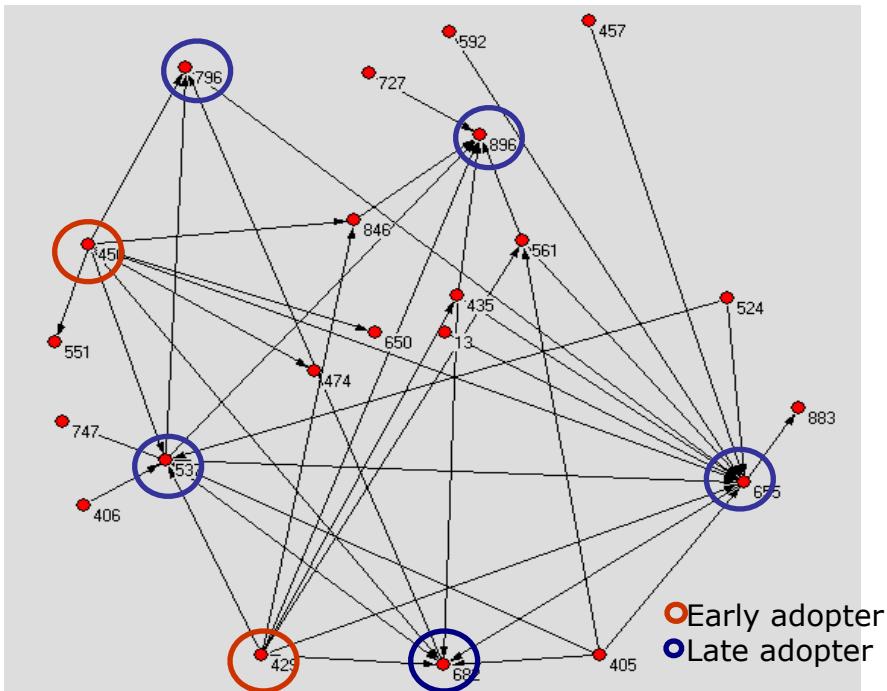
- A 3rd party Knowledge Repository: 30K users and 20K documents.

Study the most active 697 users who have at least 20 download in a year.

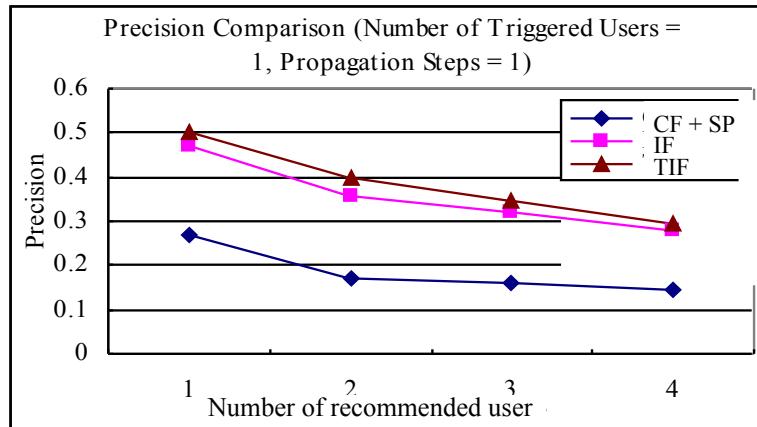
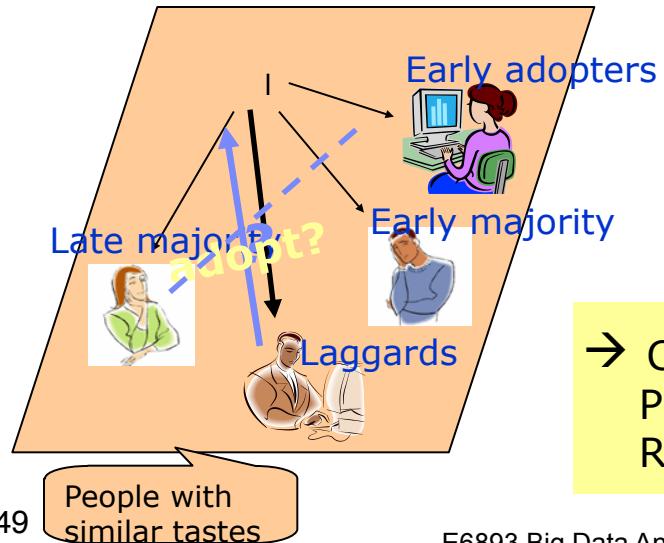
- **Results: beyond Collaborative Filtering:** (1) Collaborative + Content Filtering (53% improvement); (2) CBDR: Collaborative + Content Filtering + Graph Community Analytics (259% accuracy improvement over collaborative filtering)



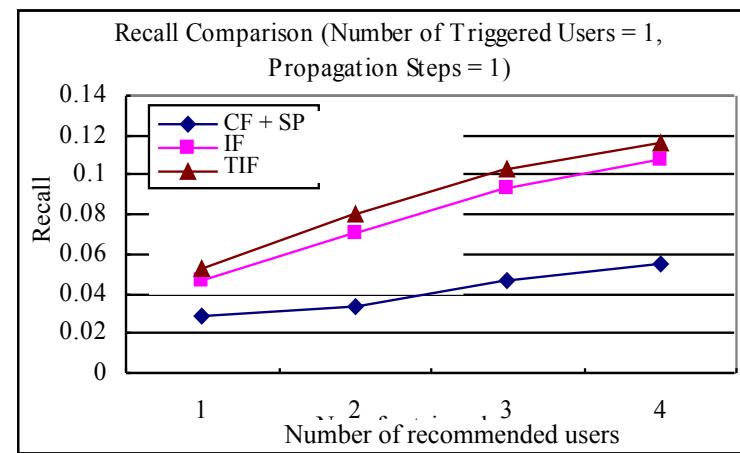
Use Case 3: Recommendation for Commerce



Innovators



Network
Info Flow



Tests:
 - 1 month
 - 586 new docs
 - 1,170 users

IF: Graphical Information Flow Model

TIF: Joint Topic Detection + Information Flow Model

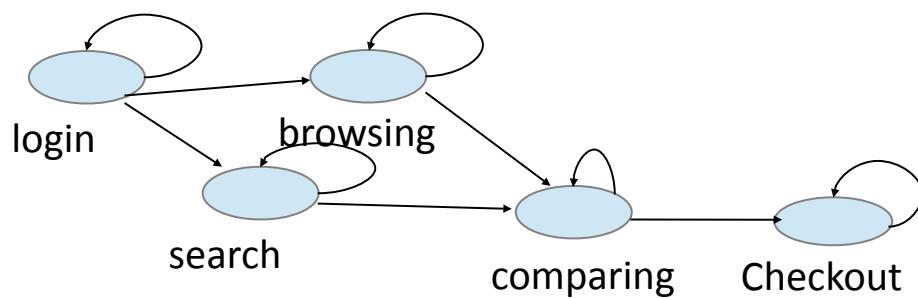
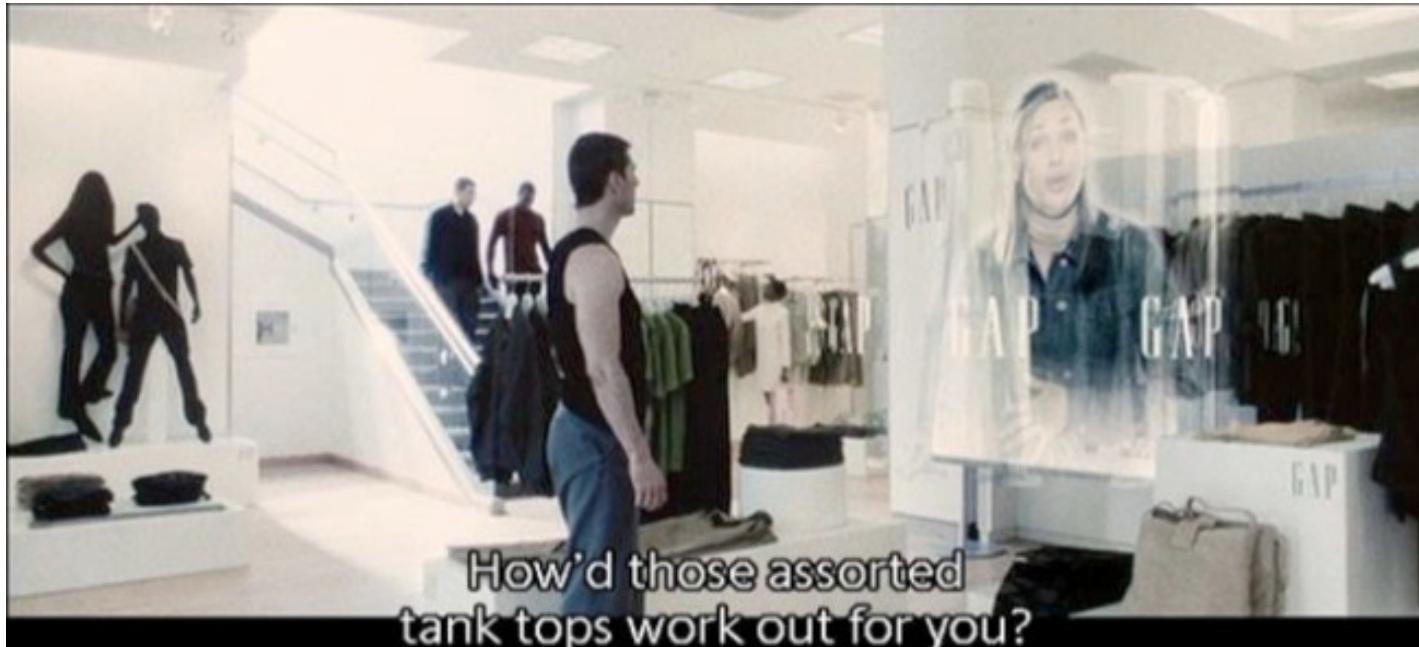
→ Comparing to Collaborative Filtering (CF) + Similar People
 Precision: IF is 91% better, TIF is 108% better
 Recall: IF is 87% better, TIF is 113% better

Customer Behavior Sequence Analytics

Markov
Network

Latent
Network

Bayesian
Network

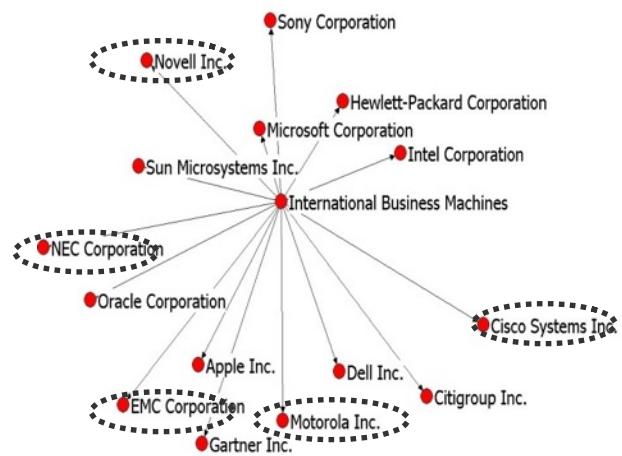


- Behavior Pattern Detection
- Help Needed Detection

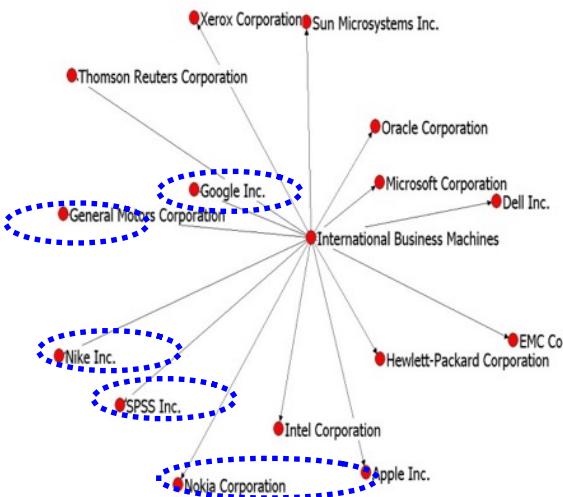
Use Case 4: Graph Analytics for Financial Analysis

Goal: Injecting Network Graph Effects for Financial Analysis. Estimating company performance considering correlated companies, network properties and evolutions, causal parameter analysis, etc.

- IBM 2003



- IBM 2009



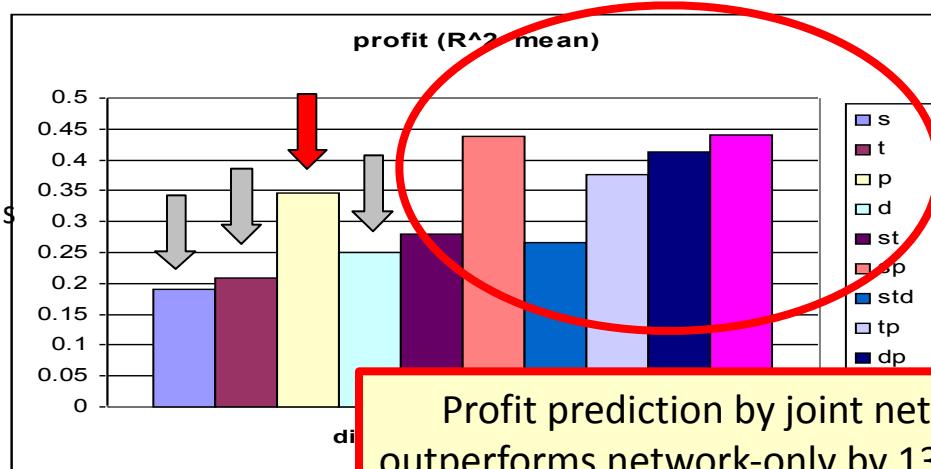
- Data Source:

- Relationships among 7594 companies, data mining from NYT 1981 ~ 2009

Targets: 20 Fortune companies' normalized Profits

Goal: Learn from previous 5 years, and predict next year

Model: Support Vector Regression (RBF kernel)



Network feature:

s (current year network feature),
 t (temporal network feature),
 d (delta value of network feature)

Financial feature:

p (historical profits and revenues)

Use Case 5: Social Media Monitoring

Ching-Yung Lin | Search www.ibm.com

System G SMISC Social Media Monitoring

Home | Live | Forensics | Research Projects | People | News

Select CIO Catetory(-ies): EXECDB BLADE HRTEANT IBM SecurityAnalysis SWG WATSON or Word: Egypt language: Arabic

Total Tweets: 231
 Positive: 35 15%
 Negative: 31 13%

EGYPT wearing @RawyaRageh beauty brutality More ||| Am Egypt's 12 police hijab Er 14 dozen sponge allege port Egypt than Cairo you my Egyptian Said egypt lady call

Saloom Butilla @SaloomButilla RT @Lion_King_Bhr: إثناء الصنفرين الغرفة في 19/2/2013 #الغرف على المركب العامة ورجل الأمن #Bahrain #Egypt #KSA #UAE #News h ... Translation: RT "@Lion_King_Bhr": The traitors in Bahrain Safavid attack on public utilities and security men, 2/19/2013 "LBahrain" #Egypt "LSyria" "LKSA" "LUAE" "LNews" h * * * --Wed Feb 20 17:57:58 2013

Zenza Raggi fan-club @Zenzaclub Private Gold 64: Cleopatra 2 // A sect that worships ancient Egypt is attempting to bring Cleopatra back to life... http://t.co/TcvMDiwb --Wed Feb 20 17:57:53 2013

@SH_QalamSara RT @HebaFaroq: An #Egyptian beauty :) ▶ http://t.co/59bzb5f3 --Wed Feb 20 17:57:53 2013

Mona Metwally @monametwally RT @EgyBloodBank: مريض محتاج متبر عن نم مستكتي الجامعه بالاسمايلوريه فصله نم اب مرجعي 01024705247 #Egypt # مصر # http://t.co/5oO6mtZ5. Translation: . RT * @EgyBloodBank*: A

IBM CIO monitoring categories

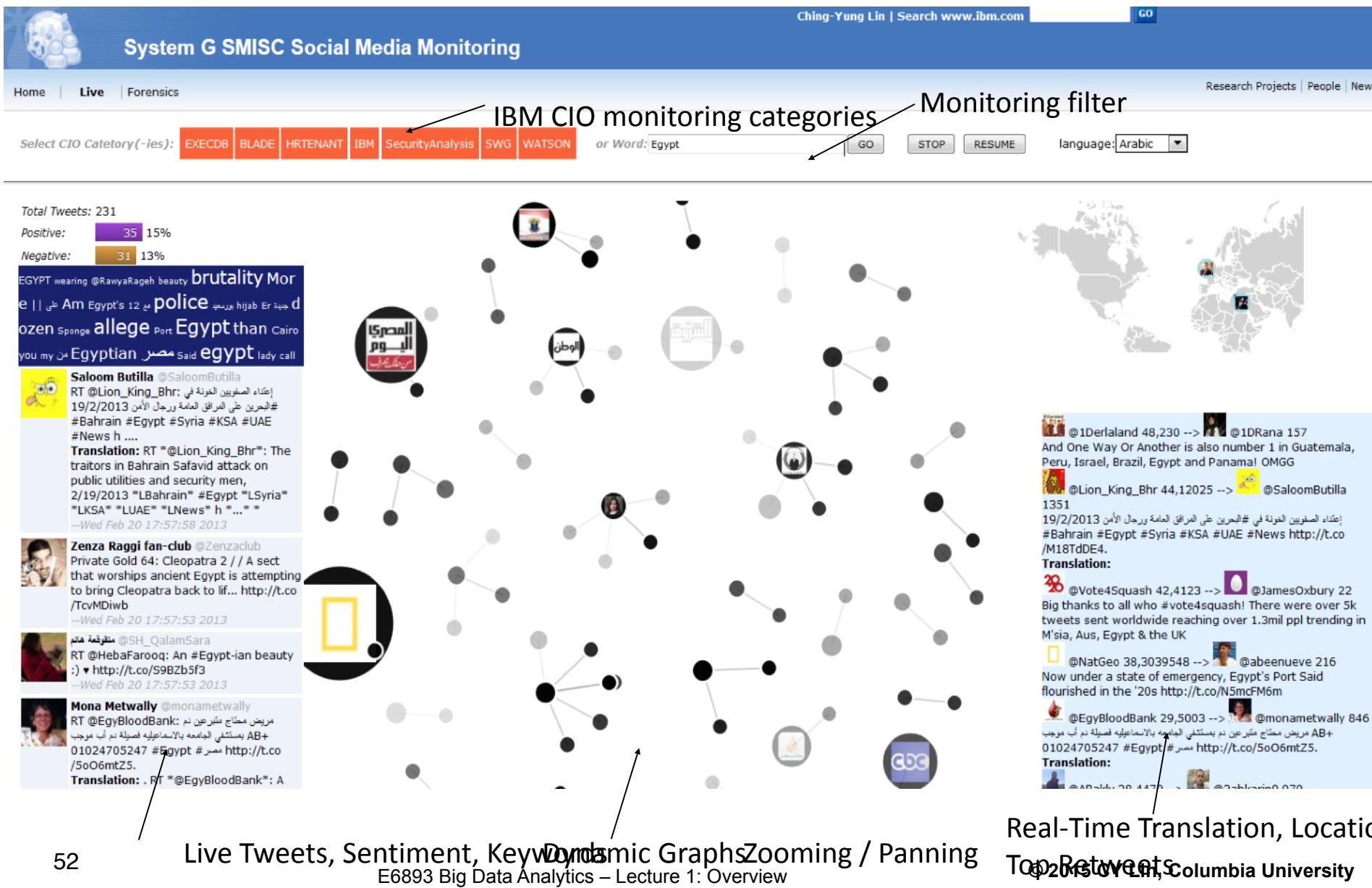
Monitoring filter

Live Tweets, Sentiment, Keywords, Dynamic GraphsZooming / Panning

E6893 Big Data Analytics – Lecture 1: Overview

Real-Time Translation, Location, Top Retweets

© 2013 IBM, Columbia University



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:

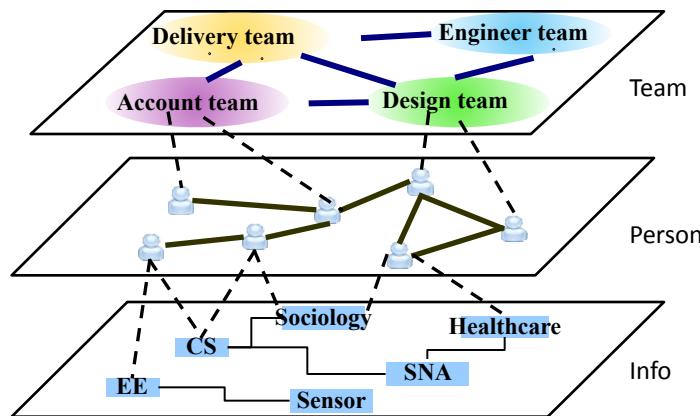
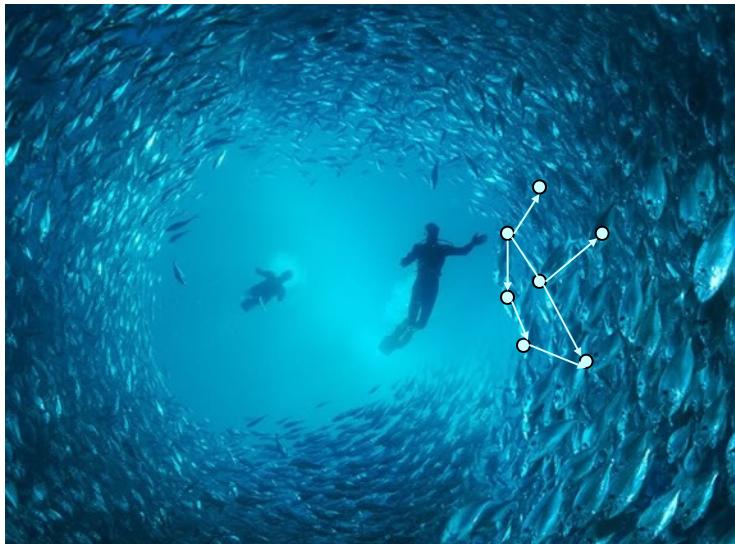
- Task 2.1. Real-Time and Large-Scale Social Media Mining
- Task 2.2. Role and Function Discovery
- Task 2.3. Detecting Malicious Users and Malware Propagation
- Task 2.4. Emergent Topic Detection and Tracking
- Task 2.5. Detecting Evolution History and Authenticity of Multimedia Memes
- Task 2.6. Synchronistic 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



Heterogeneous Synchronicity Networks Predict Performance

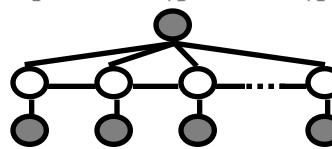
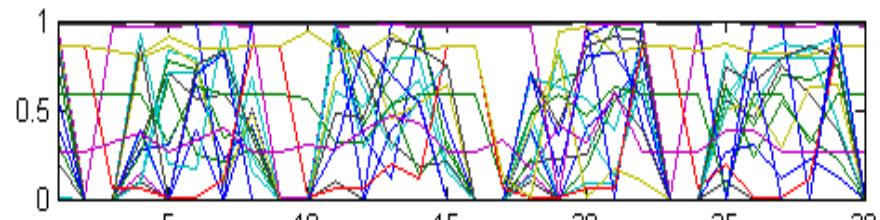


Outperform existing approaches by up to 18% (SDM 13)

One-class HCRF to detect temporal anomalies



Detected as top 1 anomaly in Sandy Tweets



Outperform existing approaches by up to 180% (IJCAI 13)

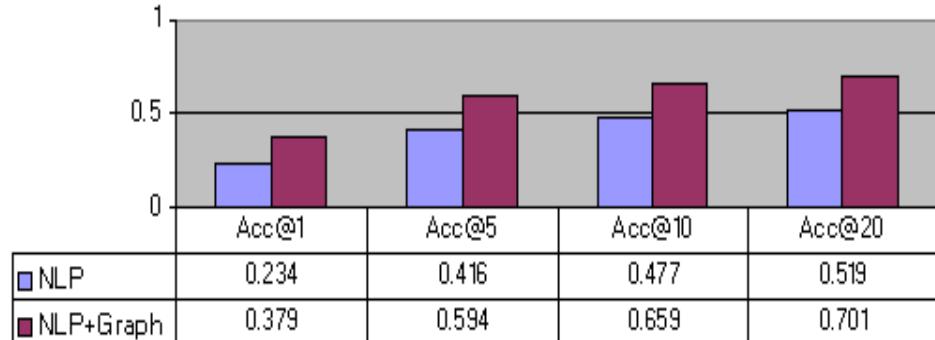
- Motivation:

- Info morph: new links keep emerging to give new meaning to existing phrases

- Approach:

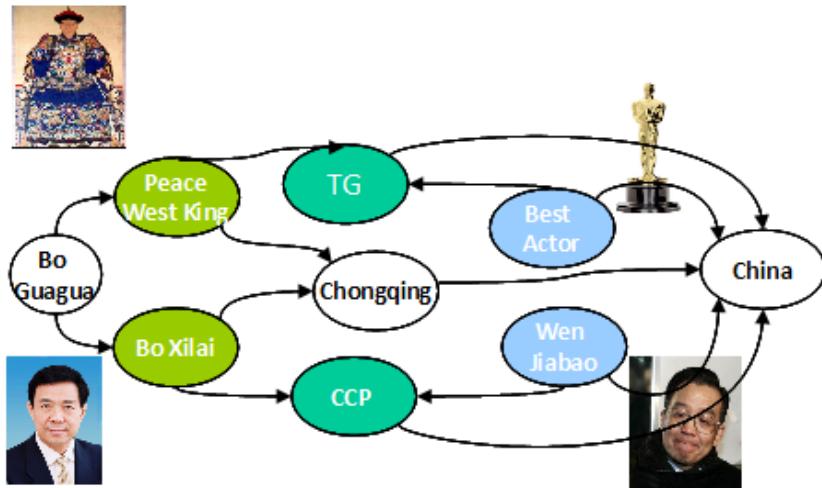
- Compare characteristics of meta-paths between nodes in heterogeneous networks

Entity morph resolution accuracy
(ACL 2013)



Peace West King from *Chongqing* fell from power, still need to *sing red songs*?

- *Bo Xilai* led *Chongqing* city leaders and 40 district and county party and government leaders to *sing red songs*.



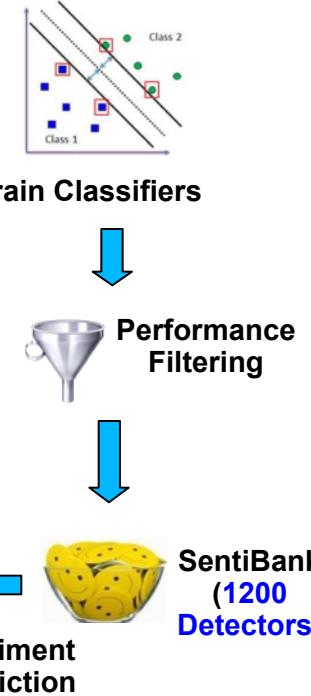
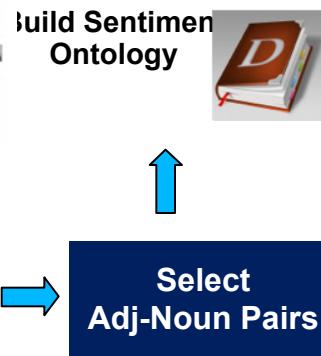
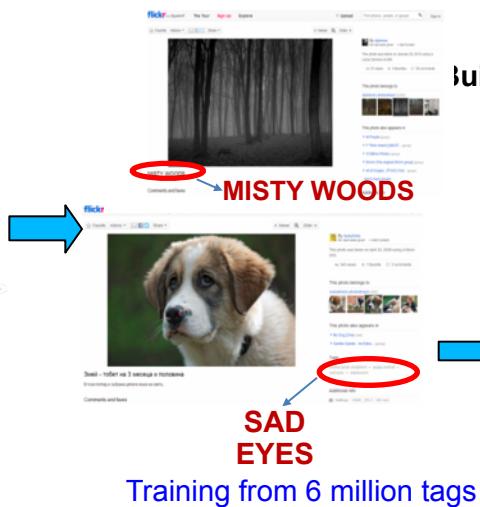
$$\sum_{i=1}^N p_m(x_i) \log \frac{p_m(x_i)}{p_e(x_i)} + p_e(x_i) \log \frac{p_e(x_i)}{p_m(x_i)}$$

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



Sentiment Prediction

Experiment on Sentiment Detection Accuracy on Twitter



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



Text	0.43
Visual	0.70
T+V	0.72

Cognitive Feeling Detection on Images

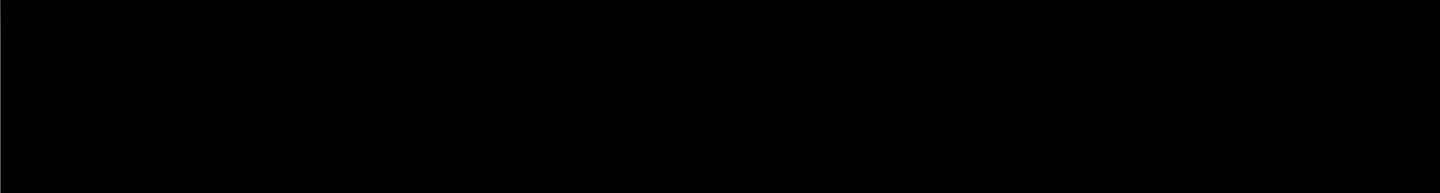


1CAj-9Na4M5-aTuwj4-cdx7Fu-bg7CiV-9PTDrZ-8vrfYC-8XwuK...   



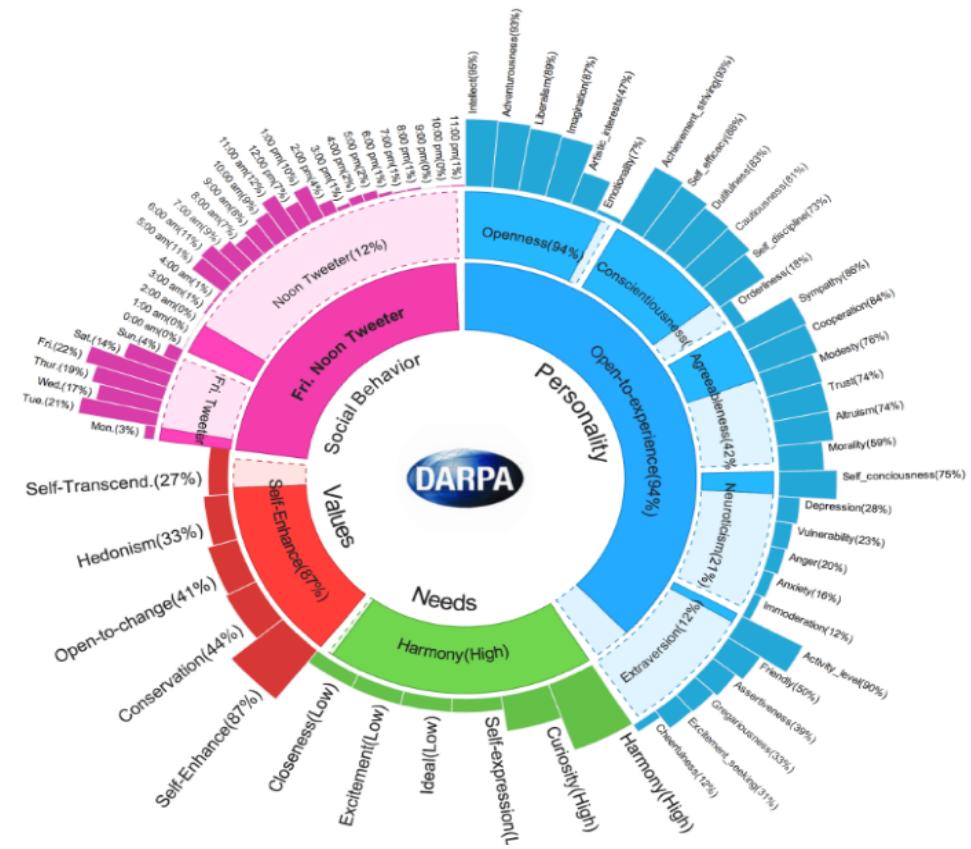
Nice pictures, interesting writing. A beautiful little girl.
 Nice treatment of a fantastic capture. A wonderful picture. Have a good day and keep smiling.
 Excellent portrait. Beautiful look. Fantastic light.

[Make a Comment!](#) [More Specific](#) [More Generic](#) [Cancel](#)

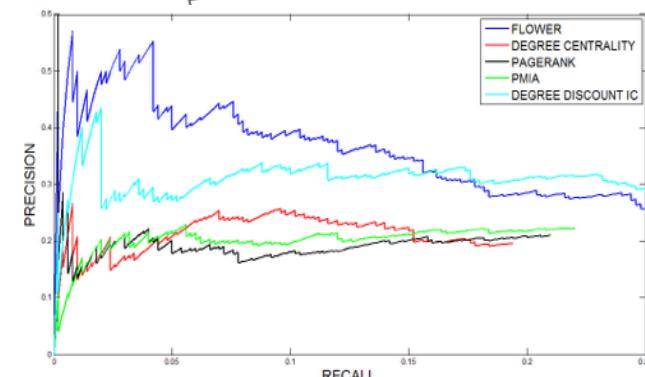


Measuring Human Essential Traits in Social Media

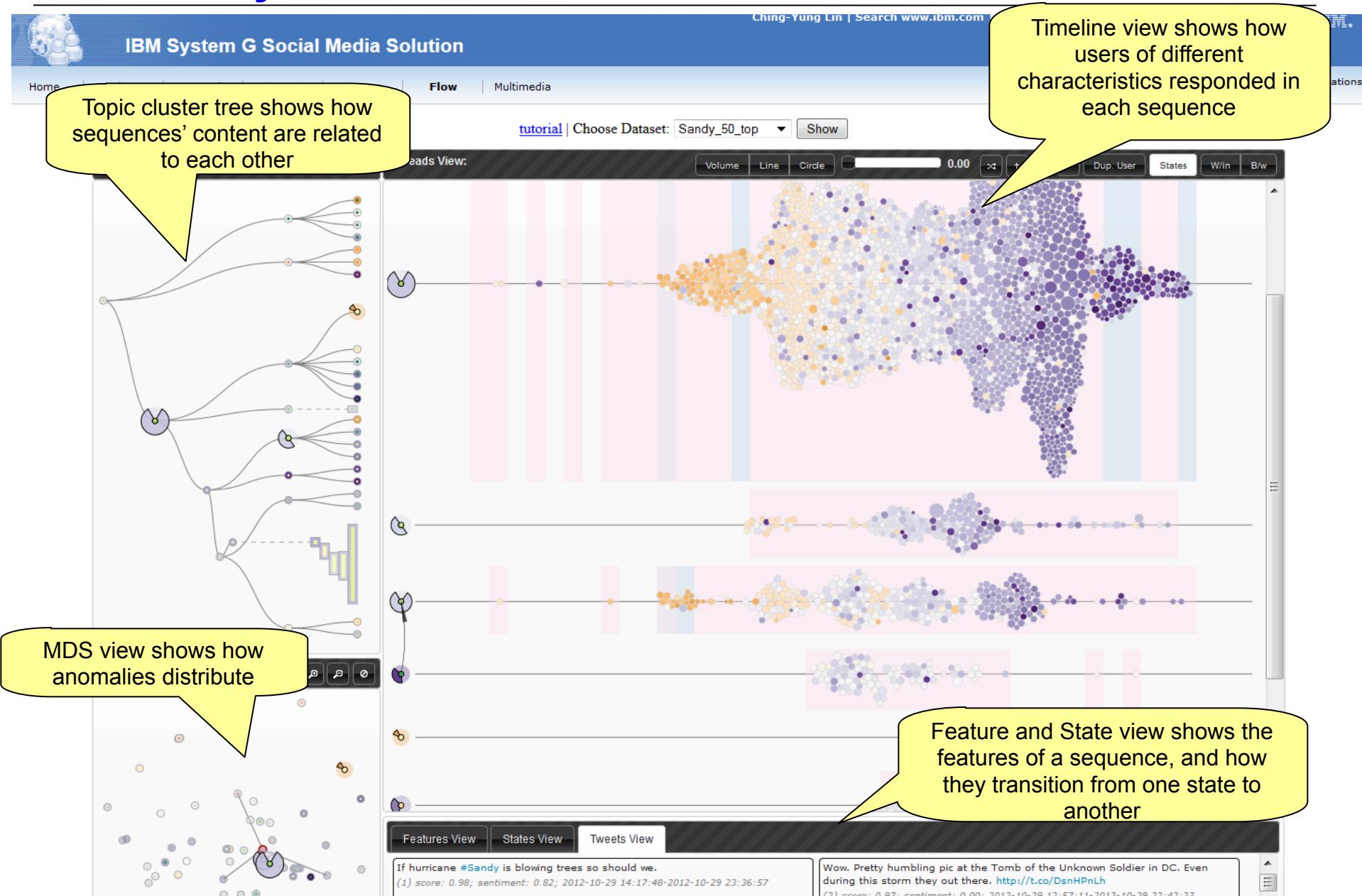
- **Personality:** Mapping personal/organizational social media postings to scores of BIG 5 Personality (*Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism*)
- **Needs:** Mapping personal/organizational social media postings to scores of *Harmony, Curiosity, Self-expression, Ideal, Excitement, and Closeness*.
- **Values:** Mapping personal/organizational social media postings to scores of *Self-Enhance, Conservation, Open-to-Change, Hedonism, and Self-Transcend*.
- **Trustingness and Trustworthness:** 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)



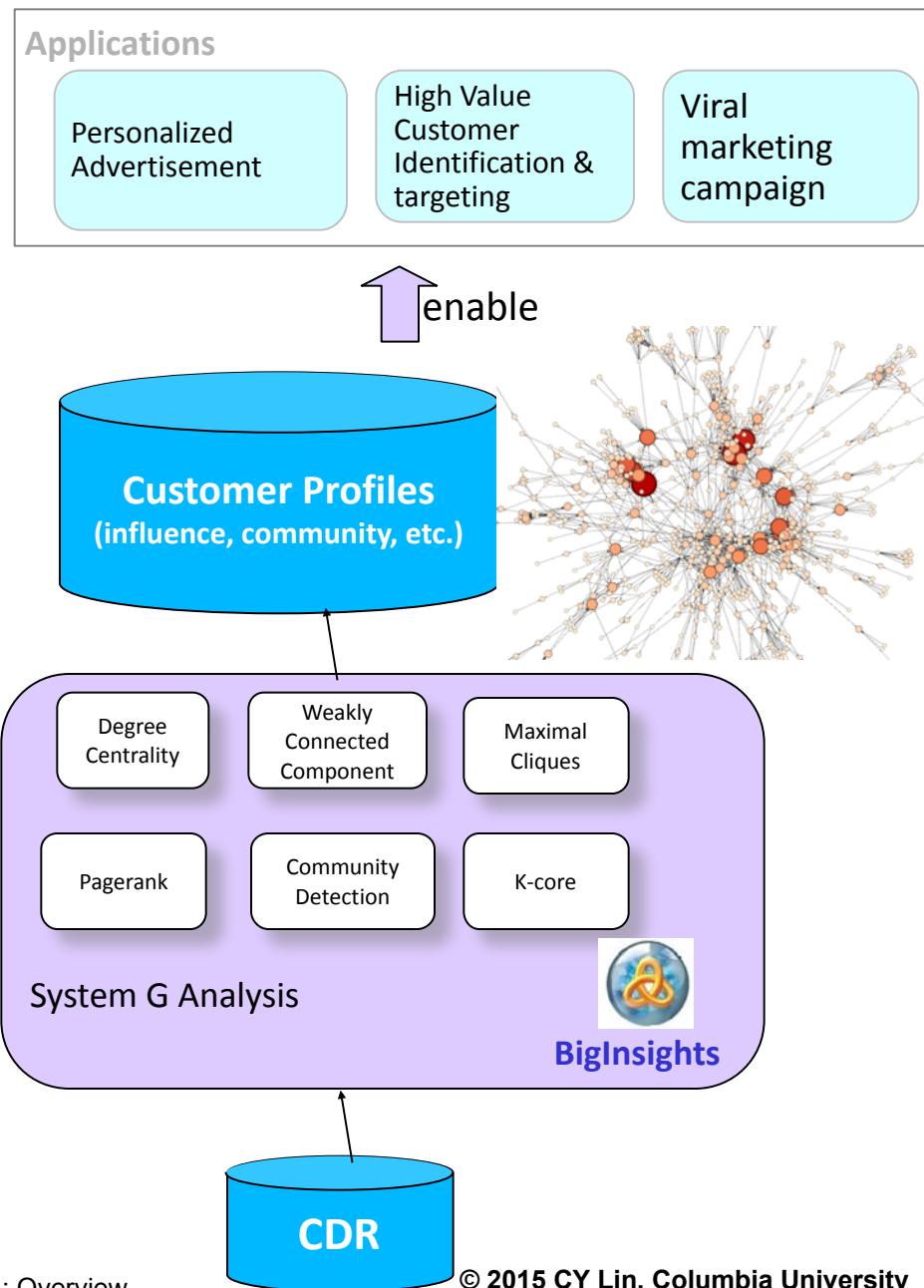
Flow Analytics - I



Use Case 6: Customer Social Analysis for Telco

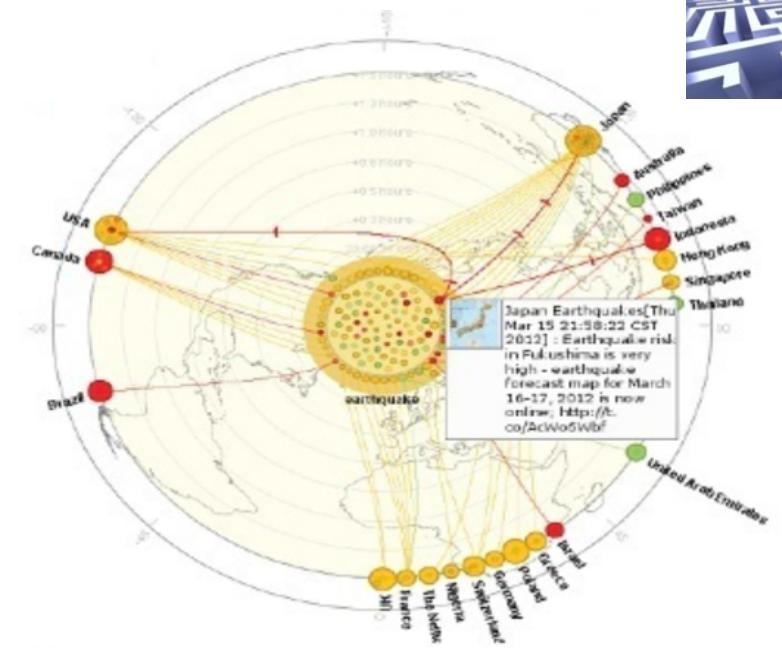
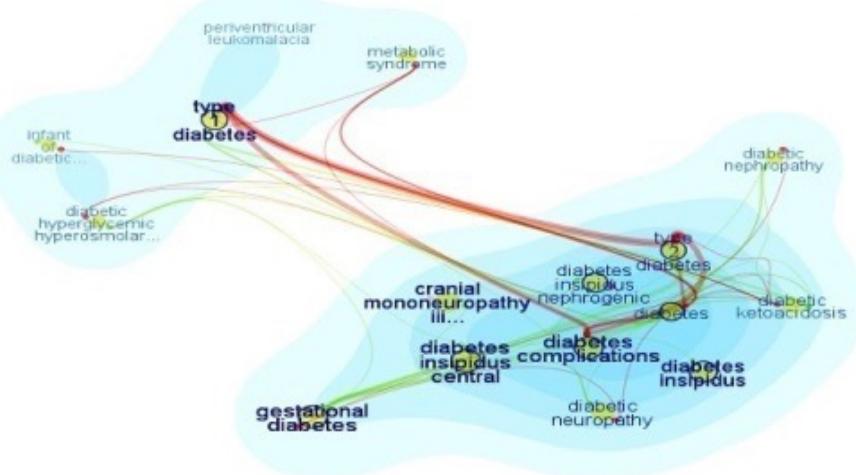
Goal: Extract customer social network behaviors to enable Call Detail Records (CDRs) data monetization for Telco.

- Applications based on the extracted social profiles
 - Personalized advertisement (beyond the scope of traditional campaign in Telco)
 - High value customer identification and targeting
 - Viral marketing campaign
- Approach
 - Construct social graphs from CDRs based on {caller, callee, call time, call duration}
 - Extract customer social features (e.g. influence, communities, etc.) from the constructed social graph as customer social profiles
 - Build analytics applications (e.g. personalized advertisement) based on the extracted customer social profiles

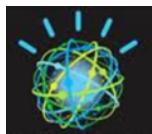




Category 2: Data Exploration



Enhancing:



Vivísmo®

cúram®
SOFTWARE

Huge Network
Visualization

Network
Propagation

I2 3D Network
Visualization

Geo Network
Visualization

Graphical Model
Visualization

Communities

Graph Search

Network Info Flow

Bayesian Networks

Centralities

Graph Query

Shortest Paths

Latent Net Inference

Ego Net Features

Graph Matching

Graph Sampling

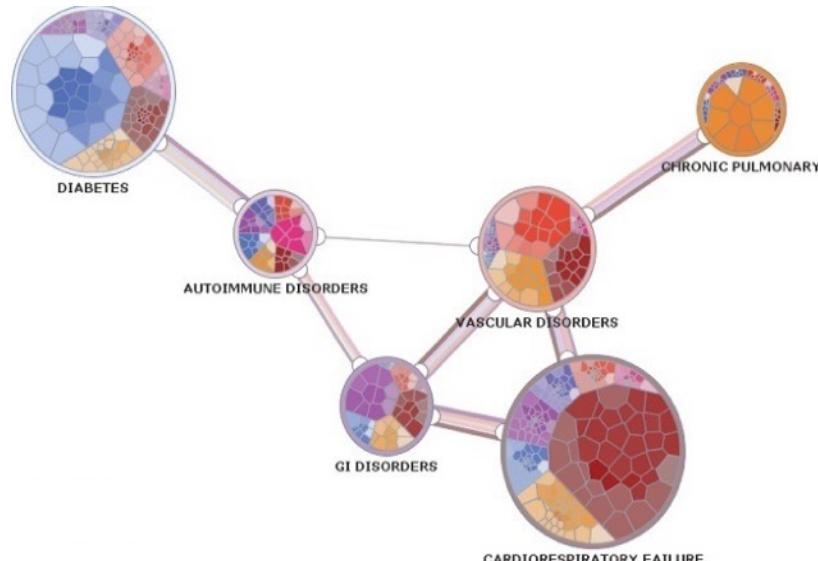
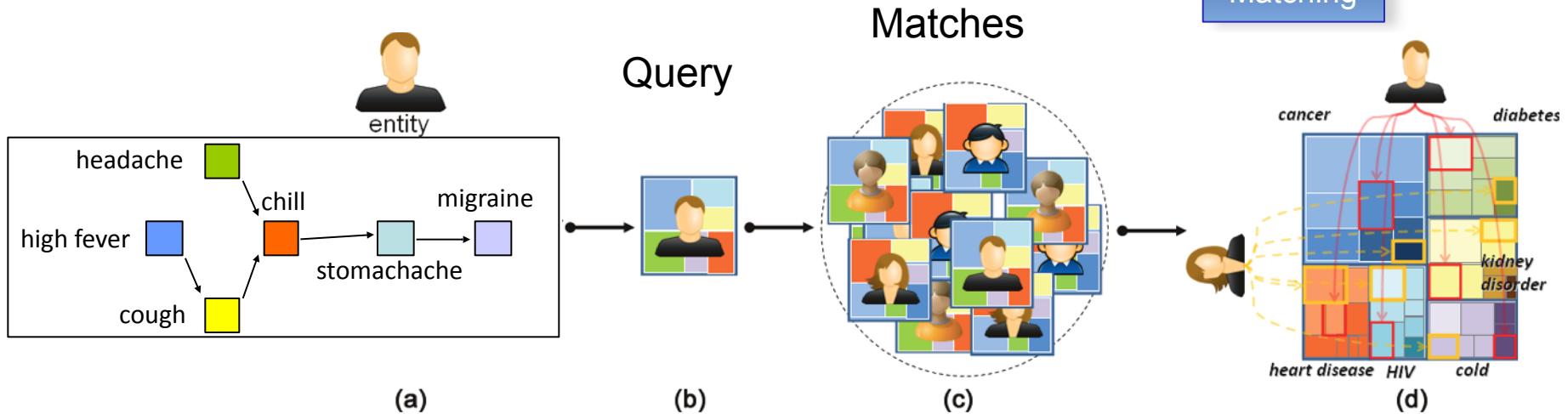
Markov Networks

Middleware and Database

Use Case 7: Graph Analytics and Visualization for Watson



Graph
Matching

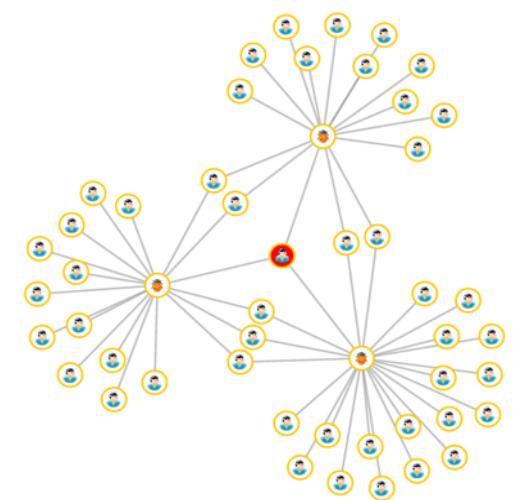
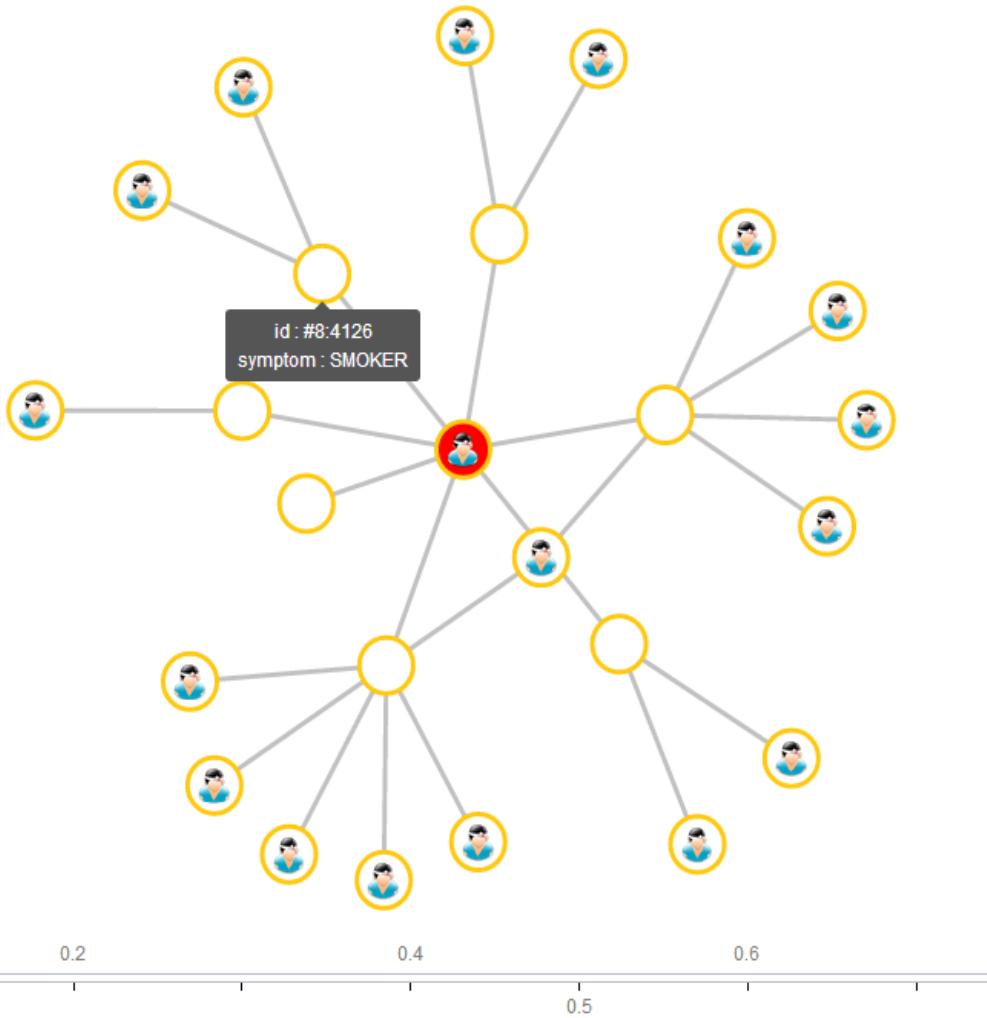


Graph
Communities

Graph Analytics for Watson



query : 8:232 symptom[slots: node imag [attributes: [filter: [by: [| reset

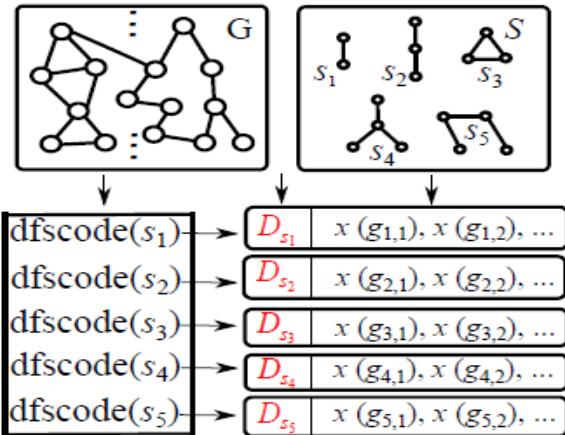




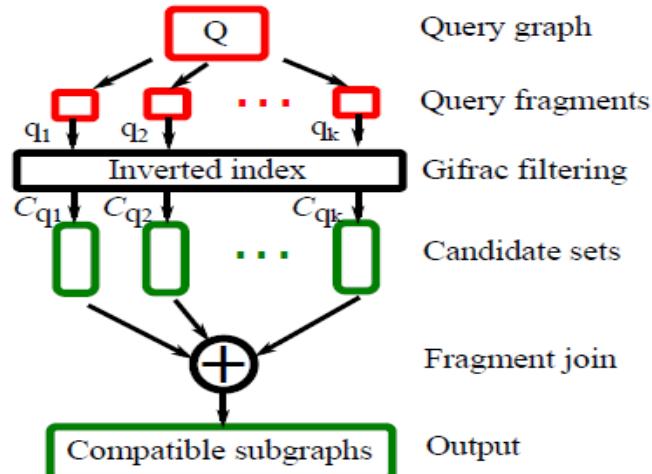
Graph Matching

Fast Graph Matching Algorithm

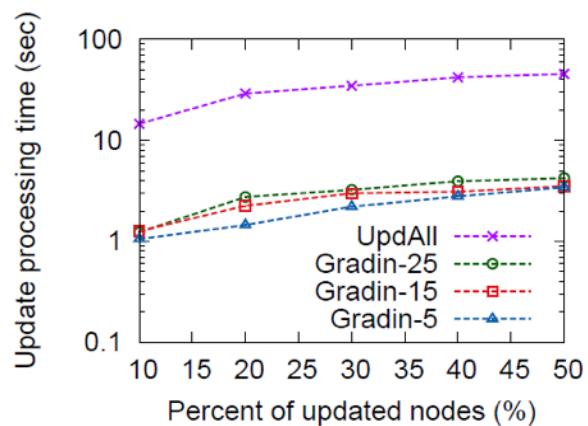
- Data: (CAIDA) 26.5K nodes and 106.8K edges
- Index construction: 13-20 times faster than the prior state-of-the-art
- Query time: close to UpdAll (upper bound) and ~8x faster than UpdNo and NaiveGrid



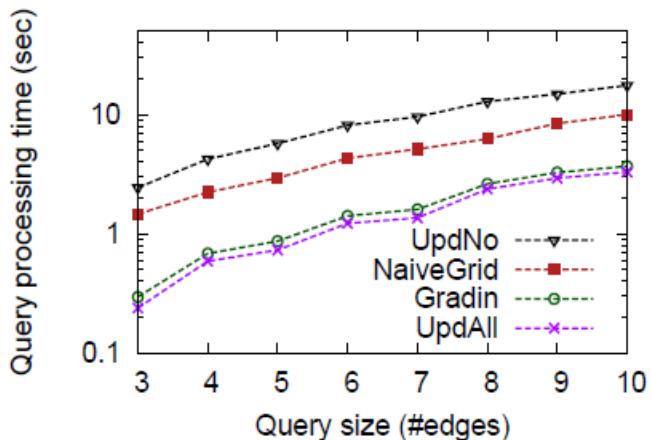
(a) Offline index building



(b) Online query processing

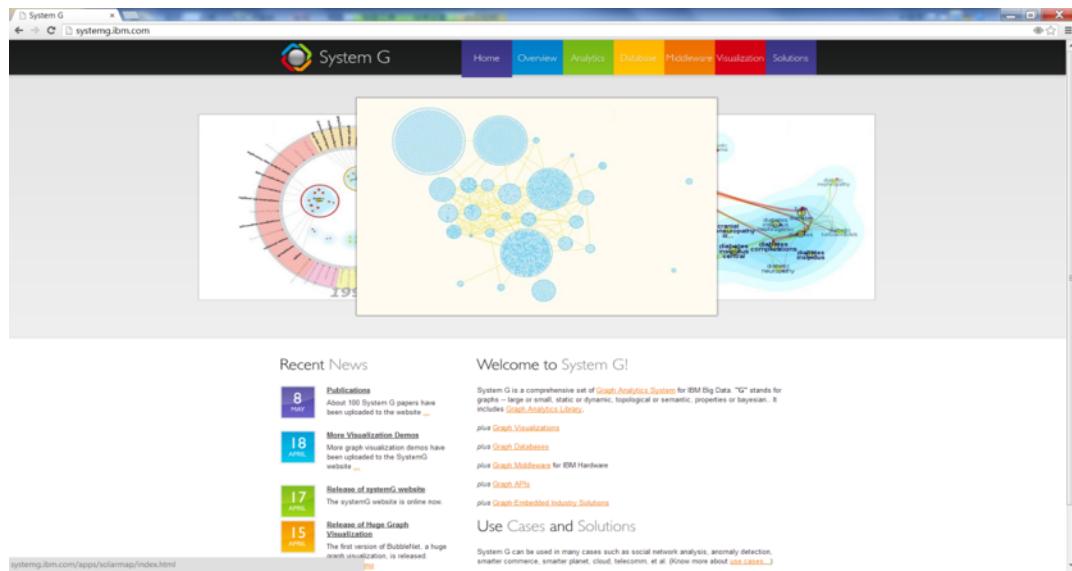


Indexing time



Query processing time

User Case 8: Visualization for Navigation and Exploration



The screenshot shows the System G website interface. At the top, there's a navigation bar with links: Home, Overview, Analytics, Database, Middleware, Visualization, and Solutions. Below the navigation bar, there are three main visualization components: a circular sunburst chart, a cluster-based graph visualization, and a network diagram with a highlighted path.

Recent News:

- 8 MAY Publications: About 100 System G papers have been uploaded to the website ...
- 18 APRIL More Visualization Demos: More graph visualization demos have been uploaded to the SystemG website ...
- 17 APRIL Release of systemG website: The systemG website is online now.
- 15 APRIL Release of huge Graph Visualization: The first version of SolderNet, a huge graph visualization, is released.

Welcome to System G!

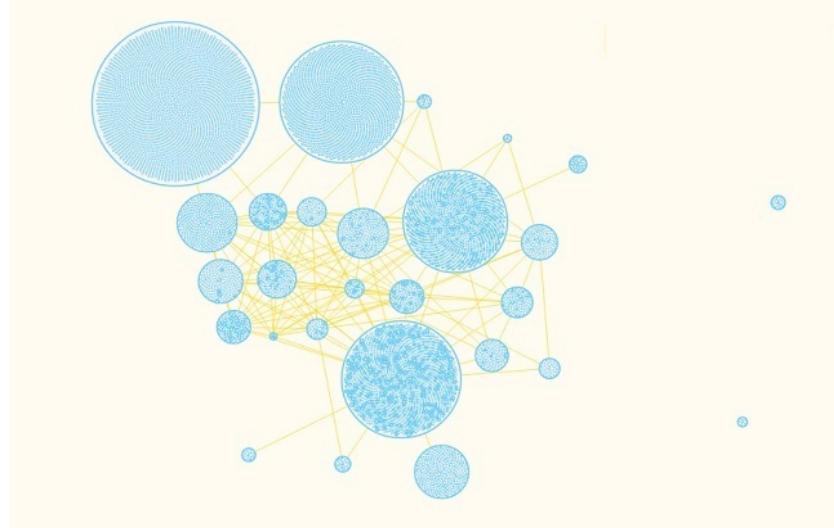
System G is a comprehensive set of [Graph Analytics System](#) for IBM Big Data. "G" stands for graphs – large or small, static or dynamic, topological or semantic, properties or behavior. It includes [Graph Analytics Library](#):

- + [Graph Visualizations](#)
- + [Graph Databases](#)
- + [Graph Middleware for IBM Hardware](#)
- + [Graph APIs](#)
- + [Graph Embedded Industry Solutions](#)

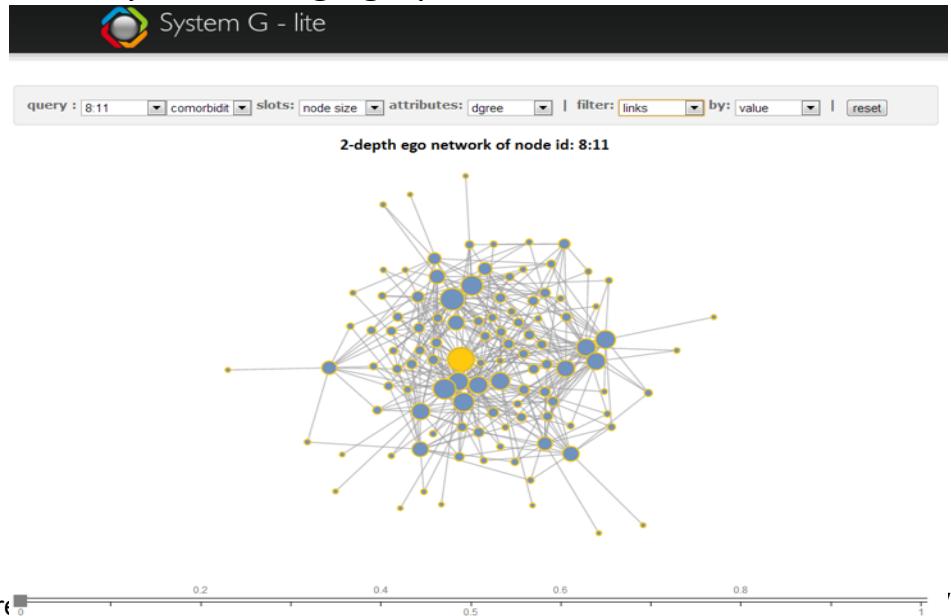
Use Cases and Solutions

System G can be used in many cases such as social network analysis, anomaly detection, smarter commerce, smarter planet, cloud, telecomm, et al. (Know more about [use_cases...](#))

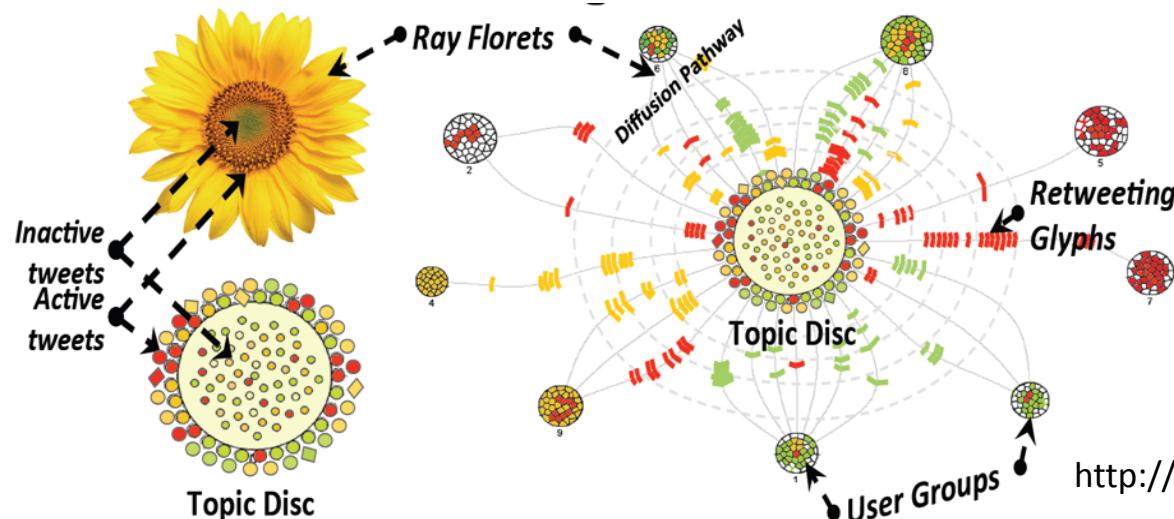
Cluster based huge graph visualization



Query based huge graph visualization



Visualizing Information Diffusion and Divergence

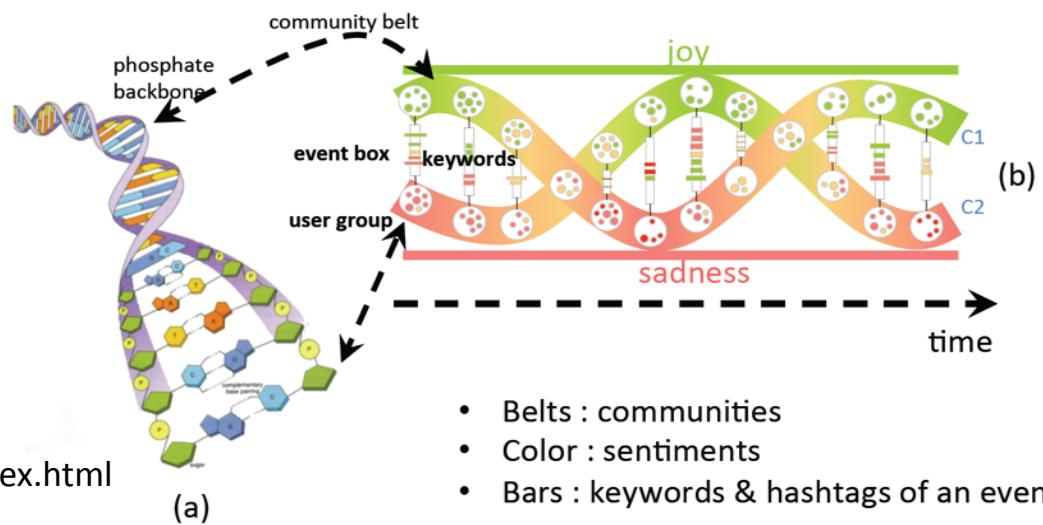


Whisper : Tracing the information diffusion in Social Media

<http://systemg.ibm.com/apps/whisper/index.html>

SocialHelix: Visualizaiton of Sentiment Divergence in Social Media

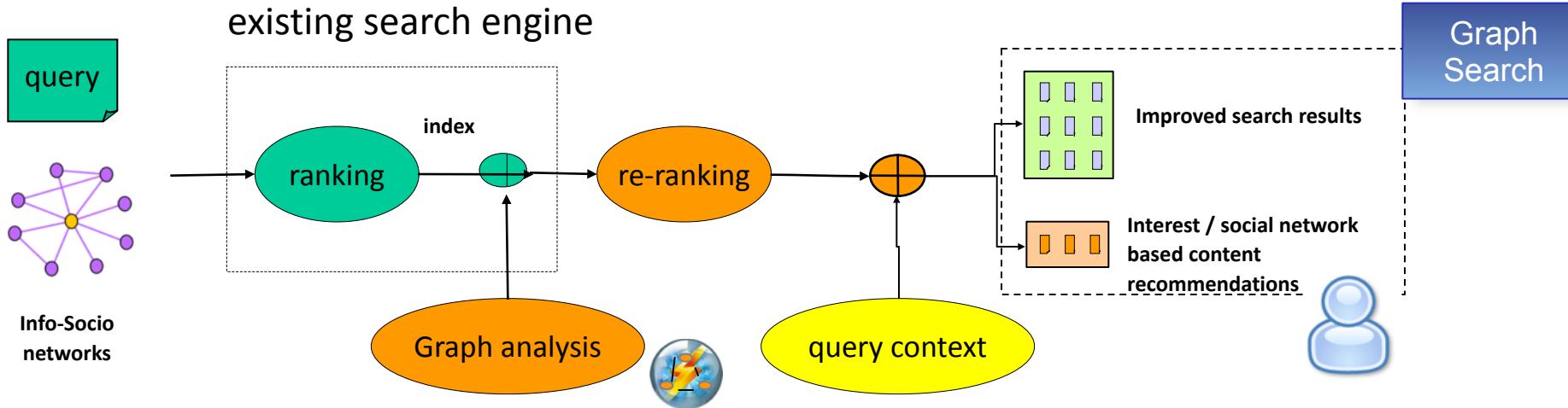
<http://systemg.ibm.com/apps/socialhelix/index.html>



- Belts : communities
- Color : sentiments
- Bars : keywords & hashtags of an event



Use Case 9: Graph Search



Practitioner Portal Translate this page: English

< Return to starting page

Refine Results ?

- ▼ By Tag

Select a tag to filter search results i

View as: cloud | list

 more — less

Search criteria

Go Search within results [Search results](#)

Use "", AND or NOT for better results (default in phrases is AND). E.g. "HR" AND "Human Resource"

► Top search terms, pages and tags
Search keywords: **social business** i

All results Social network results i [Subscribe to s](#)

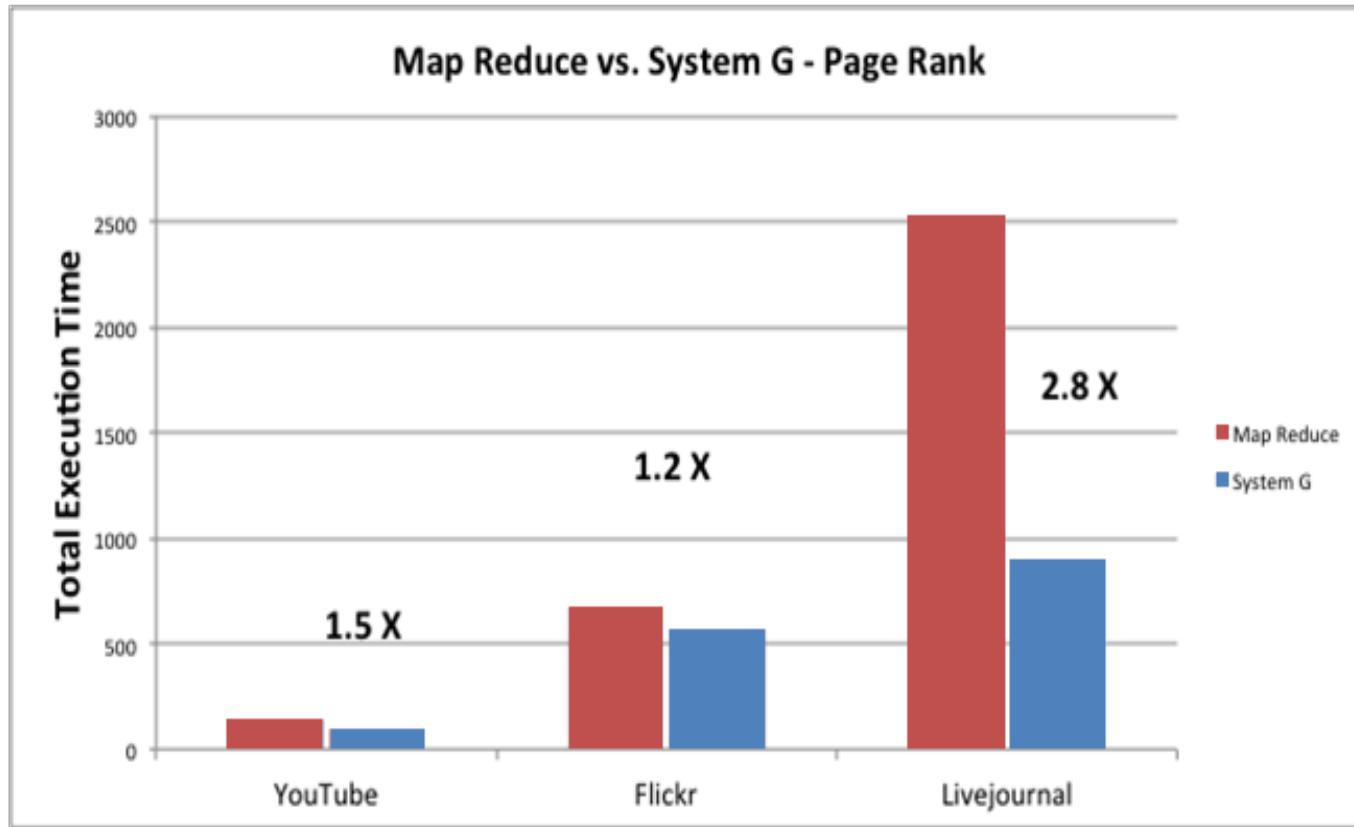
18,577 results found

Title	Relevance	Modified	Bookmarks
IBM Social Business Adoption Quick Start (U.S. English) - Proposal Insert [in Proposal and Presentation Accelerator (PPX)] i	100 %	29 Aug 2012	0
Drive the successful launch and adoption of social business software throughout your organization with a structured engagement comprised of assessments, planning and design consultation, onsite workshops, and team- and skills-building activities.			
Sales Support Information(SSI) i DAGE@stibo.com			

Graph DB and Analytics Co-Processing



Centrailities



System G

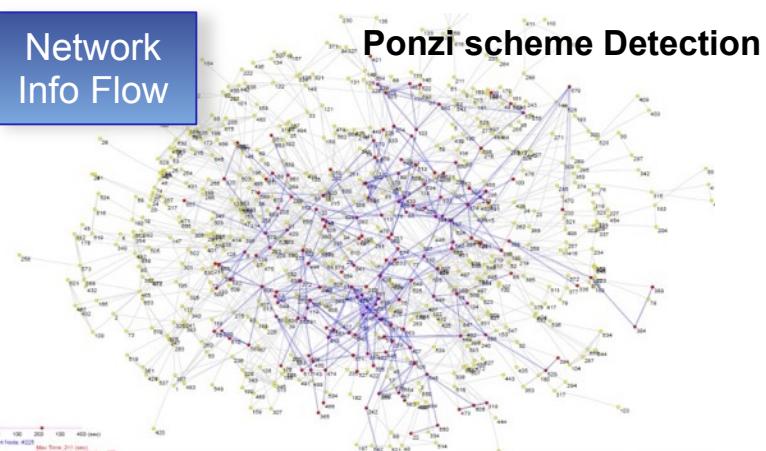


MapReduce

Execution Time in seconds.

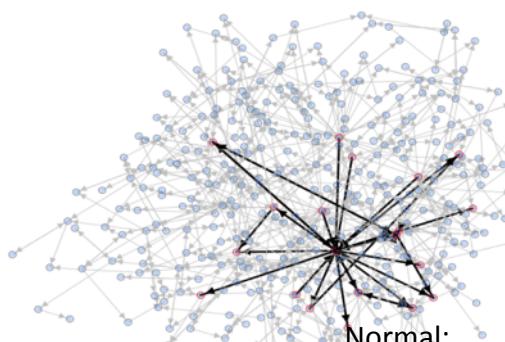
Category 3: Security

Network
Info Flow



Ponzi scheme Detection

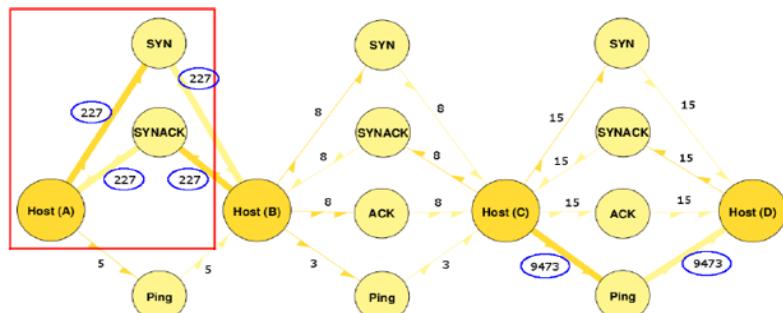
Ego Net
Features



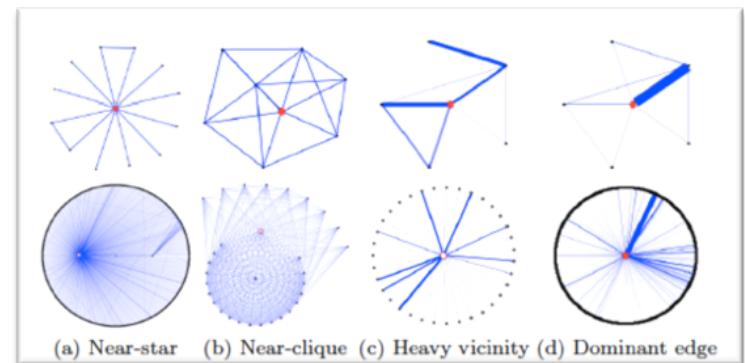
Normal:
(1) Clique-like
(2) Two-way links

Attacker:
Near-Star

Detecting DoS attack



(a) Single large graph representing TCP SYN and ICMP PING network traffic, with two Denial of Service (DoS) attacks taking place.



Graph Visualizations

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 and Database

Use Case 10: Anomaly Detection at Multiple Scales

Based on President Executive Order 13587

Goal: System for Detecting and Predicting Abnormal Behaviors in Organization, through **large-scale social network & cognitive analytics and data mining**, to decrease insider threats such as espionage, sabotage, colleague-shooting, suicide, etc.





THE WALL STREET JOURNAL

Many Past Espionage Cases Had Links to China

To Catch Worker Misconduct, Companies Hire Corporate Detectives

by AILSA CHANG

January 10, 2013 6:25PM



npr

news > business

What's emerged is a multibillion dollar detective industry

npr Jan 10, 2013

“Enterprise Information Leakage Impacted economy and jobs” Feb 2013

Emails

Instant Messaging

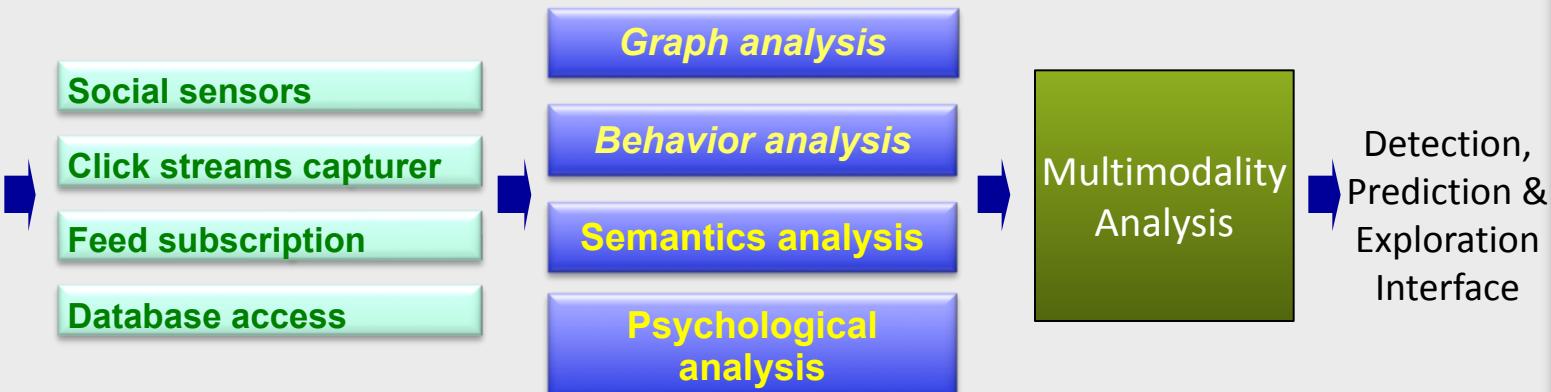
Web Access

Executed Processes

Printing

Copying

Log On/Off



Infrastructure + ~ 70 Analytics

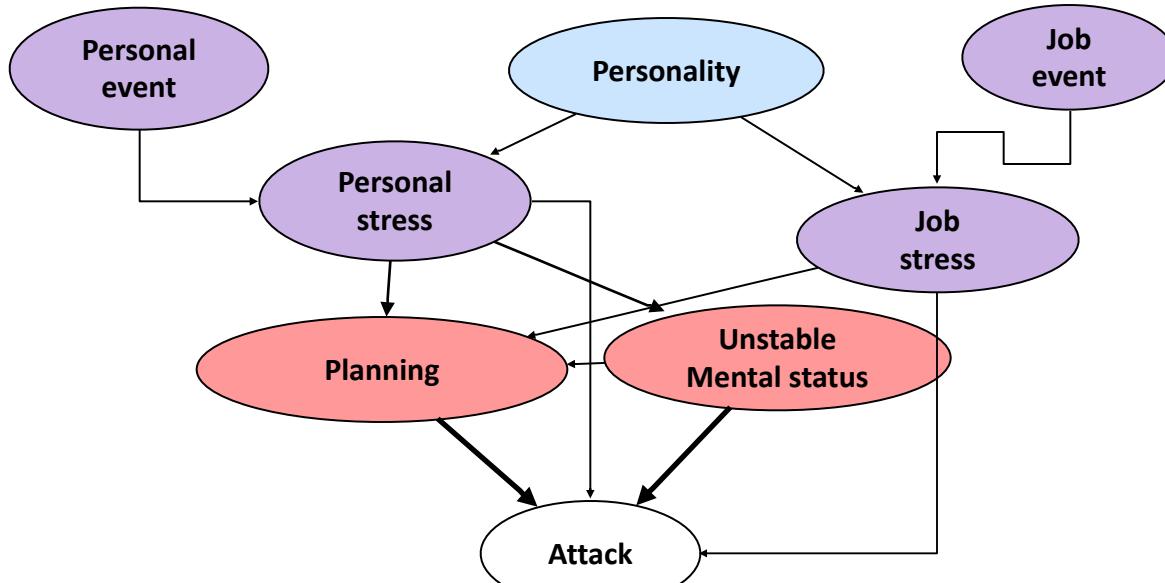
Story – Espionage Example

(1) Personal stress:

- (1) Gender identity confusion
- (2) Family change (termination of a stable relationship)

(2) Job stress:

- Dissatisfaction with work
 - Job roles and location (sent to Iraq)
 - long work hours (14/7)



(1) Unstable Mental Status:

- (1) Fight with colleagues, write complaining emails to colleagues
- (2) Emotional collapse in workspace (crying, violence against objects)
- (3) Large number of unhappy Facebook posts (work-related and emotional)

(2) Planning:

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

(1) Attack:

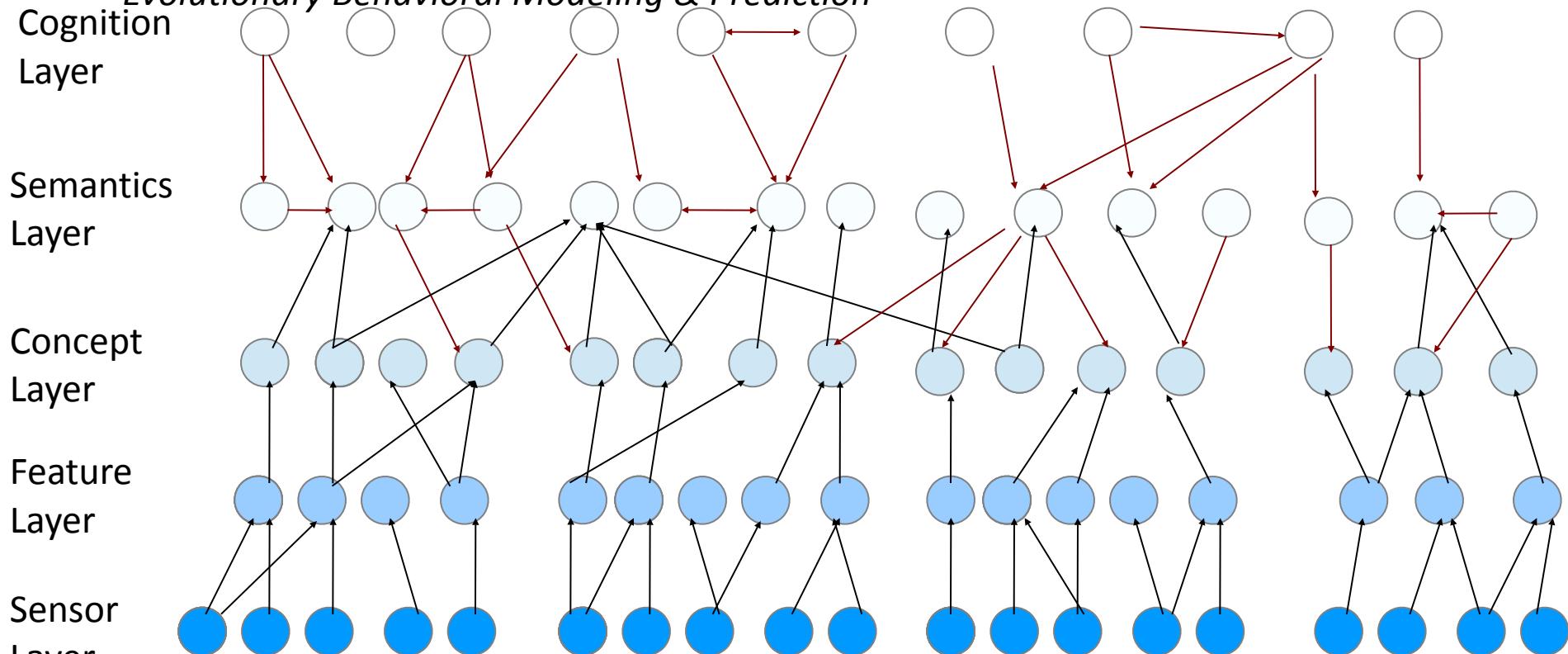
- Brought music CD to work and downloaded/copied documents onto it with his own account



Multi-Modality Multi-Layer Understanding of Human



- Mapping Espionage, Sabotage, and Fraud Use Cases into Five Layers of Classifiers
- Structure Learning
- Evolutionary Behavioral Modeling & Prediction



HR records, Travel records,
Badge/Location records,
Phone records, Mobile records

Transmitted images,
speech content, video
content

future additions?

Example of Graphical Analytics and Provenance

Markov
Network

Latent
Network

Bayesian
Network



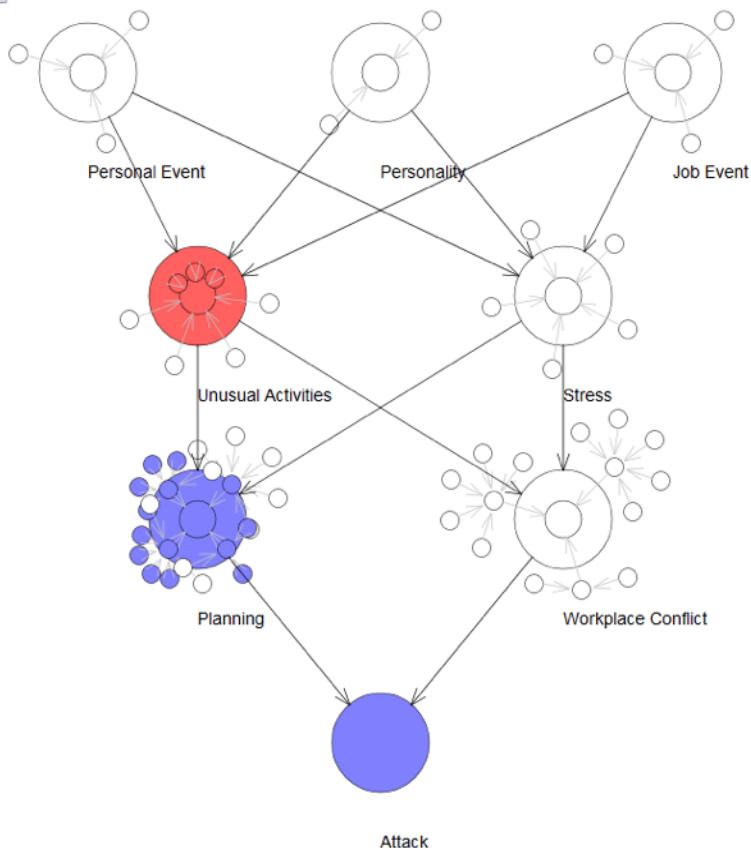
IBM Anomaly Detection at Multiple Scales (ADAMS) Analytics

 Sam Chang
Security Specialist
IT Department
samchang@company.com

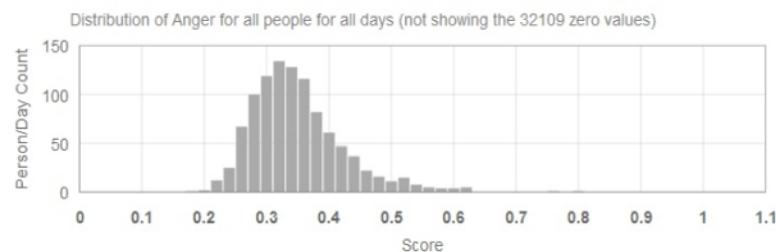


Day 4

Timeline:



Email (Anger)



Top Person/Days -- Anger, group Self

Userid	Date	Anger
0A40E96890D1ED5E62F9C7F19191108D	2012-07-18	81
7B131181E78AF8C08F3A9E605E58227B	2012-07-17	77
41C50A4433E2CC2C34E6DCDFD1290C3F	2012-07-10	63
369730CC2965FD97E8435A4365AF6FFF	2012-07-17	63
2FF547A39E2D9E4D0B5DC6ACB1165441	2012-07-27	62
6826DFDB7FFE74364C4FF7C58089E795	2012-07-10	62
CBD7C1659A41C74998837310639E2817	2012-07-24	62
C00A0A9DB436F130DCC5671D2936CD45	2012-07-13	62
9C2568C8AF6809EAAAAC53E7B267AB00	2012-07-10	62
215291F411748EDD54E2AA43966615A8	2012-07-24	61
BCE7365D13A8396290BFB4CCB1D9DE36	2012-07-25	60

Evaluations on the Real-World Data in Vegas Lab (Oct 2013)

- Each month, 3 cases were inserted (1 abnormal person per case) in the real data.
- Each performer system retrieved top abnormal people out of the 5,500 people per month.
- This chart showed where the 3 IBM 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. Performer 2 did not report results. Performer 3 reported: 3 of the 12 cases Top #50-#100, 6 cases Top #101-#500, and 3 cases beyond Top #501.

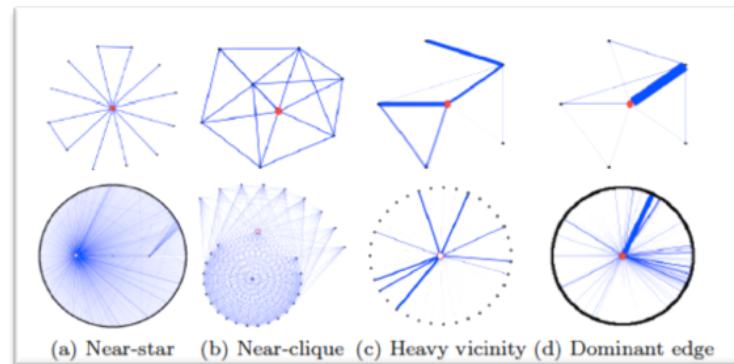
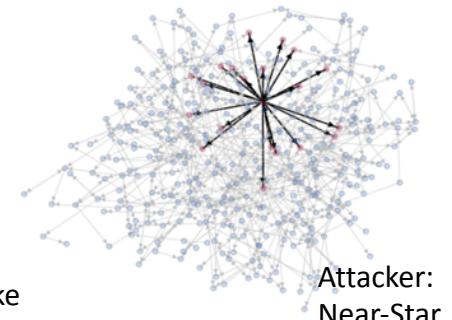
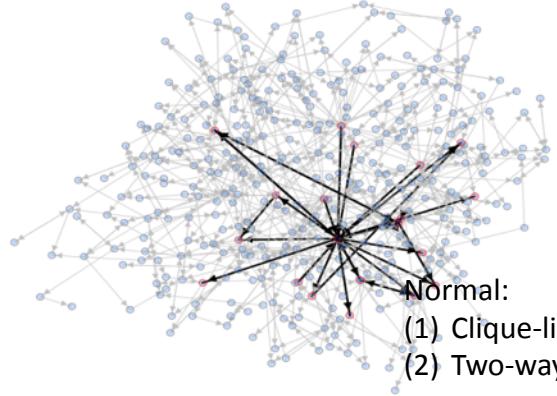
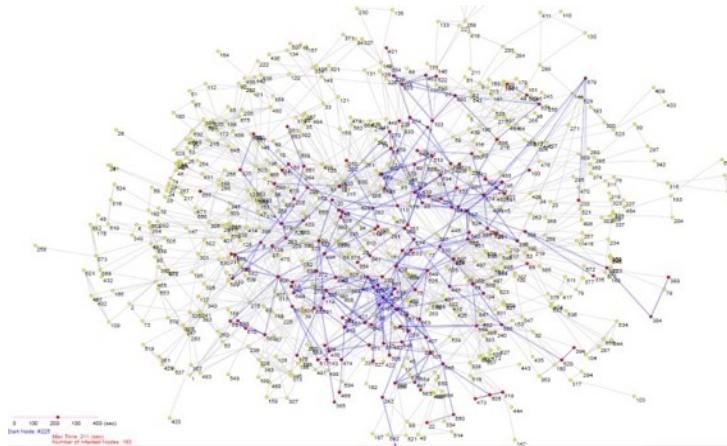
Use Case 11: Fraud Detection for Bank

Network
Info Flow

Ego Net
Features



Ponzi scheme Detection



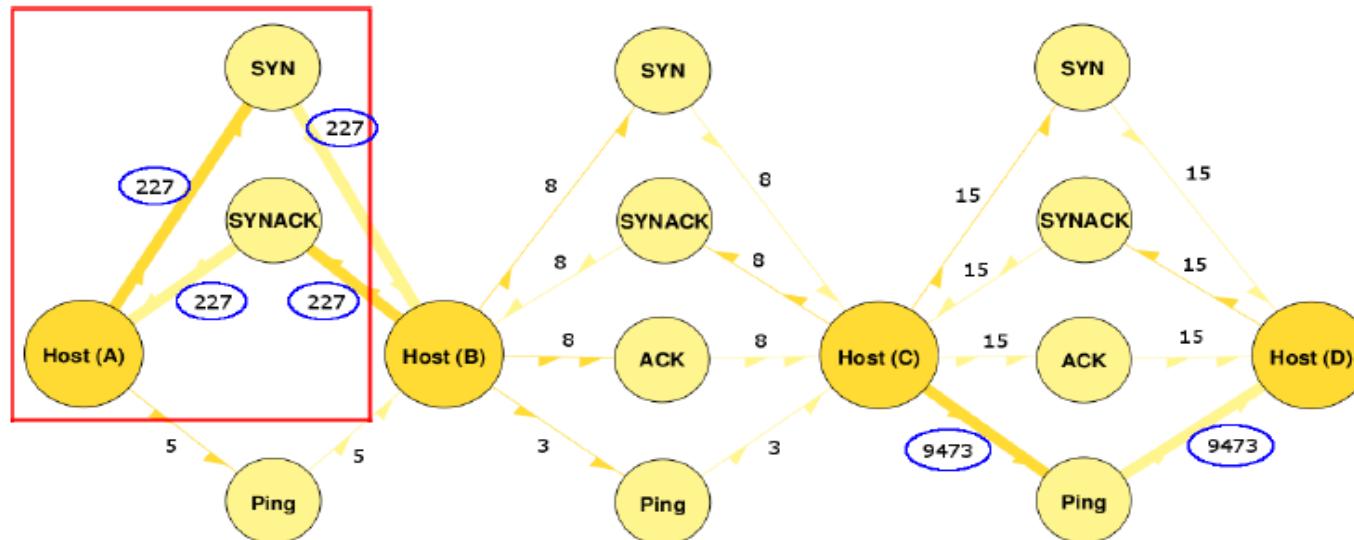
Use Case 12: Detecting Cyber Attacks

Network
Info Flow

Ego Net
Features

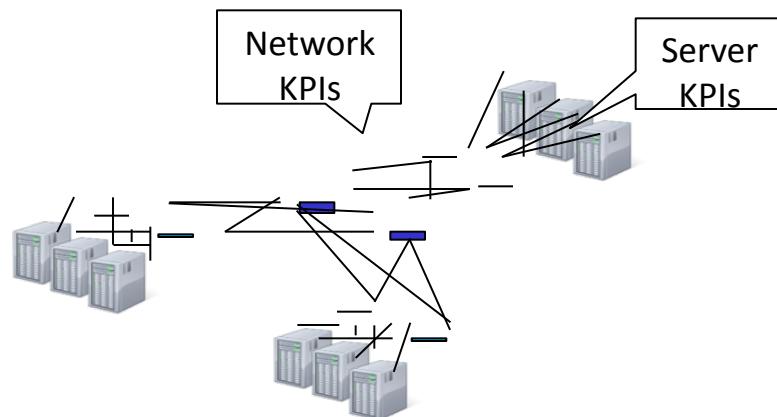


Detecting DoS attack



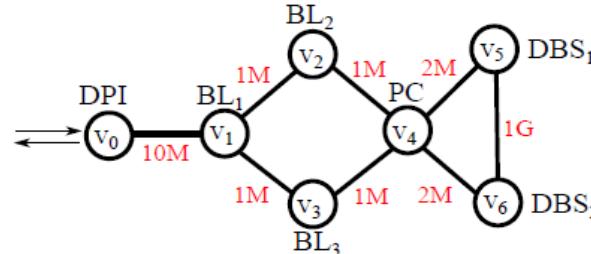
(a) Single large graph representing TCP SYN and ICMP PING network traffic, with two Denial of Service (DoS) attacks taking place.

Category 4: Operations Analysis



Cloud Service Placement

DPI - Deep Package Inspector BL - Business Logic
 PC - Package classifier DBS - DB Server



Memory requirements

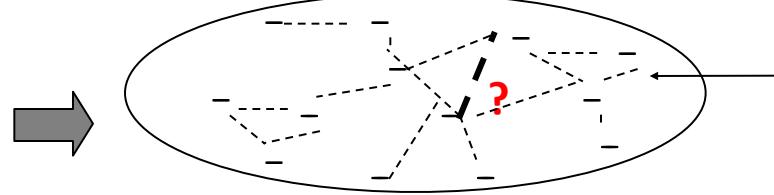
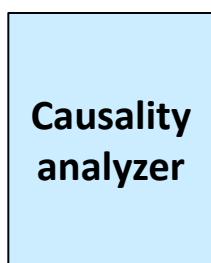
v_0	8G
v_1	2.5G
v_2	2G
v_3	2G
v_4	12G
v_5	20G
v_6	32G



Graph Matching

Bayesian Network

KPI time series (e.g., server performance/load, network performance/load)



- KPI (a time series)
- (potential) pairwise relationship (e.g., causality)

Graph Visualizations

Communities

Graph Search

Network Info Flow

Bayesian Networks

Centralities

Graph Query

Shortest Paths

Latent Net Inference

Ego Net Features

Graph Matching

Graph Sampling

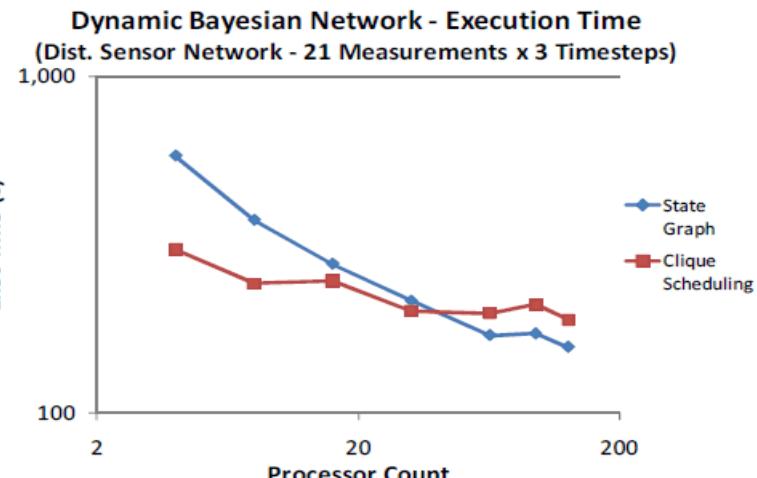
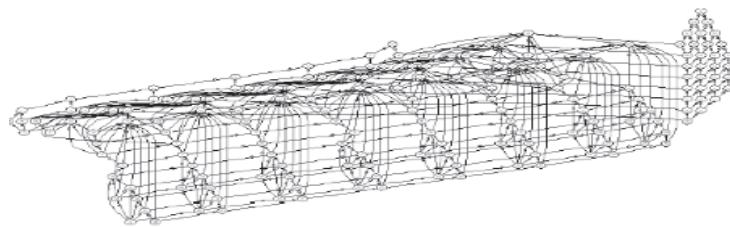
Markov Networks

Middleware and Database

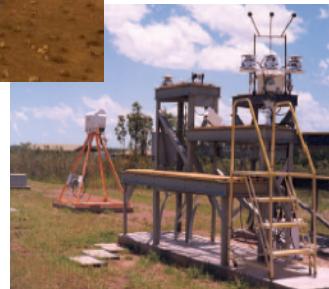
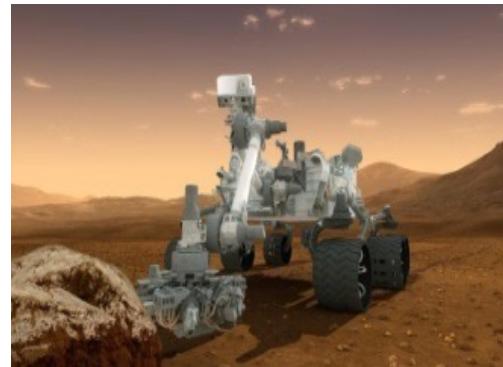
Use Case 13: Smarter *another* Planet

Goal: Atmospheric Radiation Measurement (ARM) climate research facility provides 24x7 *continuous field observations* of cloud, aerosol and radiative processes. **Graphical models** can automate the validation with improvement efficiency and performance.

Approach: BN is built to represent the dependence among sensors and replicated across timesteps. BN parameters are learned from over 15 years of ARM climate data to support distributed climate sensor validation. Inference validates sensors in the connected instruments.



Bayesian Network



Bayesian Network

- * 3 timesteps * 63 variables
- * 3.9 avg states * 4.0 avg indegree
- * 16,858 CPT entries

Junction Tree

- * 67 cliques
- * 873,064 PT entries in cliques

Use Case 14: Cellular Network Analytics in Telco Operation

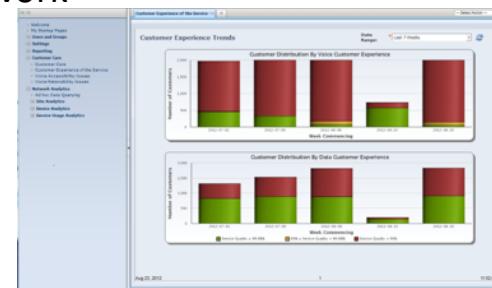
Goal: Efficiently and uniquely identify *internal* state of Cellular/Telco networks (e.g., performance and load of network elements/links) using probes between monitors placed at selected network elements & endhosts

- Applied Graph Analytics to telco network analytics based on CDRs (call detail records): estimate traffic load on CSP network with low monitoring overhead

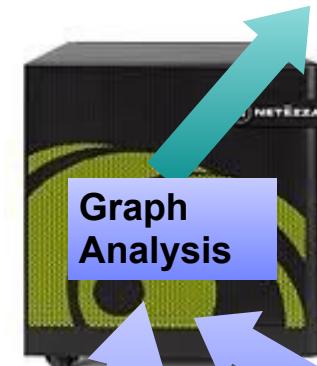
- (1)CDRs, already collected for billing purposes, contain information about voice/data calls
- (2)Traditional NMS* and EMS** typically lack of end-to-end visibility and topology across vendors
- (3)Employ graph algorithms to analyze network elements which are not reported by the usage data from CDR information

- Approach

- Cellular network comprises a hierarchy of network elements
- Map CDR onto network topology and infer load on each network element using graph analysis
- Estimate network load and localize potential problems

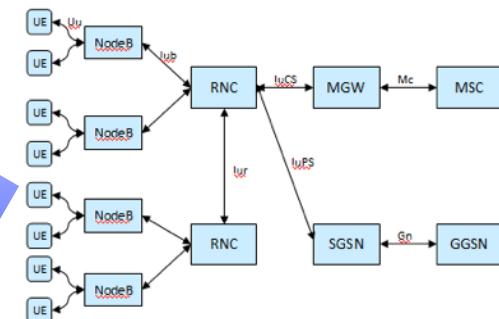


Network load level report



CDR

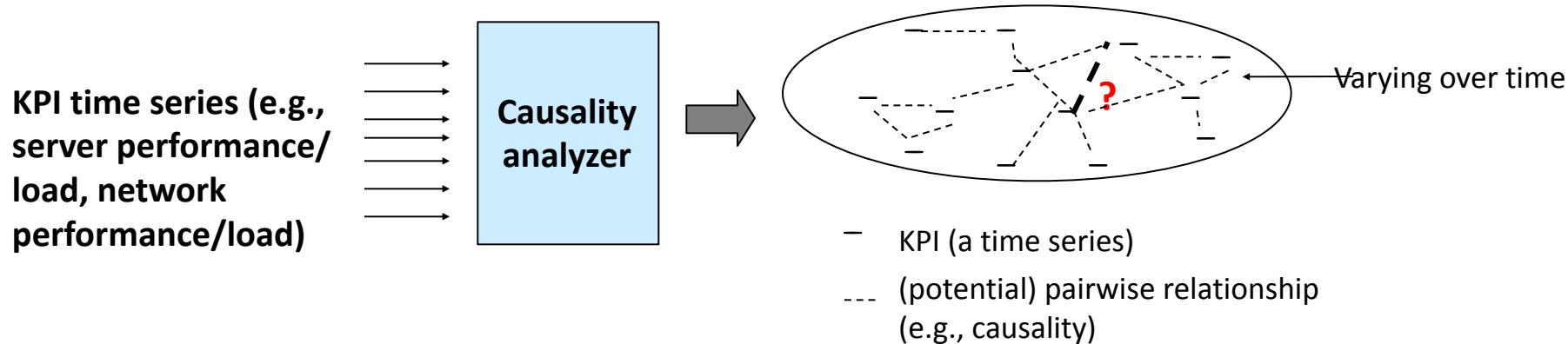
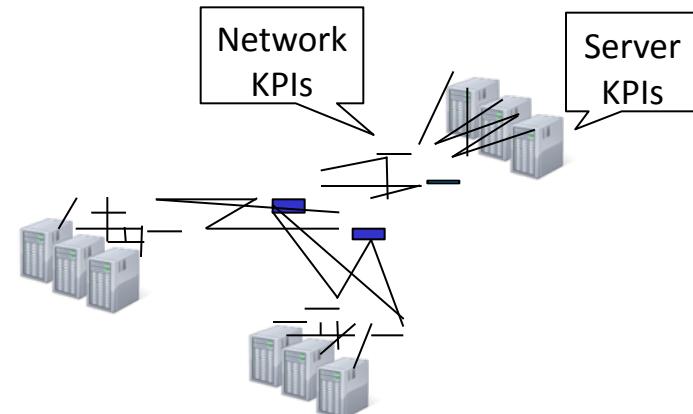
Network topology



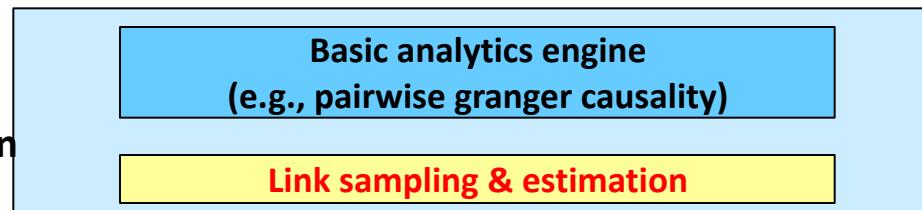
Use Case 15: Monitoring Large Cloud

Goal: Monitoring technology that can track the time-varying state (e.g., causality relationships between KPIs) of a large Cloud when the processing power of monitoring system cannot keep up with the scale of the system & the rate of change

- *Causality relationships (e.g., Granger causality) are crucial in performance monitoring & root cause analysis*
- *Challenge: easy to test pairwise relationship, but hard to test multi-variate relationship (e.g., a large number of KPIs)*



Our approach:
Probabilistic monitoring via sampling & estimation

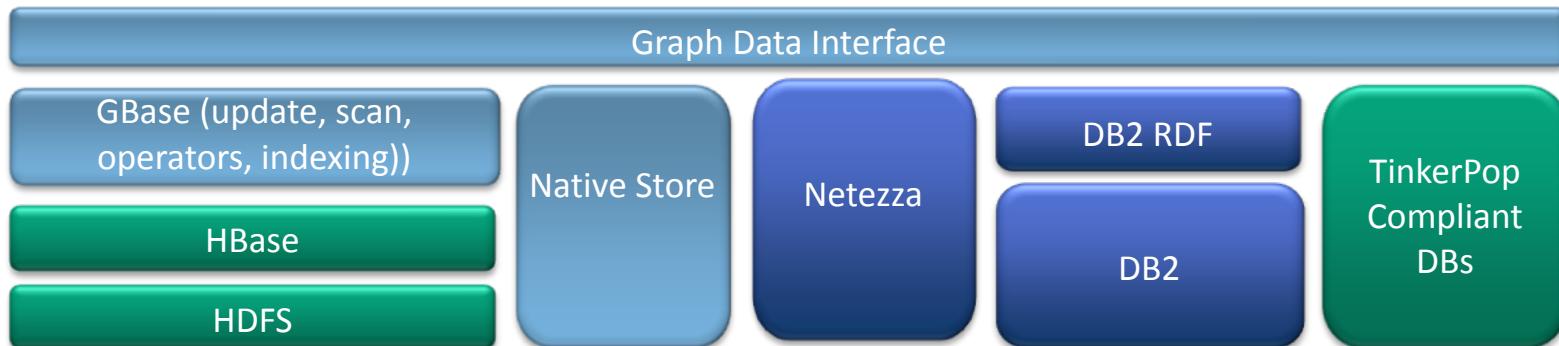
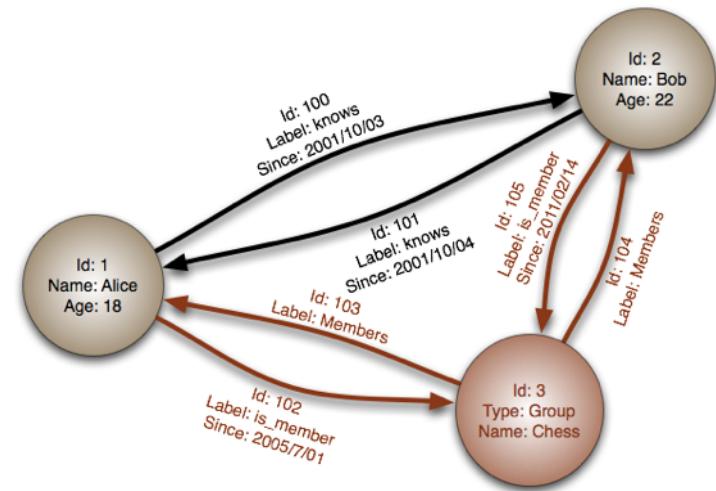




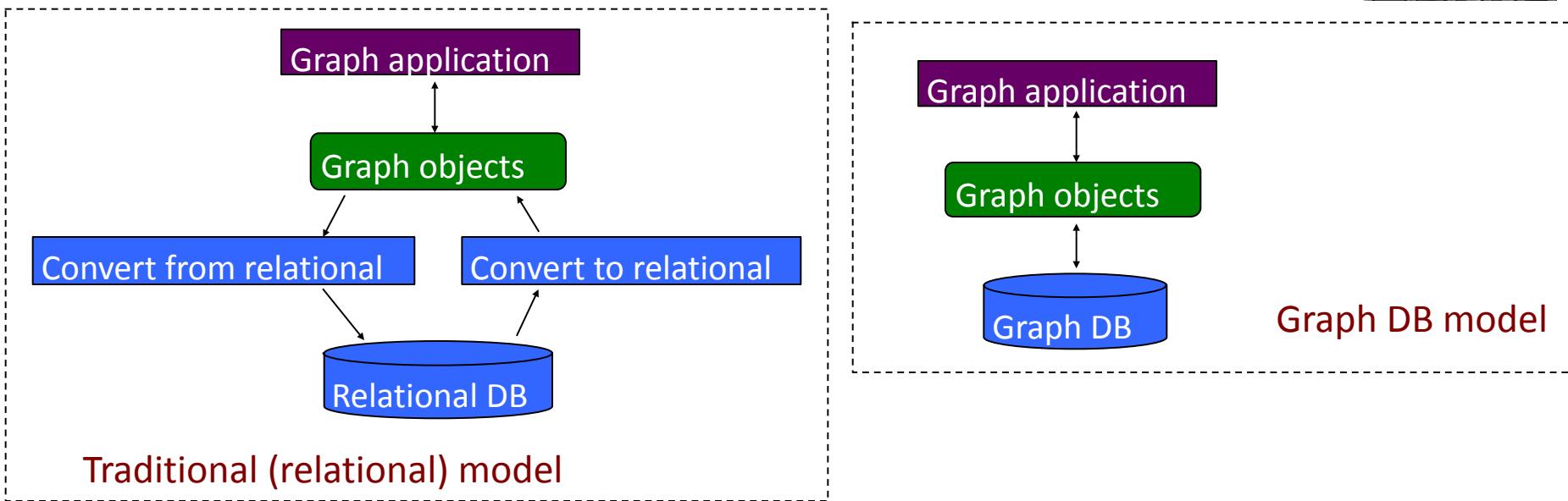
Category 5: Data Warehouse Augmentation

(1) System G currently has 4 supported graph db backends:

- (1) A relational based architecture (DB2RDF) for enterprise specific graph query based workloads.
 - (2) A HBase based architecture, with Hadoop support for graph analytic workloads.
 - (3) A Native Store which is not a full database, but is optimized for storing and retrieving graphs
 - (4) Compliant with open source Graph databases that can be accessed through TinkerPop API (e.g., Neo4j, Titan).
- System G creates API to allow Analytics, Middleware & Visualization to be swappable with different DB options.
 - Netezza implements were started but not finished yet.



Use Case 16: Code Life Cycle Improvement



- Advantages of working directly with graph DB for graph applications
 - (1) Smaller and simpler code
 - (2) Flexible schema → easy schema evolution
 - (3) Code is easier and faster to write, debug and manage
 - (4) Code and Data is easier to transfer and maintain

Graph Query in DB2RDF



- Novel mechanisms to store sparse data in RDBMS (SIGMOD 2013 paper)
- 4-6 times better than other open source graph stores on 4 benchmarks
- Only store to handle all queries correctly
- Scalability tested up to 2.3 billion triples on real datasets (Uniprot). Worst performance was ~180 s for 2 queries, 2 queries took ~100 s, 14/31 query performance was under a second, others were under ~10 s. Worst performing queries had huge result sets (~100M)



Rational Jazz workload, single query

Systems	Average (secs)	Standard Dev.	Geometric mean	Min	Max	#Queries
Jena TDB	5.7	13.8	.06	.001	56.6	29/29
Virtuoso	3.9	8.6	.26	.003	39.7	25/29
DB2RDF	1.0	1.6	.27	.004	5.0	29/29

LUBM benchmark workload, single query

Systems	Average (secs)	Standard Dev.	Geometric mean	Min	Max	#Queries
Jena TDB	35.1	51.1	.29	.002	149.8	12/12
Sesame	164.7	328.9	.37	.001	658.0	4/12
Virtuoso	16.8	28.4	.19	.001	83.3	12/12
RDF-3X	2.8	4.7	.08	.001	12.9	11/12
DB2RDF	6.9	12.8	.14	.003	33.7	12/12

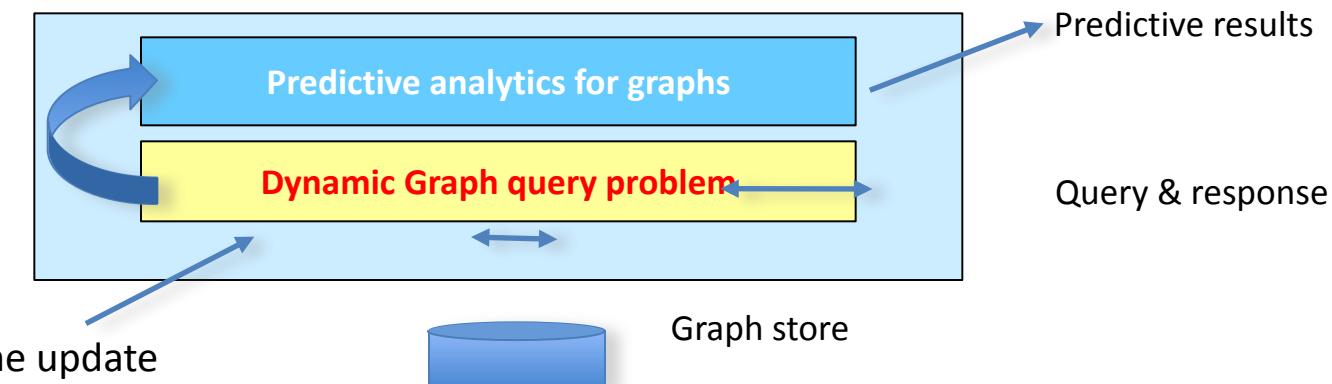
Use Case 17: Smart Navigation Utilizing Real-time Road Information

Goal: Enable unprecedented level of accuracy in **traffic scheduling** (for a fleet of transportation vehicles) and navigation of individual cars utilizing the **dynamic real-time information** of changing road condition and predictive analysis on the data

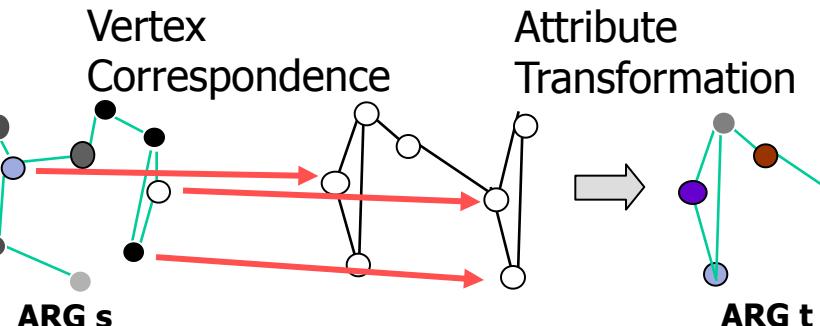
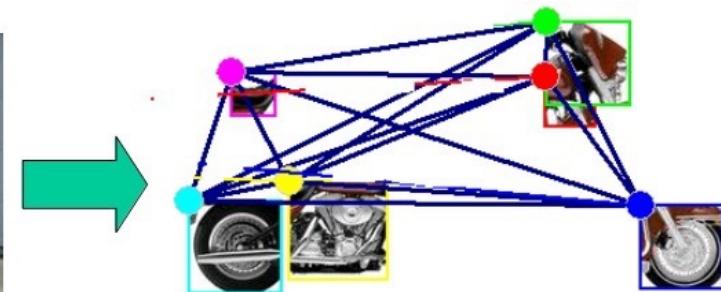
- Dynamic graph algorithms implemented in System G provide **highly efficient graph query computation** (e.g. shortest path computation) on time-varying graphs (order of magnitudes improvement over existing solutions)
- High-throughput **real-time predictive analytics** on graph makes it possible to estimate the future traffic condition on the route to make sure that the decision taken now is optimal overall



Our approach: Querying over dynamic graph + predictive analytics on graph properties



Use Case 18: Graph Analysis for Image and Video Analysis



Use Case 19: Graph Matching for Genomic Medicine

- Ongoing discussions

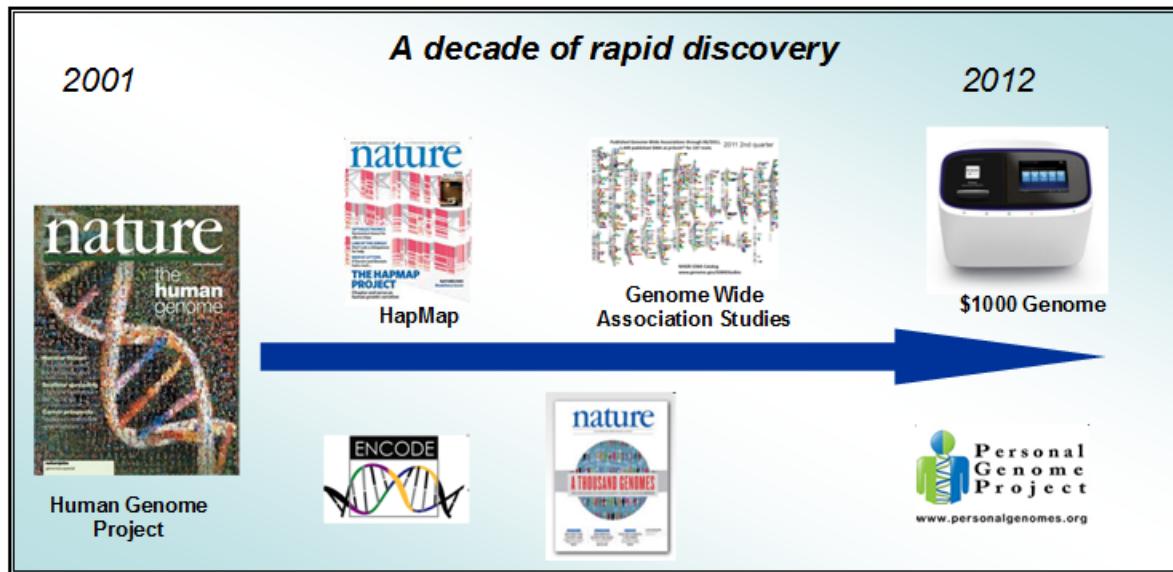
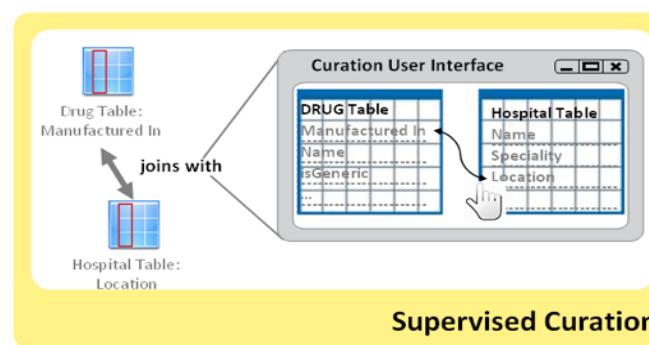
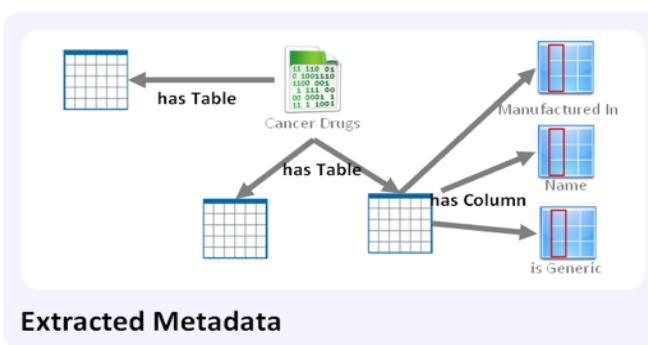
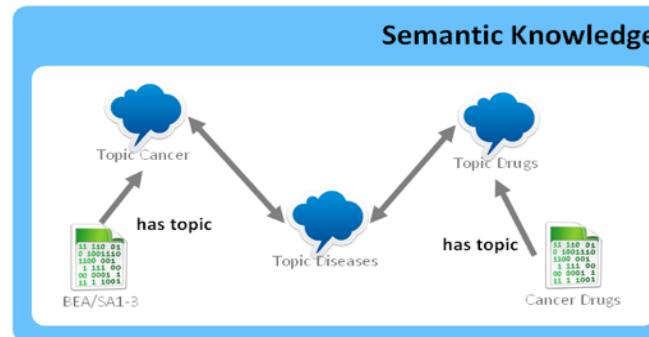
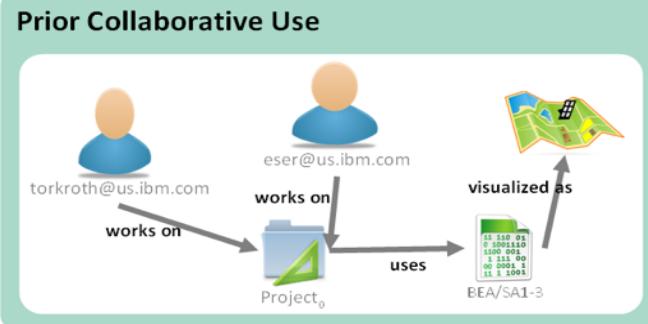


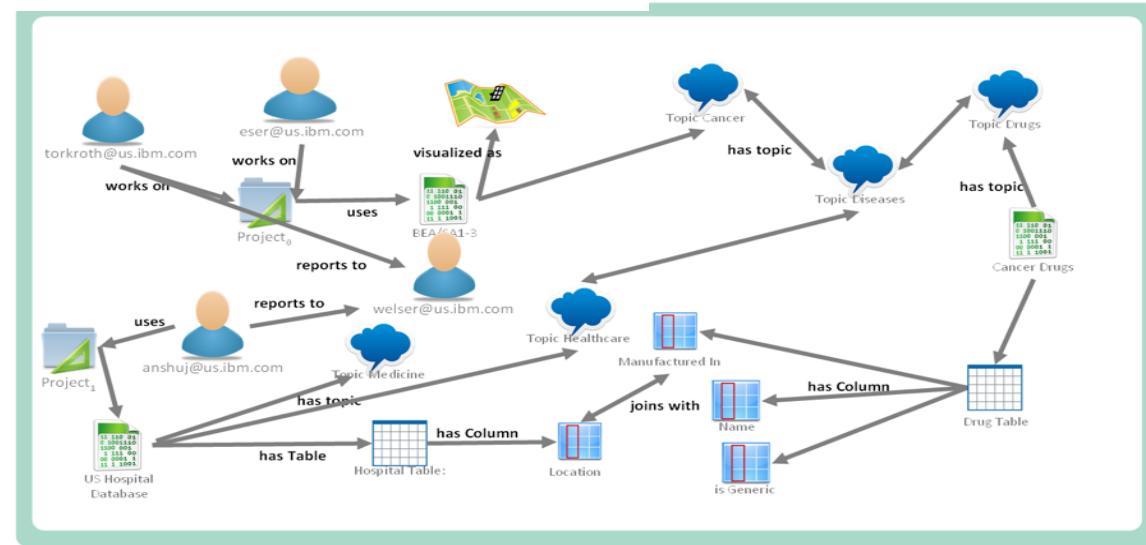
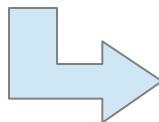
Figure 1: Since the Human Genome Project, various projects have started to reveal the mysteries of genomes and the \$1000 Genome is almost reality.

Use Case 20: Data Curation for Enterprise Data Management 

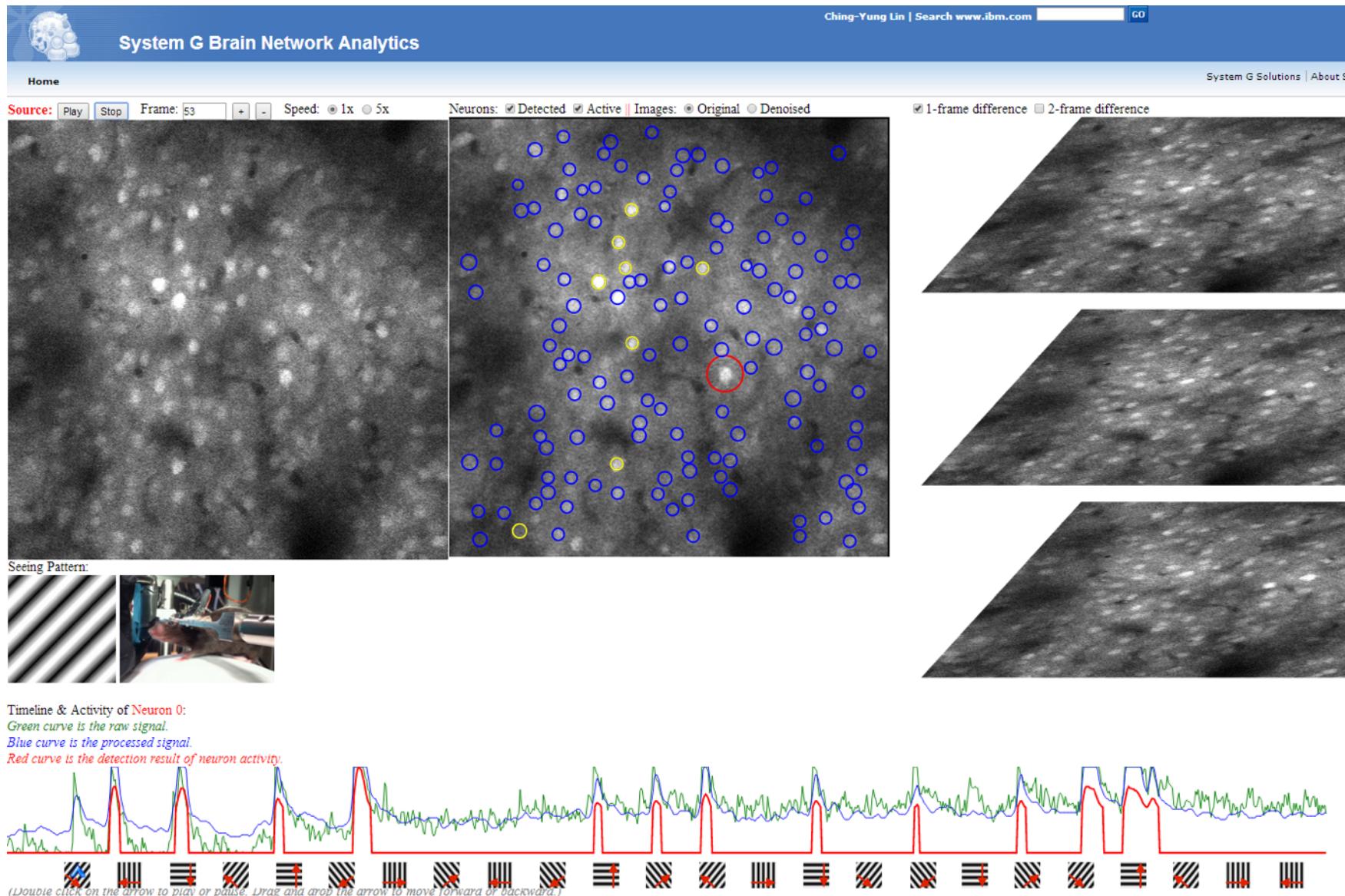


Extracted Metadata

Supervised Curation

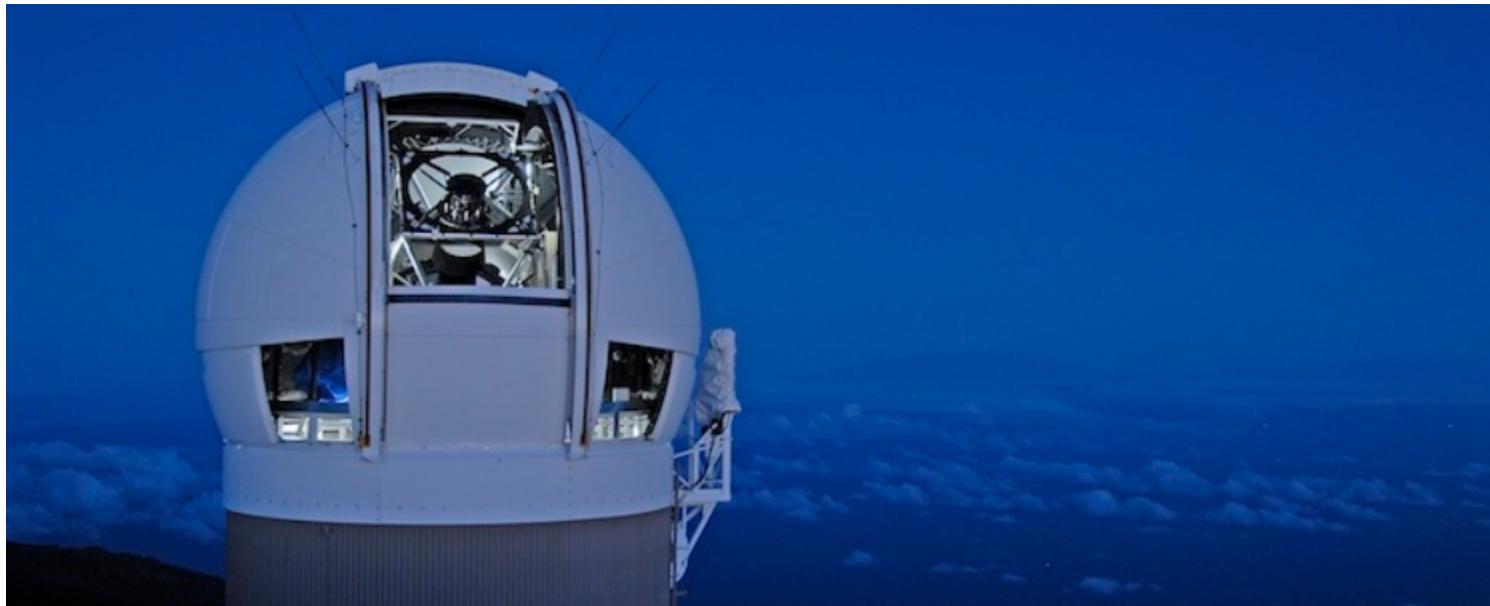


Use Case 21: Understanding Brain Network



Use Case 22: Planet Security

- Big Data on Large-Scale Sky Monitoring



Photograph by Rob Ratkowski for the PS1SC

Dangers from space Learn about the threat to Earth from asteroids & comets and how the Pan-STARRS project is designed to help detect these NEOs. Learn more... 	1,400,000,000 pixels Pan-STARRS has the world's largest digital cameras. Read about them here... 	The PS1 Prototype PS1 goes operational and begins science mission PS1 Science Consortium formed... PS1SC Blog PS1 image gallery 
--	---	--

Questions?

Sign List



Name (Last, First)	UNI	Department	Degree (yr)	Prior School or Company
--------------------	-----	------------	-------------	-------------------------