Applications in Social Networks

Jason Riedy, David Bader, David Ediger, ...

Georgia Tech College of Computing
Computational Science and Engineering

Outline

- Background on applications in social network analysis
  - Static applications
  - Growing need for dynamic analysis
- Quick, high-level description of two algorithm areas with potential for acceleration.
  - *k*-Betweenness Centrality
  - Agglomerative clustering / community identification

Graph –theoretic problems in social networks

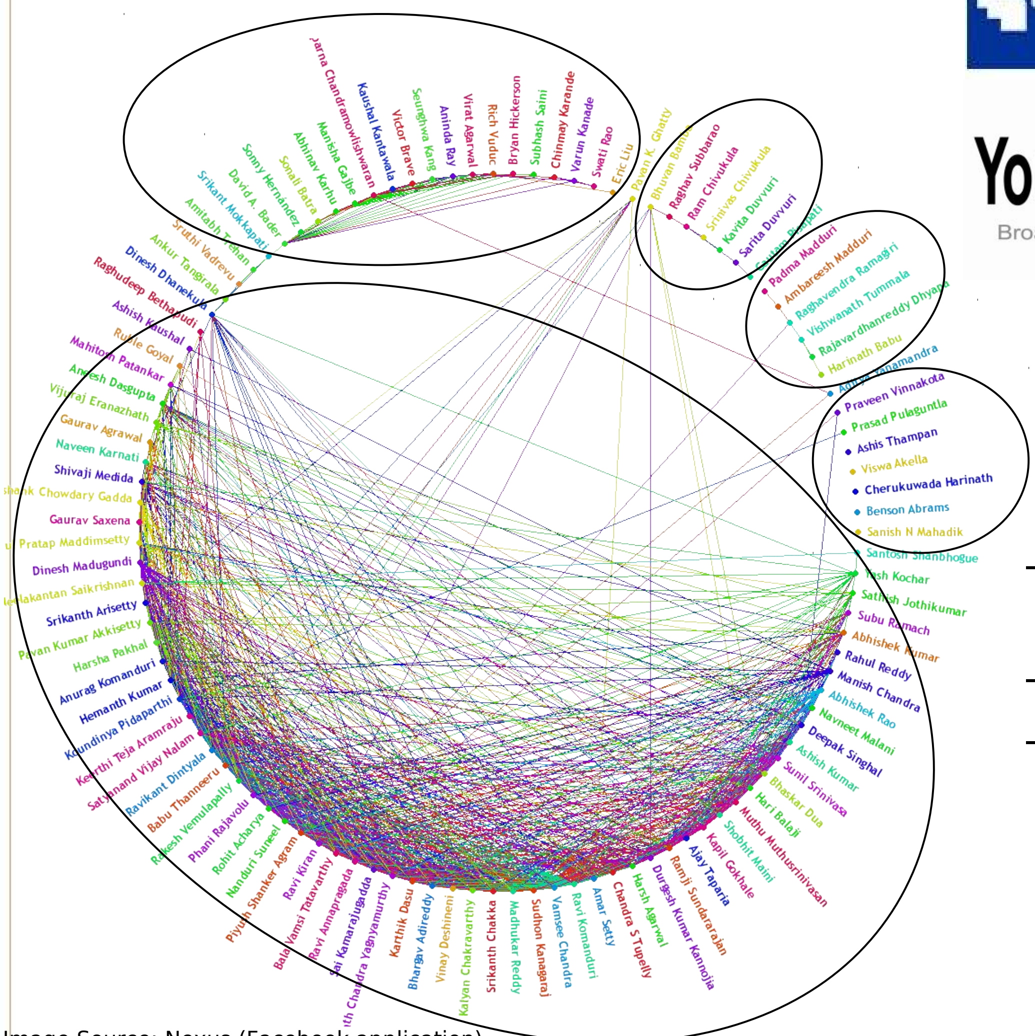


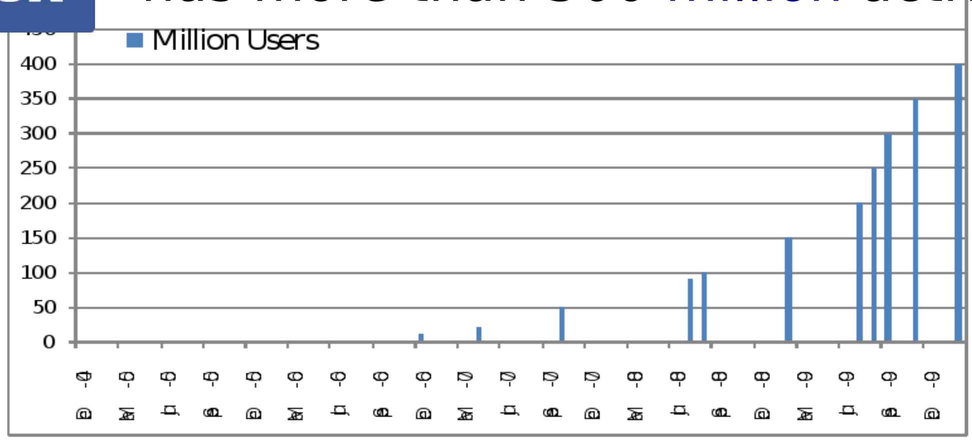
Image Source: Nexus (Facebook application)

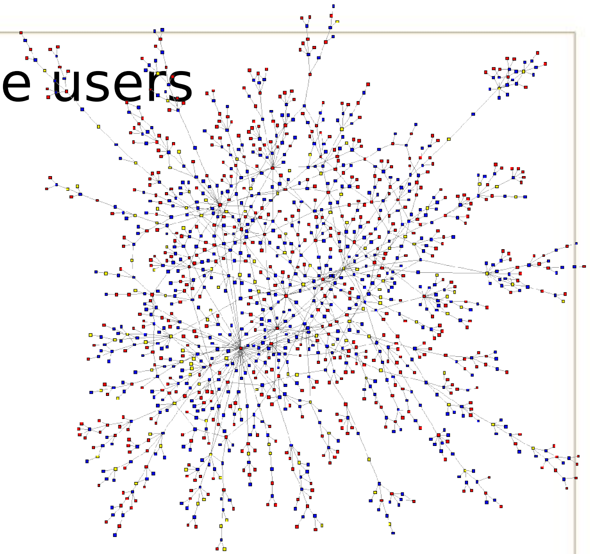
- Community identification: clustering
- Targeted advertising: centrality
- Information spreading: modeling

Driving Forces in Social Network Analysis

- facebook has more than 500 million active users

3 orders of magnitude growth in 3 years!



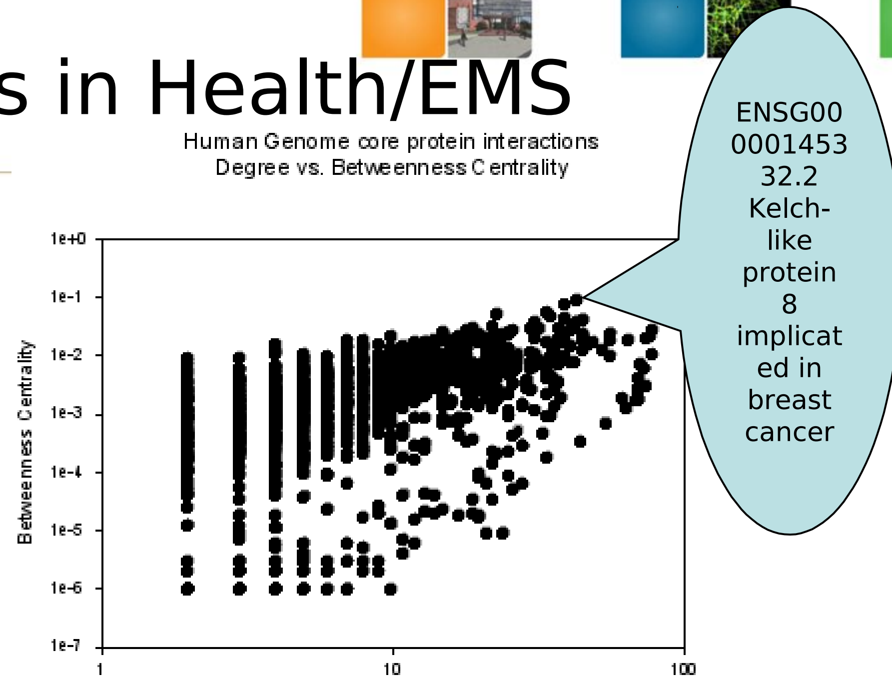


- Note the graph is **changing** as well as growing.
- Traditional graph partitioning often fails:
  - **Topology**: Interaction graph is low-diameter, and has no good separators
  - **Irregularity**: Communities are not uniform in size
  - **Overlap**: individuals are members of one or more communities
- Currently recompute ad targeting once per hour. Accelerate?
  - Now consider targeting usage on a power grid, *etc*.
  - Similar size, but static. Dynamic identification of issues definitely needs accelerated.

Massive Data Analytics in Health/EMS

Public Health

- CDC / Nation-scale surveillance of public health
- Cancer genomics and drug design
  - computed Betweenness Centrality of Human Proteome



Rank	H1N1	atlflood
1	**@CDCFlu**	**@ajc**
2	@addthis	@driveafastercar
3	@Official_PAX	@ATLCheap
4	**@FluGov**	**@TWCi**
5	**@nytimes**	@HelloNorthGA
6	@tweetmeme	**@11AliveNews**
7	@mercola	**@WSB_TV**
8	**@CNN**	@shaunking

(Collaboration w/PNNL)

- Identify locally important news and information sources.
  - Spread correct information.
  - Prevent misinformation.
- Similar uses: Identify regions being affected by disaster / disease.

Georgia Tech College of Computing

Social/Economic Policy

- NYSE “Flash crash” of 6 May 2010:
  - Dropped 700 pts in 20 minutes.
  - Simple “circuit breakers” were of no use.
    - "A number of the [regulatory pauses] in effect on May 6 were resolved in less than one second, [...]" – Congressional Testimony of Larry Leibowitz, CEO NYSE Euronext
  - Breakers based on levels, not structure.
- 1 Oct: Finally announce the reason:
  - **One single large trade** triggered a network of reactions.
- Regulators need accelerated analysis.

Current Example Data Rates

- Financial / regulatory:
  - NYSE processes 1.5TB daily, maintains 8PB
- Social:
  - Facebook adds >100k users, 55M “status” updates, 80M photos daily; report more than 500M active users with an average of 130 “friend” connections each.
  - Foursquare, a *new* service, reports 1.2M location check-ins per week
- Scientific:
  - MEDLINE adds from 1 to 140 publications a day

Shared features: All data is rich, irregularly connected to other data. All is a mix of “good” and “bad” data... And much real data may be missing or inconsistent.

Current Unserved Applications

- Separate the “good” from the “bad”
  - Spam. Frauds. *Irregularities*.
  - Pick news from world-wide events tailored to interests *as the events & interests change*.
- Identify and track changes
  - Disease outbreaks. Social trends. Utility & service changes during weather events.
- Discover new relationships
  - Similarities in scientific publications.
- Predict upcoming events
  - Present advertisements *before* a user searches.

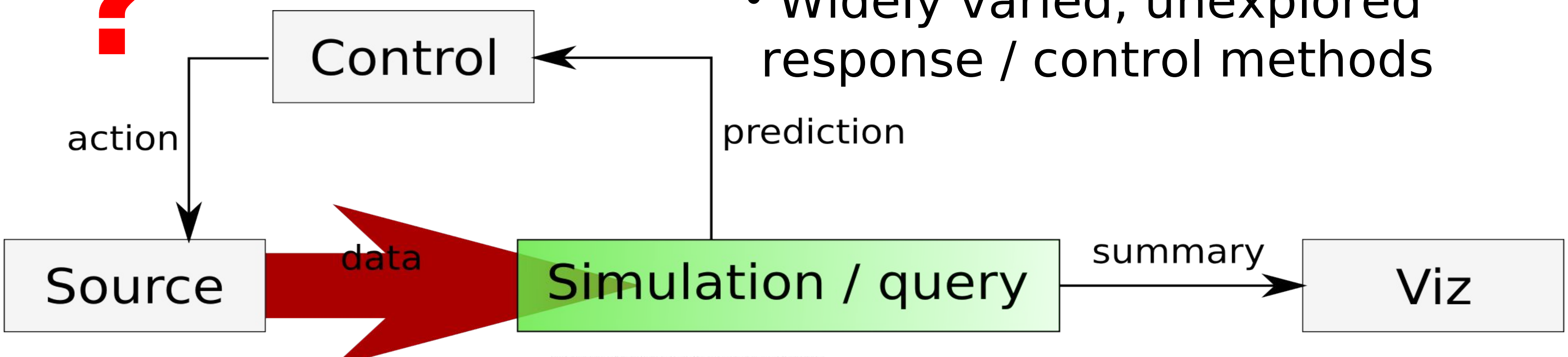
Shared features: Relationships are abstract. Physical locality is only one aspect, unlike physical simulation.

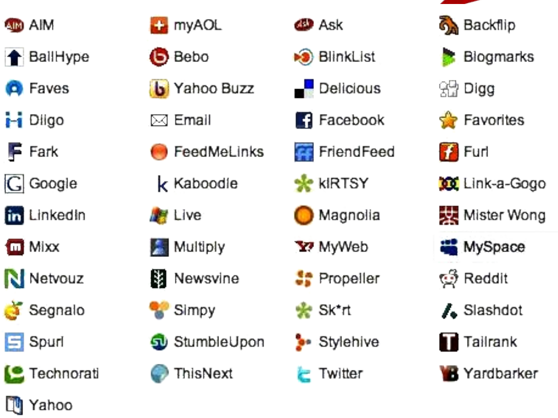
Streaming Data Analysis

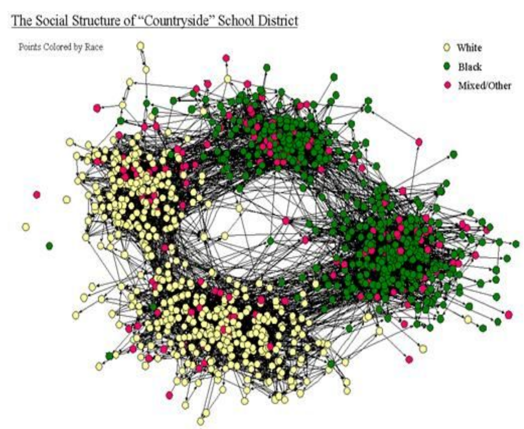
Current needs, future knowledge:

?

- Massive, ***irregularly structured*** input data.
- New simulation, analysis methods
- Widely varied, unexplored response / control methods





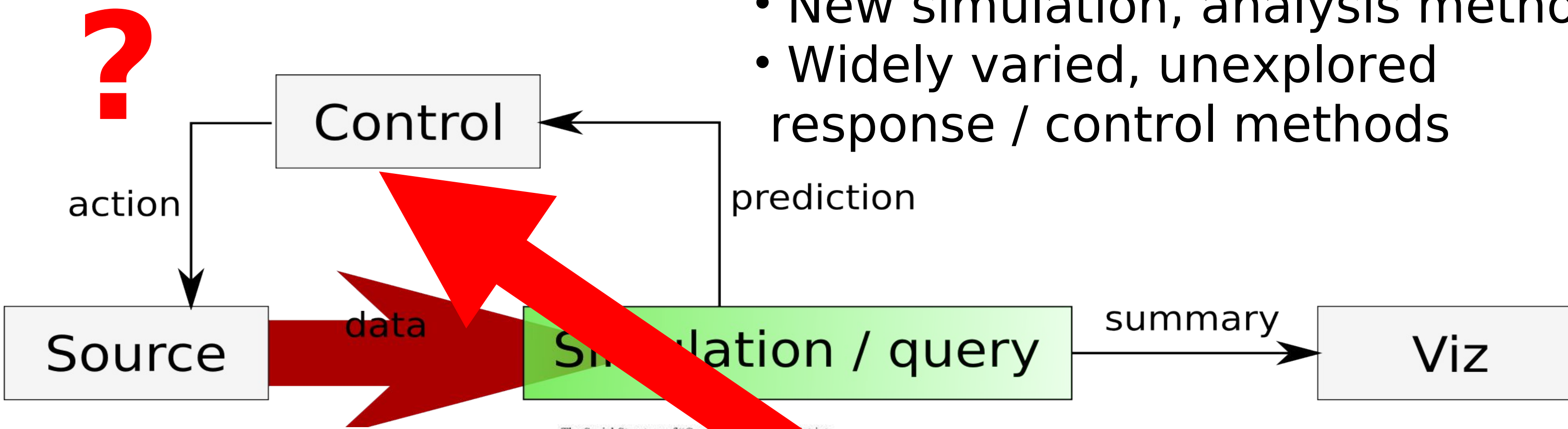


?

Streaming Data Analysis

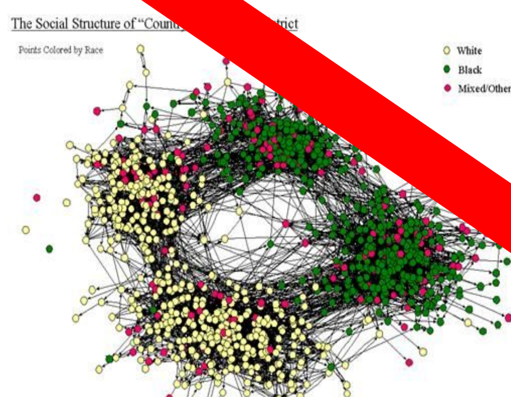
Current needs, future knowledge:

- Massive, ***irregularly structured*** input data.
- New simulation, analysis methods
- Widely varied, unexplored response / control methods



AIM	myAOL	Ask	Backflip
BallHype	Bebo	BlinkList	Blogmarks
Faves	Yahoo Buzz	Delicious	Digg
Diigo	Email	Facebook	Favorites
Fark	FeedMeLinks	FriendFeed	Furl
Google	Kaboodle	kIRTSY	Link-a-Gogo
LinkedIn	Live	Magnolia	Mister Wong
Mixx	Multiply	MyWeb	MySpace
Netvouz	Newsvine	Propeller	Reddit
Segnalo	Simpy	Sk*rt	Slashdot
Spurl	StumbleUpon	Stylehive	Tailrank
Technorati	ThisNext	Twitter	Yardbarker
Yahoo			

NYSE CDC Centers for Disease Control and Prevention • • •

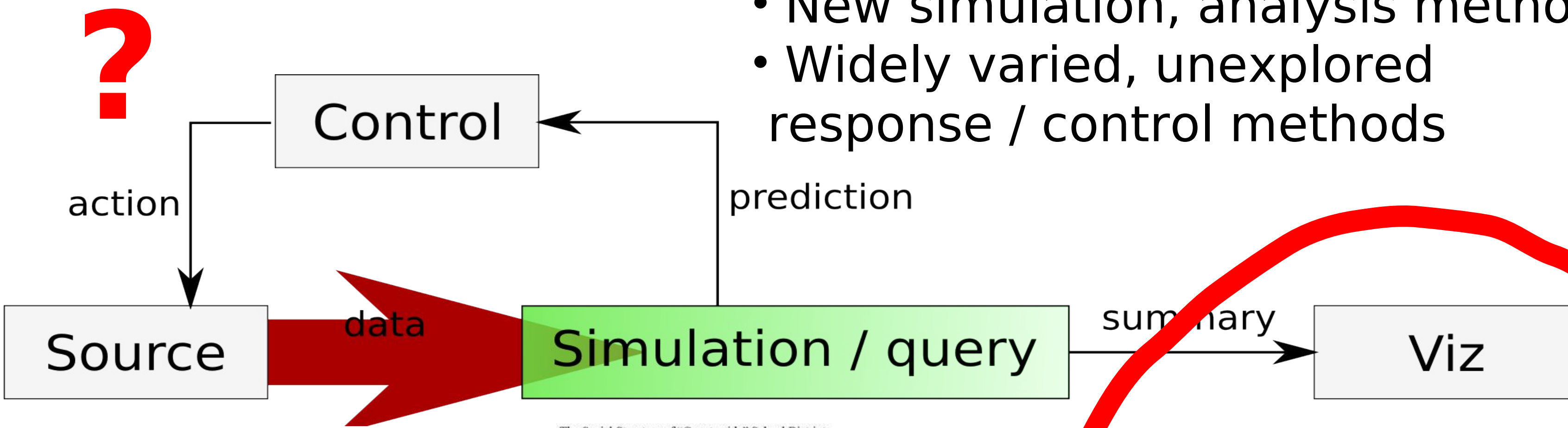


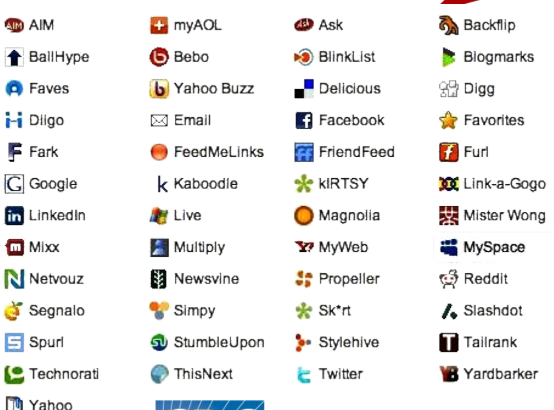
Analysts need us here. Yesterday.
Closing the loop needs acceleration.

Streaming Data Analysis

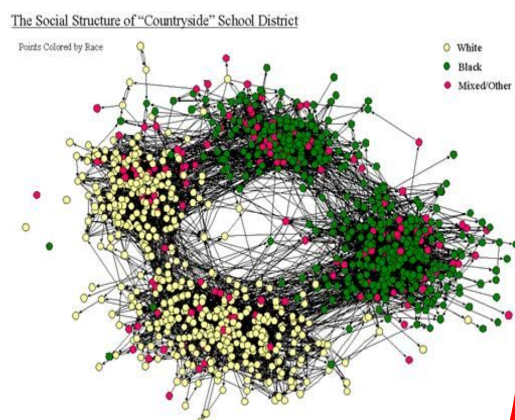
Current needs, future knowledge:

- Massive, ***irregularly structured*** input data.
- New simulation, analysis methods
- Widely varied, unexplored response / control methods





NYSE CDC Centers for Disease Control and Prevention …



Facebook friendship graph: 30k edges *per pixel* on 1600x1200 screen!

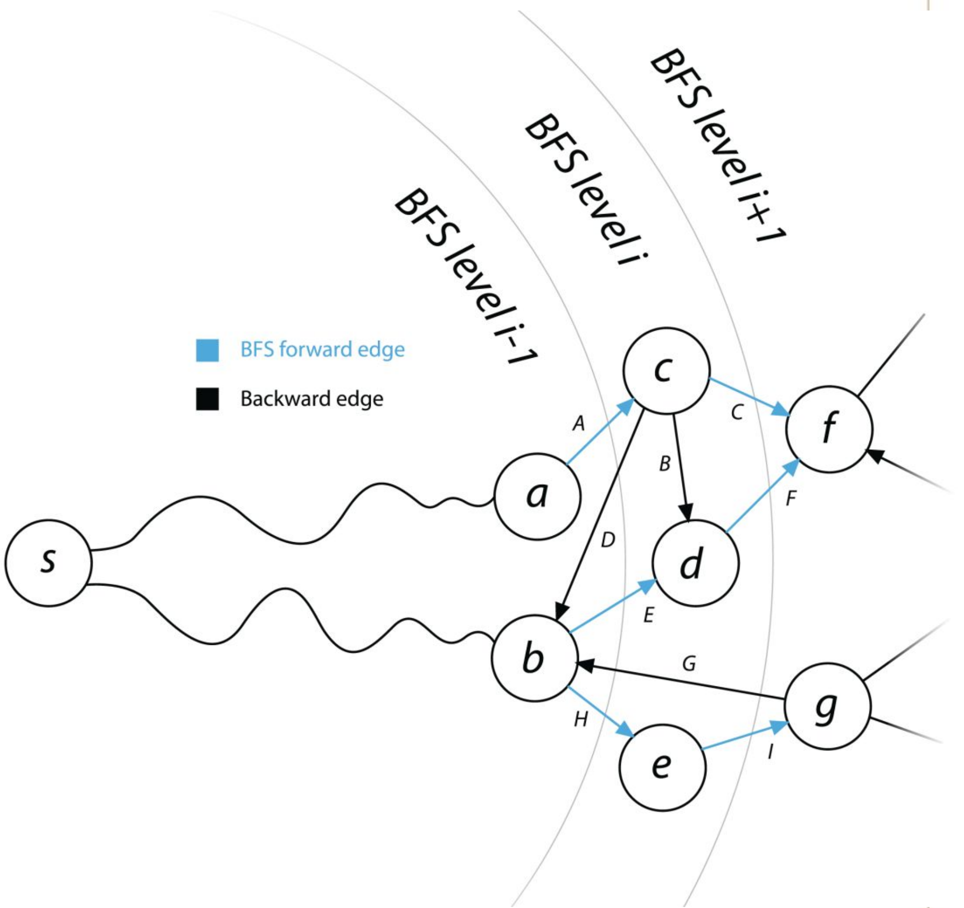
Georgia Tech College of Computing

Algorithms

- Just to set the stage for discussing accelerators:
  - *k*-Betweenness Centrality
    - (showing effectiveness on one alternative architecture, the Cray XMT)
  - Agglomerative community identification
    - (could be very useful to assist “acceleration” by data filtering)

K-Betweenness Centrality

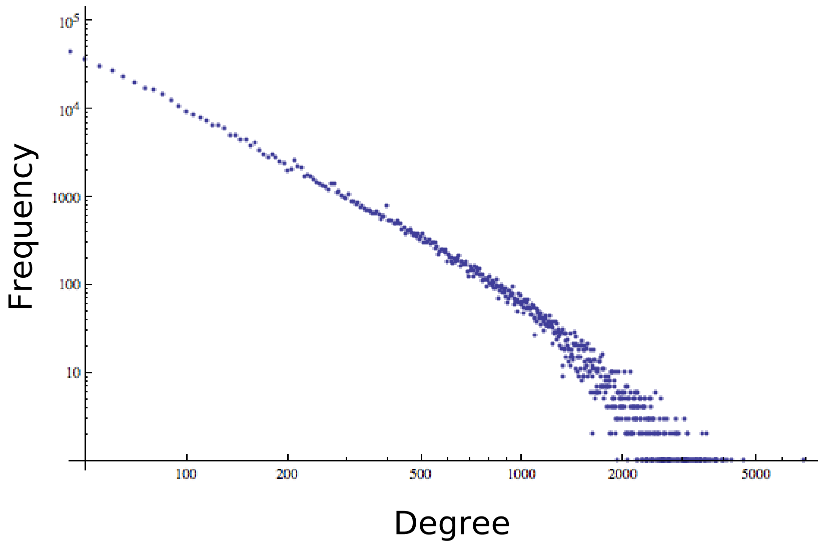
- Count short paths (shortest + *k*) relative to all paths.
- Maintain multiple BFS fronts (single: Dijkstra's algorithm).
- Each vertex has a forward, lock-free queue of relevant edges.
- Distances and BC_k values can be computed by a second pass along queued edges.
- Not directly expressible (by our attempts) in linear algebra or mathematical optimization.
- Cost exponential in *k*, *O(mn)* for fixed *k*. Can be approx.

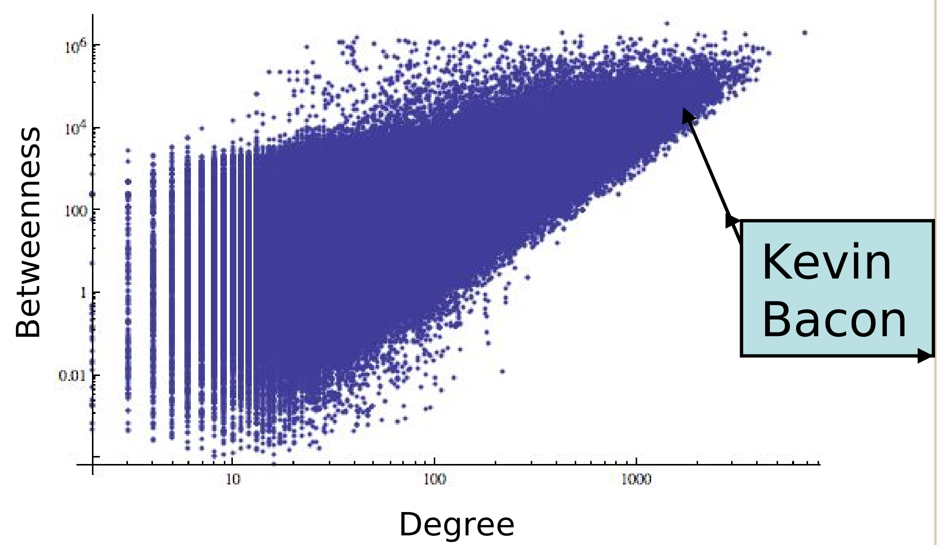


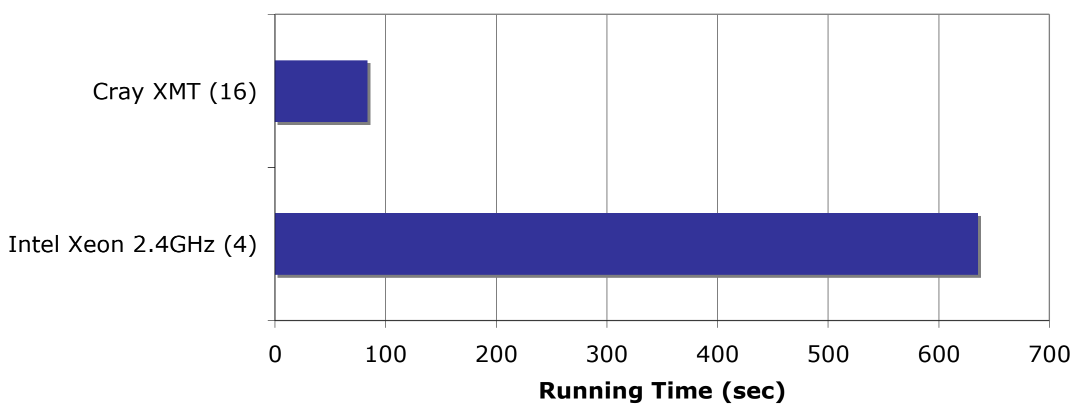
Brandes, 2001; Bader, *et al.* 2009 & 2009.

IMDB Movie Actor Network (Approx BC_0)

An undirected graph of 1.54 million vertices (movie actors) and 78 million edges. An edge corresponds to a link between two actors if they have acted together in a movie.







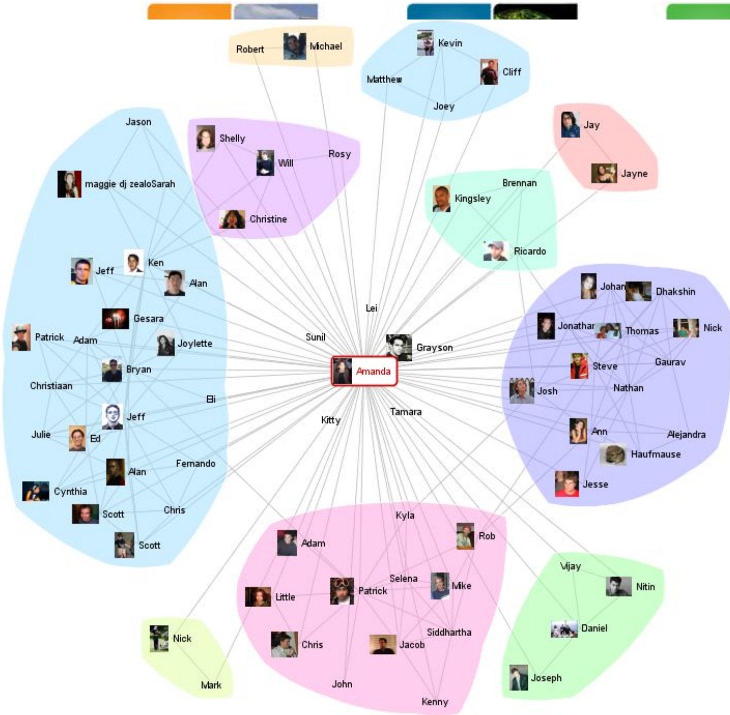
Disparity grows with *k*…
So what is the XMT?

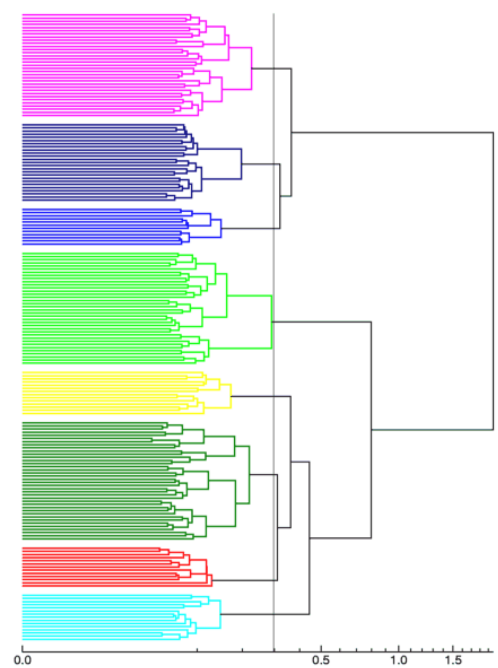
Alternative architecture: Cray XMT

- Tolerates latency by extreme multithreading
  - Each processor supports 128 hardware threads
  - Context switch in a single tick
  - No cache or local memory
  - Context switch on memory request
  - Multiple outstanding loads
- Remote memory requests do not stall processors
  - Other streams work while the request gets fulfilled
- Light-weight, word-level synchronization
  - Minimizes access conflicts
- Hashed global shared memory
  - 64-byte granularity, minimizes hotspots
- **High-productivity graph analysis!**

Community Identification

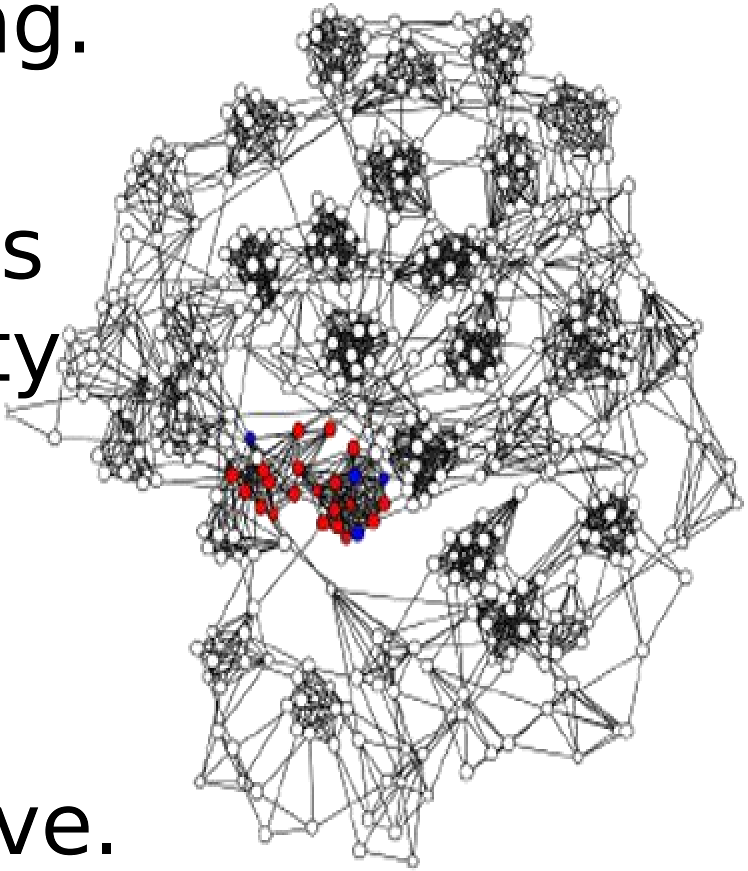
- Implicit communities in large-scale networks are of interest in many cases.
  - WWW
  - Social networks
  - Biological networks
- Formulated as a graph clustering problem.
  - Informally, identify/extract “dense” sub-graphs.
- Several different objective functions exist.
  - Metrics based on intra-cluster vs. inter-cluster edges, community sizes, number of communities, overlap …
- Agglomerative, bottom-up:
  - Evaluate metric change, merge (independent) sets to maximize change.





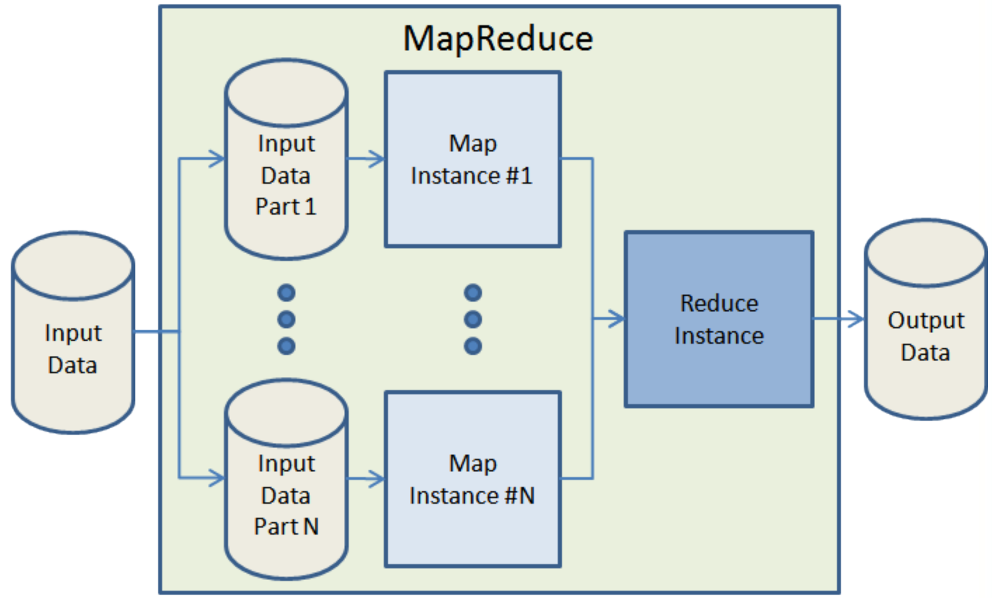
Seed Set Expansion

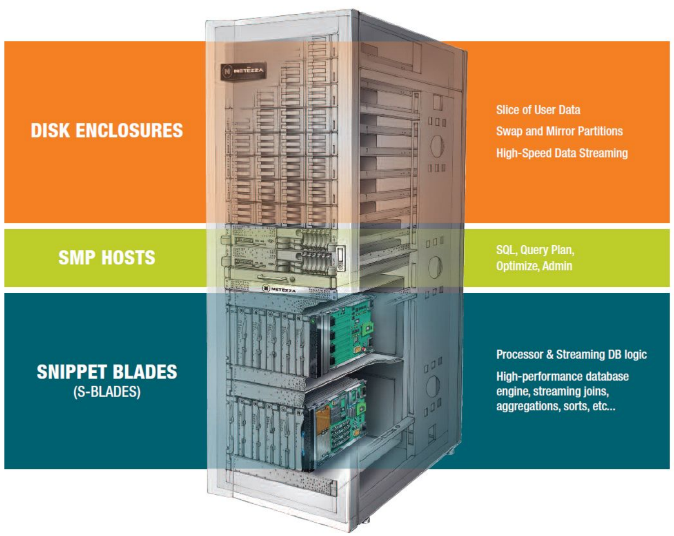
- Useful to find communities to which several vertices belong.
- Blue vertices are
  are seeds, red vertices belong to a community of interest.
- Selection for viz,
  analysis...
- Consider agglomerative.



Possible accelerators

- Map-Reduce: A large aggregate disk capacity with data replication support (Hadoop File System)....
- Netezza Twin-Fin: FPGAs to filter/reduce data...





Source: Netezza

- Currently experimenting with agglomerative methods on multithreaded architectures.
- **Can we express clustering / community detection as a selection rule instead?**

Acknowledgment of Support

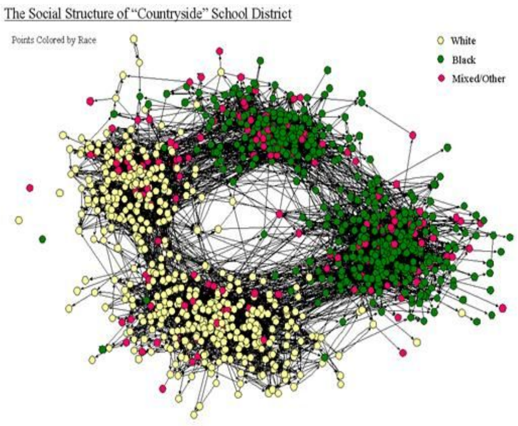
Extra information on sizes / rates

Data Volumes in Commercial Sector

- **EBay** (April 2009) has a pair of data warehouses
  - **>2 PB**, traditional
  - **6.5PB,** 17 trillion records, 1.5B records/day, each web click is 50-150 details
  - Source: http://www.dbms2.com/2009/04/30/ebays-two-enormous-data-warehouses/
- **Facebook** (May 2009):
  - Estimate of **2.5PB** of user data
  - 15 TB of new data per day
  - Queries to develop targeted ads are run hourly
  - Source: http://www.dbms2.com/2009/05/11/facebook-hadoop-and-hive/
- In 2008: http://www.dbms2.com/2008/10/15/teradatas-petabyte-power-players/
  - **Walmart**: **2.5PB**
  - **Bank of America**: **1.5PB**
  - **Dell**: **1PB**

Data Volumes: *Current* data sets

- NYSE: 1.5TB daily, 8PB maintained
- Google: “Several dozen” 1PB sets (CACM Jan 2010)
- LHC: 15PB per year (avg 41TB/day)
  - (http://public.web.cern.ch/public/en/lhc/Computing-en.html)
- LSST: 13TB nightly
  - (http://www.lsst.org/Project/docs/data-challenge.pdf)
- Wal-Mart: 536TB, 1B entries daily (2006)
- Facebook: 350M users, 3.5B shared items/week

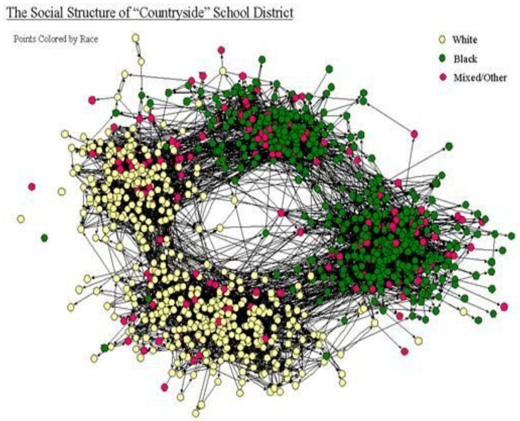


- All data is **rich.**
- Data rates do not include building **relationships**.

Data Volumes: *Current* data rates

- NYSE: 1.5TB daily
- LHC: 41TB daily
- LSST: 13TB daily

- 1 Gb Ethernet: 8.7TB daily at 100%, 5-6TB daily realistic
- Multi-TB storage on 10GE: 300TB daily read, 90TB daily write

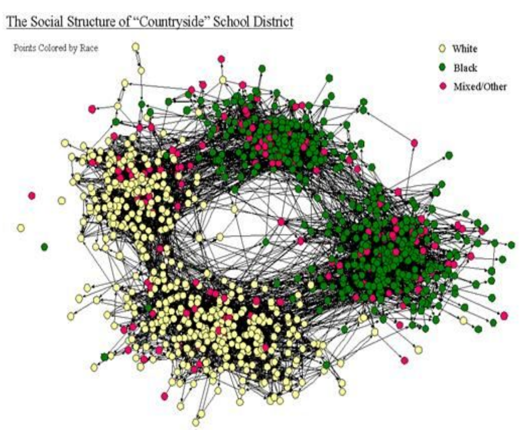


- **Current** data is at the limit of **current** systems.
- Not counting relationships…

Data Volumes: *Future* data rates

- Facebook: >2x yearly
- Twitter: >10x yearly
- Growing sources:
  - Bioinformatics
  - Nano-scale devices
  - Security

- Ethernet: 4x in next 2 years. Maybe.
- Flash storage, direct: 10x write, 4x read. Huge cost for multi-PB storage.

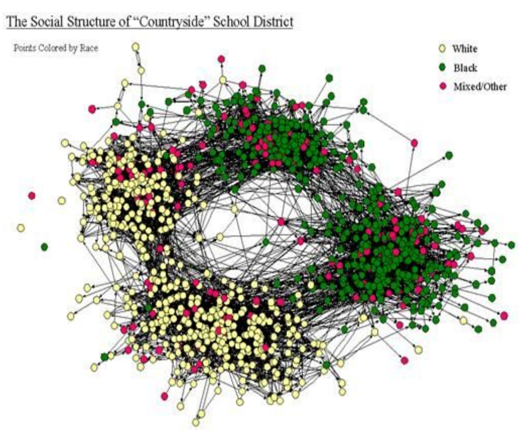


- **Data rate growth** is outstripping technology.
- Then consider: latency, ingest, processing, response…

Data Volumes: *Current* data sets

- NYSE: 8PB
- Google: >12PB
- LHC: >15PB

- CPU ↔ Memory:
  – QPI,HT: 2PB/day@100%
  – Power7: 8.7PB/day
- Mem:
  – NCSA Blue Waters target: 2PB



➔ Even with parallelism, current (in-progress) systems cannot handle more than a few passes... per day.