

Course Grading

- **Final Project:**
 - **Teamwork:** 2 - 3 students per team
 - **Proposal** (slides — short presentation in the class)
 - **Progress Presentation** (slides — short presentation in the class)
 - **Progress Report** (report)
 - **Final Report** (paper, up to 10 pages)
 - **Workshop Presentation** (Oral and Demo)
 - **Open Source Codes**
 - **Video Presentation**

5 Example 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 Examples -- Application Use Cases

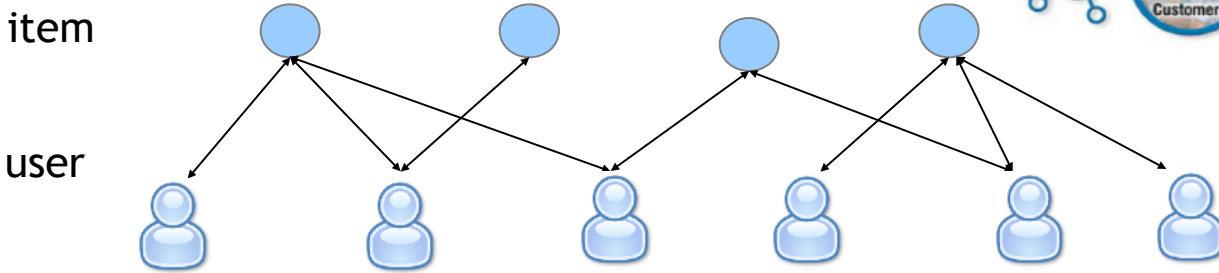
1. Expertise Location
2. Recommendation
3. Commerce
4. Financial Analysis
5. Social Media Monitoring
6. Telco Customer Analysis
7. Healthcare Analysis
8. Data Exploration and Visualization
9. Personalized Search
10. Anomaly Detection
11. Fraud Detection
12. Cybersecurity
13. Sensor Monitoring (Smarter 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

A screenshot of the Amazon.com website showing recommendations for the user 'Ching Yung Lin'. The page includes a search bar, navigation links like 'Cart' and 'Help', and a 'Find Gifts' button. The main content area displays three recommended books: 'Spikes [Reprint]' by Fred Rieke, 'Spiking Neuron Models' by Wulfram Gerstner, and 'Methods In Neuronal Modeling - 2nd Edition' by Christof Koch. On the left, there's a sidebar for 'Your Favorites' and 'Featured Stores'.



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)

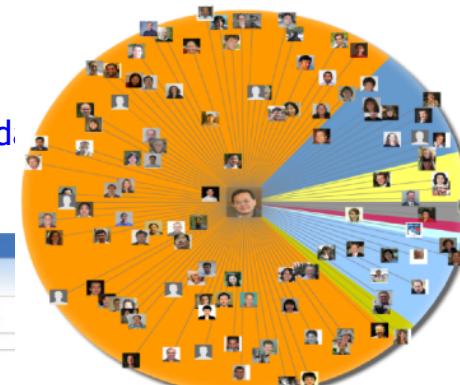
1,000,000+ Learning clicks; 14M KnowledgeView, SalesOne, ..., access d...

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

200,000 people's consulting project & earning d...

The screenshot shows the SmallBlue Suite interface with a search bar for 'subject keywords' set to 'healthcare'. Below the search bar, there are dropdown menus for 'Country' (all) and 'Division' (Advanced search). A 'Find Expert' button is visible. To the right, there is a network visualization titled 'SmallBlue Net' with the subtext 'Click to see results as a Social Network'. On the left, a list of six search results is displayed:

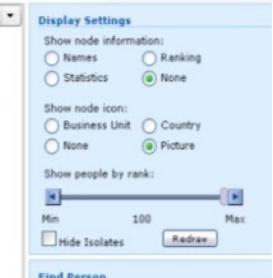
1. Patricia (Patti) Okita: Global Business Services Associate Partner, Healthcare Integration Other Consultant. Ask: MARTHA E. (Martha) GIBSON > Amy D. (AMY) Berk
2. Michael Hehenberger: IBM Research Life Sciences Business Development Category Sales. Ask: Ravi B. Konuru > Vanessa L.
3. Todd (T.H.) Kalyniuk: Global Business Services GBS Partner, Healthcare and Public Health -- Practice Administrator is Shirley Carkner Other Consultant. Ask: Chung Sheng Li > Robert (R.) Tarok
4. Susan E. (SUSAN) Rivers: Global Business Services Healthcare Knowledge Manager Market Insights. Ask: MARTHA E. (Martha) GIBSON
5. M.C. (Mark) Effingham: IBM Global Distribution Public Sector
6. Paul (P.E.) Van Aggelen: Global Business Services



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
- APQC (WW leader in Knowledge Practice) April 2013:

"The Industry Leader and Best Practice in Expertise Location"

Use Case 2: Personalized 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, Ching-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

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 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

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 [W3] ★★★★★

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 [S] ★★★★★

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

Junos Pulse for Android Smartphone [S] ★★★★★

Project Management Orientation [W3] ★★★★★

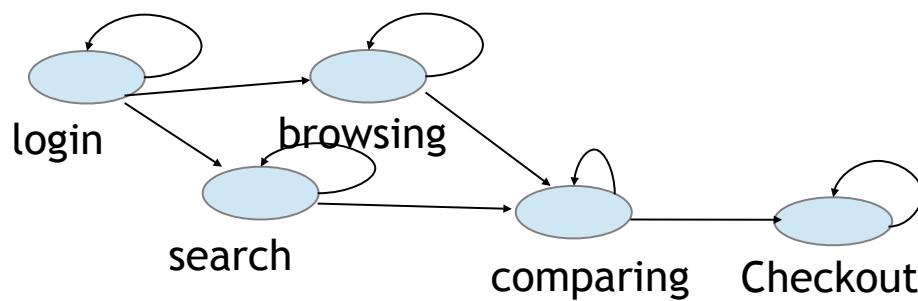
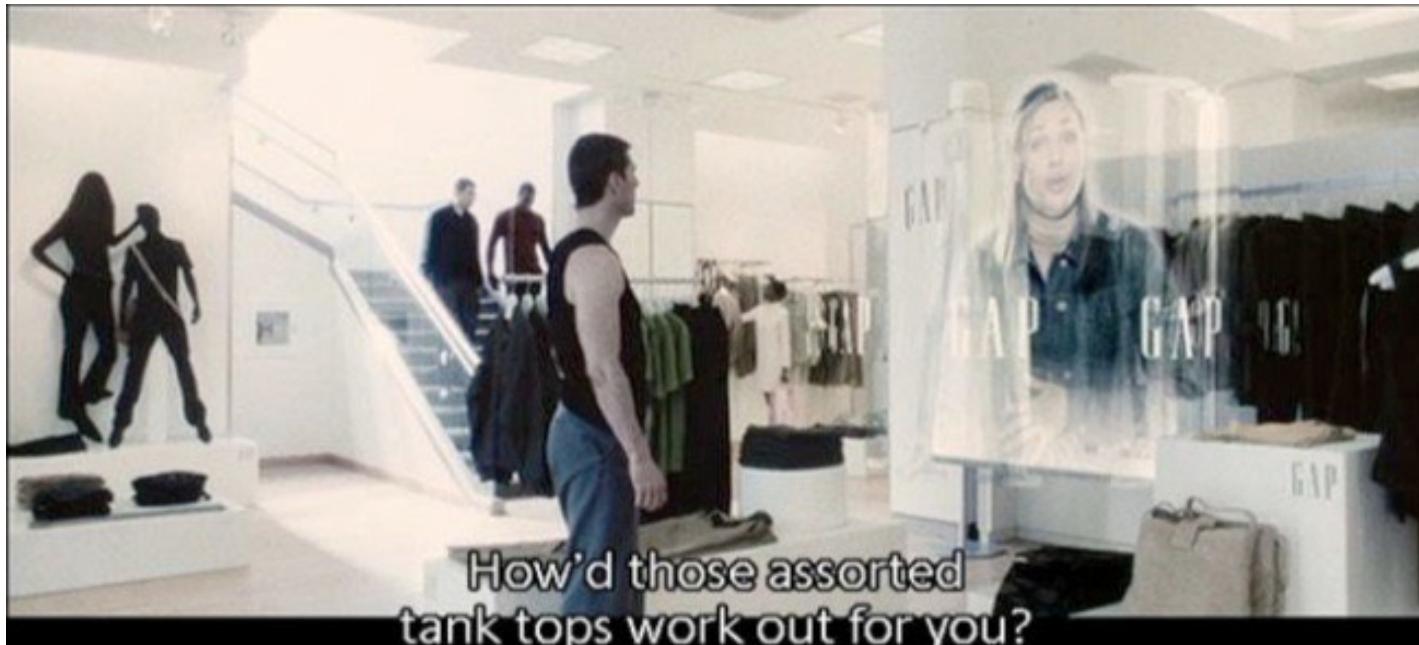
Show more

Use Case 3: Customer Behavior Sequence Analytics

Markov
Network

Latent
Network

Bayesian
Network

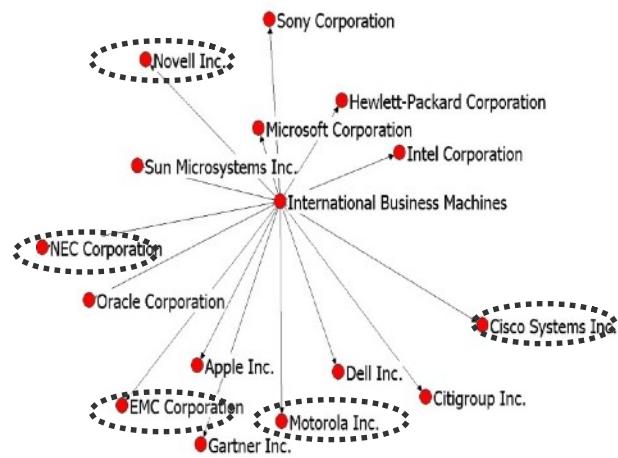


- Behavior Pattern Detection
- Help Needed Detection

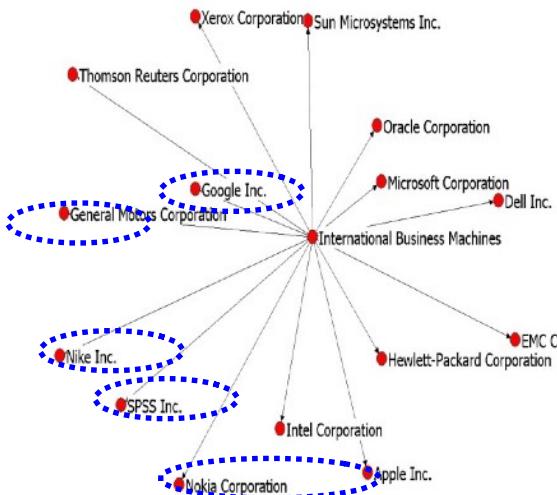
Use Case 4: Graph Analytics for Financial Analysis

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



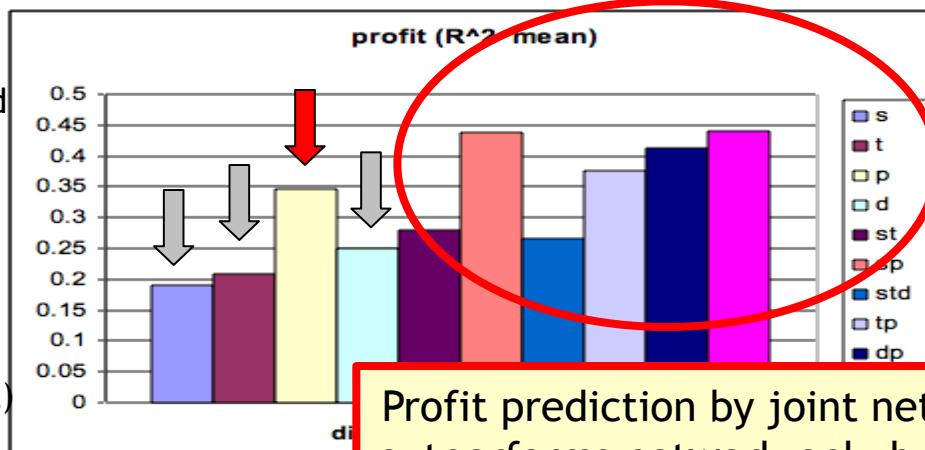
- Data Source:

- 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

Profit prediction by joint network and financial analysis outperforms network-only by 130% and financial-only by 33%.

Use Case 5: Social Media Monitoring

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

Select CIO Category(-ies): EXECDB BLADE HRTEANT IBM SecurityAnalysis SWG WATSON or Word: Egypt GO STOP RESUME language: Arabic

monitoring categories

Monitoring filter

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

EGYPT wearing @RawyaRageh beauty **brutality** Mor
 e ||| Am Egypt's 12 police hijab Er
 dozen allege port Egypt than Cairo
 you my Egyptian said egypt lady call

Saloom Butilla @SaloomButilla
 إنكاء المتنبئين الغرفة في
 RT @Lion_King_Bhr: إنكاء المتنبئين الغرفة في
 19/2/2013 #Bahrain #Egypt #Syria #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 @Zenzadub
 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 @SH_QalamSara
 مفترقة هات
 RT @HebaFaroog: An #Egypt-ian beauty
 :) ▶ http://t.co/S9BZb5f3
 --Wed Feb 20 17:57:53 2013

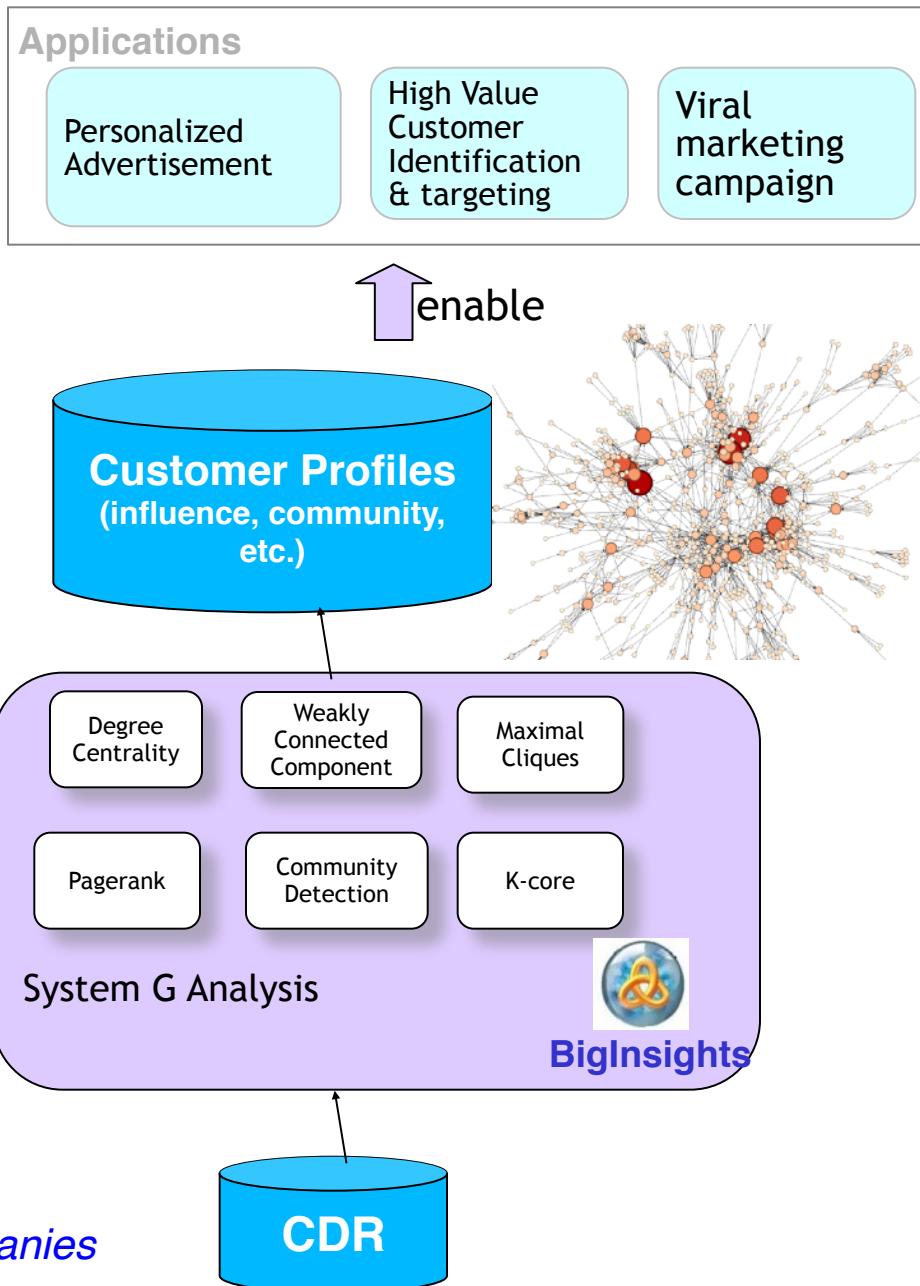
Mona Metwally @monametwally
 معرض محتاج مفترقين دم
 RT @EgyBloodBank: معرض محتاج مفترقين دم بمستشفي الجامعه بالاسكندرية فضله دم اب موجي
 AB+ 01024705247 #Egypt # مصر http://t.co/
 /5oO6mtZ5.
 Translation: . RT *@EgyBloodBank*: A

Real-Time Translation, Location, Top Retweets

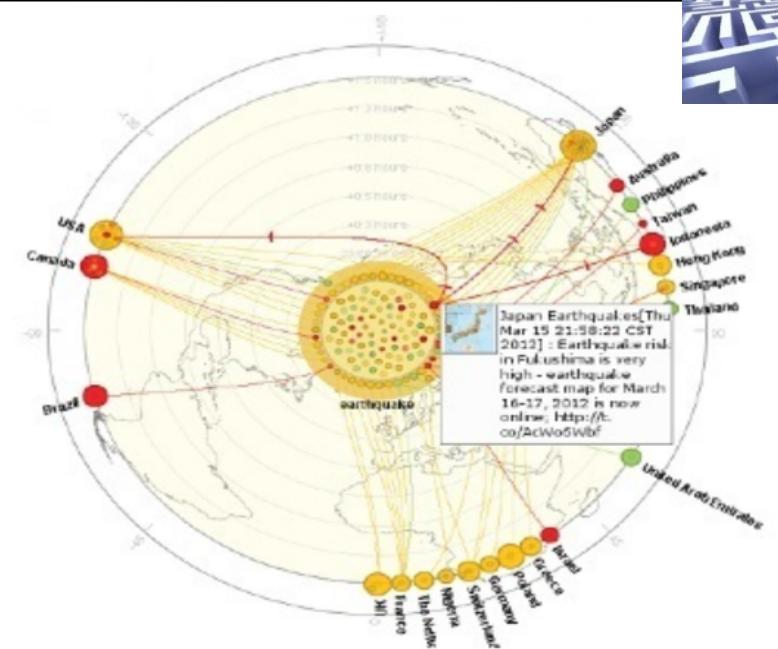
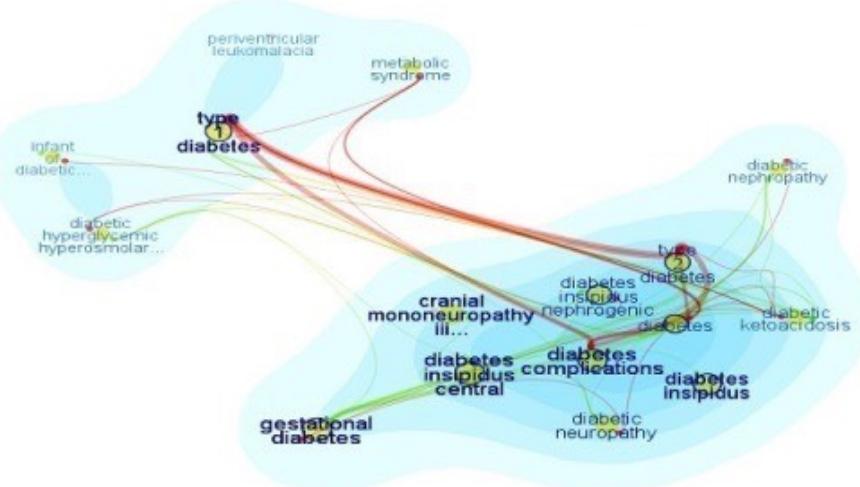
Use Case 6: Customer Social Analysis for Telco

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

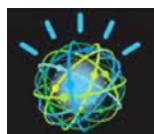
- Applications based on the extracted social profiles
 - Personalized advertisement (beyond the scope of traditional campaign in Telco)
 - High value customer identification and targeting
 - Viral marketing campaign
- Approach
 - Construct social graphs from CDRs based on {caller, callee, call time,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

Communities

Graph Search

Network Info Flow

Bayesian Networks

Centralities

Graph Query

Shortest Paths

Latent Net Inference

Ego Net Features

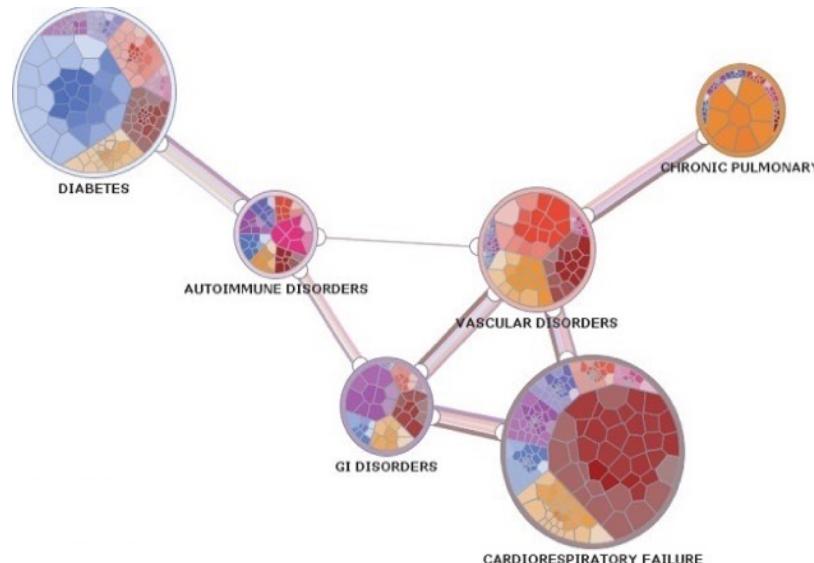
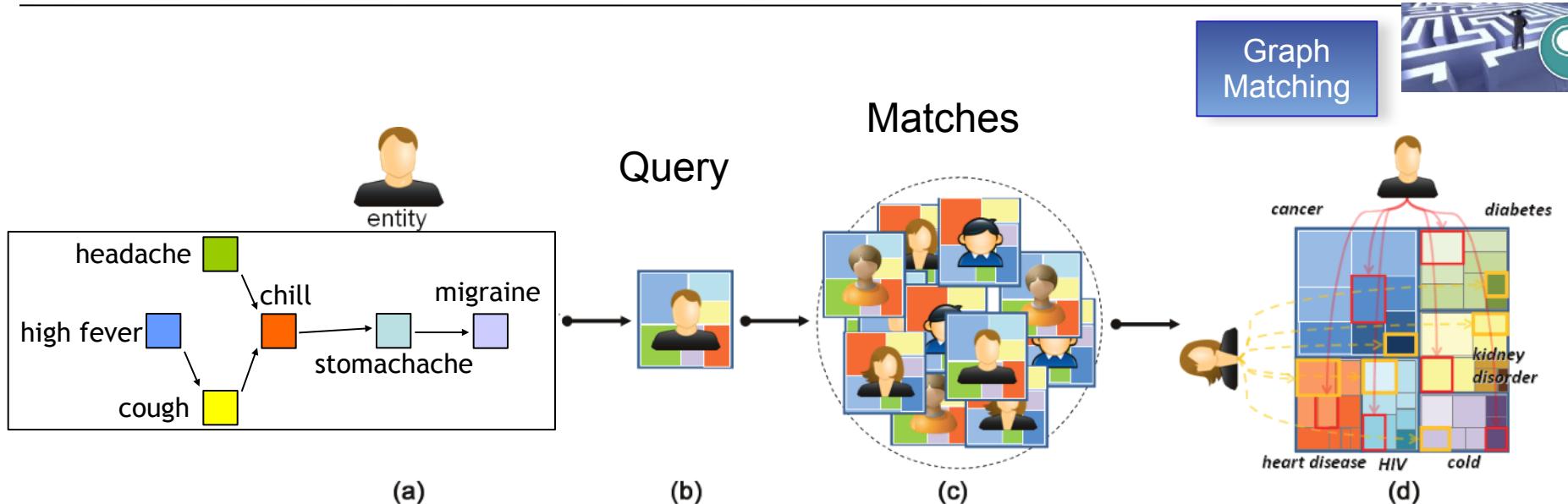
Graph Matching

Graph Sampling

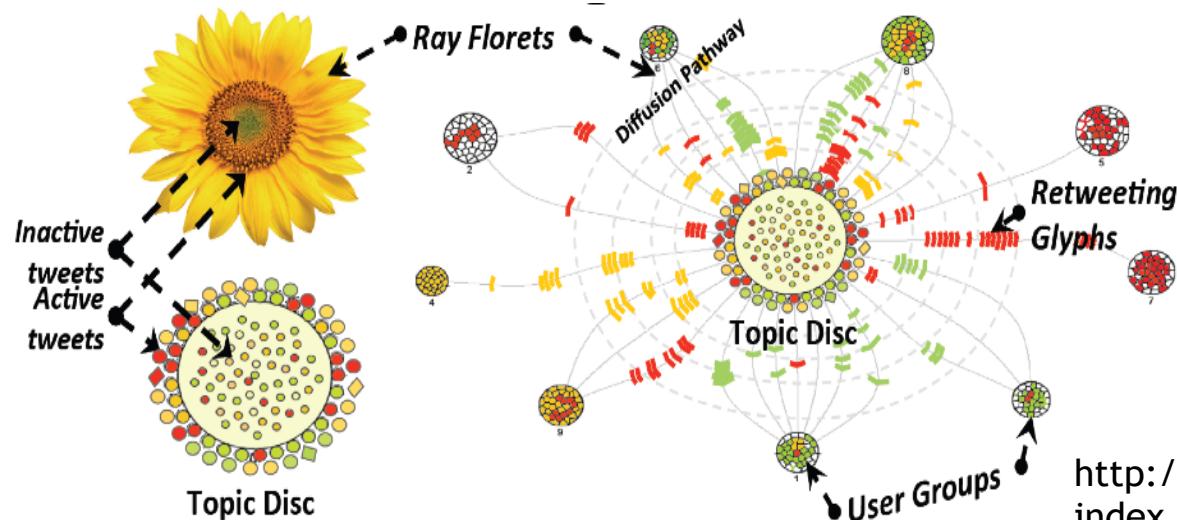
Markov Networks

Middleware and Database

Use Case 7: Graph Analytics and Visualization



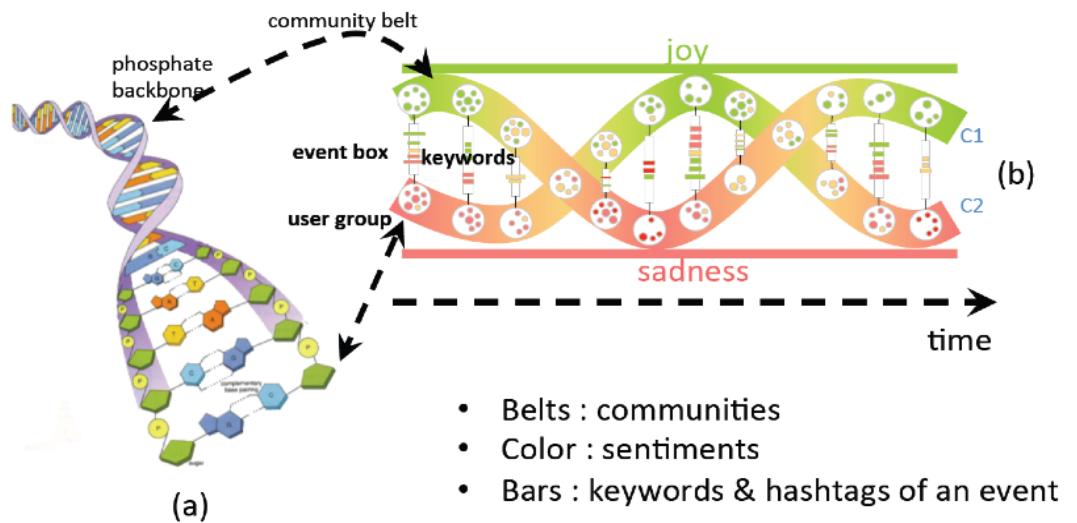
User Case 8: Visualization for Navigation and Exploration



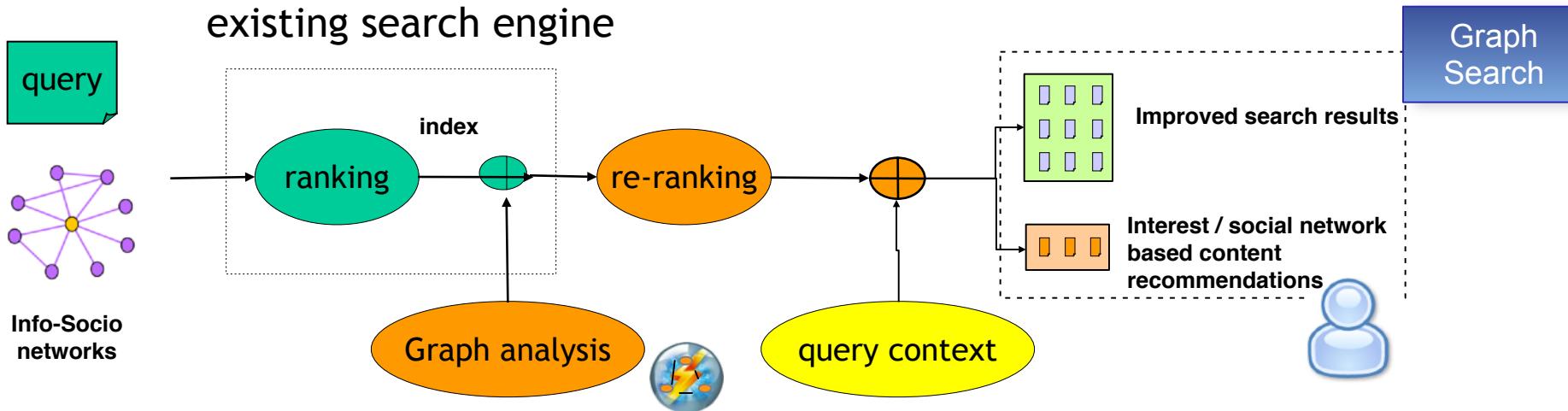
Whisper : Tracing the information diffusion in Social Media

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

SocialHelix: Visualizaiton of Sentiment Divergence in Social Media



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 [?](#)
View as: cloud | list

more — less

2012 analyst_report analytics bao baseline csp deliverable europe forrester fccol gartner gbs gmu government kh leader_priority na proposal public_sector retail sales sales tools sandt social social_business telecommunications

By Category
Select a category to filter search results [?](#)
[Expand all](#) [Collapse all](#)

- ▶ Asset Type
- ▶ Audience
- ▶ Business Topics
- ▶ Client Value Method (CVM)
- ▶ Geography
- ▶ IBM Business Unit
- ▶ Industry
- ▶ Language

Search criteria

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** [?](#)

18,577 results found

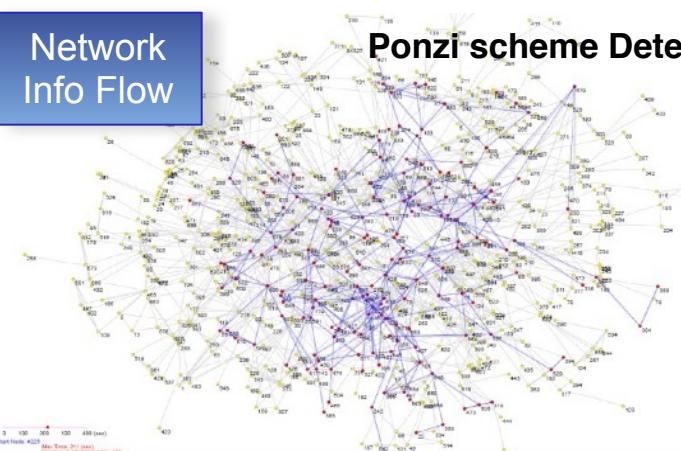
1 to 25 shown

Title	Relevance	Modified	Bookmarks
IBM Social Business Adoption QuickStart (U.S. English) - Proposal Insert [in Proposal and Presentation Accelerator (PPX)] ?	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) ? DAGE@stibo.com			

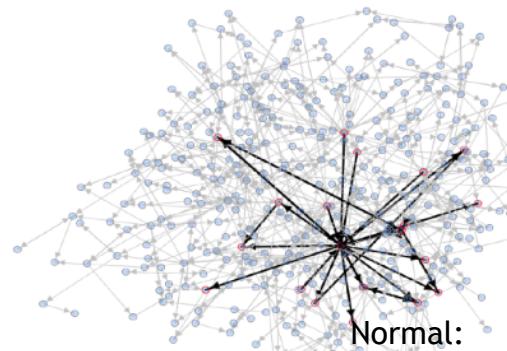
Category 3: Security

Network
Info Flow

Ponzi scheme Detection



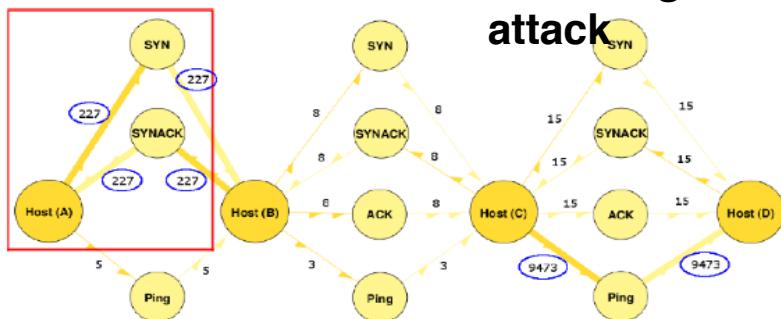
Ego Net
Features



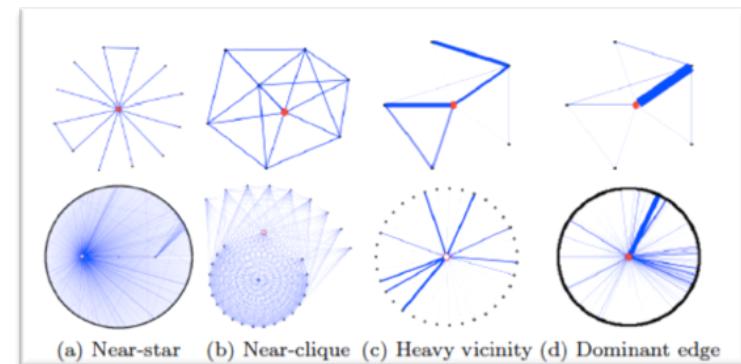
Normal:
(1)Clique-like
(2)Two-way links

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 U.S. Ups Ante for Spying

To Catch Worker Misconduct, Companies Hire Corporate Detectives

“What's emerged is a multibillion dollar detective industry”
npr Jan 10, 2013

Emails

Instant Messaging

Web Access

Executed Processes

Printing

Copying

Log On/Off

Social sensors

Click streams capturer

Feed subscription

Database access

Graph analysis

Behavior analysis

Semantics analysis

Psychological analysis

Multimodality Analysis

Detection,
Prediction
&
Exploration
Interface

Infrastructure + ~ 490 Analytics

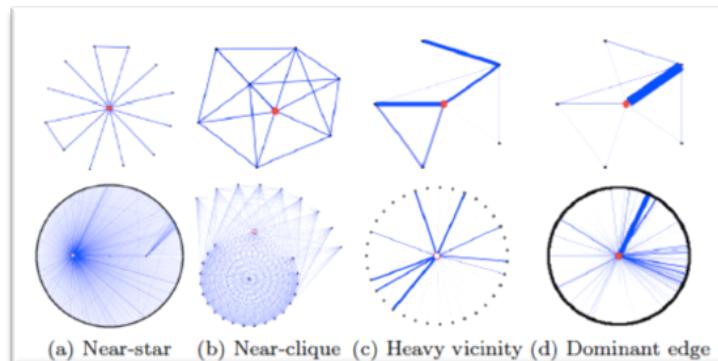
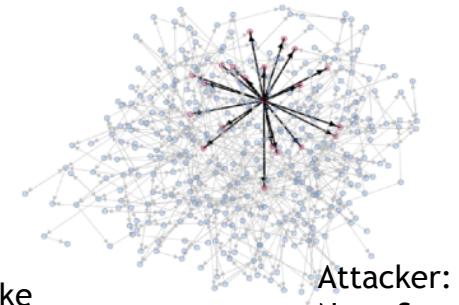
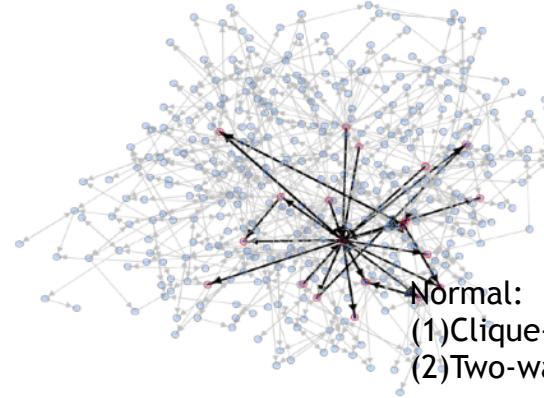
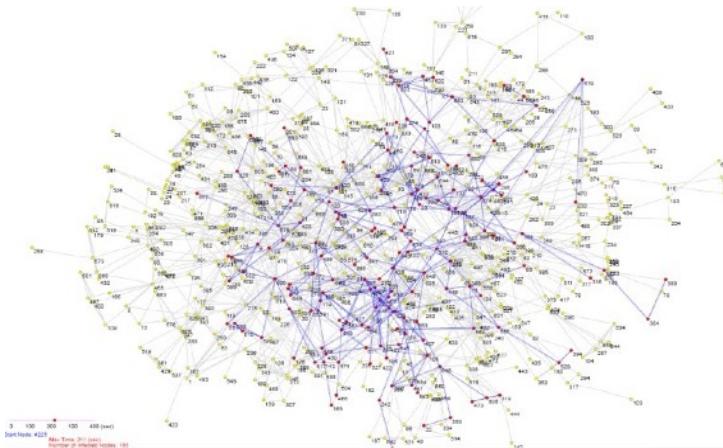
Use Case 11: Fraud Detection for Bank

Network
Info Flow

Ego Net
Features



Ponzi scheme Detection



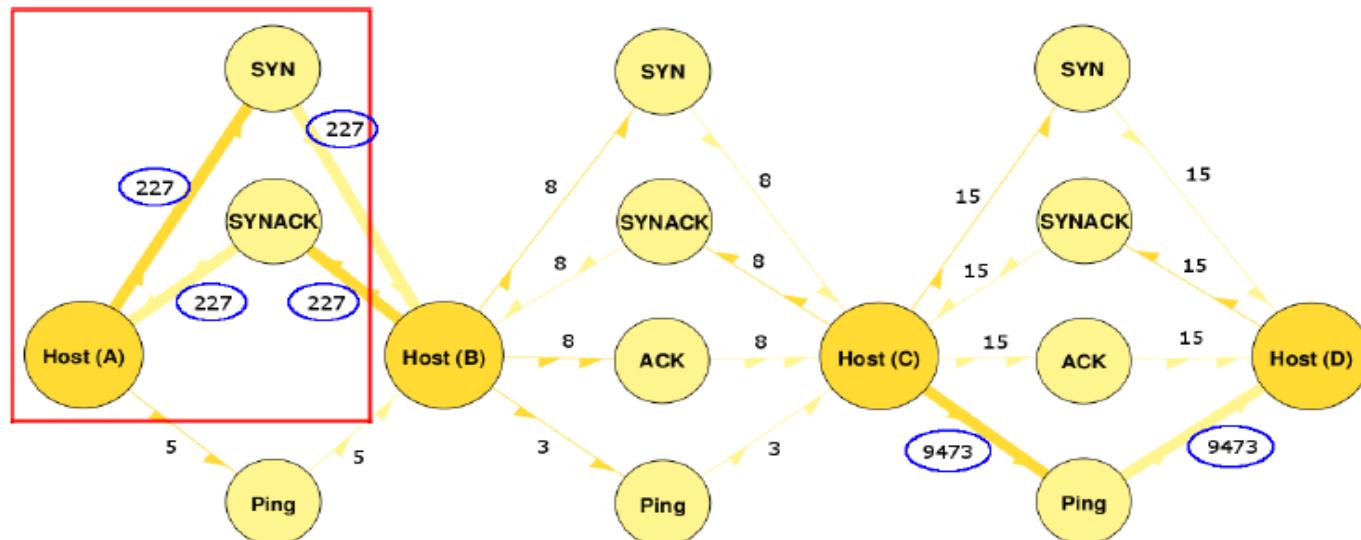
Use Case 12: Detecting Cyber Attacks

Network
Info Flow

Ego Net
Features

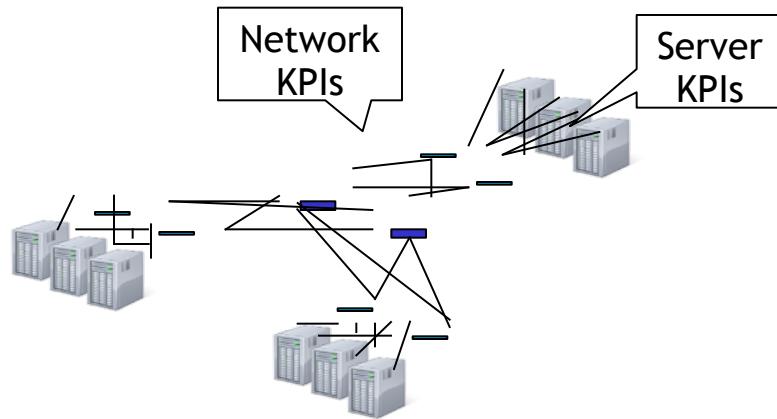


Detecting DoS
attack



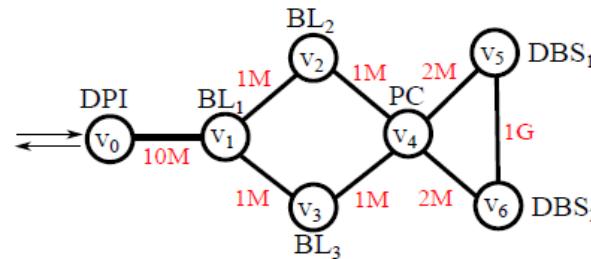
(a) Single large graph representing TCP SYN and ICMP PING network traffic, with two Denial of Service (DoS) attacks taking place.

Category 4: Operations Analysis



Cloud Service Placement

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



Memory requirements

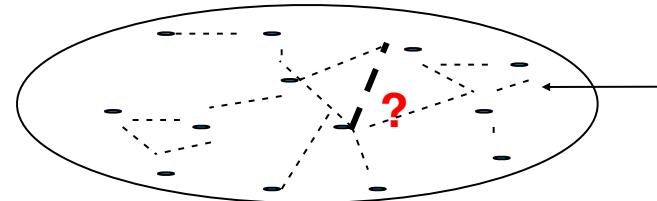
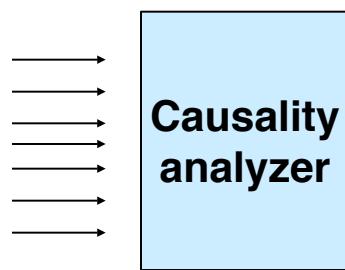
v ₀	8G
v ₁	2.5G
v ₂	2G
v ₃	2G
v ₄	12G
v ₅	20G
v ₆	32G



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

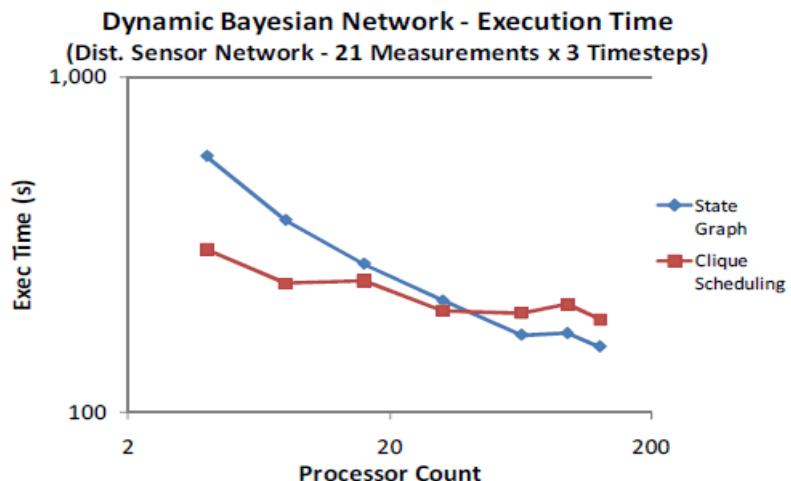
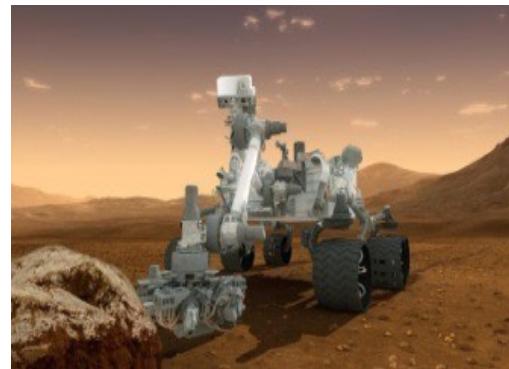
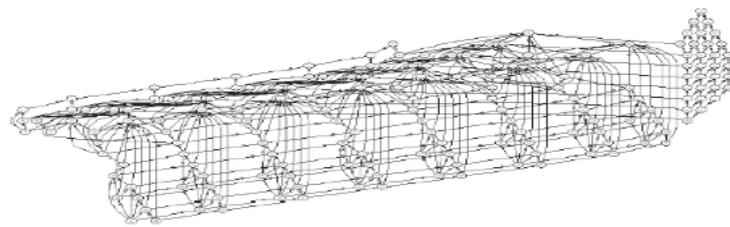
Use Case 13: Smarter *another* Planet

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.

Bayesian Network



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

- * 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

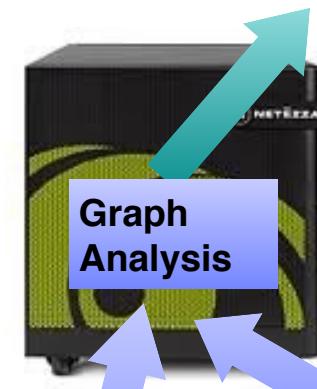
- (1)CDRs, already collected for billing purposes, contain information about voice/data calls
- (2)Traditional NMS* and EMS** typically lack of end-to-end visibility and topology across vendors
- (3)Employ graph algorithms to analyze network elements which are not reported by the usage data from CDR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

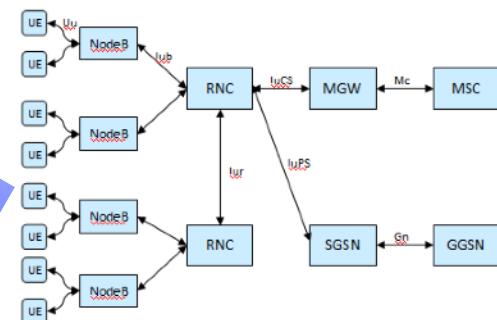


Network load level report



CDR

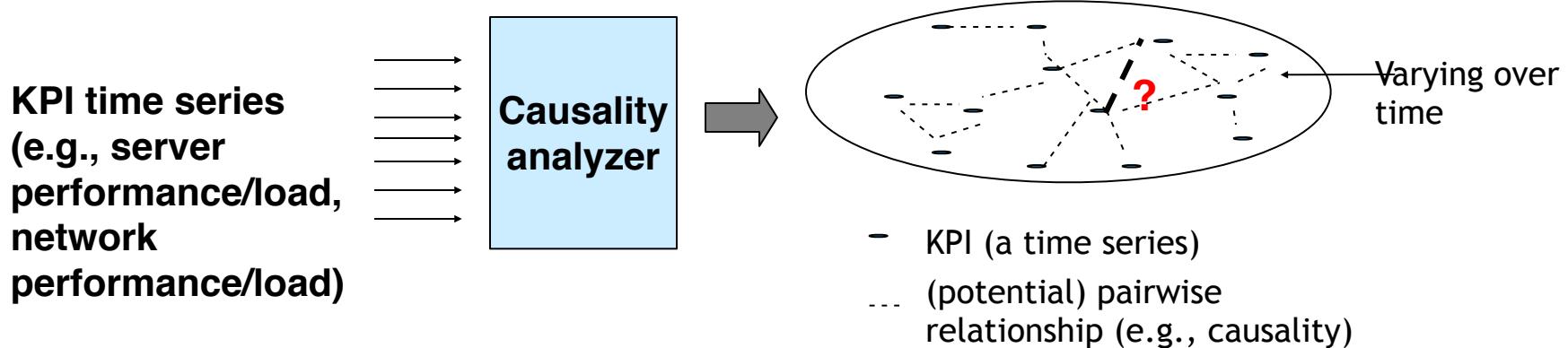
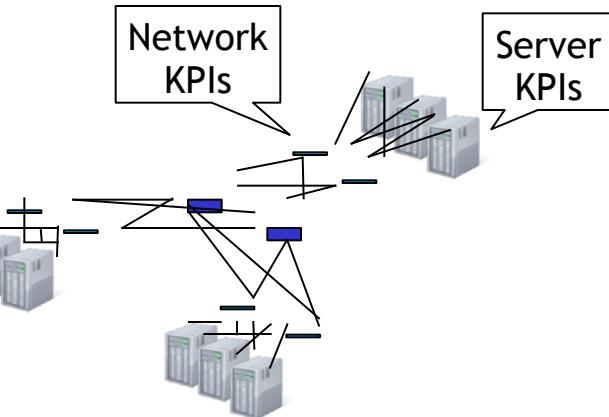
Network topology



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 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)*



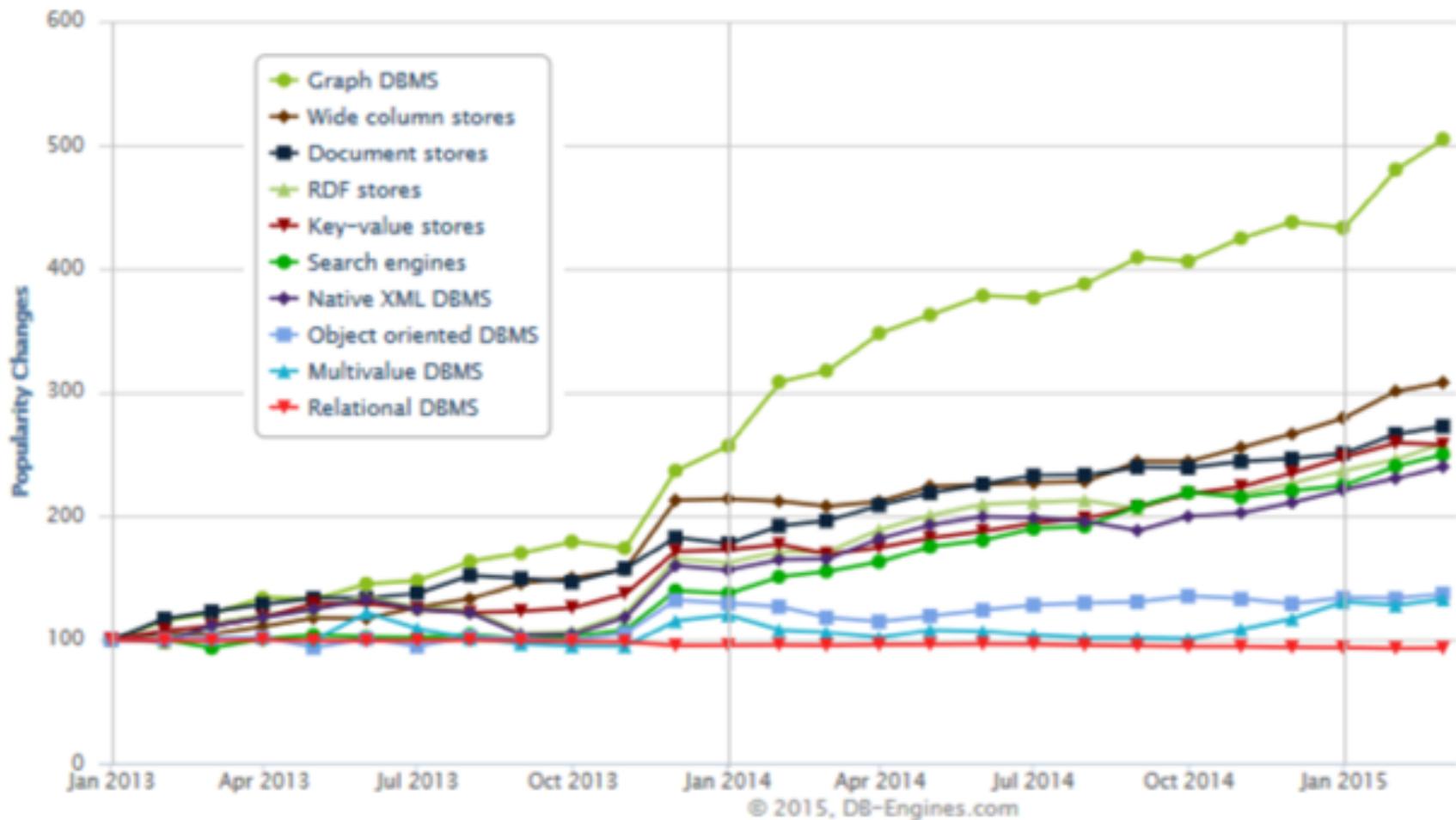
Our approach:
Probabilistic monitoring via sampling & estimation

Basic analytics engine
(e.g., pairwise granger causality)

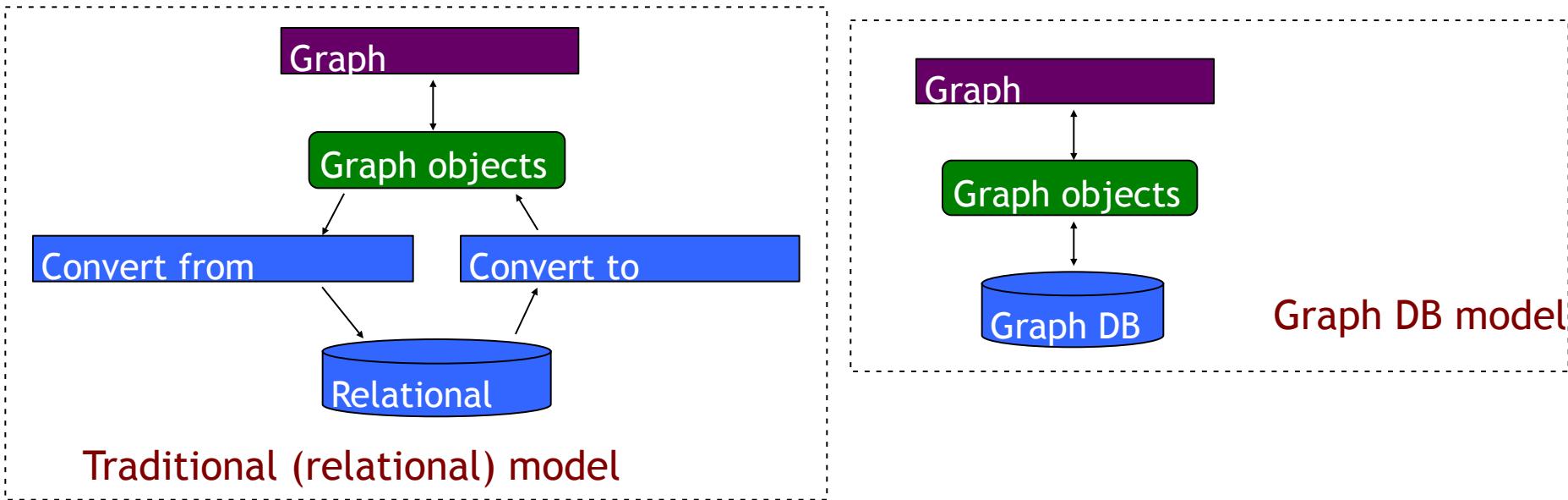
Link sampling & estimation

Select KPI pairs (sampling) → Test link existence → Estimate unsampled links based on history
50 → Overall graph

Category 5: Data Warehouse Augmentation



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

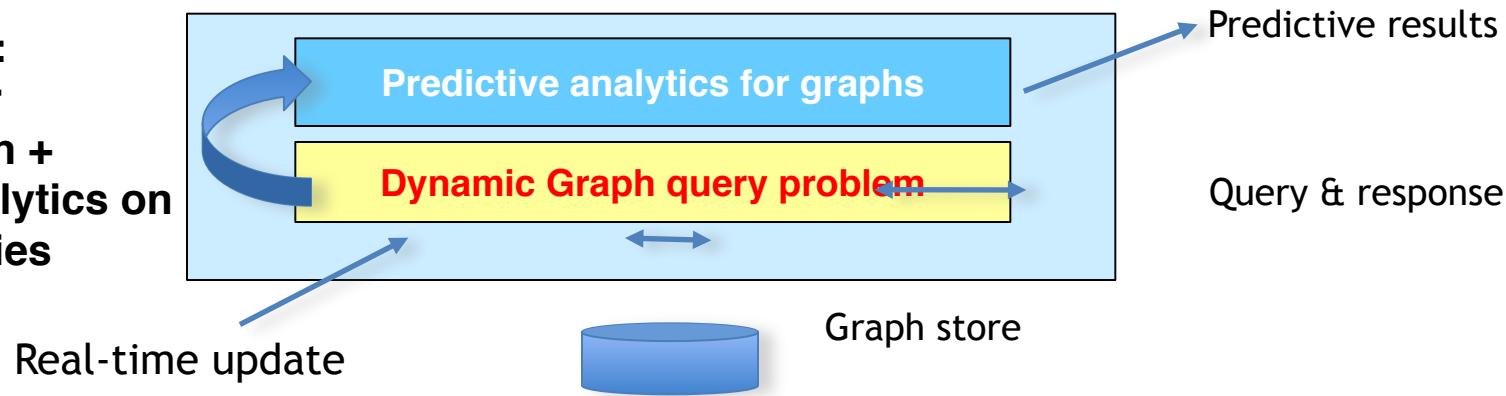
Use Case 17: Smart Navigation Utilizing Real-time Road Information

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

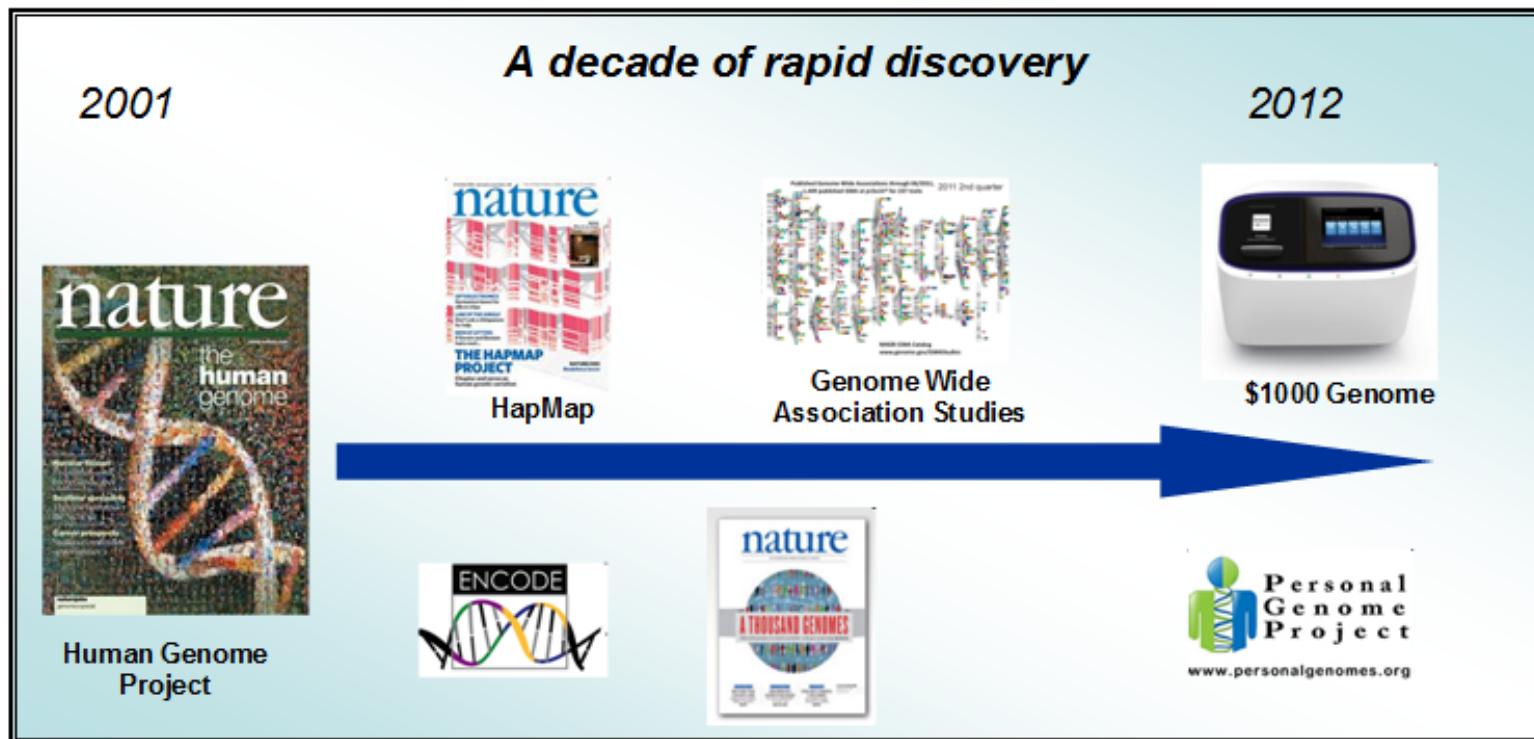
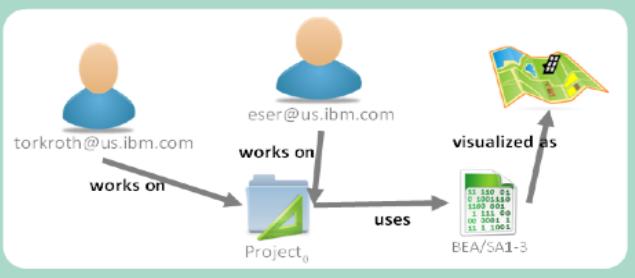


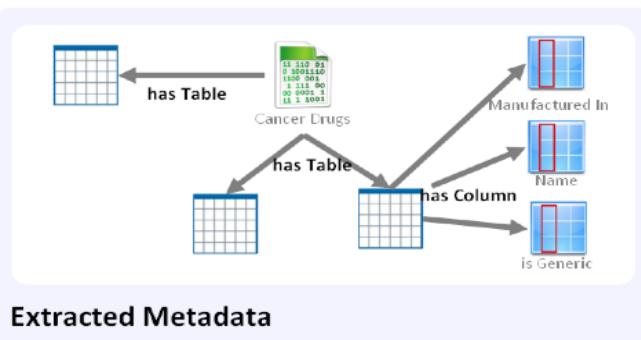
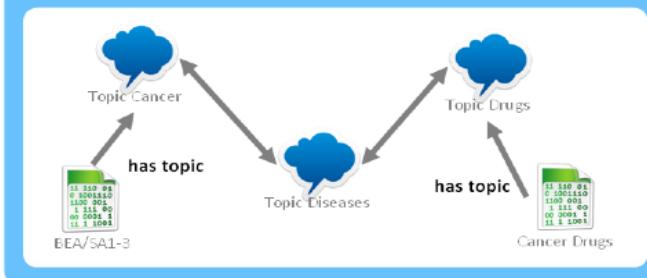
Figure 1: Since the Human Genome Project, various projects have started to reveal the mysteries of genomes and the \$1000 Genome is almost reality.

Use Case 20: Data Curation for Enterprise Data Management

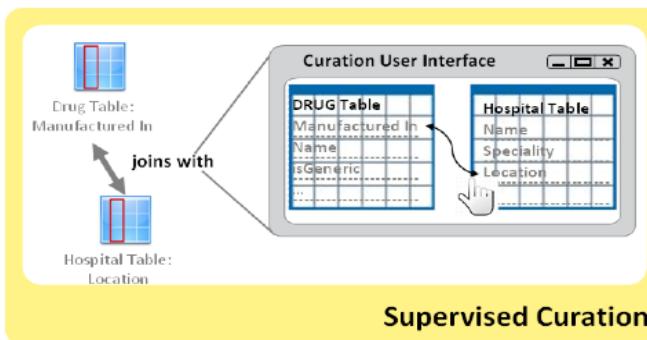
Prior Collaborative Use



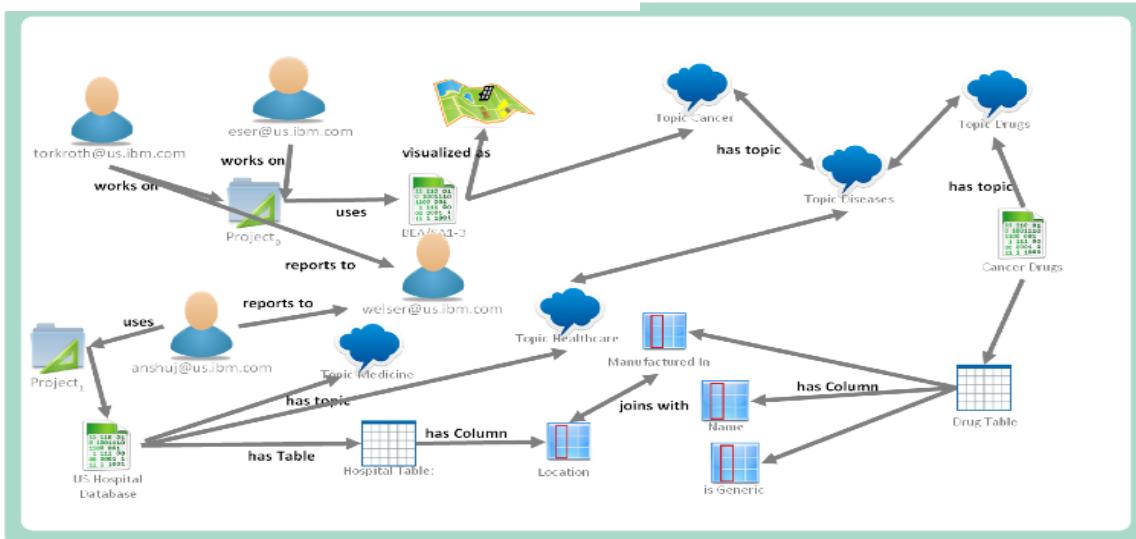
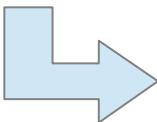
Semantic Knowledge



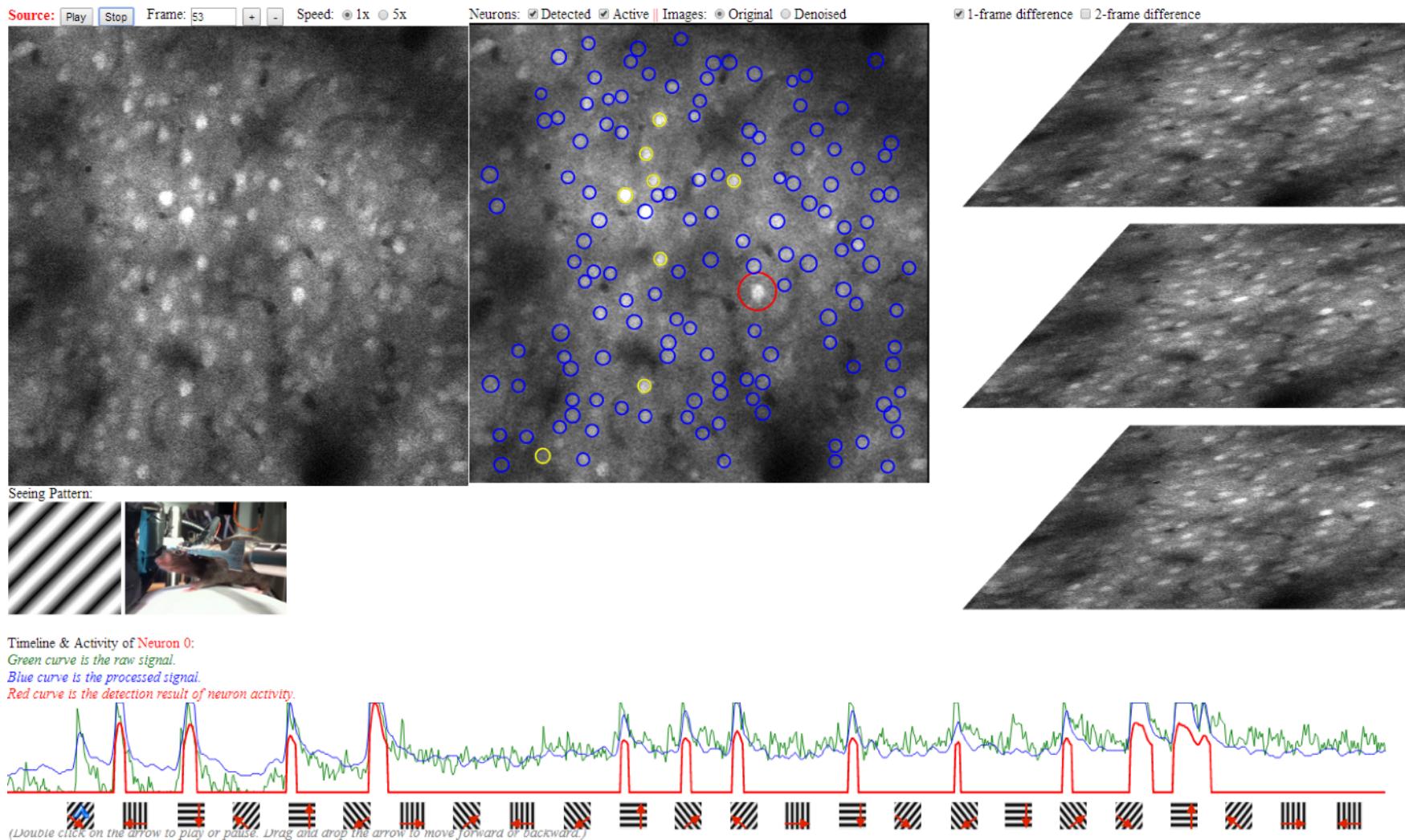
Extracted Metadata



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...](#)



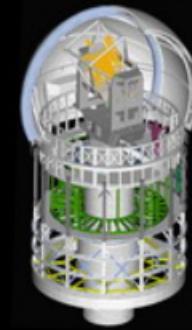
The PS1 Prototype

PS1 goes operational and begins science mission

PS1 Science Consortium formed...

[PS1SC Blog](#)

[PS1 image gallery](#)



Final Project Proposal Scoring

Proposal — preparing about 5 pages (each item 1/5 of the proposal score):

Goal — novel? challenging?

Data — 3Vs? New dataset? Existing dataset?

Methods — planning of methodologies and algorithms? Feasible?

System — an overview of system. What will be implemented?

Schedule — what to achieve by what time, and by whom?

Final Project Report Scoring

- **Title, Author(s)**
- **Abstract:** Briefly describe your problem, approach, and key results.
- **Introduction (5%):** Describe and define the problem you are working on. Why is it important? Include an overview of your methods and results.
- **Related Work (5%):** Discuss published works or approaches that are related to your project. What's the benefit or drawback of the previous works? What kind of problems have they solved? How is your approach similar or different from others?
- **Data (10%):** Describe the data you are working with for your project. What type of data is it? Where did it come from? How much data are you working with? Did you have to do any preprocessing, filtering, feature engineering or other special treatment to use this data in your project?
- **Methods (25%):** Discuss your approach for solving the problems that you set up in the introduction. Why is your approach the right thing to do? Did you consider alternative approaches? Have you tried some methods that didn't work out? It may be helpful to include figures, diagrams, or tables to describe your method or compare it with other methods.
- **Experiments (20%):** Discuss the experiments that you performed to demonstrate your approach solves the problem. The experiments will vary depending on the project, but you might compare with previously methods, determine the impact of the components of your system, experiment with different hyper-parameters, architectures, or algorithms, use visualization techniques to gain insight of how your model works, etc. Graphs, tables, and figures are highly recommended to be included to illustrate your experimental results.
- **System Overview (25%):** Describe the software architecture and tech stacks of your application. Discuss potential bottlenecks and improvements that could be made. Mention the software packages that you used. Mention how to use your application. You could provide screenshots to your application.
- **Conclusion (5%):** Summarize your key results. What have you learned? What problems have you discovered and solved? Suggest ideas for future extensions or new applications.
- **Writing / Formatting (5%)** Is your paper clearly written and nicely formatted?