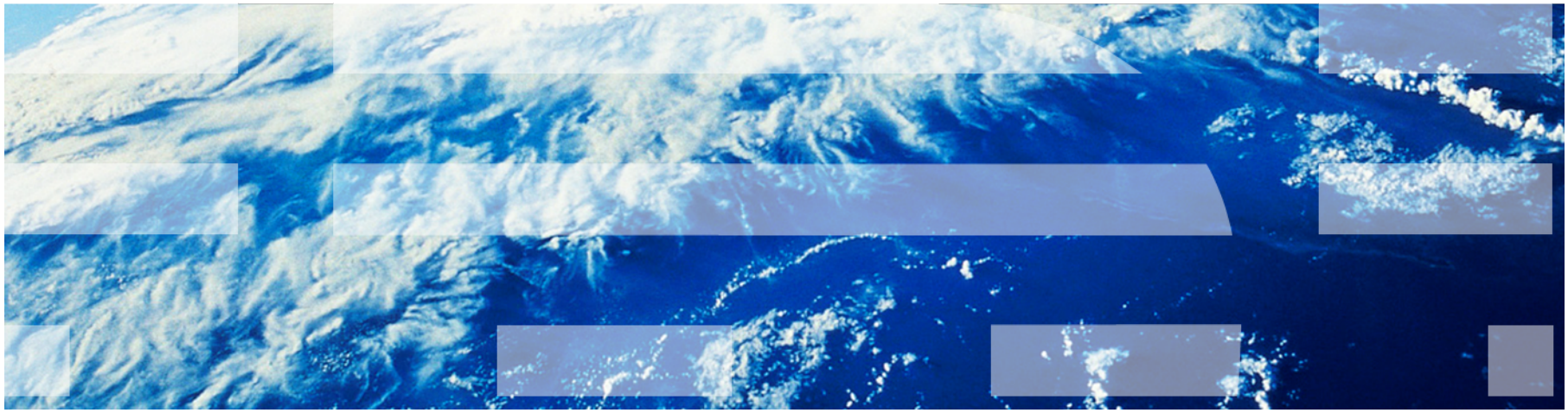
Columbia University

E6893 Big Data Analytics Lecture 5:

Linked Big Data Graph Analytics

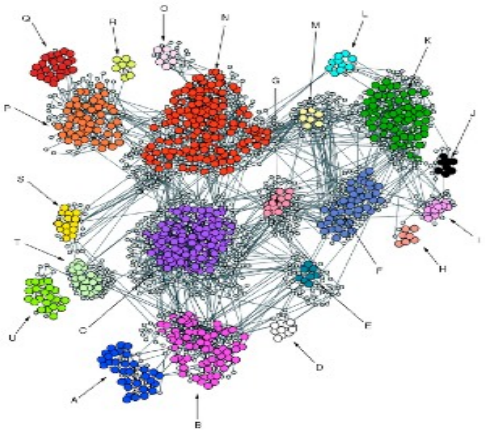
Ching-Yung Lin, Ph.D.

Adjunct Professor, Dept. of Electrical Engineering and Computer Science



October 4, 2019

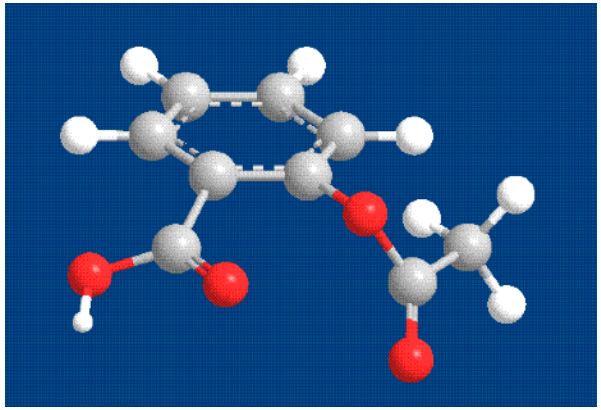
Networks Everywhere



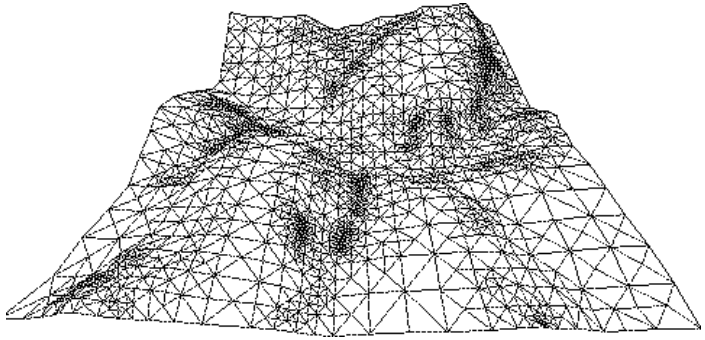
Magwene *et al. Genome Biology* 2004 **5**:R100

Gene Coexpression Network

Scientific Workflow



Chemical Compound



Mesh

First Reported Social Network Analysis

The New York Times.

Copyright, 1933, by The New York Times Company.

Entered as Second-Class Matter, Postoffice, New York, N. Y. NEW YORK, THURSDAY, APRIL 6, 1933. TWO CENTS

EMOTIONS MAPPED BY NEW GEOGRAPHY

Charts Seek to Portray the Psychological Currents of Human Relationships.

FIRST STUDIES EXHIBITED

Colored Lines Show Likes and Dislikes of Individuals and of Groups.

MANY MISFITS REVEALED

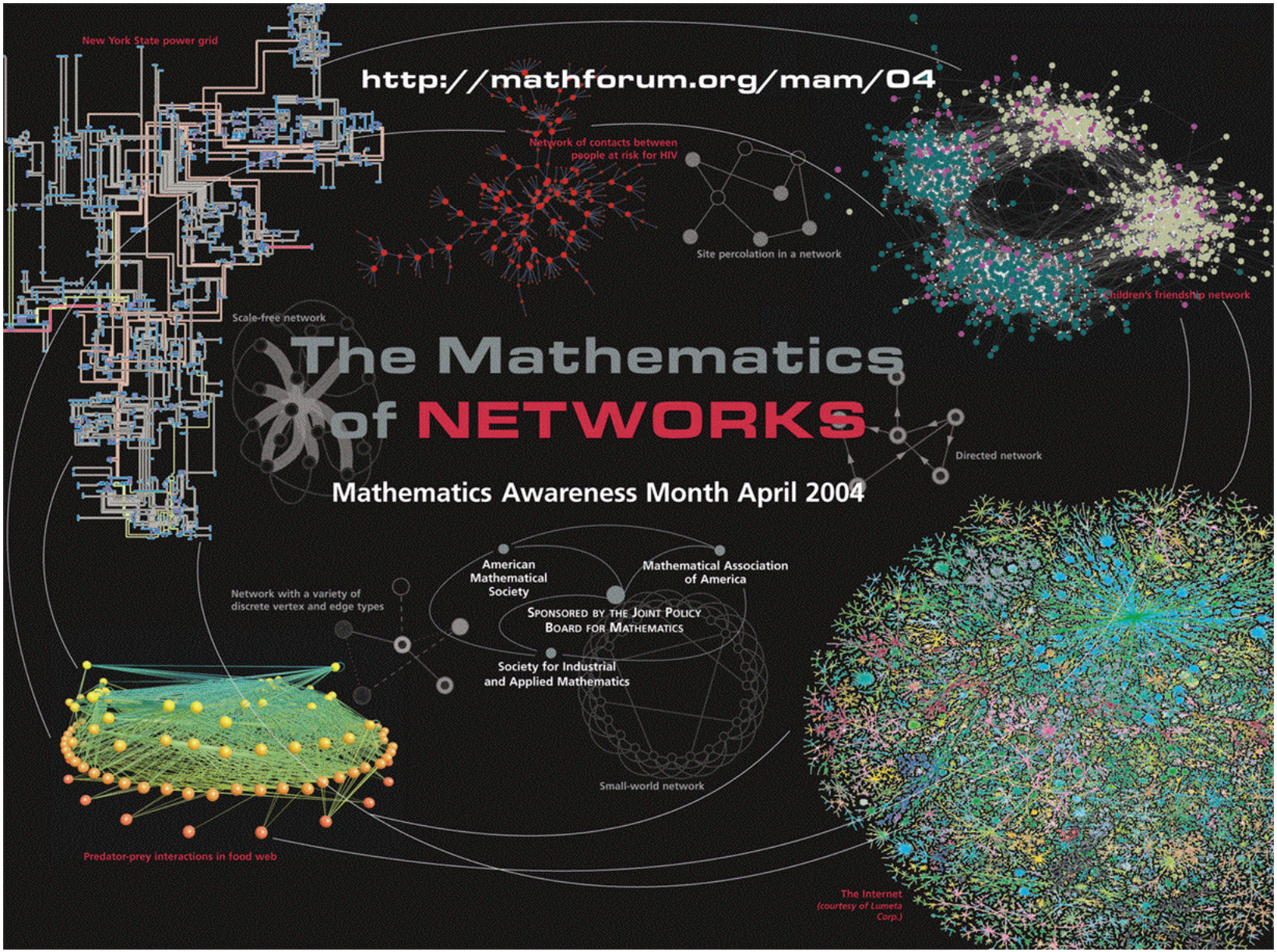
Dr. J. L. Moreno Calculates There Are 10 to 15 Million Isolated Individuals in Nation.

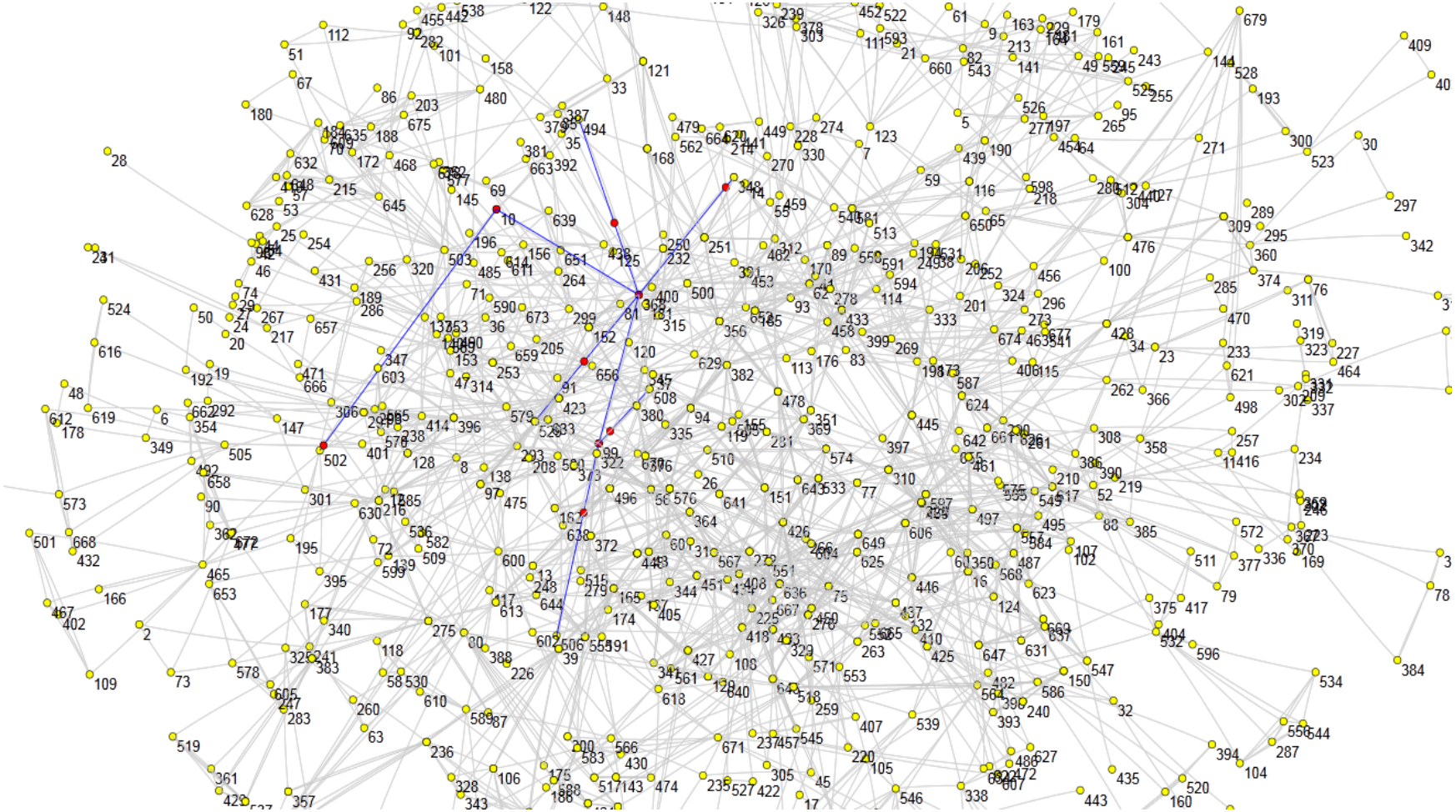
A new science, named psychological geography, which aims to chart the emotional currents, cross-currents and under-currents of human relationships in a community, was introduced here yesterday at the scientific exhibit of the Medical Society of the State of New York, which opens its 127th annual meeting here today at the Waldorf-Astoria.

The first series of maps of the new human geography were shown by Dr. Jacob L. Moreno of New York, consulting psychiatrist of the National Committee of Prisons and Prison Labor and director of research, New York State Training School for Girls, Hudson, N. Y. The maps represent studies of the forces of attraction and repulsion of individuals within a group toward one another and toward the group, as well as the attitude of the group as a whole toward its individual members, and of one group toward another group.

Emotions are represented on these psychological maps by various colored lines. Red stands for liking, black for disliking. If individual A likes B a red line with an arrow points from A to B. If B reciprocates a similar red line points from him. If he dislikes A this is indicated by a black line with an arrow pointing toward A. If B is merely indifferent the feeling is shown by a blue line.

Group of 500 Girls Studied.





Network Science <=> Graph Technology

Characteristics of Network Data

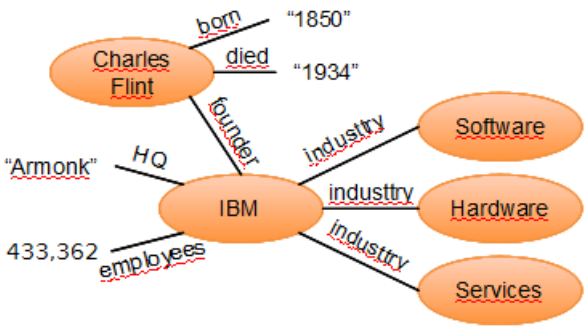
- High-Dimensional
- Dependent
- Massive

Types of Graph

Columbia University

RDF / Property Graph

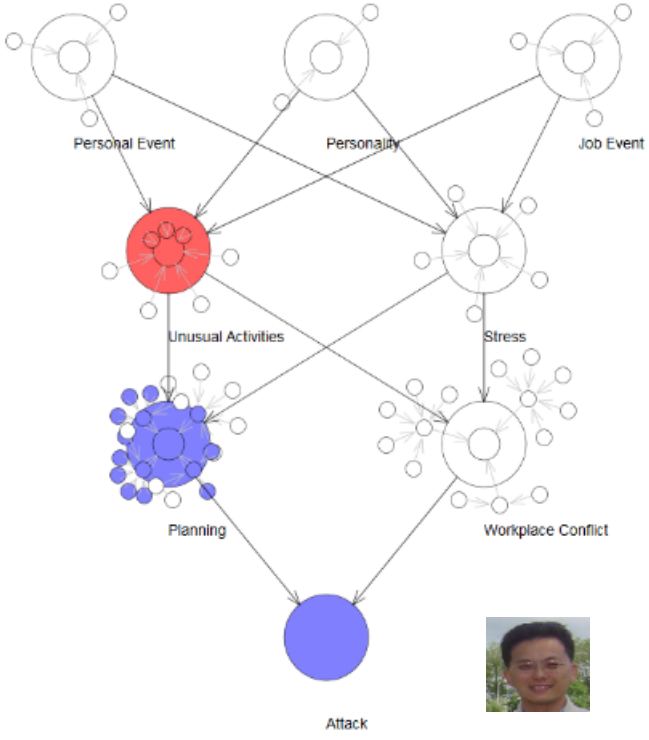
Attributes



subject	predicate	object
Charles Flint	born	"1850"
Charles Flint	died	"1934"
Charles Flint	founder	IBM
IBM	HQ	"Armonk"
IBM	employees	433,362
IBM	industry	Software
IBM	industry	Hardware
IBM	industry	Services

Activity Graph

Micro



Collective Graph

Macro

The Emergence of Network Science

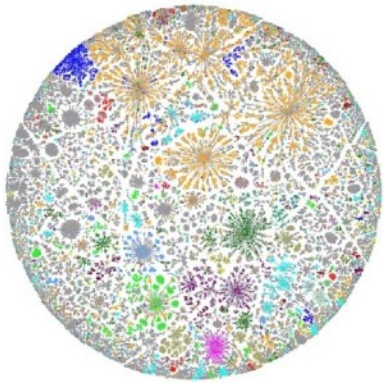
- Science <=> Observable systematic empirical data
- Facility of large-scale data collection, storage and management.

The Emergence of Network Science

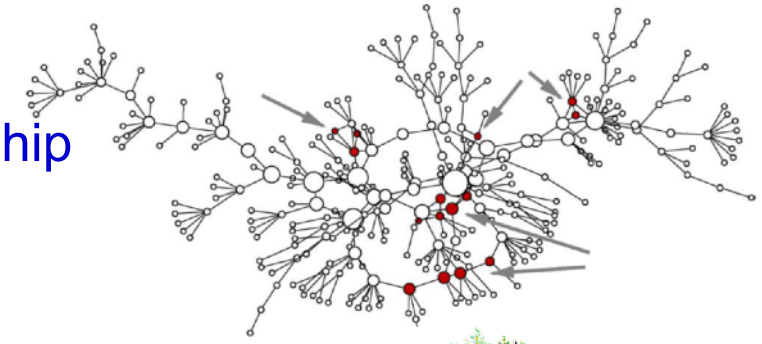
- Science <=> Observable systematic empirical data
- Facility of large-scale data collection, storage and management.
  - *Statistical Methodologies to combine behavior understanding, link analysis, multi-variant modeling, machine learning, graph theory, and non-parametric statistics for complex network analysis*

Contributions made by Physicists and Computer Scientists, have greatly expanded the discipline over the past 15 years.

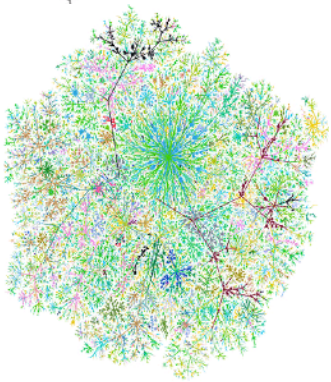
Example 1: Internet Map
Nodes: ISPs; Edges: Connection
(33K Nodes, 290K edges)



Example 2: Social Network
Nodes: People; Edges: Friendship
(FaceBook has 500M+ Users)



Example 3: Web Graph
Nodes: Web Pages; Edges: Hyperlinks
(Yahoo Web: 1.4B nodes, 6.6B edges)



Multiple Scales, Multiple Disciplines

Multi-disciplinary Research Issues

- Formation of Network
  - Communications
  - Information
  - People
  - Companies / Organizations
  - Nations
- Network Data Collection
- Network Science Infrastructure
- Network Applications
- Network Visualization
- Network Sampling, Indexing and Compression
- Network Flow
- Network Evolution and Dynamics
- Network Impact
- Cognitive Networks

Multi-disciplinary Research Issues

- Formation of Network
  - Communications ← *Electrical Engineering*
  - Information ← *Computer Science*
  - People ← *Sociology, Public Health*
  - Companies / Organizations ← *Economics, Management, Politics*
  - Nations ← *International Relationships, History*
- Network Data Collection ← *Law*
- Network Science Infrastructure
- Network Applications
- Network Visualization ← *Arts, Math*
- Network Sampling, Indexing and Compression ← *Math*
- Network Flow ← *Physics*
- Network Evolution and Dynamics
- Network Impact
- Cognitive Networks ← *Bio, Cognition, Behavior Science*

Example: Network Science Consortium (2009 – 2019)

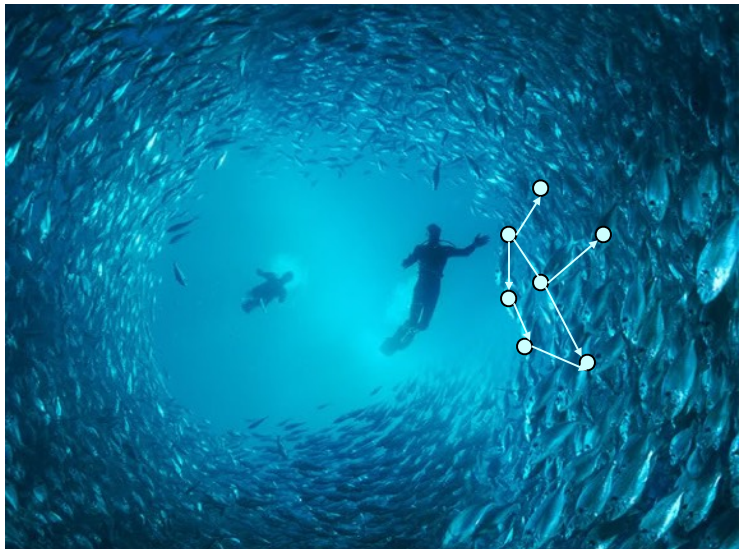
- $172 million in 10 years to fund 4 *Network Science* academic research centers:
  - Communication Network Academic Research Center (CNARC)
  - Social and Cognitive Network Academic Research Center (SCNARC)
  - Information Network Academic Research Center (INARC)
  - Interdisciplinary Research Center (IRC)

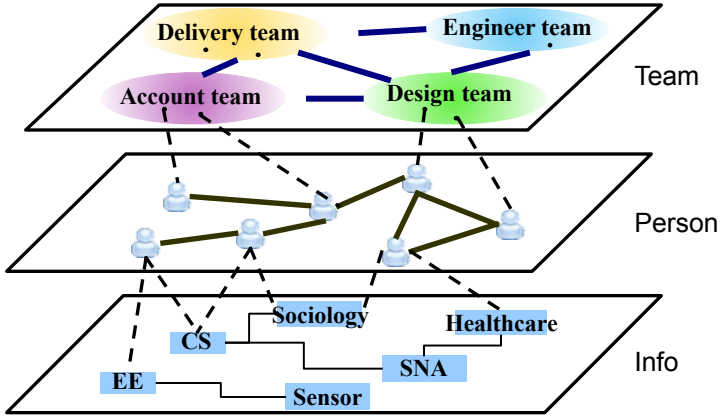
~ 100 Professor/Researcher Principle Investigators + 250 RAs, Postdocs

- Objectives:
  - Improve Decision Making
  - Develop measures, metrics and models that describe and predict human-network interaction and exchange within & across network layers
  - Develop and validate theory of human-system interaction in network-centric environments
  - Explore techniques for dynamic, flexible, adaptive, and adaptable interaction

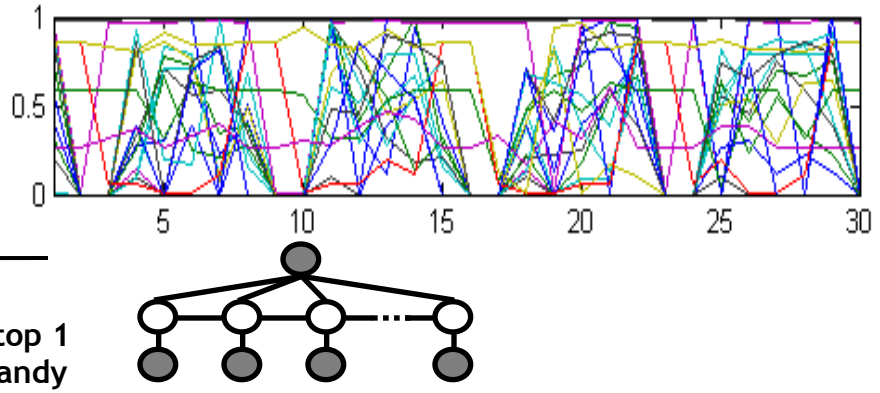
1. Characterizing and Measuring Networks
2. Understanding Networks for Analysis
3. Controlling and Managing Networks
4. Using Networks

Heterogeneous Synchronicity Networks Predict Performance





One-class HCRF to detect temporal anomalies



Detected as top 1 anomaly in Sandy Tweets

Large-Scale People Modeling and Social Network Analysis

Emails
Chats
Meetings
Web Page Clicks
Server Logs

Social sensors

Click streams capturer

Feed subscription

Database access

Network analysis

People analysis

Content analysis

Applications

Live Data, Live Production System

20,000,000 emails & SameTime messages

1,000,000 Learning click data

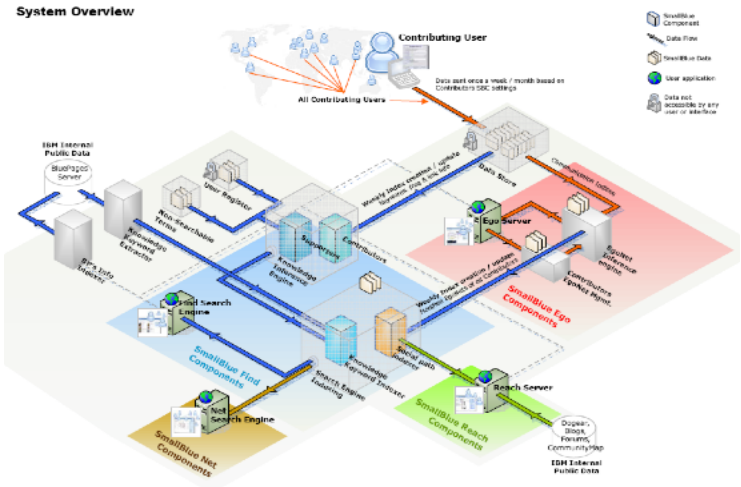
14,000,000 KnowledgeView, SalesOne, …, access data

1,000,000 Lotus Connections (blogs, flie sharing, bookmark) data

200,000 people's consulting financial databases

400,000 organization/demographic data

100,000 intranet w3 searches per day

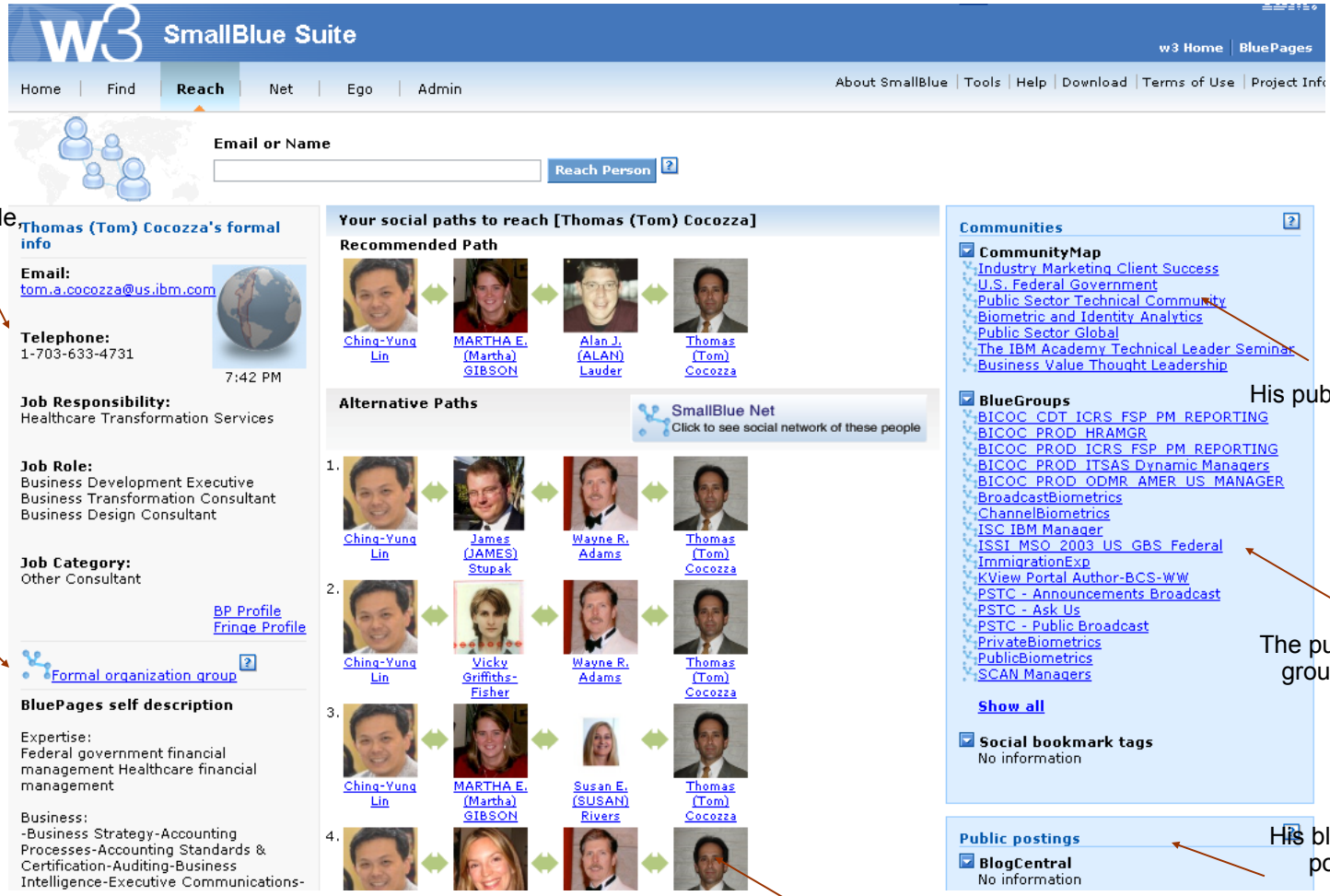


Find knowledgeable colleagues

- **E.g.: Search for the most knowledgeable colleagues within my 3-degree network for who knows 'healthcare'. (or within a country, a division, a job role, or any group/community)**

Reach – social dashboards

- **Is Tom a right person to me?**



His official job role, title, contact info

His self-described expertise

His public communities

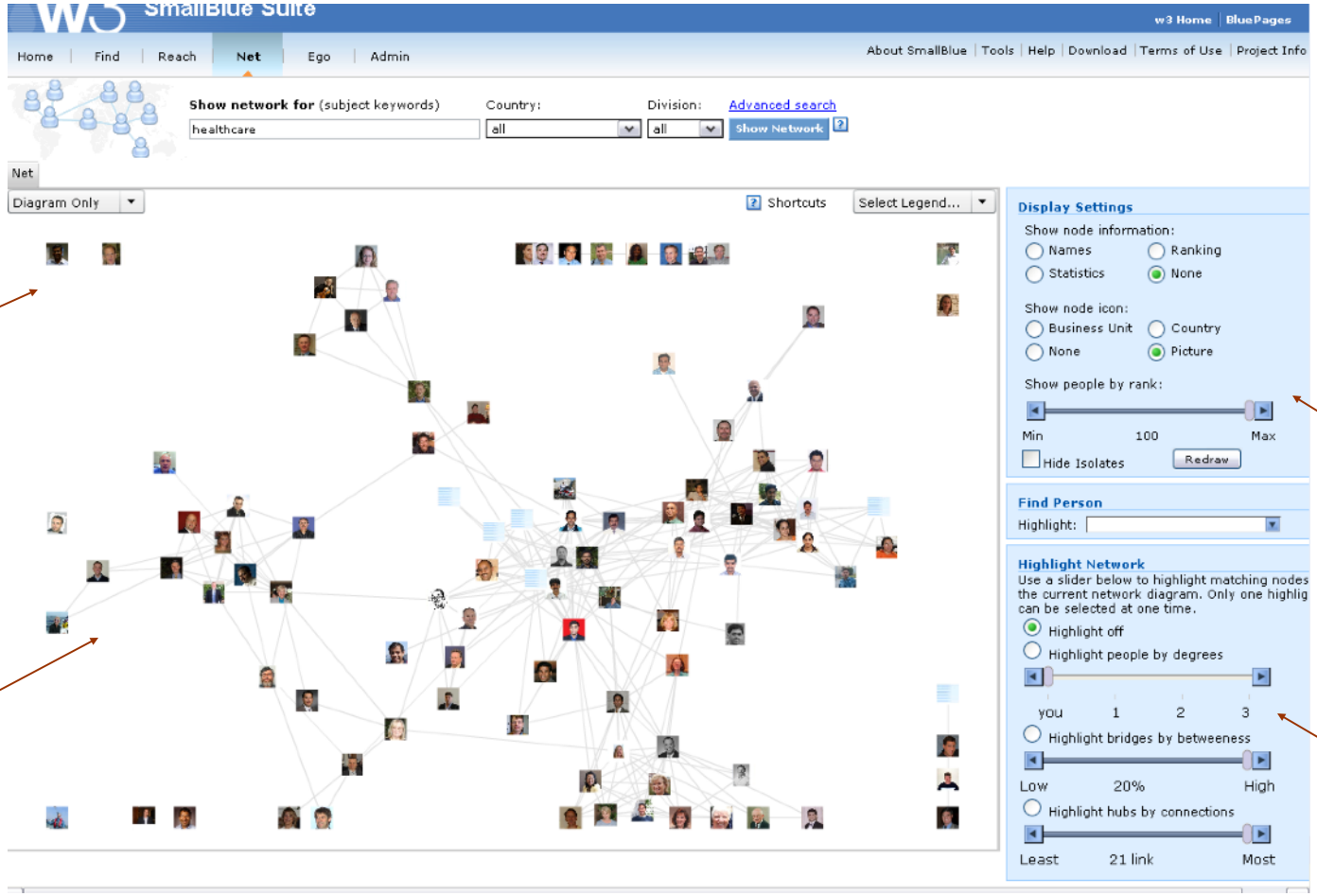
The public interest groups he is in

His blogs, forum, postings..

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

Net – corporate social network analysis

- **How are company's top healthcare experts link with each other? Who are the key bridges? Who have the most connections? How do these experts cluster?**



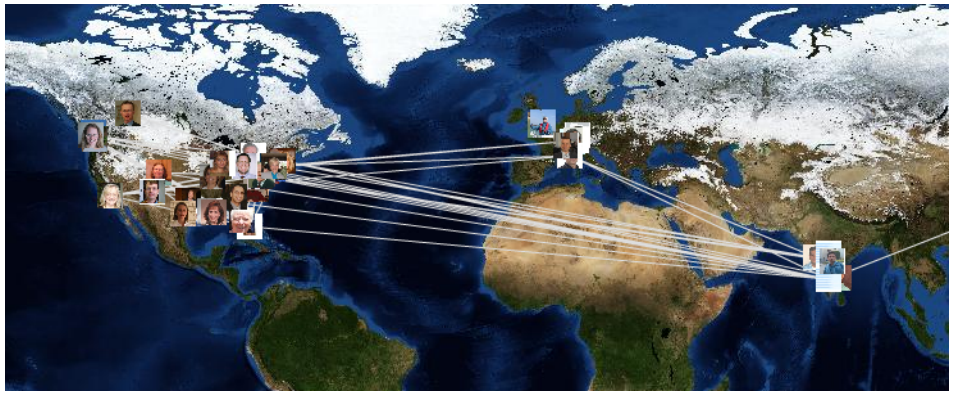
Independent experts on healthcare

A cluster of healthcare experts

Top-N experts on healthcare

Highlight experts based on my social proximity, the number of experts she connects, or the 'social bridges' importance

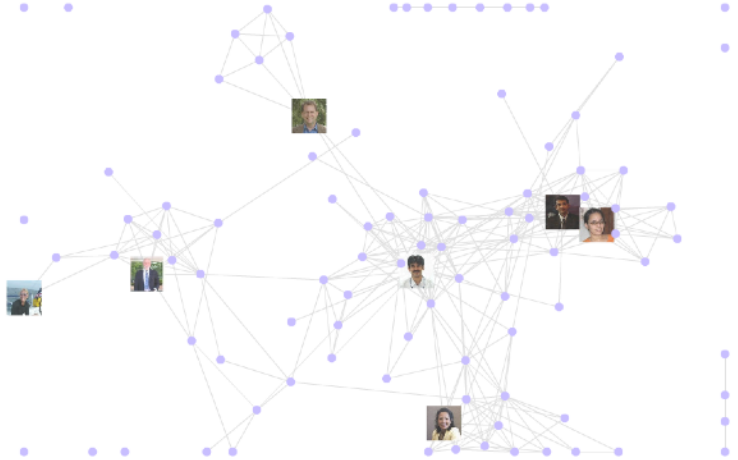
Net (cont'd)



Healthcare experts in the world

Healthcare experts in the U.S.

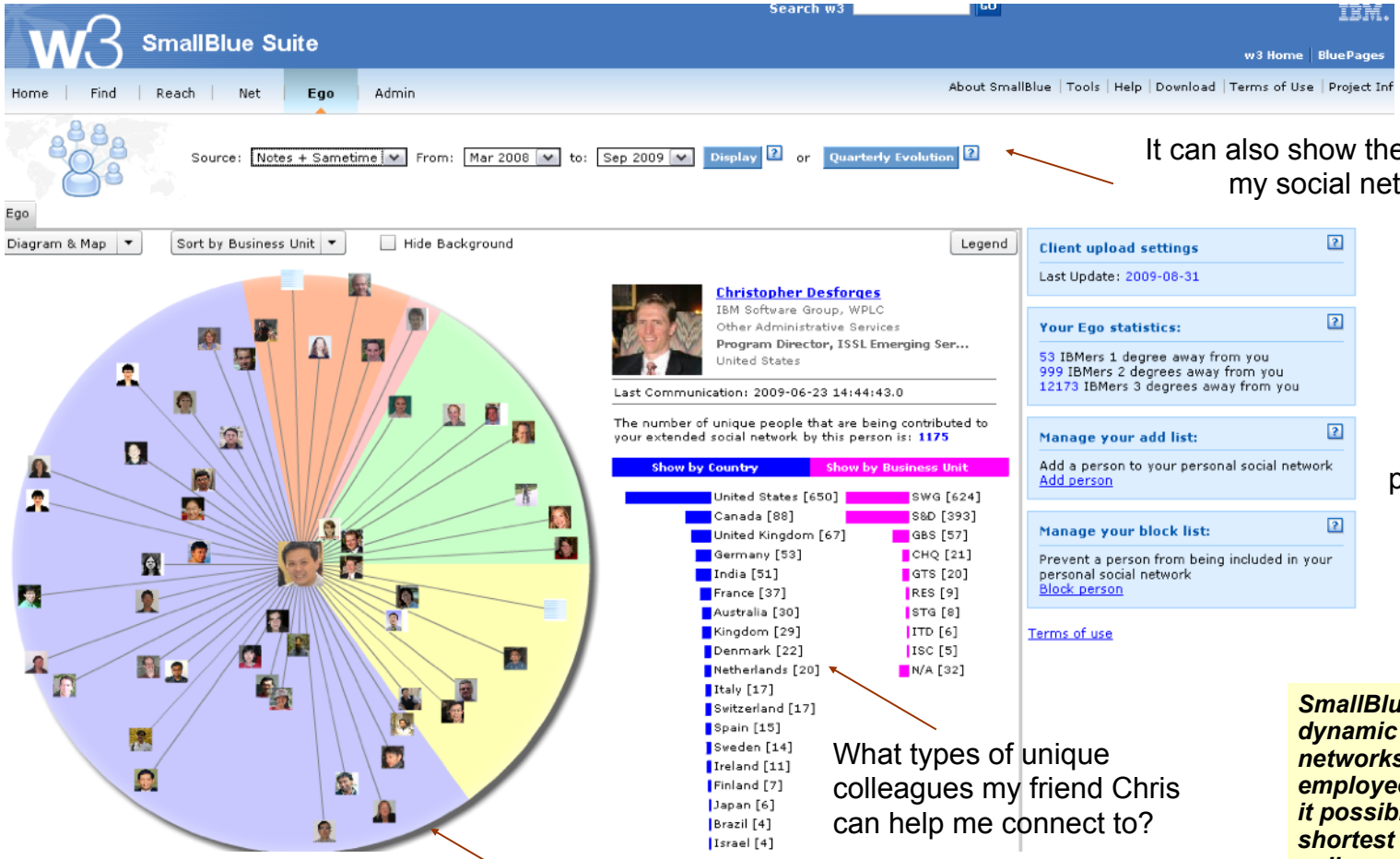
Connections between different divisions



Key social bridges

Ego – personal social network capital management [an application only visible to the user himself]

- What is a friend's social capital to me?



It can also show the evolution of my social network..

How many people in my personal networks?

What types of unique colleagues my friend Chris can help me connect to?

My personal social network automatically found by SmallBlue with social distance

SmallBlue finds dynamic existing social networks of every employee That makes it possible to find the shortest path to any colleague..

Personalized Content Recommendation and Search

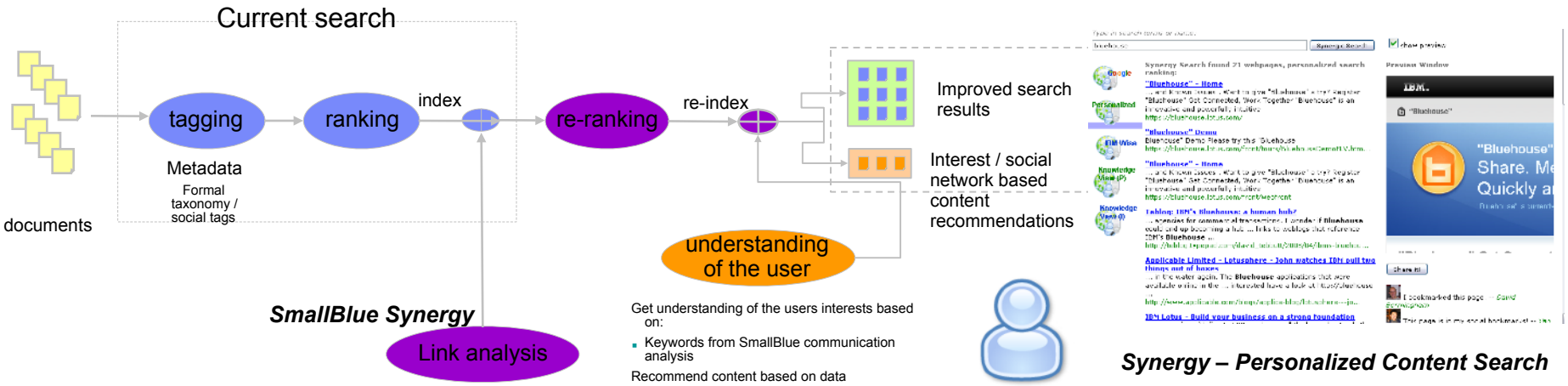
- What your friends know become what you know.. Your friends are your window to the world – *Confucius ~600 B.C.*
- Utilizing the unique large-scale weighted social network inferred by SmallBlue, personalized ranking becomes possible.
- Fusion of Recommenders: Social Filtering, Collaborative Filtering, Latent Semantic Filtering, Popularity & Freshness Filtering, etc.
- Deployed on IBM KnowledgeView, IBM Learning, and IBM TAP

SmallBlue Whisper

Ching-Yung Lin
IBM Research
Complex and Social Network Analy

46 IBMers are 1 degree away from you.
1278 IBMers are 2 degrees away from you.
14289 IBMers are 3 degrees away from you.

- **SmallBlue Whisper and Synergy usages:**
  - **On Nov. 19, 2008: 11,108**
  - **On Nov. 20, 2008: 9,567**



Synergy – Personalized Content Search

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

Productivity effect from network variables

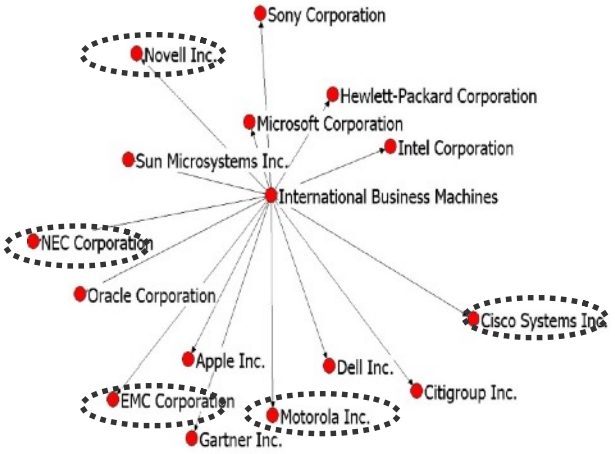
- An additional person in network size ~ $986 revenue per year
- Each person that can be reached in 3 steps ~ $0.163 in revenue per month
- A link to manager ~ $1074 in revenue per month
- 1 standard deviation of network diversity (1 - constraint) ~ $758
- 1 standard deviation of btw ~ -$300K
- 1 strong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

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

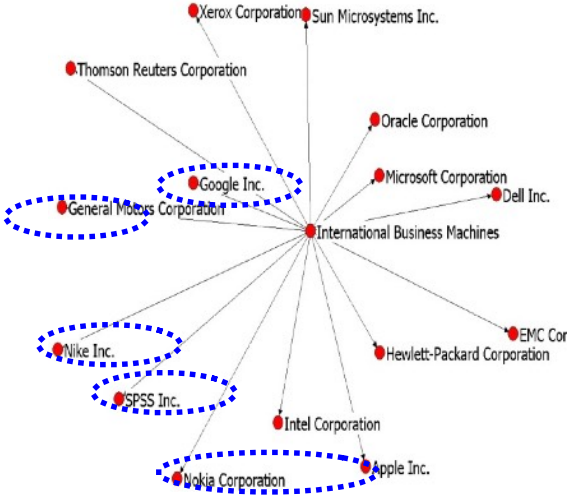
Graph Analytics for Financial Analysis

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

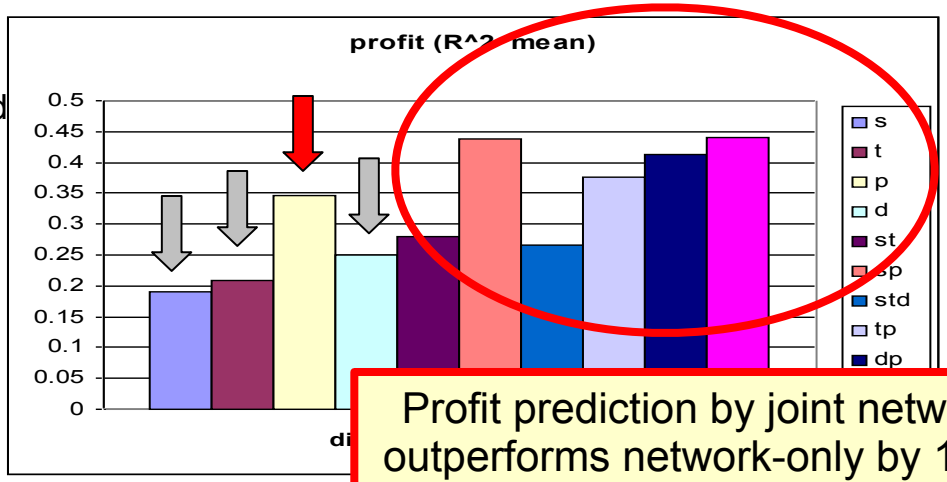


- Data Source:
  - Relationships among 7594 companies, data mining from NYT 1981 ~ 2009

Targets: 20 Fortune companies' normalized Profits

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

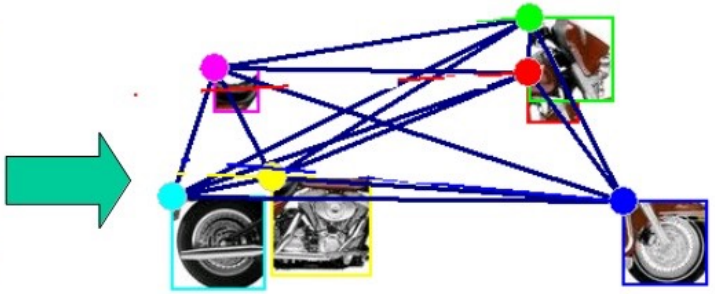
Model: Support Vector Regression (RBF kernel)

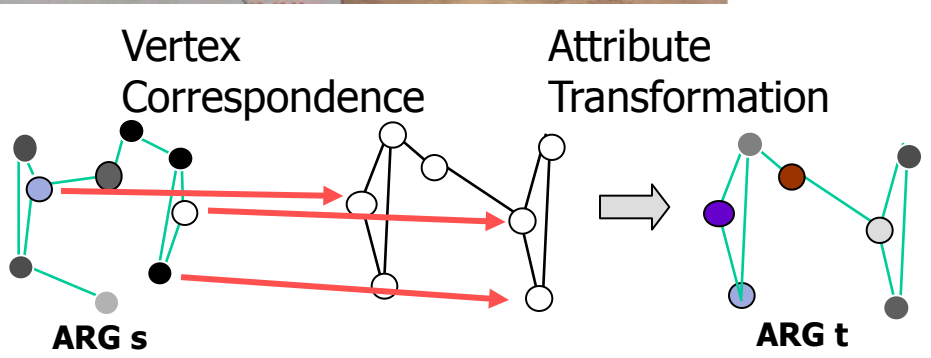


Network feature:
s (current year network feature),
t (temporal network feature),
d (delta value of network feature)
Financial feature:
p (historical profits and

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

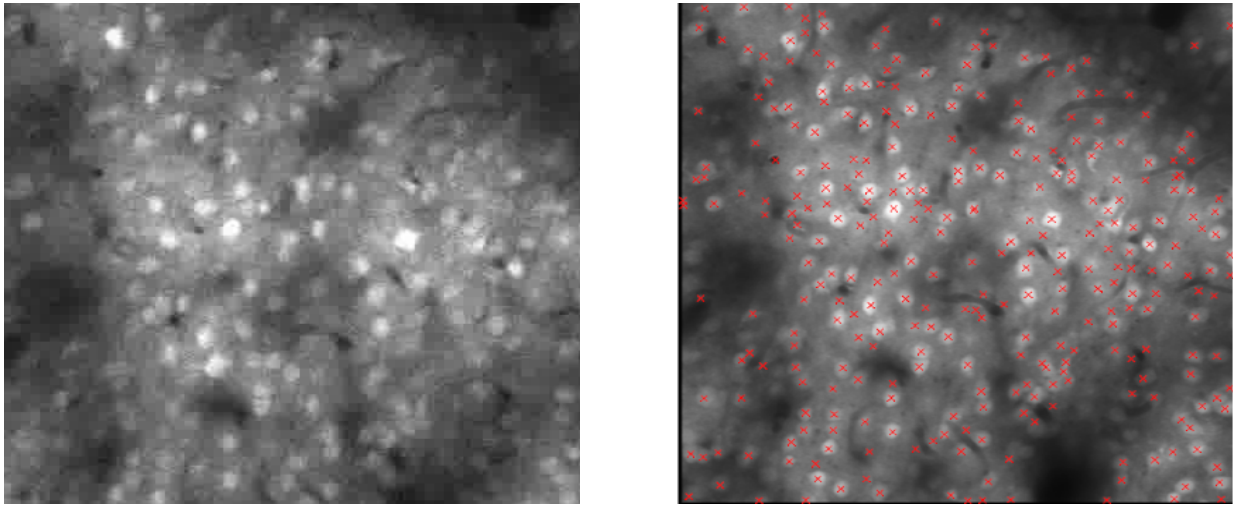
Graph Analysis for Image and Video Analysis





Mapping Brain into Big Network

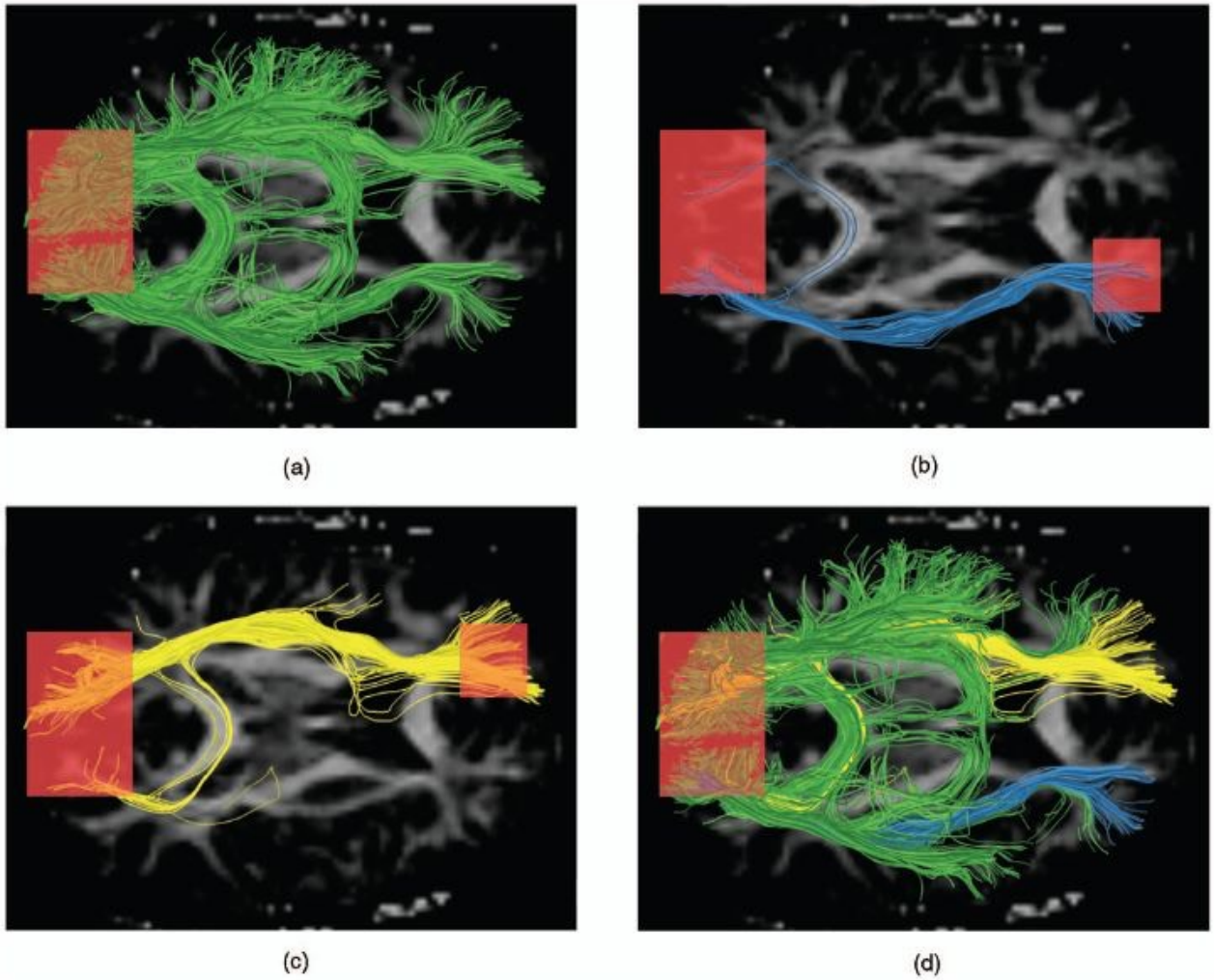
Contributing to the Brain Activity Mapping project, later renamed to BRAIN initiative



Signal capturing up to the neuron-level resolution

Demo: neuron detection

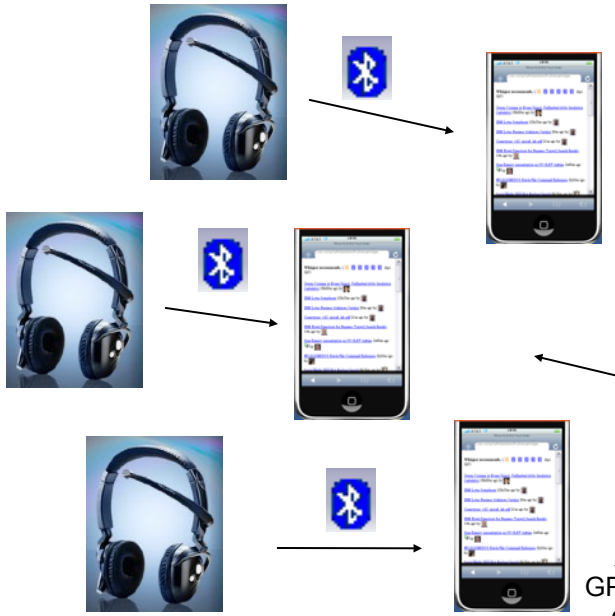
Cognitive Networks



Cognitive Network

- Cognitive [illegible]
  - 30,000 nodes of dynamic brain MRI functional networks

Composite Social-Cognitive-Info Networks

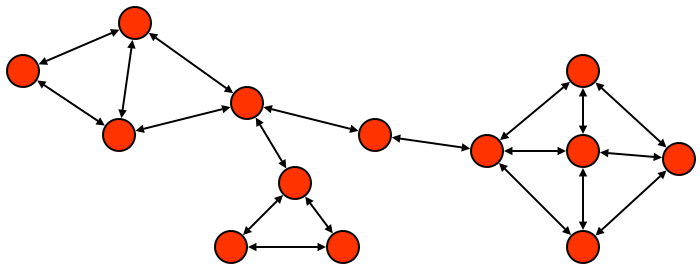


Status Monitoring,

Visualization, Personalized Information Recommendation, Routing, etc.

- A graph:

$$G = (V, E)$$



- V = Vertices or Nodes
- E = Edges or Links

- The number of vertices: "Order"

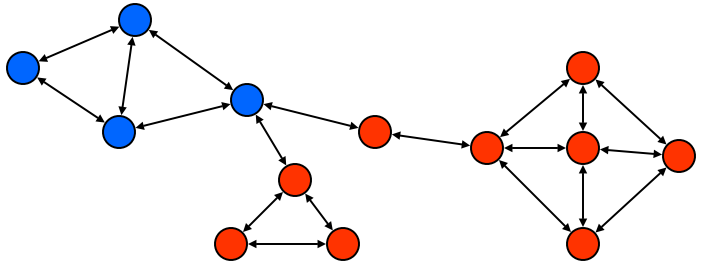
$$N_v = |V|$$

$$N_e = |E|$$

Subgraph

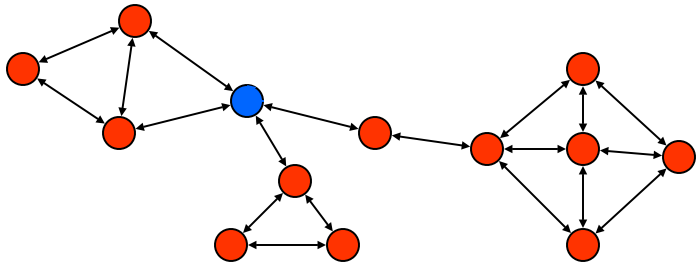
- A graph H is a subgraph of another graph G, if:

$$V_H \subseteq V_G \quad \text{and} \quad E_H \subseteq E_G$$

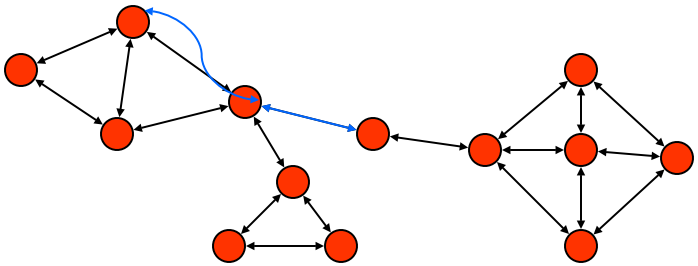


Multi-Graph vs. Simple Graph

- Loops:



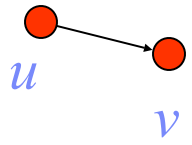
- Multi-Edges:



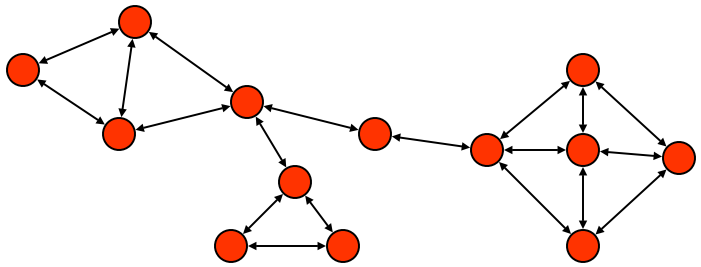
Directed Graph vs. Undirected Graph

- Directed Edges = Arcs:

$$\{u, v\}$$

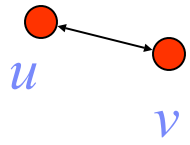


- Mutual arcs:

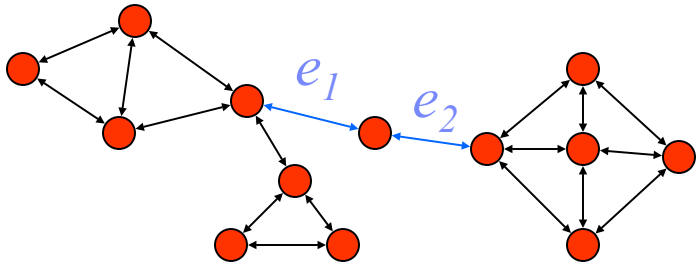


Adjacency

- u and v are adjacent if joined by an edge in E:

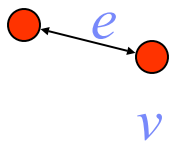


- Two edges are adjacent if joined by a common endpoint in V:

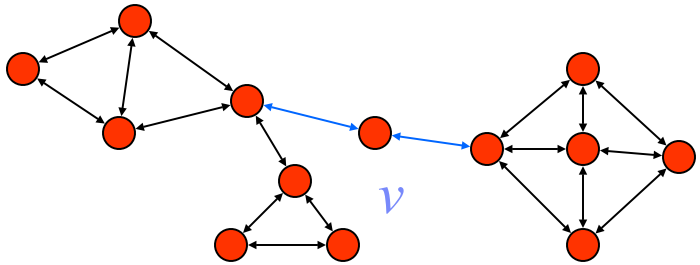


Incident and Degree

- A vertex $v \in V$ is incident on an edge $e \in E$ if v is an endpoint of e.



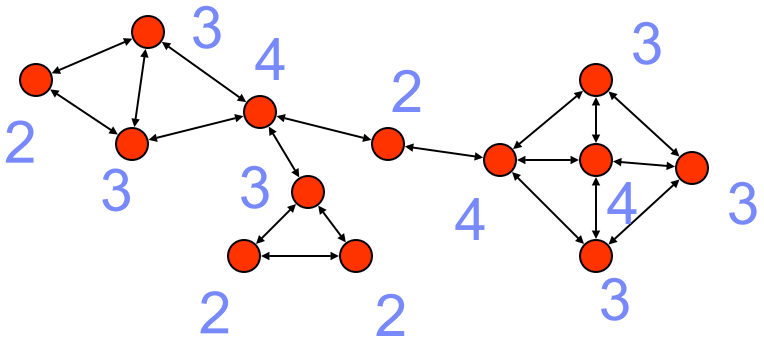
- The degree of a vertex v, say d_v, is defined as the number of edges incident on v.



$d_v = 2$

Degree Sequence

- The degree sequence of a graph G is the sequence formed by arranging the vertex degrees d_v in non-decreasing order.

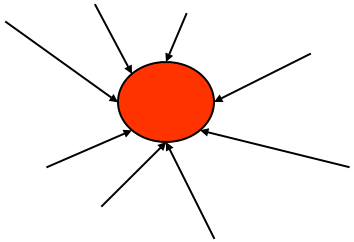


$$\{2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 4, 4, 4\}$$

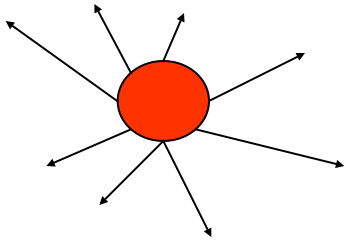
- The sum of the elements degree sequence equals to twice the number of edges in the graph (i.e. twice the size of the graph).

In-degrees and out-degrees

- For Directed graphs:



In-degree = 8



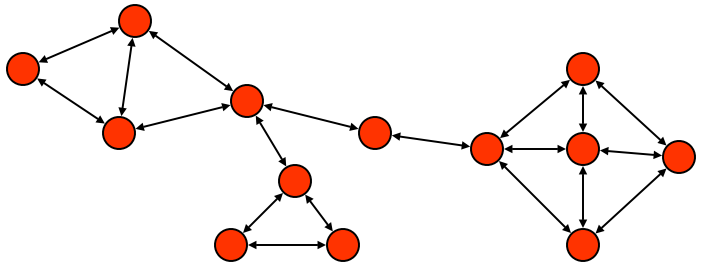
Out-degree = 8

Walk

- A walk on a graph G, from v_0 to v_l, is an alternating sequence:

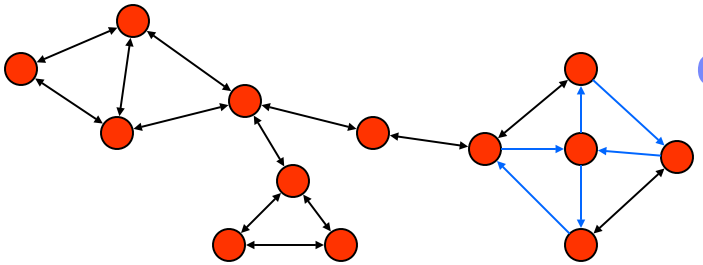
$$\{v_0, e_1, v_1, e_2, \ldots, v_{l-1}, e_l, v_l\}$$

- The length of this walk is l.
- A walk may be:
  - Trail --- no repeated edges
  - Path --- trails without repeated vertices.

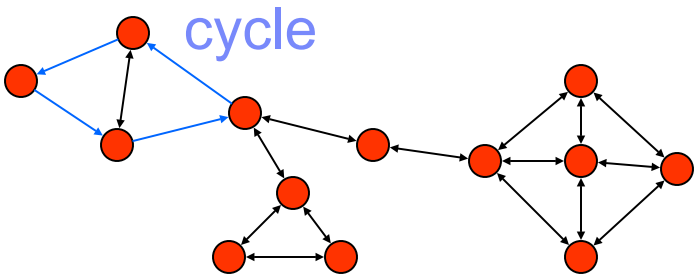


Circuit, Cycle, and Acyclic

- Circuit: A trail for which the beginning and ending vertices are the same.



  - Cycle: a walk of length at least three, the beginning node = ending node, all other nodes are distinct



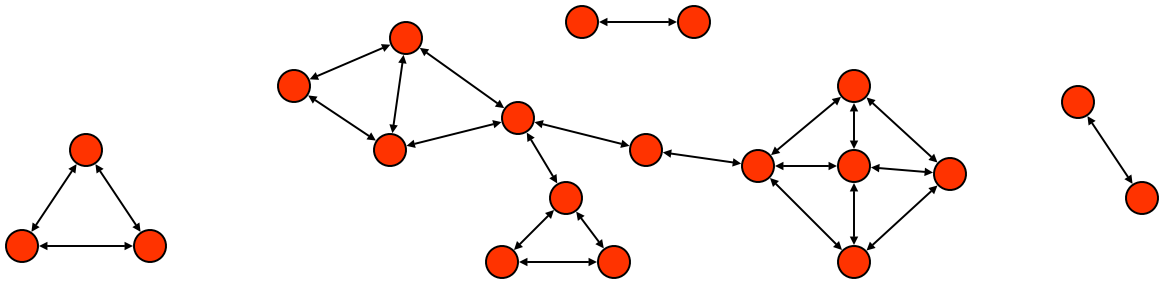
  - Acycle: graph contains no cycle

Reachable, Connected, Component

- Reachable: A vertex v in a graph G is said to be reachable from another vertex u if there exists a walk from u to v.

- Connected: A graph is said to be connected if every vertex is reachable from every other.

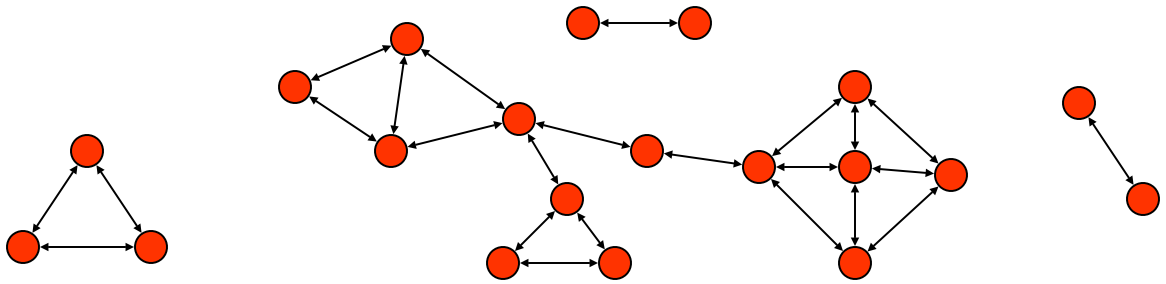
- Component: A component of a graph is a maximally connected subgraph.



Connection in a digraph

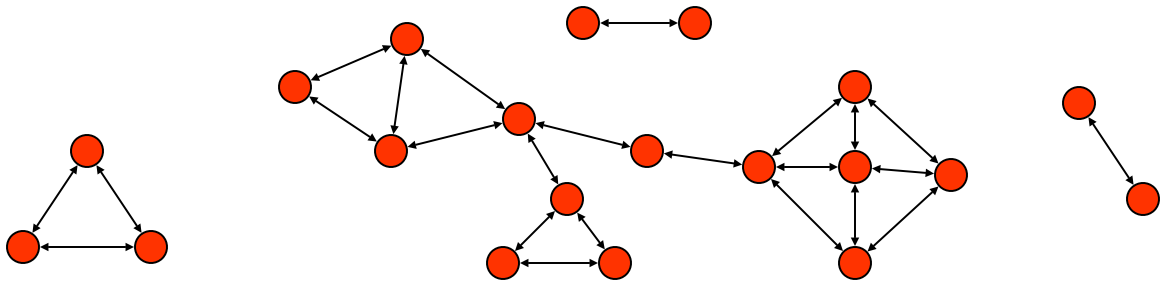
- Weakly connected: If its underlying graph is connected after stripping away the direction.

- Strongly connected: every vertex is reachable from every other vertex by a directed walk.



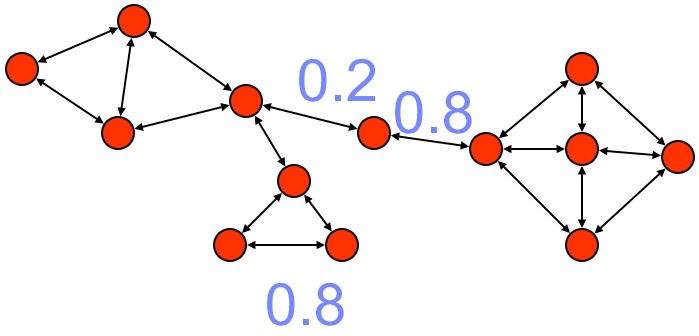
Distance

- Distance of two vertices: The length of the shortest path between the vertices.
- Geodesic: another name for shortest path.
- Diameter: the value of the longest distance in a graph



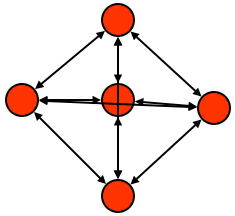
Decorated Graph

- Weighted Edges

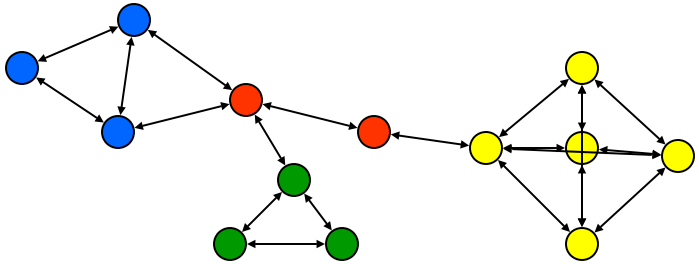


Families of Graphs

- Complete Graph: every vertex is linked to every other vertex.

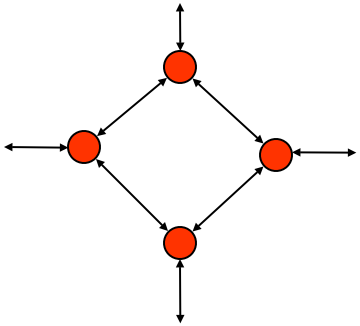


- Clique: a complete subgraph.



Regular Graph

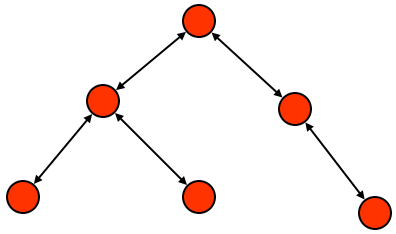
- Regular Graph: a graph in which every vertex has the same degree.



a 3-regular graph

Tree and Forest

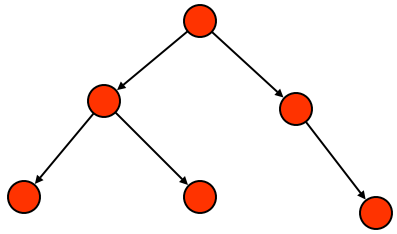
- Tree: a connected graph with no cycle.



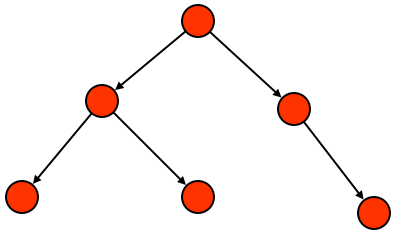
- Forest: a disjoint union of trees is called a forest.

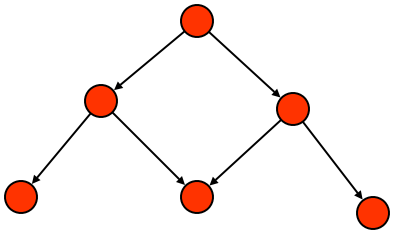
Labels in a directed tree

- Root
- Ancestor
- Descendant
- Parent
- Children
- Leaf: a vertex without children



Rooted Tree vs Directed Acyclic Graph (DAG)

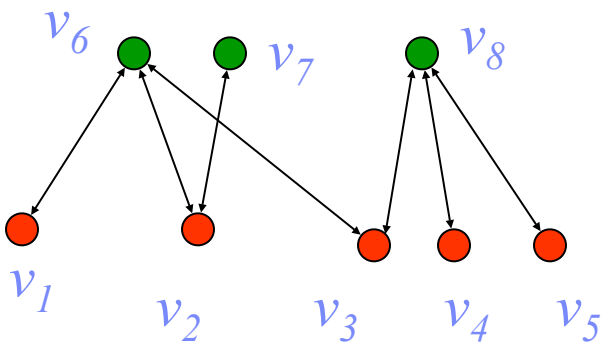




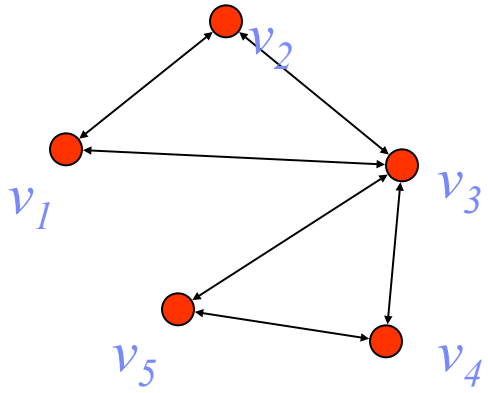
- DAG: Directed Acycle Graph. Underlining undirected graph has cycle.

Bipartite Graph

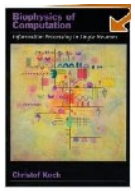
- Bipartite Graph: Vertices are partitioned into two sets. Edges link only between these two sets.



- Induced Graph (Collaborative Filtering):



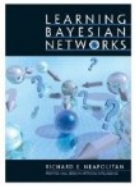
Recommendation Technique – Collaborative Filtering



Customers who bought this item also bought

Theoretical Neuroscience: Computational and Mathematic*a*
Dayan

Biophysics of Computation : Information Processing in Sir
Neuroscience Series) by *Christof Koch*

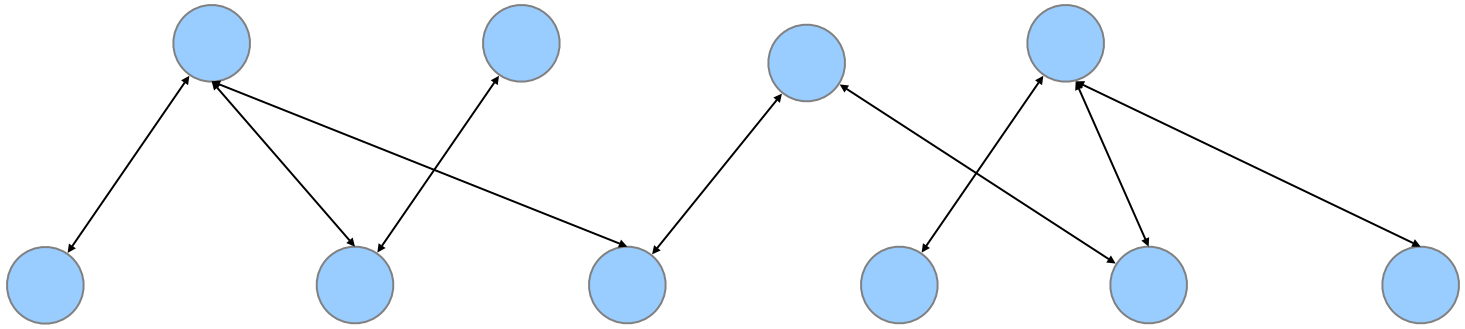


Customers who bought this item also bought

Bayesian Statistics : An Introduction (A Hodder Arnold Pu
Markov Chain Monte Carlo in Practice by *W.R. Gilks*
Monte Carlo Statistical Methods (Springer Texts in Statist
Bayes and Empirical Bayes Methods for Data Analysis, Se
The Elements of Statistical Learning by *T. Hastie*

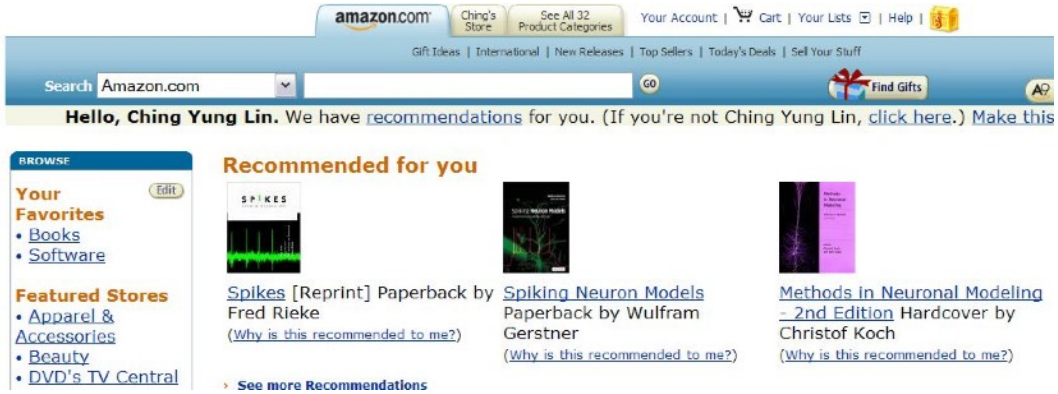
item

user



=

Recommendation



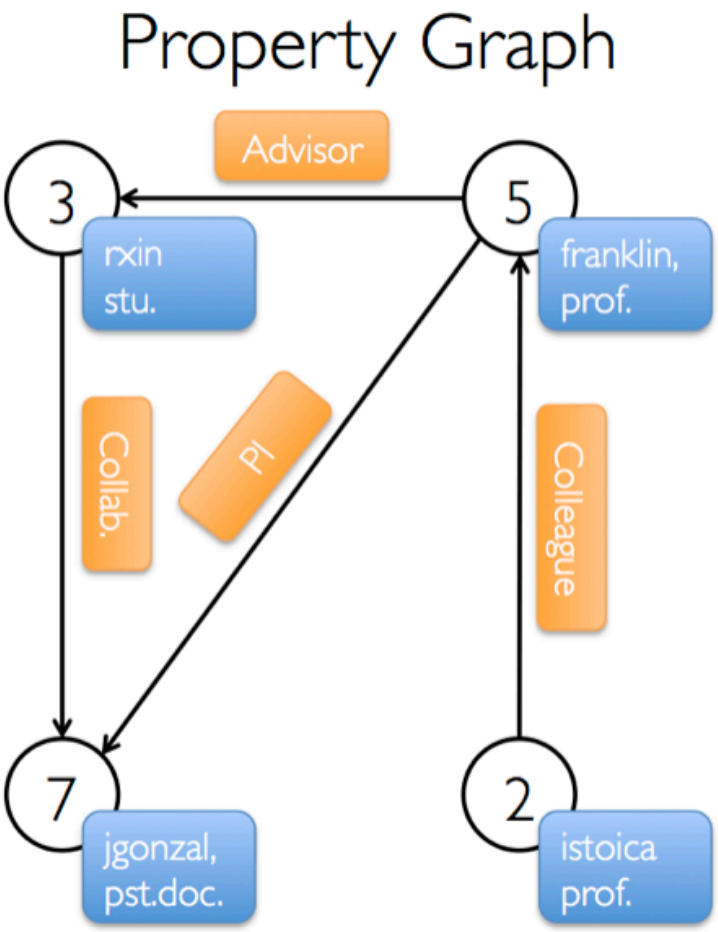
Graphs and Matrix Algebra

- The fundamental connectivity of a graph G may be captured in an $N_v \times N_v$ binary symmetric matrix $\boldsymbol{A}$ with entries:

$$A_{ij} = \begin{cases} 1, & if \{i, j\} \in E \\ 0, & otherwise \end{cases}$$

$\boldsymbol{A}$ is called the Adjacency Matrix of G

Property Graph



Vertex Table

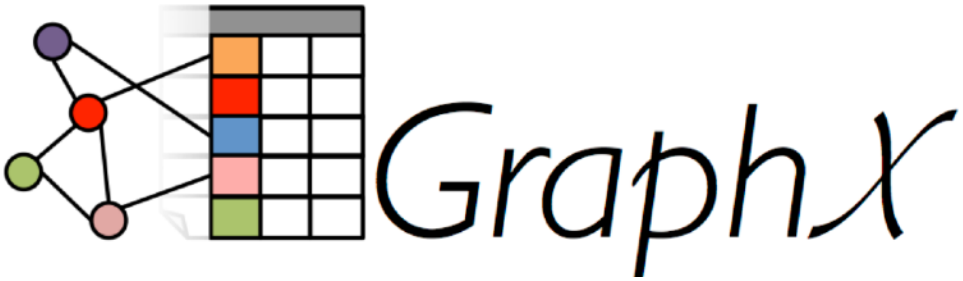
Id	Property (V)
3	(rxin, student)
7	(jgonzal, postdoc)
5	(franklin, professor)
2	(istoica, professor)

Edge Table

SrcId	DstId	Property (E)
3	7	Collaborator
5	3	Advisor
2	5	Colleague
5	7	PI

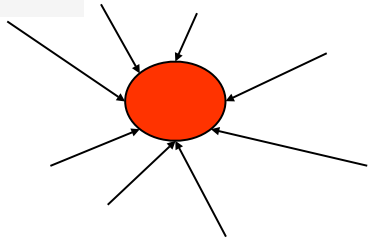
Spark GraphX

- The Property Graph
  - Example Property Graph
- Graph Operators
  - Summary List of Operators
  - Property Operators
  - Structural Operators
  - Join Operators
  - Neighborhood Aggregation
    - Aggregate Messages (aggregateMessages)
    - Map Reduce Triplets Transition Guide (Legacy)
    - Computing Degree Information
    - Collecting Neighbors
  - Caching and Uncaching
- Pregel API
- Graph Builders
- Vertex and Edge RDDs
  - VertexRDDs
  - EdgeRDDs
- Optimized Representation
- Graph Algorithms
  - PageRank
  - Connected Components
  - Triangle Counting

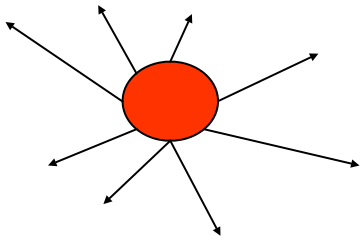


GraphX Graph Operations

```
// Information about the Graph ======
val numEdges: Long
val numVertices: Long
val inDegrees: VertexRDD[Int]
val outDegrees: VertexRDD[Int]
val degrees: VertexRDD[Int]
```

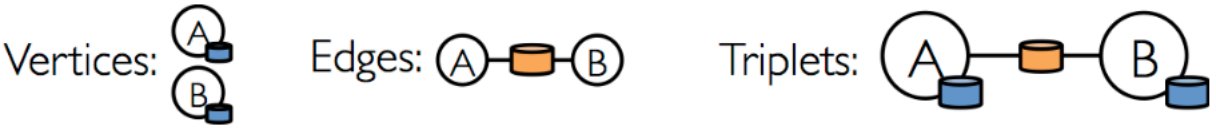


In-degree = 8



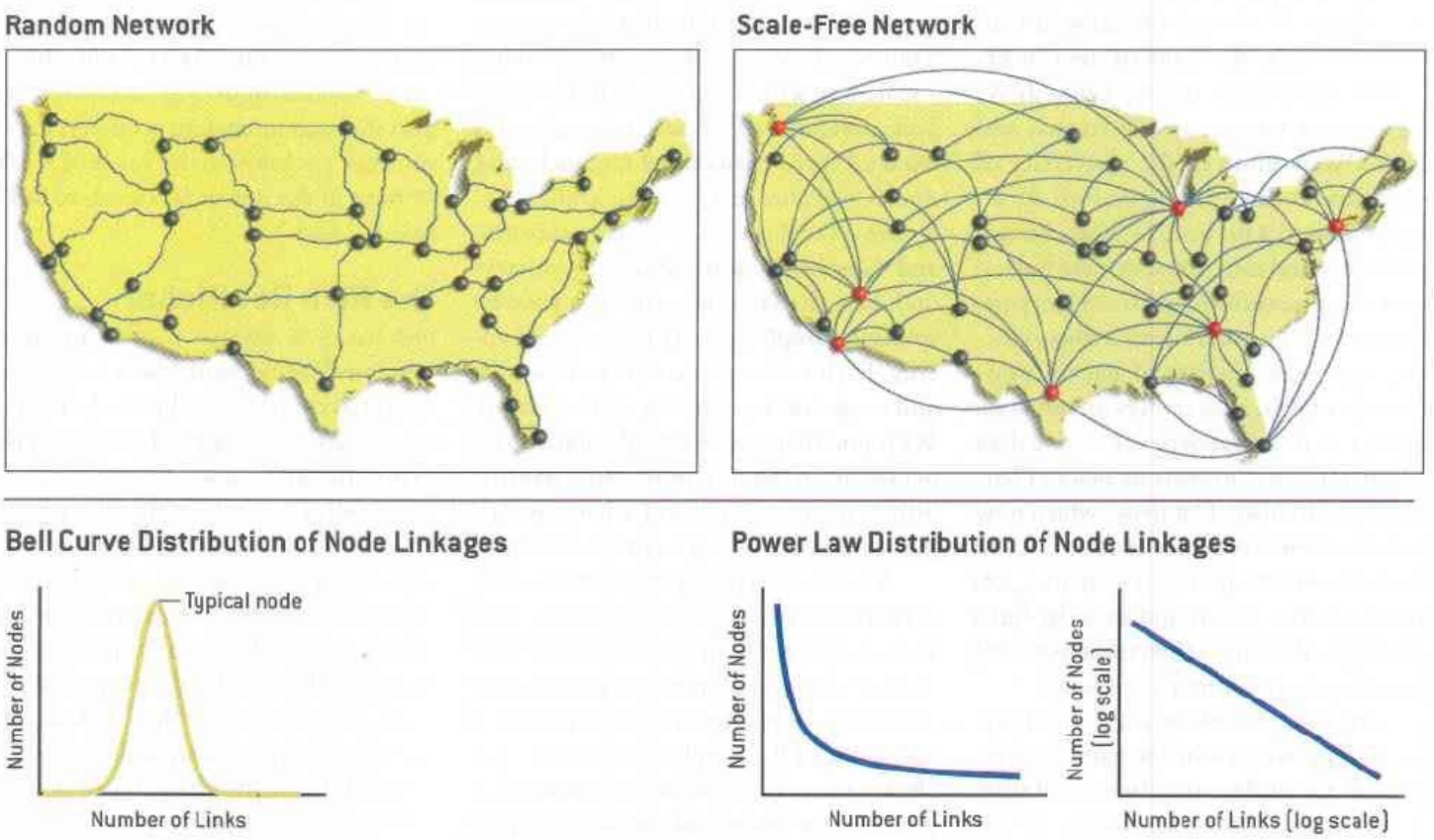
Out-degree = 8

```
// Views of the graph as collections ===
val vertices: VertexRDD[VD]
val edges: EdgeRDD[ED]
val triplets: RDD[EdgeTriplet[VD, ED]]
```



Degree Distribution Example: Power-Law Network

A. Barbasi and E. Bonabeau, “Scale-free Networks”, Scientific American 288: p.50-59, 2003.



$$p_k = e^{-m} \cdot \frac{m^k}{k!}$$

$$p_k = C \cdot k^{-\tau} e^{-k/\kappa}$$

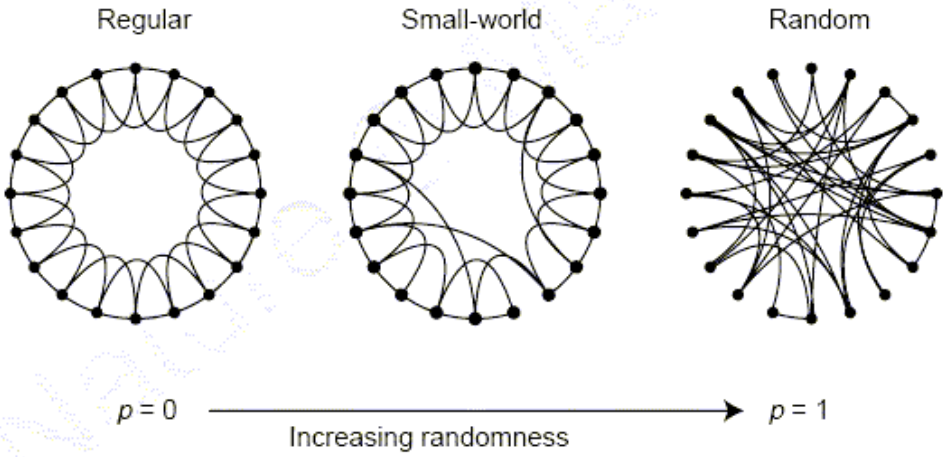
Newman, Strogatz and Watts, 2001

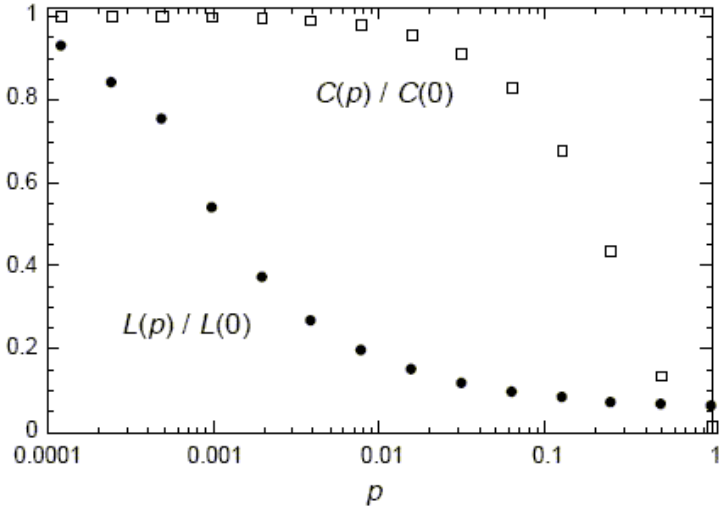
Another example of complex network: Small-World Network

Six Degree Separation:

adding long range link, a regular graph can be transformed into a small-world network, in which the average number of degrees between two nodes become small.

from Watts and Strogatz, 1998

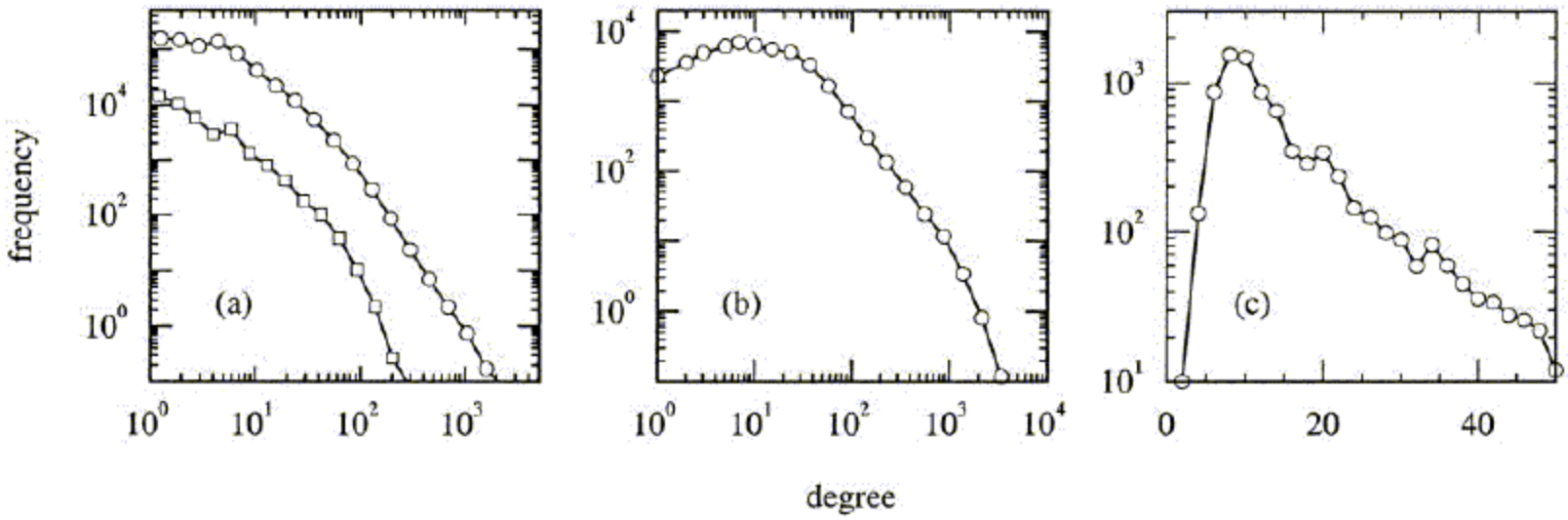




C: Clustering Coefficient, L: path length,
(C(0), L(0)): (C, L) as in a regular graph
(C(p), L(p)): (C,L) in a Small-world graph with randomness p.

Some examples of Degree Distribution

(a) scientist collaboration: biologists (circle) physicists (square), (b) collaboration of move actors, (d) network of directors of Fortune 1000 companies



Basic graph algorithms in GraphX

Columbia University

```
// Basic graph algorithms ========================================================
def pageRank(tol: Double, resetProb: Double = 0.15): Graph[Double, Double]
def connectedComponents(): Graph[VertexId, ED]
def triangleCount(): Graph[Int, ED]
def stronglyConnectedComponents(numIter: Int): Graph[VertexId, ED]
```

Centrality

"There is certainly no unanimity on exactly what centrality is or its conceptual foundations, and there is little agreement on the procedure of its measurement." – Freeman 1979.

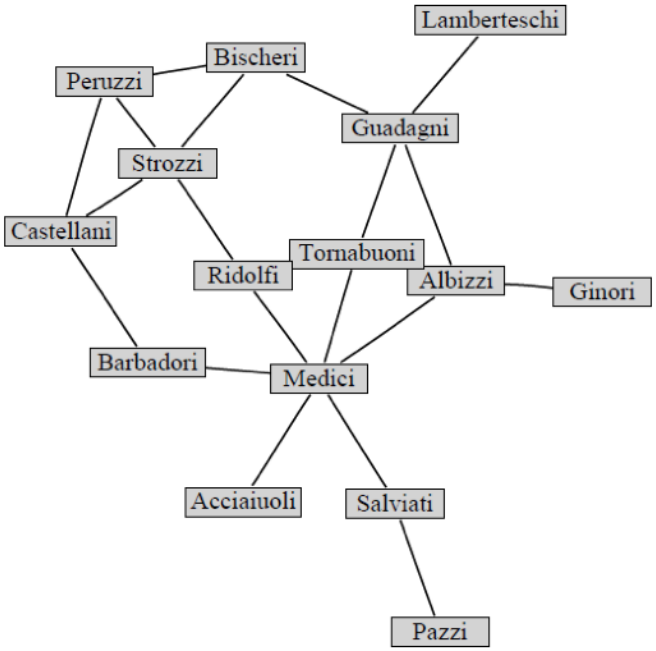
Degree (centrality)

Closeness (centrality)

Betweeness (centrality)

Eigenvector (centrality)

Network Analysis Example Centrality Ranking in Large Networks



[15th Century Florentine Family]

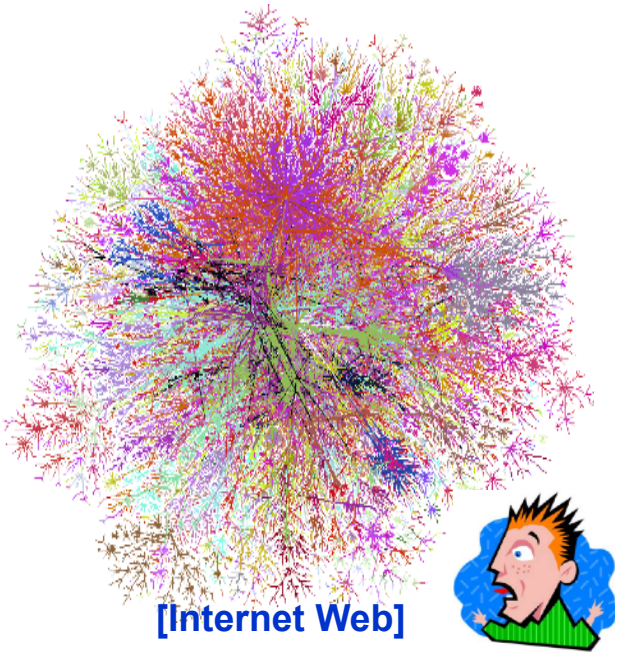
"Who are the most important actors?"

Three centralities

Degree: # of neighbor
Closeness: avg. shortest path length
Betweenness: # of times a node sits between shortest path

Application

Measuring the financial company value
Network attack monitoring



[Internet Web]

|V| = 15 ***|E| = 19***

Degree : Easy ***O(|E|)***

Closeness : Easy ***O(|V|3)***

Betweenness : Easy ***O(|V|2log|V|)***

|V| = Billions ***|E| = Billions***

Degree : Easy

Closeness : Hard

Betweenness : Hard

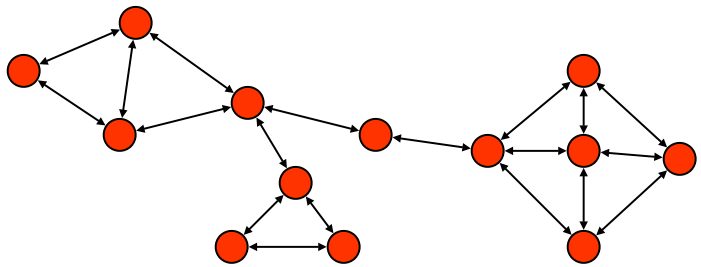
For 2 Billon Edges,
- standard closeness: 30,000 years

Closeness

Closeness: A vertex is ‘close’ to the other vertices

$$c_{CI}(v) = \frac{1}{\sum_{u \in V} dist(v,u)}$$

where dist(v,u) is the geodesic distance between vertices v and u.



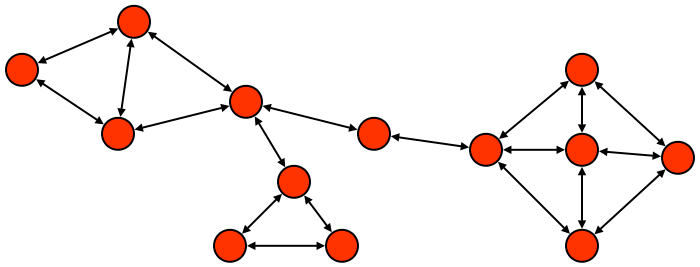
Betweenness

Betweenness measures are aimed at summarizing the extent to which a vertex is located 'between' other pairs of vertices.

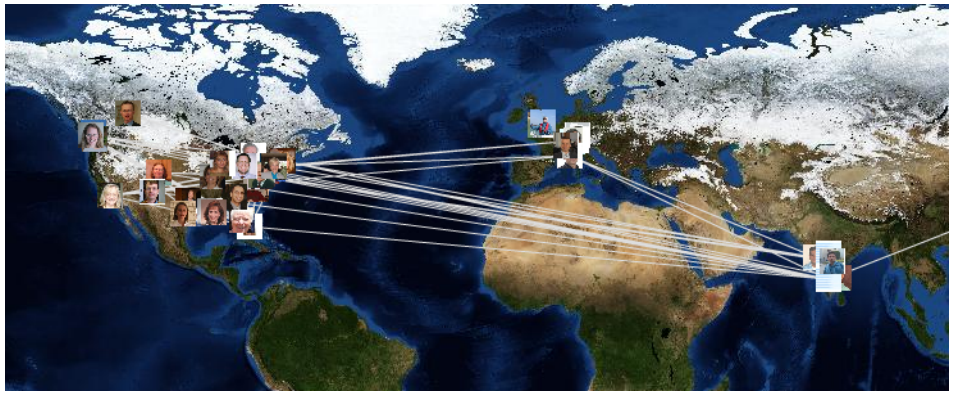
Freeman's definition:

$$c_B(v) = \sum_{s \neq t \neq v \in V} \frac{\sigma(s,t \mid v)}{\sigma(s,t)}$$

Calculation of all betweenness centralities requires
calculating the lengths of shortest paths among all pairs of vertices
Computing the summation in the above definition for each vertex



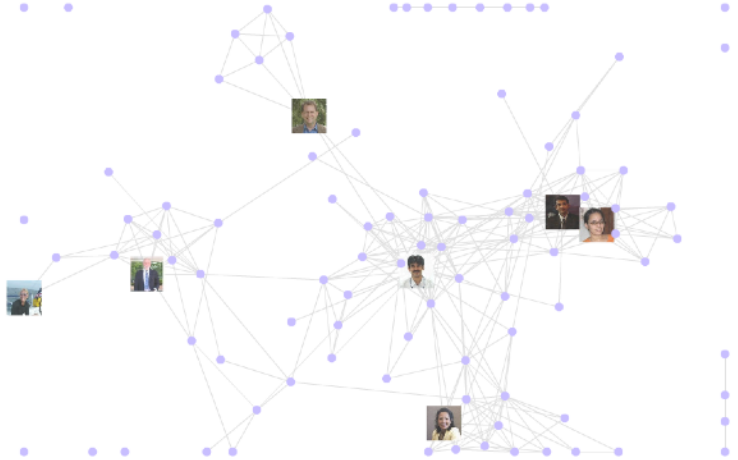
Betweenness ==> Bridges



Example: Healthcare experts in the world

Example: Healthcare experts in the U.S.

Connections between different divisions



Key social bridges

Eigenvector Centrality

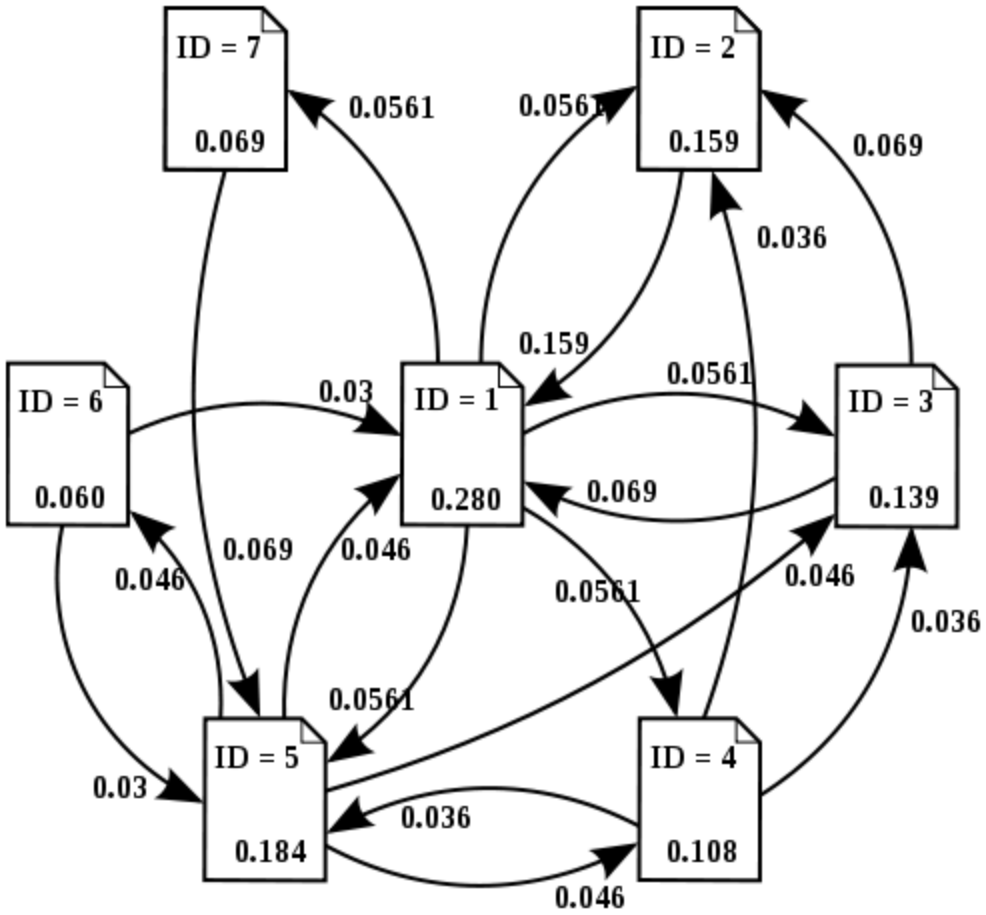
Try to capture the ‘status’, ‘prestige’, or ‘rank’.
More central the neighbors of a vertex are, the more central the vertex itself is.

$$c_{Ei}(v) = \alpha \sum_{\{u,v\} \in E} c_{Ei}(u)$$

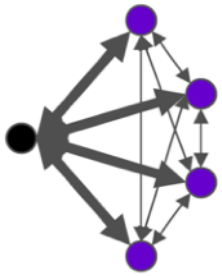
The vector $\mathbf{c}_{Ei} = (c_{Ei}(1), \ldots, c_{Ei}(N_v))^T$ is the solution of the eigenvalue problem:

$$\mathbf{A} \cdot \mathbf{c}_{Ei} = \alpha^{-1} \mathbf{c}_{Ei}$$

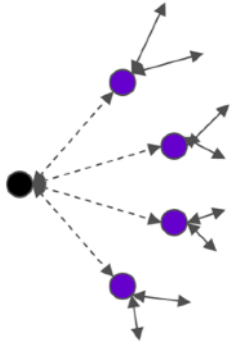
PageRank Algorithm (Simplified)



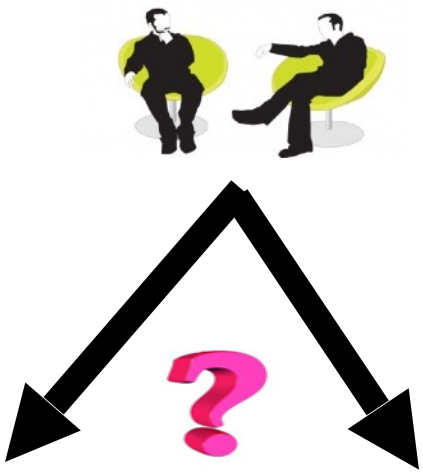
- Topological point of views
  - What type of network structure is beneficial?



Cohesive Network
- Trust
- Absorptive capacity
- Precision, Reliability



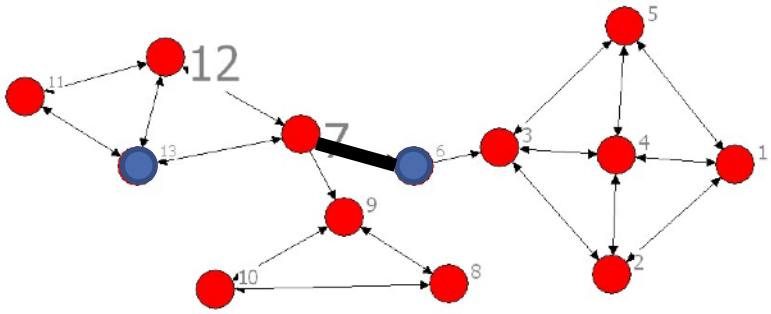
Structurally Diverse Network
- Brokering position
- Access to many pools of diverse, novel information



What type of network structure is most beneficial in a electronic network for consultants?

- Importance of Direct Contacts?
- Importance of Indirect Contacts?
- Constrained vs. unconstrained?

Network Topology Measures



Direct Contacts	Size(7) = 4 Size(12)= 3	+ No information distortion - High maintenance cost	Network size → strong work performance (?)
Indirect Contacts	Btw(7)= 33 Btw(12)=6 3steps(7) =11 3steps(12)=8	+ Access diverse information - Information distortion	Btw-centrality → Strong work performance (?) 3-step Reach →Strong work performance (?)
Structural Diversity	Div(7)=.53 Div12)=0.16	+Transfer complex knowledge - Access diverse knowledge	Diversity→ Strong work performance (?)

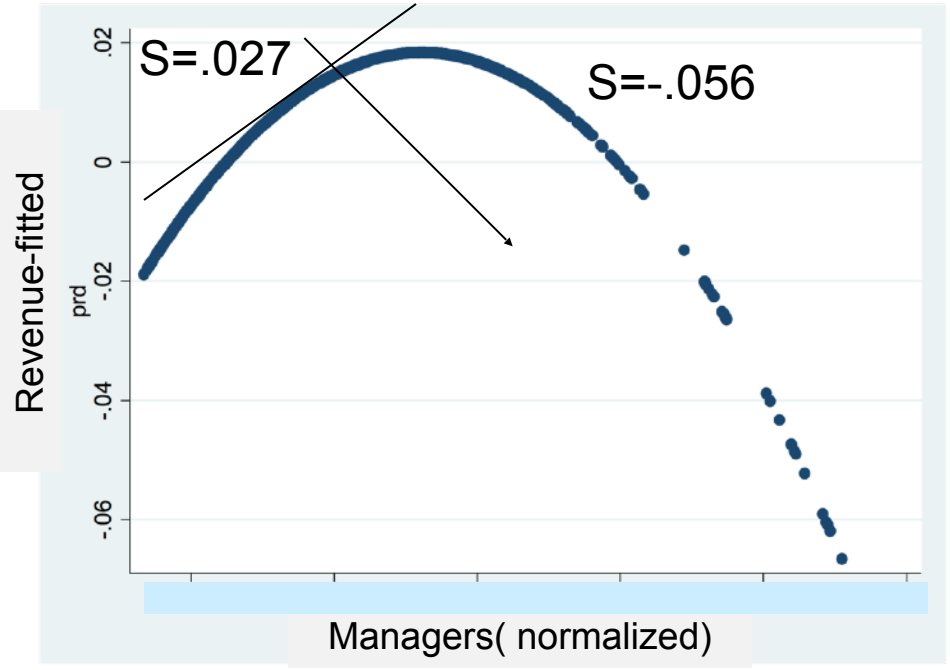
Project Team Composition—Managers

The number of managers in a project exhibit an inverted-U shaped curve.

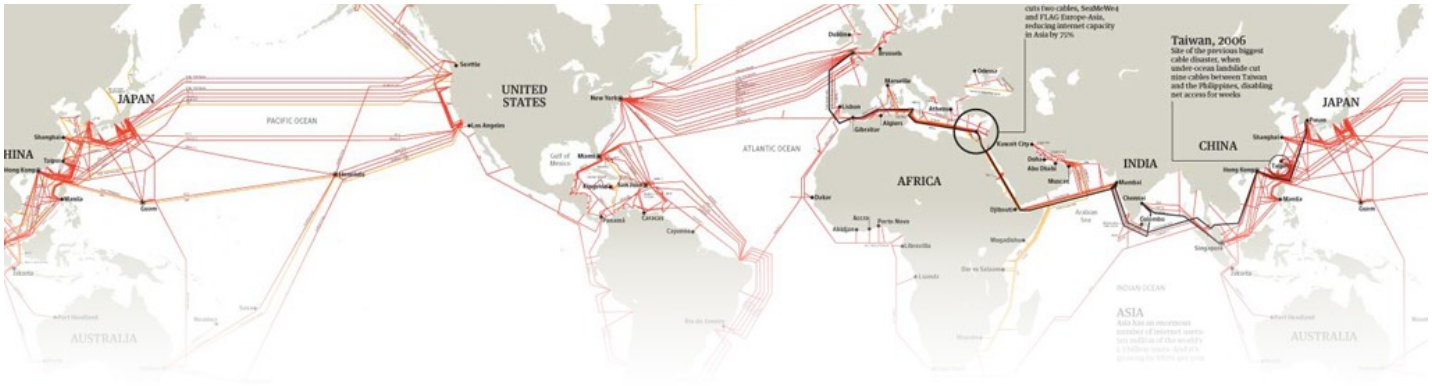
1. Having managers in a project is correlated with team performance initially.
2. Too many managers in a project is negatively associated with team performance.

$$revenue = \alpha + \beta_1 \cdot mgr + \beta_2 \cdot mgr^2 + \gamma_1 \cdot otherfactor_1 + ... + \gamma_k \cdot otherfactor_k + \varepsilon$$

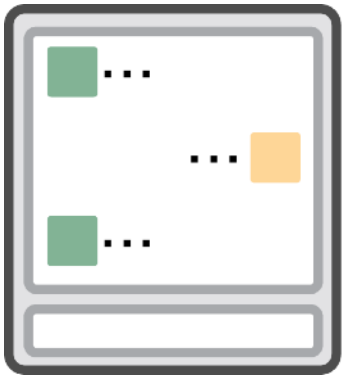
# Managers in project	β_1	2733.9*** (537.5)
(# Managers in project) ^2	β_2	-682.02*** (215.3)



Culture Factor in CMC-based Communications



Collaborating Globally:

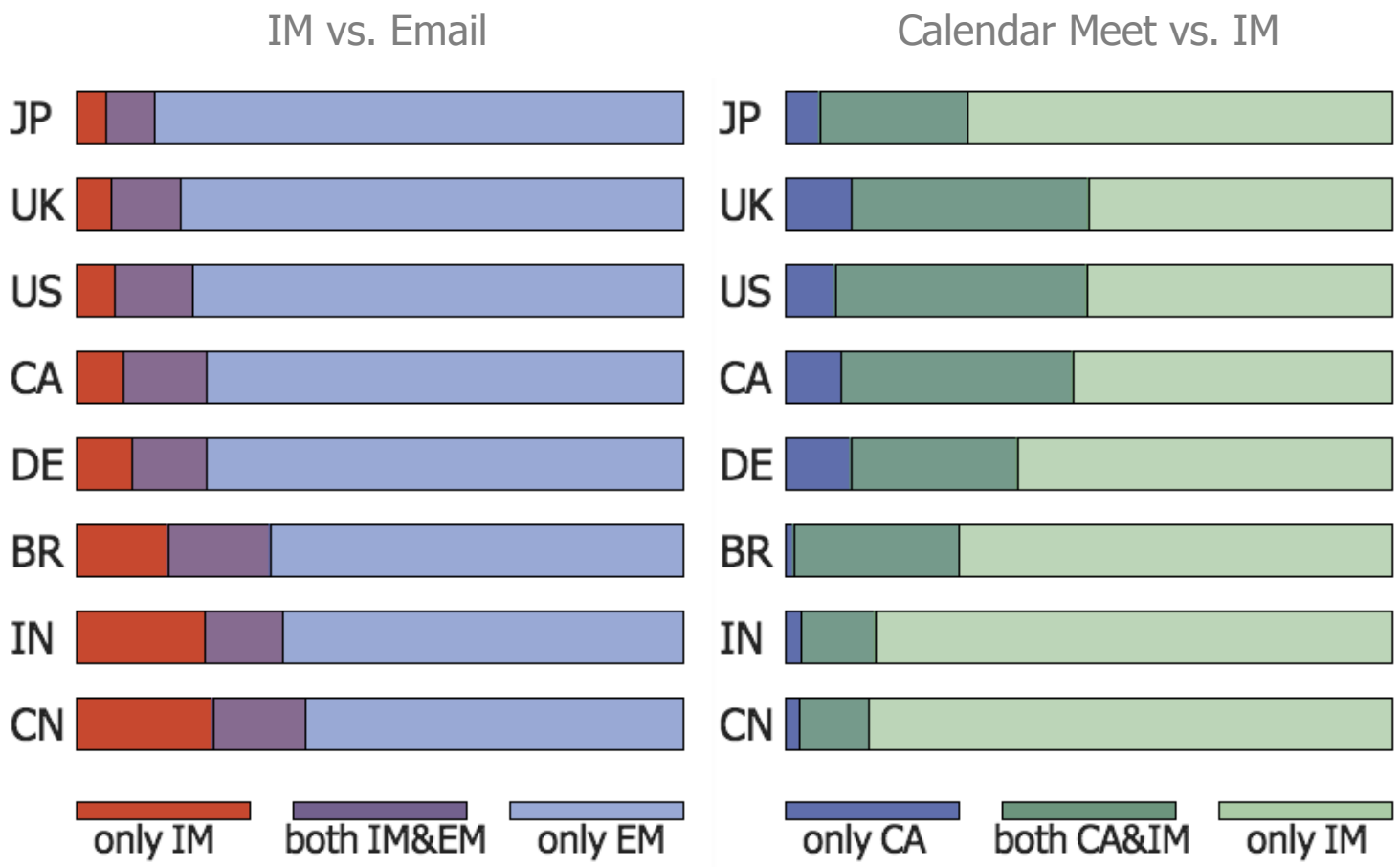


preferences of CMC tools

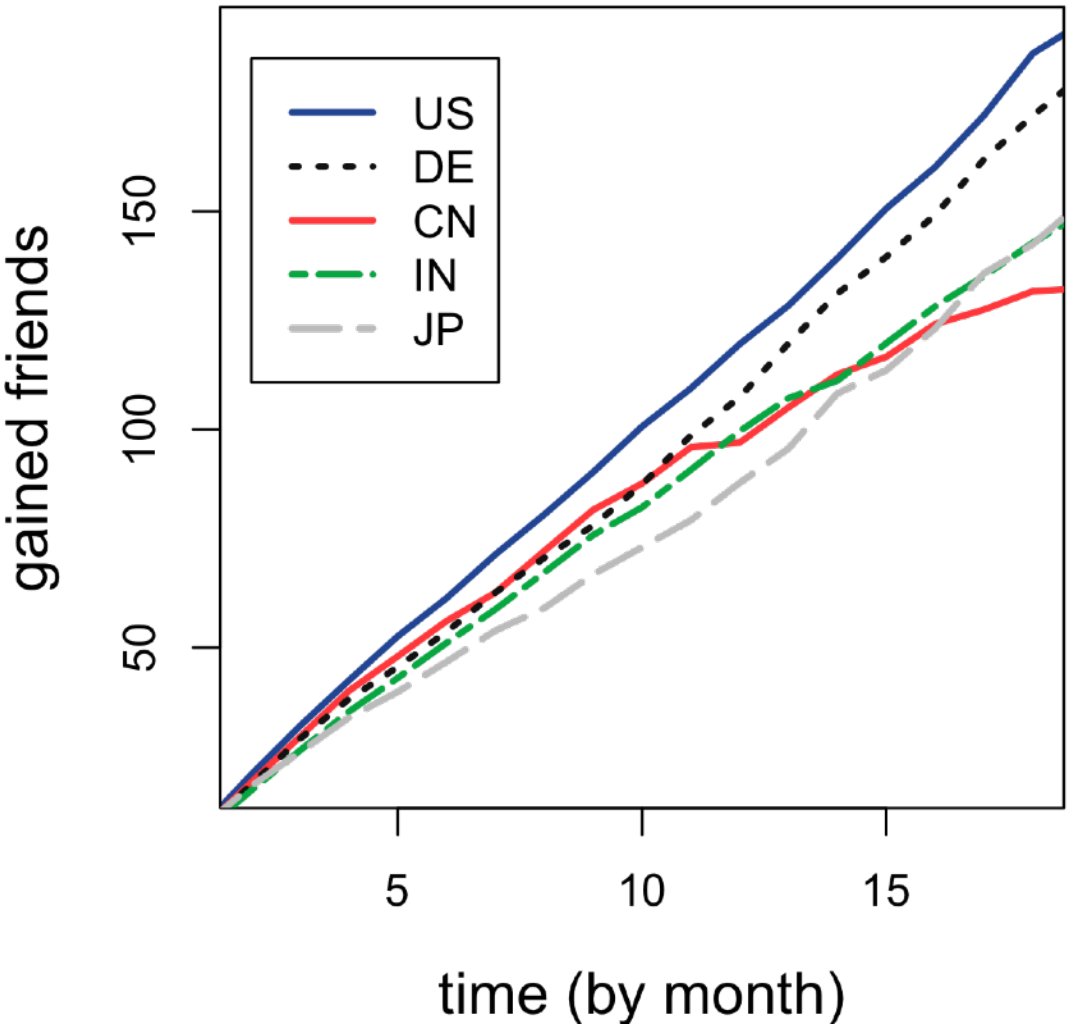
patterns of growing social network

sentiments in conversations

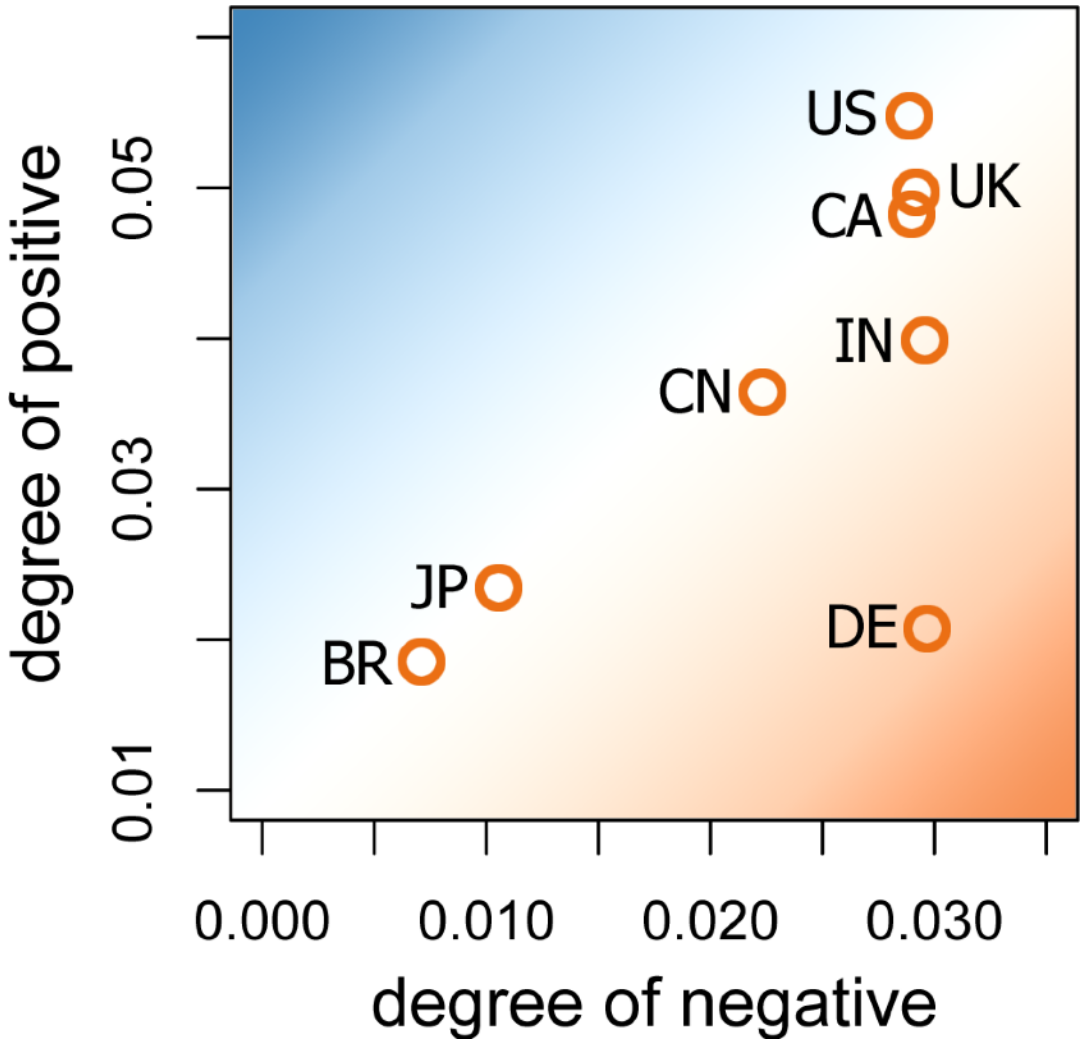
Preferences of CMC Tools

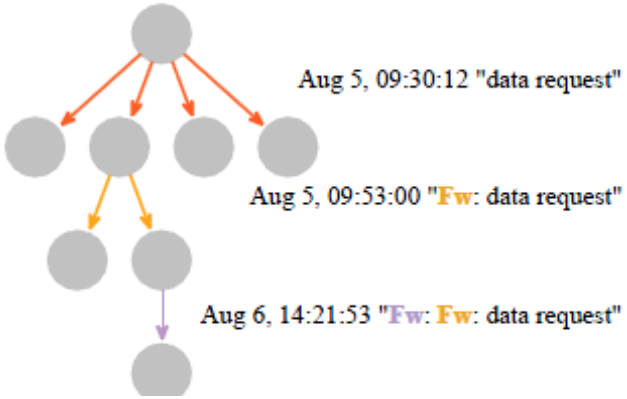


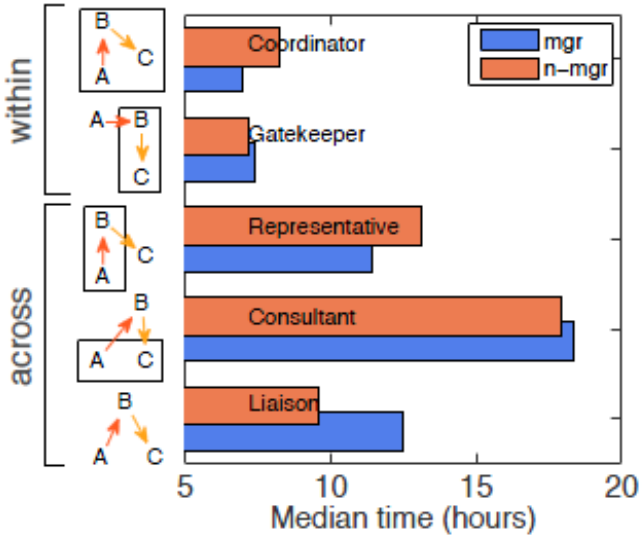
Growing one's Social Networks



Sentiments in Conversation





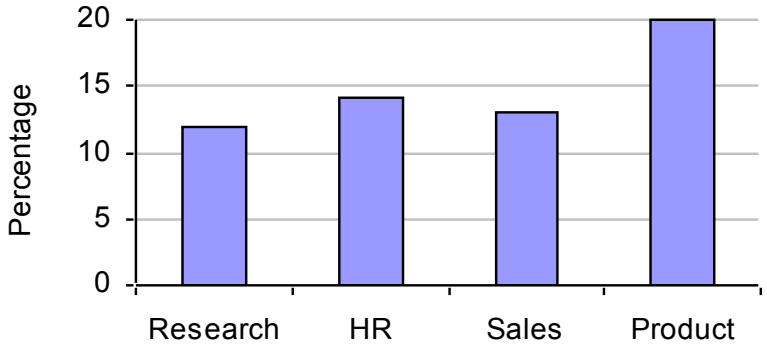


Role difference of normal behavior

Information Reuse Behavior (CHI '11)

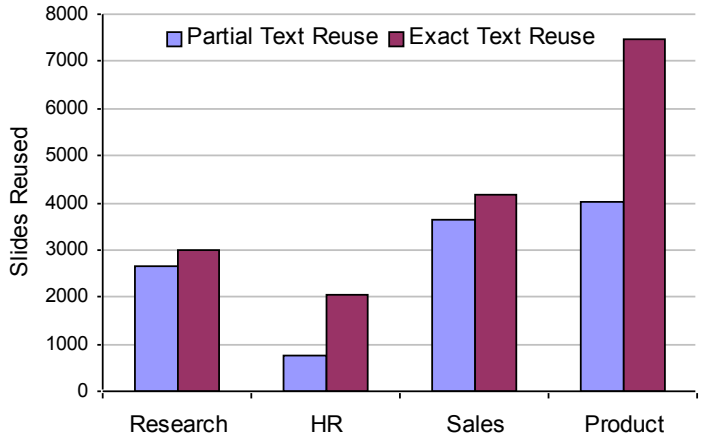
Columbia University

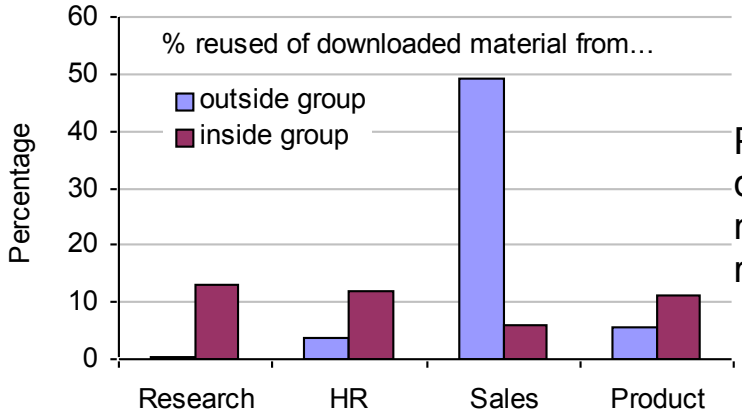
Percentage of slides with reused content



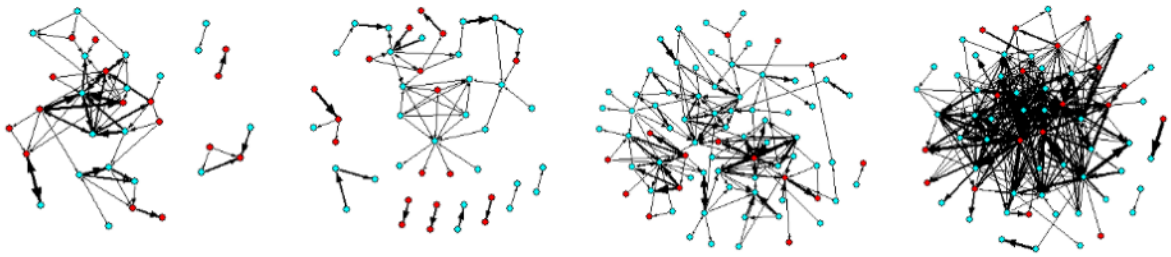
Percentage of reused slides that were reused by the same author vs. by a different author

Number of slide pairs with exact vs. partial text reuse





Percentage of downloaded material being reused

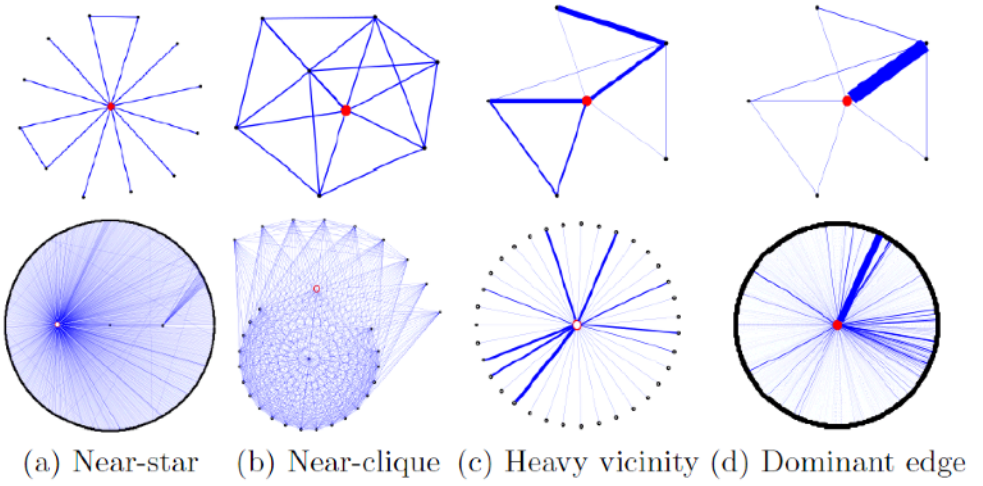


Anomaly Detection – algorithms and infrastructure

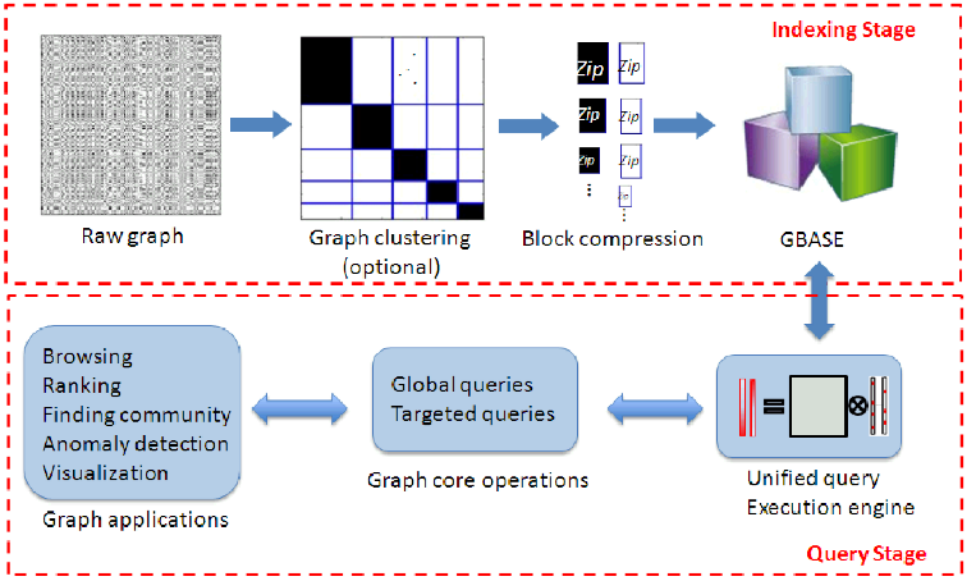
- **Thrust 1:** Anomaly Detection Algorithms
  -- New algorithms to detect abnormal humans (nodes) as well as abnormal contacts (edges) from social networks.
  -- Explore the structure feature and incorporate content (semantic) features.

- **Thrust 2:** Anomaly Usability
  -- Address the 'lack-of-the ground-truth' issue by
  (1) Interpretation friendly properties (e.g., non-negativity, sparseness, etc) into the current anomaly detection matrix factorization; and
  (2) providing some concise summarization to perform anomaly attribution.

- **Thrust 3:** Infrastructure Support
  -- General and scalable graph/network management system to process large

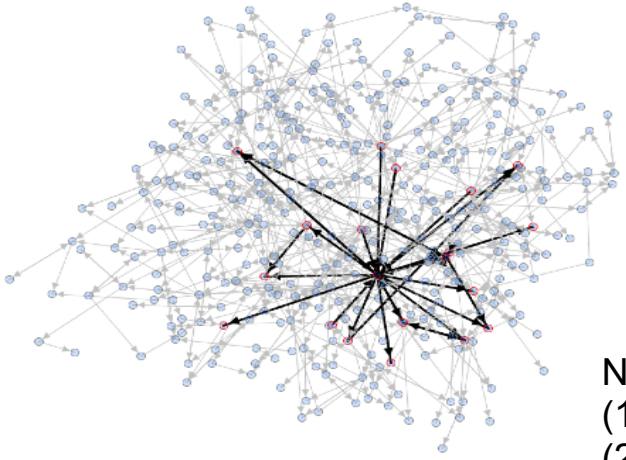


Typical abnormal nodes and their local ego-net structures

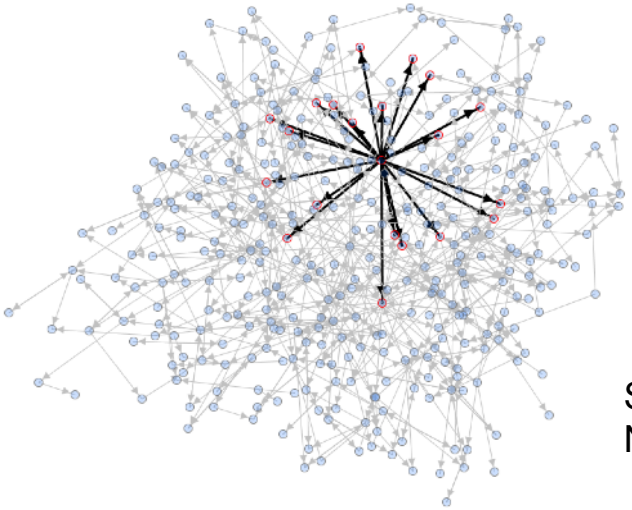


The overall flowchart of the graph management system

Use Case: Utilizing Social Network Analysis for Spam Detection

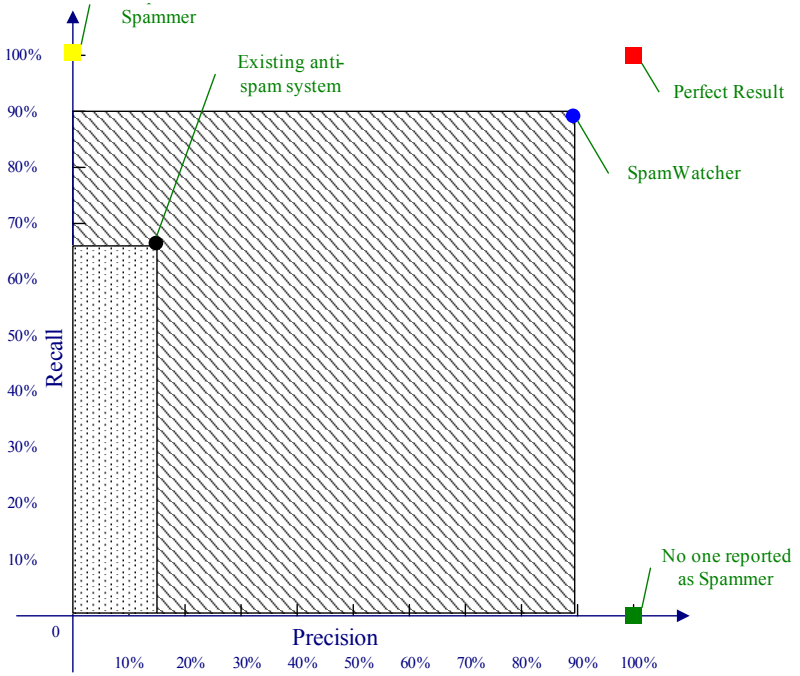


Normal:
(1) Clique-like
(2) Two-way links

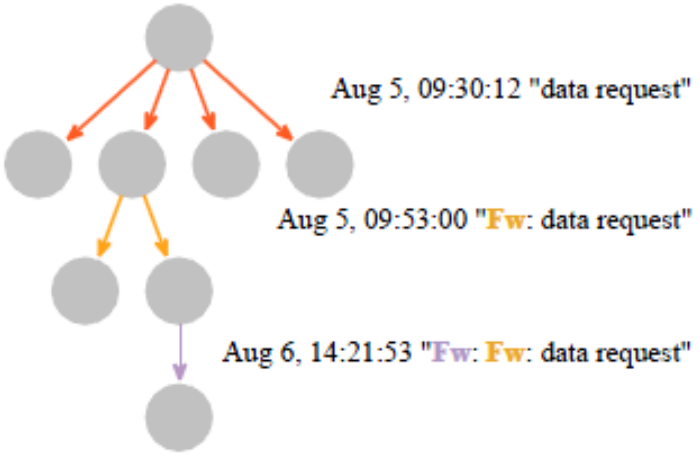


Spamming:
Near-Star

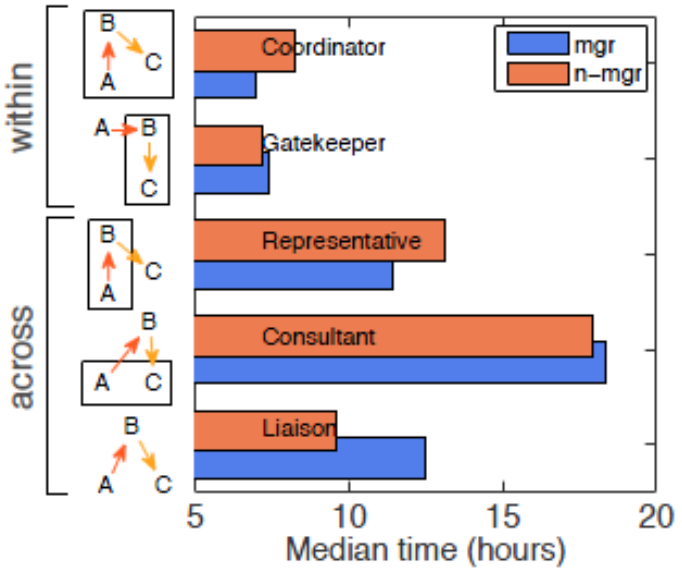
- An analysis in a telecomm area of 6 million users in 2009.
- In experiment
  - Social Network Analysis is with recall of 89.97% and precision of 88.17% while comparison system is with 66.77% recall and 14.85% precision.
  - SNA's precision/recall area is 8 times larger

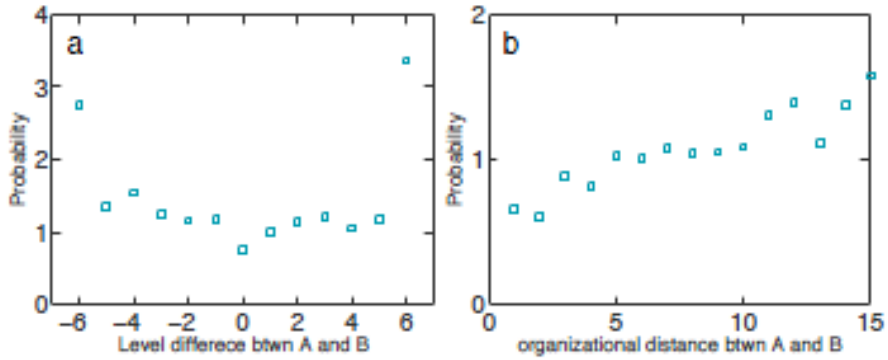


Anomaly Detection – information flow-based approach

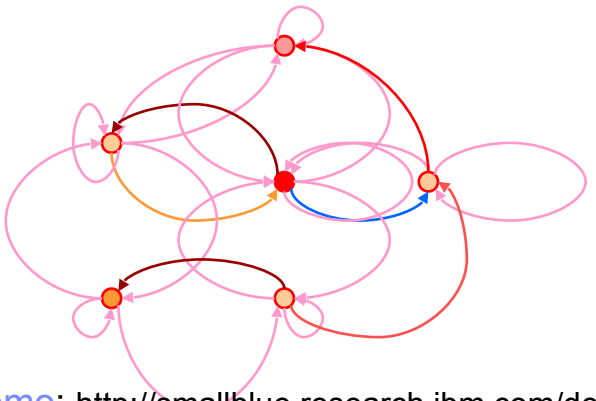


An illustrative example of an informa spreading tree. This tree is of size 8, width 4, depth 3.



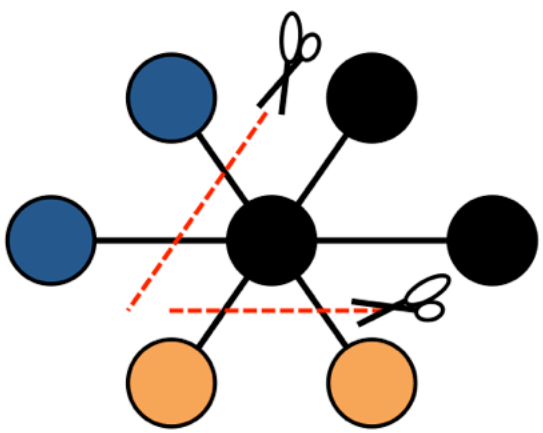


Probability ratio of email forwarding as a function of (a) hierarchical level difference and (b) organizational distance between initiators and ...readers. The information spreading exhibits some ...-homophily effect.

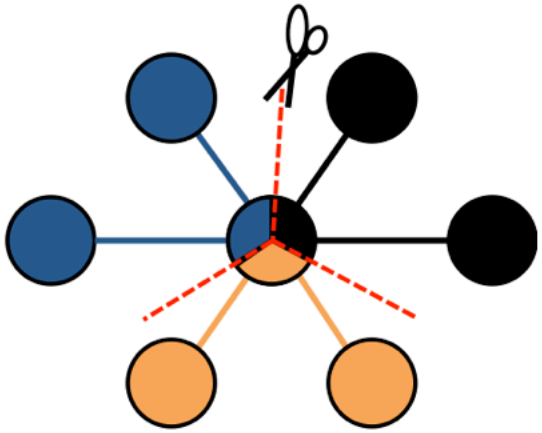


Video demo: http://smallblue.research.ibm.com/demos/

Graph Partitioning

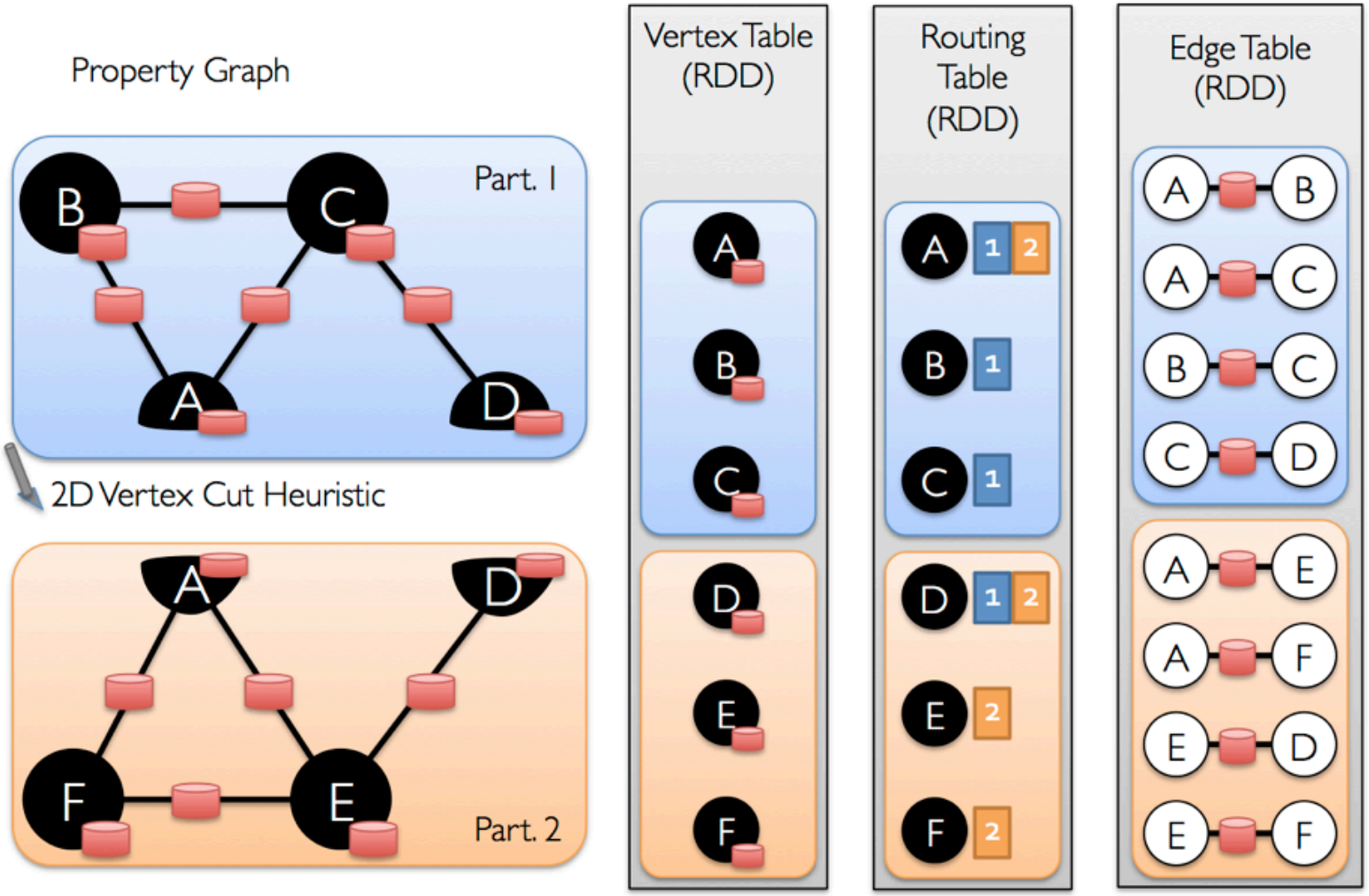


Edge Cut



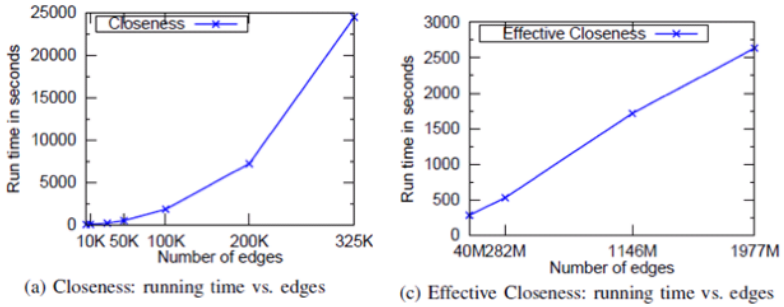
Vertex Cut

Distributed Graph Computation in GraphX



Network Analysis -- Effectiveness & Efficiency (GBase)

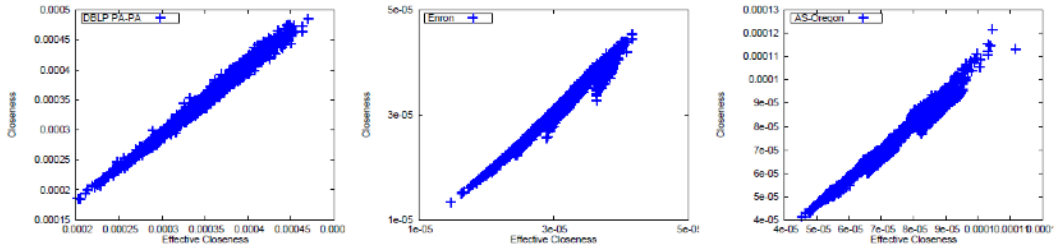
- Example -- we proposed two new centralities (`effective closeness' and `LineRank'), and efficient large scale algorithms for billion-scale graphs.



(a) Closeness: running time vs. edges (c) Effective Closeness: running time vs. edges

Scalability Results

(Near-linear scalability)

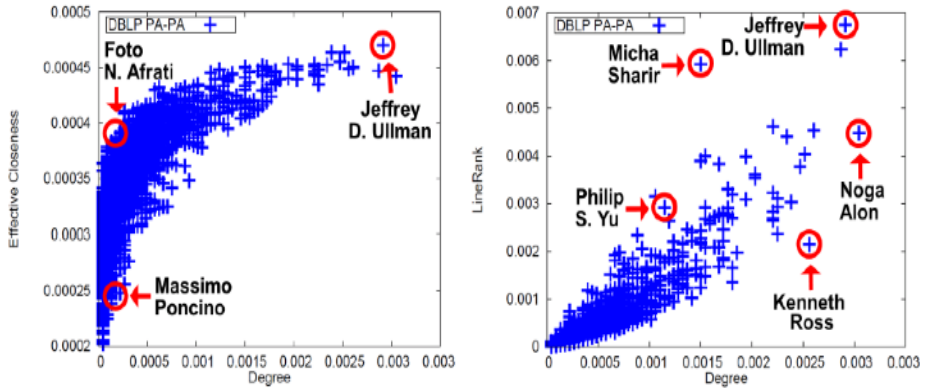


(a) DBLP Authors (b) Enron (c) AS-Oregon

Effective Closeness vs. Closeness

(Near-linear correlation (≥97.8%)

For 2 Billon Edges,
- standard closeness: 30,000 years
- effective closeness: ~ 1 day !
1,000,000 times faster!



(a) DBLP Authors: Effective Closeness vs. Degree (b) DBLP Authors: LineRank vs. Degree

Analysis of Real-World Graph

Kang, Tong, Sun, Lin, and Faloutsos, "GBase: A Scalable and general graph management system", KDD 2011

Homework #2 (Due 10/18/2018, 5pm)

Please see more detailed information at Canvas.