

# Introduction to Big Data Analytics

## Chap 1 – Overview of Big Data Analytics

Yanwei Fu

School of Data Science, Fudan University



# Something about the course

- ① 10% class, 30% about the project, 60% final exam.
- ② TA: Yibin Sheng (wechat: syb3181, syb3181@gmail.com, 14210240021@fudan.edu.cn)
- ③ Office hours: Wednesday afternoon 4:15-6:00pm, Zibin building N210
- ④ Homepage:  
[http://yanweifu.github.io/courses/Data\\_analytics/index.html](http://yanweifu.github.io/courses/Data_analytics/index.html)



- ① Every three weeks one component: linux\_shell, OS, Spark, GPU computing, etc.
- ② We will also do some tutorial.
- ③ may invite friends from industry to talk about their commercial systems for Big data analysis;
- ④ Pre-required course: nope.



## Academic Integrity (学术诚信)

- **Academic integrity** is the moral code or ethical policy of academia. This includes values such as avoidance of cheating or plagiarism; maintenance of academic standards; honesty and rigor in research and academic publishing. ([https://en.wikipedia.org/wiki/Academic\\_integrity](https://en.wikipedia.org/wiki/Academic_integrity))
- No cheating and plagiarism,
  - How to define *Plagiarism*? We follow [ACM Policy on Plagiarism](#).
  - 抄袭和被抄袭双方的成绩都将被取消.
- 作业、报告、期末论文的署名原则：署你名字的工作必须由自己完成；允许讨论，但作业必须独立完成，并在作业中列出所有参与讨论的人。不允许其他任何形式的合作——尤其是与已经完成作业的同学“讨论”。
- 这是学术底线。



Plagiarism manifests itself in a variety of forms, including

- ① Verbatim copying, near-verbatim copying, or purposely paraphrasing portions of another author's paper;
- ② Copying elements of another author's paper, such as equations or illustrations that are not common knowledge, or copying or purposely paraphrasing sentences without citing the source;
- ③ and Verbatim copying of portions of another author's paper with citing but not clearly differentiating what text has been copied (e.g., not applying quotation marks correctly) and/or not citing the source correctly.
- ④ Self-plagiarism is a related issue. In this document we define self-plagiarism as the verbatim or near-verbatim reuse of significant portions of one's own copyrighted work without citing the original source[2]. Note that self-plagiarism does not apply to publications based on the author's own previously copyrighted work (e.g., appearing in a conference proceedings) where an explicit reference is made to the prior publication[3]



- ① Something about OS and linux shell. We need to know how to use linux do some simple preprocessing and data analysis on single machine;
- ② Something about cloud computing and Spark;
- ③ Something about other advanced Big Data Analytics.

## Enjoy

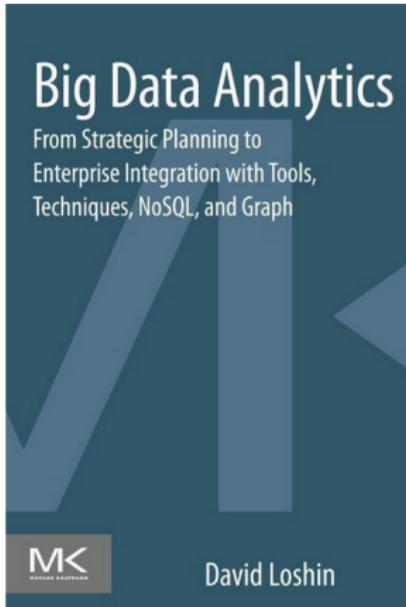
Yes, we need to know OS, linux commands and other stuffy, before really learning Spark; but not that difficulty.



- 我找不到普通家庭也需要计算机的理由。 – Ken Olsen, co-founder of DEC Corp.
- 很多人预测1996年互联网产业将大规模增长。但我的预测是1996年互联网产业由于增长过于快速，将像超新星一样爆炸后而走向崩溃。 – Robert Metcalfe, co-founder of 3Com
- 全球垃圾邮件问题将在今后两年内得到解决。 – Bill Gates, MS.
- 电视节目的流行时间不会超过半年，公众每晚会面对着一个小盒子，他们将对此感到厌倦。 – Darryl Zanuck, 20 Century Fox
- 苹果已死。 – Nathan Myhrvold, CTO of Microsoft
- 我觉得全球市场大概只需要5台计算机。 – Thomas Watson, CEO of IBM



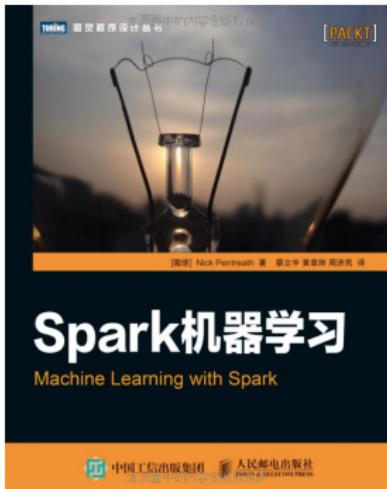
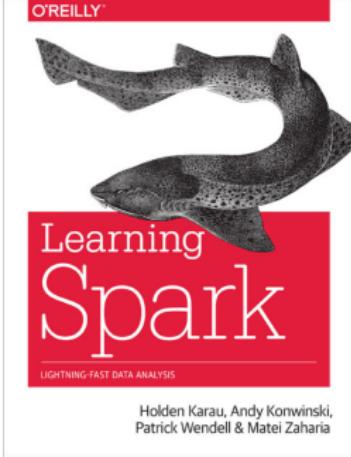
# Reading Reference for Lecture 1



- 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



# References









## Big Data



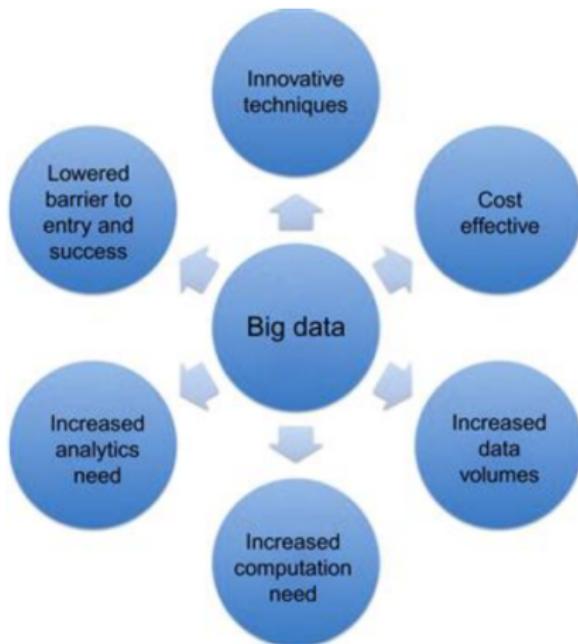
*“Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.”* -- Gartner

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



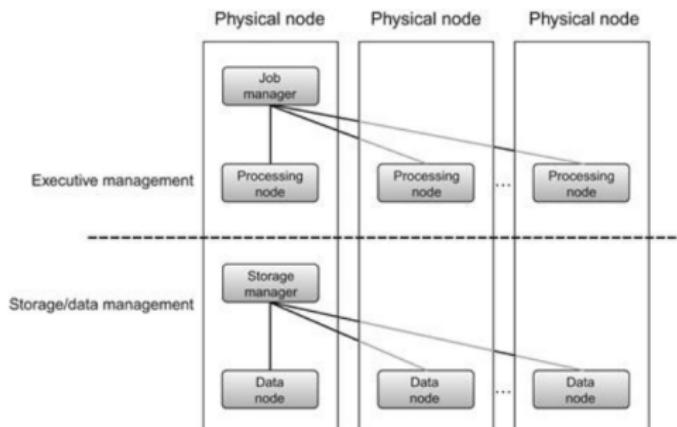
# What made Big Data needed?



"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

- 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



# Contrasting Approaches in Adopting High-Performance Capabilities

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



# 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



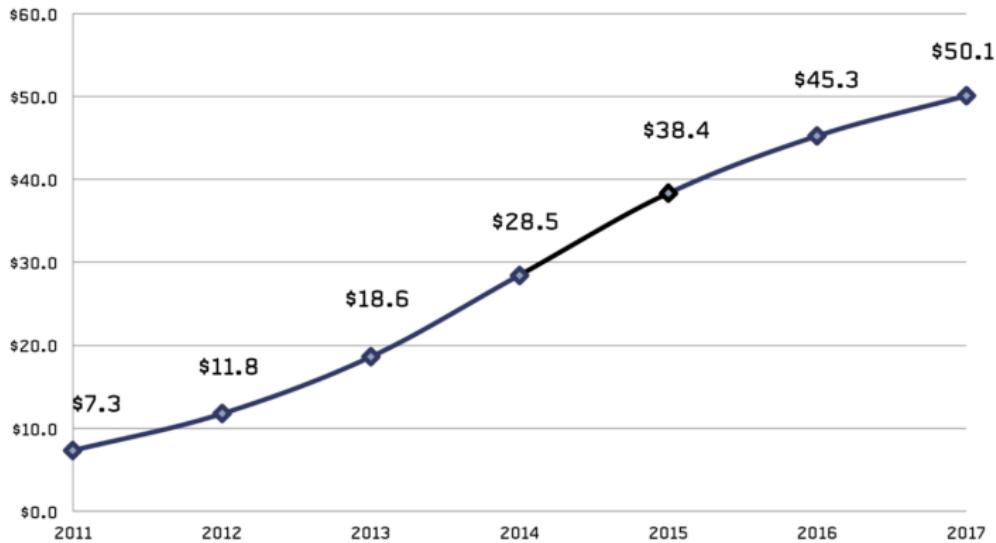
## Data Warehouse Augmentation

Integrate big data and data warehouse capabilities to increase operational efficiency

# Big Data Market



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



[http://wikibon.org/wiki/v/Big\\_Data\\_Vendor\\_Revenue\\_and\\_Market\\_Forecast\\_2013-2017](http://wikibon.org/wiki/v/Big_Data_Vendor_Revenue_and_Market_Forecast_2013-2017)



# Big Data Market further breakdown

[http://wikibon.org/wiki/v/  
Big\\_Data\\_Database\\_Revenue\\_and\\_Market\\_Forecast\\_2012-2017](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



# Sapphirine Big Data Analytics Open Source Applications

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

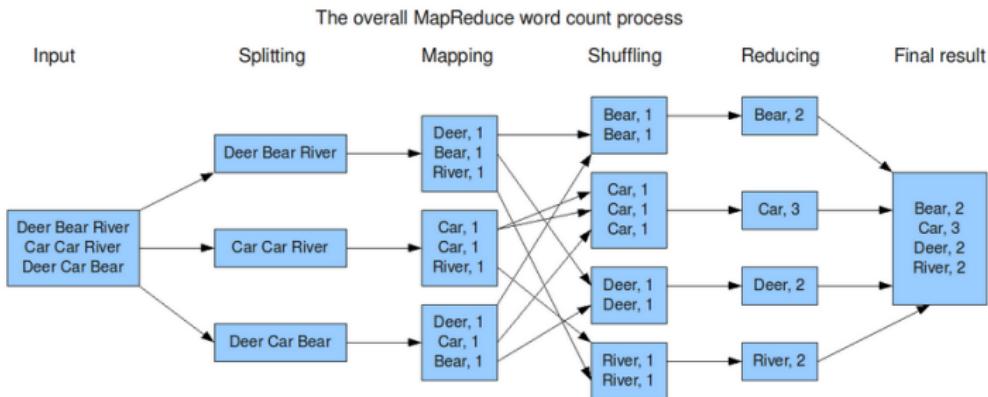


- **Dataset and Use Cases:** Welcome!!

Crowdsourcing of our collective effort!!

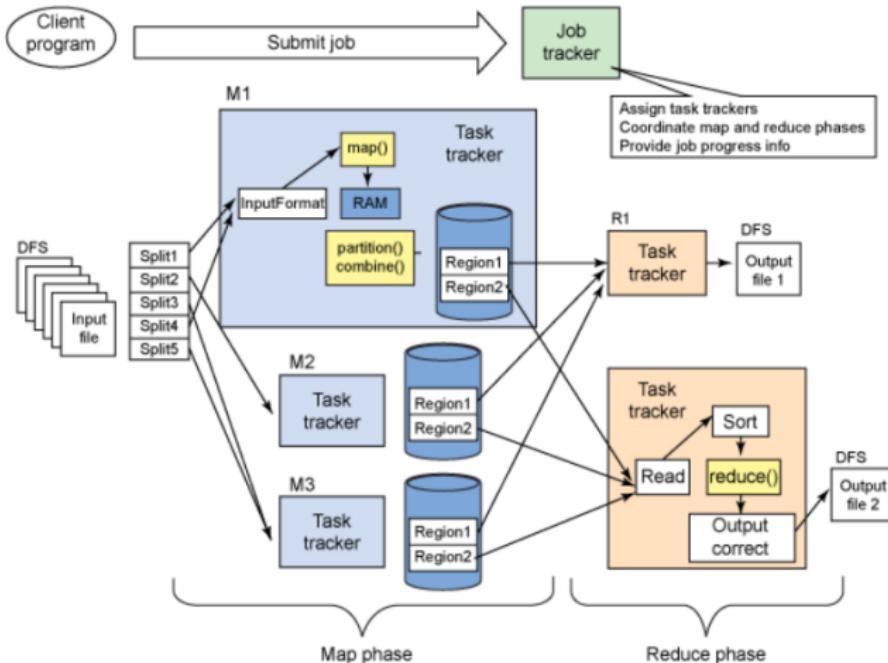


# MapReduce example



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

# MapReduce Data Flow



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



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.

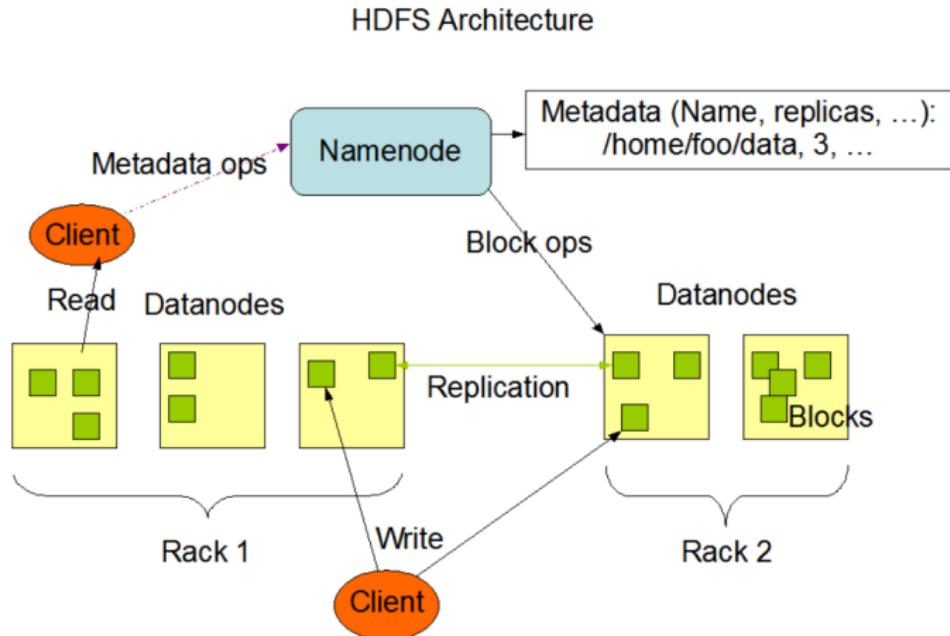
<http://hadoop.apache.org>

# 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/>

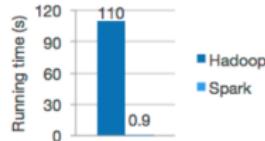
# In-Memory Computing — Apache Spark

## 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.

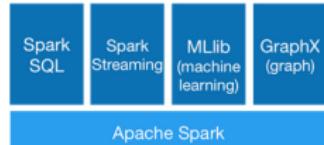
```
file = spark.textFile("hdfs://...")  
file.flatMap(lambda line: line.split())  
.map(lambda word: (word, 1))  
.reduceByKey(lambda a, b: a+b)
```

Word count in Spark's Python API

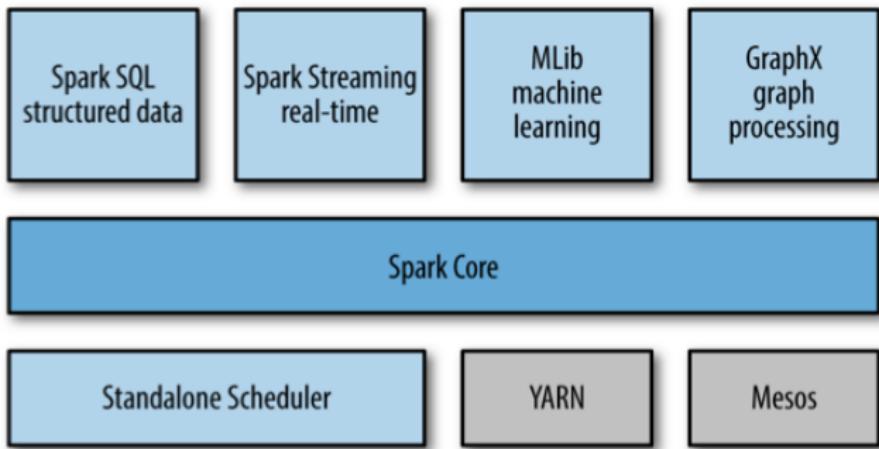
### Generality

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.



## Spark Stack



## Spark Core

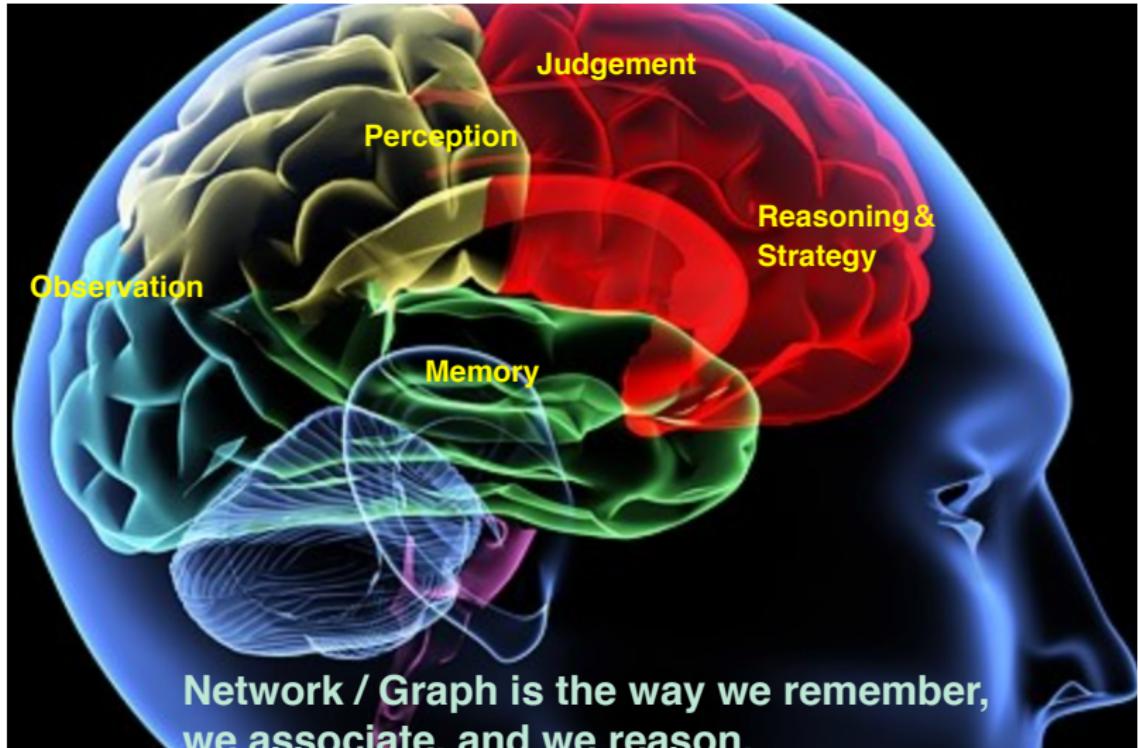
Basic functionality of Spark, including components for:

- Task Scheduling
- Memory Management
- Fault Recovery
- Interacting with Storage Systems
- and more

Home to the API that defines resilient distributed datasets (RDDs) - Spark's main programming abstraction.

RDD represents a collection of items distributed across many compute nodes that can be manipulated in parallel.





**Network / Graph is the way we remember,  
we associate, and we reason**

<http://spark.apache.org/docs/latest/quick-start.html>



Now is the Appendix.



# Linked Big Data — IBM System G

The page features a top navigation bar with tabs: Home, Overview, Toolkits, Solutions, Cloud, Documents, and Resource. Below the navigation is a large world map background. Overlaid on the map are several small portrait photos of people connected by thin white lines, forming a network. In the center, there is a large, semi-transparent graphic showing a complex network of nodes and connections. To the right of this graphic, the text "Graph Analytics" is displayed in large, bold, white letters, with the subtitle "Linked data analysis for intelligence" in a smaller font below it. At the bottom left, there is another network visualization with orange and blue nodes and connecting lines. To the right of this is a table titled "The Graph 500 List" for November 2013, listing the top 10 systems. The table includes columns for Rank, System, Institution, Location, Rating, and Entries.

Rank	System	Institution	Location	Rating	Entries
1	DOE/NERSC-8	Lawrence Livermore National Laboratory	CA, USA	17.063	40
2	DOE/OSU/Argonne Blue Waters	University of Tennessee, Knoxville; Argonne National Laboratory	TN, IL, USA	16.052	78,652
3	Blue Gene/P	IBM Research - Almaden	CA, USA	15.946	24,036
4	DOE/OSU/Argonne Blue Waters	University of Tennessee, Knoxville; Argonne National Laboratory	IL, TN, USA	15.844	38
5	Blue Gene/P	IBM Research - Almaden	CA, USA	15.742	32,088
6	DOE/OSU/Argonne Blue Waters	University of Tennessee, Knoxville; Argonne National Laboratory	IL, TN, USA	15.641	32,088
7	Blue Gene/P	IBM Research - Almaden	CA, USA	15.539	32,088
8	DOE/OSU/Argonne Blue Waters	University of Tennessee, Knoxville; Argonne National Laboratory	IL, TN, USA	15.438	32,088
9	Blue Gene/P	IBM Research - Almaden	CA, USA	15.337	32,088
10	DOE/OSU/Argonne Blue Waters	University of Tennessee, Knoxville; Argonne National Laboratory	IL, TN, USA	15.235	32,088

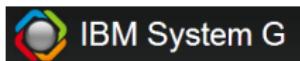


# System G — Graph Computing for Machine Intelligence



*"IBM System G": a brand name approved by HQ – April 2014.*

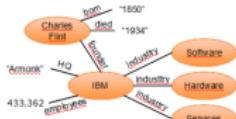
*Based on 30+ graph-related projects; 150+ papers; ~40 patents; ~10 best paper awards; ~\$25M Research funding*



*Connecting the Dots and Reasoning Big Data*

<http://systemg.research.ibm.com>

## Graph Database



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

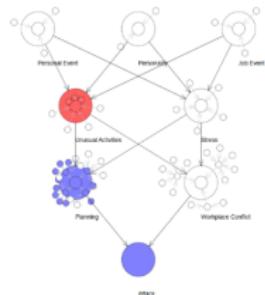
Memory

## Graph Analytics



Relationship,  
Perception &  
Contextual Analysis

## Graphical Models



Machine Reasoning &  
Deep Learning

# System G Tools — Building Blocks for Cognitive Solutions

- **Graph Middleware:**
    - Parallel Prog. Lib.
    - Power Optimization
    - GPU Optimization
  - **Graph Analytics:**
    - Topological Analysis
    - Matching and Search
    - Path and Flow
  - **Spatiotemporal Analytics:**
    - Spatiotemporal Mining
    - Spatiotemporal Indexing
  - **Graph Database:**
    - Native Store
    - GBase
  - **Graph Visualization:**
    - Multivariate Graph
    - Dynamic Graph
    - Big Graph
  - **Machine Learning Classifiers:**
    - Deep Learning Tools
    - Visual and Text Sentiment Tools
    - Anomaly Detection Tools
  - **Mobile Cognition:**
    - iOS Cognition Tools
    - Robot Cognition Tools
  - **Machine Reasoning:**
    - Bayesian Networks
    - Game Theory Tools
    - Multimodal Analysis Platform
- 
- 1 Graph Database Technology
- 2 Network Analysis Technology
- 3 Machine Judgment Topology
- 4 Machine Reasoning Technology
- IBM System G



# IBM System G Graph Tools

## Visualization



## Analytics

### Graph Computing Tools APIs and Query Language Support



## Middleware



## Database

- System G Assets
- Open Source
- Hardware

### Graph Data Interface (TinkerPop)



Other Graph Store

GBase

HBase

Performance Driven System G Native Graph Store

### File System (Linux FS, Hadoop HDFS, etc.)

## Hardware

Server  
(Linux & OS X)

Cluster  
(CPU, CPU+GPU)

Cloud

Mobile  
(iOS)

Mainframe  
(System Z & Power)

Super Computer

- Enterprise Social Solution v3.9.0 (SmallBlue)
- Insider Threat Solution v1.5.0 (ADAMS)
- Social Media Solution v2.0.0 (SMISC)
- Surveillance Insight v1.0.1
- Graph Tools v1.5.0
- Mobile Vision v0.5
- Mobile Security Solution
- Non-Performing Loans Solution
- Anti-Money Laundering Solution
- Investment Advisory Solution



# 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

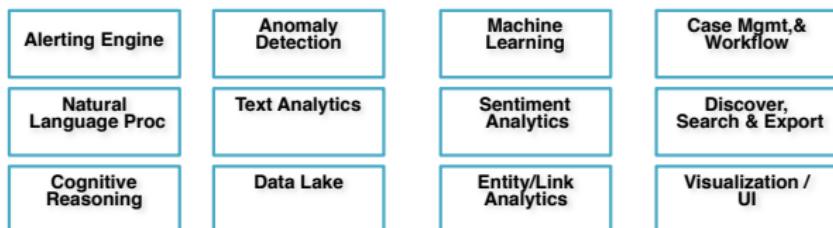


## Data Warehouse Augmentation

Integrate big data and data warehouse capabilities to increase operational efficiency

# Financial Anomaly Detection

<i>Model validation</i>	<i>Entity process</i>	<i>Social net analysis</i>	<b>Analytic Engines</b>			<i>Predictive analysis</i>	<i>Peer group analysis</i>	<i>Sim. what-if analysis</i>
AML	Anti-fraud	Non-Performing Loans	ABC	Trade Surveillance	PRG	Control room	Employee compliance	GRC
Transaction monitoring	Internal fraud	Payment screening	Gifts & entertainment	Communication monitoring	Position disclosure	Conflict management	Employee trading	Risk assessment
Client screening	External fraud	List management	Bus dev consultants	Digital surveillance		List management	G&E	Control backtesting
CRE (list rating)	Unauthorized trading		Hiring practices	Information barrier monitoring		Research clearance	Annual attestations	Monitor testing
List management				Market abuse			Business interests	Regulatory charge

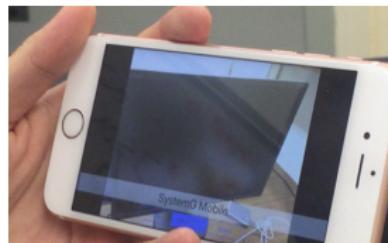


# Mobile Cognition — Enabling AI right on the Edge

- Created novel graph computing and deep learning framework on iOS devices and NAOqi robots including:
  - generic object recognition, event recognition, face recognition, visual sentiment recognition, and document recognition
  - graph database
- Prototype summer 2016 and first version release 3Q2016



IBM System G



Novel Deep Learning works that Speed Up image computation utilizing the GPUs on iOS devices: 195x or 1657x faster

	iPad Pro	iPhone 6s
Classification rate (on ~1000 classes)	~13 frames/sec	~7 frames/sec

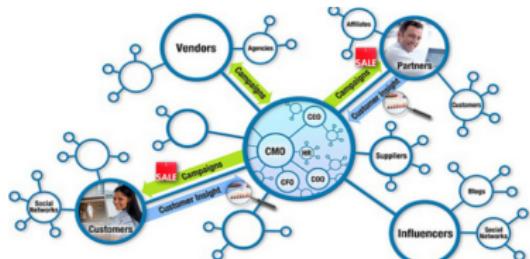


Task 1. Market Data Analysis and Investment Targets

Task 2. Advanced Dynamic 'Know Your Customer'

Task 3. Optimized Personalized Investment Strategy

Task 4. Bank-Customer Interaction Strategy



## Graph Visualizations

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

## Middleware and Database

# Social Media Solution

The screenshot displays the IBM System G Social Media Solution interface, featuring a navigation bar at the top with links like Home, Live, Trend, Multimedia, Geo, Scope, Concept, Link, Impact, Story, Person, Target, and Forensic.

The main content area is organized into several sections:

- Live Monitoring:** Shows a live stream of a person bungee jumping. Description: "Monitoring real-time tweets on keyword: #bungee". Buttons: Monitor live tweets >.
- Trend Monitoring:** Shows a sunset over water. Description: "Analyzing trend of conversations based on hashtags". Buttons: View trends >.
- Multimedia Monitoring:** Shows a purple vinyl record. Description: "Recognizing visual content and analyzing visual sentiments". Buttons: View multimedia >.
- Geo Monitoring:** Shows a person flying a kite over a city. Description: "Monitoring the places that people are sending out tweets". Buttons: View places >.
- Scope Identification:** Shows a woman looking at a computer screen. Description: "Define user-specified sets of keywords for monitoring and analytics". Buttons: Define scopes >.
- Concept Analytics:** Shows a crowd of people. Description: "Analyzing statistics of groups based on time, topics, etc.". Buttons: Concept searches >.
- Link Exploration:** Shows a network graph of nodes. Description: "Visualizing relationships, discussion sequences and graphs". Buttons: View relationships >.
- Impact Prediction:** Shows a lamp. Description: "Analyzing conversations and predicting their impact to business". Buttons: View impacts >.
- Story Detection:** Shows a woman sitting at a desk. Description: "Detecting live developing stories on social media and their evolution". Buttons: View stories >.
- Person Analytics:** Shows a person's face. Description: "Analyzing a person's personality, trustworthiness, etc.". Buttons: View person >.
- Target Discovery:** Shows a dart hitting a bullseye. Description: "Inspecting potential users for bot detection, marketing, or influencing". Buttons: Inspect targets >.
- Forensic Analytics:** Shows a firework display. Description: "Analyzing retweet sequences and displaying anomalies". Buttons: View anomalies >.

Below these sections, there is a search bar: "Select a Channel: , or a temporary channel with Keywords: , , ".

Annotations with green arrows point from specific sections to detailed examples:

- An arrow points from the "Story Detection" section to a screenshot of a news article about US midterm foreign policy.
- An arrow points from the "Person Analytics" section to a screenshot of a tweet from Hurricane Ida.
- An arrow points from the "Forensic Analytics" section to a map of the Middle East with highlighted regions and a timeline chart below it.

On the right side, there is a small thumbnail image of a news broadcast and some status information:

Sun Asia 23 02:00:04 +0000 2013 GMT  
RT @engr\_helphate: Plus - Large cover of KH heading to undefined location (3 of 3) <https://t.co/4TzJPlmsC>  
Retweet Count: 1  
Followers Count: 207  
Tweet Sender Location: N/A  
Automatic Tagging: densely\_crowded, horizontal\_text, excellent\_fish, annoying\_reflections, busy\_bridge  
Visual Sentiment Detection: Negative



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



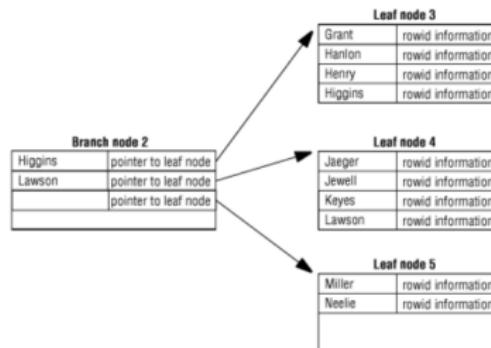
# Key Value Store

## 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(*key<sub>1</sub>, key<sub>2</sub>, ..., key<sub>N</sub>*), 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

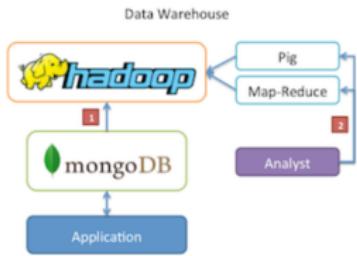


IBM Informix K-V store



Example Application: Spatio-Temporal Analys

# Document Store



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

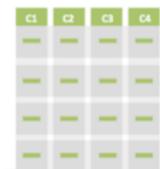
```
db.users.insert ( ← collection
{ ← field: value
  name: "sue", ← field: value
  age: 26, ← field: value
  status: "A" ← field: value
}
} } document
```

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**

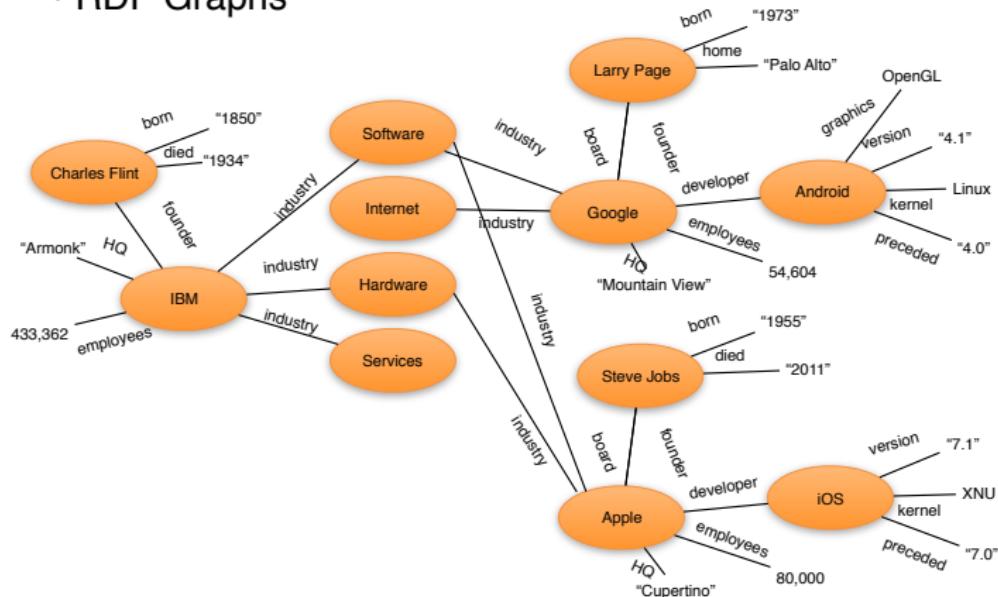
Highly-structured table organization  
with rigidly-defined data formats and  
record structure.



**Document data model**

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

- Property Graphs
- RDF Graphs



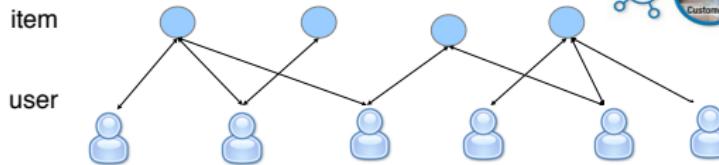
# Big Data Analytics Example Use Cases

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*



Enhancing:



## Graph Visualizations



# 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

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

200,000 people's consulting project & earning da

The screenshot shows the SmallBlue Suite interface with a search bar for "Healthcare". Below it, a network diagram displays a cluster of nodes representing IBMers, with connections indicating relationships. On the left, a sidebar lists six specific users with their profiles and titles:

1. Patricia (Patte) Okie  
Global Business Services  
Associate Partner, Healthcare Integration  
Other Consultant
2. Michael Hildebrandt  
IBM Research  
Life Sciences Business Development  
Category Sales  
Link: [View Profile](#)
3. Zolt (Z.T.) Vajnaik  
Global Business Services  
GBS Partner, Healthcare and Public Health --  
Practice Administrator is Shirley Carkner  
Other Consultant  
Link: [View Profile](#)
4. Susie L. COUGAN Rivers  
Global Business Services  
Healthcare Knowledge Navigator  
Market Insights  
Link: [View Profile](#)
5. Paul (P.C.) Van Asperen  
Global Business Services
6. Fan (F.Z.) Li  
Global Business Services

On the right, there are sections for "Dynamic networks of 400,000+ IBMers:" and a sidebar with navigation links for "Shortest Paths", "Centralities", and "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
- 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.

The screenshot shows the SmallBlue Suite interface with a network visualization titled "Healthcare". The graph consists of numerous small blue nodes representing individuals, connected by a dense web of thin grey lines representing relationships. Several nodes are larger and feature small profile pictures, indicating they are "Influencers". A red arrow points from one such influencer node to the text: "Influencers are the one with high 'Betweenness' and 'Degree' value". Another red arrow points to a specific cluster of nodes labeled "A cluster of XYZ experts". A third red arrow points to a node labeled "Independent experts on healthcare". The interface includes a top navigation bar with tabs like Home, Find, Reach, Net, Ego, Admin, and a search bar for "subject/keywords". On the right, there's a "Display Set" panel with options for "Show node i" (Names or Statistics), "Show node j" (Business or Name), "Show people" (Min or Hide link), and a "Find Person" section with a slider for "Highlight Network". A yellow banner at the bottom states: "SmallBlue analyzes underlining dynamic network structure in enterprise".

Independent experts on healthcare

A cluster of XYZ experts

Influencers are the one with high 'Betweenness' 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



# 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) Country: Division: Advanced search

healthcare all all Find Experts

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 4 degrees 5 degrees 6 degrees 7 degrees 8 degrees 9 degrees 10 degrees

SmallBlue Net Click to see results as a Social Network

As on 9/29/2009, SmallBlue is indexing/infering 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.

1. Patricia (Patrice) Orlita Global Business Services Associate Partner, Healthcare Integration Other Consultant Ask: MARTHA E. (Martha) GIBSON > Amy D. (AmyD) Basile

2. Michael Hahnberger IBM Research Life Sciences Business Development Category Sales Ask: Ravi B. Komuru > Vanessa L. Johnson

3. Todd (T.M.) Kalyanuk Global Business Services GBS Partner, Healthcare and Public Health -- Global Administrator Shirley Canfield Other Consultant Ask: Chong Sheng Li > Robert (R.) Terek

4. Susan E. (SUSAN) Rivers Global Business Services Healthcare Knowledge Manager Market Insights Ask: MARTHA E. (Martha) GIBSON

5. M.C. (Mack) Eiffingham IBM Sales & Distribution, Public Sector Client Technical Advisor Ask: Ari Fishkind > Julie A. Reid

6. Paul (P.E.) Van Akenen 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) Caceaza Global Business Services Healthcare Transformation Services Ask: MARTHA E. (Martha) GIBSON > Alan J. (ALAN) Leudier

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

Settings 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.

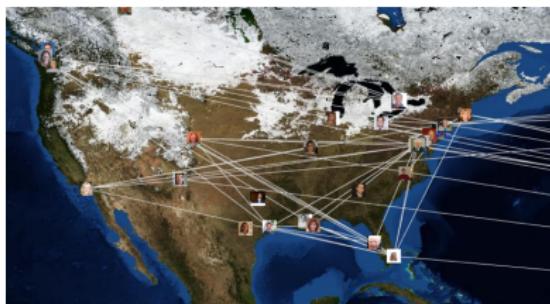
Click a name to see their profile (SmallBlue Reach)



# 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?

The screenshot shows the SmallBlue Suite interface with the following annotations:

- His official job role, title contact info**: Points to the "Email" field containing "tom.cocozza@us.ibm.com".
- His self-described expertise**: Points to the "Formal organization group" section.
- My various paths to Tom. SmallBlue can show the paths to any colleagues up to 6-degree away**: Points to the "Recommended Path" and "Alternative Paths" sections.
- His public communities**: Points to the "CommunityMap" section.
- The public interest groups he is in**: Points to the "BlueGroups" section.
- The public bookmark tags he is using**: Points to the "Social bookmark tags" section.

**Email or Name**  Reach Persons [F]

**SmallBlue Suite**

Home | Find | **Reach** | Net | Ego | Admin

About SmallBlue | Tools | Help | Download | Terms of Use | Project Info

**Your social paths to reach [Thomas (Tom) Cocozza]**

**Recommended Path**

Ching-Yen Lin → MARTHA E. (MARTHA) GIBSON → Alan J. (ALAN) Lander → Thomas (Tom) Cocozza

**Alternative Paths**

Click to see social network of these people

- Ching-Yen Lin → JAMES (JAMES) Stachak → Martin P. Adams → Thomas (Tom) Cocozza
- Ching-Yen Lin → Vicki Sofiles-Pilat → Hayes R. Adams → Thomas (Tom) Cocozza
- Ching-Yen Lin → MARTHA E. (MARTHA) GIBSON → Susan E. (SUSAN) Eusebio → Thomas (Tom) Cocozza
- Ching-Yen Lin → Vicki Sofiles-Pilat → Hayes R. Adams → Thomas (Tom) Cocozza

**Communities**

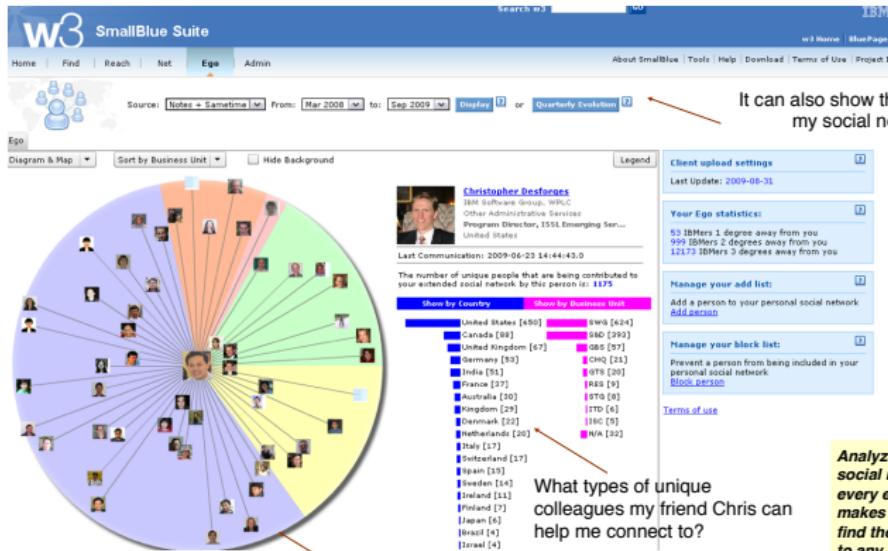
- CommunityMap
  - Industry Marketing Client Success
  - Value Based Marketing
  - Public Sector Technical Community
  - Biometric and Identity Analytics
  - Public Sector Global
  - IBM Business Leader Seminar
  - Business Value Thought Leadership
- BlueGroups
  - BLG-COM-DIV-ICBS-FSP-PM REPORTING
  - BLG-COM-PROD-HBAMG
  - BLG-COM-PROD-ICBS-FSP-PM REPORTING
  - BLG-COM-ITAS Dynamic Managers
  - BLG-COM-PROD-COMS-US-MANAGER
  - Value Based Marketing
  - ChannelMetrics
  - ISC-IBM Manager
  - ISBL-1000-2021-US-GBS-Federal
  - Investment Alpha
  - KView Portal Author-BCE-WW
  - ESTC - Announcements Broadcast
  - ESTC - Broadcast
  - ESTC - Public Broadcast
  - PrivateBiometrics
  - PublicBiometrics
  - SCAM Managers
- Social bookmark tags
  - No information
- Public postings
  - BlogCentral
    - No information

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



# Personal social network capital management

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



Evolutionary personal social network



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



## 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 Page(s) [ ]

Practitioner Portal

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

### People in your network

Network for: Un\_China-Yunnan

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 prominence



[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:

18M 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 [ID] July 09 10:50 AM Comment

RSS Feed Perfect 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 Wards  Practitioner Portal  Media Library  LX  BLX

5 of top 10 Sort by: Social Proximity | Date | Source

Network: direct Sources: All

Welcome to Graph Technologies [ID] 09 Jul 2013

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

ibm-and-smartphone-data-at-a-glance-lm-ju2013.pdf [ID] 30 Jul 2013

### Popular in the Practitioner Portal

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

- Top document searches  IBM, system, bay, signature, solutions, bob, ac, KM and KB case studies
- Top accessed content
- Top Bookmarks
- Portlet social rating information

### 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  LX  BLX

5 of top 30 Sort by: Popularity | Source

Sources: All

Leadership in a Project Team Environment  ★★★★★ PMHO eShareNet June 13, 2013 - Worldwide Project Management Conference, San Francisco, CA, USA. Improving PM Method Adaptable, Presented by Stacey Lopez and Todd Fredrickson - IBM Rational Asset Manager  ★★★★★ New2Blue - Mid-Year Review - Personal Business Commitments (Reagan National) New Business Experience 2013 Event!  ★★★★ \*

Junos Pulse for Android Smartphone  ★★★★★ Project Management Orientation  ★★★★★ Show more

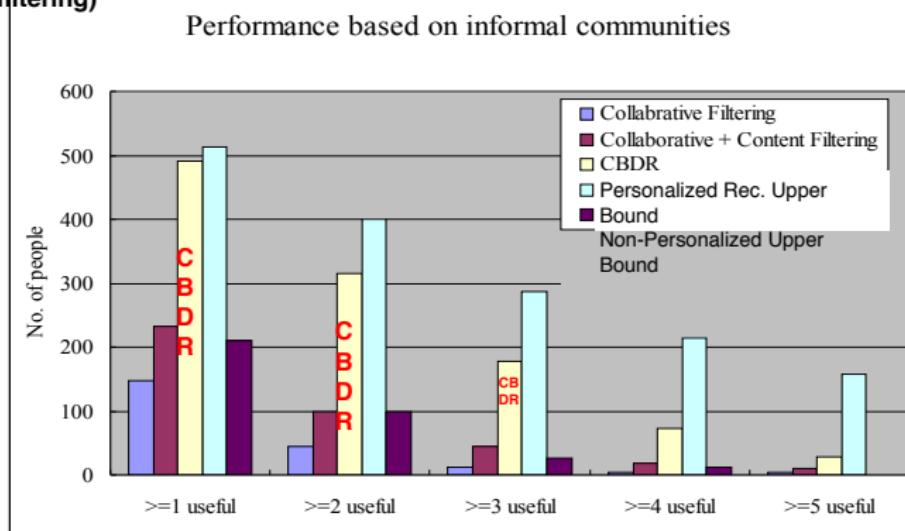
- Integrated Practitioner Portal, KnowledgeView, Media Library, Lotus Connections, and Learning@IBM and for a personalized ranking



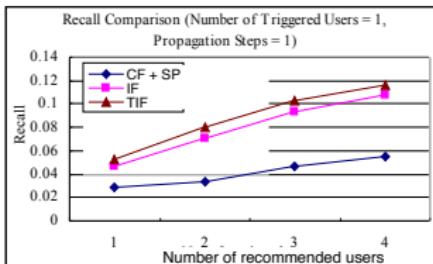
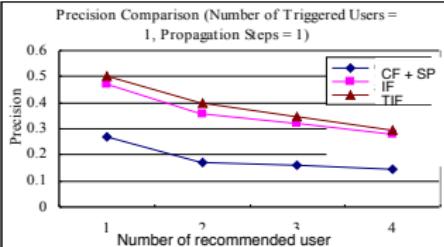
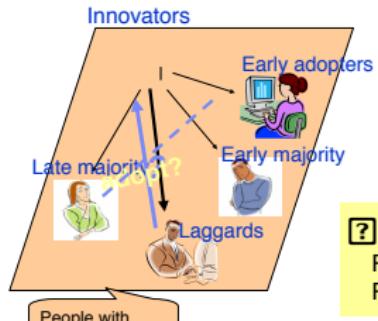
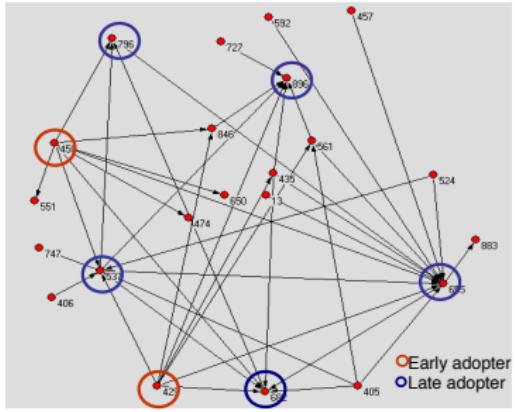
# Improving Recommendation Quality by Graph Community Analytics

- A 3<sup>rd</sup> 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



Network  
Info Flow

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

IF: Graphical Information Flow Model

TIF: Joint Topic Detection + Information Flow Model

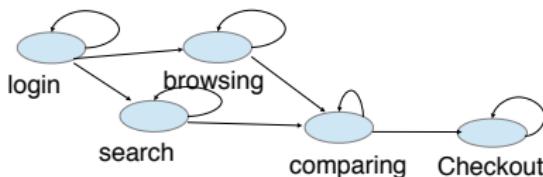
❑ Comparing to Collaborative Filtering (CF) + Similar People  
Precision: IF is 91% better, TIF is 108% better  
Recall: IF is 87% better, TIF is 113% better

# Customer Behavior Sequence Analytics

Markov  
Network

Latent  
Network

Bayesian  
Network

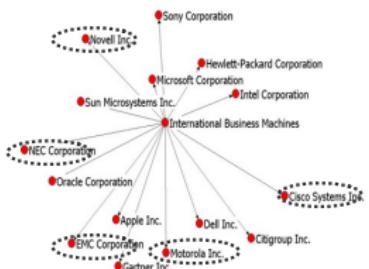


- Behavior Pattern Detection
- Help Needed Detection

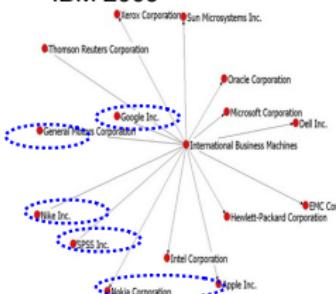
# Use Case 4: Graph Analytics for Financial Analysis

**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



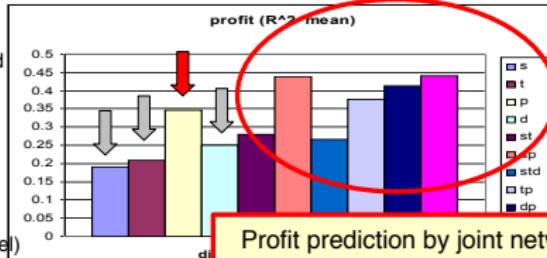
- Data Source:

- 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)

Profit prediction by joint network and financial analysis outperforms network-only by 130% and financial-only by 20%

## Use Case 5: Social Media Monitoring

Ching-Yung Lin | Search www.libnu.com

## System G SMISC Social Media Monitoring

Home | Live | Forensics

Select CIO Category(-ies): EXECOB BLADE HRTEANAN IBM SecurityAnalysis SWG WATSON or Word: egypt

IBM CIO monitoring categories

Monitoring filter

Total Tweets: 231

Positive: 33 15%  
Negative: 33 13%

**brutality Mor**  
EGYPT wants #Egyptian beauty  
e ||| Am Egypt's 12 police  
dozen Spring allege Egypt than Care  
you my Egyptian مصر said egypt last call

**Solomon Buttli** @SalomonButtli  
عذاء العذريون لـ@Solomon\_Btrli  
RT @Ulion\_King\_Bhr: [redacted] على المراق امسا ورجل ابر 10/2/2013  
#Bahrain #Egypt #KSA #UAE #News h...  
Translates: RT @Ulion\_King\_Bhr: The traitors in Bahrain Sabot attack on public utilities and security men, 2/19/2013 "Bahrain" #Egypt, "Lydia" "LKA" "UAE" "News" "h..."  
---Wed Feb 20 17:57:59 2013

**Zenza Raggi fan-club** @Zenzacub  
Private Gold 4d: Cleopatra 2 / a sect that worships ancient Egypt is attempting to bring Cleopatra back to life... http://t.co/TovDmWb  
Mon Feb 18 2013 17:57:53 2013

**مطهية** @Metehe  
RT @HesarParooz: An Egypt-ian beauty  
) • http://t.co/9f82bf3  
Mon Feb 18 2013 17:57:53 2013

**Mona Metwally** @monametwally  
عمر مصطفى مطران - د. مصطفى العقاد - ابراهيم عبد الله بن عرب  
RT @EgyBloodBank: [redacted] AB+ 01024705247 #Egypt #سرير http://t.co/5j06mtZ5.  
Translation: RT @EgyBloodBank: A

IBM CIO monitoring categories

Monitoring filter

GO STOP RESUME language: Arabic

Real-Time Translation, Loca

# 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

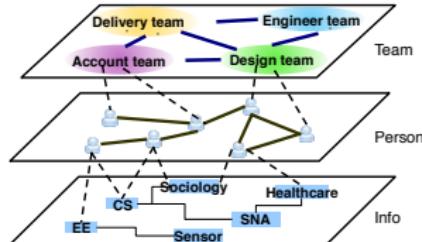
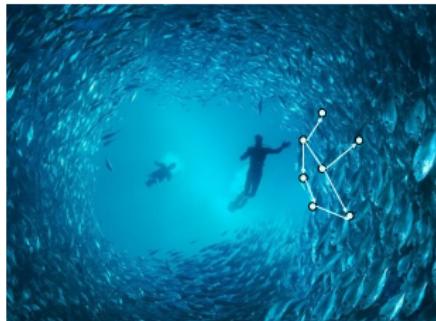
## Thrust 3. Affecting Information Dissemination:

- 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



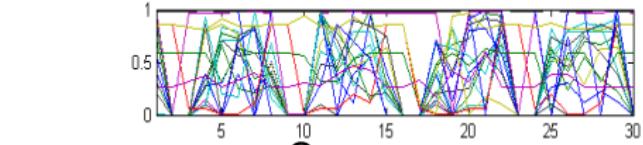
# Dynamics in Graphs

## 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)

# Dynamics of Information Graphs in Social Media

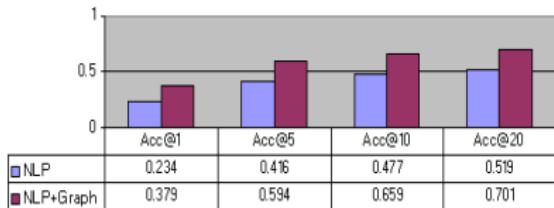
## Motivation:

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

## Approach:

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

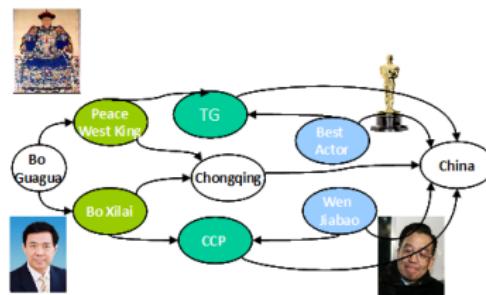
Entity morph resolution accuracy  
(ACL 2013)



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

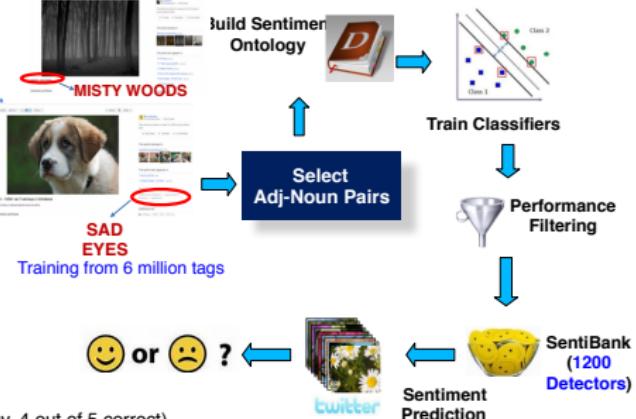


"For content to go viral, it needs to be emotional," Dan Jones, 2012

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

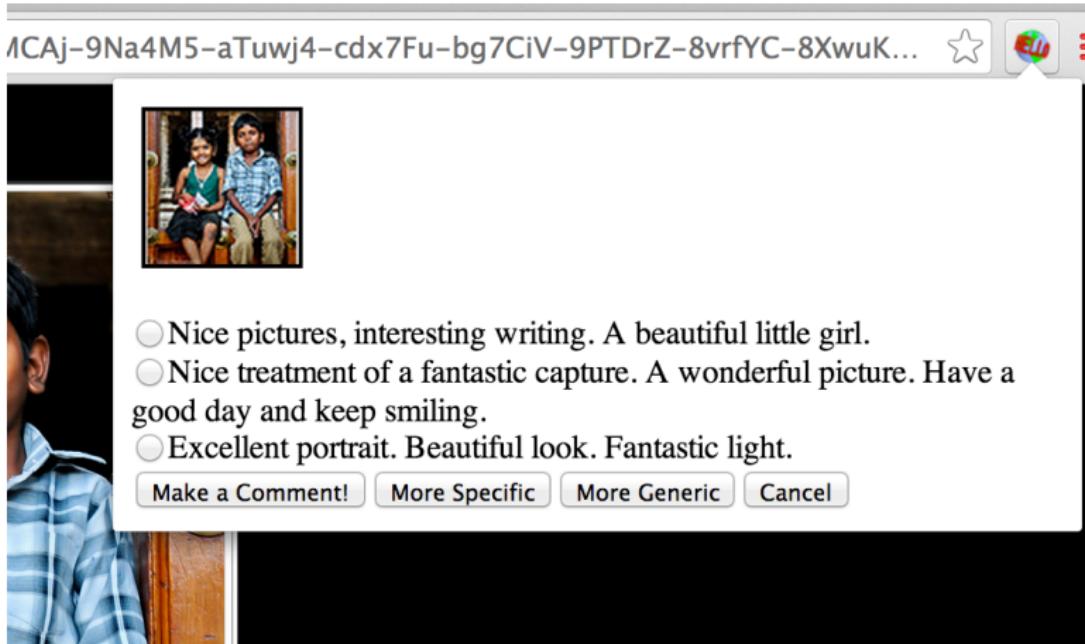
Text	0.43
Visual	0.70
T+V	0.72

# Cognitive Feeling Detection on Images

## Cognitive Feeling Detection on Images

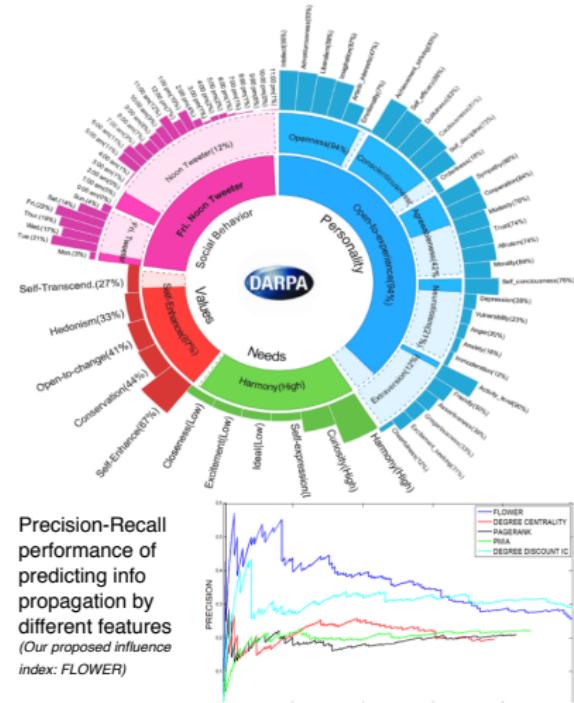


# Automatic Comments on Images

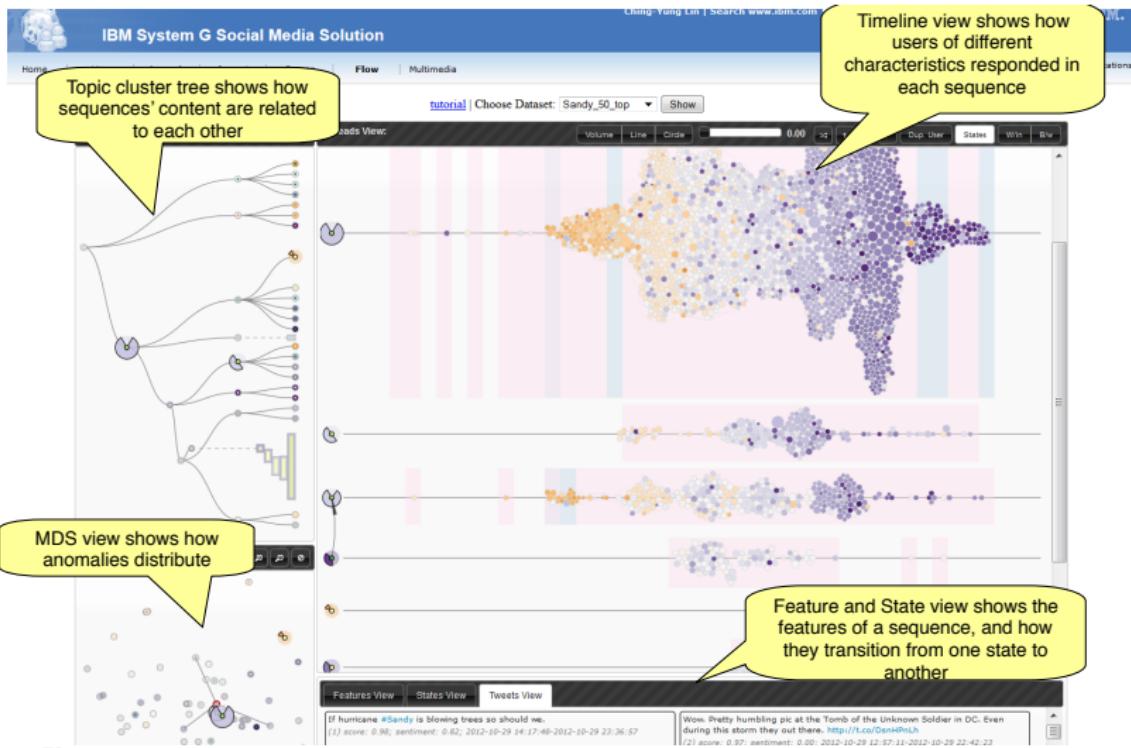


# Measuring Human Essential Traits in Social Media

- **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, 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.



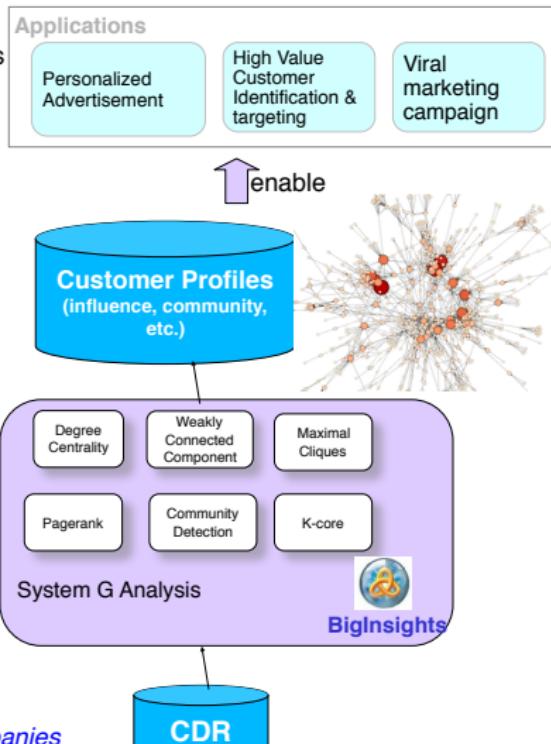
# Flow Analytics - I



# Use Case 6: Customer Social Analysis for Telco

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

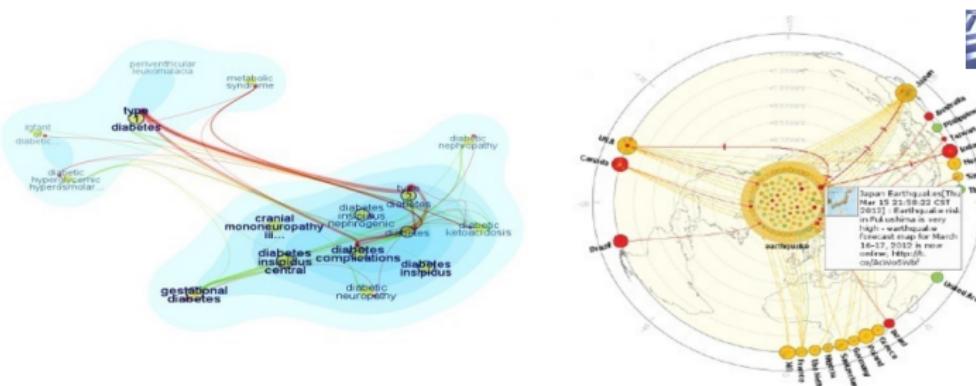
- Applications based on the extracted social profiles
  - Personalized advertisement (beyond the scope of traditional campaign in Telco)
  - High value customer identification and targeting
  - Viral marketing campaign
- Approach
  - Construct social graphs from CDRs based on {caller, callee, call time, 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



PoCs with Chinese and Indian Telecomm companies



# Category 2: Data Exploration



Enhancing:



Huge Network Visualization

Network Propagation

I2 3D Network Visualization

Geo Network Visualization

Graphical Model

Communities

Graph Search

Network Info Flow

Bayesian Networks

Centralities

Graph Query

Shortest Paths

Latent Net Inference

Ego Net Features

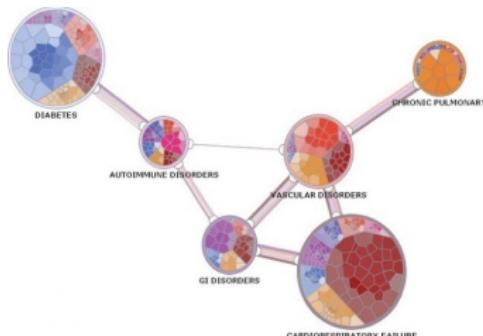
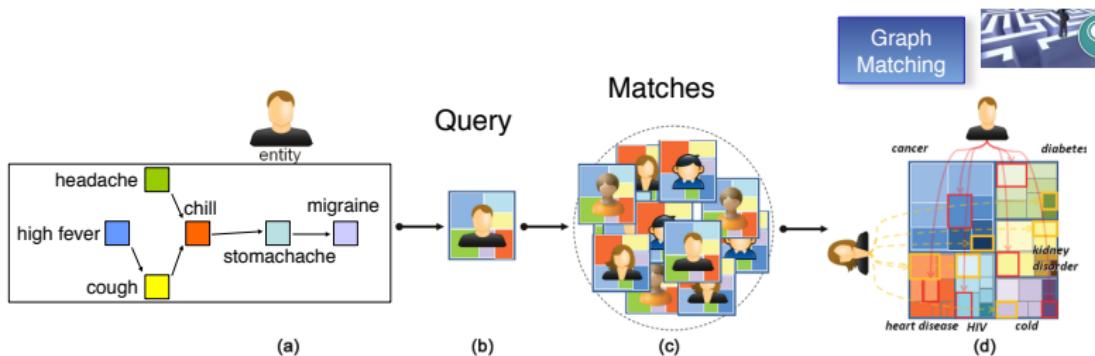
Graph Matching

Graph Sampling

Markov Networks

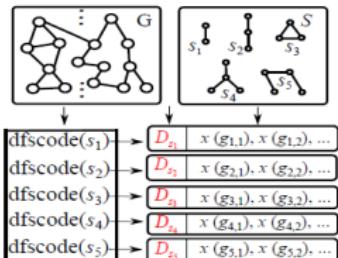
Middleware and Database

# Use Case 7: Graph Analytics and Visualization for Watson

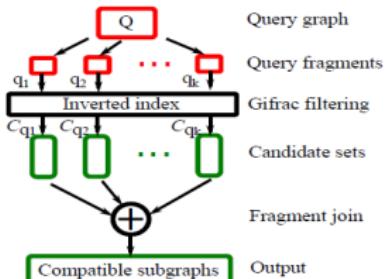


# Fast Graph Matching Algorithm

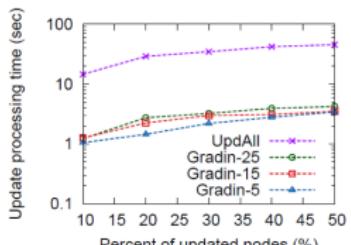
- Data: (CAIDA) 26.5K nodes and 106.8K edges
- Index construction: 13-20 times faster than the prior state-of-the-art
- Query time: close to UpdAll (upper bound) and ~8x faster than UpdNo and NaiveGrid



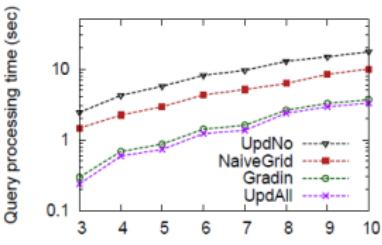
(a) Offline index building



(b) Online query processing

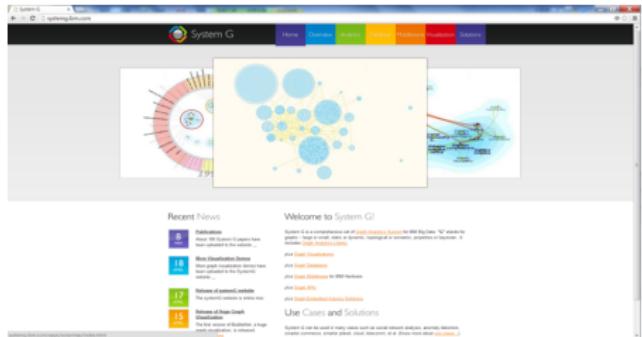


Indexing time

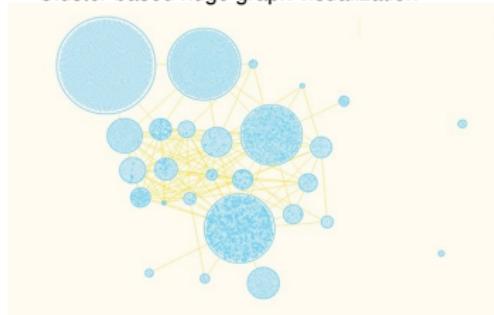


Query processing time

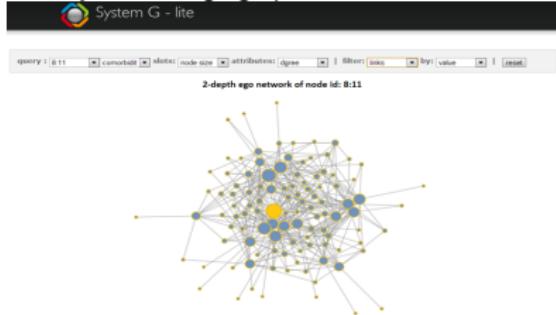
# User Case 8: Visualization for Navigation and Exploration



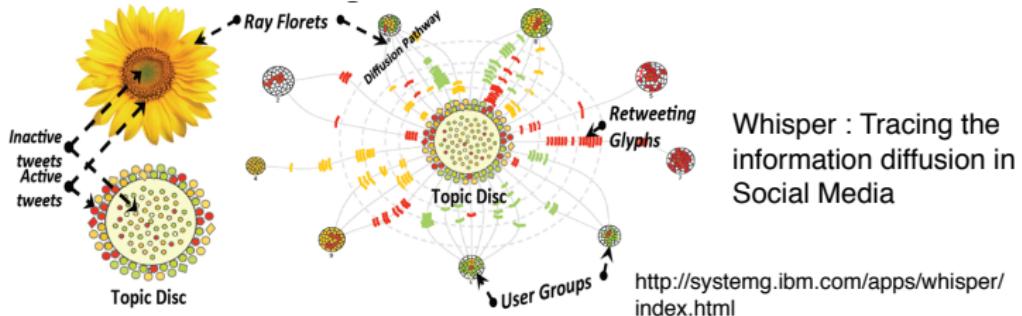
Cluster based huge graph visualization



Query based huge graph visualization

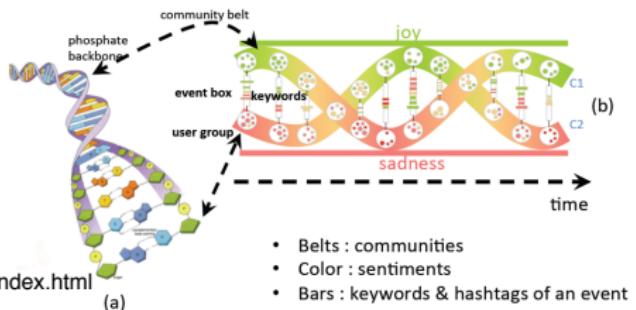


# Visualizing Information Diffusion and Divergence

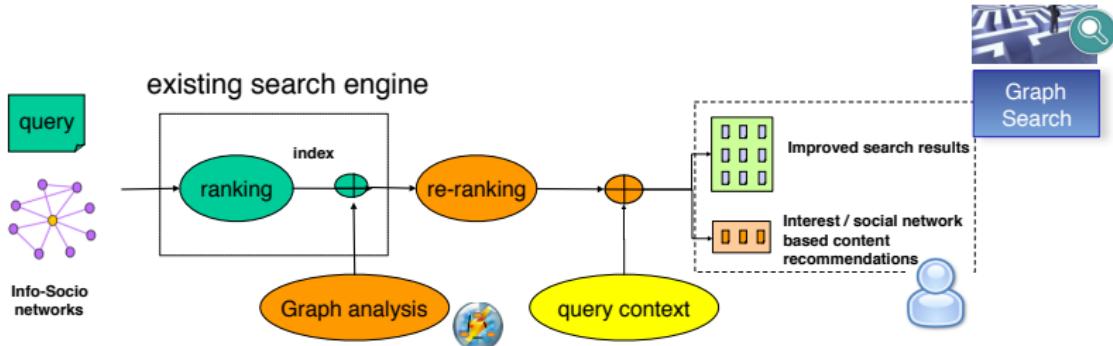


SocialHelix: Visualizaiton of Sentiment Divergence in Social Media

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



# Use Case 9: Graph Search



**Practitioner Portal** Translate this page English

< Return to starting page

**Search Results** 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**

All results Social network results Subscribe to i

18,677 results found

1 to 25 shown 1 2 3 4 5 6 7 8 9 10 ...

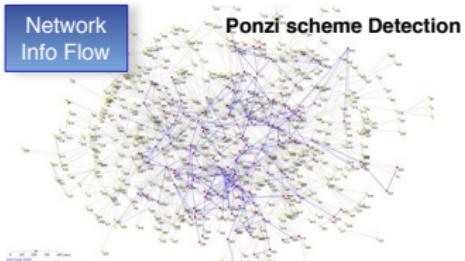
Title	Reference	Modified	Bookmarks
IBM Social Business Adoption QuickStart (U.S. English) <span style="float: right;">100 %</span>	29 Aug 2012		
IBM Social Business Accelerator (PPX) <span style="float: right;">100 %</span>			
Drive the successful launch and adoption of social business software throughout your organization with a structured engagement comprised of assessments, planning and design consultation, onsite workshops, and team- and skills-building activities.			

Sales Support Information (13) DADE@ibm.com

# Category 3: Security

**Network Info Flow**

**Ponzi scheme Detection**

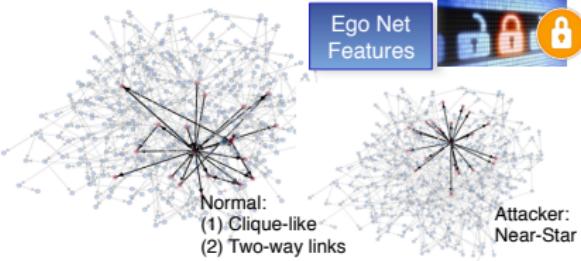


**Detecting DoS attack**



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

**Ego Net Features**



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

Attacker: Near-Star

**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.



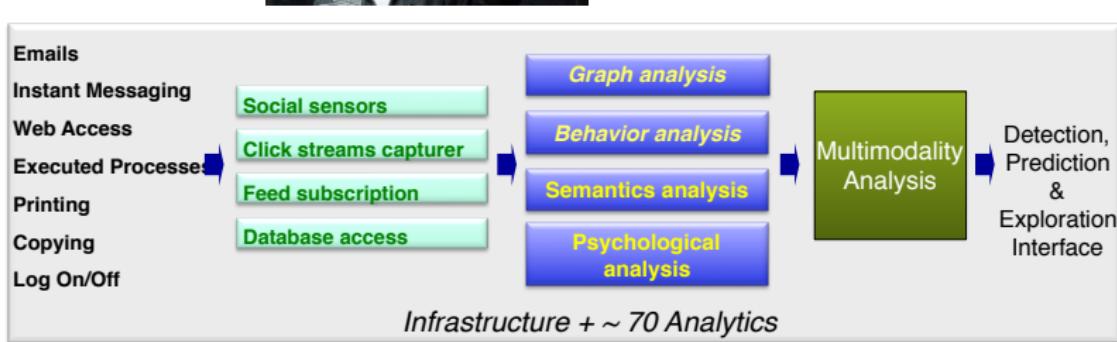
"Enterprise Information Leakage Impacted economy and jobs" Feb 2013

To Catch Worker Misconduct, Companies Hire Corporate Detectives

By ALICE CHENG  
January 10, 2013 8:25 PM

"What's emerged is a multibillion dollar detective industry"

npr Jan 10, 2013



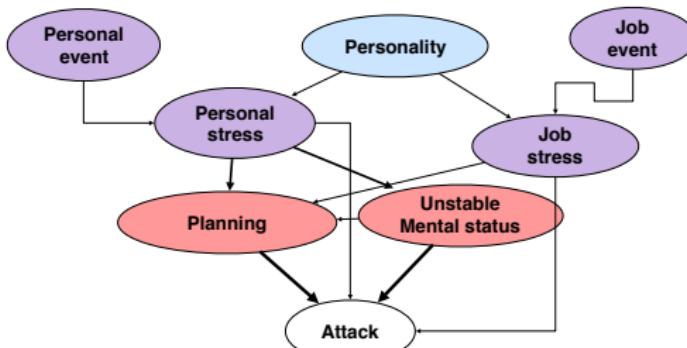
# Story – Espionage Example

## (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

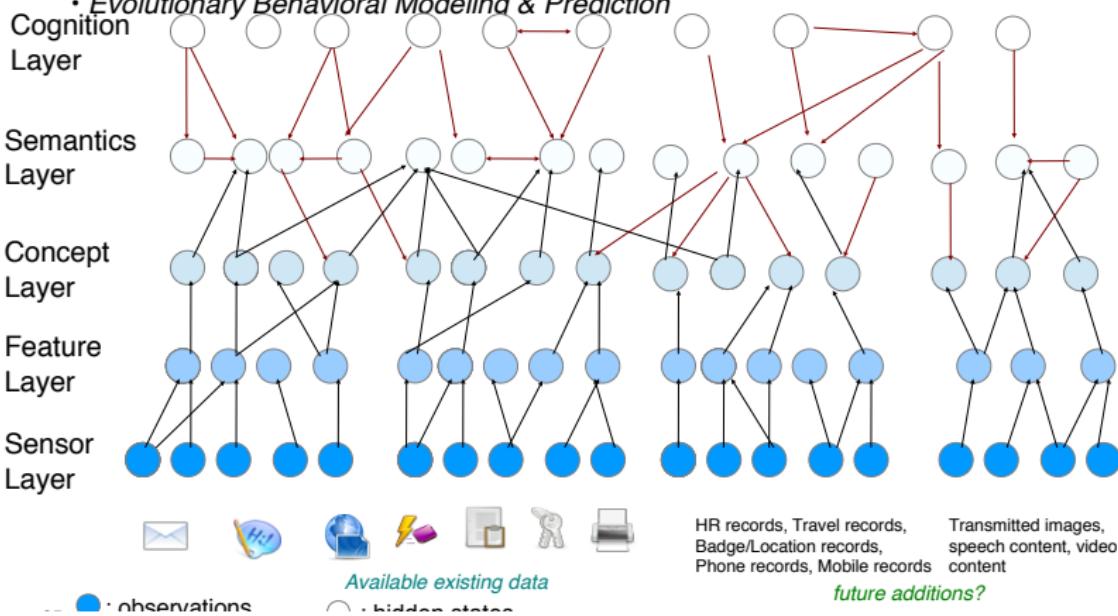
## (1) Attack:

- 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



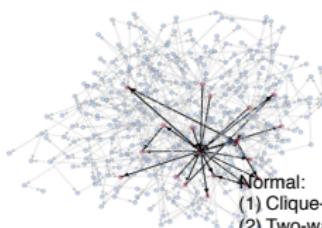
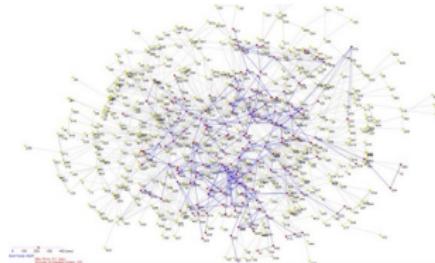
# Use Case 11: Fraud Detection for Bank

Network  
Info Flow

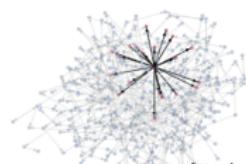
Ego Net  
Features



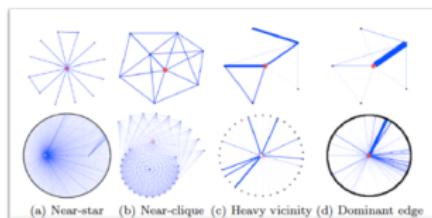
Ponzi scheme Detection



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



Attacker:  
Near-Star



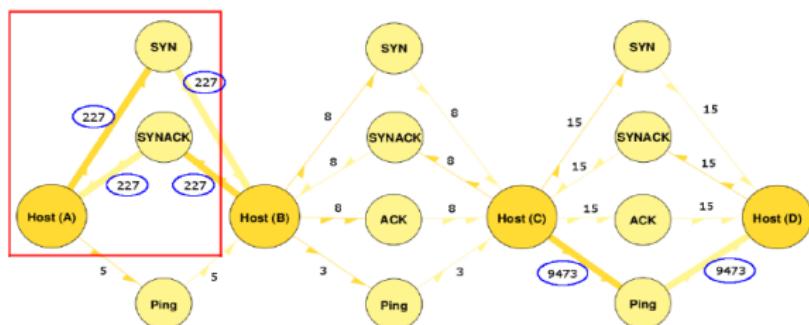
# Use Case 12: Detecting Cyber Attacks

Network  
Info Flow

Ego Net  
Features

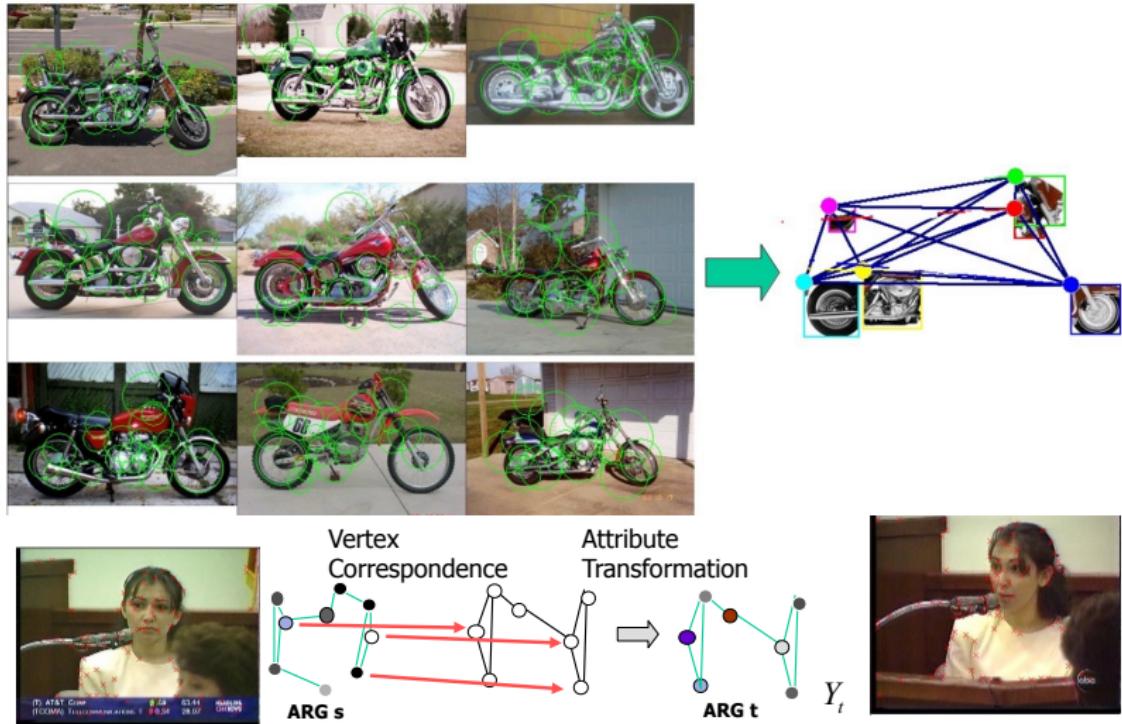


## Detecting DoS attack



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

# Use Case 13: Graph Analysis for Image and Video Analysis



# Use Case 14: 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](#)



