Sensors, networks, and massive data

Michael W. Mahoney

Stanford University

May 2012

(For more info, see:
http:// cs.stanford.edu/people/mmahoney/
or Google on "Michael Mahoney")

Lots of types of “sensors”

Examples:

- Physical/environmental: temperature, air quality, oil, etc.
- Consumer: RFID chips, SmartPhone, Store Video, etc.
- Health care: Patient Records, Images & Surgery Videos, etc.
- Financial: Transactions for regulations, HFT, etc.
- Internet/e-commerce: clicks, email, etc. for user modeling, etc.
- Astronomical/HEP: images, experiments, etc.

Common theme: easy to generate A LOT of data

Questions:

- What are similarities/differences i.t.o. funding drivers, customer demands, questions of interest, time sensitivity, etc. about “sensing” in these different applications?
- What can we learn from one area and apply to another area?

BIG data??? MASSIVE data????

NYT, Feb 11, 2012: "The Age of Big Data"

- "What is Big Data? A meme and a marketing term, for sure, but also shorthand for advancing trends in technology that open the door to a new approach to understanding the world and making decisions. …"

Why are big data big?

- Generate data at different places/times and different resolutions
- Factor of 10 more data is not just more data, but *different data*

BIG data??? MASSIVE data????

MASSIVE data:

- Internet, Customer Transactions, Astronomy/HEP = "Petascale"
- One Petabyte = watching 20 years of movies (HD) = listening to 20,000 years of MP3 (128 kbits/sec) = way too much to browse or comprehend

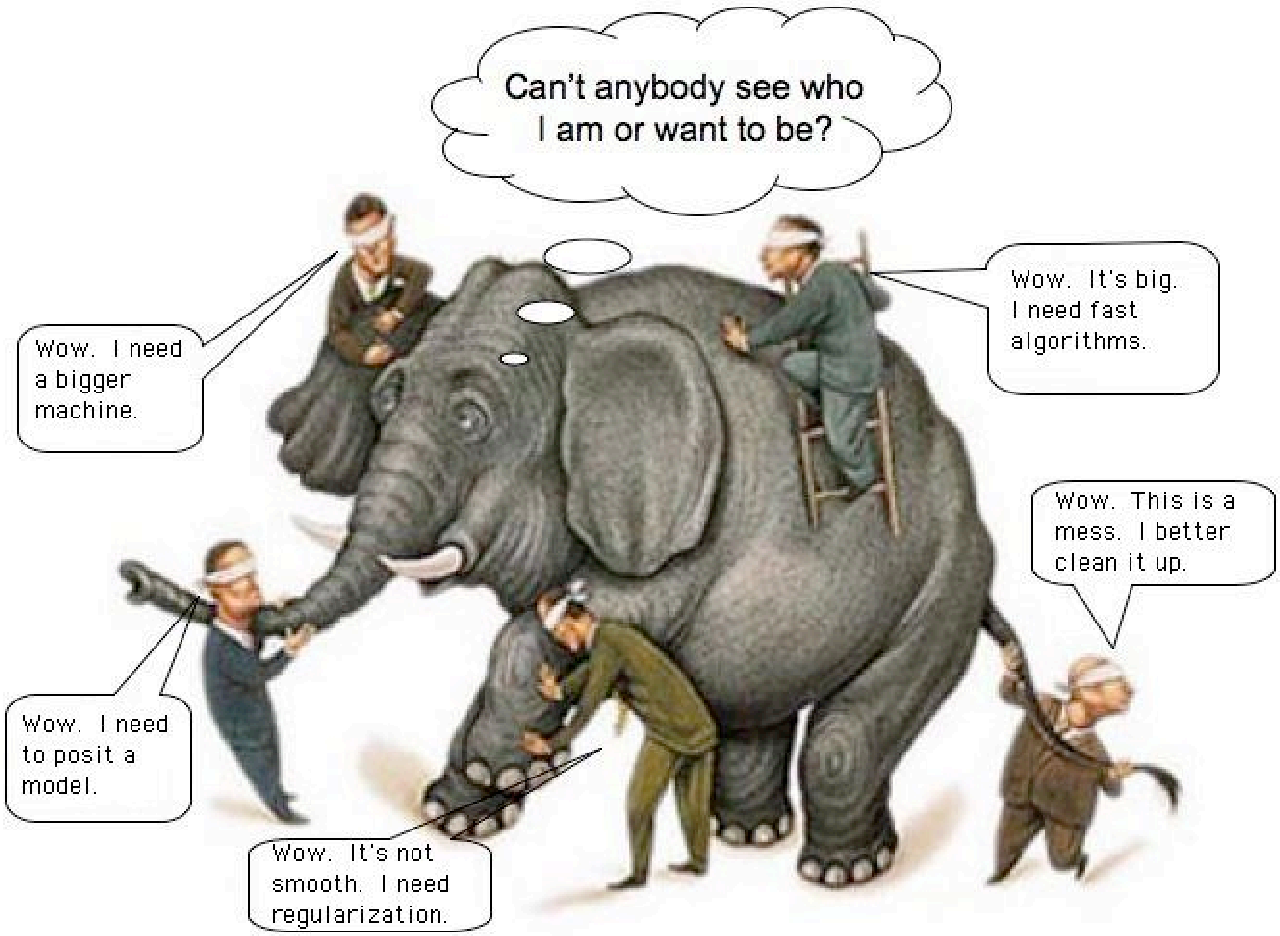
massive data:

- 10^5 people typed at 10^6 DNA SNPs; 10^6 or 10^9 node social network; etc.

In either case, main issues:

- Memory management issues, e.g., push computation to the data
- Hard to answer even basic questions about what data "looks like"

How do we view BIG data?



Algorithmic vs. Statistical Perspectives

Lambert (2000), Mahoney (2010)

Computer Scientists

• *Data*: are a record of everything that happened.

• *Goal*: process the data to find interesting patterns and associations.

• *Methodology*: Develop approximation algorithms under different models of data access since the goal is typically computationally hard.

Statisticians (and Natural Scientists)

• *Data*: are a particular random instantiation of an underlying process describing unobserved patterns in the world.

• *Goal*: is to extract information about the world from noisy data.

• *Methodology*: Make inferences (perhaps about unseen events) by positing a model that describes the random variability of the data around the deterministic model.

Thinking about large-scale data

Data generation is modern version of microscope/telescope:

- See things couldn't see before: e.g., movement of people, clicks and interests; tracking of packages; fine-scale measurements of temperature, chemicals, etc.
- Those inventions ushered new scientific eras and new understanding of the world and new technologies to do stuff

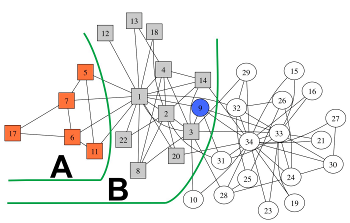
Easy things become hard and hard things become easy:

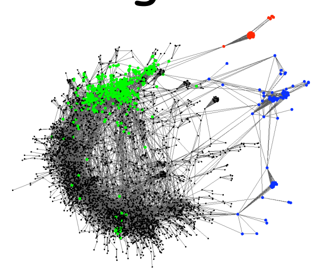
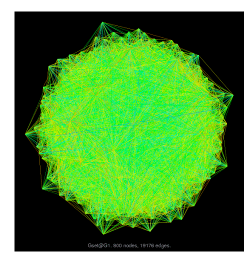
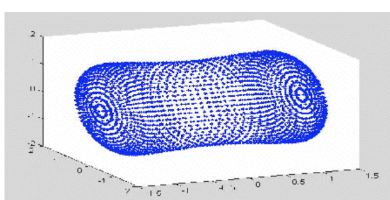
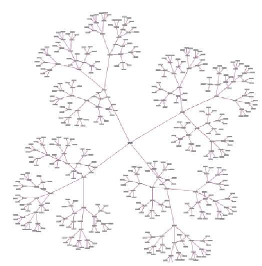
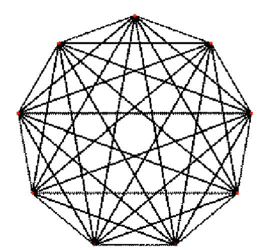
- Easier to see the other side of universe than bottom of ocean
- Means, sums, medians, correlations is easy with small data

Our ability to generate data far exceeds our ability to extract insight from data.

Many challenges ...

- Tradeoffs between prediction & understanding
- Tradeoffs between computation & communication,
- Balancing heat dissipation & energy requirements
- Scalable, interactive, & inferential analytics
- Temporal constraints in real-time applications
- Understanding "structure" and "noise" at large-scale (*)
- Even meaningfully answering "What does the data look like?"

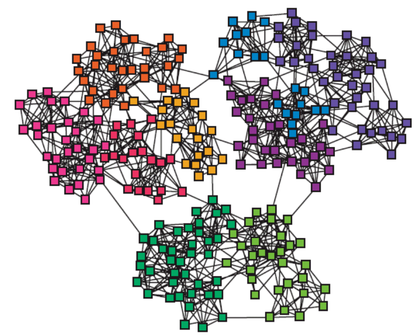


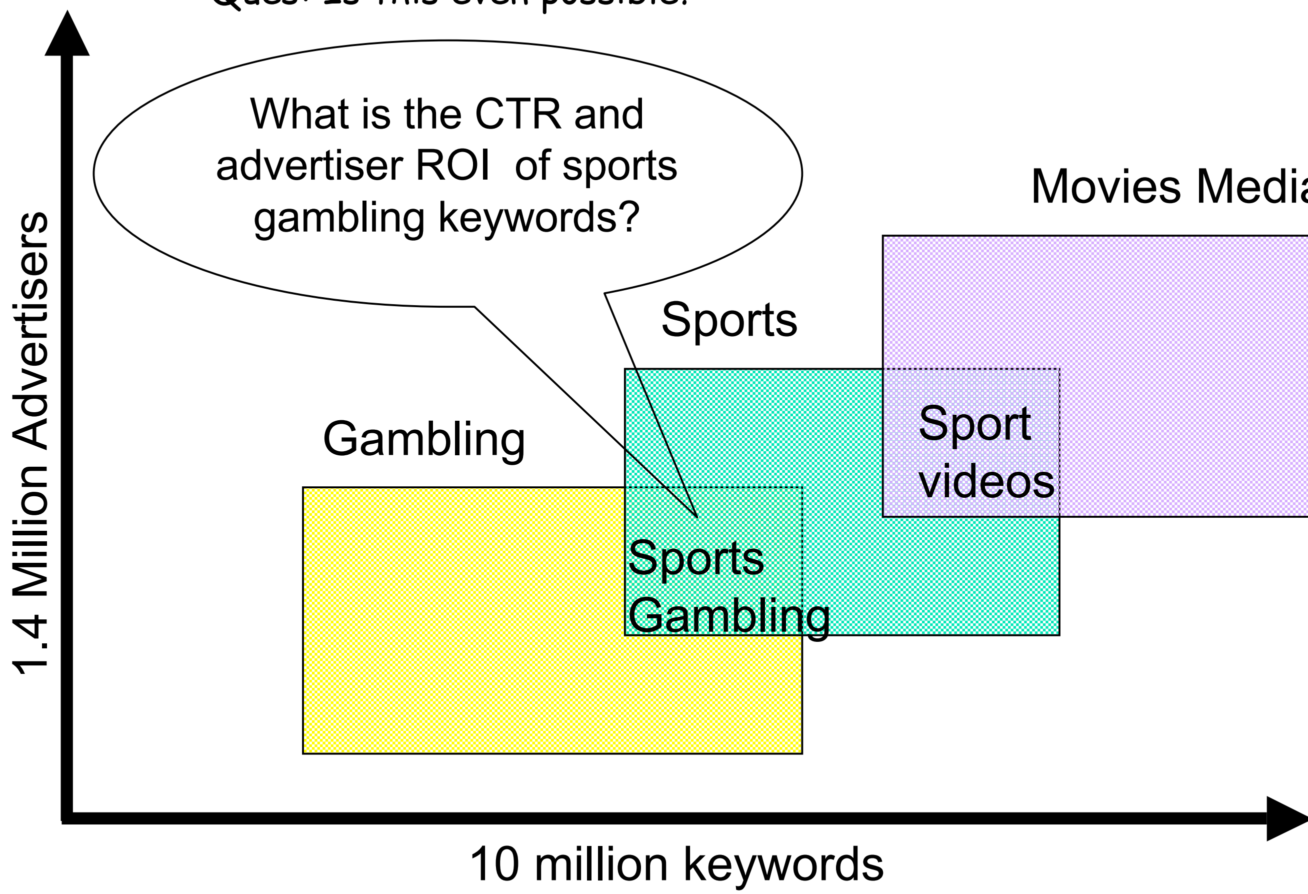






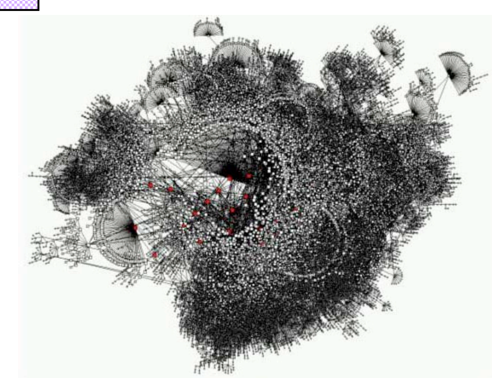
Micro-markets in sponsored search

Goal: Find *isolated* markets/clusters (in an advertiser-bidded phrase bipartite graph) with *sufficient money/clicks* with *sufficient coherence.*

Ques: Is this even possible?







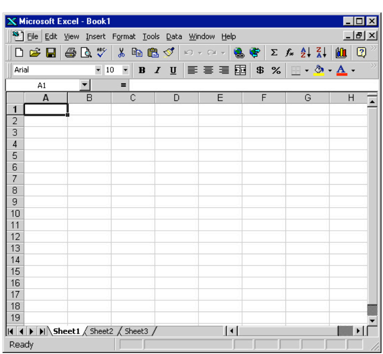
What about *sensors?*

Vector space model - *analogous* to "bag-of-words" model for documents/terms.

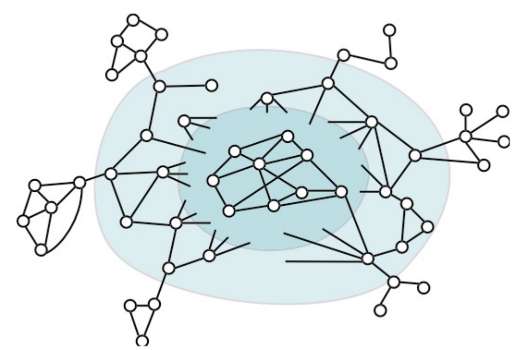
- Each sensor is a "document," a *vector in a high-dimensional Euclidean space*
- Each measurement is a "term", describing the *elements of that vector*
- (Advertisers and bidded-phrases--and many other things--are also analogous.)

Can also define sensor-measurement graphs :

- Sensors are nodes, and edges are between sensors with similar measurements



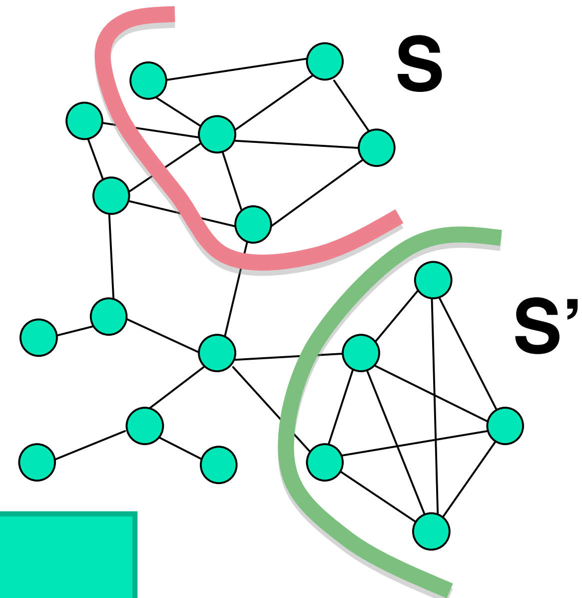
$$= \begin{array}{cc} & n \text{ terms (measurements)} \\ \begin{array}{c} m \\ \text{documents} \\ \text{(sensors)} \end{array} & \left(\begin{array}{c} \mathcal{A} \\ A_{ij} = \text{frequency of j-th term in i-th} \\ \text{document (value of j-th measurement} \\ \text{at i-th sensor)} \end{array} \right) \end{array} =$$



Cluster-quality Score: Conductance

- How cluster-like is a set of nodes?

  Idea: balance "boundary" of cluster with "volume" of cluster

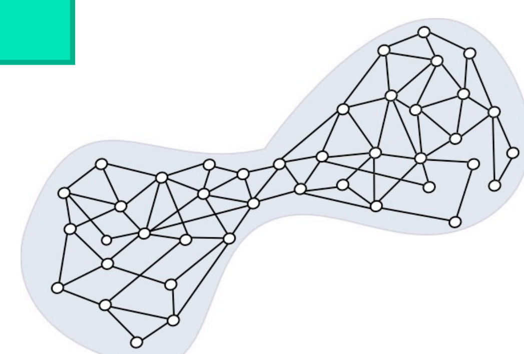


- Need a natural intuitive measure:

Conductance (normalized cut)

$$\phi(S) \approx \text{\# edges cut / \# edges inside}$$

- Small $\phi(S)$ corresponds to better clusters of nodes



Graph partitioning

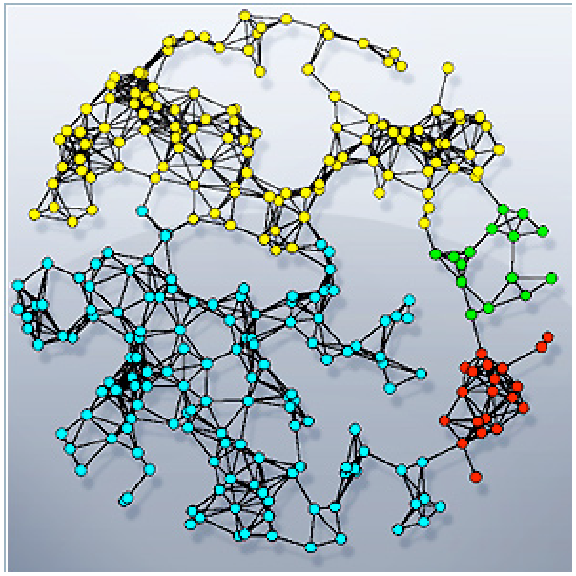
A family of combinatorial optimization problems - want to partition a graph's nodes into two sets s.t.:

- Not much edge weight across the cut (cut quality)
- Both sides contain a lot of nodes

Standard formalizations of the bi-criterion are NP-hard!

Approximation algorithms:

- Spectral methods* - (compute eigenvectors)
- Local improvement - (important in practice)
- Multi-resolution - (important in practice)
- Flow-based methods* - (mincut-maxflow)



* comes with strong underlying theory to guide heuristics

Comparison of "spectral" versus "flow"

Spectral:

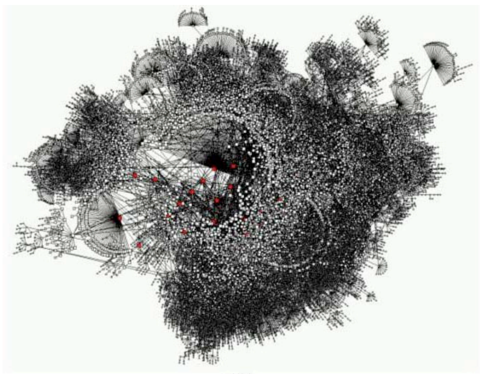
- Compute an eigenvector
- "Quadratic" worst-case bounds
- Worst-case achieved -- on "long stringy" graphs
- Embeds you on a line (or Kn)

Flow:

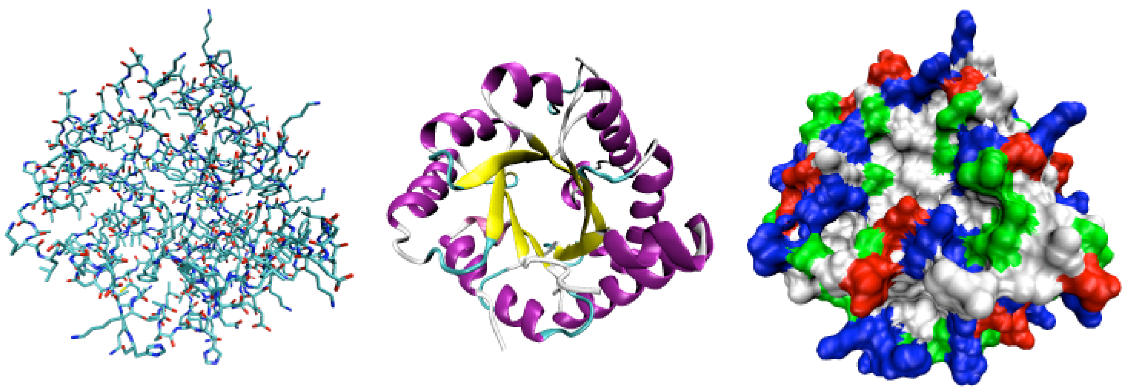
- Compute a LP
- O(log n) worst-case bounds
- Worst-case achieved -- on expanders
- Embeds you in L1

Two methods:

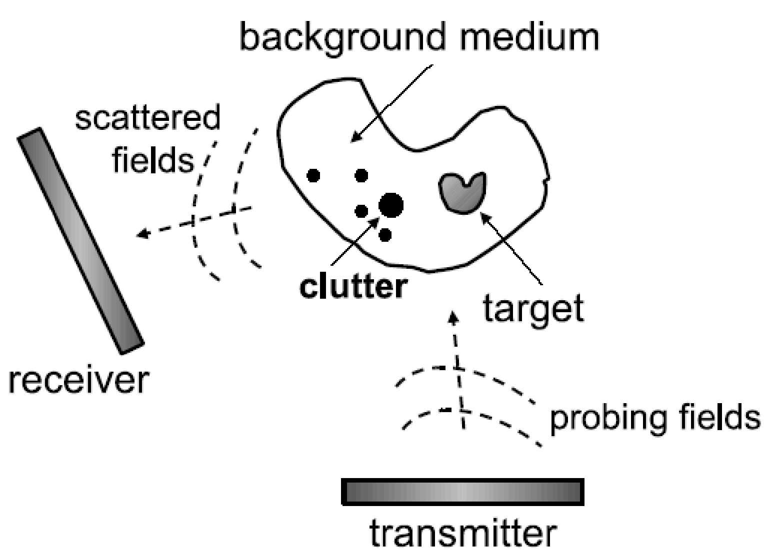
- Complementary strengths and weaknesses
- What we compute will depend on approximation algorithm as well as objective function.



Analogy: What does a protein look like?



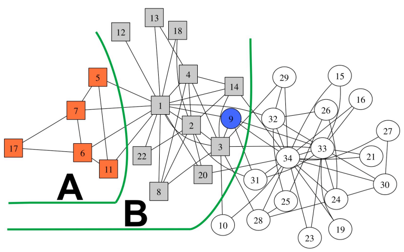
Three possible representations (all-atom; backbone; and solvent-accessible surface) of the three-dimensional structure of the protein triose phosphate isomerase.

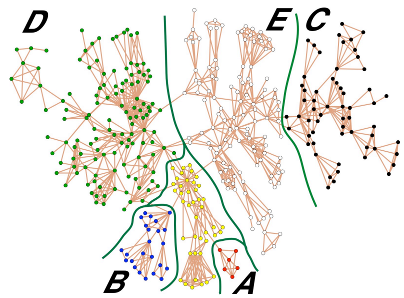


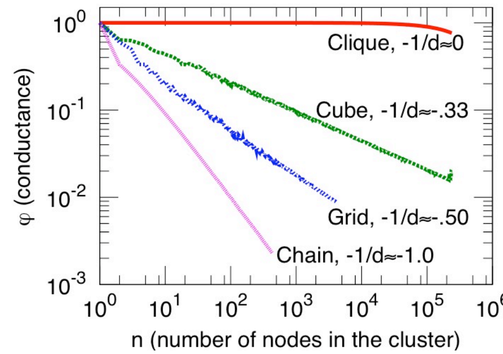
Experimental Procedure:

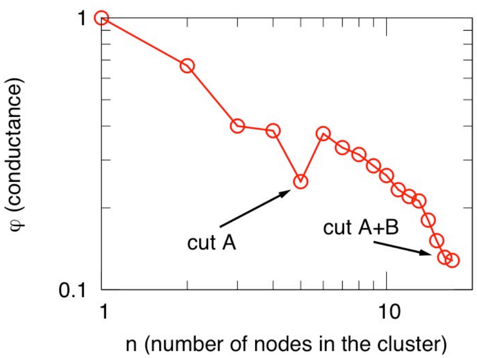
- Generate a bunch of output data by using the unseen object to filter a known input signal.
- Reconstruct the unseen object given the output signal and what we know about the artifactual properties of the input signal.

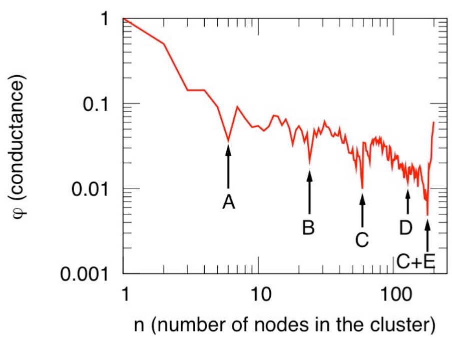
Popular small networks

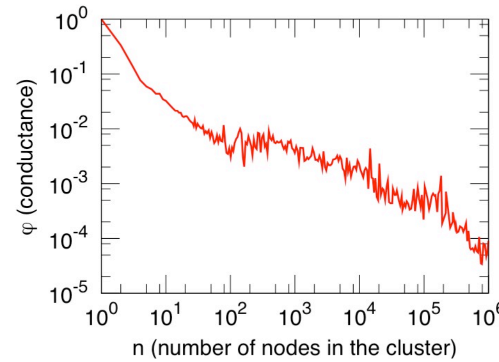












Zachary's karate club

Newman's Network Science

Meshes and RoadNet-CA

Large Social and Information Networks

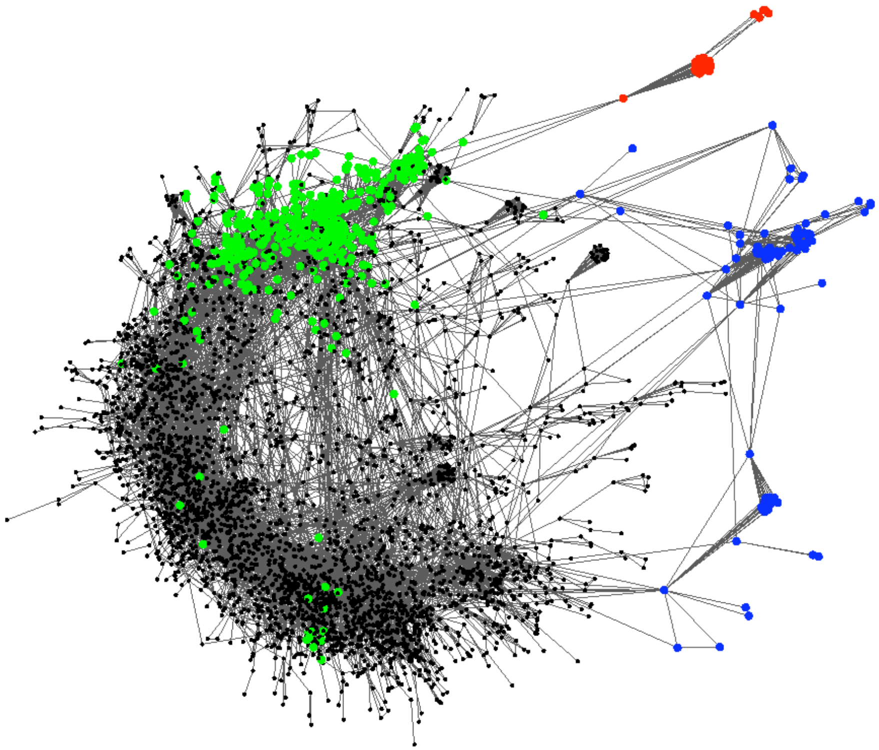
• Social nets	Nodes	Edges	Description
LiveJournal	4,843,953	42,845,684	Blog friendships [4]
Epinions	75,877	405,739	Who-trusts-whom [35]
Flickr	404,733	2,110,078	Photo sharing [21]
Delicious	147,567	301,921	Collaborative tagging
CA-DBLP	317,080	1,049,866	Co-authorship (CA) [4]
CA-cond-mat	21,363	91,286	CA cond-mat [25]
• Information networks			
Cit-hep-th	27,400	352,021	hep-th citations [13]
Blog-Posts	437,305	565,072	Blog post links [28]
• Web graphs			
Web-google	855,802	4,291,352	Web graph Google
Web-wt10g	1,458,316	6,225,033	TREC WT10G web
• Bipartite affiliation (authors-to-papers) networks			
Atp-DBLP	615,678	944,456	DBLP [25]
AtP-astro-ph	54,498	131,123	Arxiv `astro-ph` [25]
• Internet networks			
AS	6,474	12,572	Autonomous systems
Gnutella	62,561	147,878	P2P network [36]

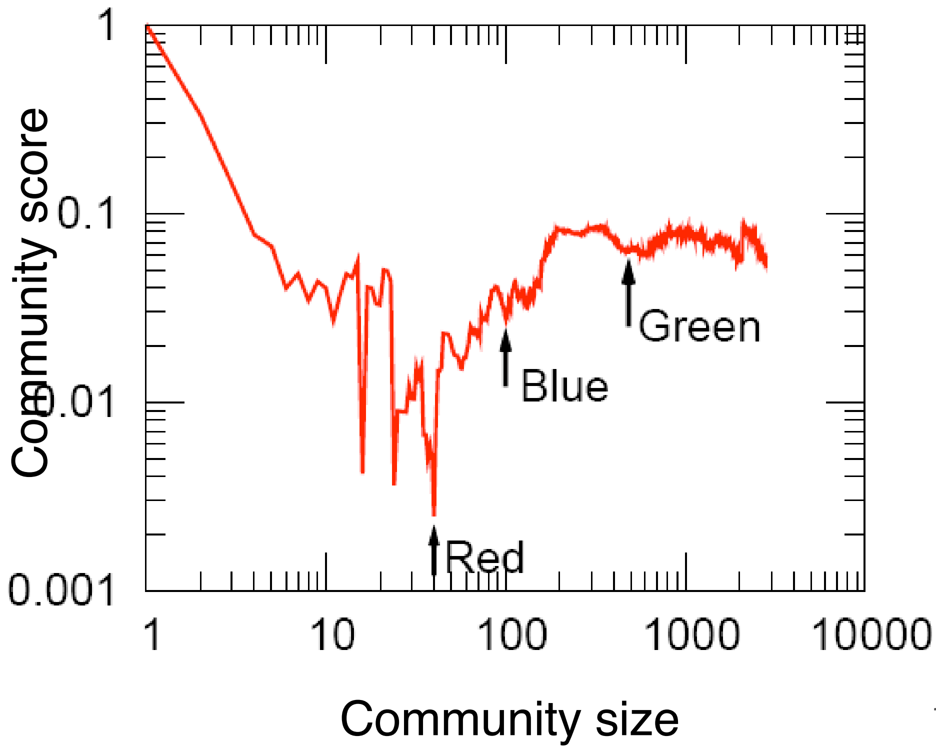
Table 1: Some of the network datasets we studied.

Typical example of our findings

Leskovec, Lang, Dasgupta, and Mahoney (WWW 2008, 2010 & IM 2009)

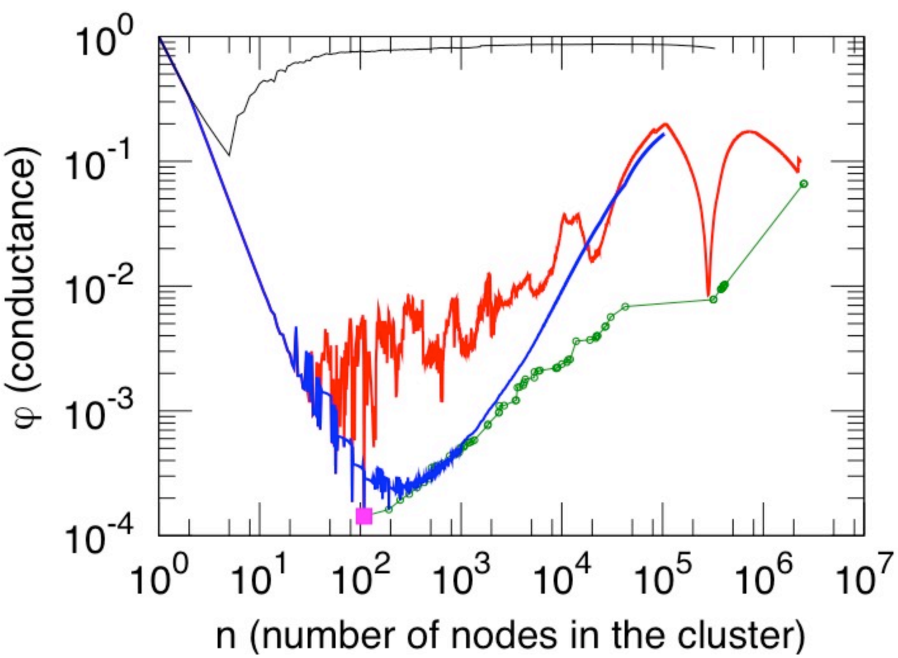
General relativity collaboration network
(pretty small: 4,158 nodes, 13,422 edges)



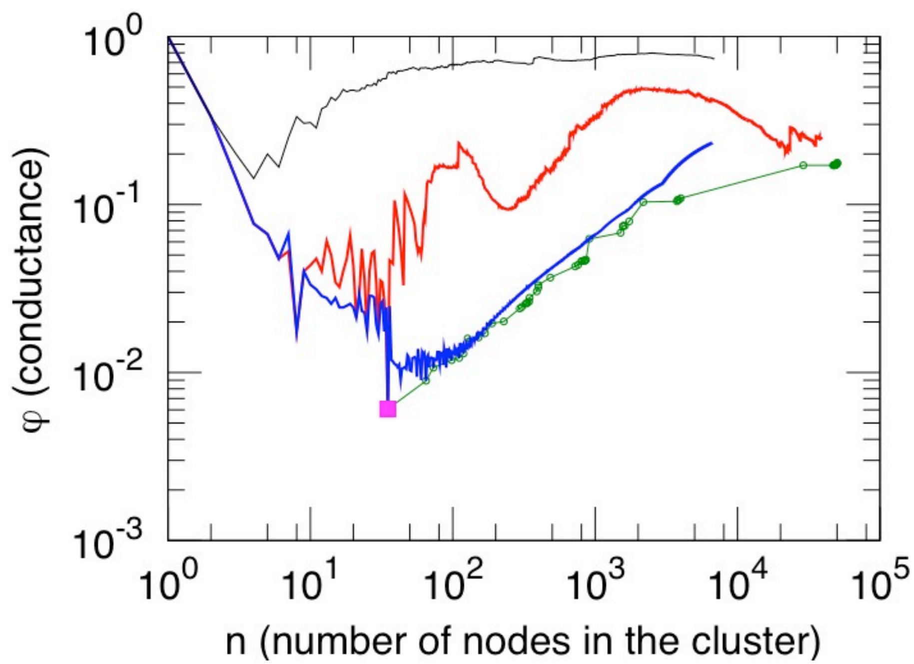


Large Social and Information Networks

Leskovec, Lang, Dasgupta, and Mahoney (WWW 2008, 2010 & IM 2009)



LiveJournal



Epinions

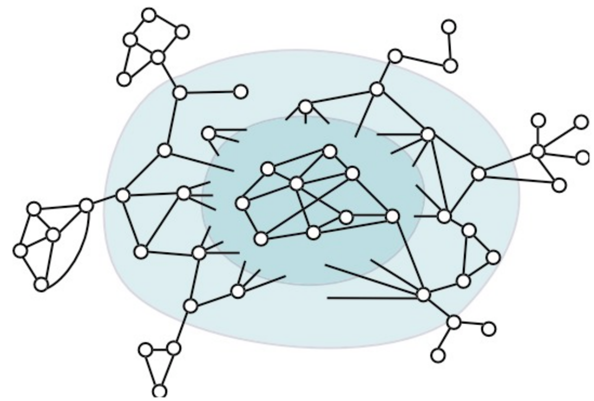
Focus on the red curves (local spectral algorithm) - blue (Metis+Flow), green (Bag of whiskers), and black (randomly rewired network) for consistency and cross-validation.

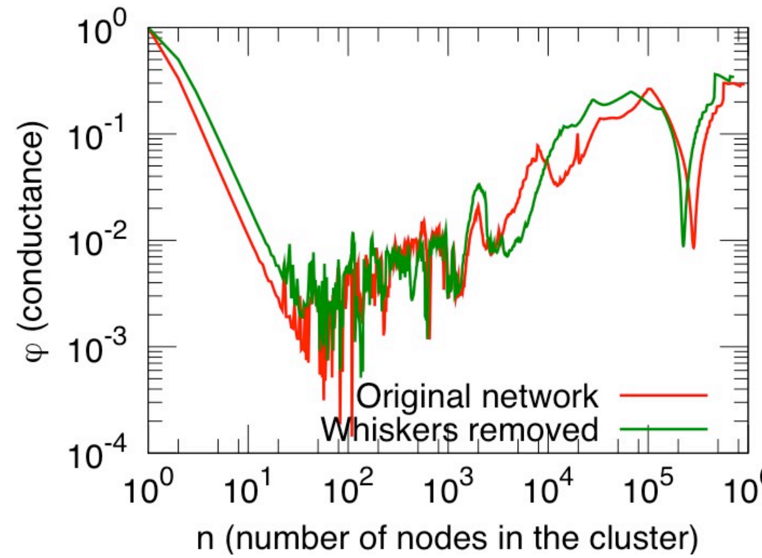
Interpretation: "Whiskers" and the "core" of large informatics graphs

Leskovec, Lang, Dasgupta, and Mahoney (WWW 2008, 2010 & IM 2009)

- "Whiskers"
  - maximal sub-graph detached from network by removing a single edge
  - contains 40% of nodes and 20% of edges
- "Core"
  - the rest of the graph, i.e., the 2-edge-connected core

- Global minimum of NCPP is a whisker
- BUT, *core itself has nested whisker-core structure*





Local "structure" and global "noise"

Many (most/all?) large informatics graphs (& massive data in general?)

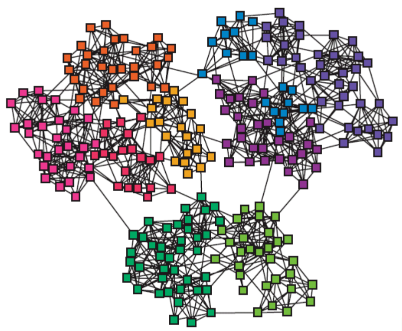
• have local structure that is meaningfully geometric/low-dimensional

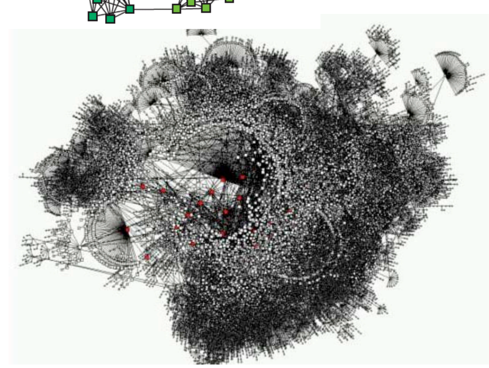
• does *not* have analogous meaningful global structure

Intuitive example:

• What does the graph of you and your 10^2 closest Facebook friends "look like"?

• What does the graph of you and your 10^5 closest Facebook friends "look like"?





Many lessons …

This is *problematic for MANY things* people want to do:

- statistical analysis that relies on asymptotic limits
- recursive clustering algorithms
- analysts who want a few meaningful clusters

More data need *not* be better if you:

- don't have control over the noise
- want "islands of insight" in the "sea of data"

How does this manifest itself in your "sensor" application?

- Needles in haystack; correlations; time series -- "scientific" apps
- *Historically*, CS & database apps did more summaries & aggregates

Big changes in the past ... and future

Consider the creation of:

- Modern Physics
- Computer Science
- Molecular Biology
- OR and Management Science
- Transistors and Microelectronics
- Biotechnology

These were driven by *new measurement techniques* and *technological advances*, but they led to:

- big new (academic and applied) questions
- new perspectives on the world
- lots of downstream applications

We are in the middle of a similarly big shift!

Conclusions

HUGE range of "sensors" are generating A LOT of data:

- will lead to a very different world in many ways

Large-scale data are *very* different than small-scale data.

- Easy things become hard, and hard things become easy
- Types of questions that are meaningful to ask are different
- Structure, noise, etc. properties are often deeply counterintuitive

Different applications are driven by different considerations

- next-user-interaction, qualitative insight, failure modes, false positives versus false negatives, time sensitivity, etc.

Algorithms can compute answers to known questions

- but algorithms can also be used as "experimental probes" of the data to form questions!

MMDS Workshop on
"Algorithms for Modern Massive Data Sets"
(http://mmds.stanford.edu)

at Stanford University, July 10-13, 2012

Objectives:

- Address algorithmic, statistical, and mathematical challenges in modern statistical data analysis.

- Explore novel techniques for modeling and analyzing massive, high-dimensional, and nonlinearly-structured data.

- Bring together computer scientists, statisticians, mathematicians, and data analysis practitioners to promote cross-fertilization of ideas.

Organizers: M. W. Mahoney, A. Shkolnik, G. Carlsson, and P. Drineas,

Registration is available now!