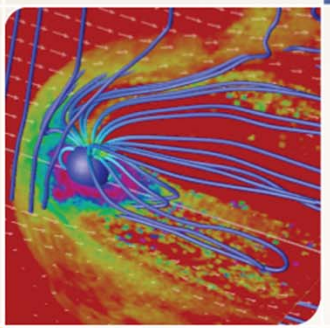
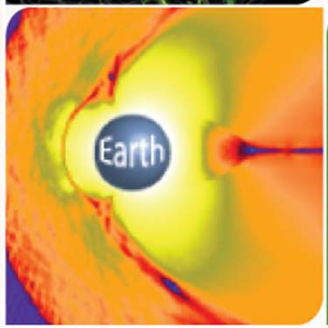
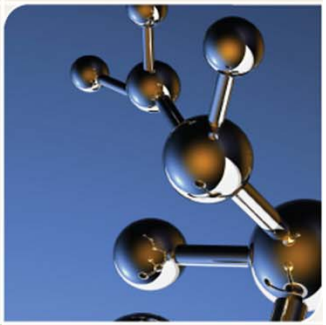
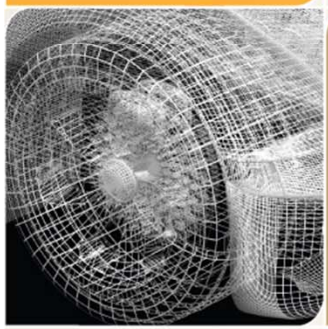
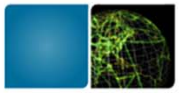




Future of High-Performance Computing Workshop: Enabling Broader Engagement and Workforce Development

David A. Bader

Georgia Tech College of Computing
Computational Science and Engineering

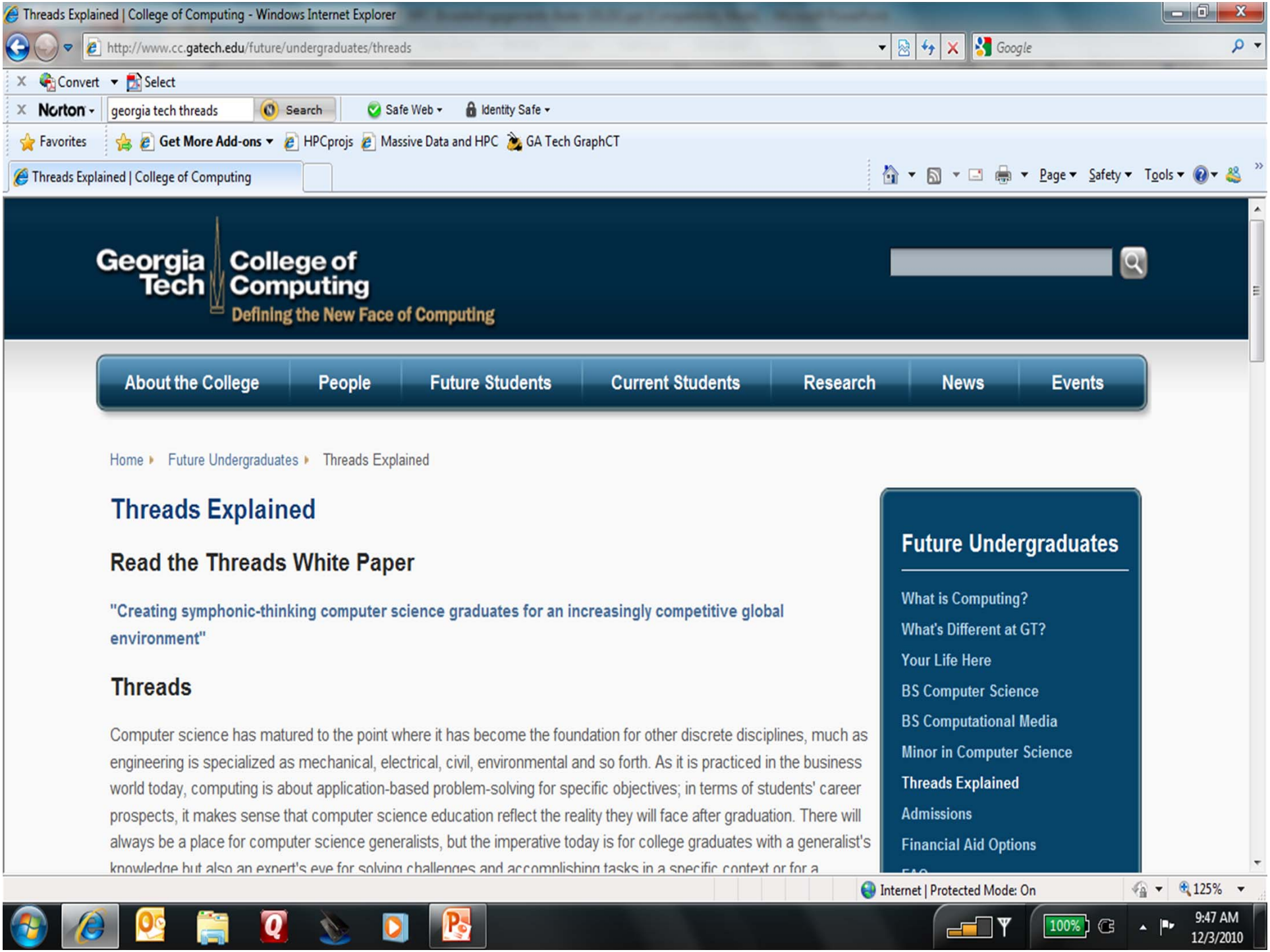




Panelists will answer the following questions:

1. What is broader engagement and why is it important?
2. What strategies should we use to engage new communities?
3. How can we train and sustain the next generation?
4. What are the challenges and how can we overcome them?

Diverse: Issues in Higher Education

- Bachelor's (2010)
  - No. 2 in engineering bachelor's degrees awarded to African American students
  - No. 2 in engineering bachelor's degrees awarded to all categories of minority students
  - No. 3 in engineering bachelor's degrees awarded to Asian American students
- Master's (2010)
  - No. 4 in engineering master's degrees awarded to African American students
  - No. 4 in physical sciences master's degrees awarded to African American students
  - No. 4 in engineering master's degrees awarded to Hispanic students
  - No. 5 in engineering master's degrees awarded to all minority students
  - No. 10 in physical sciences master's degrees awarded to all minority students
- Doctoral (2010)
  - No. 1 in engineering doctoral degrees awarded to African American students
  - No. 1 in engineering doctoral degrees awarded to Hispanic students
  - No. 1 in engineering doctoral degrees awarded to Asian American students
  - No. 1 in engineering doctoral degrees awarded to all minority students


Threads Explained | College of Computing - Windows Internet Explorer
http://www.cc.gatech.edu/future/undergraduates/threads
Google
Convert
Select
Norton
georgia tech threads
Search
Safe Web
Identity Safe
Favorites
Get More Add-ons
HPCprojs
Massive Data and HPC
GA Tech GraphCT
Threads Explained | College of Computing
Page
Safety
Tools
Georgia Tech College of Computing
Defining the New Face of Computing
About the College
People
Future Students
Current Students
Research
News
Events
Home
Future Undergraduates
Threads Explained
Threads Explained
Read the Threads White Paper
"Creating symphonic-thinking computer science graduates for an increasingly competitive global environment"
Threads
Computer science has matured to the point where it has become the foundation for other discrete disciplines, much as engineering is specialized as mechanical, electrical, civil, environmental and so forth. As it is practiced in the business world today, computing is about application-based problem-solving for specific objectives; in terms of students' career prospects, it makes sense that computer science education reflect the reality they will face after graduation. There will always be a place for computer science generalists, but the imperative today is for college graduates with a generalist's knowledge but also an expert's eye for solving challenges and accomplishing tasks in a specific context or for a
Future Undergraduates
What is Computing?
What's Different at GT?
Your Life Here
BS Computer Science
BS Computational Media
Minor in Computer Science
Threads Explained
Admissions
Financial Aid Options
Internet | Protected Mode: On
125%
100%
9:47 AM
12/3/2010

Threads Explained | College of Computing - Windows Internet Explorer

http://www.cc.gatech.edu/future/undergraduates/threads

Google

Convert Select

Norton georgia tech threads Search Safe Web Identity Safe

Favorites Get More Add-ons HPCprojs Massive Data and HPC GA Tech GraphCT

Threads Explained | College of Computing

Page Safety Tools

Devices

Creating devices embedded in physical objects that interact in the physical world

Information Internetworks

Representing, transforming, transmitting, and presenting information

Intelligence

Building top-to-bottom models of human-level intelligence

Media

Building systems in order to exploit computing's abilities to provide creative outlets

Modeling & Simulation

Representing natural and physical processes

People

Designing, building, and evaluating systems that treat the human as a central component

Platforms

Creating computer architectures, systems and languages

Theory

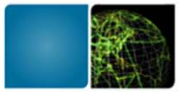
Theoretical foundations underlying a wide range of computing disciplines

Internet | Protected Mode: On

125%

100%

9:48 AM
12/3/2010


Exascale Analytics: Real-world challenges

- Health care → disease spread, detection and prevention of epidemics/pandemics (e.g. SARS, Avian flu, H1N1 "swine" flu, …)
- Massive social networks → energy conservation requires social change, modeling 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, and evolution
- Electric Power Grid → communication, transportation, energy, water, food supply
- Modeling and Simulation → Perform full-scale economic-social-political simulations

Requires dynamic Spatio-Temporal Interaction Networks and Graphs (STING)

Homeland Security: Terrorist Networks

- Certain activities are often suspicious not because of the characteristics of a single actor, but because of the interactions among a group of actors.
- Interactions are modeled through a graph abstraction where the entities are represented by vertices, and their interactions are the directed edges in the graph.

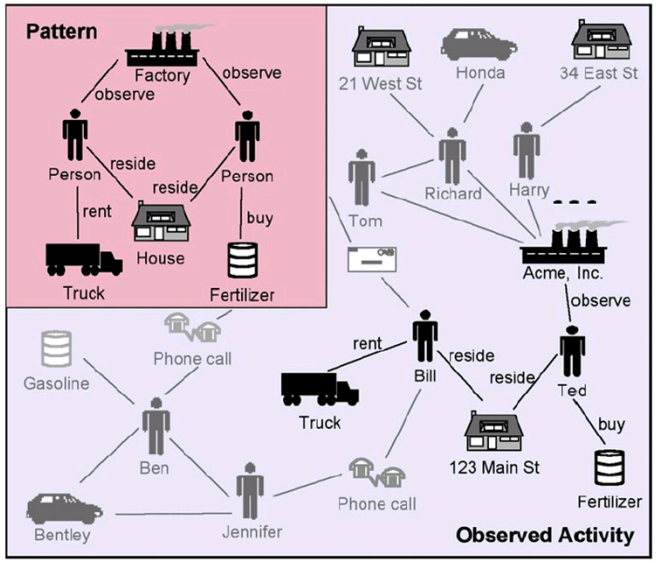


Figure Credit: Graph-based technologies for intelligence analysis,T. Coffman, S. Greenblatt, S. Marcus, Commun. ACM, 47(3):45-47, 2004.

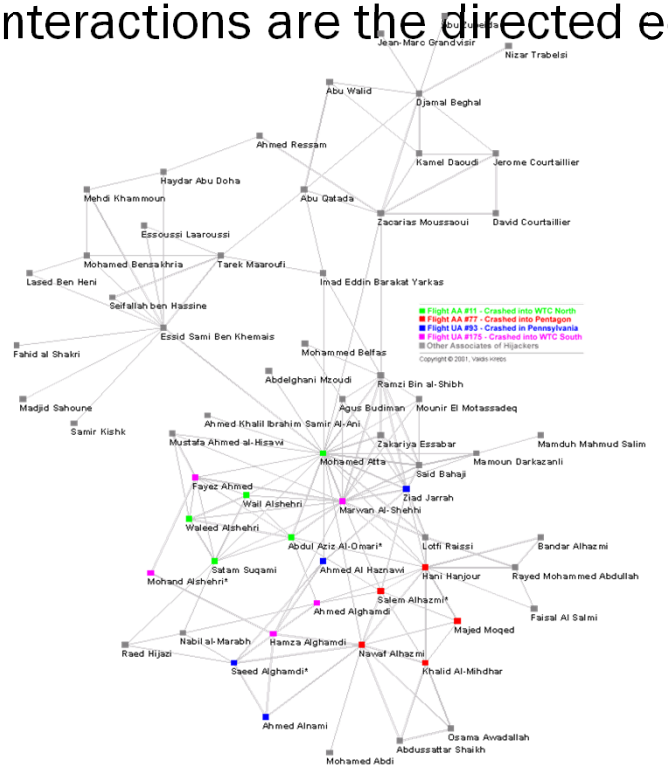
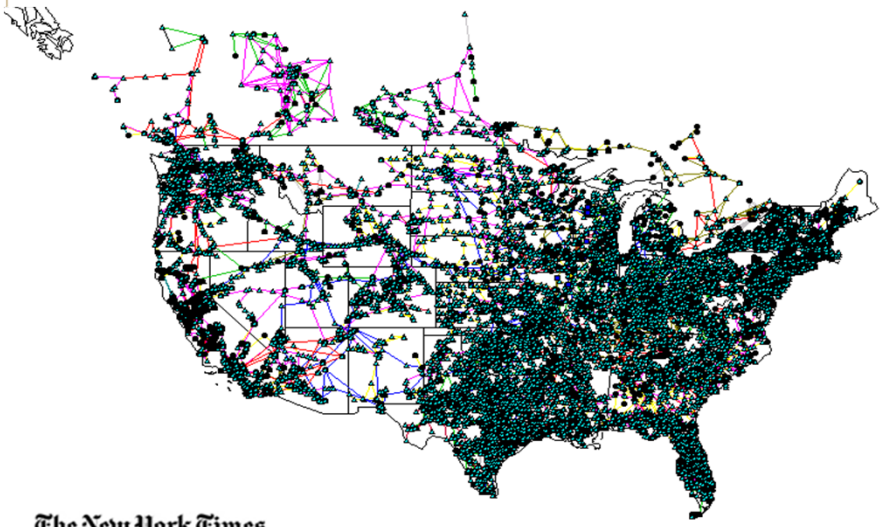


Figure Credit: Uncloaking Terrorist Networks, V.E. Krebs, First Monday, 7(4), April 2002.

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

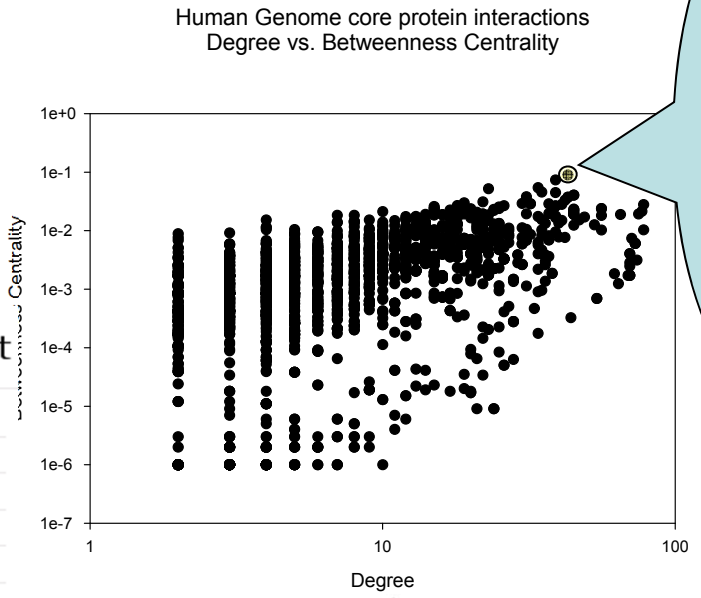
A report on the Aug. 14 blackout identifies specific lapses by various parties, including FirstEnergy's failure to react properly to the loss of a transmission line, people who have seen drafts of it say.

A working group of experts from eight states and Canada will meet in private on Wednesday to evaluate the report, people involved in the

E-MAIL
PRINT
SINGLE-PAGE
REPRINTS
SAVE
SHARE

Public Health

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

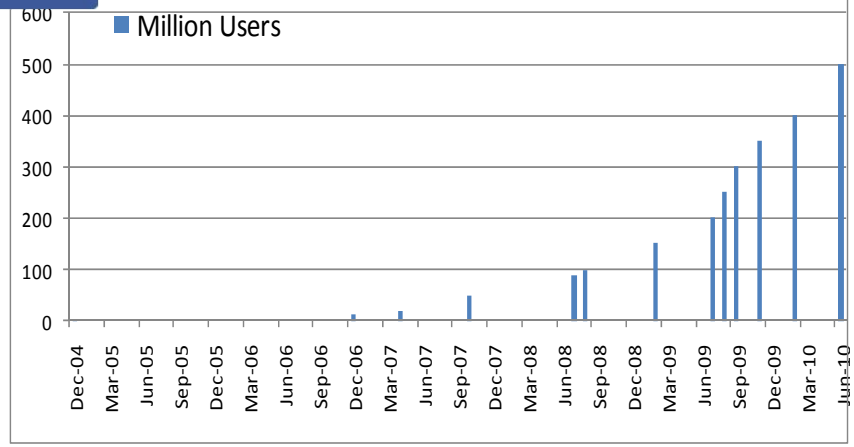


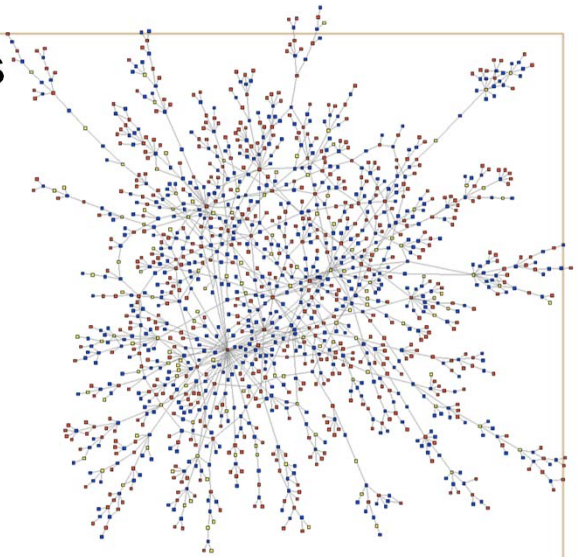
ENSG0000001455332.2 Kelch-like protein 8 implicated in breast cancer

Computational Social Sciences Driving the Need for HPC

facebook has more than 500 million active users

3 orders of magnitude growth in 3 years!





- Note the graph is **changing** as well as growing.
- What are this graph's properties? *How do they change?*
- Traditional graph partitioning often fails:
  - **Topology**: Interaction graph is low-diameter, and has no good separators
  - **Irregularity**: Communities are not uniform in size
  - **Overlap**: individuals are members of one or more communities
- 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

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

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, In
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 *T*oolkit 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 ave
120 'friendship' connections each and sharing 5
references to items each month [11].

One analysis approach treats the interactions as
and applies tools from graph theory, social n
analysis, and scale-free networks [29]. Howev
volume of data that must be processed to appl
techniques overwhelms current computational cap
Even well-understood analytic methodologies
advances in both hardware and software to proc
growing corpus of social media.

Social media provides staggering amounts

TOP 15 USERS BY BETWEENNESS CENTRALITY

Rank	Data Set H1N1	Data Set atlflood
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

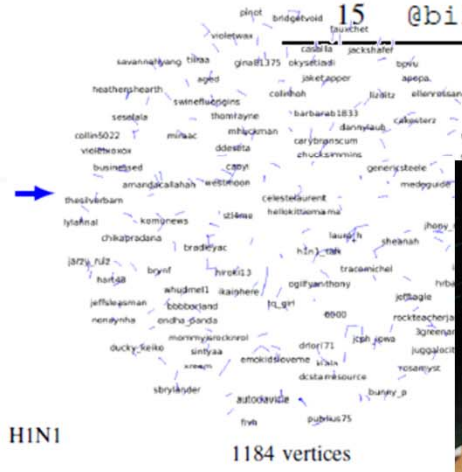


Fig. 3. Subcommunity filtering on Twitter data sets

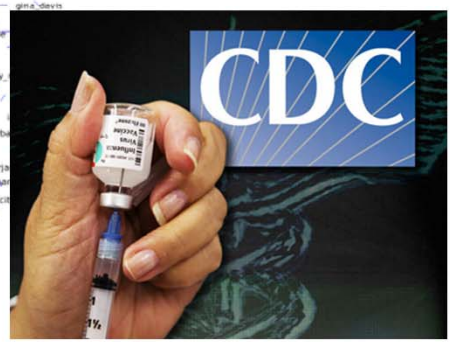


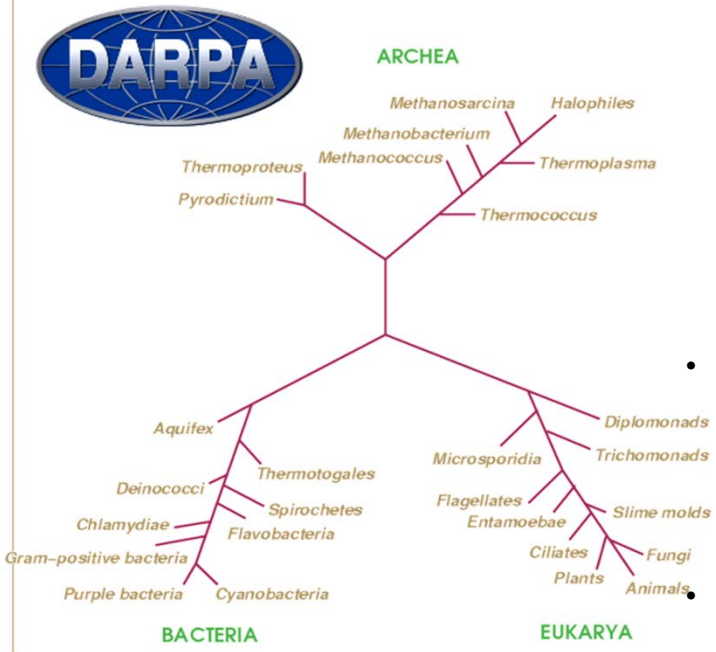
Image credit: bioethicsinstitute.org

twitter™
public tweets

Pacific Northwest NATIONAL LABORATORY

NSF PetaApps Award ($1 Million) for Phylogenetics Research on IBM Blue Waters

As part of the IBM PERCS team, we designed the IBM Blue Waters supercomputer that will sustain petascale performance on our applications, under the **DARPA High Productivity Computing Systems** program.



- GRAPPA: Genome Rearrangements Analysis under Parsimony and other Phylogenetic Algorithm
  - Freely-available, open-source, GNU GPL
  - already used by other computational phylogeny groups, Caprara, Pevzner, LANL, FBI, Smithsonian Institute, Aventis, GlaxoSmithKline, PharmCos.
- Gene-order Phylogeny Reconstruction
  - Breakpoint Median
  - Inversion Median
- over one-billion fold speedup from previous codes
- Parallelism scales linearly with the number of processors

FACULTY
David A. Bader, CSE

www.phylo.org

Georgia Tech creates a Petascale Pipeline to Accelerate New Science, Engineering, and Users

Development of new petascale users and applications

Petascale Foundation
Curricula, Training, Outreach

Broader participation and impact

Broad Participation Next-gen users, high school outreach

Under-graduate Comput Model. Threads & Research

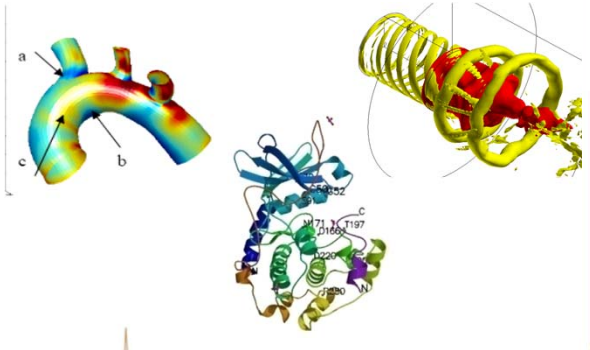
MS + PhD CSE program

Petascale-ready users, deep impact

Dual Degrees Student from Morehouse College and GA Tech

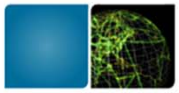
Petascale Systems

Breakthrough science and engineering at multiple levels



GA Tech grows the nation's pipeline

Broadening Participation in Petascale Science and Engineering


Petascale Foundation
Curricula, Training, Outreach

Broad Participation

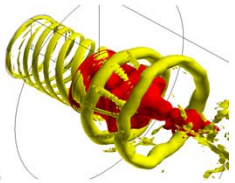
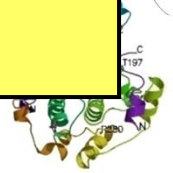
B
participation
and impact

Curricula and outreach to increase the broad participation and impact

- **Education portal** for national access to all materials
- **Engagement with and enhancement materials for under-represented groups** (Intel Opportunity Scholars and SAIC Scholars promote undergraduate research experiences and help retain underrepresented minorities and women in computer science at GA Tech.
- **Summer Undergraduate Internships** with petascale faculty users such as the CRUISE (CSE Research for Undergraduates in Summer Experience) program, with support from NSF REU, the DoD HPC JOEM minority program, and industry.


Dual Degrees Student from Morehouse College and GA Tech

gh
d
l at
els

GA Tech grows the nation's pipeline

In aggregate, we provide a comprehensive, pipelined environment for petascale science and engineering

Development of new petascale users and applicatior

Petascale Foundation
Curricula, Training, Outreach

Broader participation and impact

Broad Participation
Next-gen users, high school outreach

Under-graduate
Comput Model. Threads & Research

MS + PI
CSE
progra

GA Tech grows the nation's

Curricula and education targeted to the next-gen of petascale users

- **Undergraduate course materials and courses** in "computational modeling" thread: parallel programming for multicore, large-scale parallelism
- GA Tech is establishing a rotating **CSE seminar series** and **undergraduate research program** with Morehouse College (A. Johnson), and Spelman College (A. Lawrence).
- **Leadership in Education:** new MS & PhD graduate curriculum in CSE with **HPC**, **large-scale data analysis**, **modeling & simulation**, **num. methods**, and **real-world algorithms**.
- **Curriculum sharing** for national access to all materials through: tutorials and workshops at TG and SC; Computational Science Education Reference Desk (CSERD), an NSF-supported national resource library

Georgia Tech promotes petascale research and education

Development of new petascale users and applications

Petascale Foundation

Curricula, Training,

Strong commitment to integrate research and education

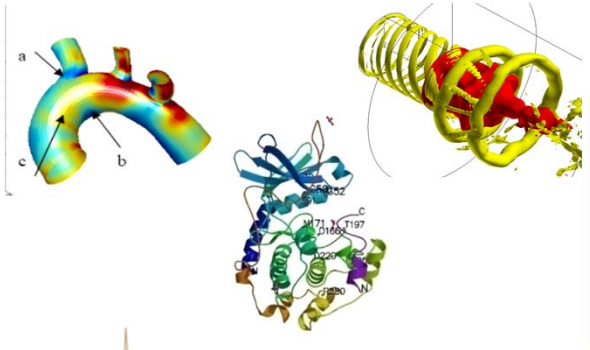
Petascale-ready users, deep impact

Broader participation and impact

Dual Degrees Student from Morehouse College and GA Tech

Petascale Systems

Breakthrough science and engineering at multiple levels

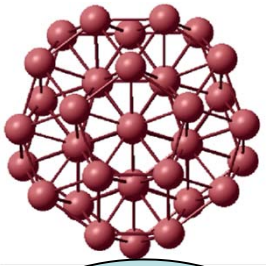


GA Tech grows the nation's pipeline

Computational Science & Engineering (CSE)

CSE is a discipline devoted to the systematic study of computer-based models of natural and engineered systems.

Astrophysics



Nanomaterials

Transportation

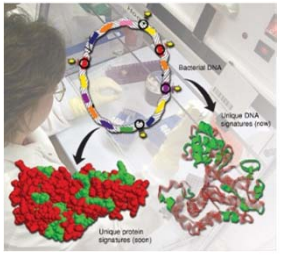
Weather and climate

Computation

CSE

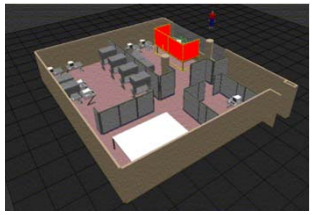
Mathematics

Aerospace



Biology (drug design, cancer treatment, phylogeny, …)

Biomedical



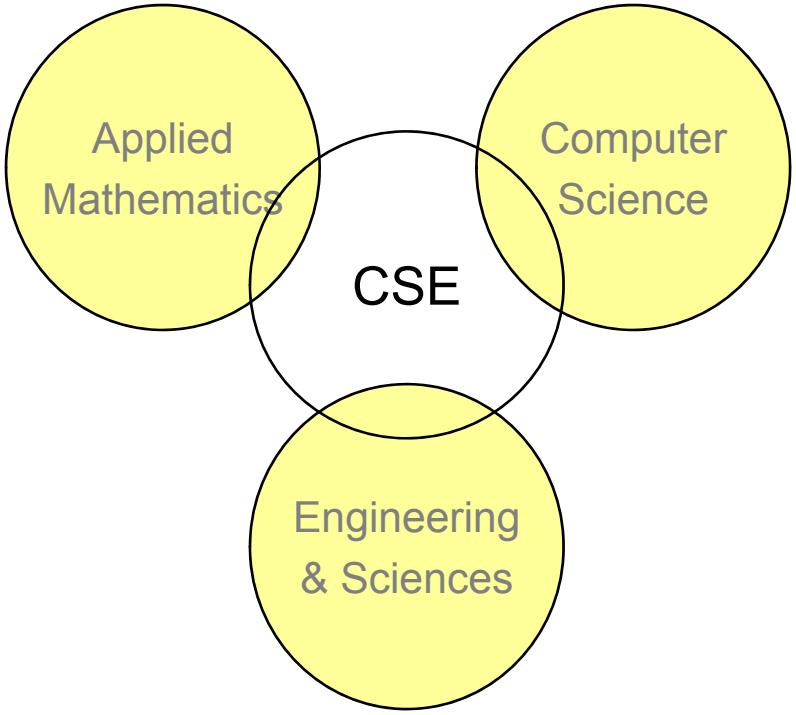
Manufacturing

Interdisciplinary collaboration with science and engineering

Computational Science & Engineering: A Discipline

Computation is often characterized as the third paradigm for advancing knowledge and practice in science and engineering, in addition to theory and experimentation.

CSE is a discipline arising from the confluence of principles from computing, applied mathematics, science and engineering.



Subfields include:

- Numerical computing
- Discrete algorithms
- Modeling and simulation
- Computational data analysis, machine learning and visualization
- High performance computing

Georgia Tech

Georgia Tech

Colleges

- Architecture
- Computing
- Liberal Arts
- Management
- Engineering
- Sciences

Schools

- Computer Science
- Interactive Computing
- Computational Science and Engineering

- High Performance Computing
- Modeling and Simulation
- Data Analysis, Machine Learning & Visualization
- Numerical Computing
- Discrete Algorithms

- CSE created in 2005; School in 2010
- Composed of new faculty + interdisciplinary joint appointments

CSE Faculty

David A. Bader
High Performance Computing
('96 Univ. Maryland)

David Sherrill
High Performance Computing
('96 UGa)
joint Chemistry

Alberto Apostolico
Bioinformatics, Pattern Matching
('76 Univ. Salerno)
joint Inter. Comp.

Guy Lebanon
Machine Learning
Data Analytics
('05 CMU)

George Biros
High Performance Computing
('00 CMU)
joint Biomed. Eng.

Jeff Vetter
High Performance Computing
('98 GT)
joint with ORNL

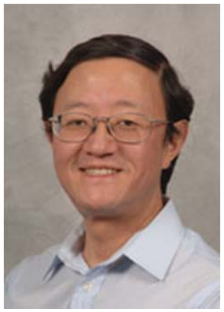
Ken Brown
Quantum Computing
('04 UC-Berkeley)
joint Chemistry

Mark Borodovsky
Bioinformatics
('76, Moscow Inst. Phs&Tech)
joint Biomed. Eng.

Haesun Park
Scientific Computing
Data Analytics
('87 Cornell)

Edmond Chow
High Performance Computing
('97 Univ. Minn.)

Rich Vuduc
High Performance Computing
('04 UC-Berkeley)



Richard Fujimoto
Parallel/Distributed Simulation
('83 UC-Berkeley)

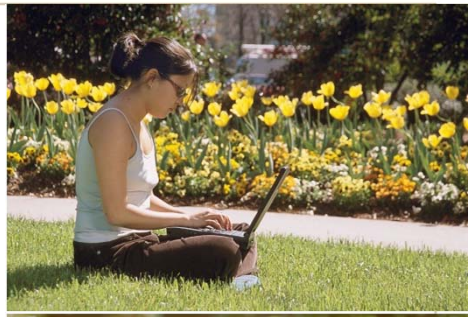
Alex Gray
Machine Learning
Data Analytics
('06 CMU)

Hongyuan Zha
Scientific Computing
Data Analytics
('93 Stanford)

Education Initiatives

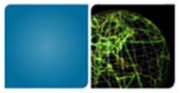
Building a pipeline of CSE professionals

- Undergraduate "thread" in CS program
  - Thread in modeling and simulation
  - Computing core + mathematics + sciences, engineering
- Multidisciplinary MS and PhD degree programs in *Computational Science and Engineering*
  - Jointly offered by three colleges: Computing, Sciences, and Engineering
  - MS offered through distance learning




CRUISE: Computing Research Undergraduate Intern Summer Experience

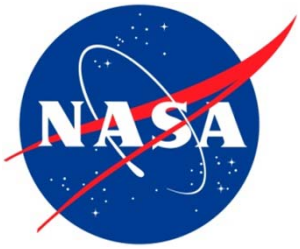
- Initiated in summer 2008
- Encourage students to consider graduate studies
- Diverse student participation
  - Multicultural, emphasizing minorities, women, international students
  - Typically 15-20 students
- Ten week summer research projects
  - High performance computing
  - Data and visual analytics (e.g., VAST challenge problem)
  - Many in interdisciplinary research projects
- CRUISE-wide events
  - Weekly seminars (technical, grad studies)
  - Social events
  - Symposium: conference-style presentations


Summary

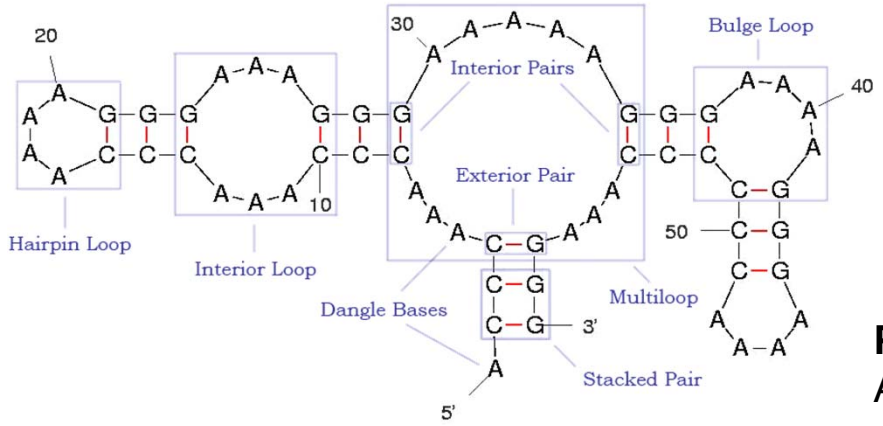
- The Georgia Tech **School of Computational Science and Engineering** was founded to establish a culture of interdisciplinary collaboration among computing, the sciences, and engineering
- Research emphases include work in core areas and a variety of disciplines
- Education initiatives include new undergraduate and graduate level programs

Acknowledgment of Support

RNA Secondary Structure Prediction

- RNA is composed of smaller building blocks called bases (Adenine, Cytosine, Guanine, Uracil)
- Pairing and non-pairing of bases is called “folding”
- Result of folding called secondary structure



Program Goals

Accurate structure of large viruses such as:

- Influenza
- HIV
- Polio
- Tobacco Mosaic
- Hanta

FACULTY

Christine Heitsch (Mathematics)
David A. Bader
Steve Harvey (Biology)

Ubiquitous High Performance Computing (UHPC)

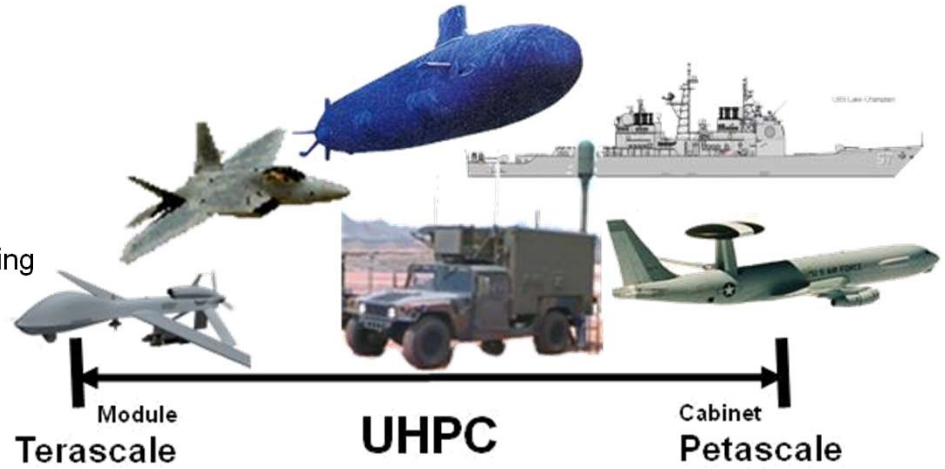
Goal: develop highly parallel, security enabled, power efficient processing systems, supporting ease of programming, with resilient execution through all failure modes and intrusion attacks

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

CRAY
THE SUPERCOMPUTER COMPANY

GEORGIA TECH

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

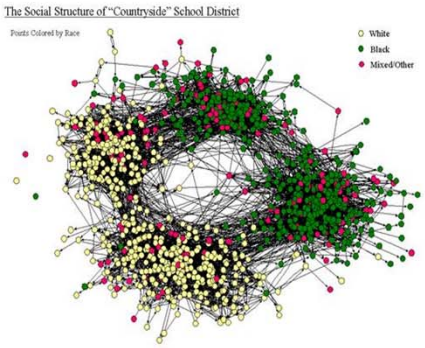
Echelon: Extreme-scale Compute Hierarchies with Efficient Locality-Optimized Nodes

Georgia Institute of Technology

CASS-MT:
Center for Adaptive Supercomputing Software

- DoD-sponsored, launched July 2008
- Pacific-Northwest Lab
  - Georgia Tech, Sandia, WA State, Delaware

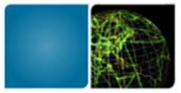
- The newest breed of supercomputers have hardware set up not just for speed, but also to better tackle large networks of seemingly random data. And now, a multi-institutional group of researchers has been awarded more than $12M to develop software for these supercomputers. Applications include anywhere complex webs of information can be found: from internet security and power grid stability to complex biological networks.



Pacific Northwest NATIONAL LABORATORY

CRAY THE SUPERCOMPUTER COMPANY

GEORGIA TECH


Bader High Performance Computing Lab

- We are the **national experts** at:
  - Multicore and Manycore Computing
  - High Performance Computing
  - Accelerated Supercomputing
  - Massive/Streaming Data Analytics
  - Extreme-scale Graph Analytics
  - Real-World Applications from Biology to Social Networks
- We work in collaboration with academia, and
  - Government: NSF, NIH, DARPA, DOE, DoD, CDC, …
  - Industry: NVIDIA, IBM, Intel, Microsoft, Cray, Northrop Grumman, LexisNexis, Netezza, Sony, Toshiba, Convey, …