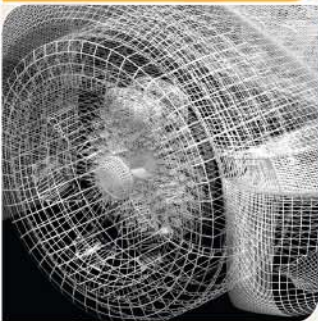


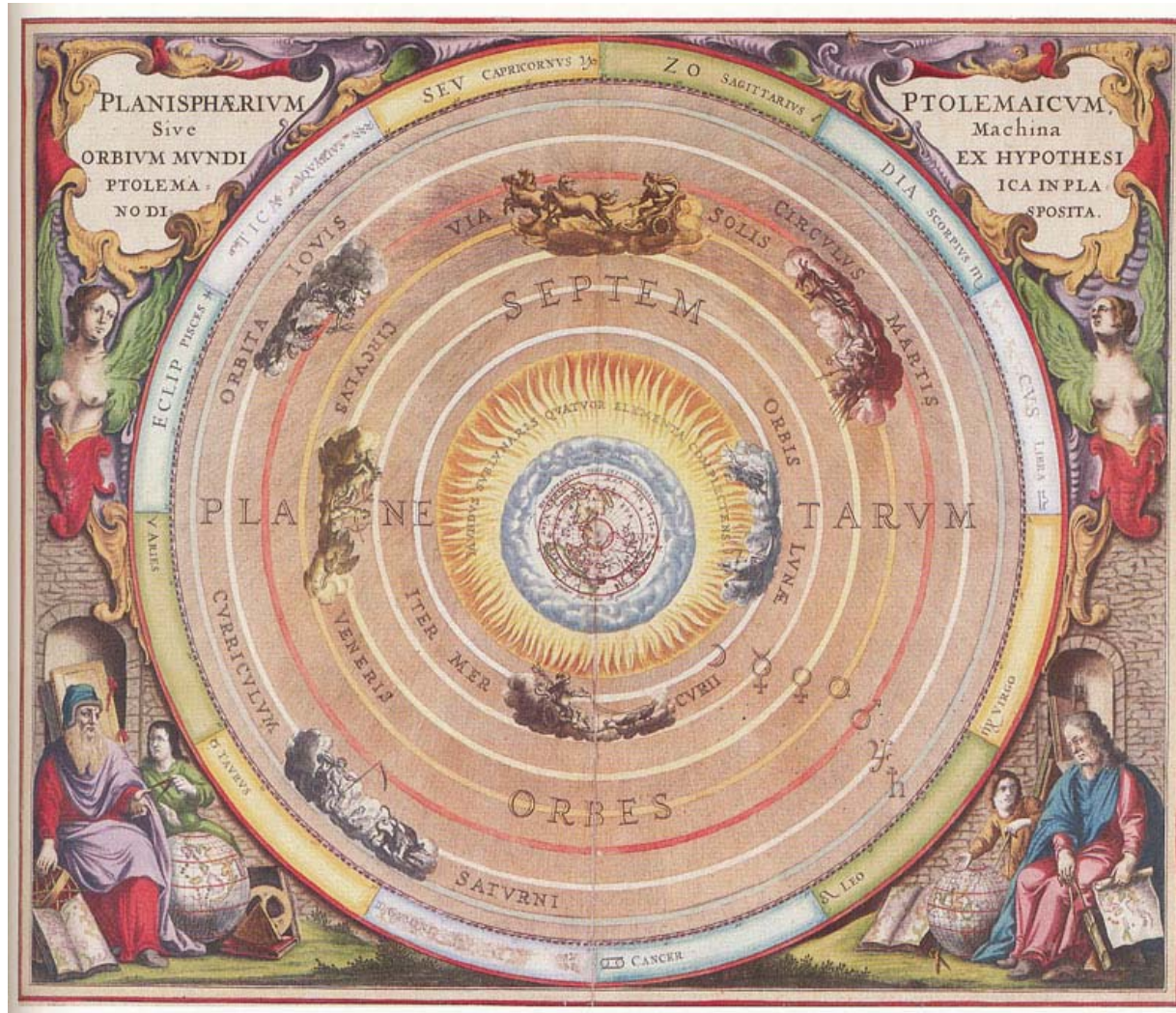
Opportunities and Challenges in Massive Data-Intensive Computing

David A. Bader



**Georgia
Tech**  **College of
Computing**
Computational Science and Engineering

Ptolemy: Floating-point centric ...



Copernicus: Data-centric



Figure 2 – This diagram from Copernicus’ original manuscript places the Sun at the centre of the universe.



OPPORTUNITIES



Opportunities

- Application-oriented Opportunities:
 - High performance computing for massive graphs
 - Streaming analytics
 - Informational Visualization techniques for massive graphs
 - Heterogeneous systems: Methodologies for combining the use of the Cloud and Manycore for high-performance computing
 - Energy-efficient high-performance computing



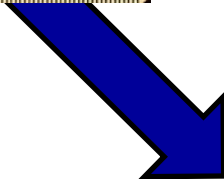
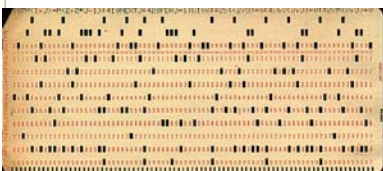
Opportunity 1: High performance computing for massive graphs

- Traditional HPC has focused primarily on solving large problems from chemistry, physics, and mechanics, using dense linear algebra.
 - HPC faces new challenges to deal with:
 - time-varying interactions among entities, and
 - massive-scale graph abstractions where the vertices represent the nouns or entities and the edges represent their observed interactions.
 - Few parallel computers run well on these problems because
 - they often lack locality required to get high performance from distributed-memory cache-based supercomputers.
 - **Case study:** Massively threaded architectures are shown to run several orders of magnitude faster than the fastest supercomputers on these types of problems!
- ➔ A focused research agenda is needed to design algorithms that scale on these new platforms.



Opportunity 2: Streaming analytics

- While our high performance computers have delivered a sustained petaflop, they have done so using the same antiquated **batch processing** style where a program and a static data set are scheduled to compute in the next available slot.
 - Today, data is overwhelming in volume *and rate*, and we struggle to keep up with these **streams**.
- ➔ Fundamental computer science research is needed in:
- ➔ the design of streaming architectures, and
 - ➔ data structures and algorithms that can compute important analytics while sitting in the middle of these torrential flows.



VS.

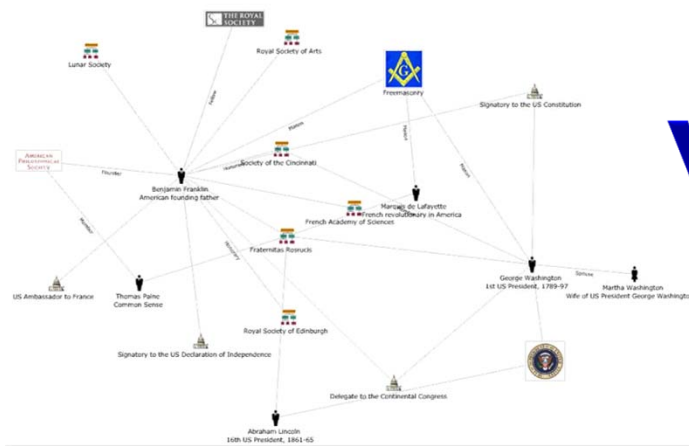


Opportunity 3: Information Visualization techniques for massive graphs



- Information Visualization today
 - addresses traditional scientific computing (fluid flow, molecular dynamics), or
 - when handling discrete data, scale to only hundreds of vertices at best.
- ➔ However, there is a strong need for visualization in the data sciences so that analytics can gain understanding from data sets with from millions to billions of interacting non-planar discrete entities.
 - Applications include: data mining, intelligence, situational awareness

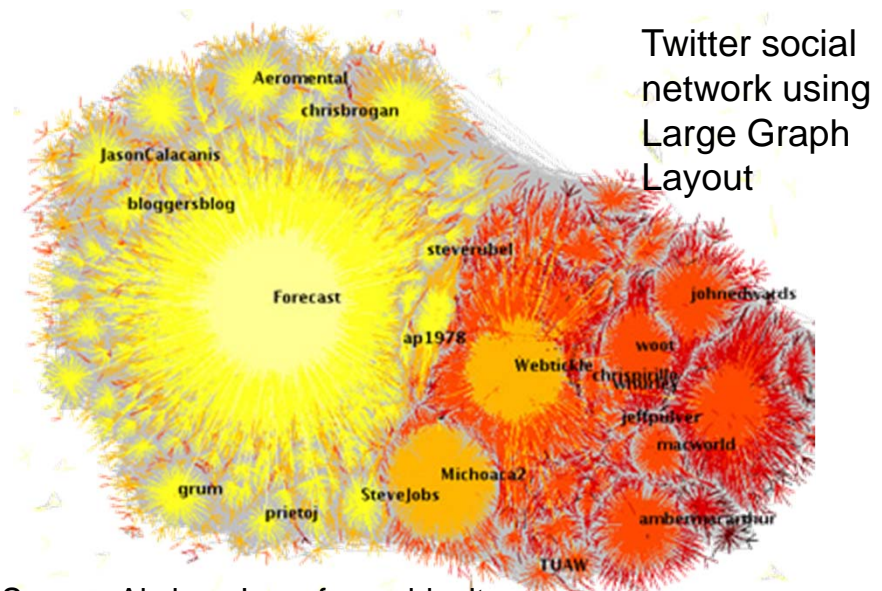
NNDB tracking the entire world



NNDB Mapper of George Washington

David A. Bader

VS.

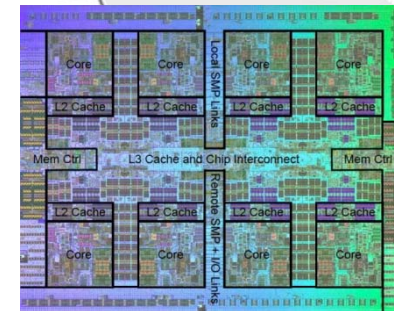
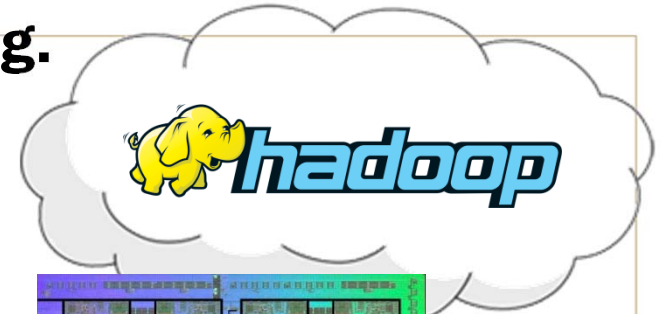


Twitter social network using Large Graph Layout

Source: Akshay Java, from ebiquity group

Opportunity 4: Heterogeneous Systems: Methodologies for combining the use of the Cloud and Manycore for high-performance computing.

- Today, there is a dichotomy between using clouds (e.g. Hadoop, map-reduce) for massive data storage, filtering, summarization, and massively parallel/multithreaded systems for data-intensive computation.
- We must develop methodologies for employing these complementary systems for solving grand challenges in data analysis.



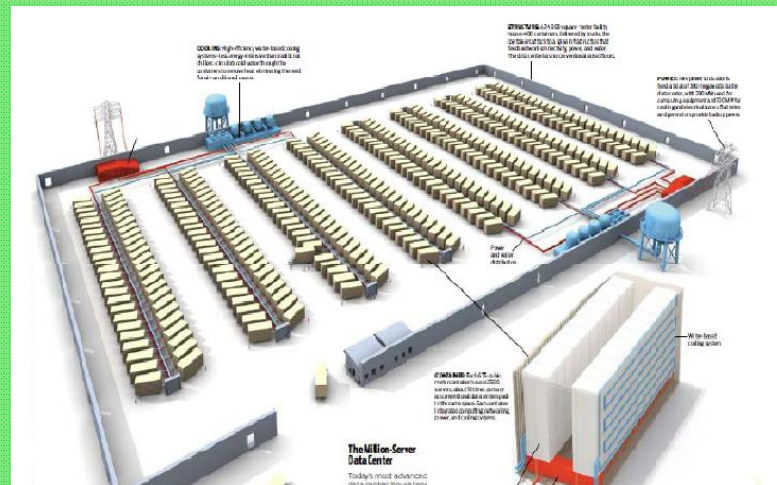
Steve Mills, SVP of IBM Software (left), and Dr. John Kelly, SVP of IBM Research, view Stream Computing technology





Opportunity 5: Energy-efficient high-performance computing

- The main constraint for our ability to compute has changed
 - from availability of compute resources
 - to the ability to power and cool our systems within budget.
- ➔ Holistic research is needed that can permeate from the architecture and systems up to the applications AND DATA CENTERS, whereby energy use is a first-class object that can be optimized at all levels.



Microsoft's Chicago Million Server DataCenter



MOTIVATION

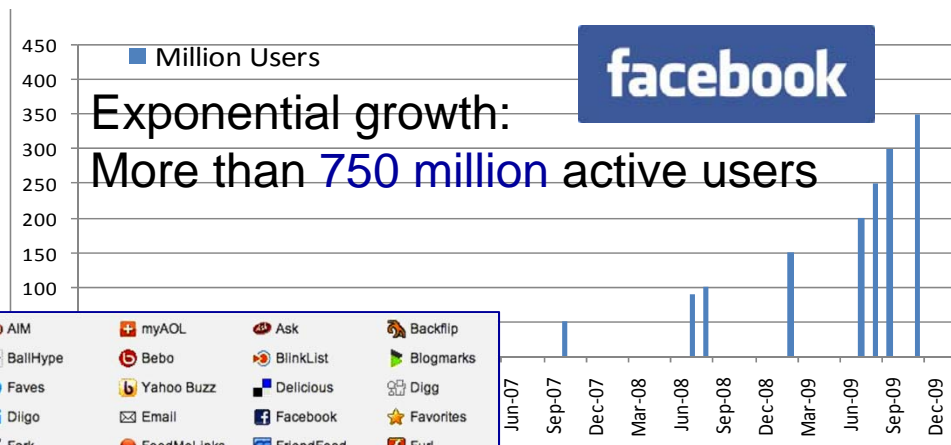
Exascale Streaming Data Analytics:

Real-world challenges



All involve analyzing massive streaming complex networks:

- **Health care** → disease spread, detection and prevention of epidemics/pandemics (e.g. SARS, Avian flu, H1N1 “swine” flu)
- **Massive social networks** → understanding communities, intentions, population dynamics, pandemic spread, transportation and evacuation
- **Intelligence** → business analytics, anomaly detection, security, knowledge discovery from massive data sets
- **Systems Biology** → understanding complex life systems, drug design, microbial research, unravel the mysteries of the HIV virus; understand life, disease,
- **Electric Power Grid** → communication, transportation, energy, water, food supply
- **Modeling and Simulation** → Perform full-scale economic-social-political simulations



Ex: discovered minimal changes in O(billions)-size complex network that could hide or reveal top influencers in the community

- Sample queries:**
- Allegiance switching:** identify entities that switch communities.
 - Community structure:** identify the genesis and dissipation of communities
 - Phase change:** identify significant change in the network structure

REQUIRES PREDICTING / INFLUENCE CHANGE IN REAL-TIME AT SCALE



Ubiquitous High Performance Computing (UHPC)



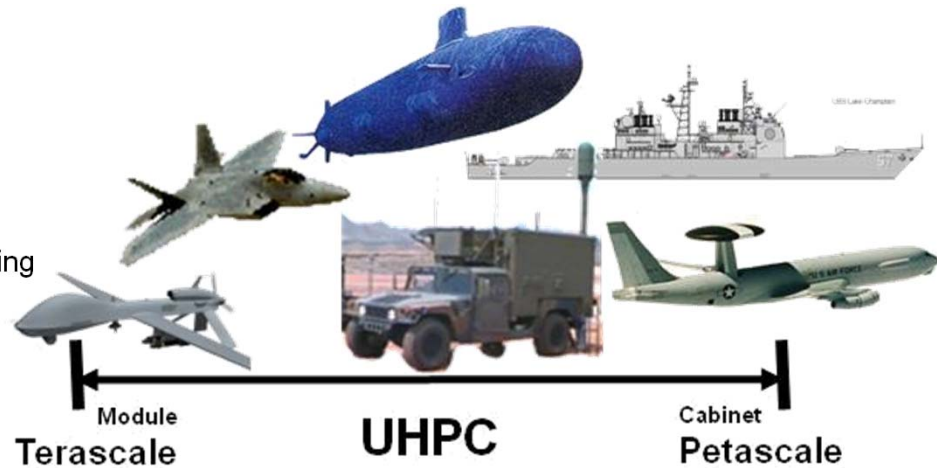
Goal: develop highly parallel, security enabled, power efficient processing systems, supporting ease of programming, with resilient execution through all failure modes and intrusion attacks

Architectural Drivers:

- Energy Efficient
- Security and Dependability
- Programmability

Program Objectives:

- One PFLOPS, single cabinet including self-contained cooling
- 50 GFLOPS/W (equivalent to 20 pJ/FLOP)
- Total cabinet power budget 57KW, includes processing resources, storage and cooling
- Security embedded at all system levels
- Parallel, efficient execution models
- Highly programmable parallel systems
- Scalable systems – from terascale to petascale



David A. Bader (CSE)
Echelon Leadership Team



“NVIDIA-Led Team Receives \$25 Million Contract From DARPA to Develop High-Performance GPU Computing Systems” -MarketWatch

Echelon: Extreme-scale Compute Hierarchies with Efficient Locality-Optimized Nodes





Information Innovation Office

PRODIGAL: *Proactive Detection of Insider Threats with Graph Analysis and Learning*

ADAMS Program Kickoff Meeting, June 6-7, 2011

SAIC

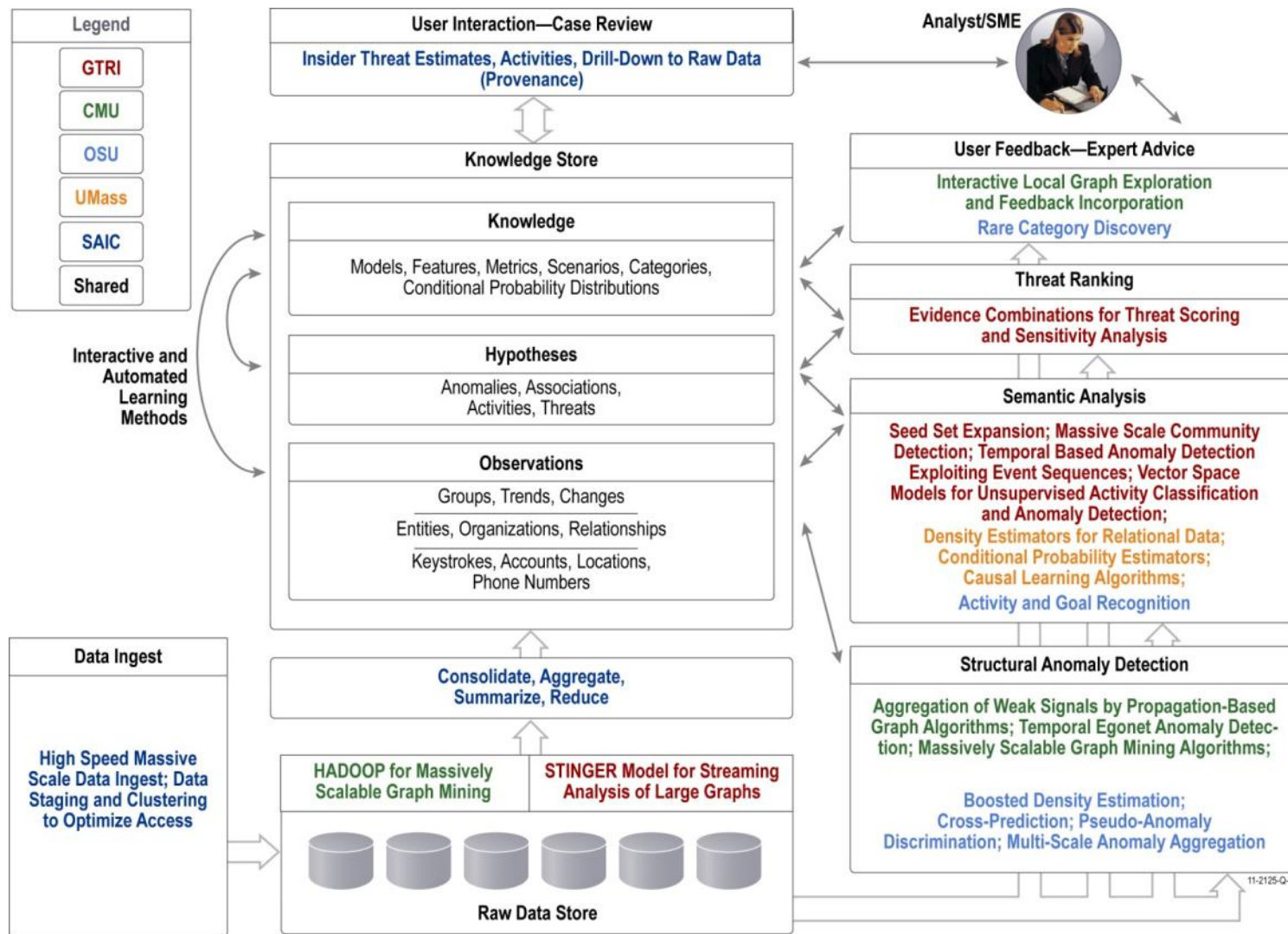
Georgia Tech Research Institute
Carnegie-Mellon University
Oregon State University
University of Massachusetts



The PRODIGAL Architecture

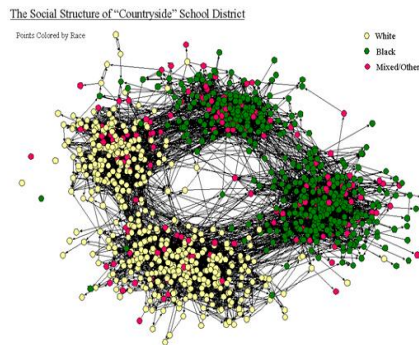


Information Innovation Office



Center for Adaptive Supercomputing Software for MultiThreaded Architectures (CASS-MT)

- Launched July 2008
- Pacific-Northwest Lab
 - Georgia Tech, Sandia, WA State, Delaware
- The newest breed of supercomputers have hardware set up not just for speed, but also to better tackle large networks of seemingly random data. And now, a multi-institutional group of researchers has been awarded over \$14 million to develop software for these supercomputers. Applications include anywhere complex webs of information can be found: from internet security and power grid stability to complex biological networks.



David A. Bader



Example: Mining Twitter for Social Good

ICPP 2010

Massive Social Network Analysis: Mining Twitter for Social Good

David Ediger Karl Jiang
Jason Riedy David A. Bader
Georgia Institute of Technology
Atlanta, GA, USA

Courtney Corley Rob Farber
Pacific Northwest National Lab.
Richland, WA, USA

William N. Reynolds
Least Squares Software, Inc
Albuquerque, NM, USA

Abstract—Social networks produce an enormous quantity of data. Facebook consists of over 400 million active users sharing over 5 billion pieces of information each month. Analyzing this vast quantity of unstructured data presents challenges for software and hardware. We present GraphCT, a Graph Characterization Toolkit for massive graphs representing social network data. On a 128-processor Cray XMT, GraphCT estimates the betweenness centrality of an artificially generated (R-MAT) 537 million vertex, 8.6 billion edge graph in 55 minutes and a real-world graph (Kwak, *et al.*) with 61.6 million vertices and 1.47 billion edges in 105 minutes. We use GraphCT to analyze public data from Twitter, a microblogging network. Twitter's message connections appear primarily tree-structured as a news dissemination system. Within the

involves over 400 million active users with an average of 120 'friendship' connections each and sharing 5 billion references to items each month [11].

One analysis approach treats the interactions as and applies tools from graph theory, social network analysis, and scale-free networks [29]. However, the volume of data that must be processed to apply techniques overwhelms current computational capabilities. Even well-understood analytic methodology advances in both hardware and software to process the growing corpus of social media.

Social media provides staggering amounts of information, but it is difficult to extract useful information from this data.

TOP 15 USERS BY BETWEENNESS CENTRALITY

Rank	H1N1	Data Set
1	@CDCFlu	@ajc
2	@addthis	@driveafaste
3	@Official_PAX	@ATLCheap
4	@FluGov	@TWCi
5	@nytimes	@HelloNorthGA
6	@tweetmeme	@11AliveNews
7	@mercola	@WSB_TV
8	@CNN	@shaunking
9	@backstreetboys	@Carl
10	@EllieSmith_x	@SpaceyG
11	@TIME	@ATLINTownPa
12	@CDCemergency	@TJsDJs
13	@CDC_eHealth	@ATLien
14	@perezhilton	@MarshallRamsey
15	@billmaher	@Kanye

twitter
public tweets

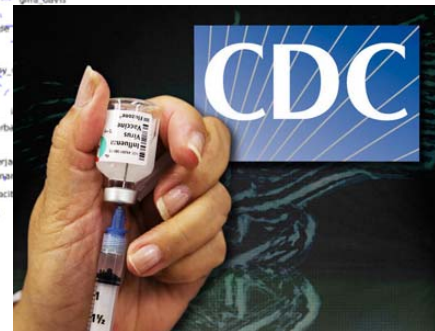
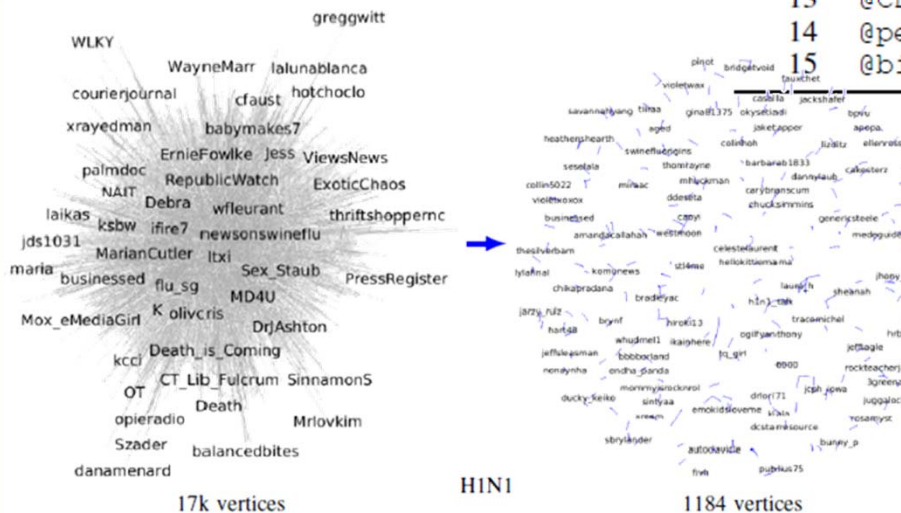


Image credit: bioethicsinstitute.org

Fig. 3. Subcommunity filtering on Twitter data sets

David A. Bader

Georgia
Tech

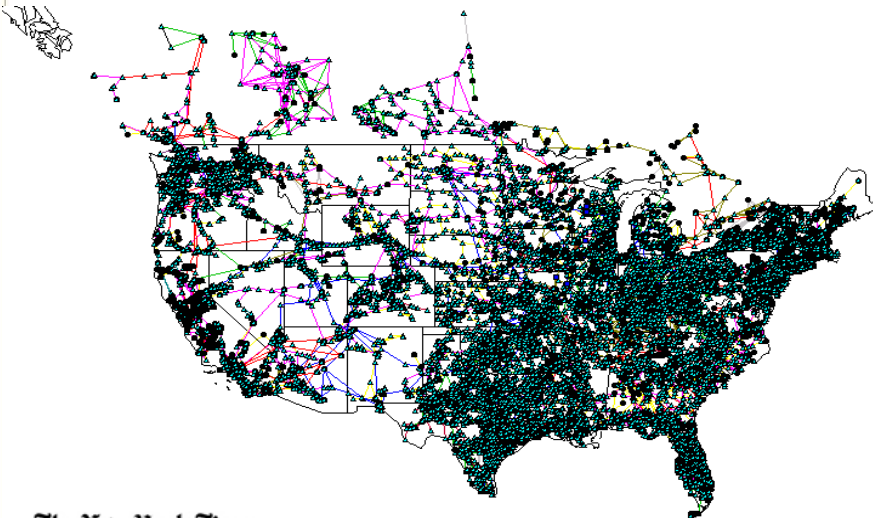
College of
Computing

Pacific Northwest
NATIONAL LABORATORY



Massive Data Analytics: Protecting our Nation

US High Voltage Transmission Grid (>150,000 miles of line)









The New York Times
Thursday, September 4, 2008

Report on Blackout Is Said To Describe Failure to React

By MATTHEW L. WALD
Published: November 12, 2003

A report on the Aug. 14 blackout identifies specific lapses by various parties, including FirstEnergy's failure to react properly to the loss of a transmission line, people who have seen drafts of it say.

A working group of experts from eight states and Canada will meet in private on Wednesday to evaluate the report, people involved in the investigation said Tuesday. The report, which the Energy Department

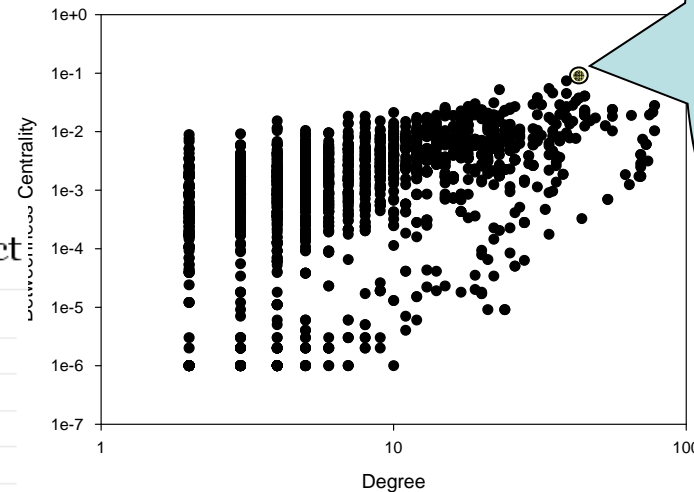
-  E-MAIL
-  PRINT
-  SINGLE-PAGE
-  REPRINTS
-  SAVE
-  SHARE

David A. Bader

Public Health

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

Human Genome core protein interactions
Degree vs. Betweenness Centrality



ENSG0000145332.2
Kelch-like protein implicated in breast cancer



Network Analysis for Intelligence and Surveillance

- [Krebs '04] Post 9/11 Terrorist Network Analysis from public domain information
- Plot masterminds correctly identified from interaction patterns: **centrality**
- A global view of entities is often more insightful
- Detect anomalous activities by exact/approximate **graph matching**

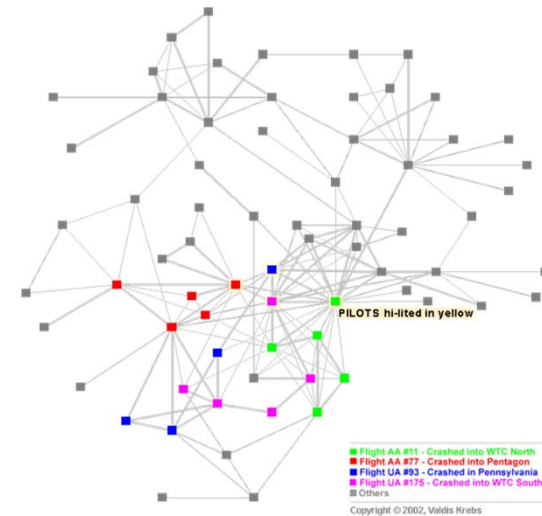


Image Source: <http://www.orgnet.com/hijackers.html>

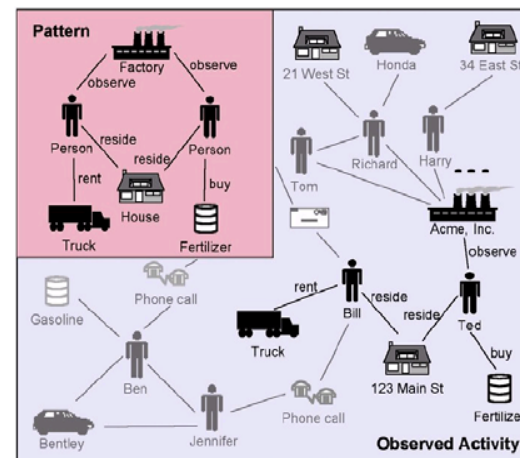


Image Source: T. Coffman, S. Greenblatt, S. Marcus, Graph-based technologies for intelligence analysis, CACM, 47 (3, March 2004): pp 45-47



Graphs are pervasive in large-scale data analysis

- **Sources** of massive data: petascale simulations, experimental devices, the Internet, scientific applications.
- **New challenges for analysis:** data sizes, heterogeneity, uncertainty, data quality.

Astrophysics

Problem: Outlier detection.

Challenges: massive datasets, temporal variations.

Graph problems: clustering, matching.



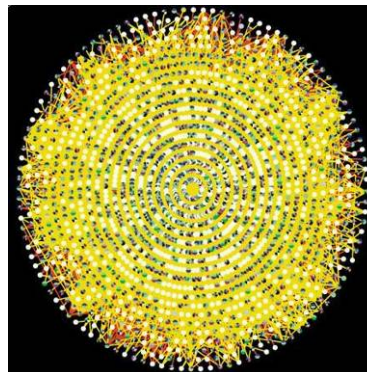
Image sources: (1) http://physics.nmt.edu/images/astro/hst_starfield.jpg
(2,3) www.visualComplexity.com

Bioinformatics

Problem: Identifying drug target proteins.

Challenges: Data heterogeneity, quality.

Graph problems: centrality, clustering.

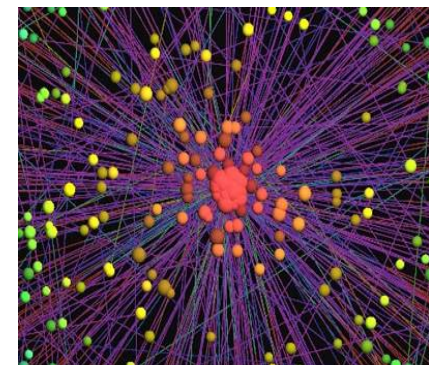


Social Informatics

Problem: Discover emergent communities, model spread of information.

Challenges: new analytics routines, uncertainty in data.

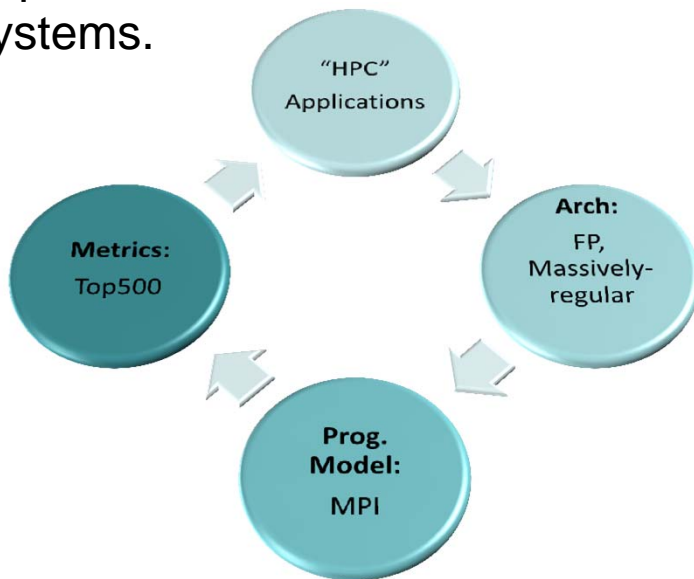
Graph problems: clustering, shortest paths, flows.





Flywheel has driven HPC into a corner

For decades, HPC has been on a vicious cycle of enabling applications that run well on HPC systems.

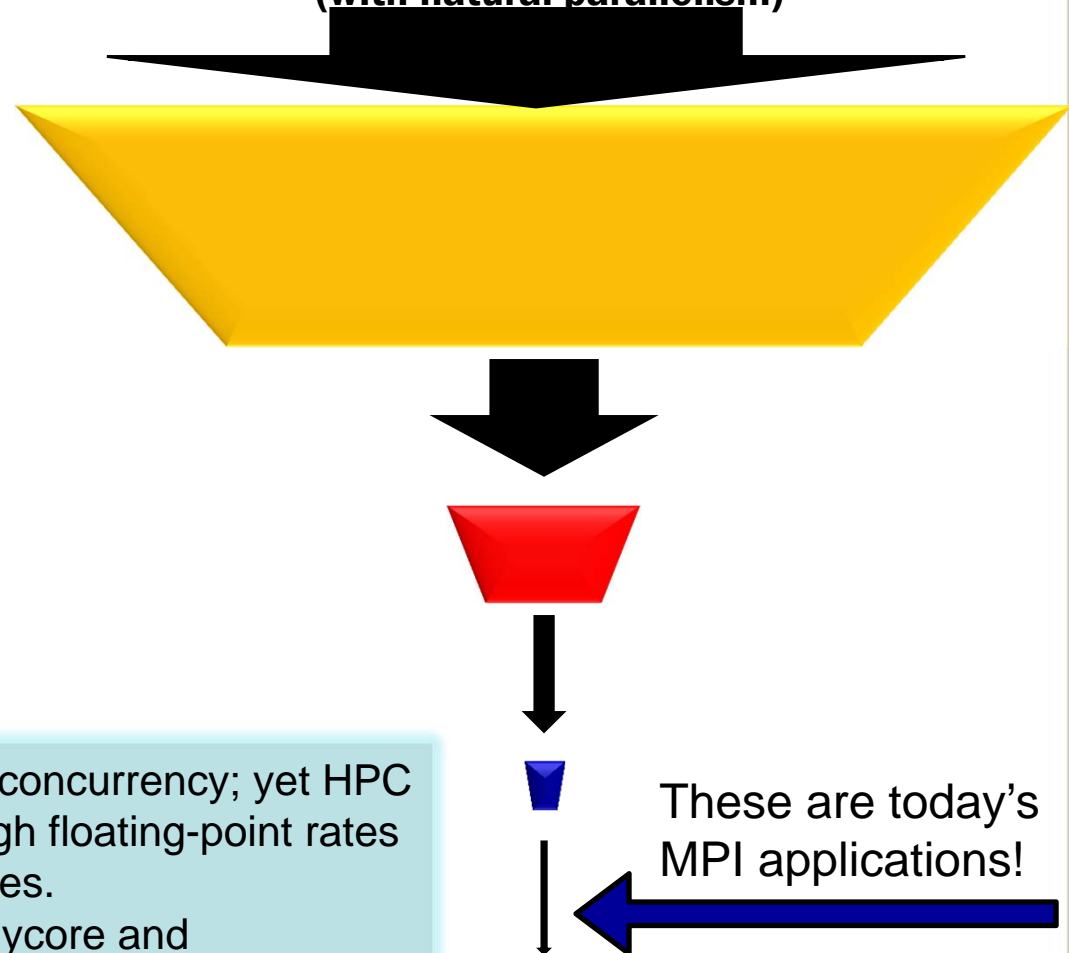


→ Data-intensive computing has natural concurrency; yet HPC architectures are designed to achieve high floating-point rates by exploiting spatial and temporal localities.

- For the first time in **decades**, manycore and multithreaded can let us rethink architecture.

Real-World Applications

(with natural parallelism)



Architectural Requirements for the Efficient Graph Analysis (**Challenges**)



- Runtime is dominated by latency
 - Random accesses to global address space
 - Perhaps many at once
- Essentially no computation to hide memory costs
- Access pattern is data dependent
 - Prefetching unlikely to help
 - Usually only want small part of cache line
- Potentially abysmal locality at **all** levels of memory hierarchy

Architectural Requirements for the Efficient Graph Analysis (Desired Features)



- A large memory capacity
- Low latency / high bandwidth
 - For small messages!
- Latency tolerant
- Light-weight synchronization mechanisms
- Global address space
 - No graph partitioning required
 - Avoid memory-consuming profusion of ghost-nodes
 - No local/global numbering conversions



Streaming Graphs

- ▶ **STINGER: A Data Structure for Graphs with Streaming Updates**
 - Light-weight data structure that supports efficient iteration *and* efficient updates.
- ▶ **Experiments with Streaming Updates to Clustering Coefficients**
 - Working with bulk updates, can handle almost 200k per second

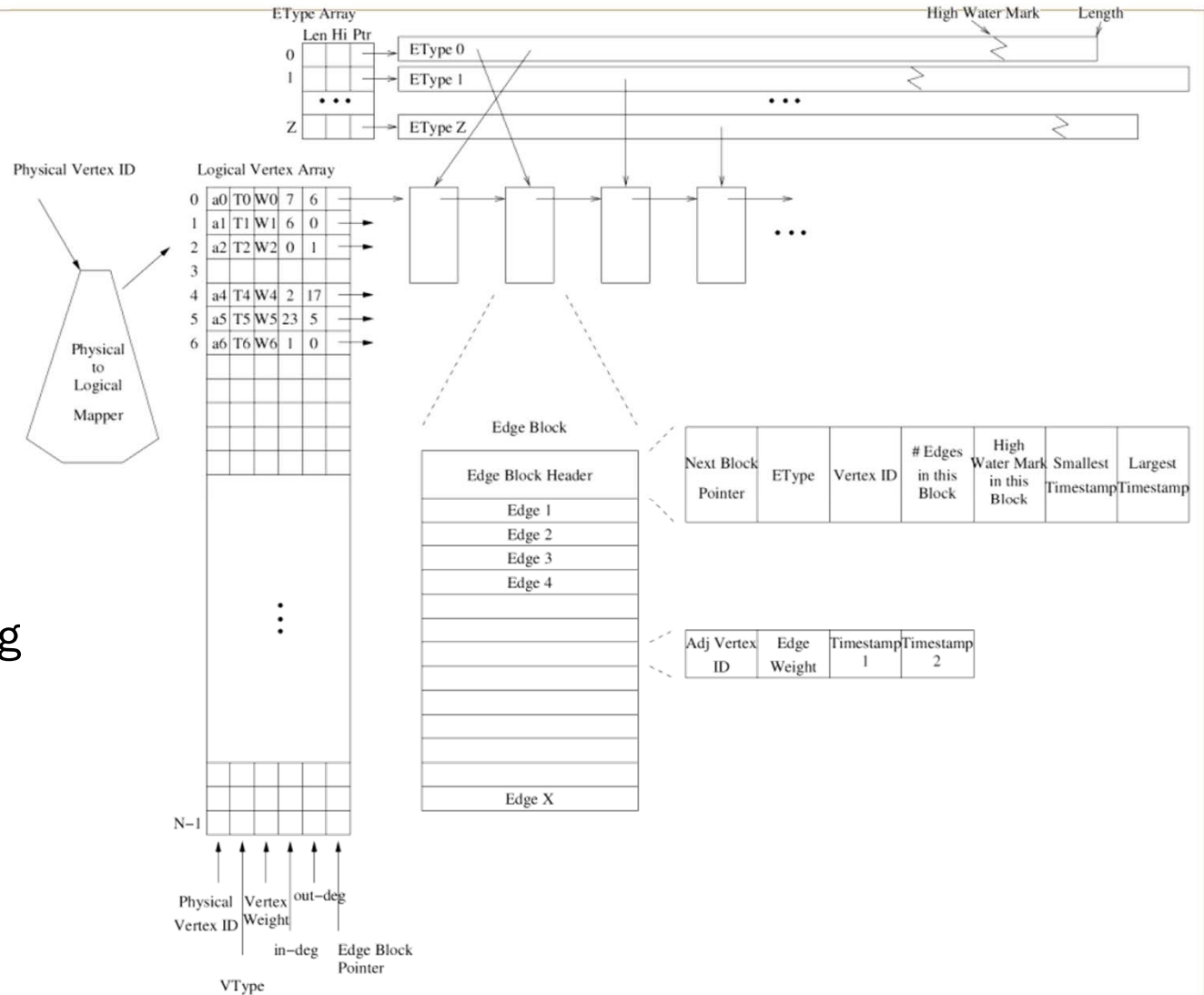


STING Extensible Representation (STINGER)

- ▶ Enhanced representation developed for dynamic graphs developed in consultation with David A. Bader, Johnathan Berry, Adam Amos-Binks, Daniel Chavarría-Miranda, Charles Hastings, Kamesh Madduri, and Steven C. Poulos.
- ▶ Design goals:
 - Be useful for the entire “large graph” community
 - Portable semantics and high-level optimizations across multiple platforms & frameworks (XMT C, MTGL, etc.)
 - Permit good performance: No single structure is optimal for all.
 - Assume globally addressable memory access
 - Support multiple, parallel readers and a single writer
- ▶ Operations:
 - Insert/update & delete both vertices & edges
 - Aging-off: Remove old edges (by timestamp)
 - Serialization to support checkpointing, etc.

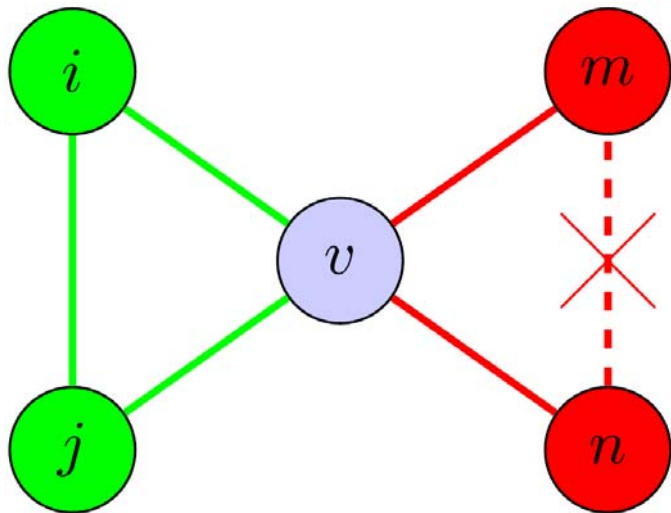
STING Extensible Representation

- ▶ Semi-dense edge list blocks with free space
- ▶ Compactly stores timestamps, types, weights
- ▶ Maps from application IDs to storage IDs
- ▶ Deletion by negating IDs, separate compaction



Testbed: Clustering Coefficients

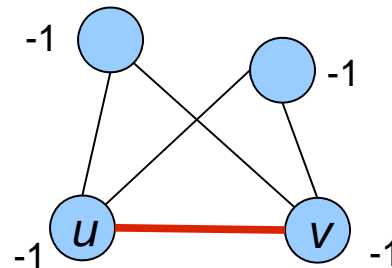
- ▶ Roughly, the ratio of actual triangles to possible triangles around a vertex.



- ▶ Defined in terms of **triplets**.
- ▶ $i-j-v$ is a **closed triplet** (triangle).
- ▶ $m-v-n$ is an **open triplet**.
- ▶ Clustering coefficient
closed triplets / # all triplets
- ▶ Locally, count those around v .
- ▶ Globally, count across entire graph.
 - Multiple counting cancels ($3/3=1$)

Streaming updates to clustering coefficients

- ▶ Monitoring clustering coefficients could identify anomalies, find forming communities, *etc.*
- ▶ Computations stay **local**. A change to edge $\langle u, v \rangle$ affects only vertices u, v , and their neighbors.



- ▶ Need a fast method for updating the triangle counts, degrees when an edge is inserted or deleted.
 - Dynamic data structure for edges & degrees: STINGER
 - Rapid triangle count update algorithms: exact and **approximate**
- ▶ “Massive Streaming Data Analytics: A Case Study with Clustering Coefficients.” Ediger, David, Karl Jiang, E. Jason Riedy, and David A. Bader. MTAAP 2010, Atlanta, GA, April 2010.

Updating clustering coefficients

- ▶ Using RMAT as a graph and edge stream generator.
 - Mix of insertions and deletions
- ▶ Result summary for single actions
 - Exact: from 8 to 618 actions/second
 - Approx: from 11 to 640 actions/second
- ▶ Alternative: Batch changes
 - Lose some temporal resolution within the batch
 - Median rates for batches of size B:

Algorithm	B = 1	B = 1000	B = 4000
Exact	90	25 100	50 100
Approx.	60	83 700	193 300

- ▶ STINGER overhead is minimal; most time is spent in metric.



CHALLENGES

Hierarchy of Interesting Analytics

- ▶ **Extend single-shot graph queries to include time.**
 - Are there s - t paths between time T_1 and T_2 ?
 - What are the important vertices at time T ?
- ▶ **Use persistent queries to monitor properties.**
 - Does the path between s and t shorten drastically?
 - Is some vertex suddenly very central?
- ▶ **Extend persistent queries to fully dynamic properties.**
 - Does a small community stay independent rather than merge with larger groups?
 - When does a vertex jump between communities?
- ▶ **New types of queries, new challenges...**



Graph Analytics for Social Networks

- Are there new graph techniques? Do they parallelize? Can the computational systems (algorithms, machines) handle massive networks with millions to billions of individuals? Can the techniques tolerate noisy data, massive data, streaming data, etc. ...
- Communities may overlap, exhibit different properties and sizes, and be driven by different models
 - Detect communities (static or emerging)
 - Identify important individuals
 - Detect anomalous behavior
 - Given a community, find a representative member of the community
 - Given a set of individuals, find the best community that includes them



Suddenly, the flock became suspicious:
How come the newcomer wasn't shorn?



Open Questions for Massive Data Analytic Apps

- How do we **diagnose** the health of streaming systems?
- Are there **new analytics** for massive spatio-temporal interaction networks and graphs (STING)?
- Do current methods **scale up** from thousands to millions and billions?
- How do I **model** massive, streaming data streams?
- Are algorithms **resilient** to noisy data?
- How do I **visualize** a STING with $O(1M)$ entities? $O(1B)$? $O(100B)$? with scale-free power law distribution of vertex degrees and diameter = 6 ...
- Can **accelerators** aid in processing streaming graph data?
- How do we leverage the benefits of multiple architectures (e.g. **map-reduce clouds**, and **massively multithreaded architectures**) in a single platform?



10th DIMACS Implementation Challenge

- ▶ **Graph Partitioning and Graph Clustering** are ubiquitous subtasks in many application areas. Generally speaking, both techniques aim at the identification of vertex subsets with many internal and few external edges. To name only a few, problems addressed by graph partitioning and graph clustering algorithms are:
 - What are the communities within an (online) social network?
 - How do I speed up a numerical simulation by mapping it efficiently onto a parallel computer?
 - How must components be organized on a computer chip such that they can communicate efficiently with each other?
 - What are the segments of a digital image?
 - Which functions are certain genes (most likely) responsible for?
- ▶ **12-13 February 2012, Atlanta, Georgia**
 - Co-sponsored by DIMACS, by the Command, Control, and Interoperability Center for Advanced Data Analysis (CCICADA); Pacific Northwest National Laboratory; Sandia National Laboratories; and Deutsche Forschungsgemeinschaft (DFG).
 - **Paper deadline: 21 October 2011**
 - <http://www.cc.gatech.edu/dimacs10/>



Graph500 Benchmark, www.graph500.org

Defining a new set of benchmarks to guide the design of hardware architectures and software systems intended to support such applications and to help procurements. Graph algorithms are a core part of many analytics workloads.

Credit: Rich Murphy (Sandia), and Graph 500 committee



- Five Business Area Data Sets:

- Cybersecurity

- 15 Billion Log Entries/Day (for large enterprises)
- Full Data Scan with End-to-End Join Required

- Medical Informatics

- 50M patient records, 20-200 records/patient, billions of individuals
- Entity Resolution Important

- Social Networks

- Example, Facebook, Twitter
- Nearly Unbounded Dataset Size

- Data Enrichment

- Easily PB of data
- Example: Maritime Domain Awareness
 - Hundreds of Millions of Transponders
 - Tens of Thousands of Cargo Ships
 - Tens of Millions of Pieces of Bulk Cargo
 - May involve additional data (images, etc.)

- Symbolic Networks

- Example, the Human Brain
- 25B Neurons
- 7,000+ Connections/Neuron



Collaborators and Acknowledgments

- Jason Riedy, Research Scientist, (Georgia Tech)
- Graduate Students (Georgia Tech):
 - David Ediger
 - Karl Jiang
 - Pushkar Pande
 - Rob McColl
 - Anita Zakrzewska
- Bader PhD Graduates:
 - Seunghwa Kang (Pacific Northwest National Lab)
 - Kamesh Madduri (Penn State)
 - Guojing Cong (IBM TJ Watson Research Center)
- John Feo and Daniel Chavarría-Miranda (Pacific Northwest Nat'l Lab)



Bader, Related Recent Publications (2005-2008)

- D.A. Bader, G. Cong, and J. Feo, “**On the Architectural Requirements for Efficient Execution of Graph Algorithms,**” *The 34th International Conference on Parallel Processing (ICPP 2005)*, pp. 547-556, Georg Sverdrups House, University of Oslo, Norway, June 14-17, 2005.
- D.A. Bader and K. Madduri, “**Design and Implementation of the HPCS Graph Analysis Benchmark on Symmetric Multiprocessors,**” *The 12th International Conference on High Performance Computing (HiPC 2005)*, D.A. Bader et al., (eds.), Springer-Verlag LNCS 3769, 465-476, Goa, India, December 2005.
- D.A. Bader and K. Madduri, “**Designing Multithreaded Algorithms for Breadth-First Search and st-connectivity on the Cray MTA-2,**” *The 35th International Conference on Parallel Processing (ICPP 2006)*, Columbus, OH, August 14-18, 2006.
- D.A. Bader and K. Madduri, “**Parallel Algorithms for Evaluating Centrality Indices in Real-world Networks,**” *The 35th International Conference on Parallel Processing (ICPP 2006)*, Columbus, OH, August 14-18, 2006.
- K. Madduri, D.A. Bader, J.W. Berry, and J.R. Crobak, “**Parallel Shortest Path Algorithms for Solving Large-Scale Instances,**” *9th DIMACS Implementation Challenge – The Shortest Path Problem*, DIMACS Center, Rutgers University, Piscataway, NJ, November 13-14, 2006.
- K. Madduri, D.A. Bader, J.W. Berry, and J.R. Crobak, “**An Experimental Study of A Parallel Shortest Path Algorithm for Solving Large-Scale Graph Instances,**” *Workshop on Algorithm Engineering and Experiments (ALENEX)*, New Orleans, LA, January 6, 2007.
- J.R. Crobak, J.W. Berry, K. Madduri, and D.A. Bader, “**Advanced Shortest Path Algorithms on a Massively-Multithreaded Architecture,**” *First Workshop on Multithreaded Architectures and Applications (MTAAP)*, Long Beach, CA, March 30, 2007.
- D.A. Bader and K. Madduri, “**High-Performance Combinatorial Techniques for Analyzing Massive Dynamic Interaction Networks,**” *DIMACS Workshop on Computational Methods for Dynamic Interaction Networks*, DIMACS Center, Rutgers University, Piscataway, NJ, September 24-25, 2007.
- D.A. Bader, S. Kintali, K. Madduri, and M. Mihail, “**Approximating Betweenness Centrality,**” *The 5th Workshop on Algorithms and Models for the Web-Graph (WAW2007)*, San Diego, CA, December 11-12, 2007.
- David A. Bader, Kamesh Madduri, Guojing Cong, and John Feo, “**Design of Multithreaded Algorithms for Combinatorial Problems,**” in S. Rajasekaran and J. Reif, editors, *Handbook of Parallel Computing: Models, Algorithms, and Applications*, CRC Press, Chapter 31, 2007.
- Kamesh Madduri, David A. Bader, Jonathan W. Berry, Joseph R. Crobak, and Bruce A. Hendrickson, “**Multithreaded Algorithms for Processing Massive Graphs,**” in D.A. Bader, editor, *Petascale Computing: Algorithms and Applications*, Chapman & Hall / CRC Press, Chapter 12, 2007.
- D.A. Bader and K. Madduri, “**SNAP, Small-world Network Analysis and Partitioning: an open-source parallel graph framework for the exploration of large-scale networks,**” *22nd IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, Miami, FL, April 14-18, 2008.



Bader, Related Recent Publications (2009-2010)

- S. Kang, D.A. Bader, “**An Efficient Transactional Memory Algorithm for Computing Minimum Spanning Forest of Sparse Graphs,**” 14th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP), Raleigh, NC, February 2009.
- Karl Jiang, David Ediger, and David A. Bader. “**Generalizing k-Betweenness Centrality Using Short Paths and a Parallel Multithreaded Implementation.**” The 38th International Conference on Parallel Processing (ICPP), Vienna, Austria, September 2009.
- Kamesh Madduri, David Ediger, Karl Jiang, David A. Bader, Daniel Chavarría-Miranda. “**A Faster Parallel Algorithm and Efficient Multithreaded Implementations for Evaluating Betweenness Centrality on Massive Datasets.**” 3rd Workshop on Multithreaded Architectures and Applications (MTAAP), Rome, Italy, May 2009.
- David A. Bader, et al. “**STINGER: Spatio-Temporal Interaction Networks and Graphs (STING) Extensible Representation.**” 2009.
- David Ediger, Karl Jiang, E. Jason Riedy, and David A. Bader. “**Massive Streaming Data Analytics: A Case Study with Clustering Coefficients,**” Fourth Workshop in Multithreaded Architectures and Applications (MTAAP), Atlanta, GA, April 2010.
- Seunghwa Kang, David A. Bader. “**Large Scale Complex Network Analysis using the Hybrid Combination of a MapReduce cluster and a Highly Multithreaded System.,**” Fourth Workshop in Multithreaded Architectures and Applications (MTAAP), Atlanta, GA, April 2010.
- David Ediger, Karl Jiang, Jason Riedy, David A. Bader, Courtney Corley, Rob Farber and William N. Reynolds. “**Massive Social Network Analysis: Mining Twitter for Social Good,**” The 39th International Conference on Parallel Processing (ICPP 2010), San Diego, CA, September 2010.
- Virat Agarwal, Fabrizio Petrini, Davide Pasetto and David A. Bader. “**Scalable Graph Exploration on Multicore Processors,**” *The 22nd IEEE and ACM Supercomputing Conference (SC10)*, New Orleans, LA, November 2010.



Acknowledgment of Support

