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

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

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

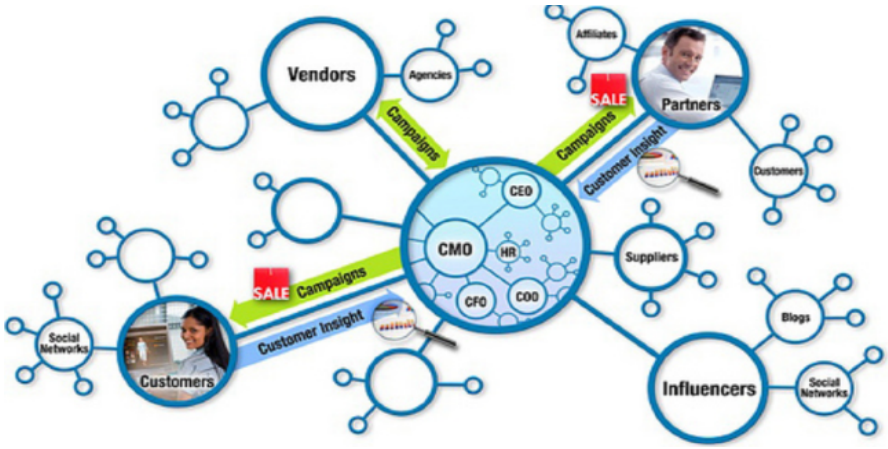
Data Warehouse Augmentation
Integrate big data and data warehouse capabilities to increase operational efficiency

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

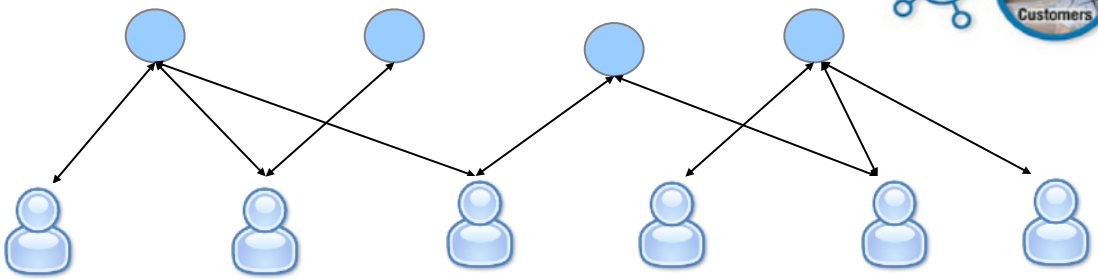
Category 1: 360° View

Recommendation



item

user



Enhancing:

Coremetrics An IBM Company

unica

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

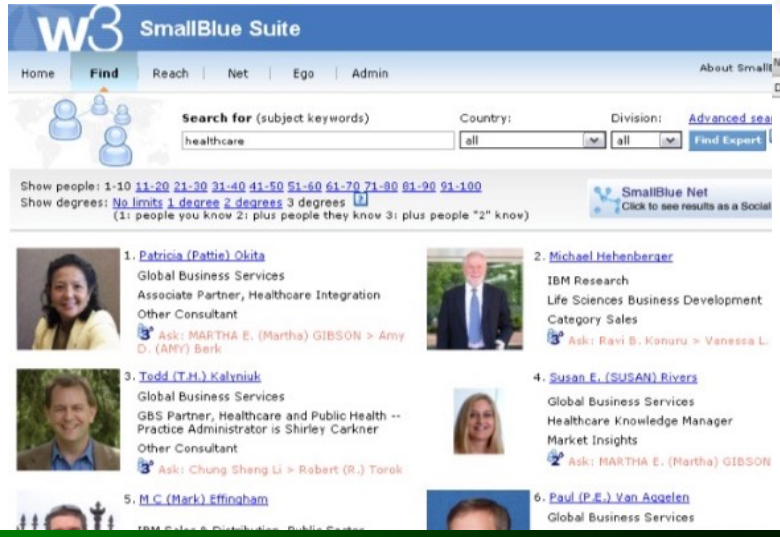
1,000,000+ Lotus Connections (blogs, file sharing, bookmark) data

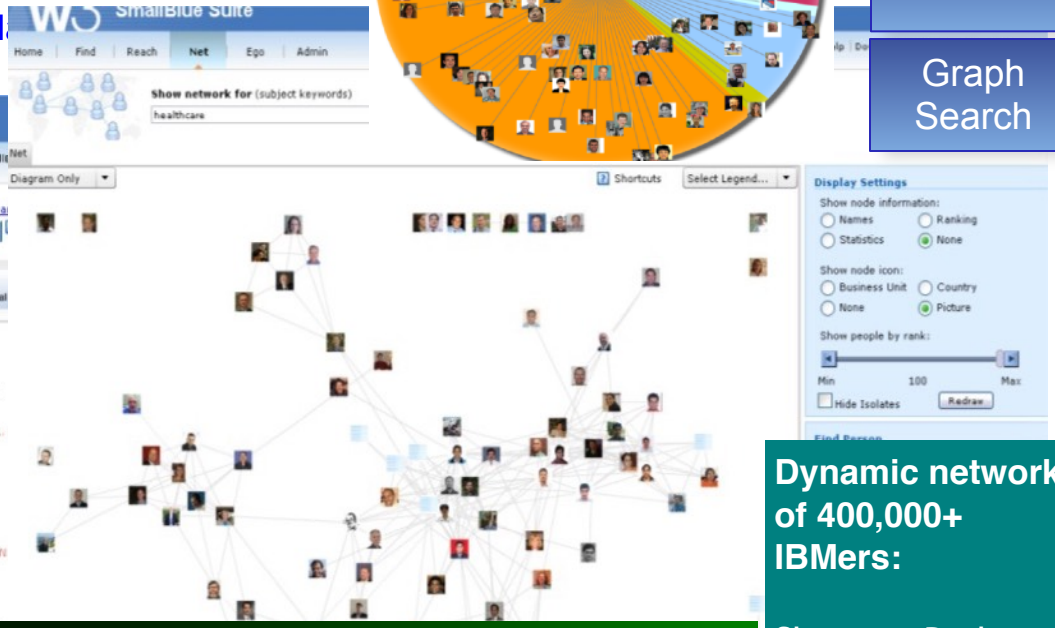
200,000 people's consulting project & earning d

Shortest Paths

Centralities

Graph Search



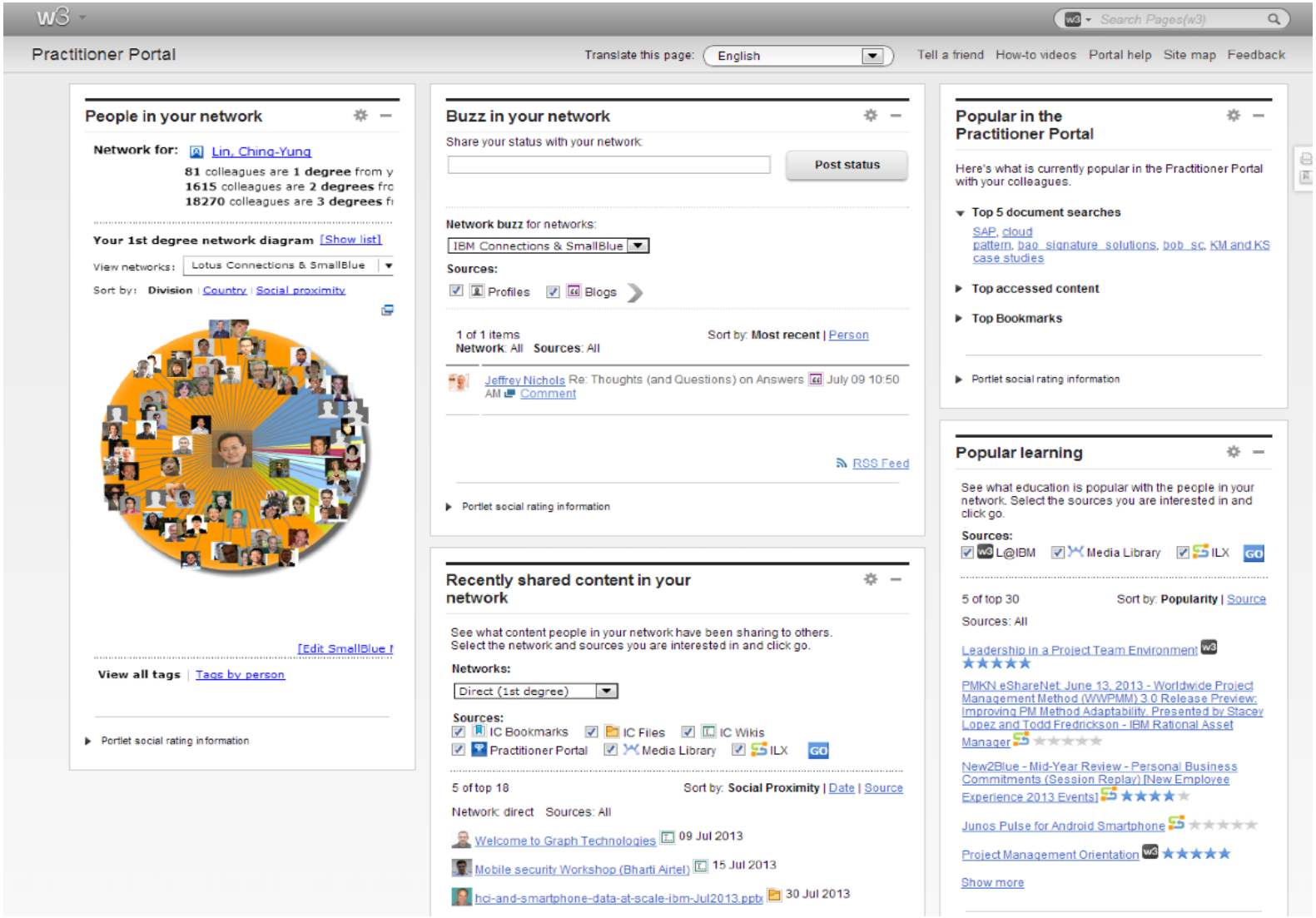


- On BusinessWeek four times, including being the Top Story of Week, April 2009
- Help IBM earned the 2012 Most Admired Knowledge Enterprise Award
- Wharton School study: $7,010 gain per user per year using the tool
- In 2012, contributing about 1/3 of GBS Practitioner Portal $228.5 million savings an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

Dynamic networks of 400,000+ IBMers:

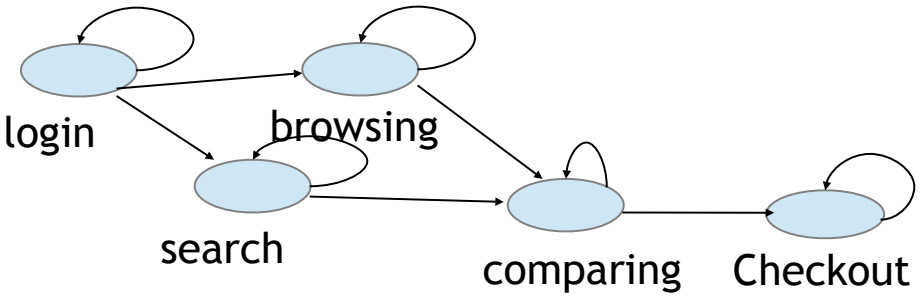
Shortest Paths
Social Capital
Bridges
Hubs
Expertise Search
Graph Search
Graph Recomm.

Use Case 2: Personalized Recommendation



Use Case 3: Customer Behavior Sequence Analytics

Markov Network | Latent Network | Bayesian Network

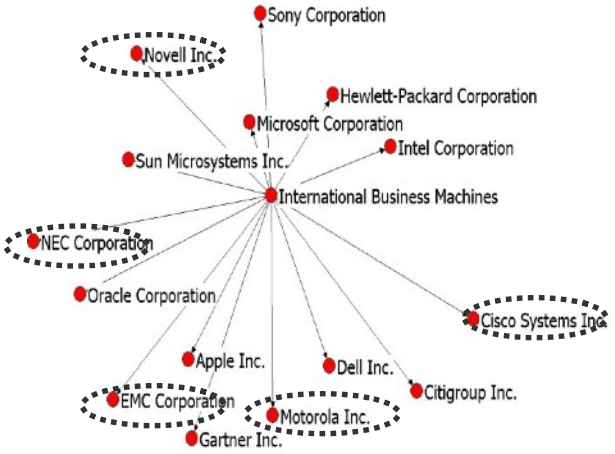


- Behavior Pattern Detection
- Help Needed Detection

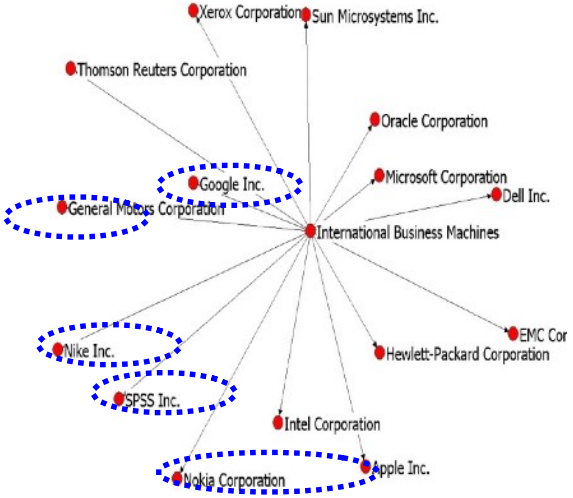
Use Case 4: Graph Analytics for Financial Analysis

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

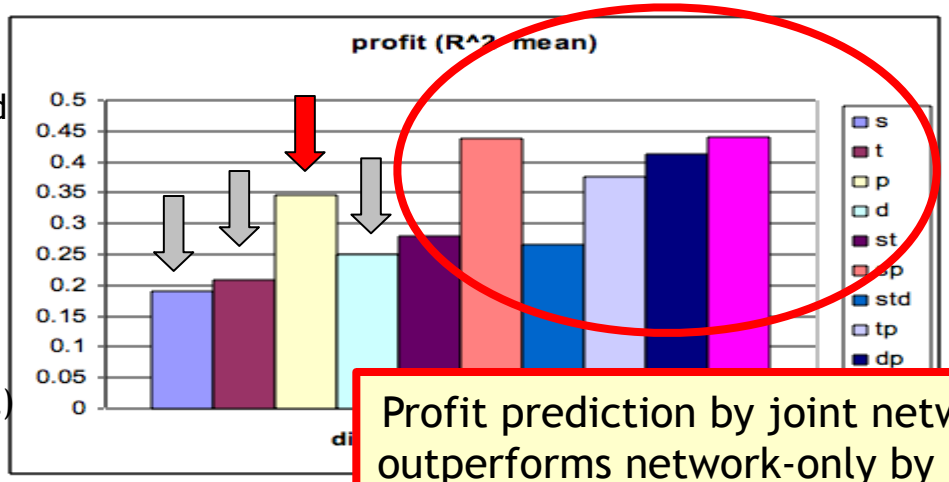


- Data Source:
  - Relationships among 7594 companies, data mining from NYT 1981 ~ 2009

Targets: 20 Fortune companies' normalized Profits

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

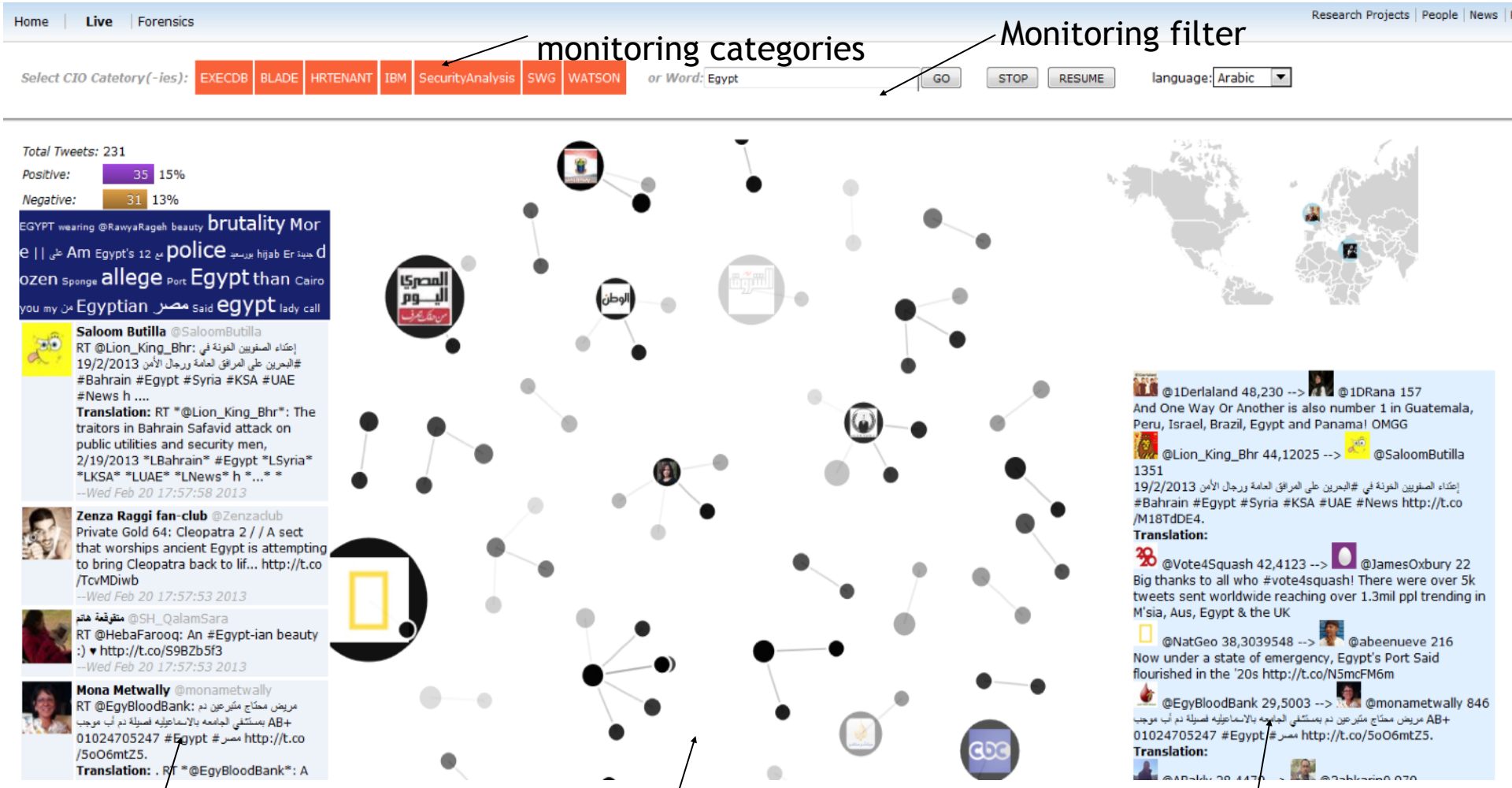
Model: Support Vector Regression (RBF kernel)



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%.

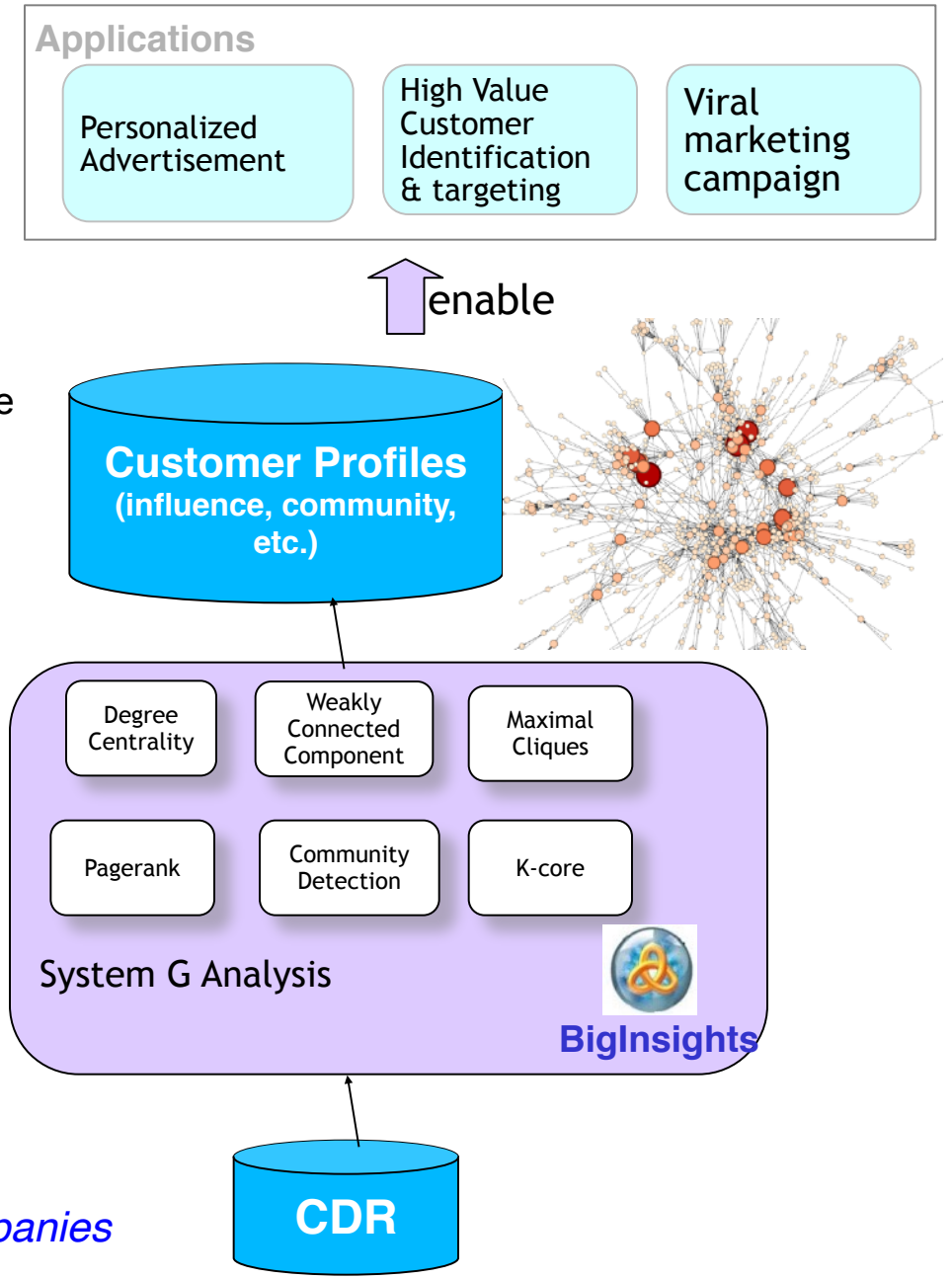
Use Case 5: Social Media Monitoring



Use Case 6: Customer Social Analysis for Telco

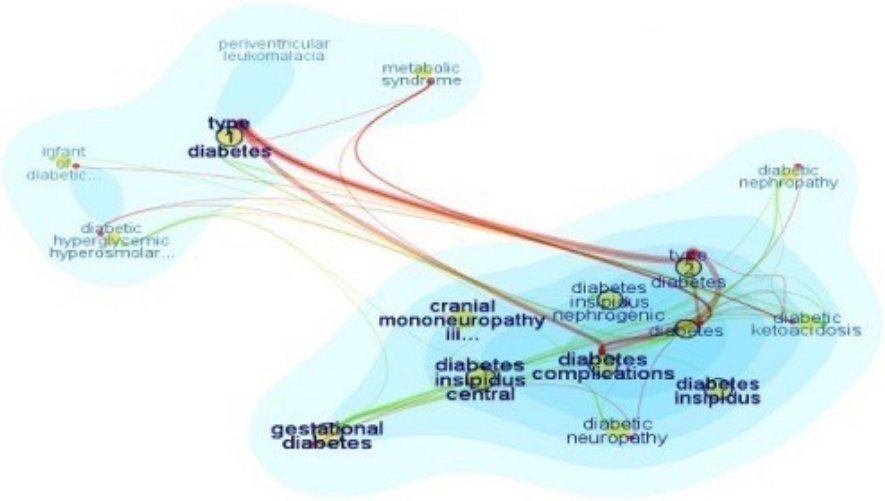
Goal: Extract customer social network behaviors to enable Call Detail Records (CDRs) data monetization for Telco.

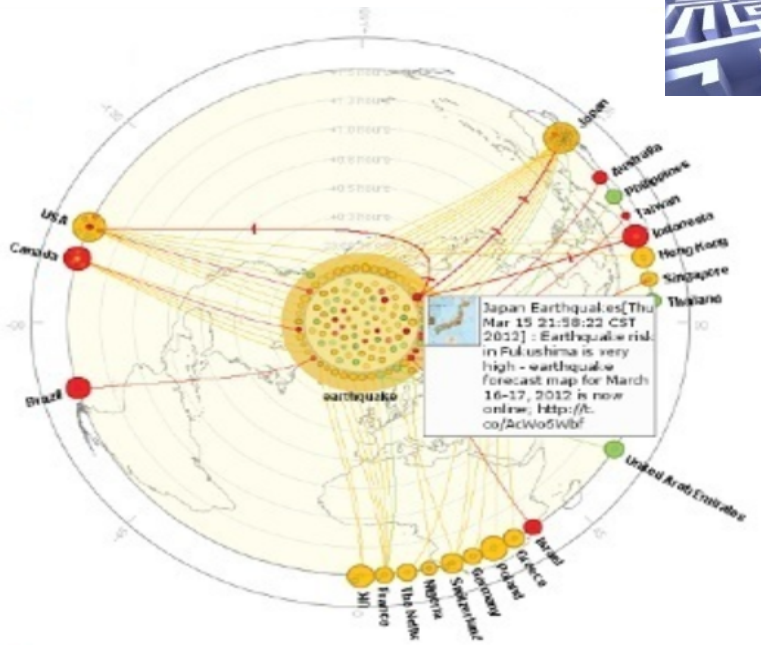
- Applications based on the extracted social profiles
  - Personalized advertisement (beyond the scope of traditional campaign in Telco)
  - High value customer identification and targeting
  - Viral marketing campaign
- Approach
  - Construct social graphs from CDRs based on {caller, callee, call time,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

Vivisimo

CÚRAM SOFTWARE

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

Communities | Graph Search | Network Info Flow | Bayesian Networks

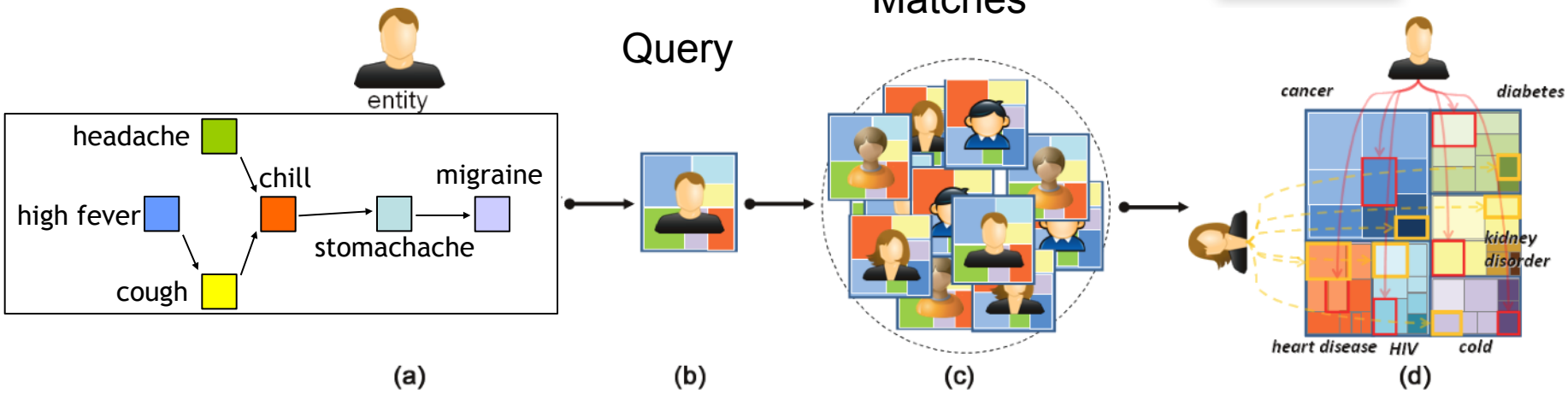
Centralities | Graph Query | Shortest Paths | Latent Net Inference

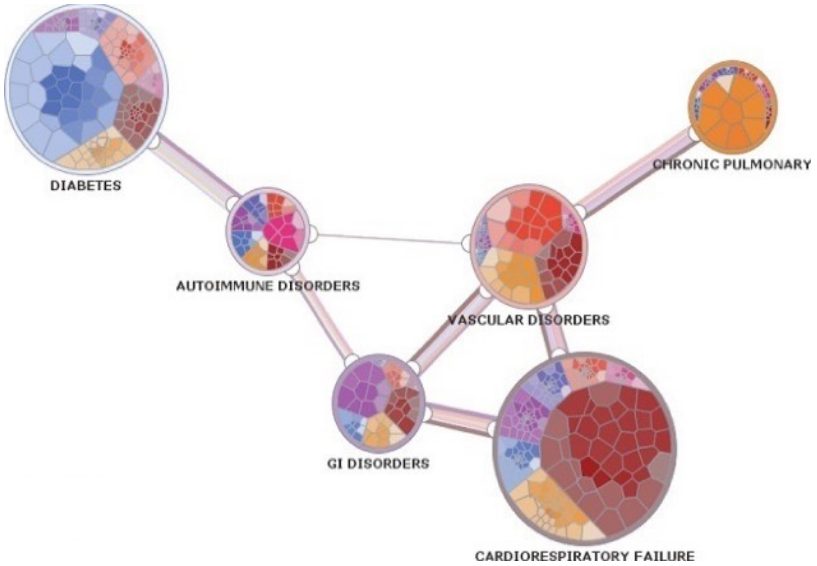
Ego Net Features | Graph Matching | Graph Sampling | Markov Networks

Middleware and Database

Use Case 7: Graph Analytics and Visualization

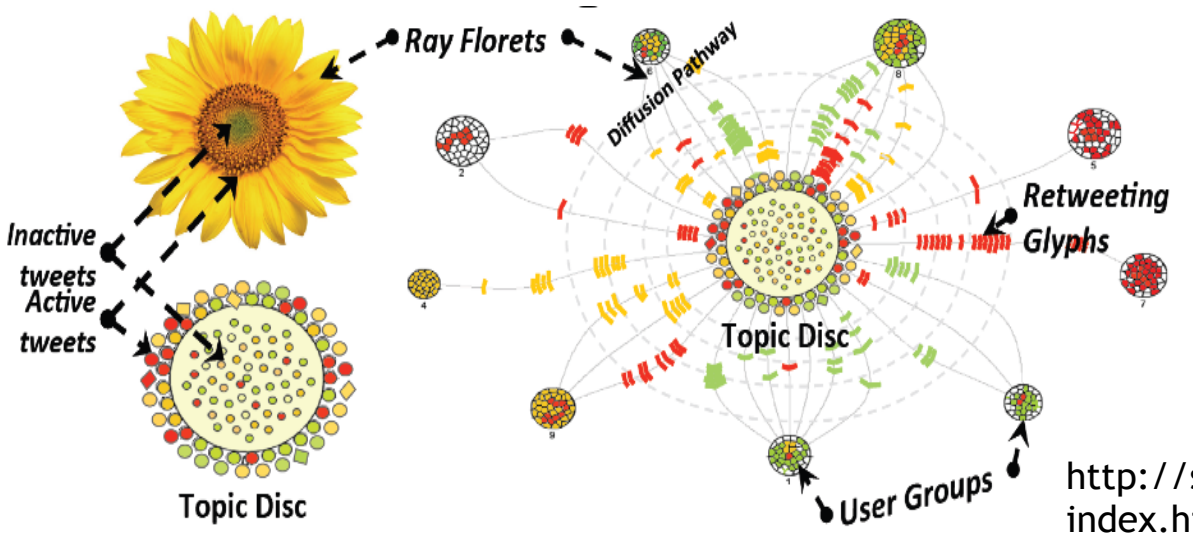
Graph Matching





Graph Communities

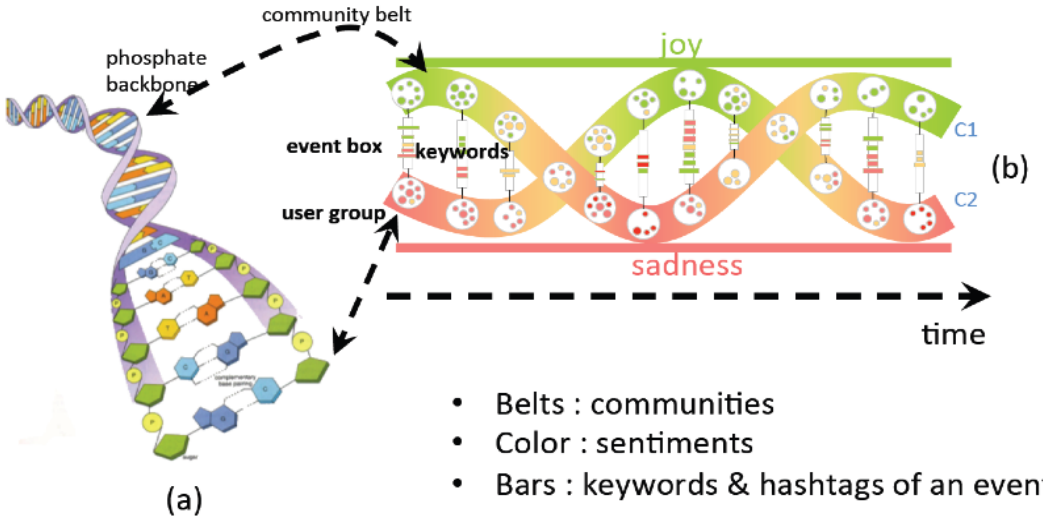
User Case 8: Visualization for Navigation and Exploration



Whisper : Tracing the information diffusion in Social Media

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

SocialHelix: Visualizaiton of Sentiment Divergence in Social Media



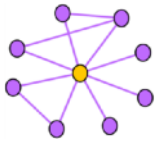
- Belts : communities
- Color : sentiments
- Bars : keywords & hashtags of an event

Use Case 9: Graph Search

Graph Search

existing search engine

query



Info-Socio networks

index

ranking

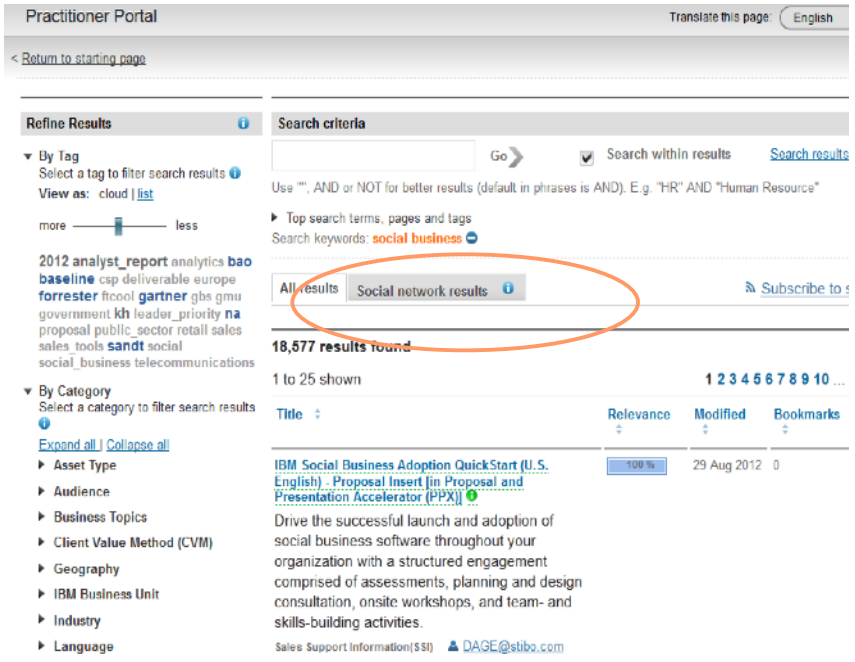
re-ranking

Improved search results

Interest / social network based content recommendations

Graph analysis

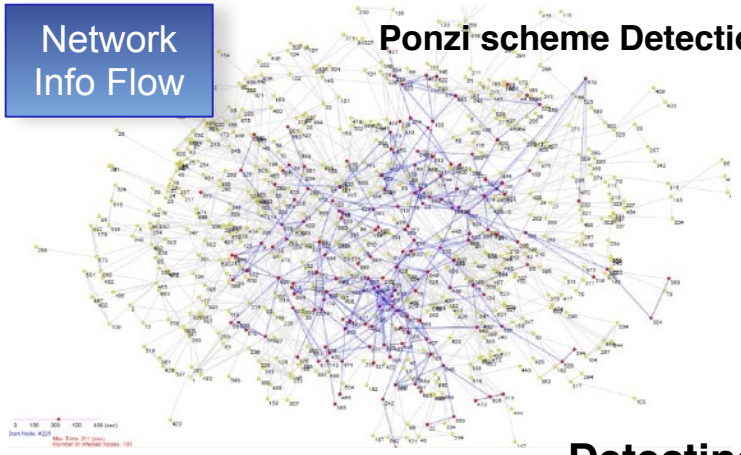
query context



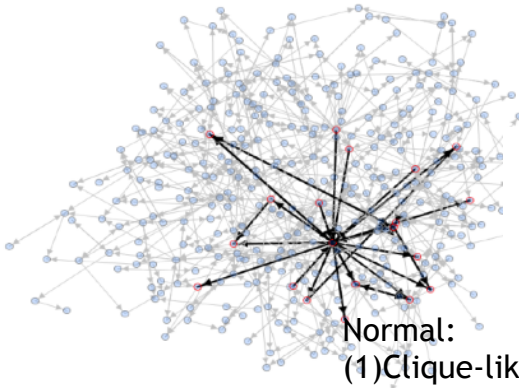
Category 3: Security

Network Info Flow

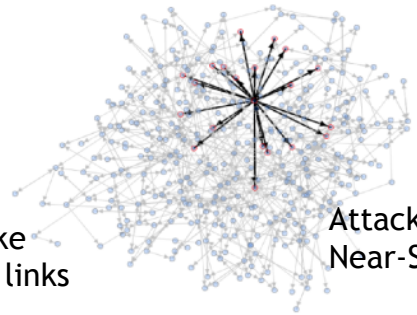
Ponzi scheme Detection



Ego Net Features

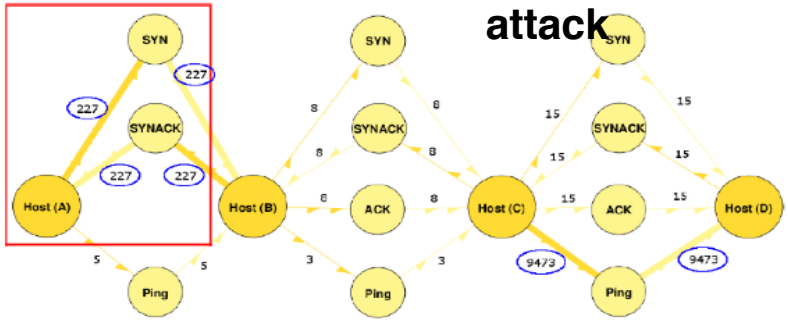


Normal:
(1)Clique-like
(2)Two-way links

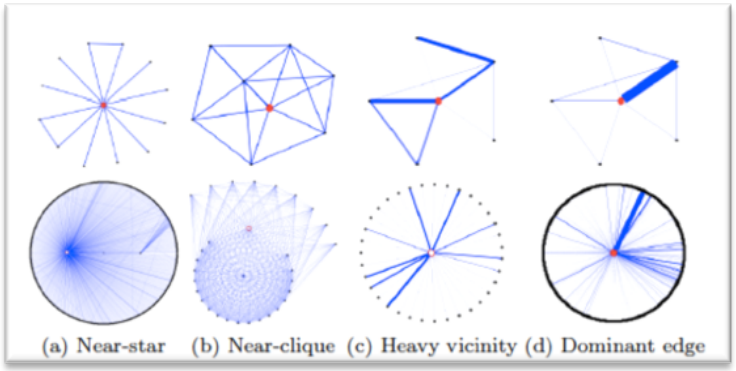


Attacker:
Near-Star

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.

"Enterprise Information Leakage Impacted economy and jobs" Feb 2013

"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

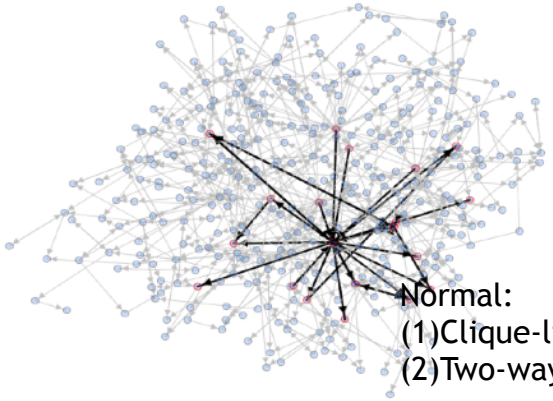
Infrastructure + ~ 490 Analytics

Use Case 11: Fraud Detection for Bank

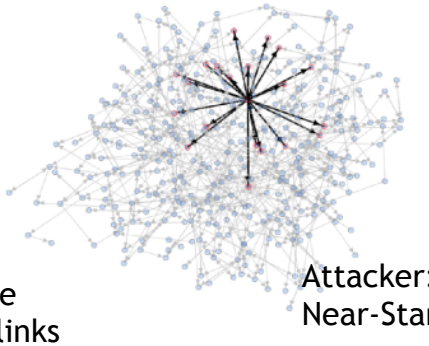
Network Info Flow

Ego Net Features

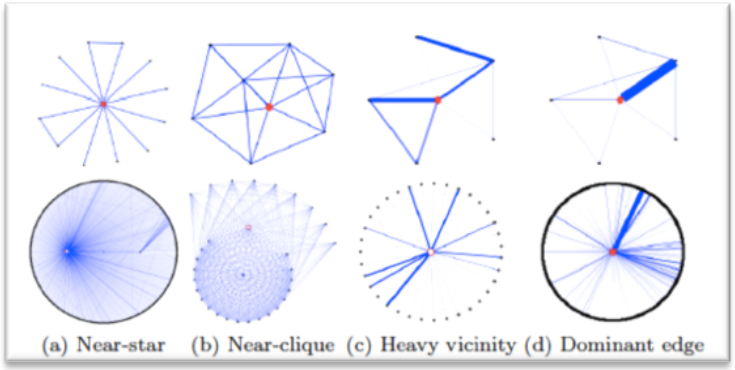
Ponzi scheme Detection



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



Attacker:
Near-Star

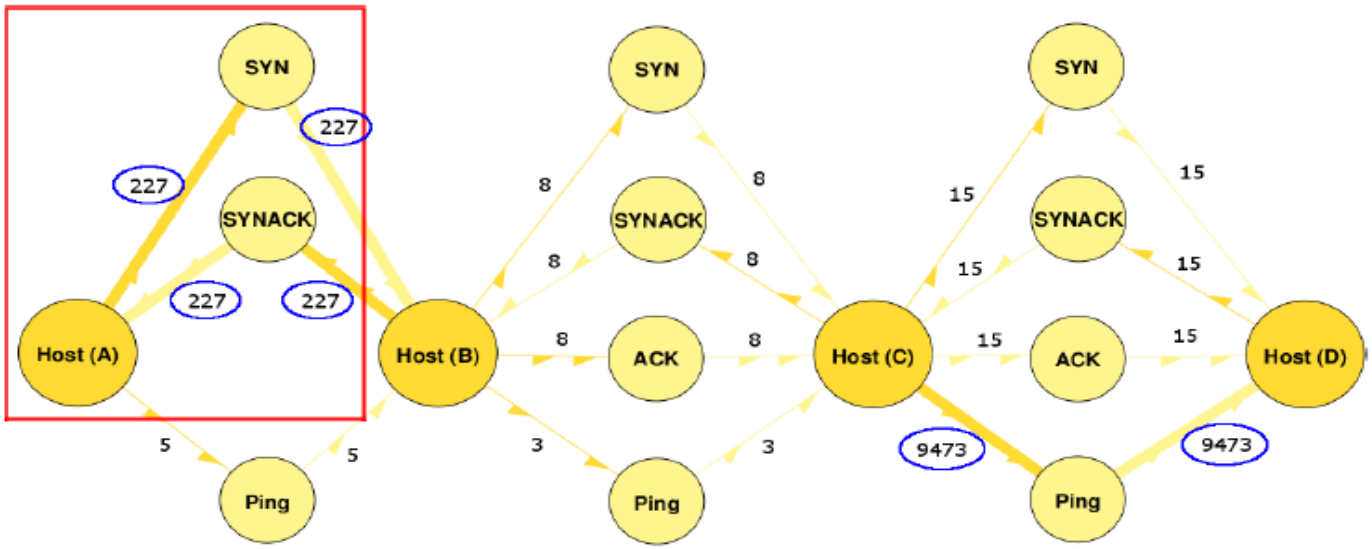


Use Case 12: Detecting Cyber Attacks

Network Info Flow

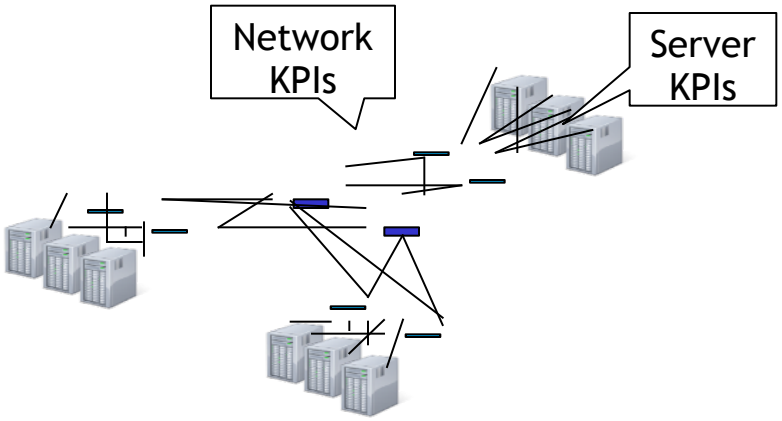
Ego Net Features

Detecting DoS attack



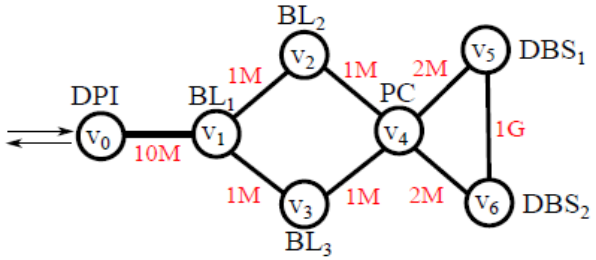
(a) Single large graph representing TCP SYN and ICMP PING network traffic, with two Denial of Service (DoS) attacks taking place.

Category 4: Operations Analysis



Cloud Service Placement

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



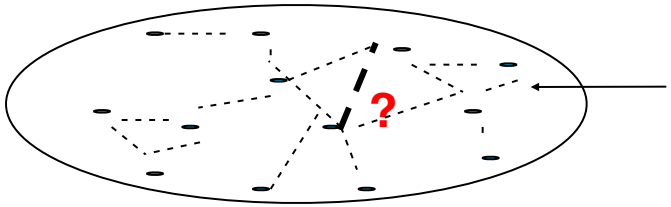
Memory requirements	
v_0	8G
v_1	2.5G
v_2	2G
v_3	2G
v_4	12G
v_5	20G
v_6	32G

Graph Matching

Bayesian Network

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

Causality analyzer



Varying over time

– KPI (a time series)

--- (potential) pairwise relationship (e.g., causality)

Graph Visualizations

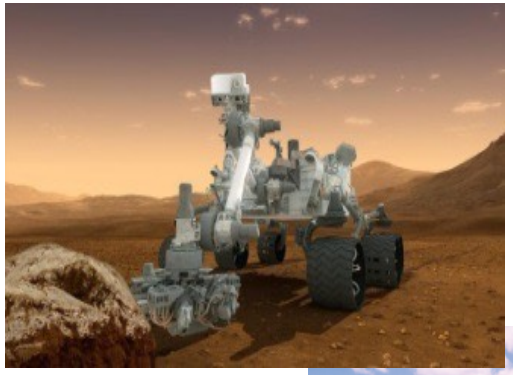
Communities	Graph Search	Network Info Flow	Bayesian Networks
Centralities	Graph Query	Shortest Paths	Latent Net Inference
Ego Net Features	Graph Matching	Graph Sampling	Markov Networks

Middleware and Database

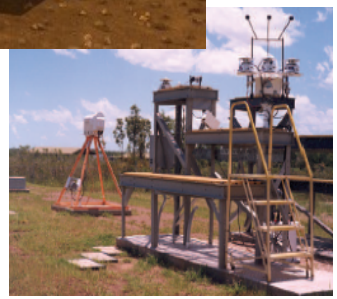
Use Case 13: Smarter *another* Planet

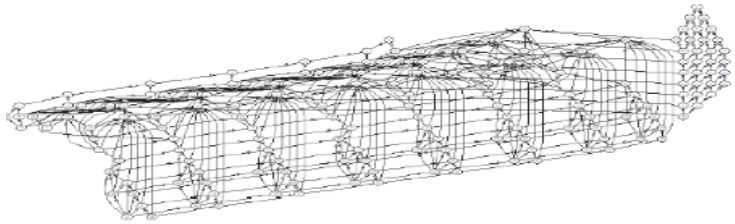
Bayesian Network

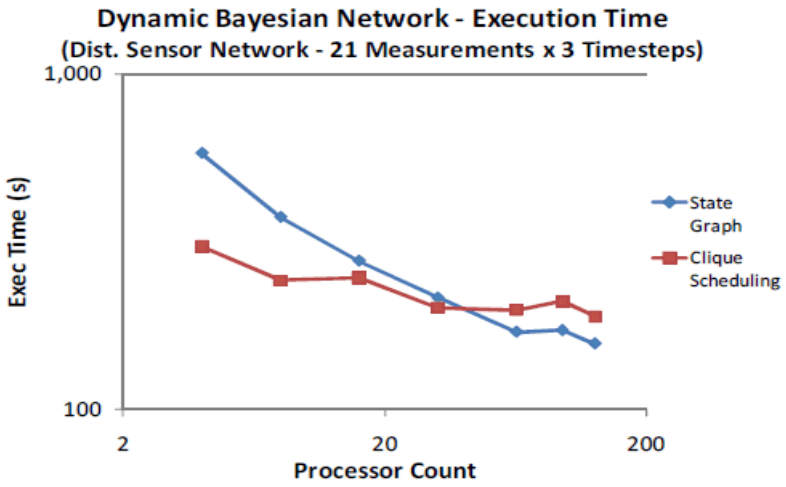
Goal: Atmospheric Radiation Measurement (ARM) climate research facility provides *24x7 continuous field observations* of cloud, aerosol and radiative processes. **Graphical models** can automate the validation with improvement efficiency and performance.



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







Bayesian Network

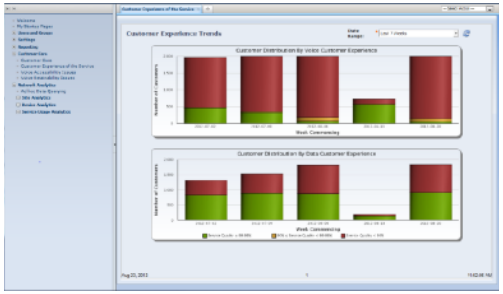
* 3 timesteps * 63 variables
* 3.9 avg states * 4.0 avg indegree
* 16,858 CPT entries

Junction Tree

* 67 cliques
* 873,064 PT entries in cliques

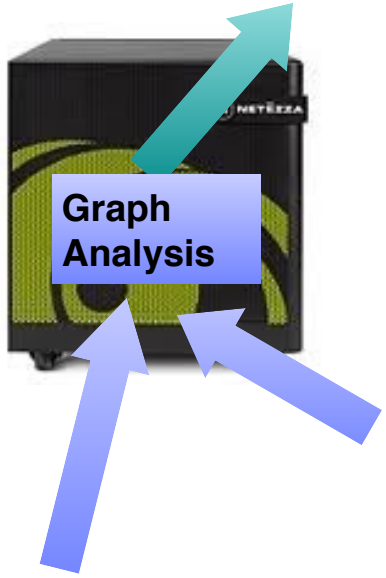
Use Case 14: Cellular Network Analytics in Telco Operation

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

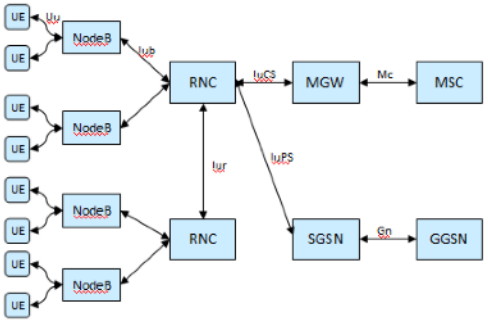


Network load level report

- 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



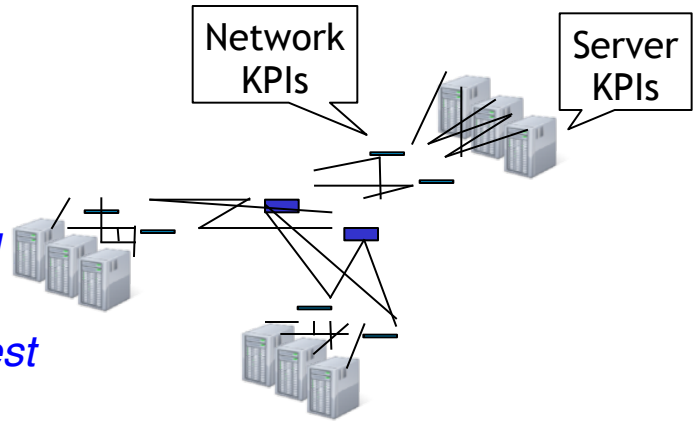
Network topology



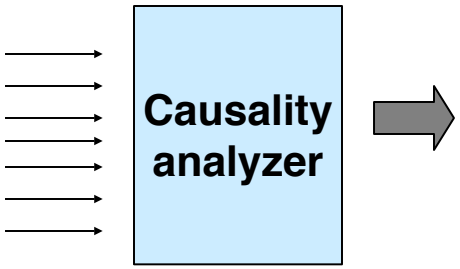
Use Case 15: Monitoring Large Cloud

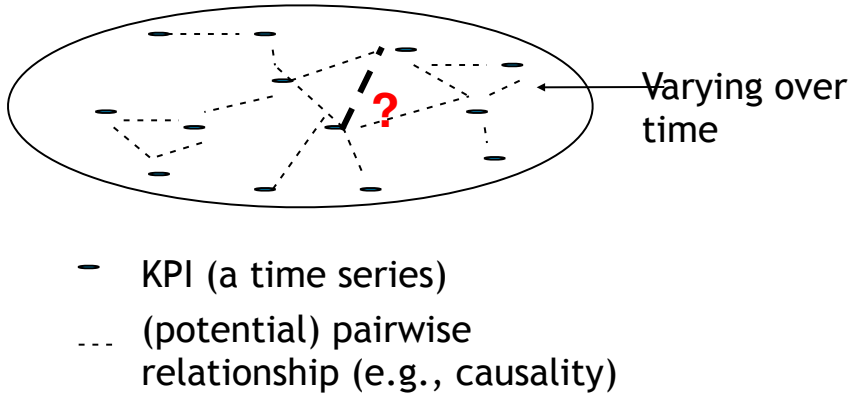
Goal: *M*onitoring technology that can track the time-varying state (e.g., causality relationships between KPIs) of a large Cloud when the processing power of monitoring system cannot keep up with the scale of the system & the rate of change

- *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)*

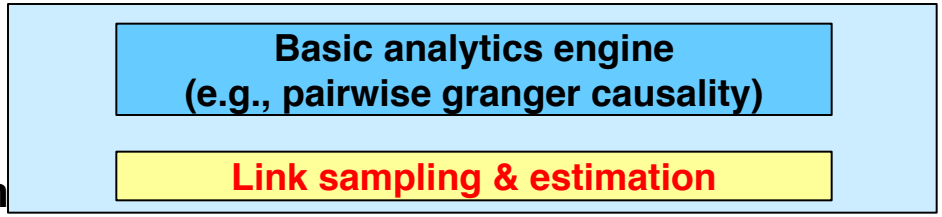


KPI time series (e.g., server performance/load, network performance/load)



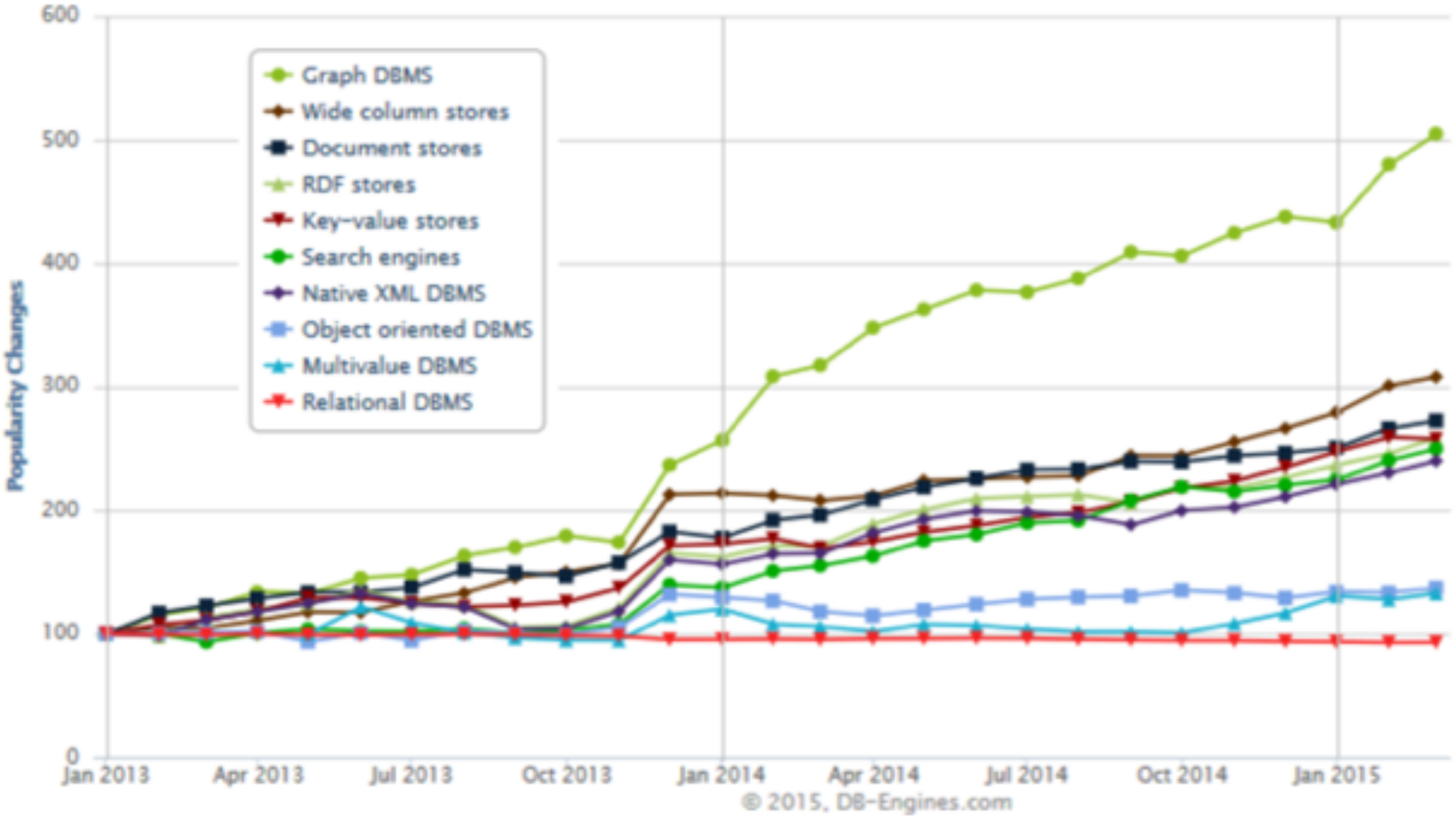


Our approach: Probabilistic monitoring via sampling & estimation

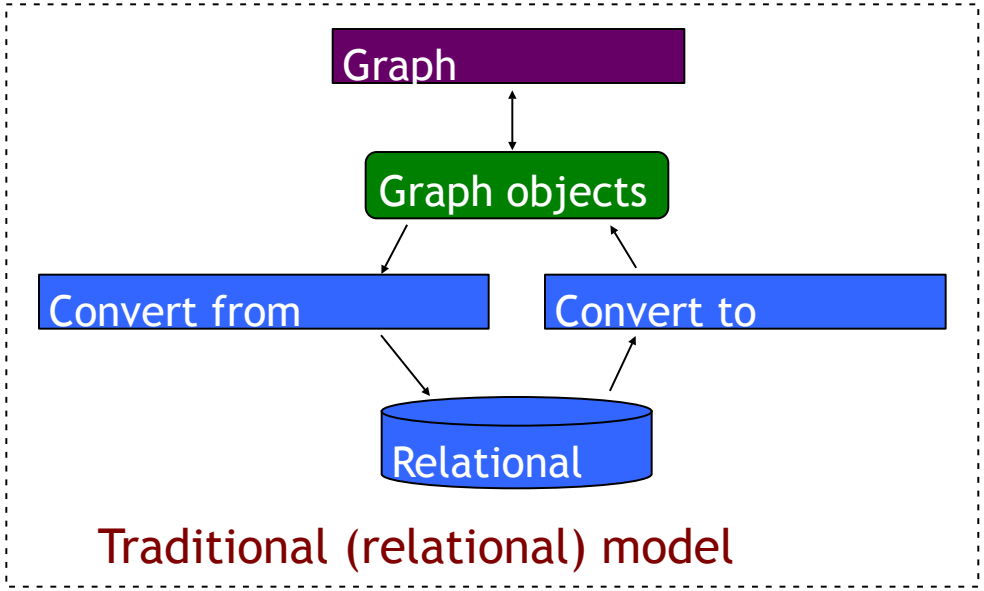


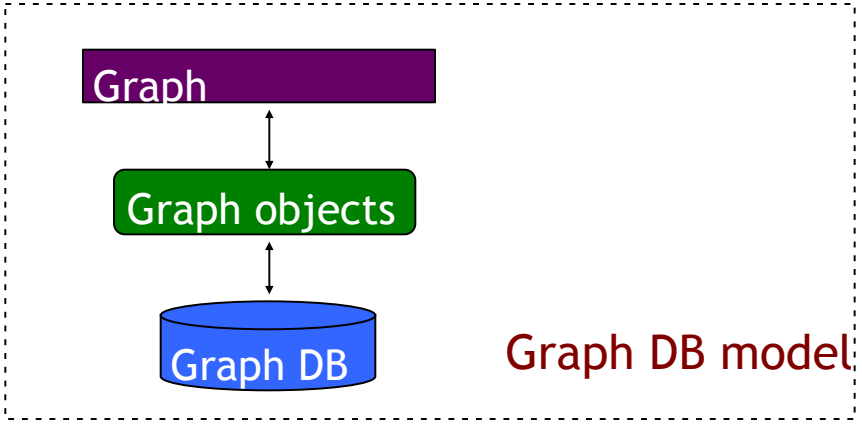
Select KPI pairs (sampling)→ Test link existence → Estimate unsampled links based on history → Overall graph

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

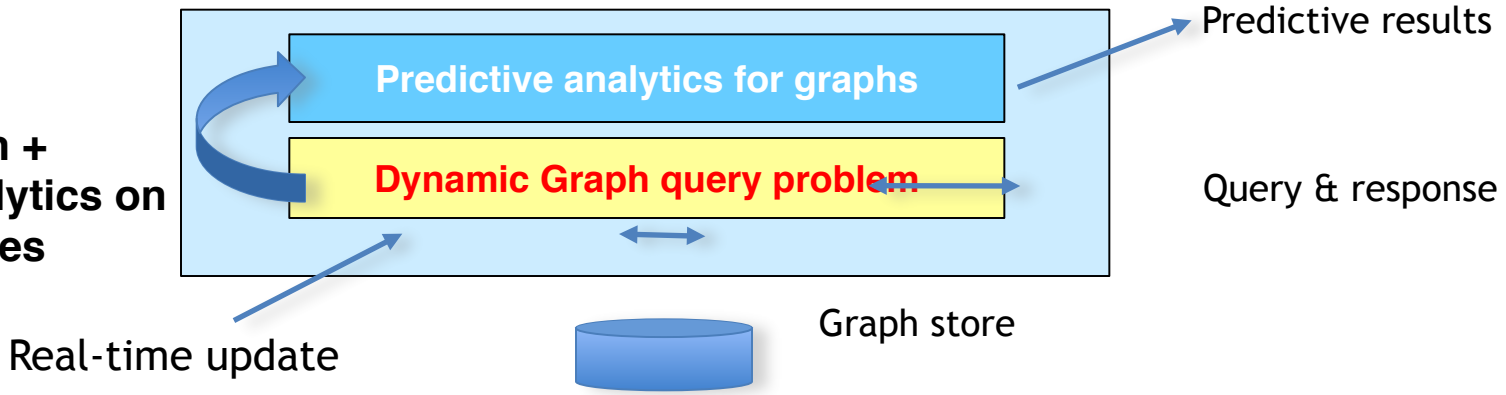
Use Case 17: Smart Navigation Utilizing Real-time Road Information

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*

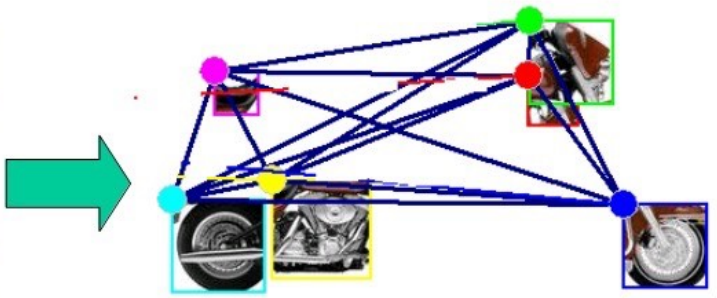
- *Dynamic graph algorithms implemented in System G provide **highly efficient graph query computation** (e.g. shorted path computation) on time-varying graphs (order of magnitudes improvement over existing solutions)*

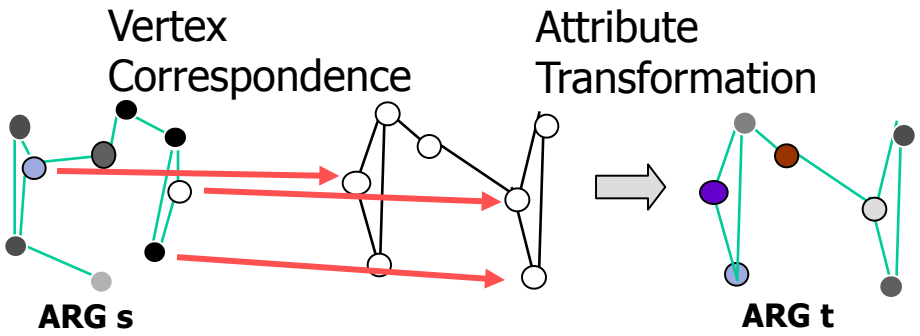
- *High-throughput **real-time predictive analytics** on graph makes it possible to estimate the future traffic condition on the route to make sure that the decision taken now is optimal overall*

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

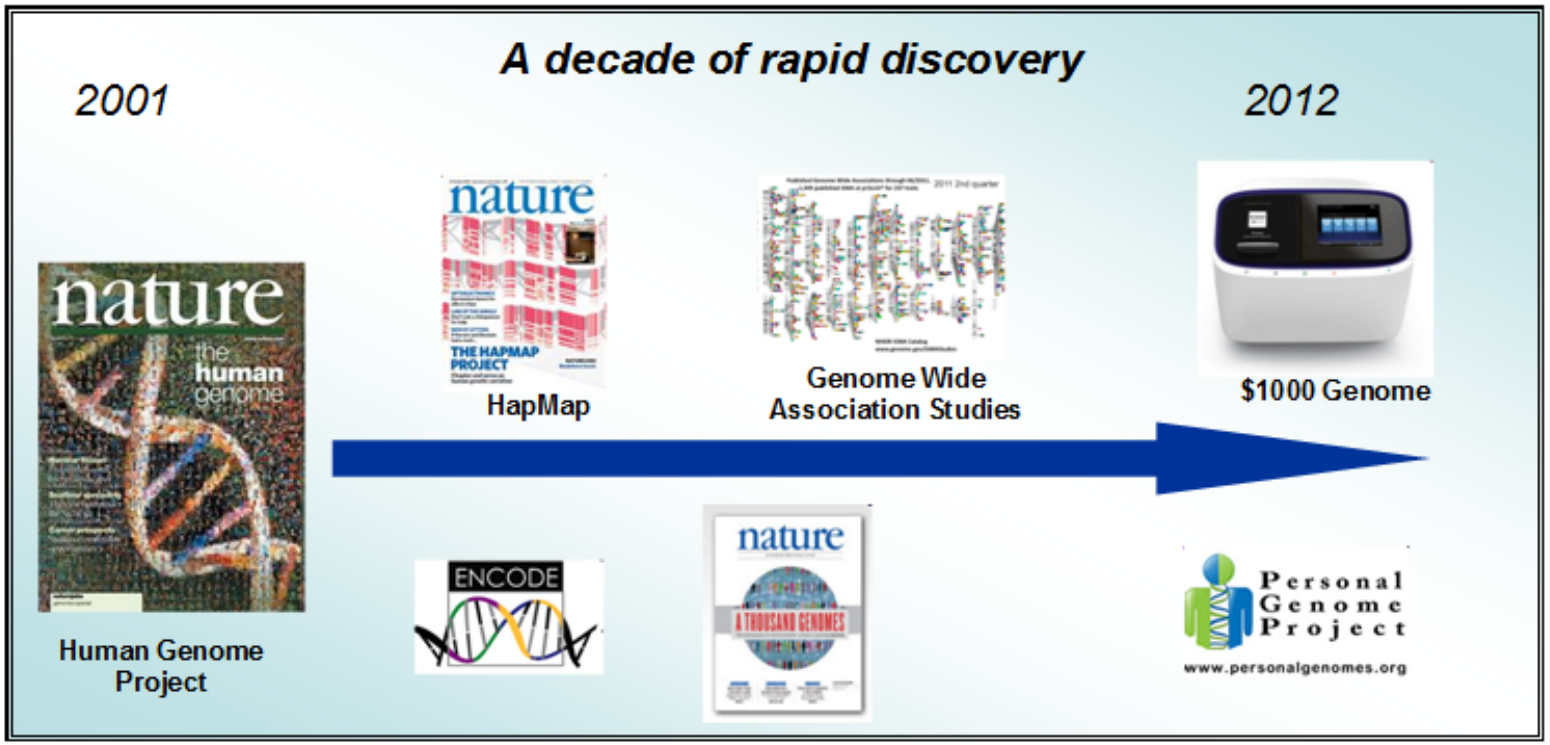
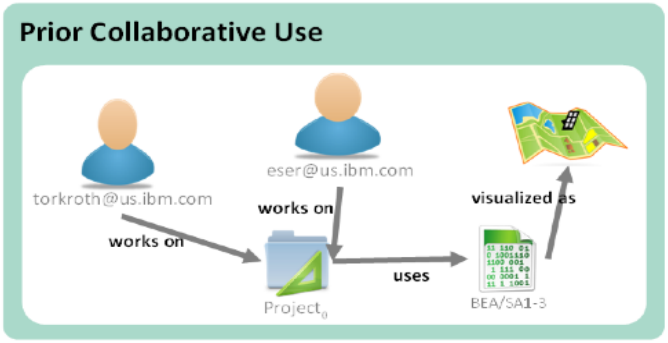
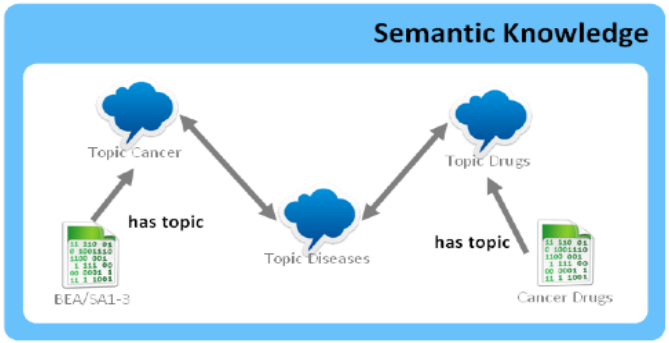
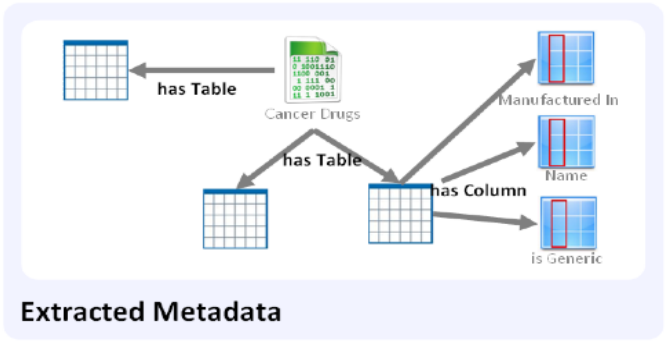


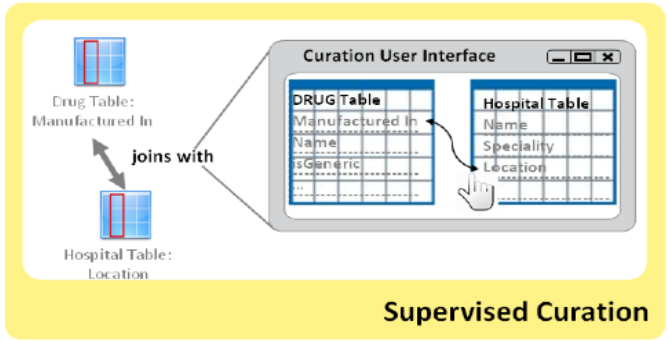
Figure 1: Since the Human Genome Project, various projects have started to reveal the mysteries of genomes and the $1000 Genome is almost reality.

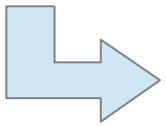
Use Case 20: Data Curation for Enterprise Data Management

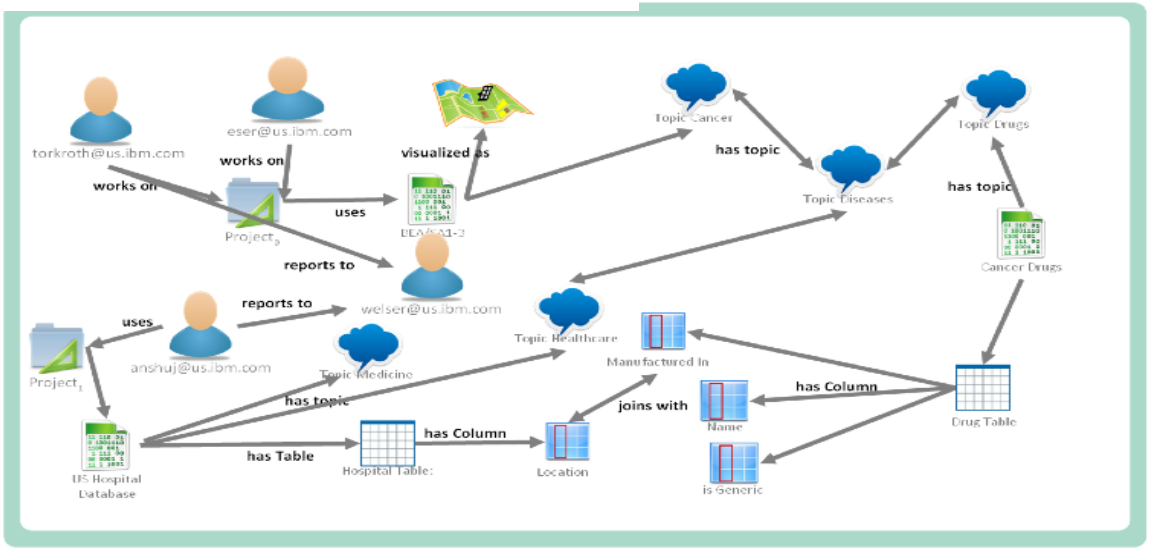




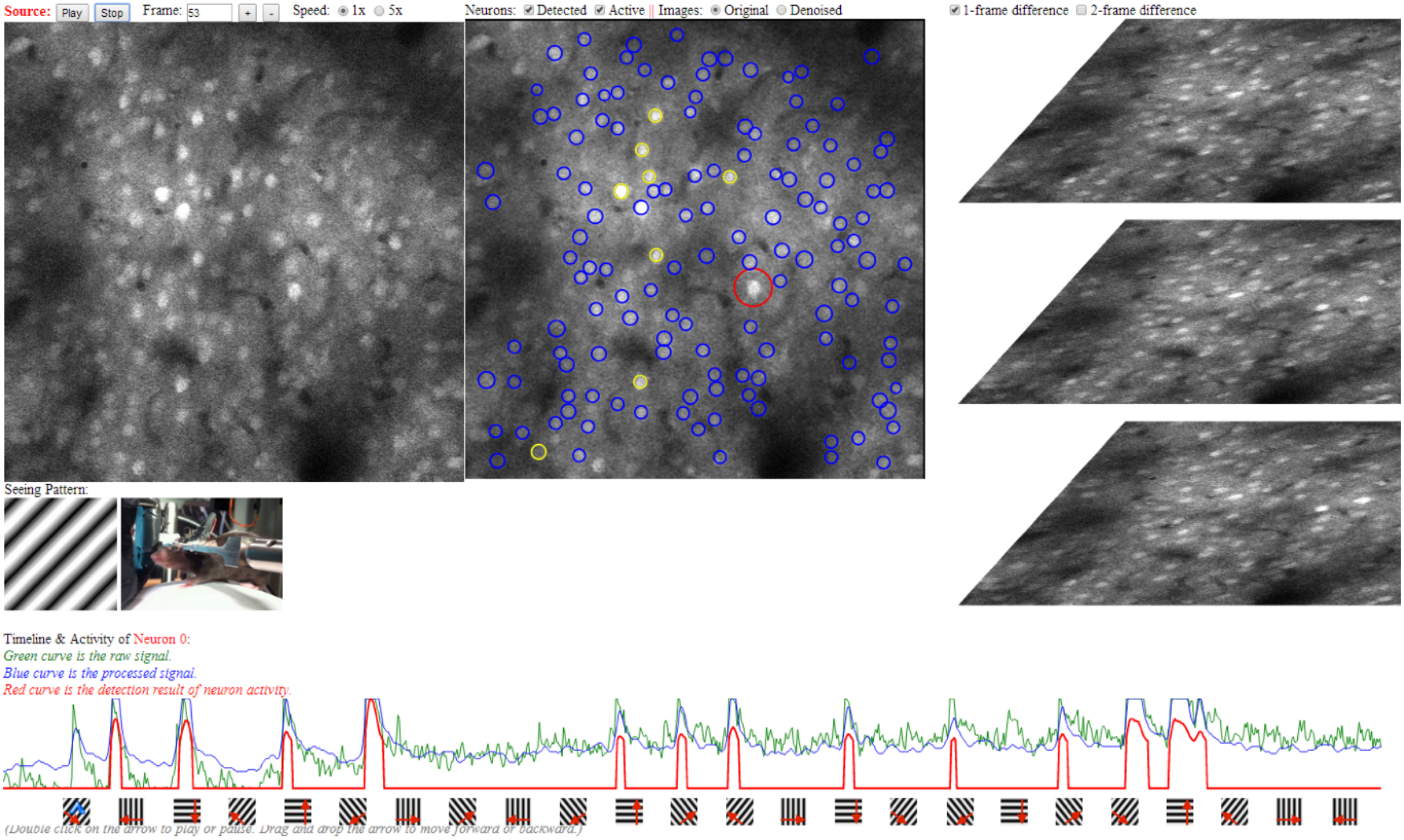








Use Case 21: Understanding Brain Network



Use Case 22: Planet Security

- Big Data on Large-Scale Sky Monitoring

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 benefi 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 engineerin 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 yo approach the right thing to do? Did you consider alternative approaches? Have your 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 a solved? Suggest ideas for future extensions or new applications.
- **Writing / Formatting (5%)** Is your paper clearly written and nicely formatted?