Massive-Scale Streaming Analytics

David A. Bader

Georgia Tech College of Computing
Computational Science and Engineering

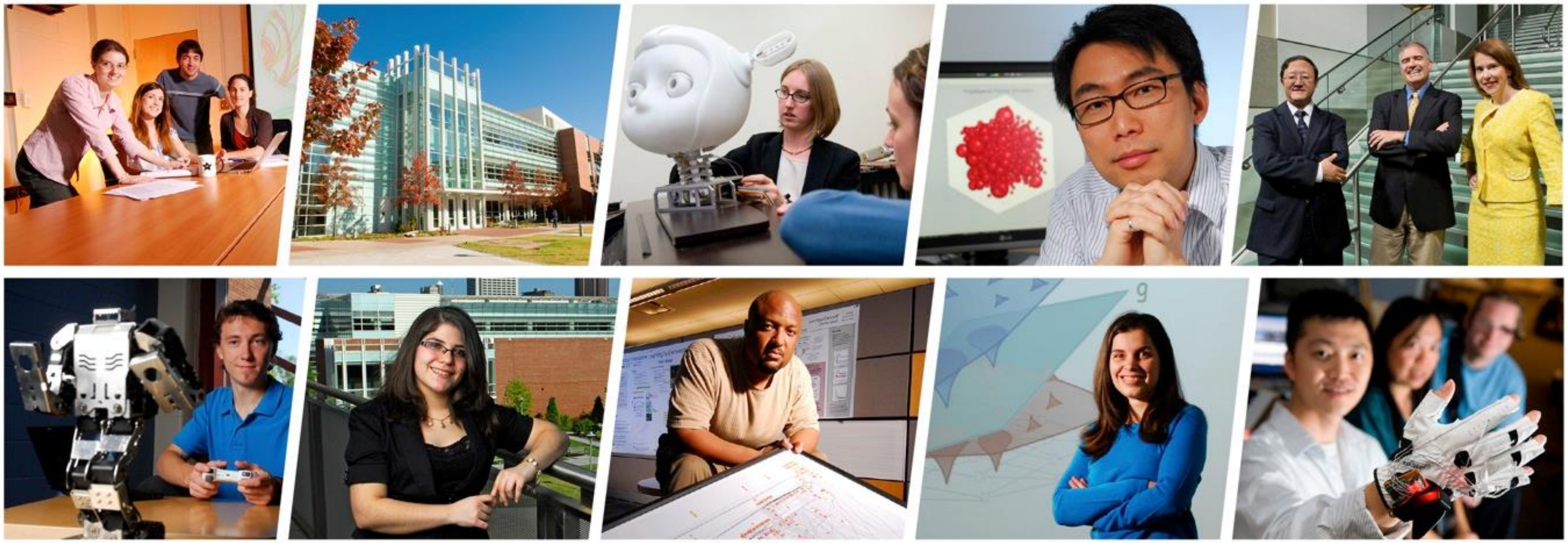
Dr. David A. Bader

- Full Professor, Computational Science and Engineering
- Executive Director for High Performance Computing.
- IEEE Fellow, AAAS Fellow
- interests are at the intersection of high-performance computing and real-world applications, including computational biology and genomics and massive-scale data analytics.
- Over $165M of research awards
- Steering Committees of the major HPC conferences, IPDPS and HiPC
- Multiple editorial boards in parallel and high performance computing
  - EIC of IEEE Transactions on Parallel and Distributed Systems
- Elected chair of IEEE and SIAM committees on HPC
- 230+ publications, ≥ 4,700 citations, *h*-index ≥ 38
- National Science Foundation CAREER Award recipient
- Directed the Sony-Toshiba-IBM Center for the Cell/B.E. Processor
- Founder of the Graph500 List for benchmarking “Big Data” computing platforms
- Recognized as a “**RockStar**” of High Performance Computing by InsideHPC in 2012 and as HPCwire's **People to Watch** in 2012 and 2014.

Outline

- Overview of Georgia Tech
- STINGER: Streaming Analytics
- Case study: Seed Set Expansion
- Future architectures
- Conclusions

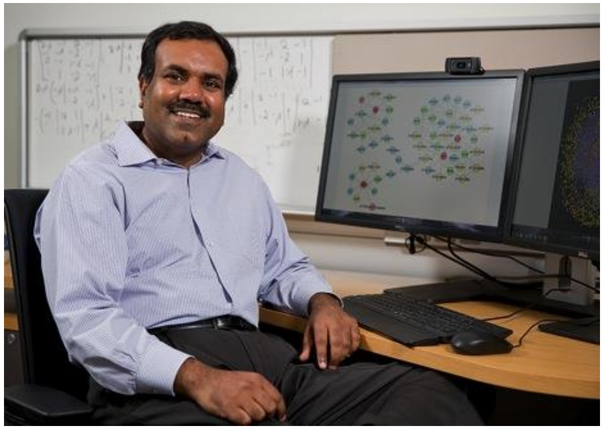
Georgia Tech | School of Computational Science and Engineering



THE CSE INNOVATION ECOSYSTEM:

CREATING SOLUTIONS AND LEADERS

Innovate. Collaborate. Problem Solved.



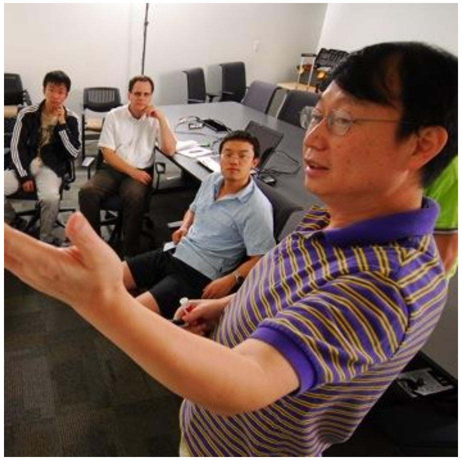
CSE is a diverse, interdisciplinary innovation ecosystem composed of award-winning faculty, researchers and students that

- Solves real-world problems and creates future leaders
- Enables breakthroughs in scientific discovery and engineering practice
- Uses the most advanced resources, techniques and ideas
- Is highly collaborative with an impressive roster of GT and industry partners

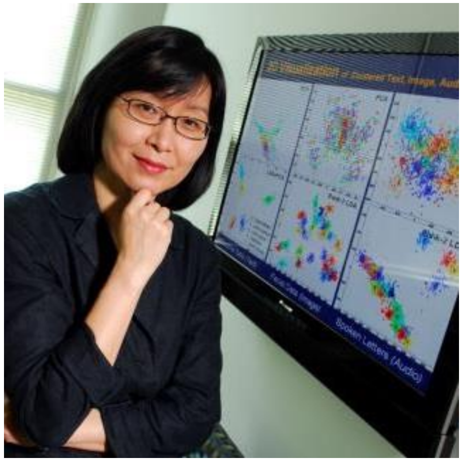
Ten Years of Success

- Founded: 2005
- Chair: David Bader
- Faculty:
  - 11 tenure track (FY 16)
  - 4 joint appointments
  - 6 adjunct faculty
  - 5 research scientists
- Administrative staff: 5
- Research expenditures: $5.6 million (FY 2015)
- High impact: $463K expenditure per faculty member

Award-Winning Faculty



- 11 tenure-track faculty members (FY 16)
- 1 Regents' professor
- 5 NSF CAREER awards
- 2 IEEE fellows, 2 AAAS fellows, and 1 SIAM fellow
- 3 recent best paper awards and 2 finalists from SIAM, IEEE, etc.
- Several recent awards from industry:



Accenture
IBM
Google
NVIDIA
Intel
Lockheed Martin
Yahoo! Labs

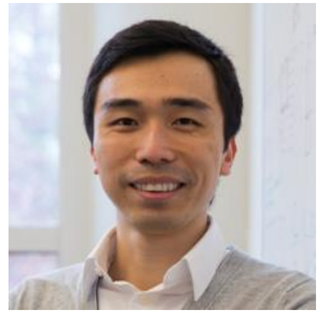
Raytheon
LexisNexis
Microsoft Research
Sony
Cray
Exxon Mobil

Faculty: Interdisciplinary Innovators

Georgia Tech School of Computational Science and Engineering

Srinivas Aluru
Professor

David Bader
Professor and Chair

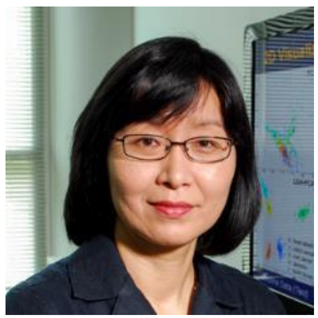


Polo Chau
Assistant Professor

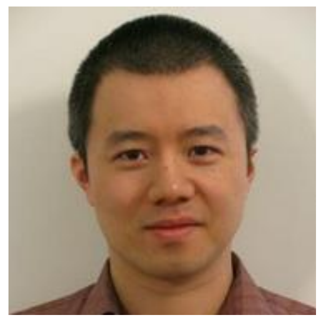
Edmond Chow
Associate Professor

Bistra Dilkina
Assistant Professor

Richard Fujimoto
Regents' Professor

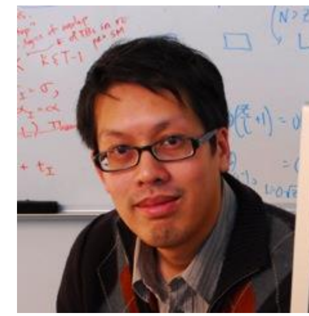


Haesun Park
Professor

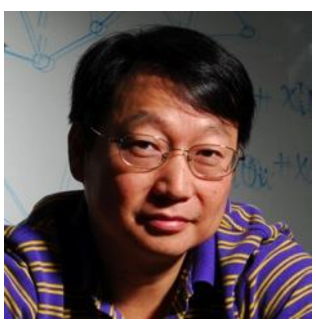


Le Song
Assistant Professor

Jimeng Sun
Associate Professor



Richard Vuduc
Associate Professor



Hongyuan Zha
Professor

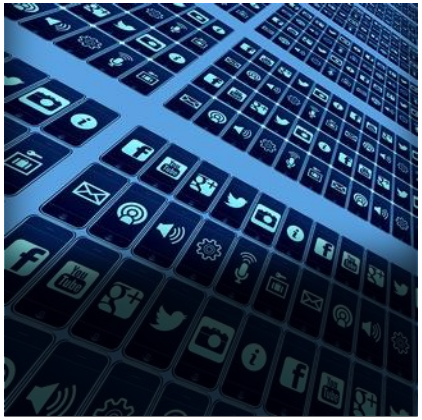
Kenneth Brown
Chemistry

Mark Borodovsky
BME

David Sherrill
Chemistry

Surya Kalidindi
Mech. Engr.

12 Pinnacle Projects > US$1M



S. Aluru (PI), W. Feng, K. Olukotun, P. Schnable, C. Sing, and J. Zola, “BIGDATA: Mid-Scale: DA: Collaborative Research: Genomes Galore - Core Techniques, Libraries, and Domain Specific Languages for High-Throughput DNA Sequencing,” NSF/NIH Bigdata Initiative, **$2M**

A. Somani, **S. Aluru** (Co-PI), R. Fox, E. Takle, and M. Gordon, “MRI: Acquisition of a HPC system for Data-Driven Discovery in Science and Engineering,” National Science Foundation, **$1.8M**

S. Aluru (PI), K. Dorman, and P.S. Schnable, “AF:Medium: Parallel Algorithms and Software for High-throughput Sequence Assembly,” National Science Foundation, **$1M**

Polo Chau (Co-PI), “Center of Excellence for Mobile Sensor Data-to-Knowledge (MD2K),” National Institute of Health, **$1.25M**

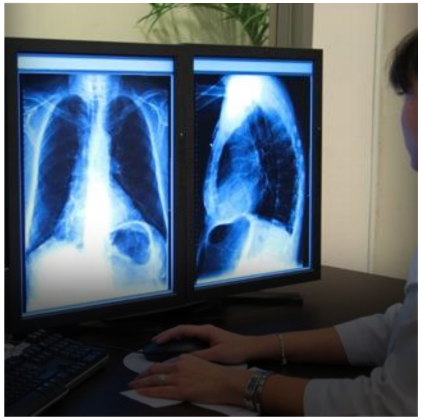
R. Fujimoto (Co-PI) and J. Crittenden (PI), “Participatory Modeling of Complex Urban Infrastructure Systems,” National Science Foundation, **$2.5M**

R. Fujimoto (PI), T. Blum, **S. Kalidindi**, W. Newstetter, and **H. Zha**, “Computation-Enabled Design and Manufacturing of High Performance Materials,” National Science Foundation, **$2.8M**

H. Park (PI), **H. Zha** (Co-PI), B. Drake (Co-PI), J. Choo (Co-PI), and J. Poulson (Co-PI), “Fast Algorithms on Imperfect, Heterogeneous, Distributed Data for Interactive Analysis,” DARPA, **$2.7M**

H. Park (PI), J. Stasko (Co-PI), A. Gray (Co-PI), J. Monteiro (Co-PI), V. Koltchinskii (Co-PI), “FODAVA-lead: Dimension Reduction and Data Reduction: Foundations for Visualization,” National Science Foundation and Department of Homeland Security, **$3.5M**

D. Bader (PI), **E.J. Riedy** (Co-PI), **R. Vuduc** (Co-PI), and V. Prasanna (PI), “SI2-SSI: Collaborative: The XScala Project: A Community Repository for Model-Driven Design and Tuning of Data-Intensive Applications for Extreme-Scale Accelerator-Based Systems,” National Science Foundation, **$1.2M**



D. Bader (PI), **E.J. Riedy** (Co-PI), "GRATEFUL: GRaph Analysis Tackling power EFficiency, Uncertainty, and Locality, Power Efficiency Revolution for Embedded Computing Technologies (PERFECT) Program,” DARPA, **$2.9M**

J. Sun, Smart Connect Health Project Award, National Science Foundation, **$2.1M**

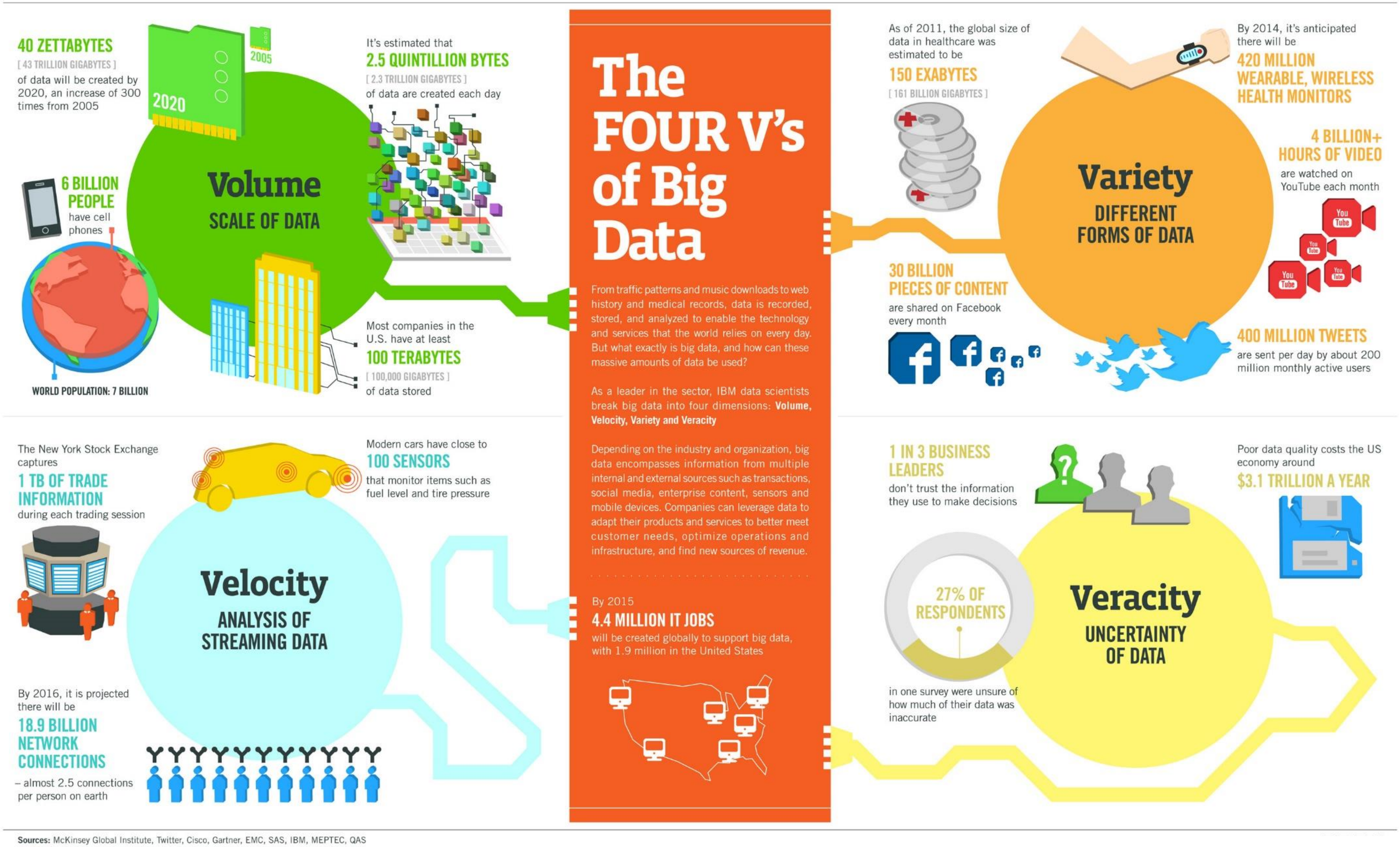
H. Zha (Co-PI), “TWC SBE: TTP Option: Medium: Collaborative: EPICA: Empowering People to Overcome Information Controls and Attacks,” National Science Foundation, **$1.1M**

…and more good news pending…

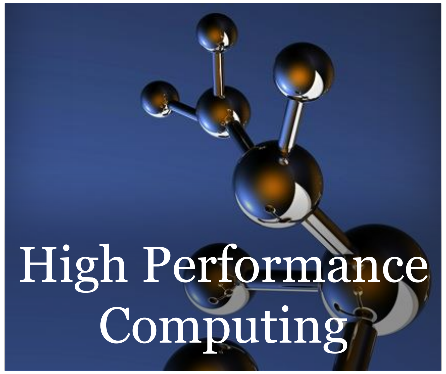
Big Data Analytics

Georgia Tech | School of Computational Science and Engineering

Answering the need for algorithms that scale to massive, complex data sets

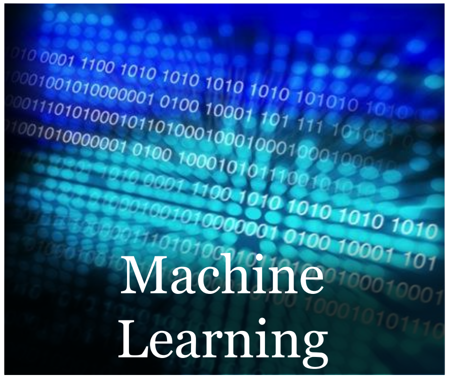


Core Research Areas



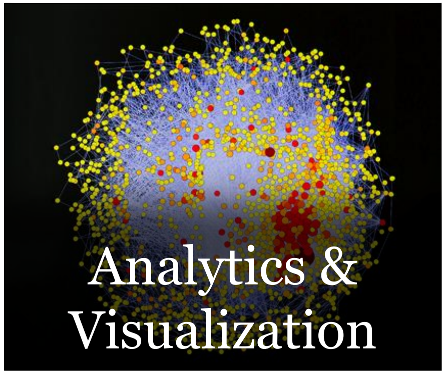
Devise computing solutions at the absolute limits of scale and speed using efficient, reliable and fast algorithms, software, tools and applications

Construct and study algorithms that build models, and make efficient data-driven predictions or decisions



Develop new methods to analyze large and complex data sets, transforming data into value and solve grand challenges

Design fast theoretic algorithms on large-scale graphs, and detect malicious activity



Present data in ways that best yield insight and support decisions as problems scale and complexity increase

Graduate Education

- **Ph.D. and MS in Computational Science and Engineering**
- Ph.D. and MS in Bioengineering, Ph.D. in Bioinformatics, MS in Analytics

Strength in Diversity: CSE Home Units

School of Aerospace Engineering
School of Biology
Coulter Department of Biomedical Engineering
School of Chemistry and Biochemistry
School of Civil and Environmental Engineering
School of Computational Science and Engineering
School of Industrial and Systems Engineering
School of Mathematics

Students select a **Home** – unit & (*if applicable*) advisor
Coursework – **Core** + **Computation** + **Application**
Research – **Dissertation**
(*MS thesis option* + *PhD only*)

Georgia Tech School of Computational Science and Engineering

2005 - 2015

Outline

- Overview of Georgia Tech
- STINGER: Streaming Analytics
- Case study: Seed Set Expansion
- Future architectures
- Conclusions

STING Initiative:
Focusing on Globally Significant Grand Challenges

- Many globally-significant grand challenges can be modeled by **Spatio-Temporal Interaction Networks and Graphs** (or “STING”).
- Emerging real-world graph problems include
  - detecting community structure in large social networks,
  - defending the nation against cyber-based attacks,
  - discovering insider threats (e.g. Ft. Hood shooter, WikiLeaks),
  - improving the resilience of the electric power grid, and
  - detecting and preventing disease in human populations.

- Unlike traditional applications in computational science and engineering, solving these problems at scale often raises new research challenges because of sparsity and the lack of locality in the massive data, design of parallel algorithms for massive, streaming data analytics, and the need for new exascale supercomputers that are energy-efficient, resilient, and easy-to-program.

Big Data problems need Graph Analysis

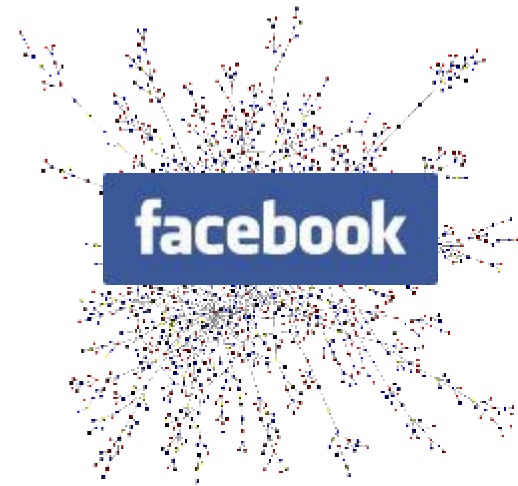
Health Care	• Finding outbreaks, population epidemiology
Social Networks	• Advertising, searching, grouping, influence
Intelligence	• Decisions at scale, regulating algorithms
Systems Biology	• Understanding interactions, drug design
Power Grid	• Disruptions, conversion
Simulation	• Discrete events, cracking meshes

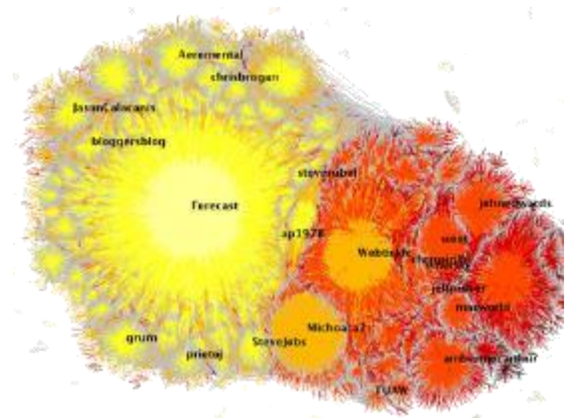
Graphs are a unifying motif for data analysis.

Changing and *dynamic* graphs are important!

Data rates and volumes are immense

- Facebook:
  - ~1 billion users
  - average 130 friends
  - 30 billion pieces of content shared / month
- Twitter:
  - 500 million active users
  - 340 million tweets / day
- Internet – 100s of exabytes / year
  - 300 million new websites per year
  - 48 hours of video to You Tube per minute
  - 30,000 YouTube videos played per second





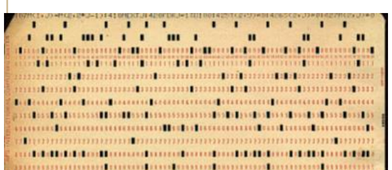
Massive-Scale Streaming Analytics

Historically, HPC uses **batch processing** style where a program and a static data set are scheduled to compute in the next available slot.

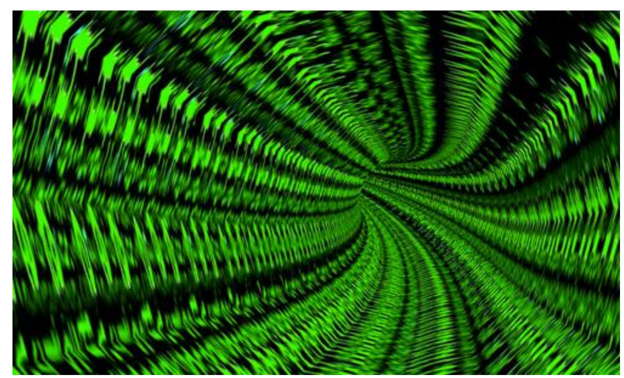
Today, data is overwhelming in volume *and rate*, and we struggle to keep up with these **streams**.

➔ **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.

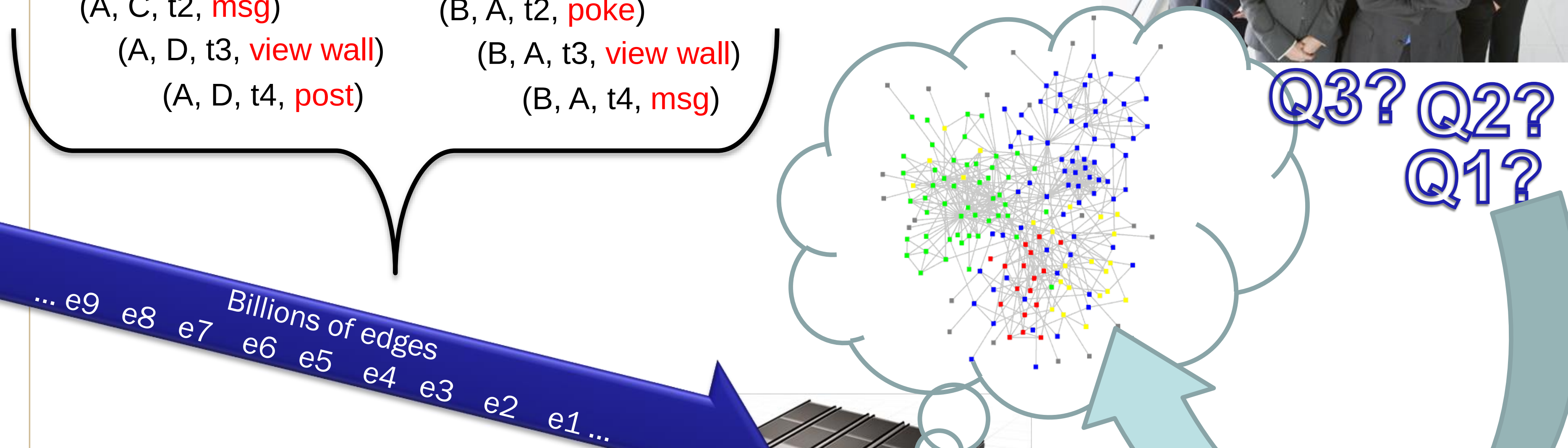


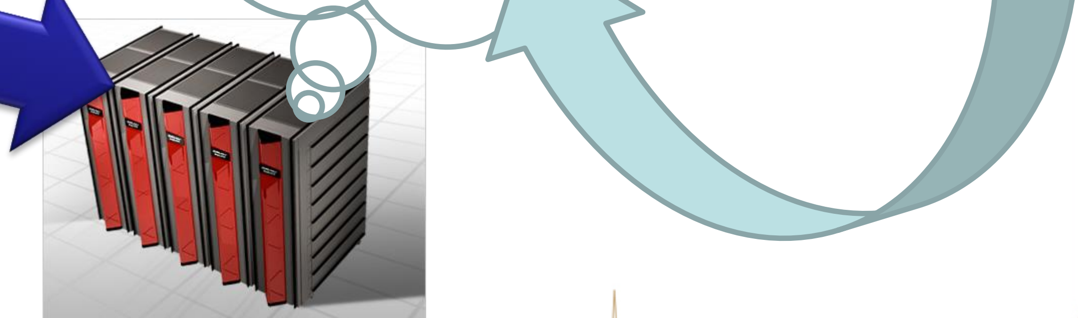
Our focus is streaming graphs

(A, B, t1, poke)
(A, C, t2, msg)
(A, D, t3, view wall)
(A, D, t4, post)

(B, A, t2, poke)
(B, A, t3, view wall)
(B, A, t4, msg)

Analysts





- Change detection
- Flows, Clustering, Centrality
- Structural change
- Key actors, anomalies

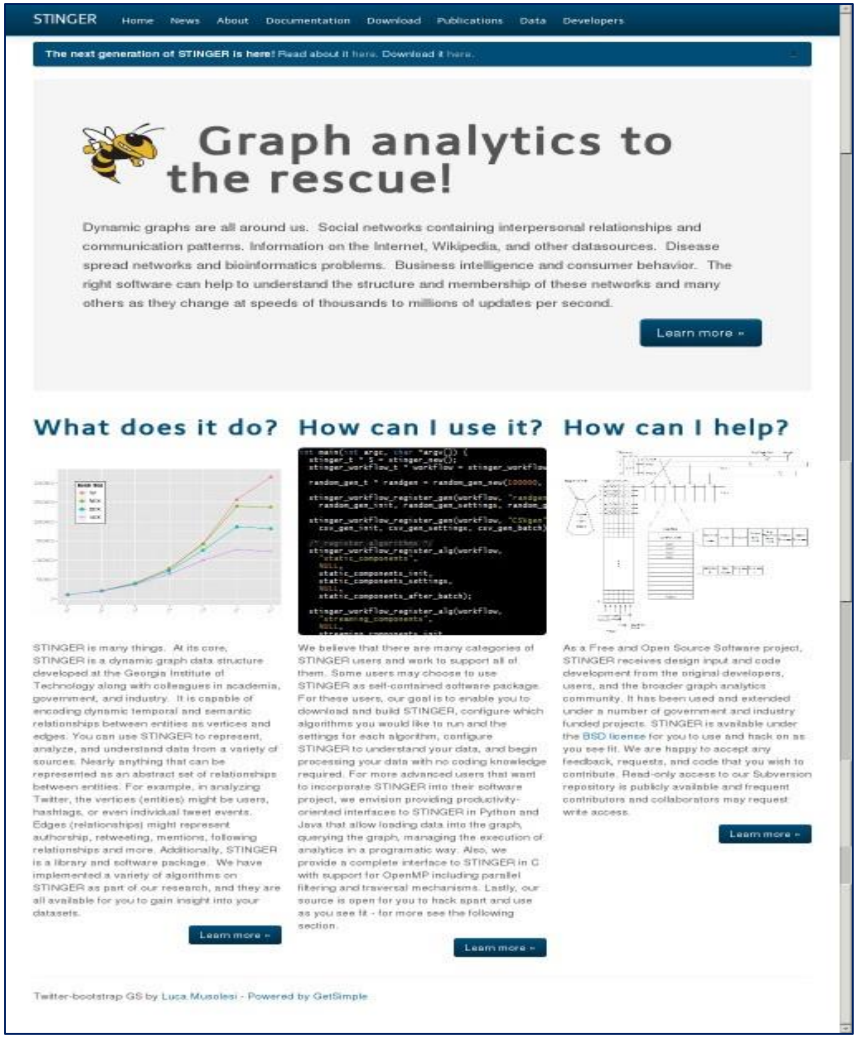
STINGER: as an analysis package

- **Streaming edge insertions and deletions:**
  Performs new edge insertions, updates, and deletions in batches or individually.
- **Streaming clustering coefficients:**
  Tracks the local and global clustering coefficients of a graph under both edge insertions and deletions.
- **Streaming connected components:**
  Accurately tracks the connected components of a graph with insertions and deletions.
- **Streaming community detection:**
  Track and update the community structures within the graph as they change.
- **Parallel agglomerative clustering:**
  Find clusters that are optimized for a user-defined edge scoring function.
- **Streaming Betweenness Centrality:**
  Find the key points within information flows and structural vulnerabilities.
- **K-core Extraction:**
  Extract additional communities and filter noisy high-degree vertices.
- **Classic breadth-first search:**
  Performs a parallel breadth-first search of the graph starting at a given source vertex to find shortest paths.

Optimized to update at rates of over 3 million edges per second on graphs of one billion edges

http://www.cc.gatech.edu/stinger

STINGER: Where do you get it?



www.cc.gatech.edu/stinger/

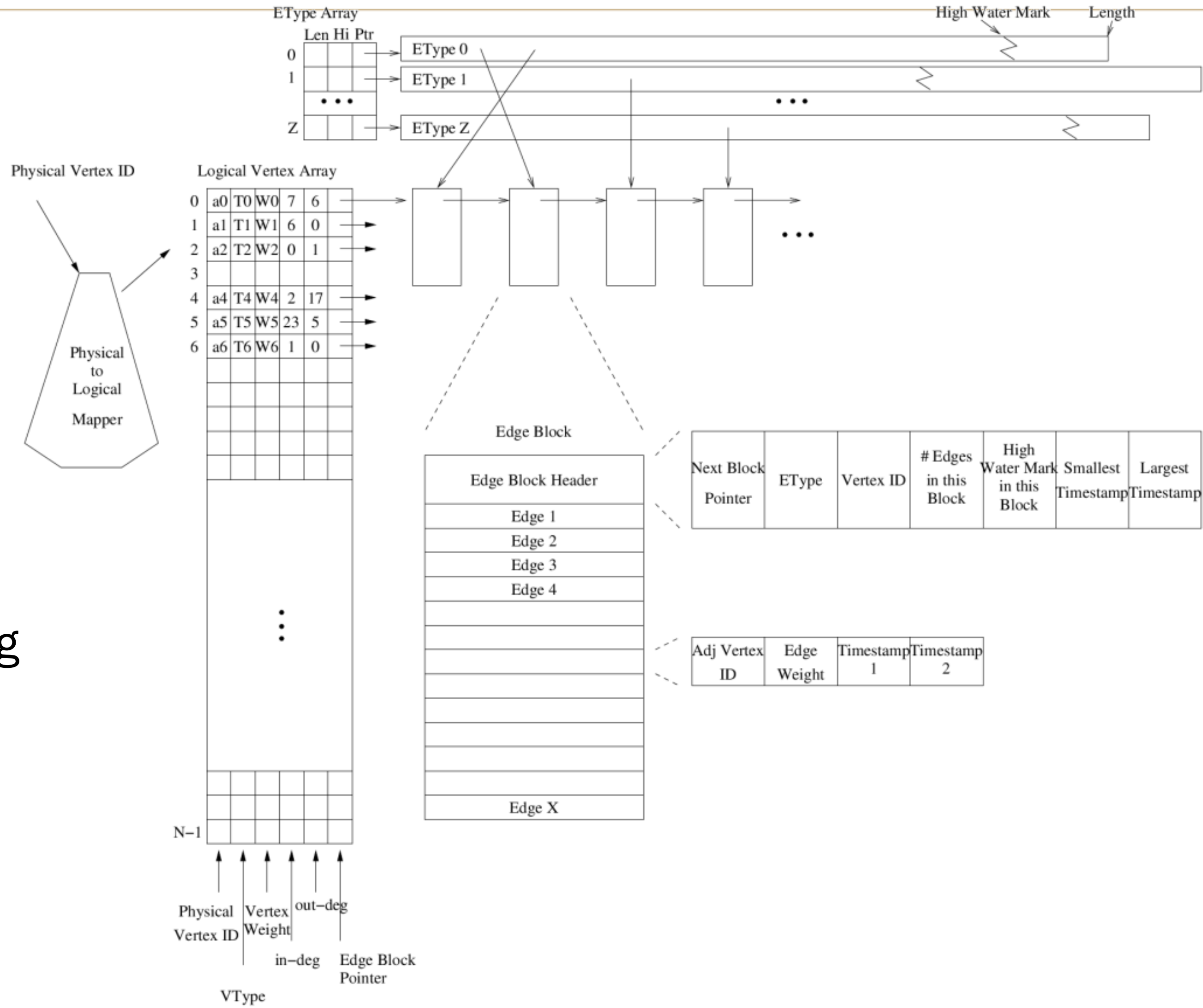
- Gateway to
  - code,
  - development,
  - documentation,
  - presentations…
- Users / contributors / questioners: Georgia Tech, PNNL, CMU, Berkeley, Intel, Cray, NVIDIA, IBM, Federal Government, Ionic Security, Citi

STING Extensible Representation (STINGER)

- Enhanced representation developed for dynamic graphs developed in consultation with David A. Bader, Jon 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.

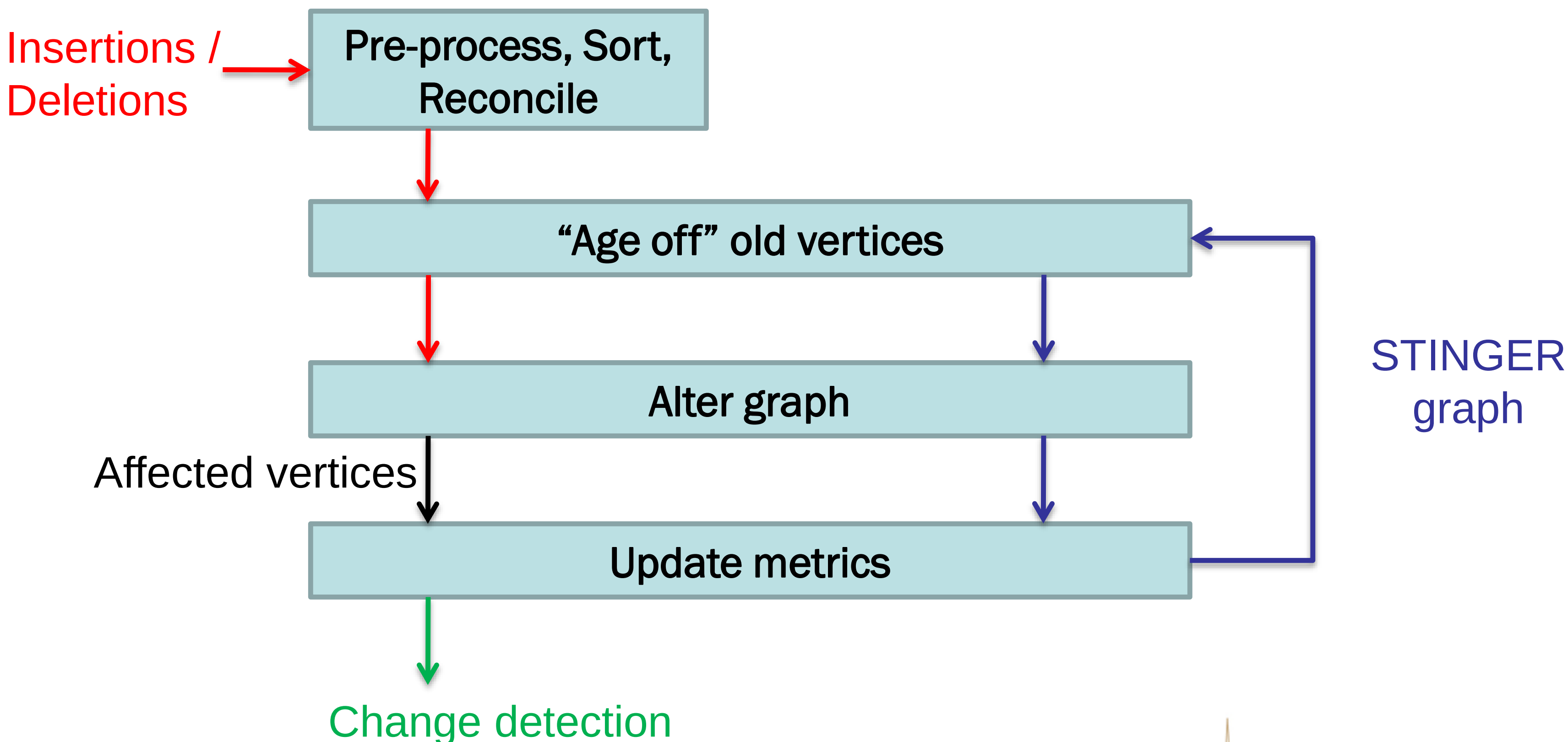
STING Extensible Representation

- 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

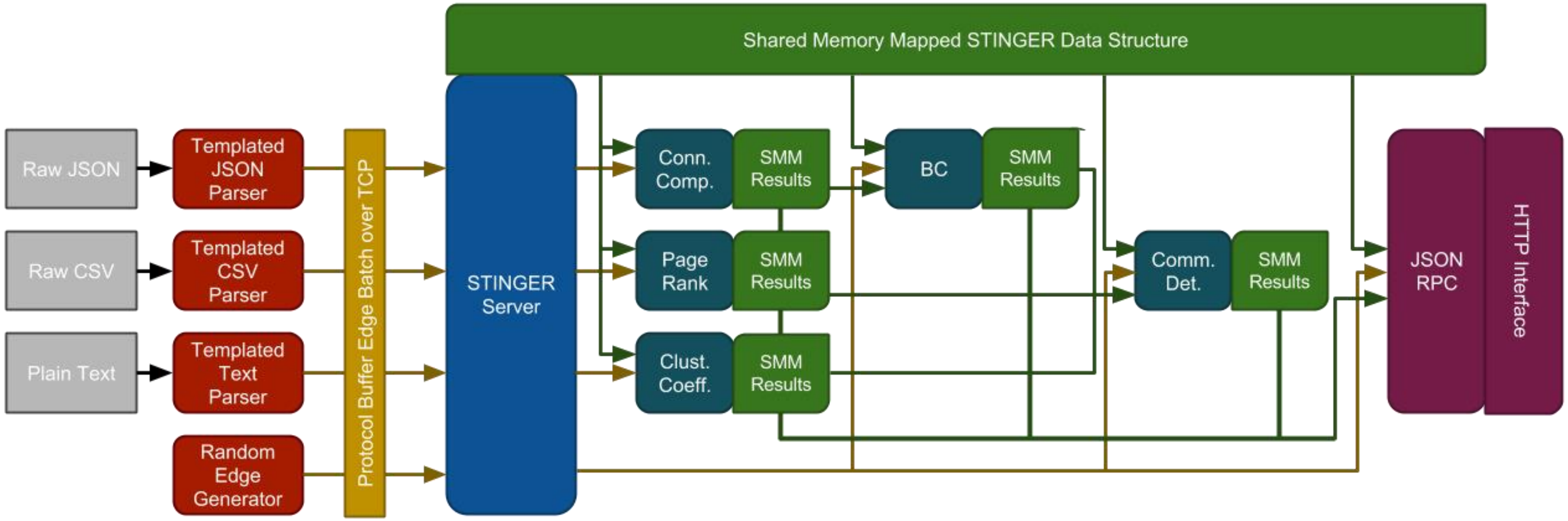


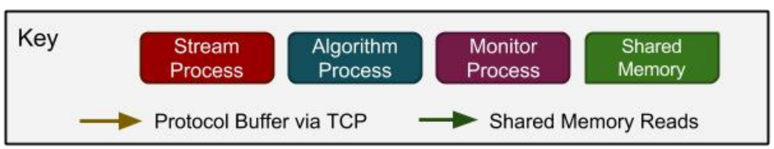
Massive Streaming Data Analytics

- Accumulate as much of the recent graph data as possible in main memory.



STING: High-level architecture





Data flows from left to right. All stages aligned vertically execute on the data in parallel. All processes in direct contact with shared memory have write access to that memory. Execution is synchronized by the server. Streams, algorithms, and monitors can join or leave the workflow at any time.

- Server: Graph storage, kernel orchestration
- OpenMP + sufficiently POSIX-ish
- Multiple processes for resilience

STINGER Streaming Graph Results

- Triangle counting / clustering coefficients
  - Up to 130k graph updates per second on X5570 (Nehalem-EP, 2.93GHz)
- Connected components & spanning forest
  - Over 88k graph updates per second on X5570
- Community detection & maintenance
  - Up to 100 million updates per second, 4-socket 40-core Westmere-EX
  - (Note: Most updates do not change communities…)
- Incremental PageRank
  - Reduce lower latency by > 2× over restarting
- Betweenness centrality
  - $O(|V| \cdot (|V| + |E|))$, can be sampled
  - Speed-ups of 40×–150× over static recomputation

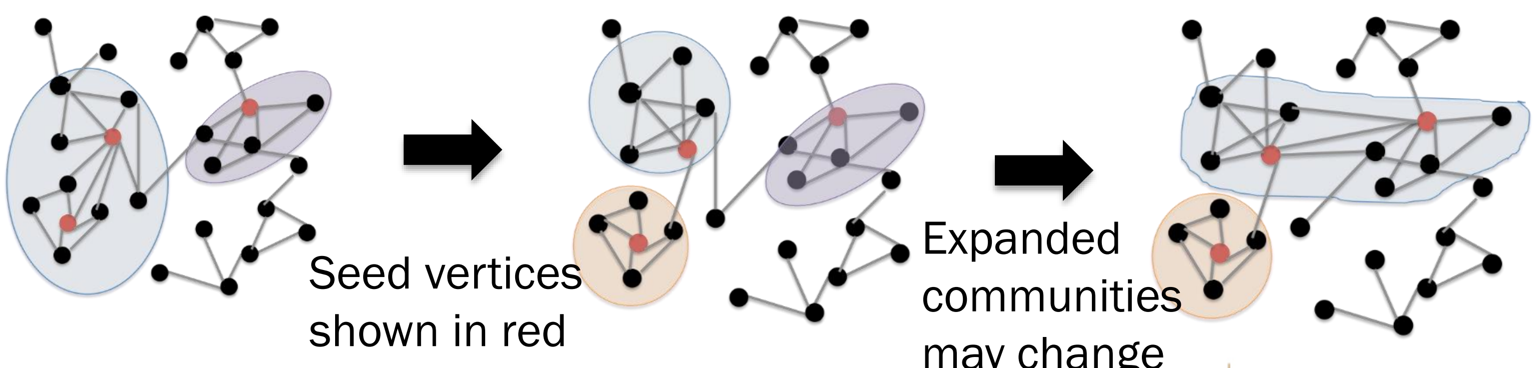
Outline

- Overview of Georgia Tech
- STINGER: Streaming Analytics
- Case study: Seed Set Expansion
- Future architectures
- Conclusions

Streaming Seed Set Expansion

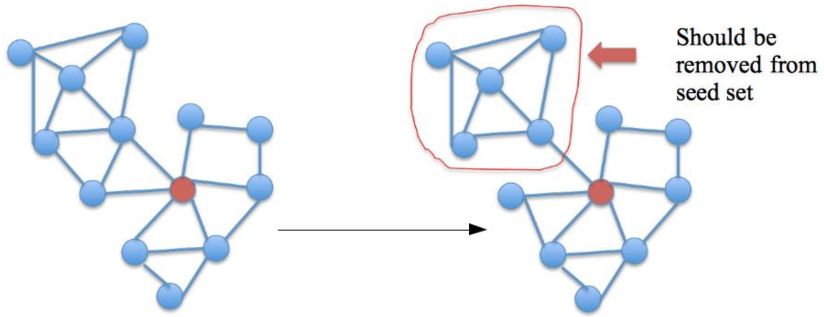
[Joint research with Anita Zakrzewska]

- Given a set of seed vertices of interest, seed set expansion finds the best subgraph or set of communities containing the seeds
- The communities found can be used to identify and track groups interacting entities
- The results also aid visualization or performing more computationally expensive analytics
- In a dynamic graph, the optimal communities may change as the graph evolves. Incremental updates can be faster than recomputing from scratch
- We can track changes over time and detect when interesting changes occur, such as the expanded communities merging or splitting



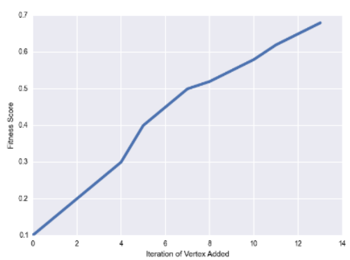
Streaming Seed Set Expansion

- To track seed vertices of interest over time, we run a seed set expansion from each seed and maintain the communities over time
- The overlap of expanded sets is used to determine when the communities merge together or split apart. This can alert us to interesting changes and events
- Each individual expansion is incrementally updated as the graph changes
  - New information may cause a community to expand or reduce in size.

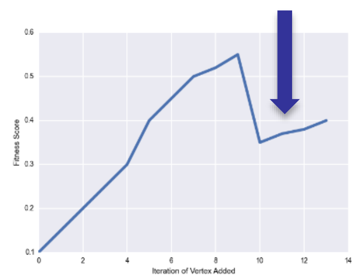


Streaming Seed Set Expansion

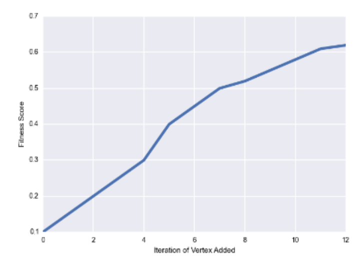
- Since seed set expansion retrieves the community of a seed vertex, the order of member selection is important. Simply re- testing vertices for membership may not detect the change in the community.
- Greedy seed set expansion results in a monotonically increasing sequence of fitness values as vertices are iteratively added to the community.
- We perform streaming greedy seed set expansion by updating the fitness sequence and detecting drops in value. After a drop in fitness, the community is be selectively pruned and the fitness scores remain increasing .



Initial sequence of fitness scores



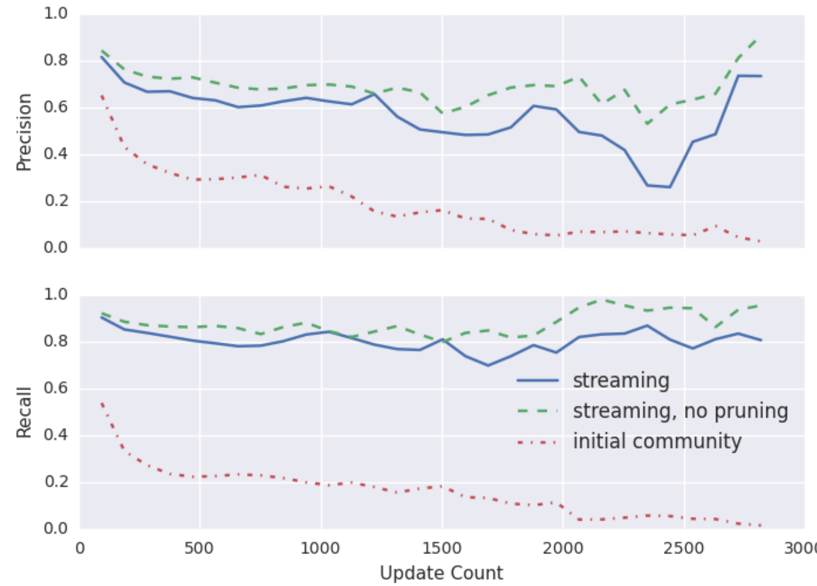
After a change in the graph

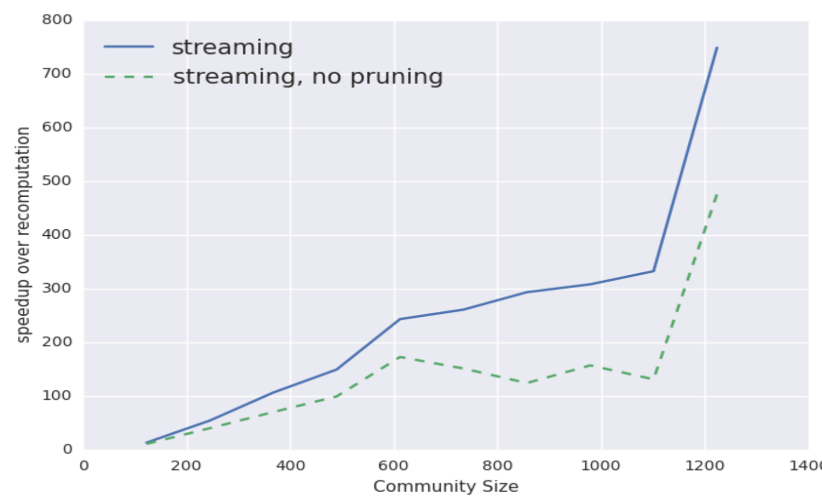


After updating the community by pruning

Streaming Seed Set Expansion

- The quality of communities produced with our incremental updating approach, compared to recomputing, is measured by precision and recall.
- Our method maintains high precision and recall as the graph changes. The red dotted line gives a baseline, showing that the communities do change significantly over time.
- Our incremental updating provides a speedup of two orders of magnitude over standard complete recomputation, depending on community size.





Outline

- Overview of Georgia Tech
- STINGER: Streaming Analytics
- Case study: Seed Set Expansion
- Future architectures
- Conclusions

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.

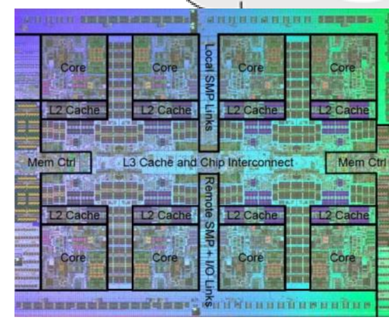
Executive Committee: D.A. Bader, R. Murphy, M. Snir, A. Lumsdaine

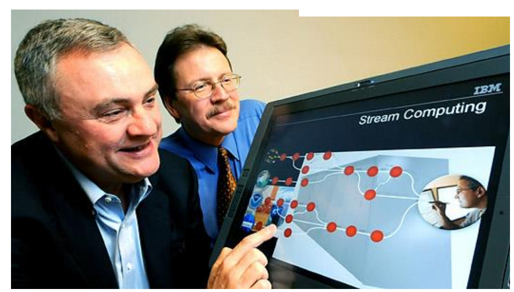
- Five Business Area Data Sets:
  - Cybersecurity
    - 15 Billion Log Entires/Day (for large enterprises)
    - Full Data Scan with End-to-End Join Required
  - Medical Informatics
    - 50M patient records, 20-200 records/patient, billions of individuals
    - Entity Resolution Important
  - Social 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

Heterogeneity in "Big Data" systems: High Performance Data Analytics

- Analytic platforms will combine:
  - Cloud (Hadoop/map-reduce)
  - Stream processing
  - Large shared-memory systems
  - Massive multithreaded architectures
  - Multicore and accelerators

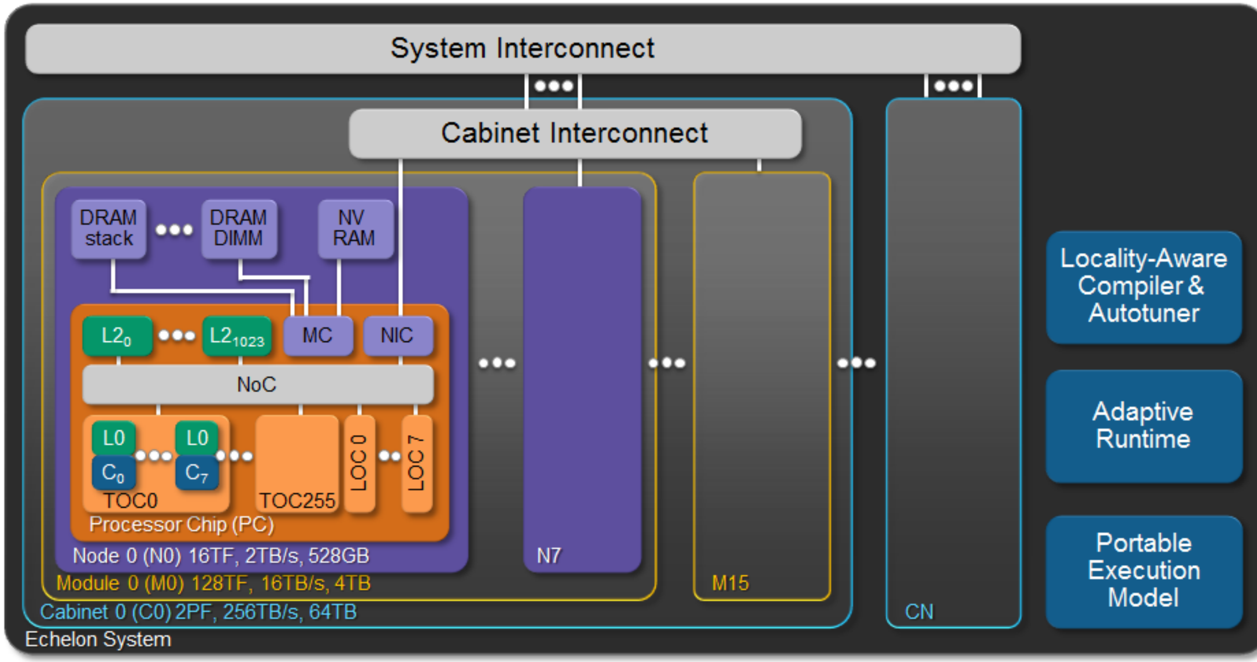
➔ **The challenge**: developing methodologies for employing these complementary systems in an enterprise-class analytics framework for solving grand challenges in massive data analysis for discovery, real-time analytics, and forensics.





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

Future Architectures



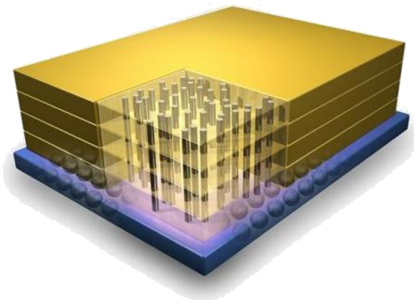
- Highly multithreaded
- High bandwidth (network and memory)
- Complex but **flexible** memory hierarchy
- Heterogeneous design in core capability and ISA

Disruptive Platform Changes

- In next 1–2 years, memory is going to change
  - 3D stacked memory (IBM, NVIDIA)
  - Hybrid memory cube (HMC Cons., Micron, Intel)
  - Programming logic layer on-chip
  - Possibly non-volatile
  - Order of magnitude higher bandwidth
  - Order of magnitude lower energy cost
- Interconnects are changing
  - Processor ⇔ memory ⇔ accelerator (NVLink, Phi)
  - Data-center networks finally may change, not just *n*GbE

Revolutionary Changes in Memory Technology

- In the next 1-2 years, memory technologies will undergo a dramatic change
  - 3-D stacked memory
  - Programming logic layer on chip
  - E.g. Hybrid Memory Cube (HMC) consortium



- Energy is a major constraint in both embedded (smartphone) and supercomputing systems, and 75% of the energy is spent moving data
- Logic layer allows operations to be performed on-memory chips without needing the round-trip to move the data from memory to processor
- This is a huge change in the relative cost of operations!

Outline

- Overview of Georgia Tech
- STINGER: Streaming Analytics
- Case study: Seed Set Expansion
- Future architectures
- Conclusions

Conclusions

- Massive-Scale Streaming Analytics will require new
  - High-performance computing platforms
  - Streaming algorithms
  - Energy-efficient implementations

  and are promising to solve **real-world challenges**!
- Mapping applications to high performance architectures may yield 6 or more orders of magnitude performance improvement

Acknowledgments

- Jason Riedy, Research Scientist, (Georgia Tech)
- Graduate Students (Georgia Tech):
  - James Fairbanks
  - Rob McColl
  - Eisha Nathan
  - Anita Zakrzewska
- Bader Alumni:
  - Dr. David Ediger (GTRI)
  - Dr. Oded Green (ArrayFire)
  - Dr. Seunghwa Kang (Pacific Northwest National Lab)
  - Prof. Kamesh Madduri (Penn State)
  - Dr. Guojing Cong (IBM TJ Watson Research Center)

Bader, Related Recent Publications (2005-2009)

- 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 Betewenness 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.
- 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.

Bader, Related Recent Publications (2010-2011)

- 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.
- Z. Du, Z. Yin, W. Liu, and D.A. Bader, "**On Accelerating Iterative Algorithms with CUDA: A Case Study on Conditional Random Fields Training Algorithm for Biological Sequence Alignment**," IEEE International Conference on Bioinformatics & Biomedicine, Workshop on Data-Mining of Next Generation Sequencing Data (NGS2010), Hong Kong, December 20, 2010.
- D. Ediger, J. Riedy, H. Meyerhenke, and D.A. Bader, "**Tracking Structure of Streaming Social Networks**," 5th Workshop on Multithreaded Architectures and Applications (MTAAP), Anchorage, AK, May 20, 2011.
- D. Mizell, D.A. Bader, E.L. Goodman, and D.J. Haglin, "**Semantic Databases and Supercomputers**," 2011 Semantic Technology Conference (SemTech), San Francisco, CA, June 5-9, 2011.
- P. Pande and D.A. Bader, "**Computing Betweenness Centrality for Small World Networks on a GPU**," *The 15th Annual High Performance Embedded Computing Workshop* (HPEC), Lexington, MA, September 21-22, 2011.
- David A. Bader, Christine Heitsch, and Kamesh Madduri, "**Large-Scale Network Analysis**," in J. Kepner and J. Gilbert, editor, *Graph Algorithms in the Language of Linear Algebra*, SIAM Press, Chapter 12, pages 253-285, 2011.
- Jeremy Kepner, David A. Bader, Robert Bond, Nadya Bliss, Christos Faloutsos, Bruce Hendrickson, John Gilbert, and Eric Robinson, "**Fundamental Questions in the Analysis of Large Graphs**," in J. Kepner and J. Gilbert, editor, *Graph Algorithms in the Language of Linear Algebra*, SIAM Press, Chapter 16, pages 353-357, 2011.

Bader, Related Recent Publications (2012)

- E.J. Riedy, H. Meyerhenke, D. Ediger, and D.A. Bader, "**Parallel Community Detection for Massive Graphs**," The 9th International Conference on Parallel Processing and Applied Mathematics (PPAM 2011), Torun, Poland, September 11-14, 2011. Lecture Notes in Computer Science, 7203:286-296, 2012.
- E.J. Riedy, D. Ediger, D.A. Bader, and H. Meyerhenke, "**Parallel Community Detection for Massive Graphs**," 10th DIMACS Implementation Challenge -- Graph Partitioning and Graph Clustering, Atlanta, GA, February 13-14, 2012.
- E.J. Riedy, H. Meyerhenke, D.A. Bader, D. Ediger, and T. Mattson, "**Analysis of Streaming Social Networks and Graphs on Multicore Architectures**," The 37th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Kyoto, Japan, March 25-30, 2012.
- J. Riedy, H. Meyerhenke, and D.A. Bader, "**Scalable Multi-threaded Community Detection in Social Networks**," 6th Workshop on Multithreaded Architectures and Applications (MTAAP), Shanghai, China, May 25, 2012.
- H. Meyerhenke, E.J. Riedy, and D.A. Bader, "**Parallel Community Detection in Streaming Graphs**," Minisymposium on Parallel Analysis of Massive Social Networks, *15th SIAM Conference on Parallel Processing for Scientific Computing* (PP12), Savannah, GA, February 15-17, 2012.
- D. Ediger, E.J. Riedy, H. Meyerhenke, and D.A. Bader, "**Analyzing Massive Networks with GraphCT**," Poster Session, *15th SIAM Conference on Parallel Processing for Scientific Computing* (PP12), Savannah, GA, February 15-17, 2012.
- R.C. McColl, D. Ediger, and D.A. Bader, "**Many-Core Memory Hierarchies and Parallel Graph Analysis**," Poster Session, *15th SIAM Conference on Parallel Processing for Scientific Computing* (PP12), Savannah, GA, February 15-17, 2012.
- E.J. Riedy, D. Ediger, H. Meyerhenke, and D.A. Bader, "**STING: Software for Analysis of Spatio-Temporal Interaction Networks and Graphs,**" Poster Session, *15th SIAM Conference on Parallel Processing for Scientific Computing* (PP12), Savannah, GA, February 15-17, 2012.
- Y. Chai, Z. Du, D.A. Bader, and X. Qin, "**Efficient Data Migration to Conserve Energy in Streaming Media Storage Systems,**" *IEEE Transactions on Parallel & Distributed Systems*, 2012.
- M. S. Swenson, J. Anderson, A. Ash, P. Gaurav, Z. Sükösd, D.A. Bader, S.C. Harvey and C.E Heitsch, "**GTfold: Enabling parallel RNA secondary structure prediction on multi-core desktops**," *BMC Research Notes*, 5:341, 2012.
- D. Ediger, K. Jiang, E.J. Riedy, and D.A. Bader, "**GraphCT: Multithreaded Algorithms for Massive Graph Analysis**," *IEEE Transactions on Parallel & Distributed Systems*, 2012.
- D.A. Bader and K. Madduri, "**Computational Challenges in Emerging Combinatorial Scientific Computing Applications**," in O. Schenk, editor, *Combinatorial Scientific Computing*, Chapman & Hall / CRC Press, Chapter 17, pages 471-494, 2012.
- O. Green, R. McColl, and D.A. Bader, "**GPU Merge Path -- A GPU Merging Algorithm**," *26th ACM International Conference on Supercomputing* (ICS), San Servolo Island, Venice, Italy, June 25-29, 2012.
- O. Green, R. McColl, and D.A. Bader, "**A Fast Algorithm for Streaming Betweenness Centrality**," *4th ASE/IEEE International Conference on Social Computing* (SocialCom), Amsterdam, The Netherlands, September 3-5, 2012.
- D. Ediger, R. McColl, J. Riedy, and D.A. Bader, "**STINGER: High Performance Data Structure for Streaming Graphs**," *The IEEE High Performance Extreme Computing Conference* (HPEC), Waltham, MA, September 20-22, 2012. **Best Paper Award.**
- J. Marandola, S. Louise, L. Cudennec, J.-T. Acquaviva and D.A. Bader, "**Enhancing Cache Coherent Architecture with Access Patterns for Embedded Manycore Systems**," *14th IEEE International Symposium on System-on-Chip* (SoC), Tampere, Finland, October 11-12, 2012.
- L.M. Munguía, E. Ayguade, and D.A. Bader, "**Task-based Parallel Breadth-First Search in Heterogeneous Environments**," *The 19th Annual IEEE International Conference on High Performance Computing* (HiPC), Pune, India, December 18-21, 2012.

Bader, Related Recent Publications (2013)

- X. Liu, P. Pande, H. Meyerhenke, and D.A. Bader, "**PASQUAL: Parallel Techniques for Next Generation Genome Sequence Assembly,**" *IEEE Transactions on Parallel & Distributed Systems*, 24(5):977-986, 2013.
- David A. Bader, Henning Meyerhenke, Peter Sanders, and Dorothea Wagner (eds.), ***Graph Partitioning and Graph Clustering***, American Mathematical Society, 2013.
- E. Jason Riedy, Henning Meyerhenke, David Ediger and David A. Bader, "**Parallel Community Detection for Massive Graphs**," in David A. Bader, Henning Meyerhenke, Peter Sanders, and Dorothea Wagner (eds.), *Graph Partitioning and Graph Clustering*, American Mathematical Society, Chapter 14, pages 207-222, 2013.
- S. Kang, D.A. Bader, and R. Vuduc, "**Energy-Efficient Scheduling for Best-Effort Interactive Services to Achieve High Response Quality**," *27th IEEE International Parallel and Distributed Processing Symposium* (IPDPS), Boston, MA, May 20-24, 2013.
- J. Riedy and D.A. Bader, "**Multithreaded Community Monitoring for Massive Streaming Graph Data**," *7th Workshop on Multithreaded Architectures and Applications* (MTAAP), Boston, MA, May 24, 2013.
- D. Ediger and D.A. Bader, "**Investigating Graph Algorithms in the BSP Model on the Cray XMT**," *7th Workshop on Multithreaded Architectures and Applications* (MTAAP), Boston, MA, May 24, 2013.
- O. Green and D.A. Bader, "**Faster Betweenness Centrality Based on Data Structure Experimentation**," *International Conference on Computational Science* (ICCS), Barcelona, Spain, June 5-7, 2013.
- Z. Yin, J. Tang, S. Schaeffer, and D.A. Bader, "**Streaming Breakpoint Graph Analytics for Accelerating and Parallelizing the Computation of DCJ Median of Three Genomes**," *International Conference on Computational Science* (ICCS), Barcelona, Spain, June 5-7, 2013.
- T. Senator, D.A. Bader, et al., "**Detecting Insider Threats in a Real Corporate Database of Computer Usage Activities**," *19th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (KDD), Chicago, IL, August 11-14, 2013.
- J. Fairbanks, D. Ediger, R. McColl, D.A. Bader and E. Gilbert, "**A Statistical Framework for Streaming Graph Analysis**," *IEEE/ACM International Conference on Advances in Social Networks Analysis and Modeling* (ASONAM), Niagara Falls, Canada, August 25-28, 2013.
- A. Zakrzewska and D.A. Bader, "**Measuring the Sensitivity of Graph Metrics to Missing Data**," *10th International Conference on Parallel Processing and Applied Mathematics* (PPAM), Warsaw, Poland, September 8-11, 2013.
- O. Green and D.A. Bader, "**A Fast Algorithm for Streaming Betweenness Centrality**," *5th ASE/IEEE International Conference on Social Computing* (SocialCom), Washington, DC, September 8-14, 2013.
- R. McColl, O. Green, and D.A. Bader, "**A New Parallel Algorithm for Connected Components in Dynamic Graphs**," *The 20th Annual IEEE International Conference on High Performance Computing* (HiPC), Bangalore, India, December 18-21, 2013.

Bader, Related Recent Publications (2014-2015)

- R. McColl, D. Ediger, J. Poovey, D. Campbell, and D.A. Bader, "A Performance Evaluation of Open Source Graph Databases," *The 1st Workshop on Parallel Programming for Analytics Applications* (PPAA 2014) held in conjunction with the *19th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming* (PPoPP 2014), Orlando, Florida, February 16, 2014.
- O. Green, L.M. Munguia, and D.A. Bader, "Load Balanced Clustering Coefficients," *The 1st Workshop on Parallel Programming for Analytics Applications* (PPAA 2014) held in conjunction with the *19th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming* (PPoPP 2014), Orlando, Florida, February 16, 2014.
- A. McLaughlin and D.A. Bader, "Revisiting Edge and Node Parallelism for Dynamic GPU Graph Analytics," *8th Workshop on Multithreaded Architectures and Applications* (MTAAP), held in conjuntion with *The IEEE International Parallel and Distributed Processing Symposium (IPDPS 2014)*, Phoenix, AZ, May 23, 2014.
- Z. Yin, J. Tang, S. Schaeffer, D.A. Bader, "A Lin-Kernighan Heuristic for the DCJ Median Problem of Genomes with Unequal Contents," *20th International Computing and Combinatorics Conference* (COCOON), Atlanta, GA, August 4-6, 2014.
- Y. You, D.A. Bader and M.M. Dehnavi, "Designing an Adaptive Cross-Architecture Combination for Graph Traversal," *The 43rd International Conference on Parallel Processing* (ICPP 2014), Minneapolis, MN, September 9-12, 2014.
- A. McLaughlin, J. Riedy, and D.A. Bader, "Optimizing Energy Consumption and Parallel Performance for Betweenness Centrality using GPUs," *The 18th Annual IEEE High Performance Extreme Computing Conference* (HPEC), Waltham, MA, September 9-11, 2014.
- A. McLaughlin and D.A. Bader, "Scalable and High Performance Betweenness Centrality on the GPU," *The 26th IEEE and ACM Supercomputing Conference* (SC14), New Orleans, LA, November 16-21, 2014. **Best Student Paper Finalist.**
- D. Dauwe, E. Jonardi, R. Friese, S. Pasricha, A.A. Maciejewski, D.A. Bader, and H.J. Siegel, "**A Methodology for Co-Location Aware Application Performance Modeling in Multicore Computing**," 17th Workshop on Advances on Parallel and Distributed Processing Symposium (APDCM), Hyderabad, India, May 25, 2015.
- A. Zakrzewska and D.A. Bader, "**Fast Incremental Community Detection on Dynamic Graphs**," 11th International Conference on Parallel Processing and Applied Mathematics (PPAM), Krakow, Poland, September 6-9, 2015.
- A. McLaughlin, J. Riedy, and D.A. Bader, "**An Energy-Efficient Abstraction for Simultaneous Breadth-First Searches**," The 19th Annual IEEE High Performance Extreme Computing Conference (HPEC), Waltham, MA, September 15-17, 2015.
- A. McLaughlin, D. Merrill, M. Garland and D.A. Bader, "**Parallel Methods for Verifying the Consistency of Weakly-Ordered Architectures**," The 24th International Conference on Parallel Architectures and Compilation Techniques (PACT), San Francisco, CA, October 18-21, 2015.

Acknowledgment of Support

National Institutes of Health

Department of Energy · United States of America

NASA

Department of Defense · United States of America

DARPA

Oak Ridge National Laboratory

Pacific Northwest National Laboratory

Sandia National Laboratories

IBM

Microsoft Research

NVIDIA

Northrop Grumman

LexisNexis

Sony

Intel

Cray

Sun microsystems

Xilinx

Toshiba

Georgia Tech College of Computing