Applications and challenges in large-scale graph analysis
David A. Bader, Jason Riedy, Henning Meyerhenke
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

**Exponential growth:**
More than 900 million active users

**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**
Current Example Data Rates

- **Financial:**
  - NYSE processes 1.5TB daily, maintains 8PB

- **Social:**
  - Facebook adds >100k users, 55M “status” updates, 80M photos daily; ~1B active users with an average of 130 “friend” connections each.
  - Foursquare 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.
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

“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
Objective: Research and develop new algorithms and software for crucial graph analysis problems in cybersecurity, intelligence integration, and network analysis.

- In-the-field, embedded processing systems have limited computational capabilities due to severe power constraints. DARPA’s Power Efficiency Revolution For Embedded Computing Technologies (PERFECT) program will address this need by increasing the computational power efficiency of these embedded systems. The project seeks a power efficiency of 75 GFLOPS/watt.

- GRATEFUL will extend high-performance graph analysis algorithms to reduce power usage and provide resilience against imperfect data. We will build a model of concurrency requirements for graph algorithms. The project will provide energy-conscious and data error-resilient algorithms on streaming data, including computing and maintaining shortest path trees and graph decompositions into communities.

**Pls: David A. Bader and Jason Riedy**

Sponsored under the DARPA PERFECT program, Contract HR0011-13-2-0001
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.
Example: Mining Twitter for Social Good

Massive Social Network Analysis: Mining Twitter for Social Good

ICPP 2010

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

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

William N. Reynolds
Los Alamos National Lab
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 XE6, GraphCT estimates the betweenness centrality of an artificially generated 1.57 billion vertex, 8.6 billion edge graph in 25 minutes and a real-world graph (Twitter, et al.) with 64.6 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 120 ‘friendship’ connections each and sharing 5 references to items each month [11].

One analysis approach treats the interactions as dynamic and applies tools from graph theory, social network analysis, and scale-free networks [26]. However, the volume of data that needs to be processed to apply techniques overwhelms current computational cap. Even well-understood analytic methodologies advance in both hardware and software to process the growing corpus of social media.

Social media provides staggering amounts of information. The data is rich and multidimensional, representing human

Fig. 3. Subcommunity filtering on Twitter data sets

Top 15 Users by Betweenness Centrality

<table>
<thead>
<tr>
<th>Rank</th>
<th>H1N1</th>
<th>Data Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>@CDCFlu</td>
<td>@ajc</td>
</tr>
<tr>
<td>2</td>
<td>@addthis</td>
<td>@driveafast</td>
</tr>
<tr>
<td>3</td>
<td>@Official_PAX</td>
<td>@ATLCheap</td>
</tr>
<tr>
<td>4</td>
<td>@FluGov</td>
<td>@TWC1</td>
</tr>
<tr>
<td>5</td>
<td>@nytimes</td>
<td>@HelloNorthGA</td>
</tr>
<tr>
<td>6</td>
<td>@tweetmeme</td>
<td>@11AliveNews</td>
</tr>
<tr>
<td>7</td>
<td>@mercola</td>
<td>@WSB_TV</td>
</tr>
<tr>
<td>8</td>
<td>@CNN</td>
<td>@shaunking</td>
</tr>
<tr>
<td>9</td>
<td>@backstreetboys</td>
<td>@Carl</td>
</tr>
<tr>
<td>10</td>
<td>@EllieSmith_X</td>
<td>@SpaceyG</td>
</tr>
<tr>
<td>11</td>
<td>@TIME</td>
<td>@ATLIntownPa</td>
</tr>
<tr>
<td>12</td>
<td>@CDCemergency</td>
<td>@TJnDJs</td>
</tr>
<tr>
<td>13</td>
<td>@CDC_eHealth</td>
<td>@ATLien</td>
</tr>
<tr>
<td>14</td>
<td>@perezhilton</td>
<td>@MarshallRamsey</td>
</tr>
<tr>
<td>15</td>
<td>@billmaher</td>
<td>@Kanye</td>
</tr>
</tbody>
</table>

Image credit: bioethicsinstitute.org
Massive Data Analytics: Protecting our Nation

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

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

The New York Times
Thursday, September 4, 2008

Report on Blackout Is Said To Describe Failure to React

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

ENS GO 000014
5332.2
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

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.

**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.

(2,3) [www.visualComplexity.com](http://www.visual Complexity.com)
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
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 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
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
STINGER Software Dissemination

- [http://www.cc.gatech.edu/stinger](http://www.cc.gatech.edu/stinger)

---

**STINGER**

Static-Temporal Interaction Networks and Graphs (STINGER) Enhances Representation

**Home**

**Motivation**

Many problems today can be formulated as dynamic spatial-temporal graph problems. For example, one may wish to track communications within social networks or Facebook as edges (friendship) are added or removed. Or more directly, one may look for people who bridge between different social communities, or switch allegiances over time.

As the computer science community increases its development of algorithms and codes for large scale graphs problems, no canonical graph representation has yet to emerge. Without a standard graph representation, algorithms that are implemented for one framework may require substantial reprogramming efforts to port to a different framework. Even worse, algorithms within a single framework may use different data structures for each type of graph. The STINGER representation integrates many types of graphs including social, temporal, and geographic graphs. STINGER as a result can represent all graphs in time, space, and possibly, must be reduced or eliminated through a canonical graph representation.

**STINGER**

STINGER accomplishes the following objectives:

- **Scalability**: Algorithms written for STINGER can easily be translated/ported between multiple languages and environments.
- **Productivity**: STINGER should provide a common abstract data structure such that the large graph community can quickly leverage and adopt STINGER.
- **Performance**: STINGER should be able to run as fast as the original data. This is not just in a speed of execution, but also in terms of data size.

---

**Georgia Tech**

College of Computing
**Bader, Related Recent Publications (2005-2008)**

Bader, Related Recent Publications (2009-2010)

- 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.
- 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.
Bader, Related Recent Publications (2011-2012)

Acknowledgment of Support
Frontiers in Large-Scale Graph Analysis

- 2:00-2:25 Applications and challenges in large-scale graph analysis
  - David A. Bader and Jason Riedy, Georgia Institute of Technology, USA; Henning Meyerhenke, Karlsruhe Institute of Technology, Germany
- 2:30-2:55 Large scale graph analytics and randomized algorithms for applications in cybersecurity
  - John Johnson, Pacific Northwest National Laboratory, USA
- 3:00-3:25 Anomaly Detection in Very Large Graphs: Modeling and Computational Considerations
  - Benjamin Miller, Nicholas Arcolano, Edward Rutledge, Matthew Schmidt, and Nadya Bliss, Massachusetts Institute of Technology, USA
- 3:30-3:55 Combinatorial and Numerical Algorithms for Network Analysis
  - Henning Meyerhenke and Christian Staudt, Karlsruhe Institute of Technology, Germany

Part II (MS179) Thursday, February 28, 9:30 AM - 11:30 AM

- 9:30-9:55 Are we there yet? When to stop a Markov chain while generating random graphs?
  - Ali Pinar, Jaideep Ray, and C. Seshadhri, Sandia National Laboratories, USA
- 10:00-10:25 Analyzing graph structure in streaming data with STINGER
  - Jason Riedy, David A. Bader, Robert C. Mccoll, and David Ediger, Georgia Institute of Technology, USA
- 10:30-10:55 High-Performance Filtered Queries in Attributed Semantic Graphs
  - John R. Gilbert, University of California, Santa Barbara, USA; Aydin Buluc, Lawrence Berkeley National Laboratory, USA; Armando Fox, University of California, Berkeley, USA; Shoaib Kamil, Massachusetts Institute of Technology, USA; Adam Lugowski, University of California, Santa Barbara, USA; Leonid Oliker and Samuel Williams, Lawrence Berkeley National Laboratory, USA
- 11:00-11:25 Large-Scale Graph-Structured Machine Learning: GraphLab in the Cloud and GraphChi in your PC
  - Joseph Gonzalez and Carlos Guestrin, Carnegie Mellon University, USA