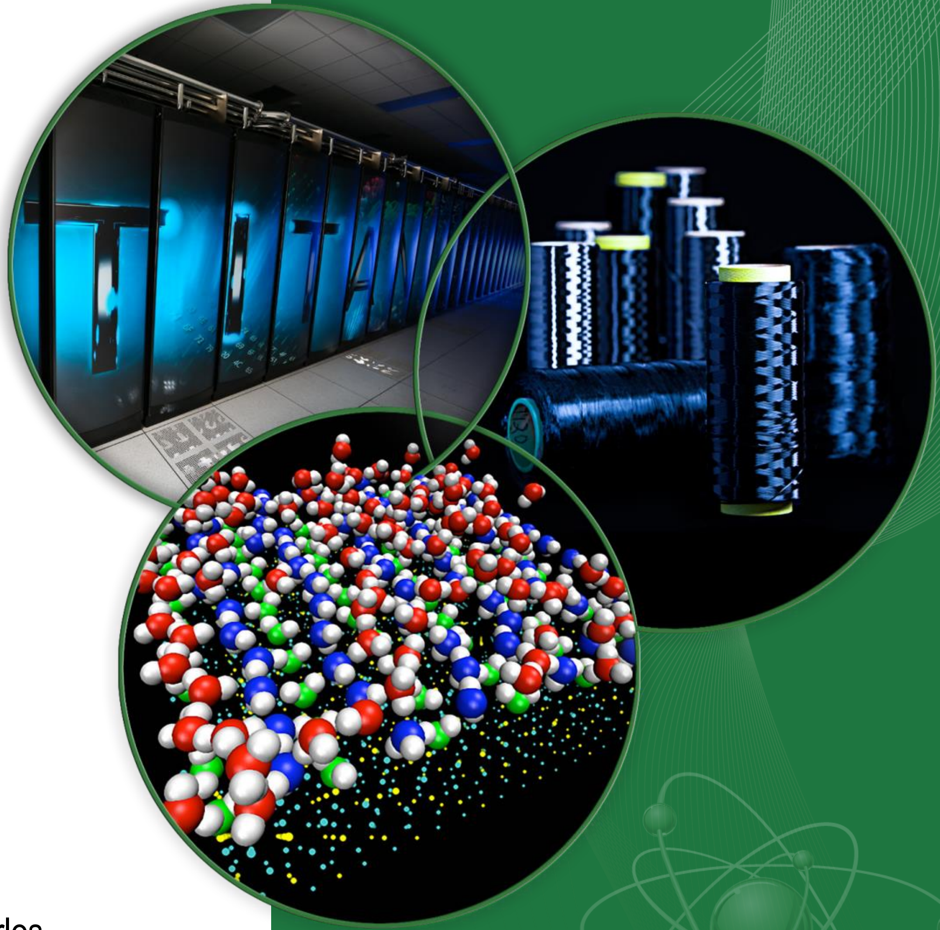
A Graph Mining "App-Store" for Urika-GD

Rangan Sukumar (Email: sukumarsr@ornl.gov)

ORNL Team: Matt Lee, Seung-Hwan Lim, Regina Ferrell, Arjun Shankar



Team of interns that have contributed to the work :

Tyler Brown[1 (HERE)], Keela Ainsworth[1(HERE)], Larry Roberts[2(SULI)], Yifu Zhao[3(ORSS)], Nathaniel Bond[4(SULI)], Carlos del-Castillo-Negrette[5(SULI)], Katie Senter[6(RAMS)] , Seokyong Hong[7(Go !)] and Gautam Ganesh[8(HERE)]

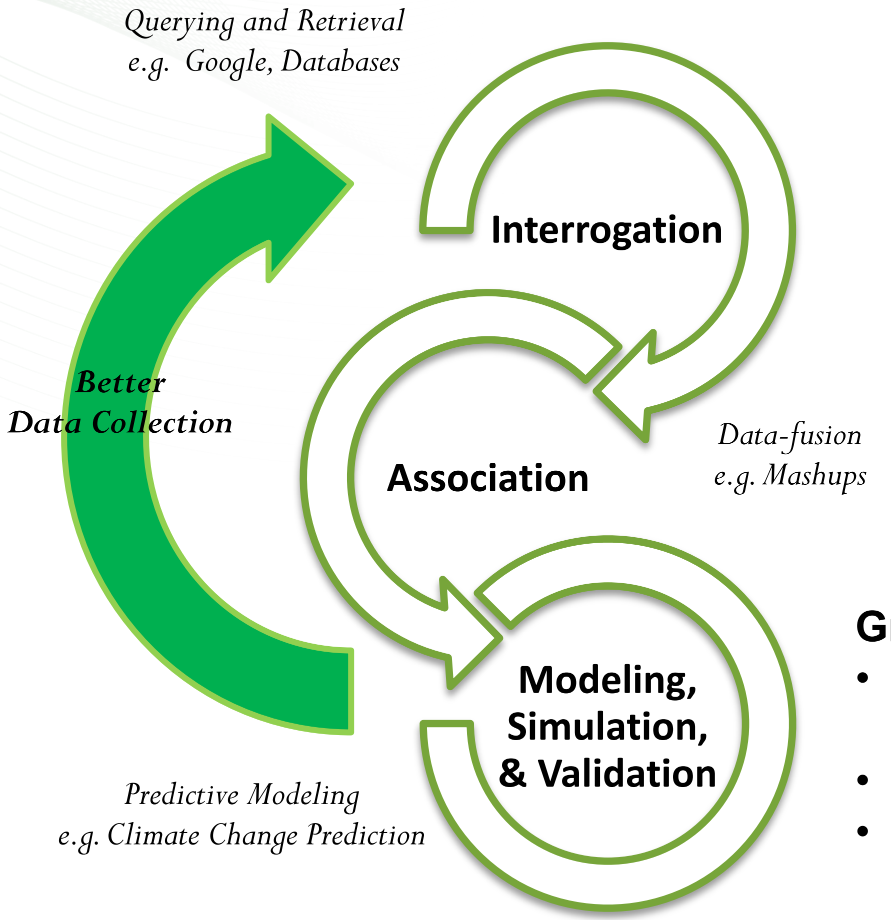
1. University of Tennessee, Knoxville, 2 Tennessee Tech University, Cookeville, 3. Denison University, Ohio, 4. Cedarville University, 5. Yale University, 6. Penn State University, 7. North Carolina State University, University of Texas, Dallas

ORNL is managed by UT-Battelle
for the US Department of Energy

Oak Ridge National Laboratory

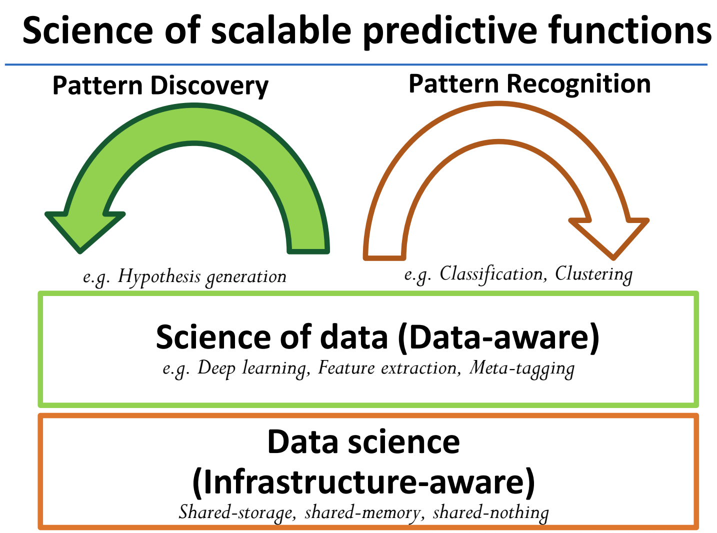
Machine Learning: Graph Computing Interest

The Lifecycle of Data-Driven Discovery



S.R. Sukumar, "Data-driven Discovery: Challenges at Scale", *in the Proc. of the Big Data Analytics: Challenges and Opportunities Workshop in conjunction with ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis (Super Computing)* , November 2014.

The Process of Data-Driven Discovery

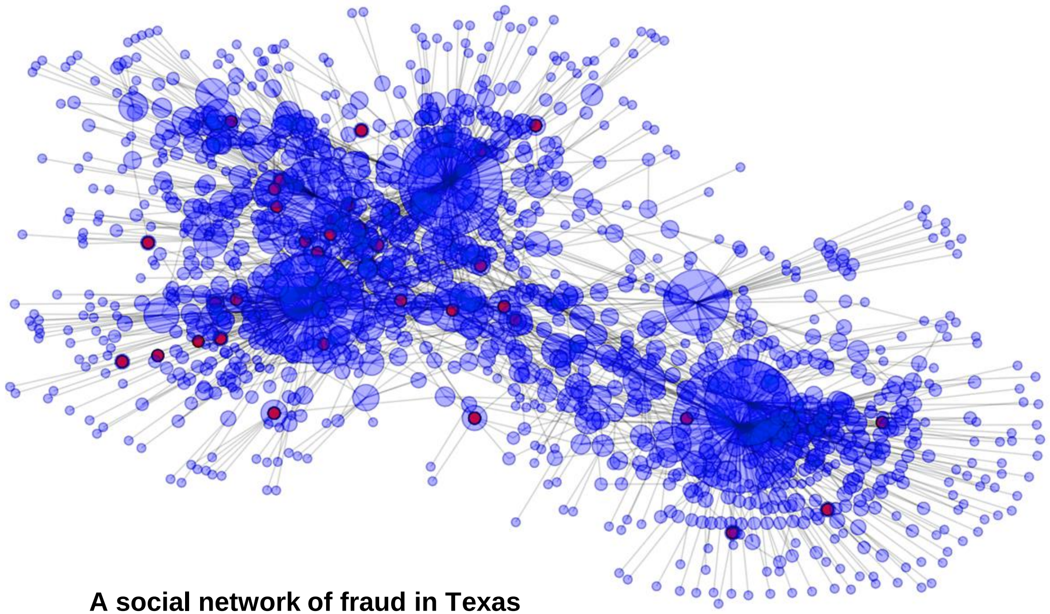


Graph Computing…

- Supports discovery by interrogation, association and predictive modeling from structured and unstructured data
- Supports discovery with evolving knowledge and incremental domain hints
- Supports exploratory and confirmatory analysis
  - Data and meta-data integrated analytics
  - Flexible data structure seamless to growth while avoiding analytical artifacts

Why ? Discovery from Big Data "Graphs"

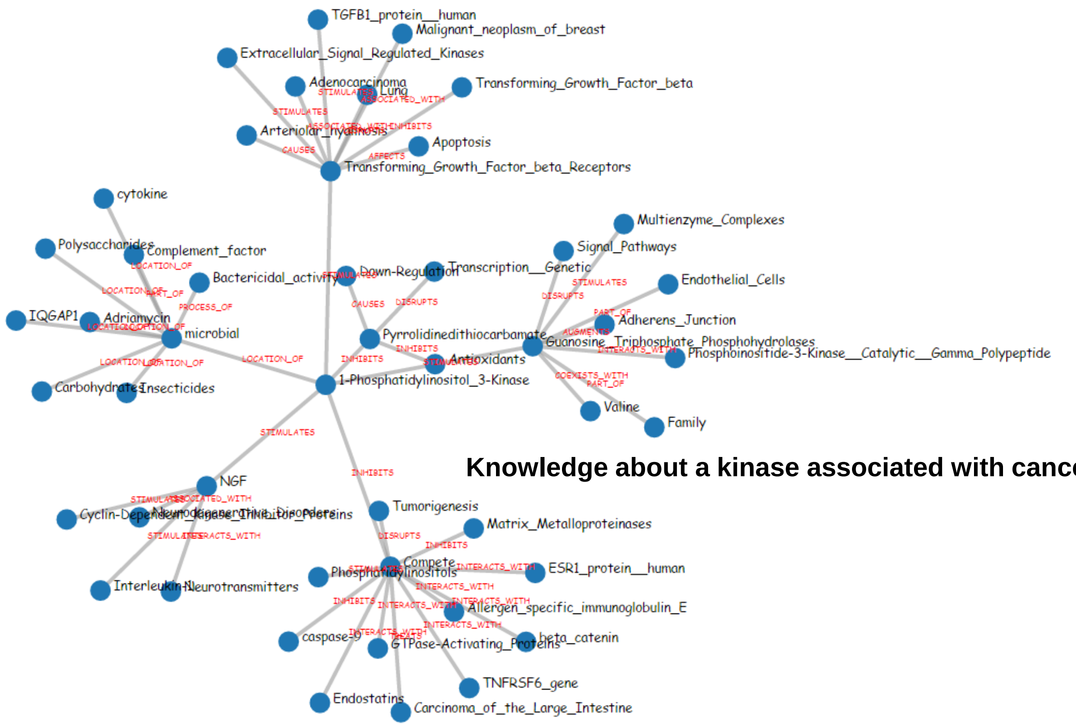
Motivation: Fraud Detection in Healthcare



A social network of fraud in Texas

V. Chandola, S.R. Sukumar and J. Schryver, "Knowledge Discovery from Massive Healthcare Claims Data", in the Proc. of the 19th ACM SIGKDD Conference on Knowledge Discovery, 2013

Motivation: Knowledge Discovery from Literature



Knowledge about a kinase associated with cancer

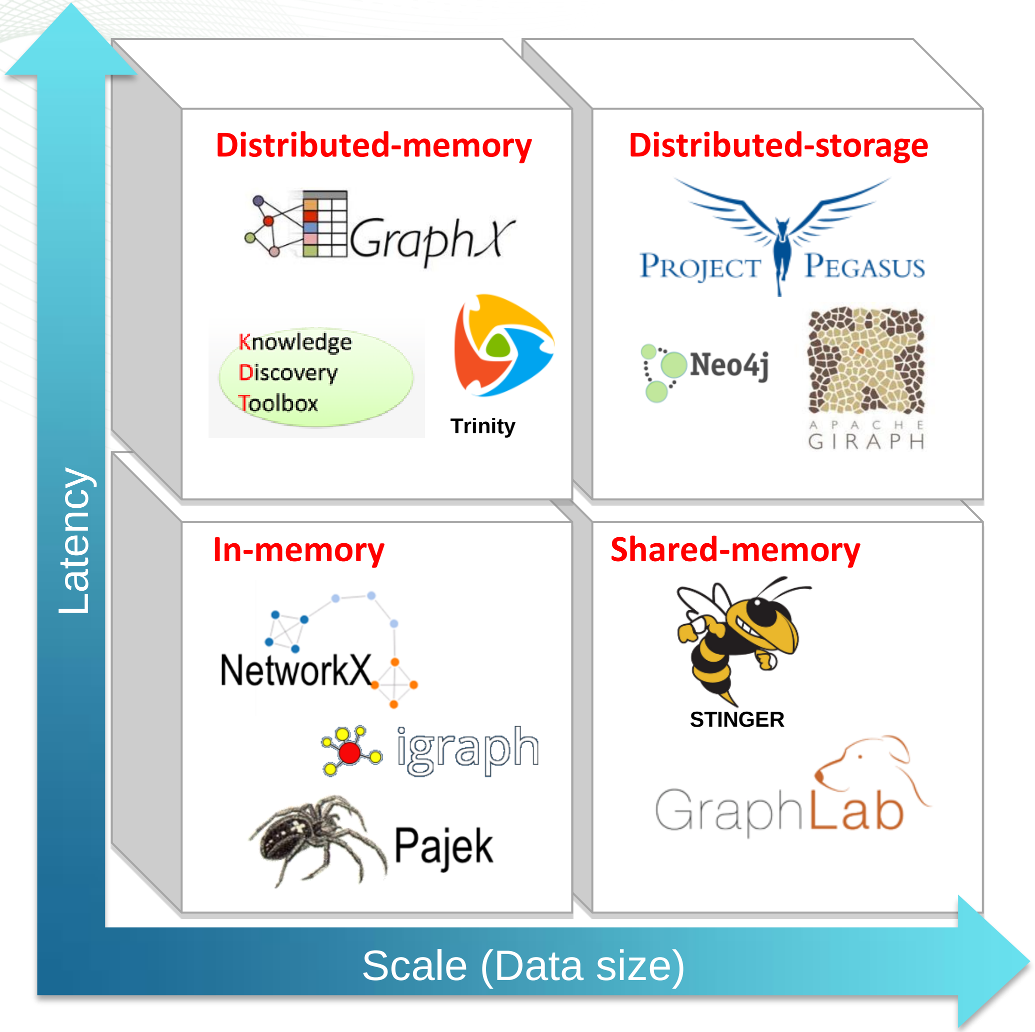
S.M. Lee, S- H. Lim, T.C. Brown, *S. R. Sukumar*, "Graph mining meets the Semantic Web", in the Proc. the Data Engineering meets the Semantic Web Workshop in conjunction with International Conference on Data Engineering, 2015.

Given a few examples of fraud (important activity), can we
(i) Automatically discover patterns typically associated with suspicious activity?
(ii) Extrapolate such high-risk patterns for investigation and fraud prevention?

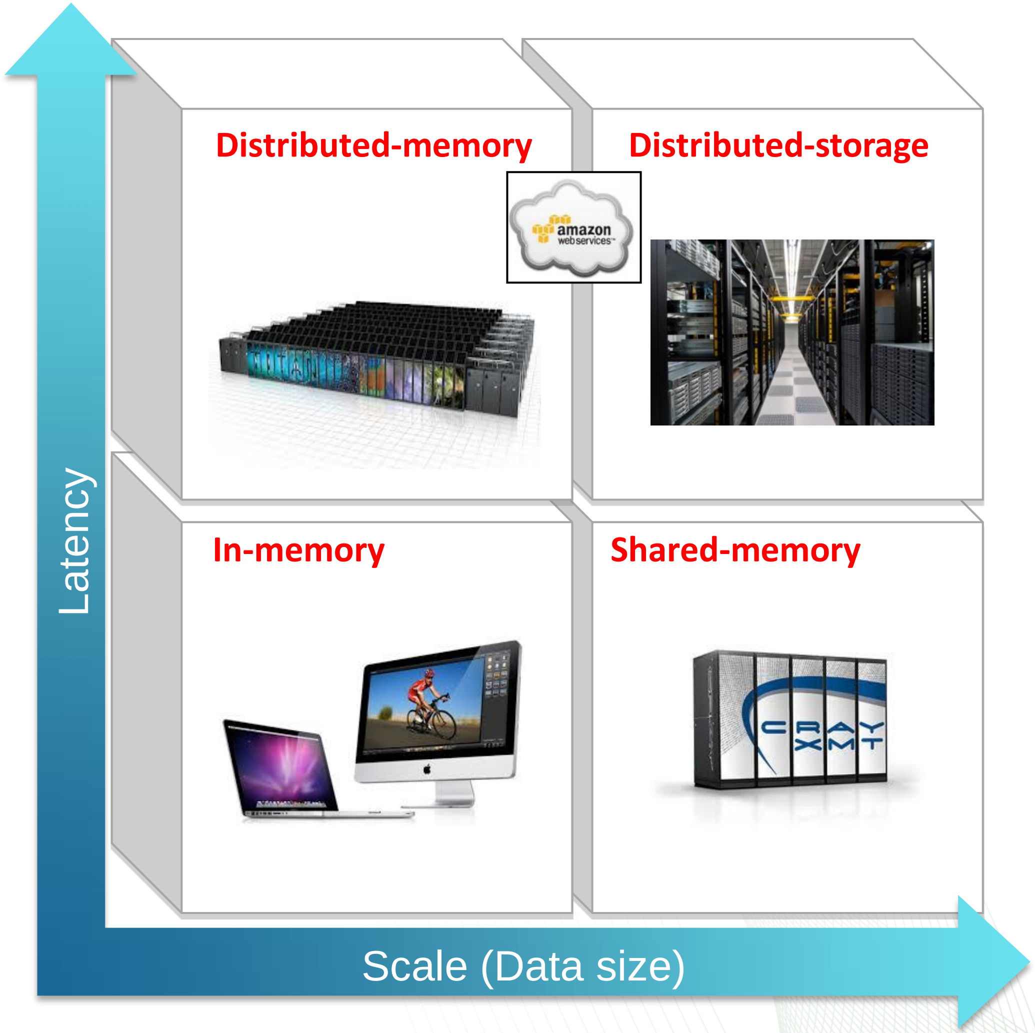
Given a knowledgebase and new clinical data/experiments, can we
(i) Find "novel" patterns of interest?
(ii) Rank and evaluate the patterns for significance?

Graph Computing at Scale: Infrastructure

Software Tools

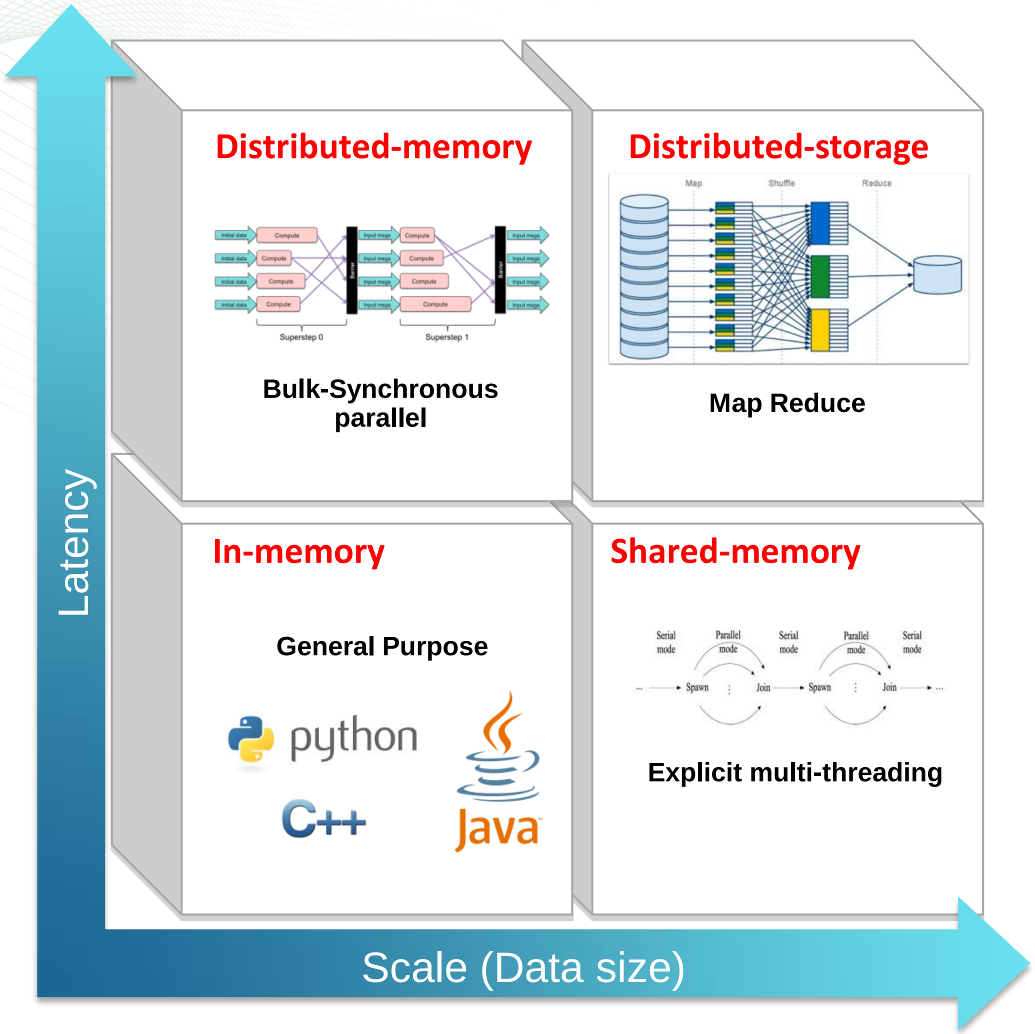


Hardware

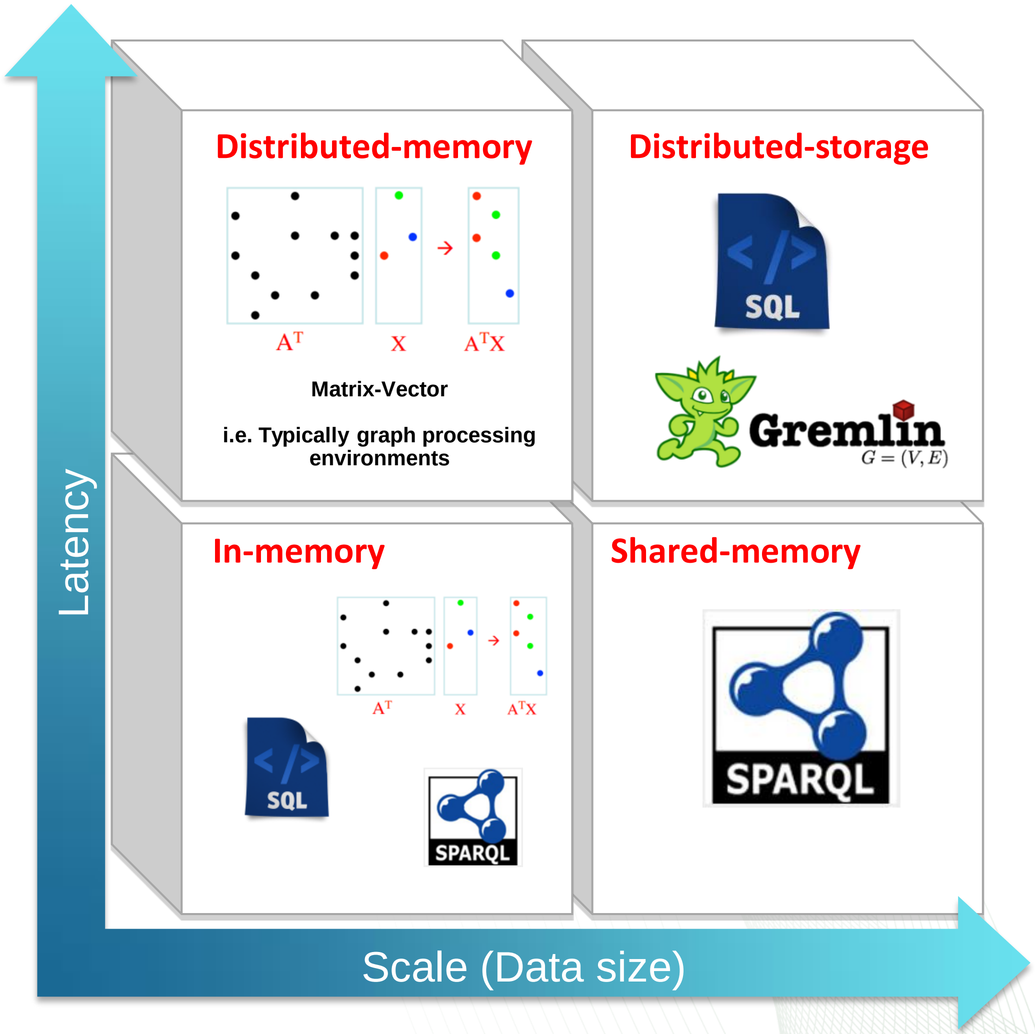


Graph Computing at Scale: Infrastructure

Programming Model

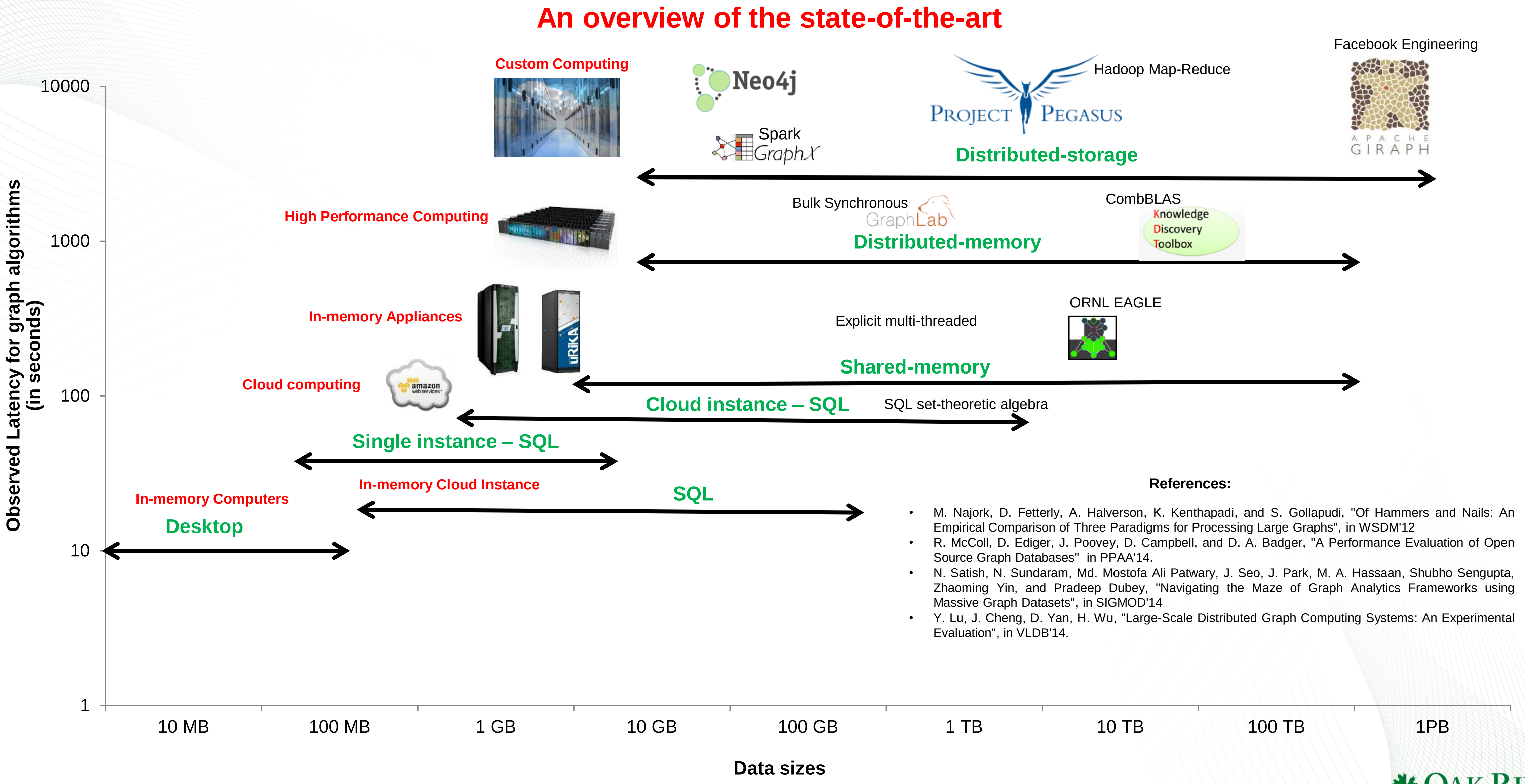


Query Language



OAK RIDGE National Laboratory

Graph Computing at Scale: Literature



References:

- M. Najork, D. Fetterly, A. Halverson, K. Kenthapadi, and S. Gollapudi, "Of Hammers and Nails: An Empirical Comparison of Three Paradigms for Processing Large Graphs", in WSDM'12
- R. McColl, D. Ediger, J. Poovey, D. Campbell, and D. A. Badger, "A Performance Evaluation of Open Source Graph Databases" in PPAA'14.
- N. Satish, N. Sundaram, Md. Mostofa Ali Patwary, J. Seo, J. Park, M. A. Hassaan, Shubho Sengupta, Zhaoming Yin, and Pradeep Dubey, "Navigating the Maze of Graph Analytics Frameworks using Massive Graph Datasets", in SIGMOD'14
- Y. Lu, J. Cheng, D. Yan, H. Wu, "Large-Scale Distributed Graph Computing Systems: An Experimental Evaluation", in VLDB'14.

OAK RIDGE National Laboratory

The Opportunity at ORNL

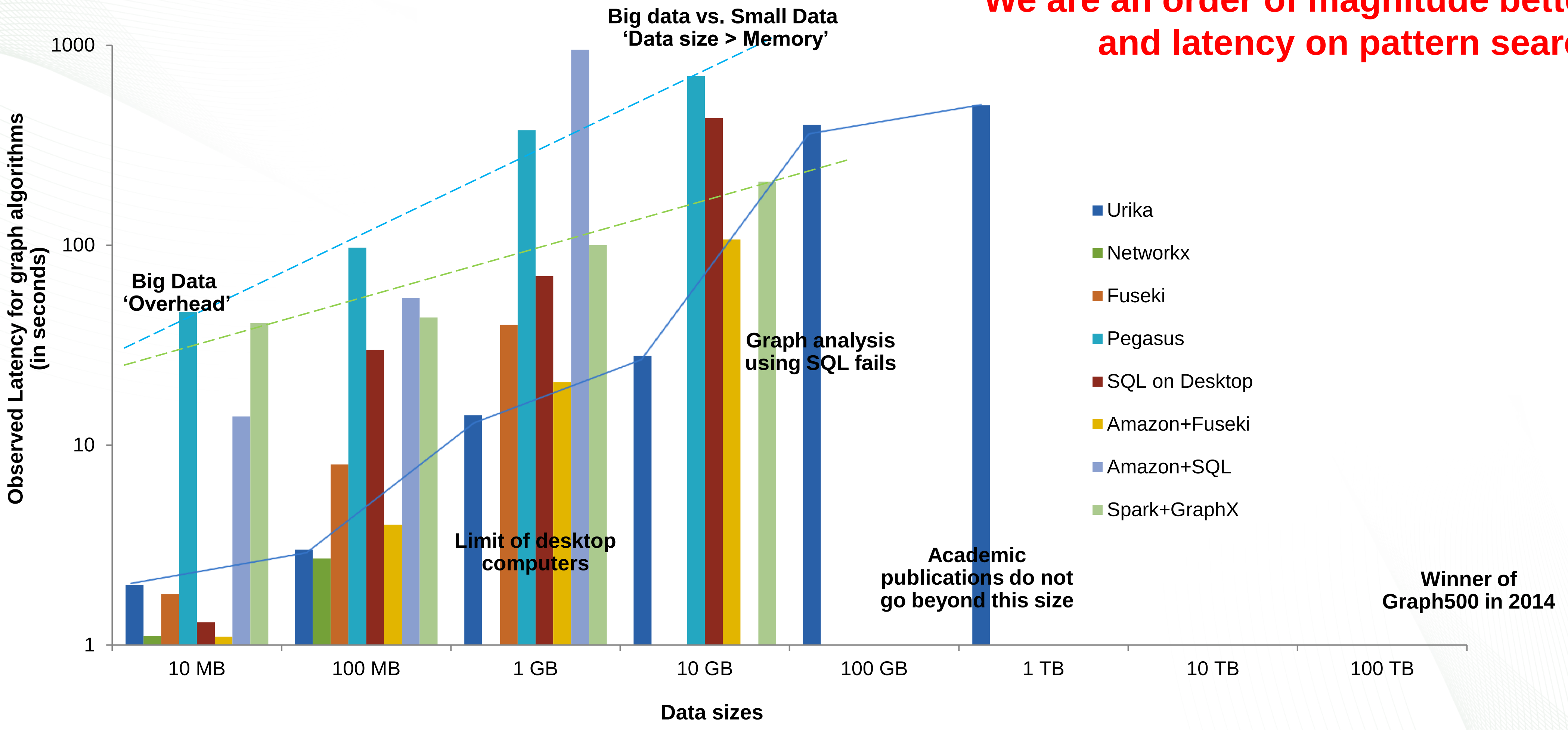
ORNL Resources

	Titan	Apollo	CADES (Cloud)
Discovery Approach	Modeling and Simulation	Association	Querying, Prediction
Architecture	Shared-compute	Shared-memory	Shared-storage
Scalability	Compute (# of cores)	Horizontal (# of datasets)	Vertical (# of rows)
Algebra	Linear	Relationship	Set-theoretic
Challenge (Pros)	Resolution	Heterogeneity	Cost
Challenge (Cons)	Dimensionality	Custom Solution	Flexibility
Leadership	#2 in the world (2013)	1 of 15 installs (2013)	--
User-interface	OpenMP, MPI, CUDA	SPARQL	SQL

S- H. Lim, S.M. Lee, G. Ganesh, T.C. Brown and S.R. Sukumar, "Graph processing platforms at scale: practices and experiences, under review to the IEEE International Symposium on Performance Analysis of Systems and Software, 2014.

Graph Computing at Scale: Pattern Search

We are an order of magnitude better on size and latency on pattern search.



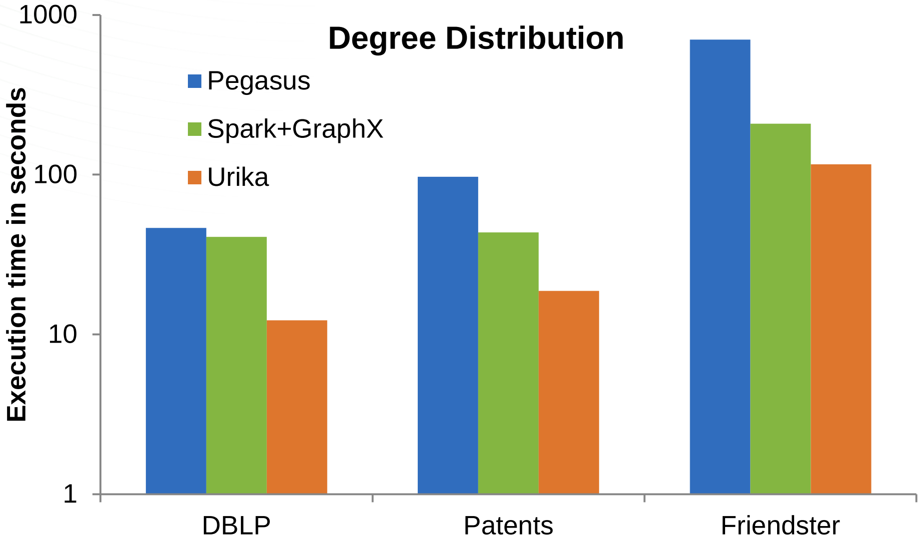
Rule of thumb: Any query that takes longer than 45 seconds (on ~ TBs) is bad code !

Graph Computing at Scale: Data Science

What is the best "programming-paradigm" for graph computing?

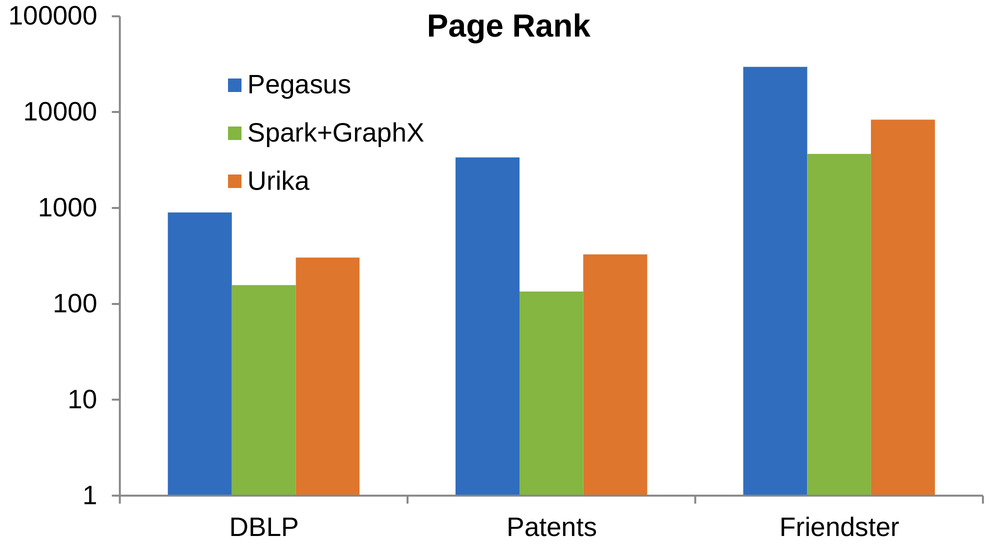
Scalability: In-disk vs. In-memory

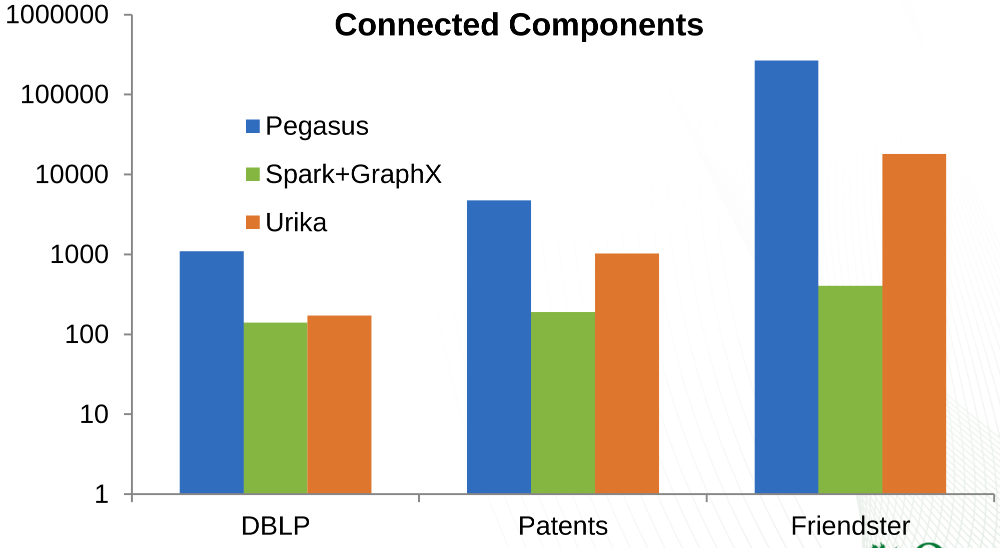
Map-Reduce vs. Spark vs. SPARQL



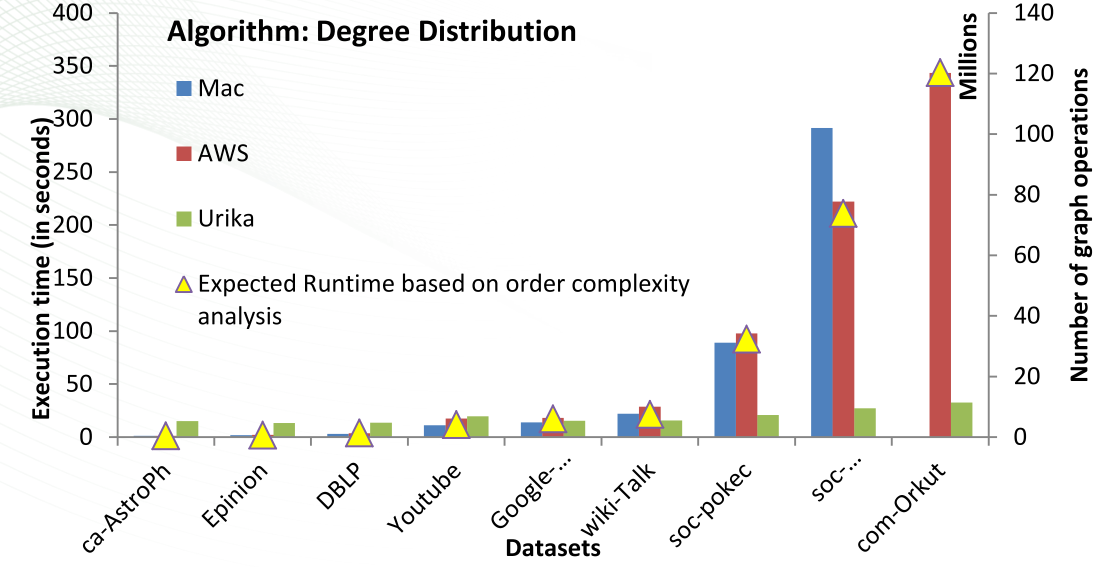
Pegasus – ICDM Best Paper 2010

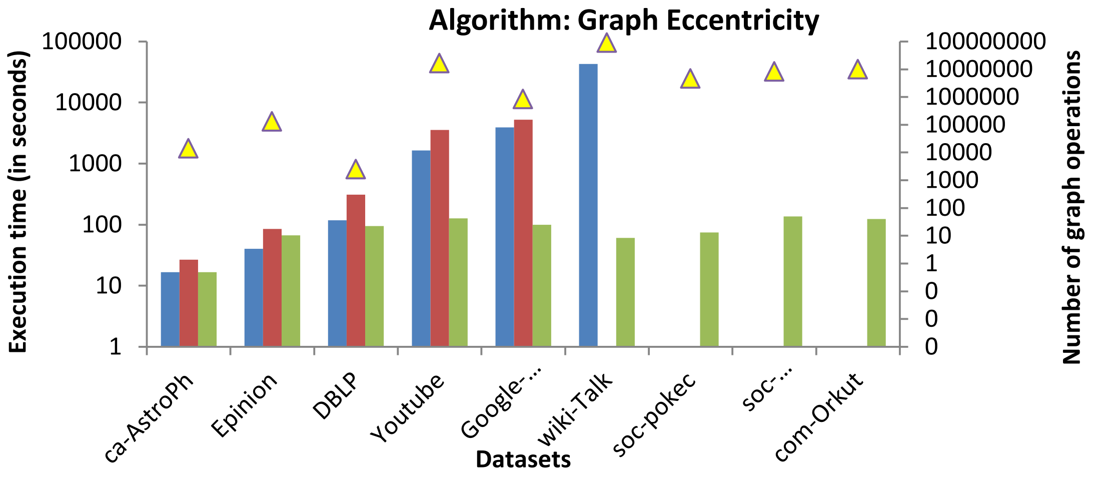
Spark+GraphX – USENIX NSDI Best Paper 2014

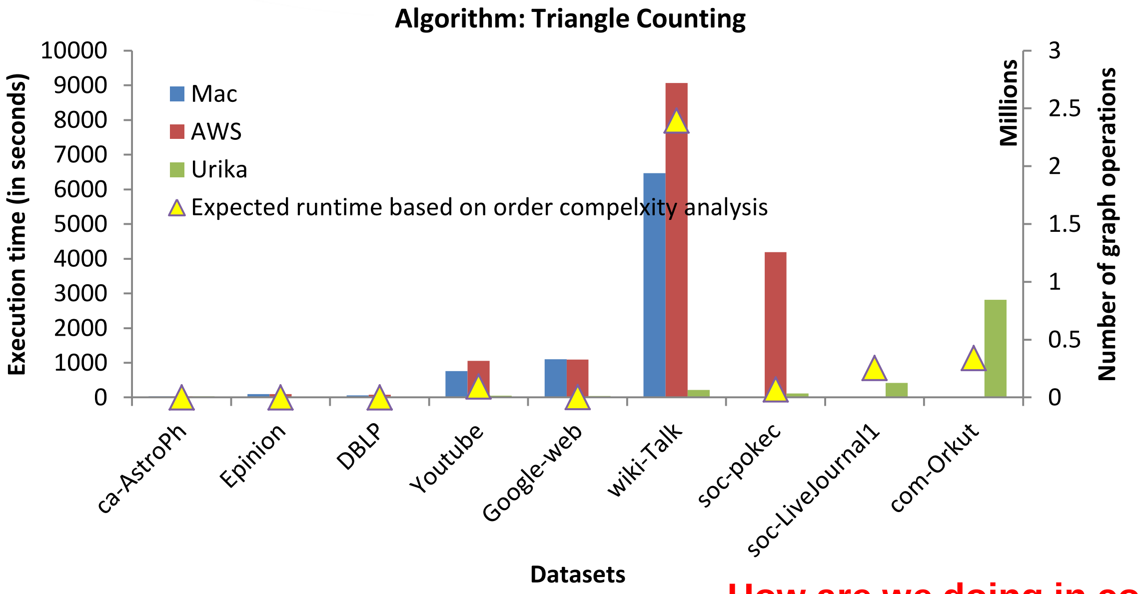


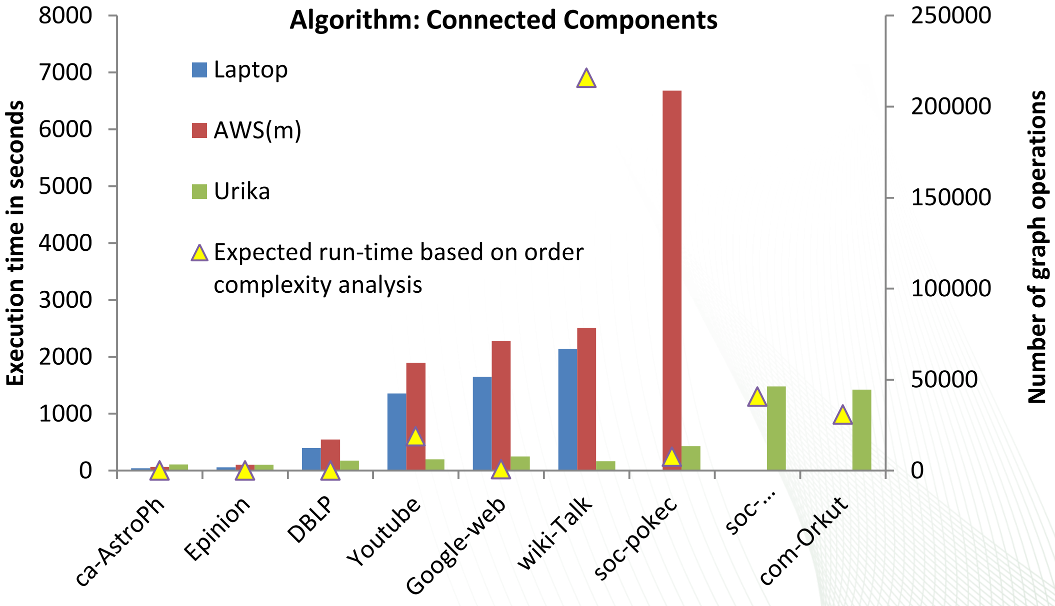


Graph Computing at Scale: Data Science







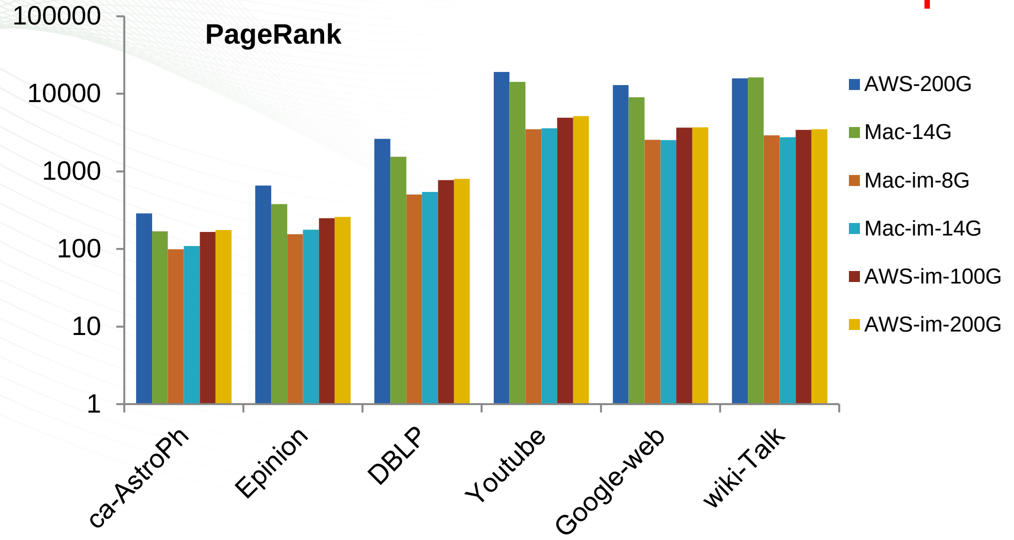


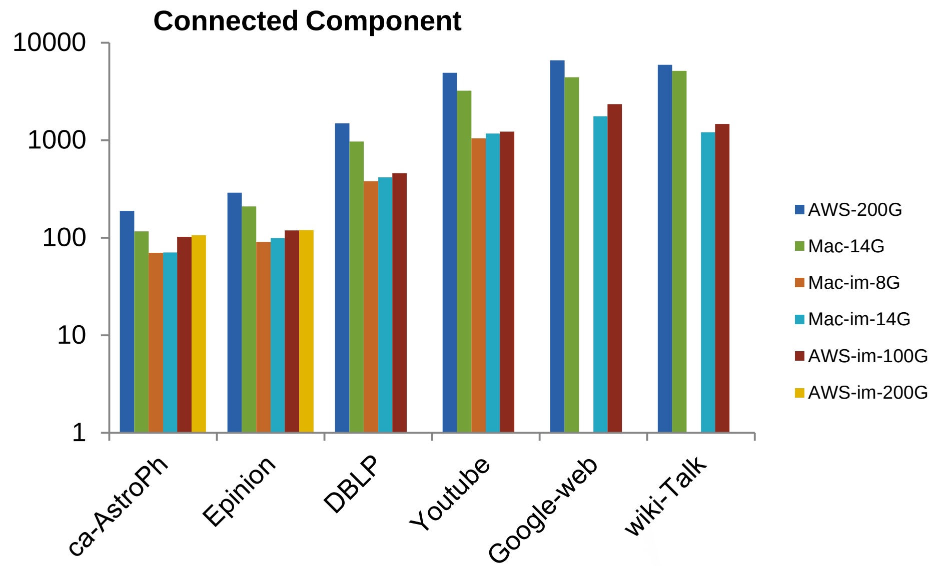
How are we doing in comparison with Amazon services?

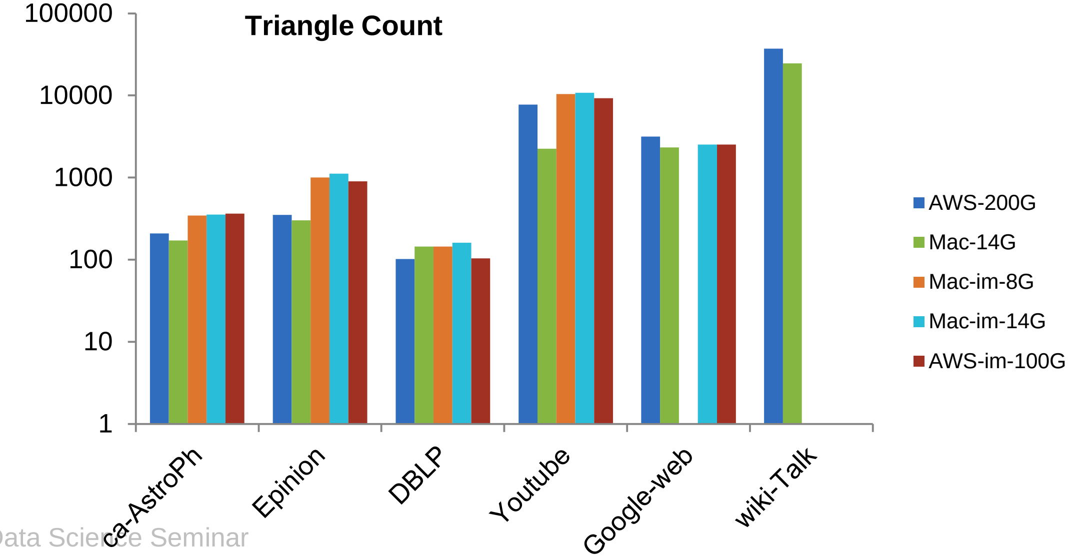
OAK RIDGE National Laboratory

Graph Computing at Scale: Algorithms Benchmark

Desktop vs. Database vs. Cloud







Lessons learned…

- Performance (feasibility) of graph algorithms are a function of the architecture and data (not just size).
  - Depends on space and time complexity of algorithm
- One size does not fit all.
- With graph analysis, scale-up does not guarantee speed-up.
  - Needs smarter re-design of algorithms.

OAK RIDGE
National Laboratory

Graph Computing at Scale: Summary

We now have one machine that is able to do both pattern search and pattern mining within "reasonable" time constraints

- Compared to CMU Pegasus (2010) – ten times speed-up.
- Compared to Berkley GraphX (2014) on a select few algorithms– 2 to 5 times speed-up.
- Compared to Desktops – 1000 times larger size for similar latency
- First of its kind handling "heterogeneous-graphs" with near real-time latency.
- First of its kind "SPARQL-based Graph-Theoretic Data Analysis" tools
  - Has huge potential with the W3C and LinkedData Community.
- Users with no knowledge of SPARQL (or linear algebra) can work with EAGLE on their domain-specific problems.

Eureka ! With Urika : ‘App’ Store

Framework of Knowledge Discovery for a future beyond the Big Data Era

PLUS

Programmatic-Python Login for Urika-like SPARQL End-points

Code Development

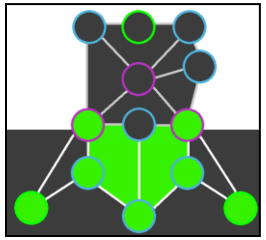
FELT

Flexible, Extract, Transform and Load Toolkit

Graph Creation

EAGLE-C

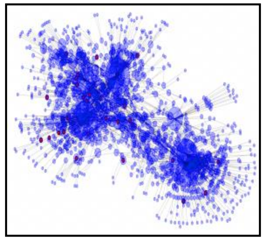
EAGLE ‘Is a’ algorithmic Graph Library for Exploratory-Analysis



Scalable Algorithms

GRAPH-IC

Graph- Interaction Console



Interactive Visualization

PAUSE

Predictive Analytics using SPARQL-Endpoints

Reasoning + Inference

KENODES

Knowledge Extraction using Network-Oriented Discovery Enabling System

Hypothesis Creation

Some parts are open-source @ https://github.com/ssrangan/gm-sparql

PLUS – Programmatic Python Login for Urika-like SPARQL Endpoints

- **What does PLUS do?**
  - Clone Urika-like developer environment
    - SPARQL end-point vs. SQL end-point
  - Deploy code in developer environment at scale on Urika with minimal changes (1 line of code change)
  - JDBC-like connection to graph database + Urika Firewalls
  - Provides programmatic API for iterative algorithms
  - Software, parallelism and query optimization unit test environment

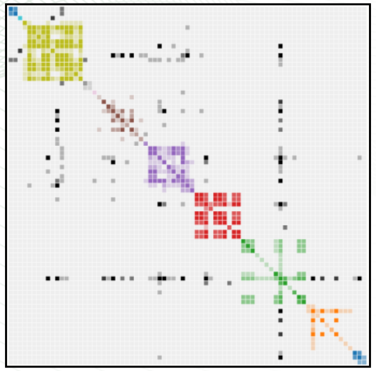
SPARQL Wrapper

Ruby

PyScripter IDE

FELT – Flexible Extract Transform and Load Toolkit

Adjacency Matrices



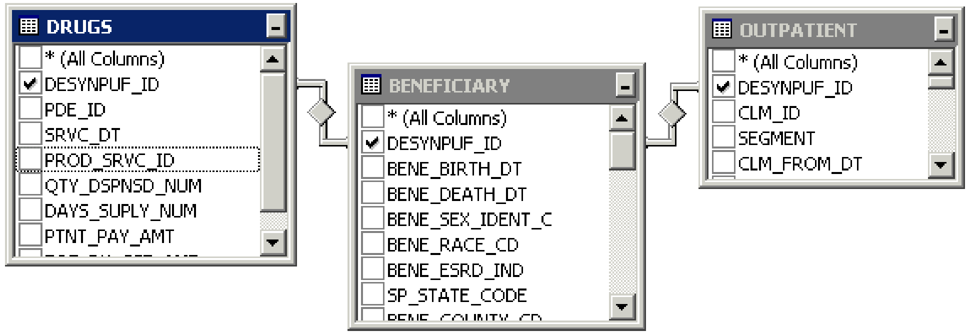
Edgelists

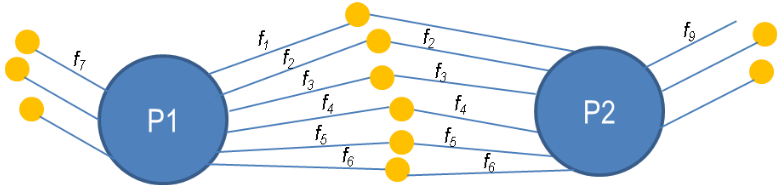
P	H
P1	H1
P1	H1
.	H1
.	.
P3	H2
.	.

Flat files

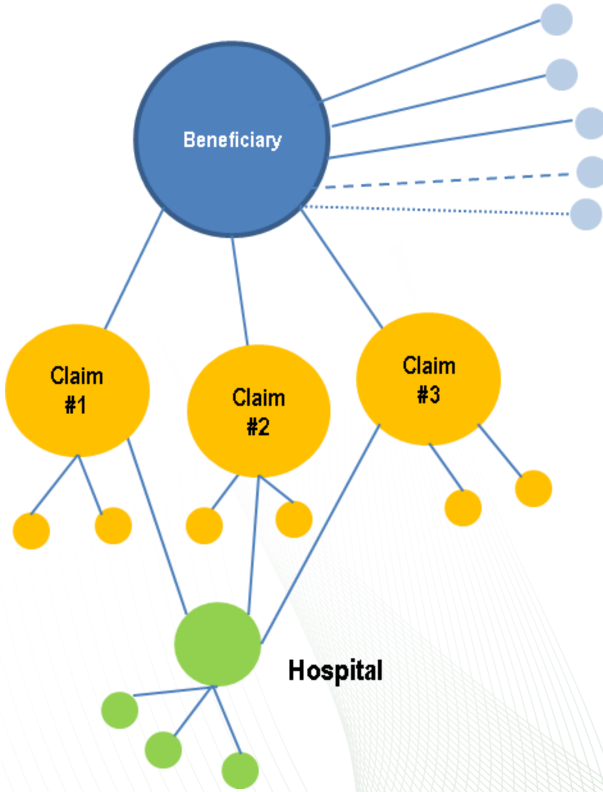
Patient	f_1	f_2	f_3	..	..	f_{100}
P1						
…						
Pn						

Relational databases



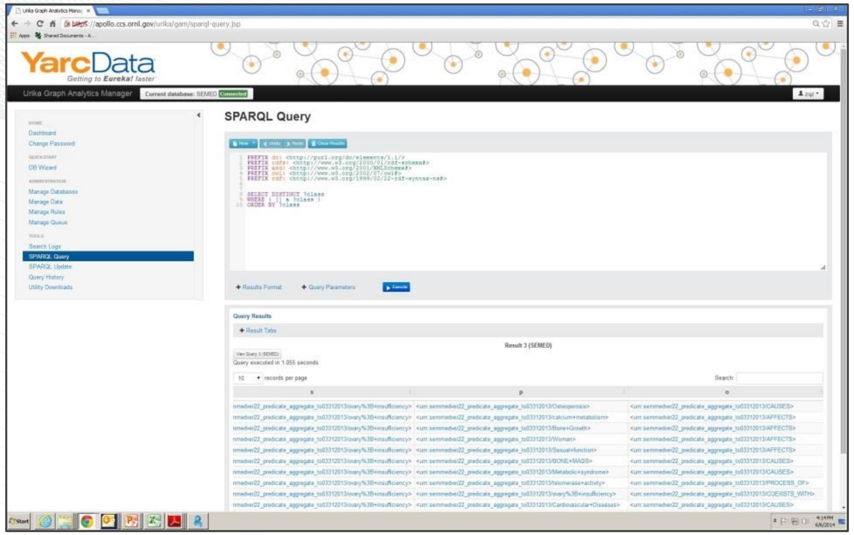


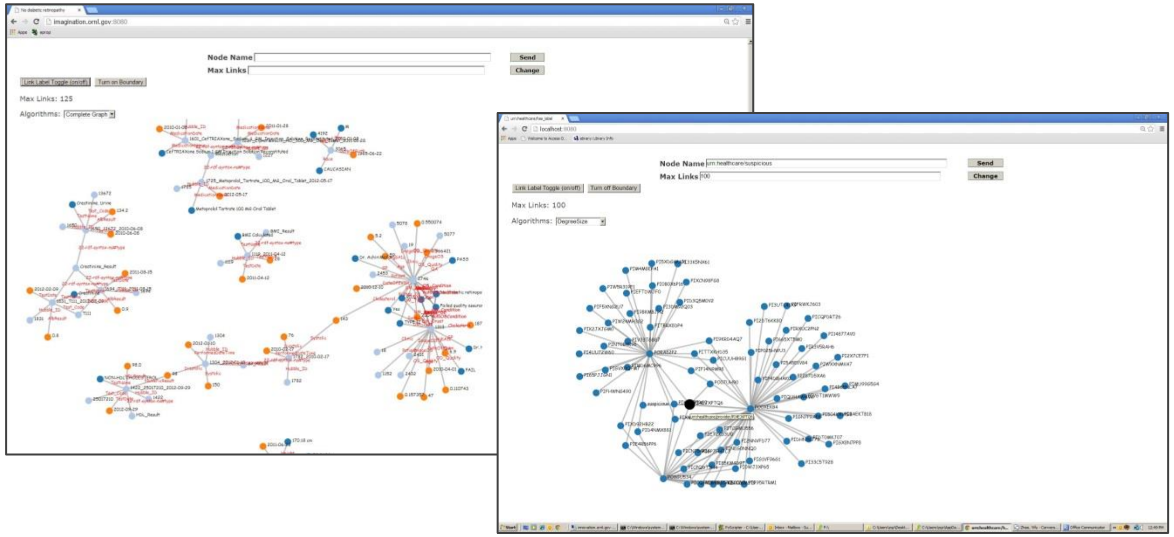
- **What does FELT do?**
  - Urika only understands RDF triples
  - Converting "graph-data" to RDF is an art that depends on the type of query we want to pose
  - Creating customized graph-models for the same dataset.
  - Map-Reduce Implementation for graph construction on Hadoop.

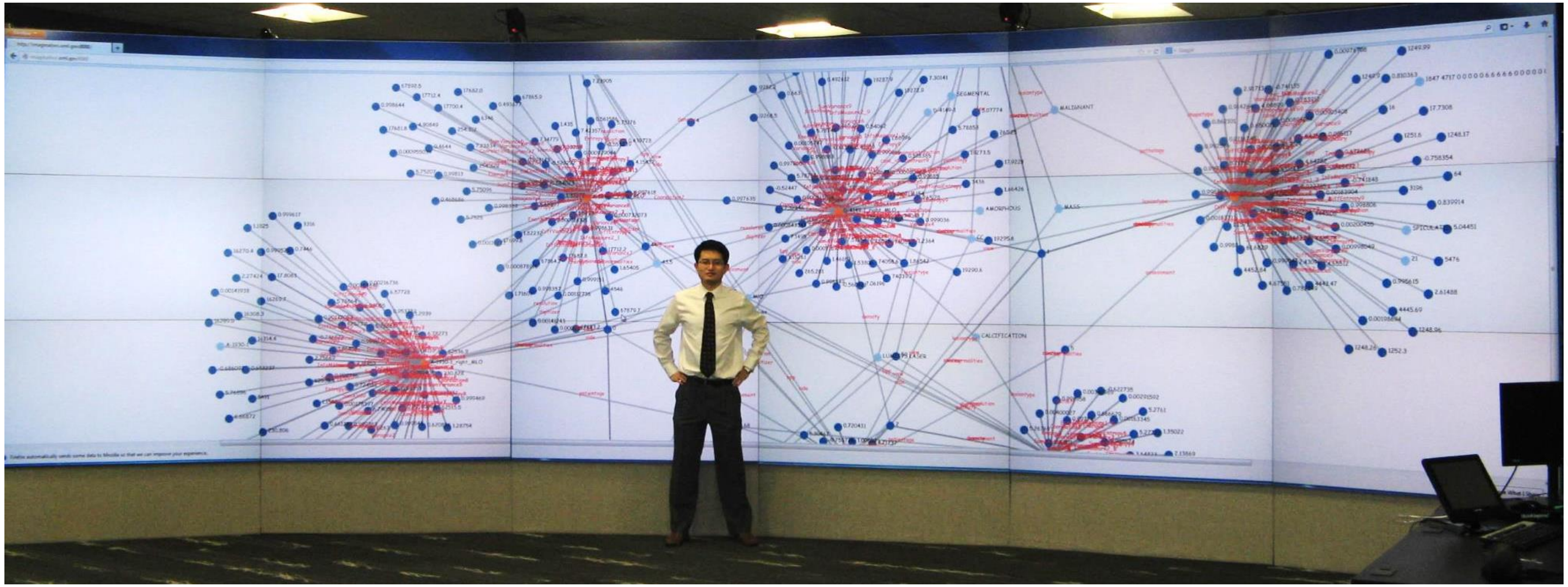


OAK RIDGE National Laboratory

GRAPHIC – Graph - Interaction Console







- **What does GRAPHIC do?**
  - Visualizes RDF triples
  - Makes pattern search easier on interactive console (particularly on EVEREST)
  - Works on iPads, EVEREST and most computers.

EAGLE: Eagle 'Is a' Algorithmic Graph Library for Exploration

- **What does EAGLE-C (C for Command-line do)?**
  - Lego-blocks for custom algorithms
  - '**First-ever**' SPARQL implementation for graph-theoretic inference

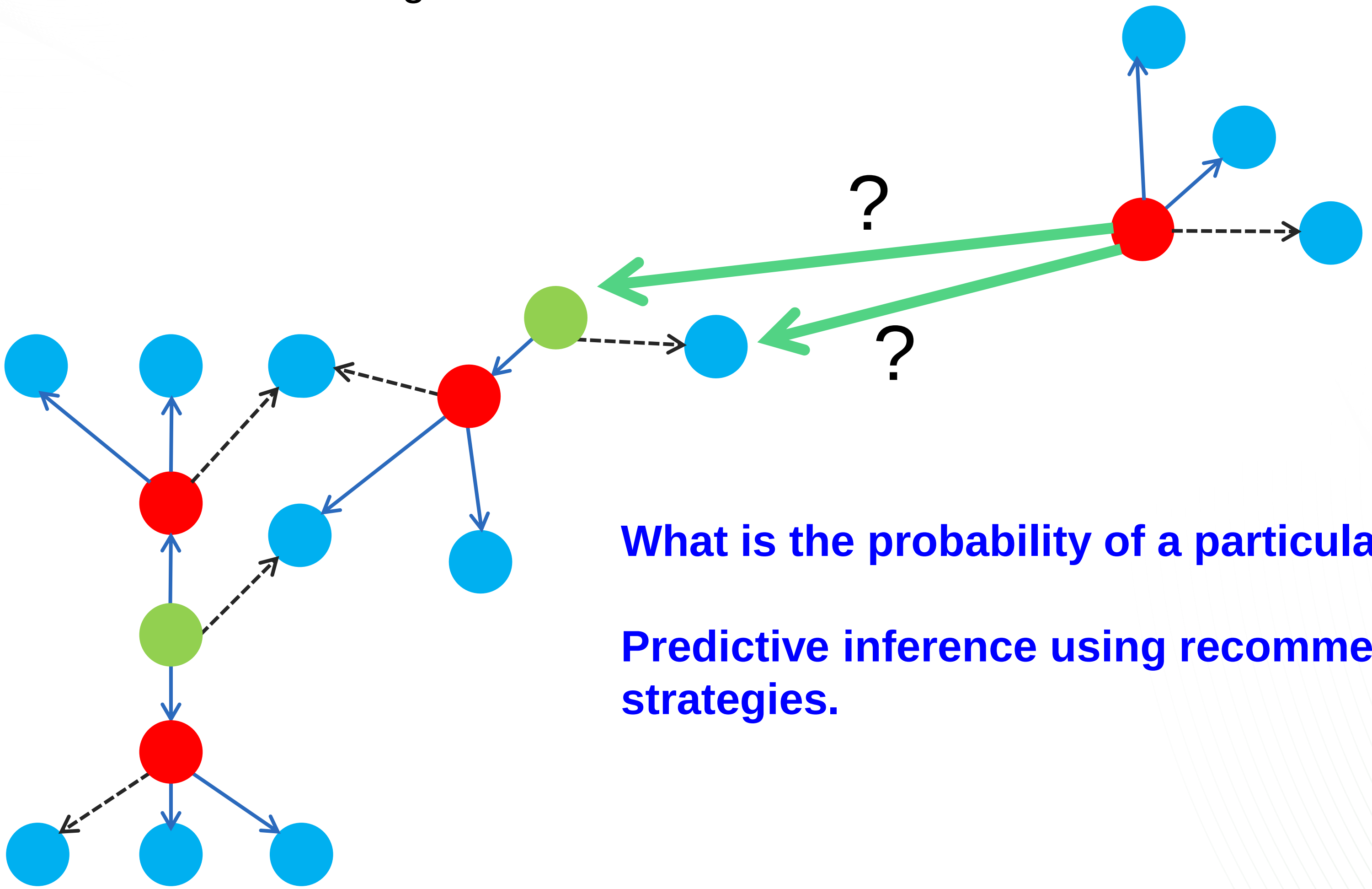
- **Some of the poplar graph-theoretic algorithms implemented and tested so far**
  - Summary metrics (~ 20 for both homogenous and heterogeneous graphs)
  - Degree (Diversity Degree)
  - Triangles (Count, Equilateral, Isosceles, Scalene)
  - N-gons
  - Shortest-path
  - PageRank (General, Personalized, BadRank, TrustRank)
  - Connected Components
  - Radius
  - Eccentricity
  - Degree-stratified clustering co-efficient
  - Peer-pressure clustering
  - Recommender systems
  - Label Propagation

Source code available:

https://github.com/ssrangan/gm-sparql

PAUSE: Predictive Analytics using SPARQL Endpoints

- **What does PAUSE do?**
  - Analyze multi-structure data (numeric data + domain knowledge/meta-data)
  - Implements similarity analysis, link prediction, simultaneous feature sub-setting and feature matching

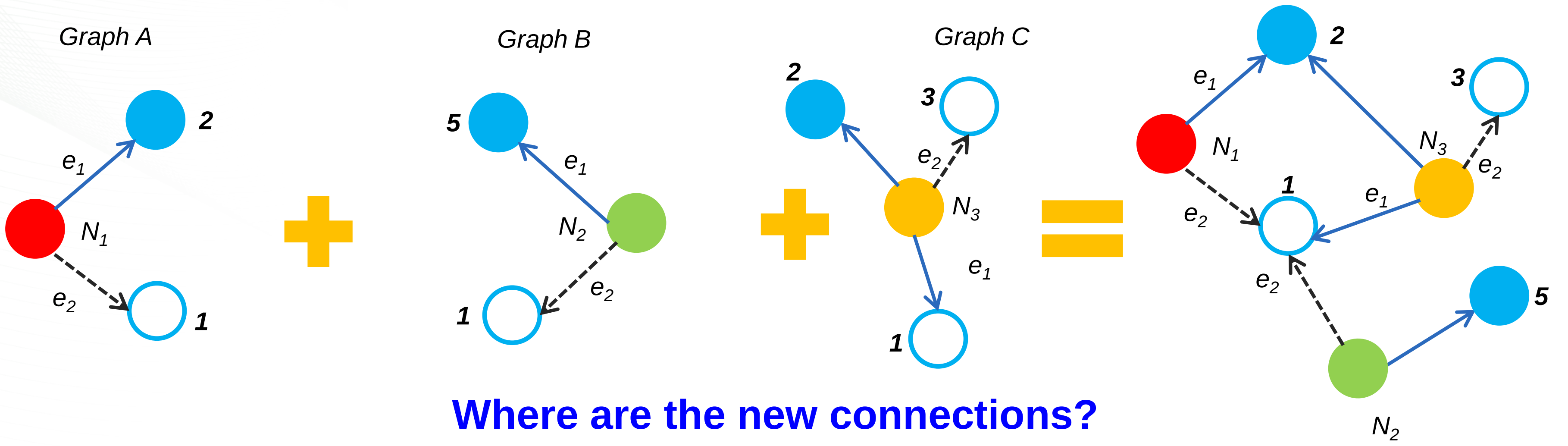


What is the probability of a particular edge to occur ?

Predictive inference using recommender system strategies.

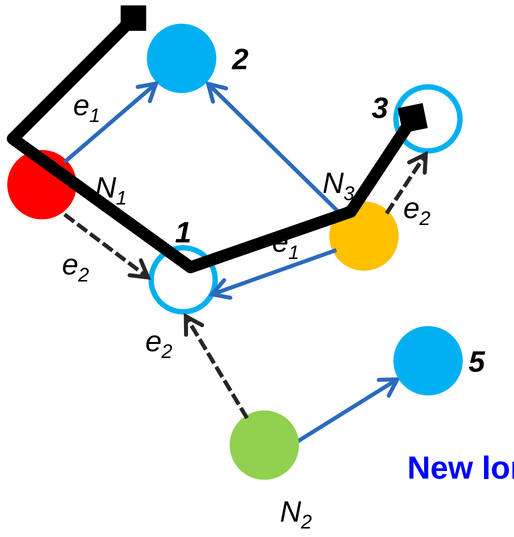
KENODES – Knowledge Extraction using Network-Oriented Discovery Enabling System

Extracting Novel and Useful Associations

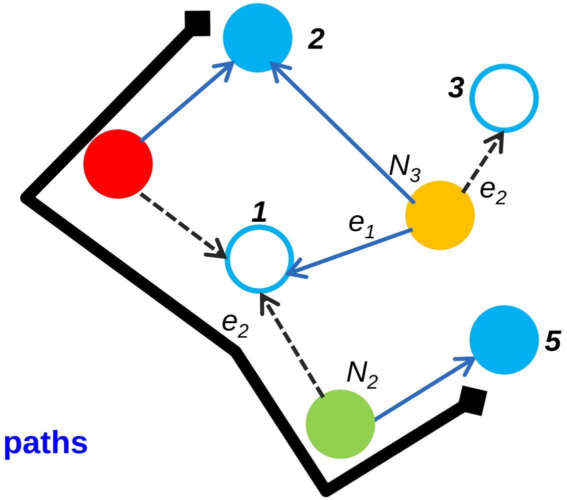


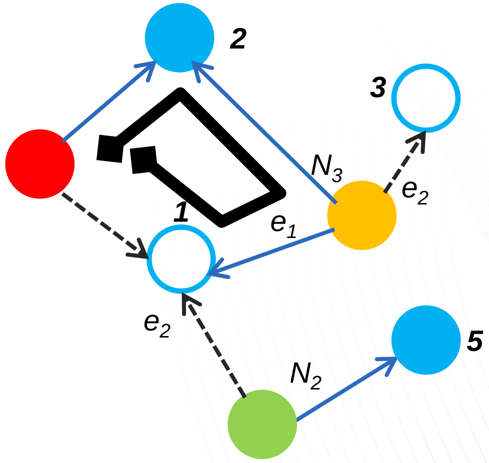
Where are the new connections?

What are the "important" new connections?



New longer paths



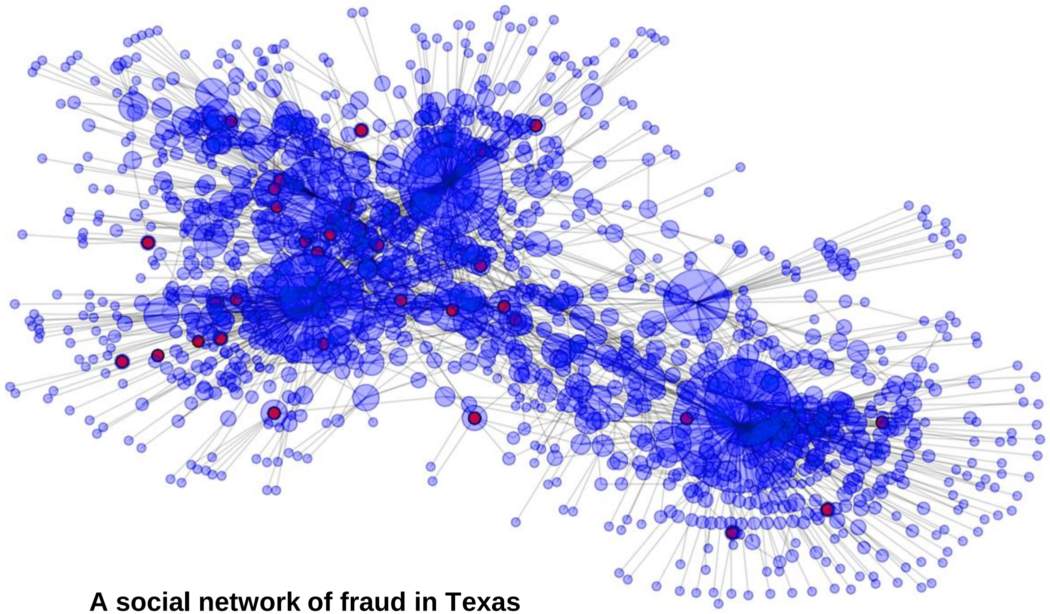


New triangles and polygons

OAK RIDGE National Laboratory

Apps @ Work

Motivation: Fraud Detection in Healthcare

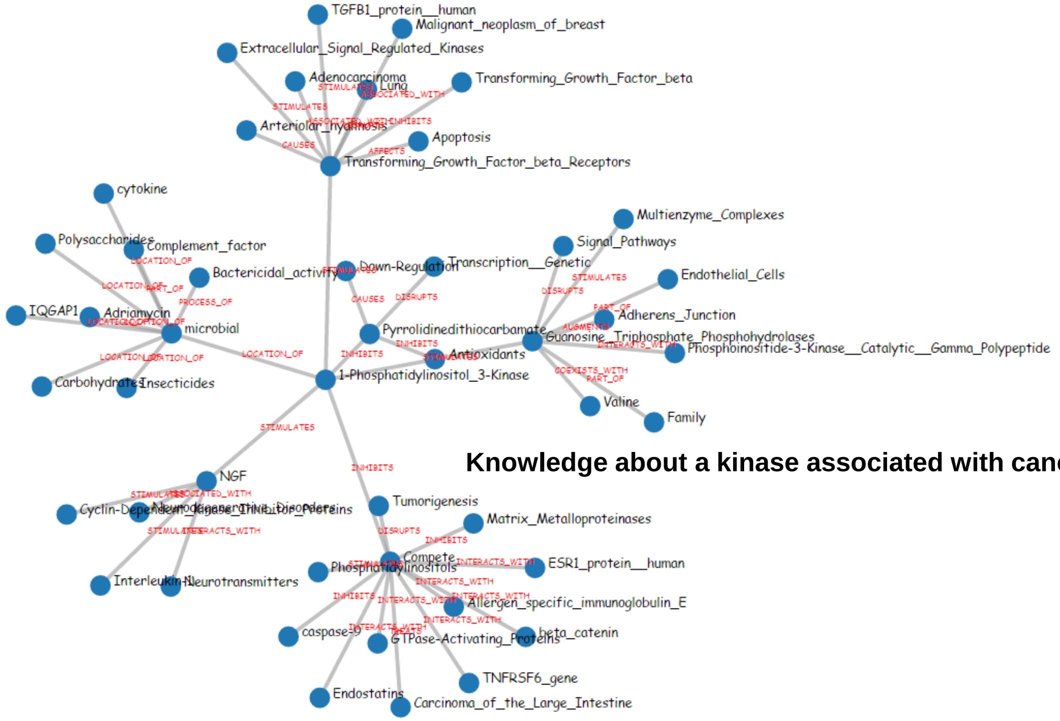


A social network of fraud in Texas

V. Chandola, S.R. Sukumar and J. Schryver, "Knowledge Discovery from Massive Healthcare Claims Data", in the Proc. of the 19th ACM SIGKDD Conference on Knowledge Discovery, 2013

Given a few examples of fraud (important activity), can we
- **(i) Automatically discover patterns typically associated with suspicious activity?**
- **(ii) Extrapolate such high-risk patterns for investigation and fraud prevention?**

Motivation: Knowledge Discovery from Literature



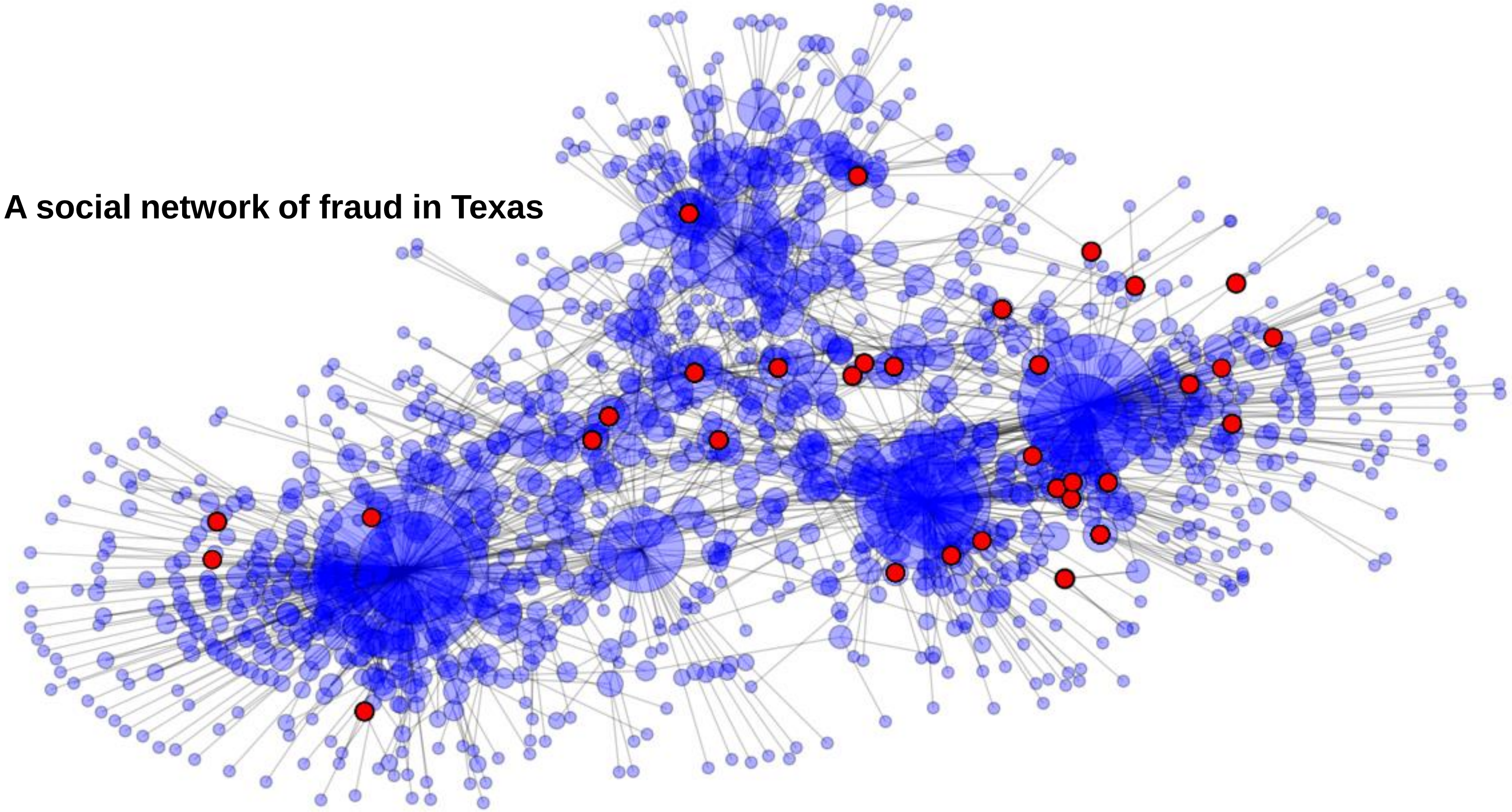
Knowledge about a kinase associated with cancer

S.M. Lee, S- H. Lim, T.C. Brown, *S. R. Sukumar*, "Graph mining meets the Semantic Web", in the Proc. the Data Engineering meets the Semantic Web Workshop in conjunction with International Conference on Data Engineering, 2015.

Given a knowledgebase and new clinical data/experiments, can we
- **(i) Find "novel" patterns of interest?**
- **(ii) Rank and evaluate the patterns for significance?**

Oak Ridge National Laboratory

Fraud Detection and Prevention



A social network of fraud in Texas

Given a few examples of fraud (important activity), can we

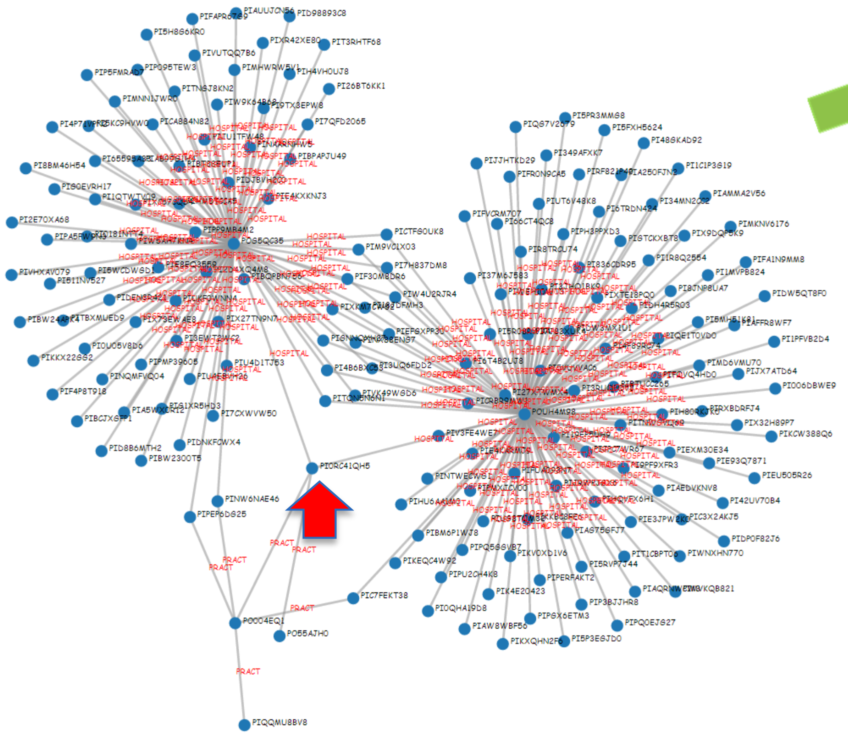
(i) **Automatically discover patterns typically associated with suspicious activity?**

(ii) **Extrapolate such high-risk patterns for investigation and fraud prevention?**

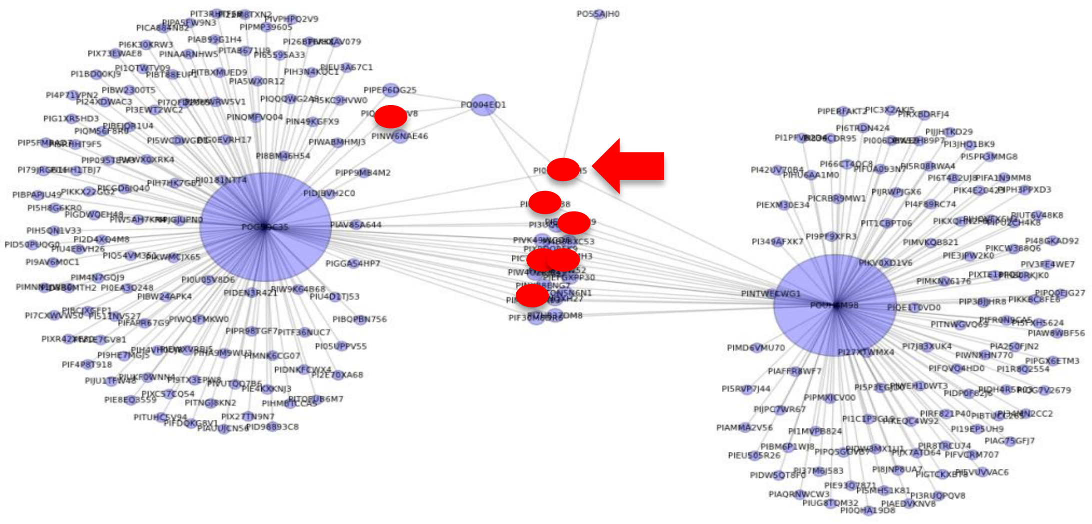
Pattern Discovery Example

Graph Pattern Search

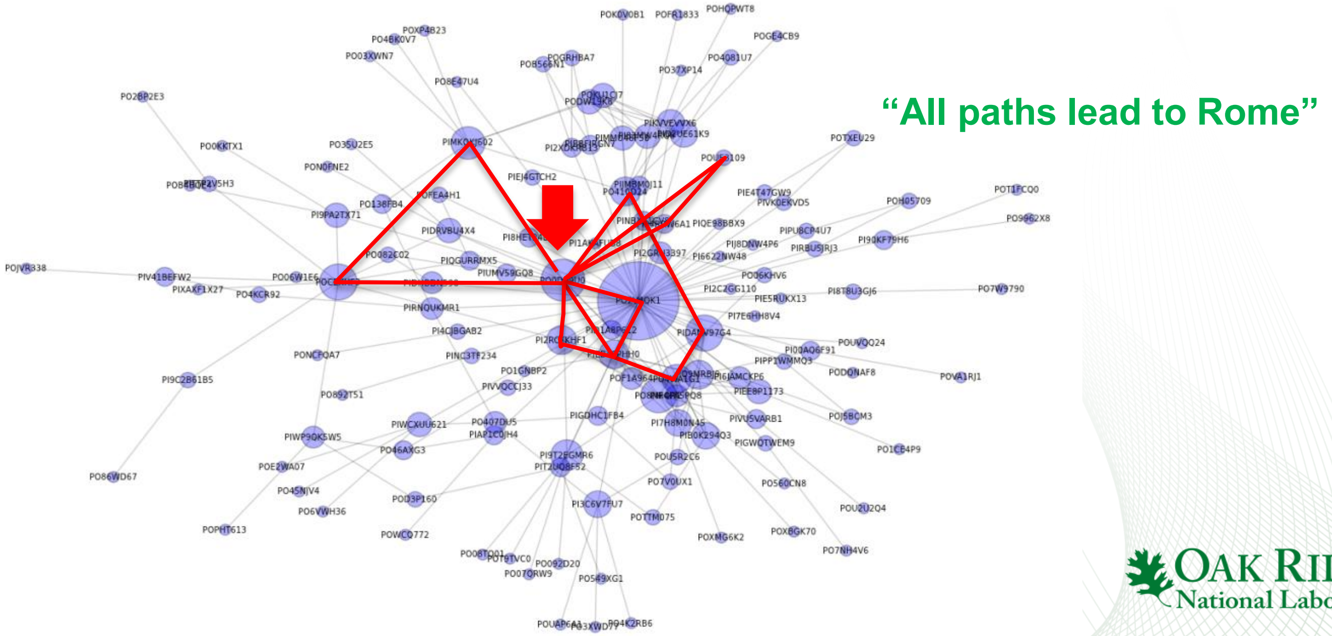
"Affiliations to multiple hospitals, owning private and group practice are strong indicators of potential suspicious activity."



"The country club phenomenon"



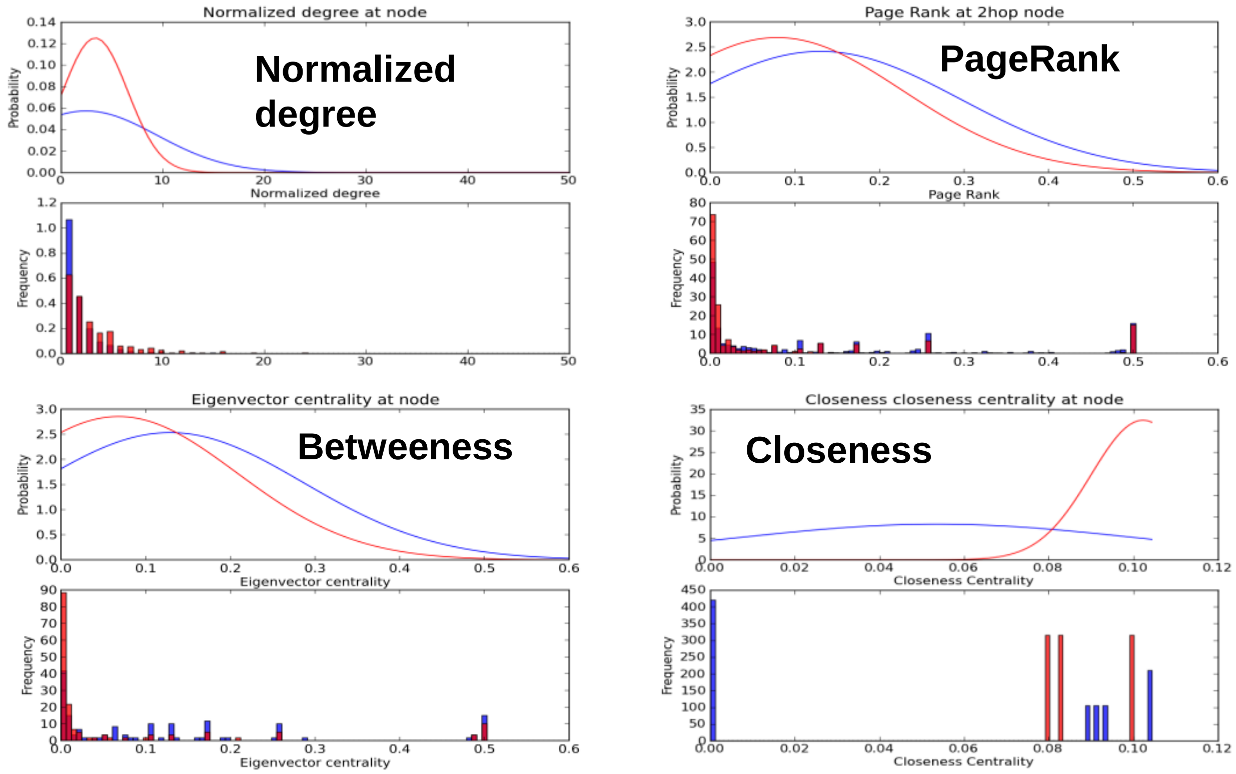
"All paths lead to Rome"



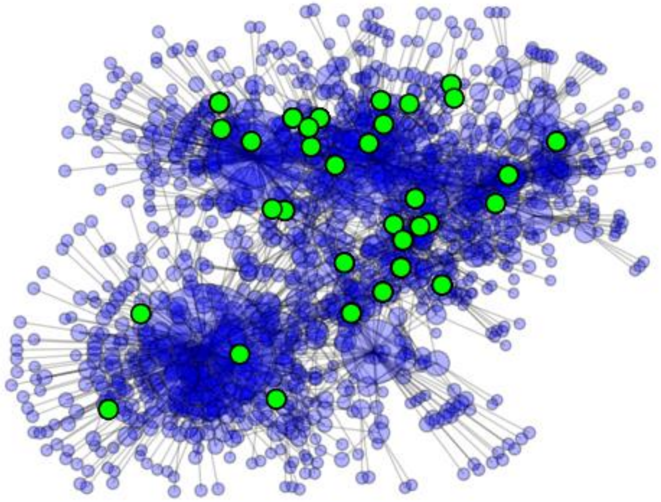
OAK RIDGE National Laboratory

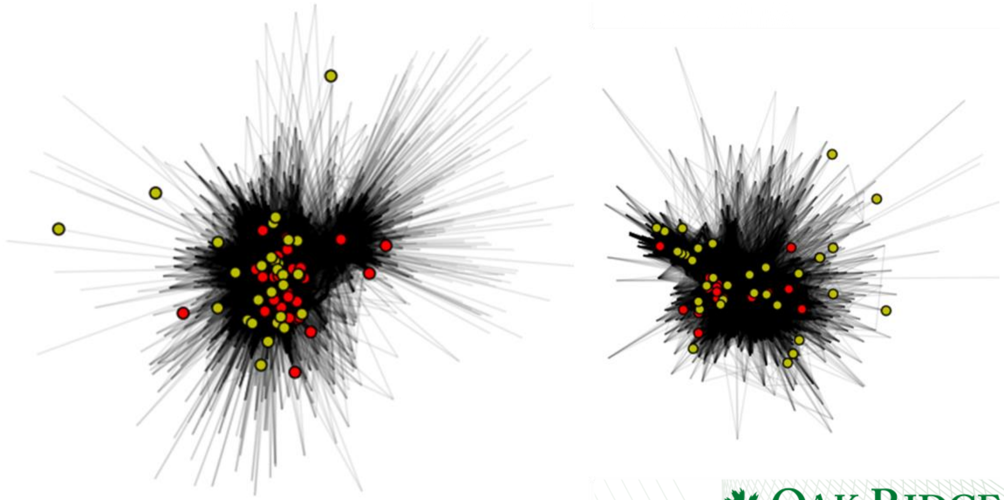
Predictive Analytics

Automated discrimination between two types of nodes based on various metrics in graphs whose generating model is not known *a priori*



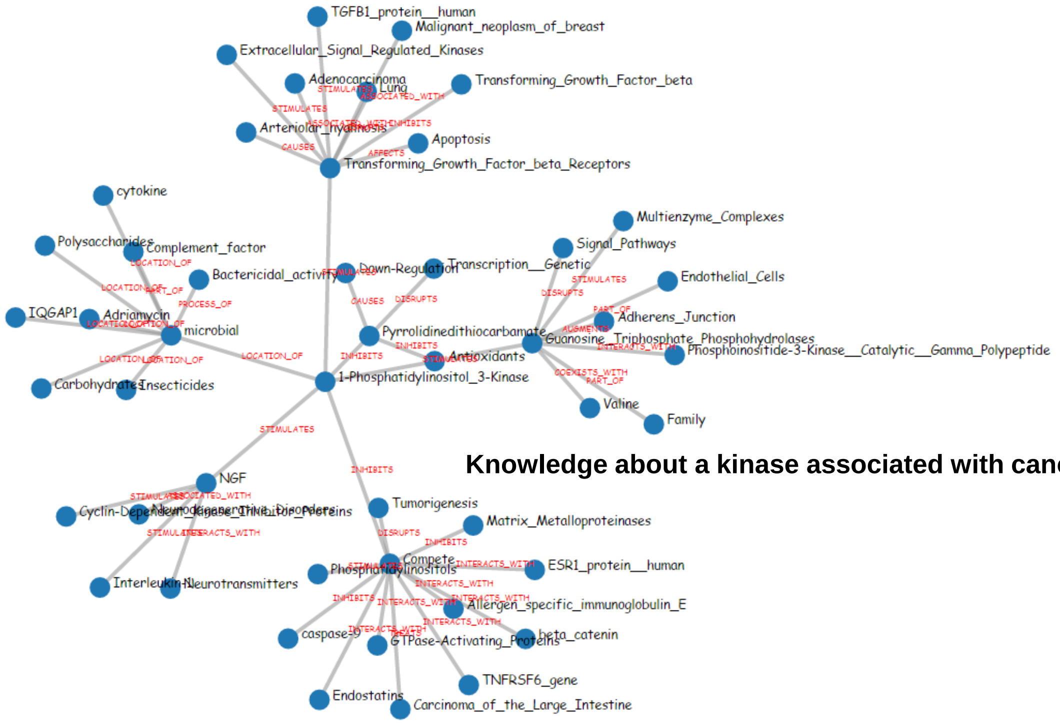
Extrapolating to unseen data





Apps @ Work

Motivation: Knowledge Discovery from Literature



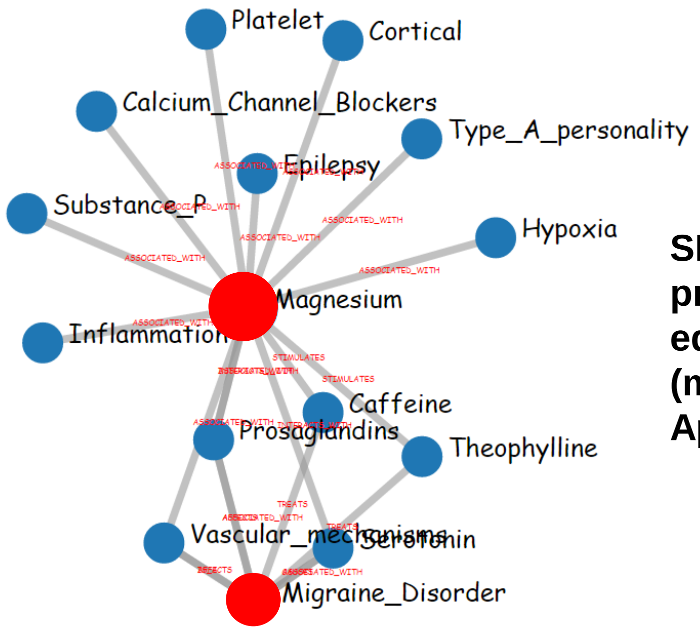
Knowledge about a kinase associated with cancer

S.M. Lee, S- H. Lim, T.C. Brown, *S. R. Sukumar,* "Graph mining meets the Semantic Web", in the Proc. the Data Engineering meets the Semantic Web Workshop in conjunction with International Conference on Data Engineering, 2015.

Given a knowledgebase and new clinical data/experiments, can we
- **(i) Find "novel" patterns of interest?**
- **(ii) Rank and evaluate the patterns for significance?**

Inspiring Motivation: Swanson's Story from 1987

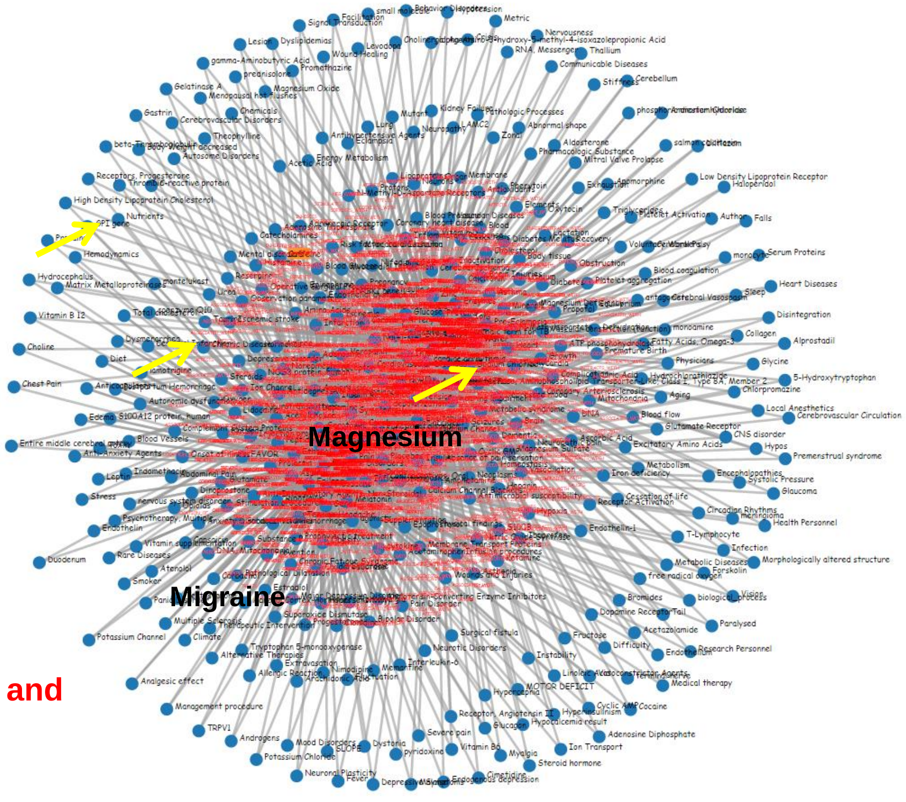
1987



Swanson, Don R. "Migraine and magnesium: eleven neglected connections." (1987).

SEMANTIC MEDLINE: 70 million predications (133 node types and 69 edge types)) from PubMed archive (more than 23.5 million citations, as of April 1st, 2014)

2014



Today: There are 133,193 connections between migraine and magnesium.

Given a knowledgebase and new clinical data/experiments, can we

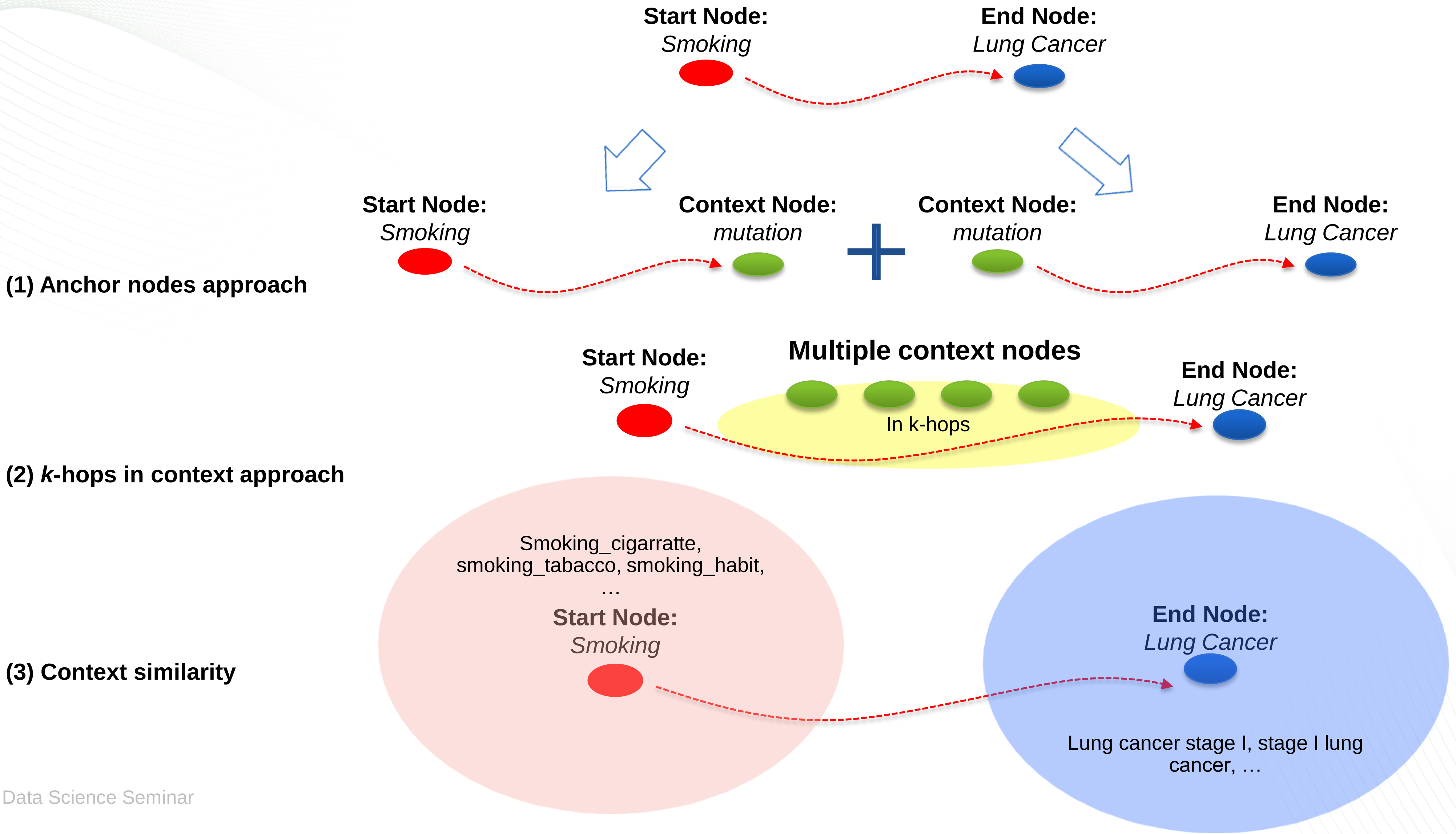
(i) Find "novel" patterns of interest?

(ii) Rank and evaluate the patterns for significance?

OAK RIDGE National Laboratory

Reasoning Apps @ Work: Information Foraging

Approach #2: Context-aware exploration



Apps @ Work: Back to Migraine and Magnesium

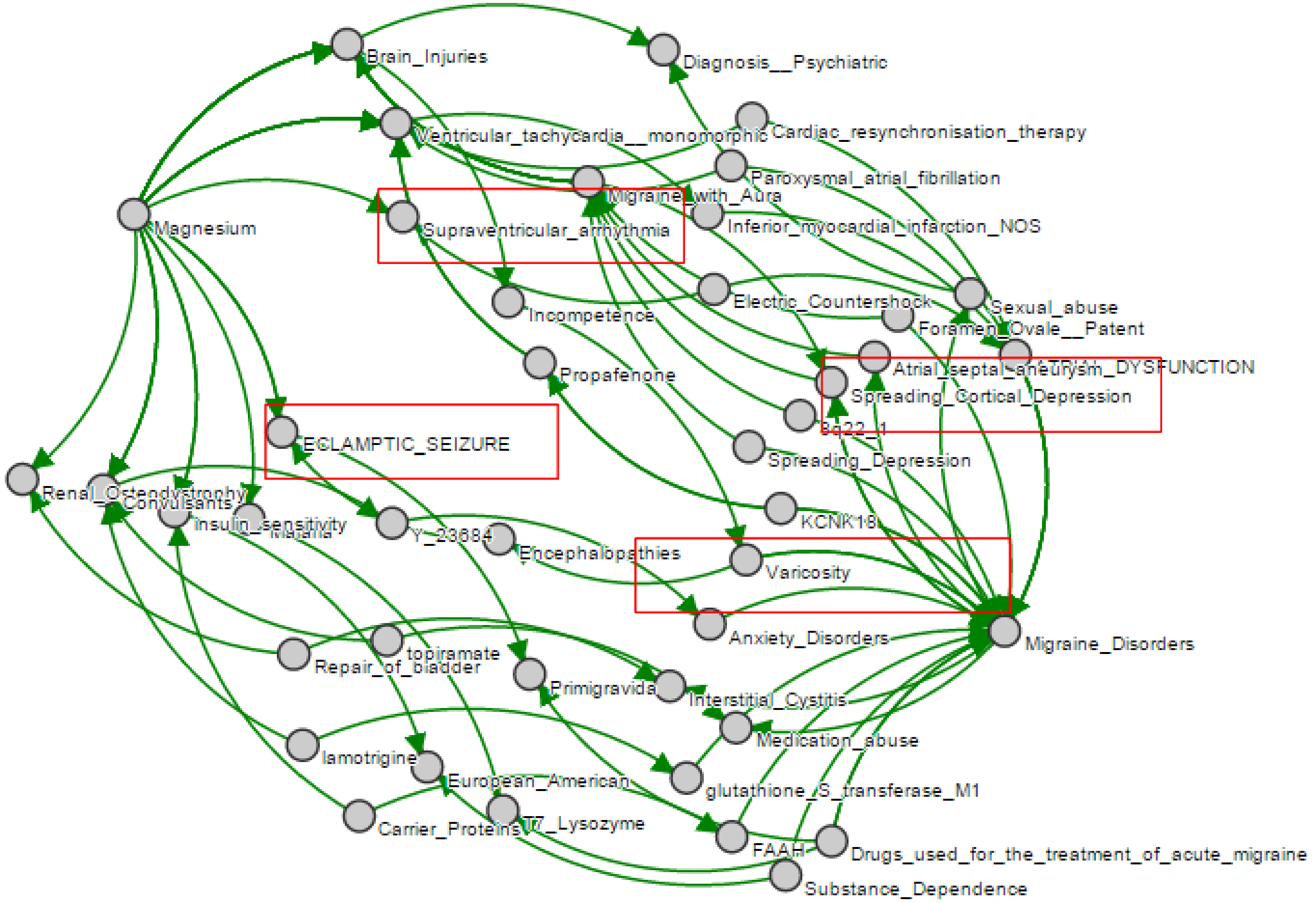
Results: Eureka ! Eureka !

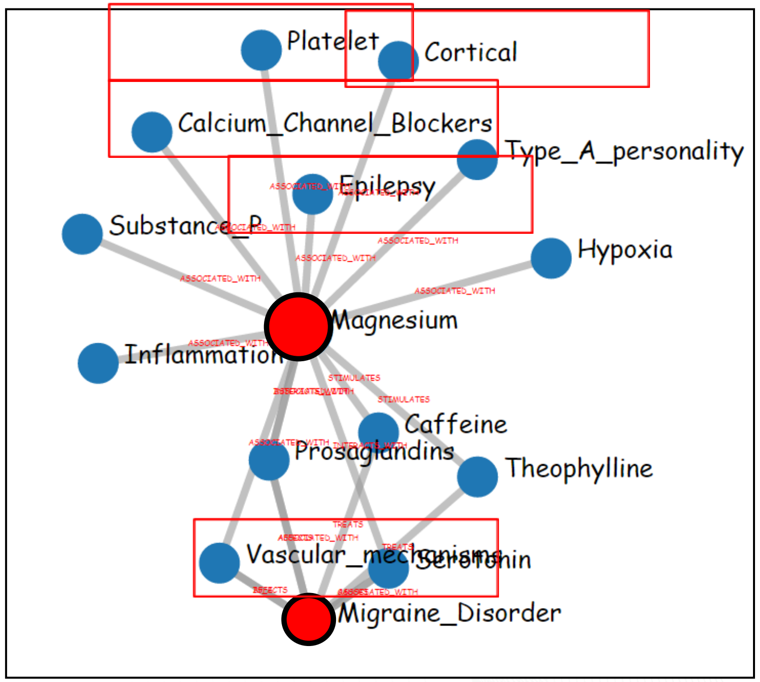
Magnesium	Rev_INHIBITS	Tantalum	AUGMENTS	Osseointegration	NEG_COEXISTS_WITH	Bone_Regeneration	Rev_COEXISTS_WITH	Platelet_function	Rev_NEG_MANIFESTATION_OF	Migraine_Disorders	0.405123
Magnesium	Rev_STIMULATES	BW35	NEG_ASSOCIATED_WITH	Renal_Insufficiency	Rev_PRECEDES	Cervix_carcinoma	Rev_NEG_PREDISPOSES	Combined_Oral_Contraceptives	NEG_COMPLICATES	Migraine_Disorders	0.355147
Magnesium	NEG_COMPLICATES	Malaria	ASSOCIATED_WITH	heme_binding	Rev_NEG_AUGMENTS	insulin_receptor_related_receptor_INSRR	Rev_CONVERTS_TO	Melatonin	NEG_DISRUPTS	Migraine_Disorders	0.351549
Magnesium	Rev_NEG_USES	kidney_preservation	DIAGNOSES	Unilateral_agenesis_of_kidney	Rev_COMPLICATES	Congenital_obstruction	CAUSES	Varicosity	NEG_COMPLICATES	Migraine_Disorders	0.335784
Magnesium	Rev_NEG_USES	kidney_preservation	TREATS	Hypertension__Renovascular	ASSOCIATED_WITH	noradrenergic	Rev_ASSOCIATED_WITH	Varicosity	NEG_COMPLICATES	Migraine_Disorders	0.304279
Magnesium	Rev_STIMULATES	hemiacidrin	Rev_USES	Percutaneous_insertion_of_nephrostomy_tube	AFFECTS	Cervix_carcinoma	Rev_NEG_PREDISPOSES	Combined_Oral_Contraceptives	NEG_COMPLICATES	Migraine_Disorders	0.303071
Magnesium	NEG_PREVENTS	Brain_Injuries	NEG_AFFECTS	Blood_Flow_Velocity	Rev_CAUSES	Dihydroergotamine	same_as	Triptans	DISRUPTS	Migraine_Disorders	0.288334
Magnesium	NEG_PREVENTS	Brain_Injuries	NEG_AFFECTS	Blood_Flow_Velocity	Rev_CAUSES	Dihydroergotamine	same_as	Triptans	AUGMENTS	Migraine_Disorders	0.288062
Magnesium	NEG_CAUSES	Hypomagnesemia	Rev_NEG_COEXISTS_WITH	Renal_Osteodystrophy	Rev_COMPLICATES	Repair_of_bladder	CAUSES	Interstitial_Cystitis	Rev_COMPLICATES	Migraine_Disorders	0.280192
Magnesium	NEG_COMPLICATES	Malaria	Rev_TREATS	Heparitin_Sulfate	NEG_PART_OF	Herpesvirus_1__Suid	INTERACTS_WITH	Varicosity	NEG_COMPLICATES	Migraine_Disorders	0.276062

OAK RIDGE National Laboratory

Apps @ Work: Back to Migraine and Magnesium

Eureka ! Eureka !

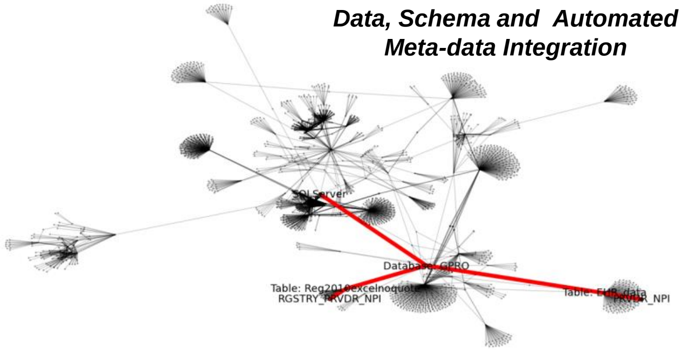




OAK RIDGE National Laboratory

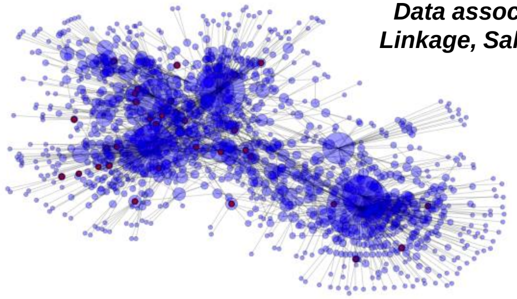
Graph Computing using Urika-GD: Summary

Data, Schema and Automated Meta-data Integration



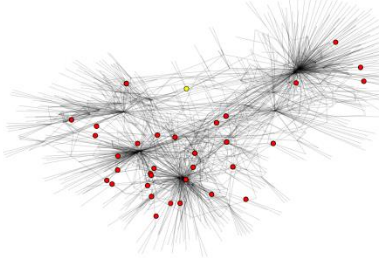
SEEKER: Schema Exploration and Evolving Knowledge Recorder

Data association, Record-Linkage, Saliency extraction

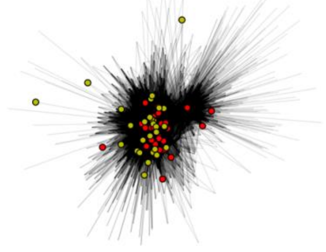


SNAKE: Social/Network Analytic Knowledge Extractor

Pattern Discovery



Pattern Recognition



PAUSE: Predictive Analytics Using Software-Endpoints

Graph Computing…

- Supports discovery by interrogation, association and predictive modeling from structured and unstructured data
- Supports discovery with evolving knowledge and incremental domain hints
- Supports exploratory and confirmatory analysis
  - Data and meta-data integrated analytics
  - Flexible data structure seamless to growth while avoiding analytical artifacts

Patents:

Sreenivas R. Sukumar, Regina K. Ferrell, and Mallikarjun Shankar. Knowledge Catalysts, US Patent Application 14/089,395, filed November 25, 2013.

Sreenivas R. Sukumar et al., Scalable Pattern Search in Multi-Structure Data, US Patent Application 62/106,342, filed January 22, 2015.

OAK RIDGE National Laboratory