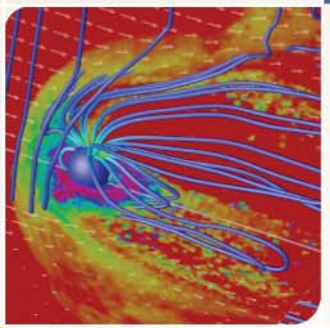
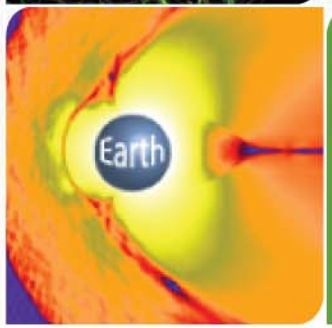
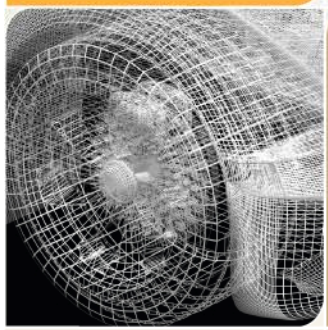




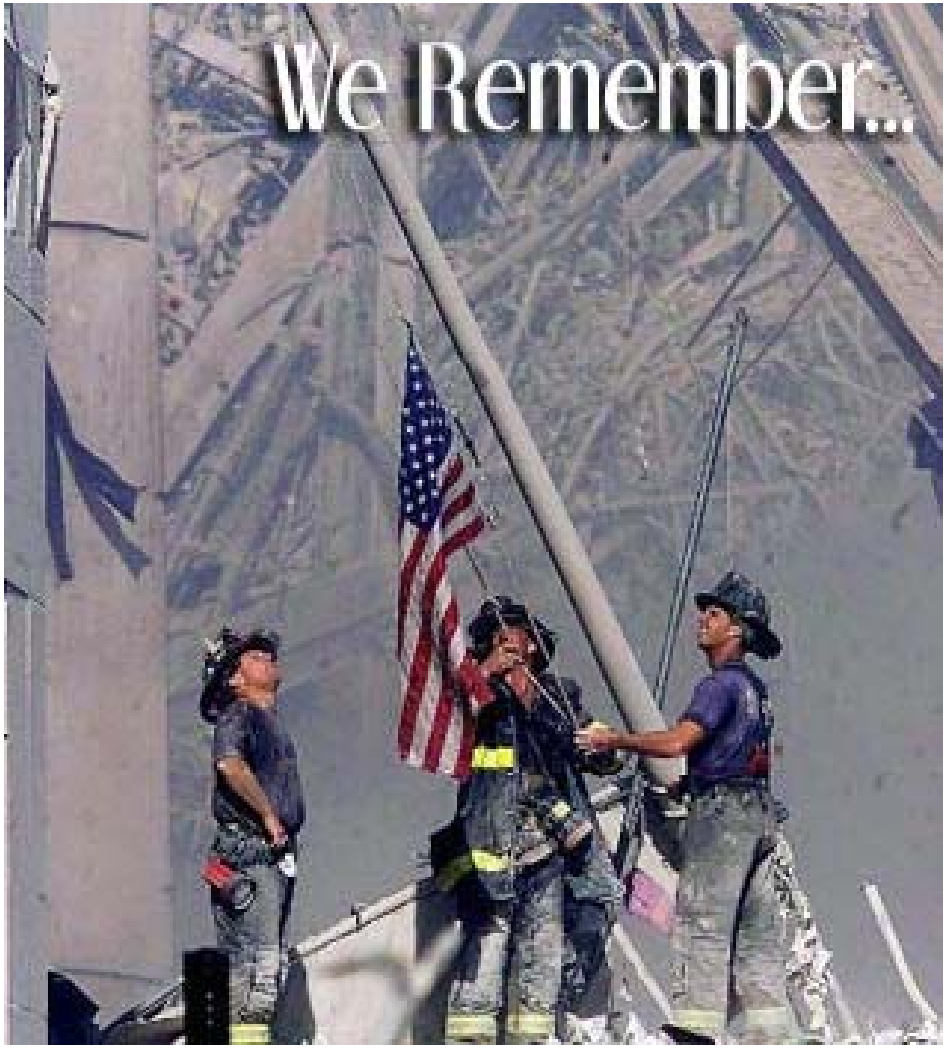

Petascale Computing for Large-Scale Graph Problems

David A. Bader

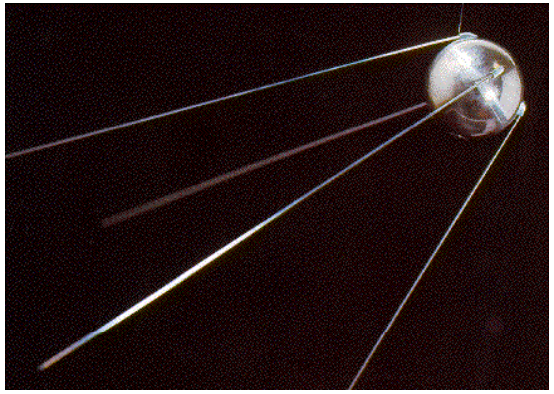



Georgia Tech College of Computing
Computational Science and Engineering

Dedication



A global race is under way ...



Sputnik (1957)

A New Arms Race to Build the World's Mightiest Computer

By JOHN MARKOFF

SAN FRANCISCO, Aug. 18 — A global race is under way to reach the next milestone in supercomputer performance, many times the speed of today's most powerful machines.

And beyond the customary rivalry in the field between the United States and Japan, there is a new entrant — China — eager to showcase its arrival as an economic powerhouse.

The new supercomputers will not be in use until the end of the decade at the earliest, but they are increasingly being viewed as crucial investments for progress in science, advanced technologies and national security.

Once the exclusive territory of nuclear weapons designers and code breakers, ultrafast computers are increasingly being used in everyday product design. Procter & Gamble used a supercomputer to study the airflow over its Pringles potato chips to help stop them from fluttering off the company's assembly lines.

Today, driven by advances in so-called parallel computing — with software making it possible to lash together arrays of tens or even hundreds of thousands of processor

ranked among the 500 fastest machines, and recent reports in Chinese newspapers stressed the importance

computer designer who is a vice president at Chiaro Networks, a technology provider for high-performance computing. "That's where

and Chinese are planning new investments in breaking the petaflop computing barrier. A petaflop is a measure of computing performance

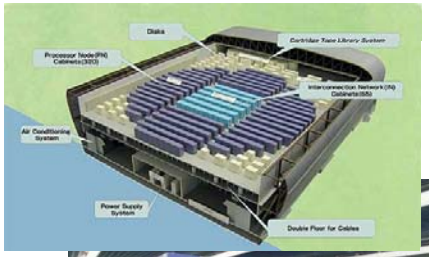
top personal computer. I.B.M. built the machine, Blue Gene/L, and plans to double its speed before the end of

Continued on Page 2

The New York Times

China joins U.S. and Japan in global race to build the fastest computer

- John Markoff, Aug 19, 2005

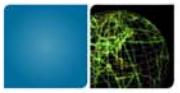


Japanese Earth Simulator (2002)

DARPA HPCS

NSF Leadership-Class System Acquisition - Creating a Petascale Computing Environment for Science and Engineering

U.S. Petascale (2008-2010)


Recent reports

- *NSF Workshop Report on Petascale Computing in the Biological Sciences*, David A. Bader, Allan Snavely, Gwen Jacobs, August 29-30, 2006, Arlington, VA.
- *Petascale Computing: Algorithms and Applications*, David A. Bader (ed.), Chapman & Hall/CRC Computational Science Series, © 2007. (ISBN: 9781584889090)

Georgia Tech and Petascale Computing

- 6th ranked academic institution in the June 2006 **Top100** List of most capable supercomputers in the world

- **Klaus Advanced Computing Building** (most advanced computing building in the world!) opened 26 October 2006

» Created a **Computational Science & Engineering** department in Fall 2005.

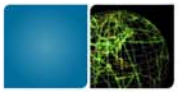
» **IBM** Shared University Research for Cell Broadband Engine

» **Cray XMT** consortium

- **Sun** Academic Excellence Grant for Sun Fire T2000 servers

» **Microsoft Research** Faculty Award for parallel programming of multicore processors

Sony-Toshiba-IBM Cell Center of Competence @ Georgia Tech

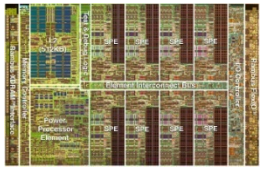

D TUESDAY, NOV. 14, 2006

The Atlanta Journal-Constitution

Business

ajc

CONTACT US: Mark Braykovich, Business editor / mbraykovich@ajc.com / 404-526-5869



"Georgia, not Austin, gets chip center," Bob Keefe, Austin American-Statesman, November 14, 2006

Ga. Tech lands research facility

By BOB KEEFE
bkeefe@ajc.com

Three of the biggest names in technology plan to announce today they will start a research center at Georgia Tech to explore ways to expand the reach of a promising new semiconductor design.

Sony Corp., IBM Corp. and Toshiba Corp. compare their new "Cell" microprocessor to a supercomputer on a chip that can handle some applications 10 times faster than traditional computer chips.

The technology that the companies jointly developed in Austin, Texas, over five years at a cost of $400 million is debuting in Sony's new PlayStation3 video game console. The $500 console went on sale in Japan last week and hits U.S. store shelves on Friday.

Now, the companies want to take the Cell technology much further.

With funding from the three Cell partners and additional money from Georgia Tech and outside grants, researchers at Tech's new STI Center of Competence will explore ways to adapt the technology for other industries, including biotech, finance and digital media creation.

Sony, Toshiba and IBM are providing an initial investment of $320,000, while Tech is putting in $230,000 and another $100,000 is coming from a National Science Foundation grant.

At the center, to be located in the school's new Christopher W. Klaus

➤ Please see RESEARCH, D6

SONY TOSHIBA IBM

Research: Tech wins center

➤ Continued from D1

Advanced Computing Building, researchers will also teach students and outside companies how to program computers and write software for the new type of chip. There will be four faculty members involved in the project.

Landing the center puts Georgia Tech at the forefront of a groundbreaking new type of semiconductor design. David Bader, executive director of the school's high-performance computing program, said he believes the center will be the only one of its kind in the United States.

"We really see this as the future of technology and innovation," Bader said. "This is so high-impact."

Austin bypassed

In picking Georgia Tech for the Center of Competence, the Cell partners sidestepped Austin as well as other high-tech hubs across the country. In addition to the University of Texas, more than a dozen schools around the country were vying to land the center, according to officials involved.

"Texas universities were absolutely part of the consideration," said Hina Shah, the Austin-based Cell development program director at IBM. But Georgia Tech won out in the end, she said, partly because its curriculum and areas of expertise matched up better with the interests of the three companies involved.

For Georgia Tech, the center is the latest in a series of big wins and increased prominence for the College of Computing.

In part, the school benefited from its extensive programs in high-performance computing, digital media and video game design.

But since the 2002 arrival as dean of Rich DeMillo, the former chief technical officer for Hewlett-Packard Co., the school has redesigned its curriculum to focus less on computer science theory and more on real-world applications.

"In many ways, we found them to be much more grounded about focusing on what's needed, not 10 years from now, but what's needed today and tomorrow," Shah said.

"That made a huge difference."

Mass Chatani, Sony's senior general manager for Cell development, said in a statement that the "collaboration with the College of Computing at Georgia Tech will create innovative applications for Cell processors."

Georgia Tech computing director **David Bader** believes the center will be the only one of its kind in the country.

The Cell chip design is only in its infancy and has a lot to prove. The chip isn't expected to make a big dent in the traditional semiconductor market controlled by Intel Corp. and Advanced Micro Devices Inc. anytime soon. Reaching into other markets won't be easy either.

Still, what makes Cell so promising is its potential power, especially when it comes to graphics-intensive programs like video games, broadband Internet video processing and other digital media applications.

Just recently, Intel released its first "dual core" and "quad core" microprocessors that essentially put two or four processors on one chip.

Cell chips have already leapfrogged that capability. The chips in Sony's PlayStation3, for instance, essentially have nine cores – eight unique sub-processors that work in connection with a central processor.

16 cores a possibility

Future Cell designs could have as many as 16 sub-processing cores, which could dramatically increase the speed and the number of applications Cell-equipped computers could handle.

"This really is a new era in performance," Jim Kahle, an IBM fellow who oversaw the chip's design in Austin, said in announcing the first Cell chips in San Francisco last year.

Sony has the most riding on Cell. The Japanese giant is counting on the chip to help it regain ground in new technology development that it lost in areas like digital music.

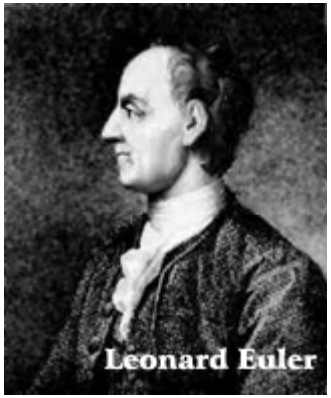
Along with its video game machines, Sony is exploring putting Cell processors into a wide array of products, including personal computers, televisions and mobile phones.

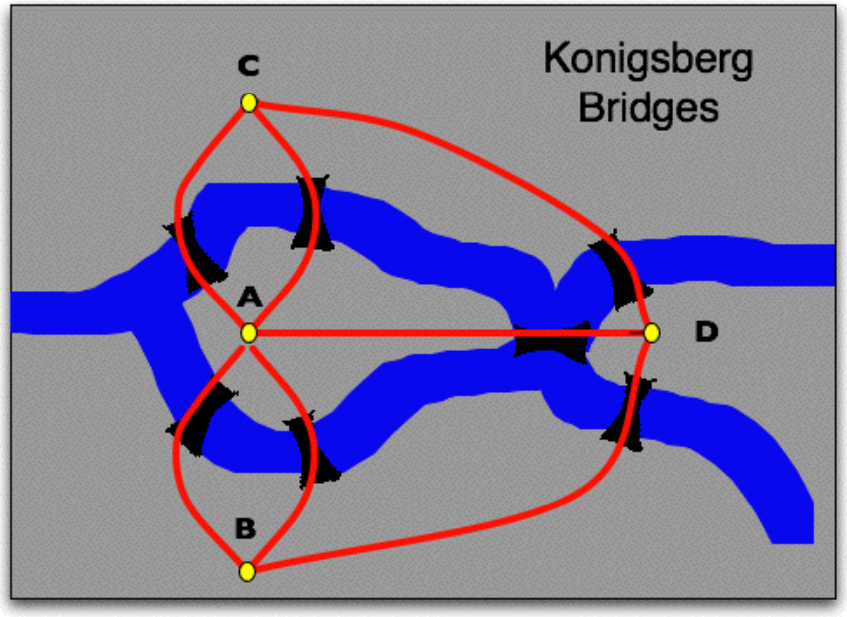
Toshiba plans to use Cell processors in its TV sets and in other products.

IBM already has introduced powerful computer servers based on the design.

“Practical” Graph Theory

- In Königsberg, a river ran through the city such that in its center was an island, and after passing the island, the river broke into two parts. Seven bridges were built so that the people of the city could get from one part to another.
- The people wondered whether or not one could walk around the city in a way that would involve crossing each bridge exactly once.
- Leonhard Euler, circa 1735

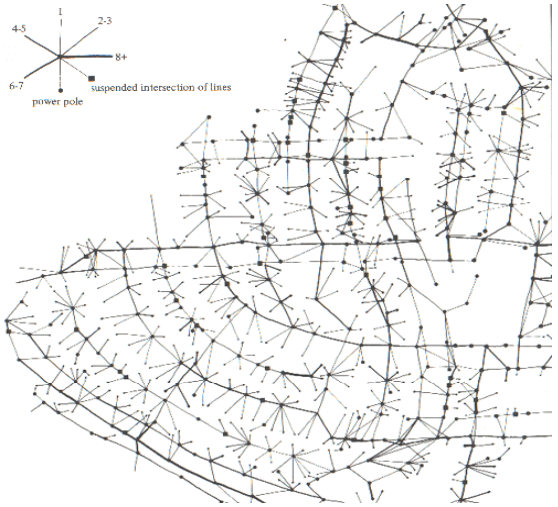




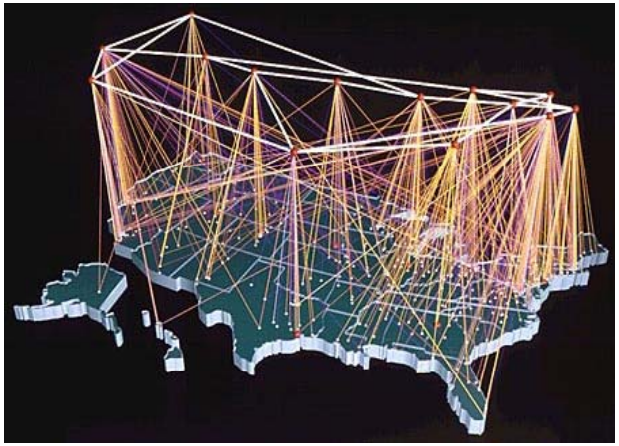
Source: The Math Forum

Graph problems arise from a variety of sources

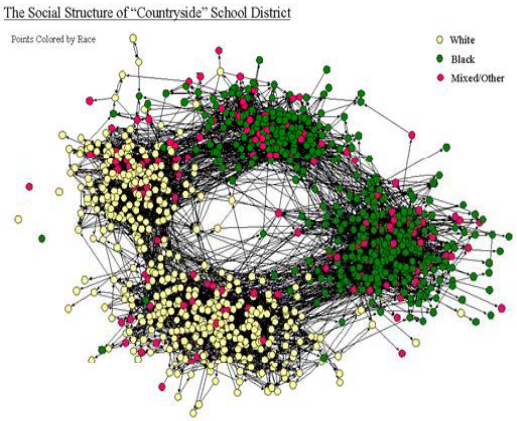
Power Distribution Networks



Internet backbone



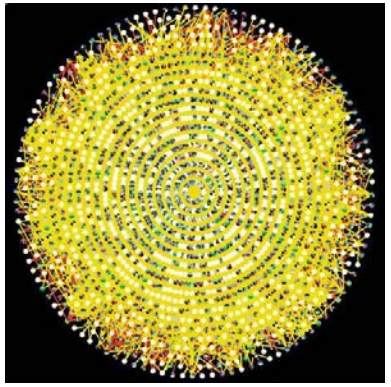
Social Networks



Graphs are everywhere!

Ground Transportation

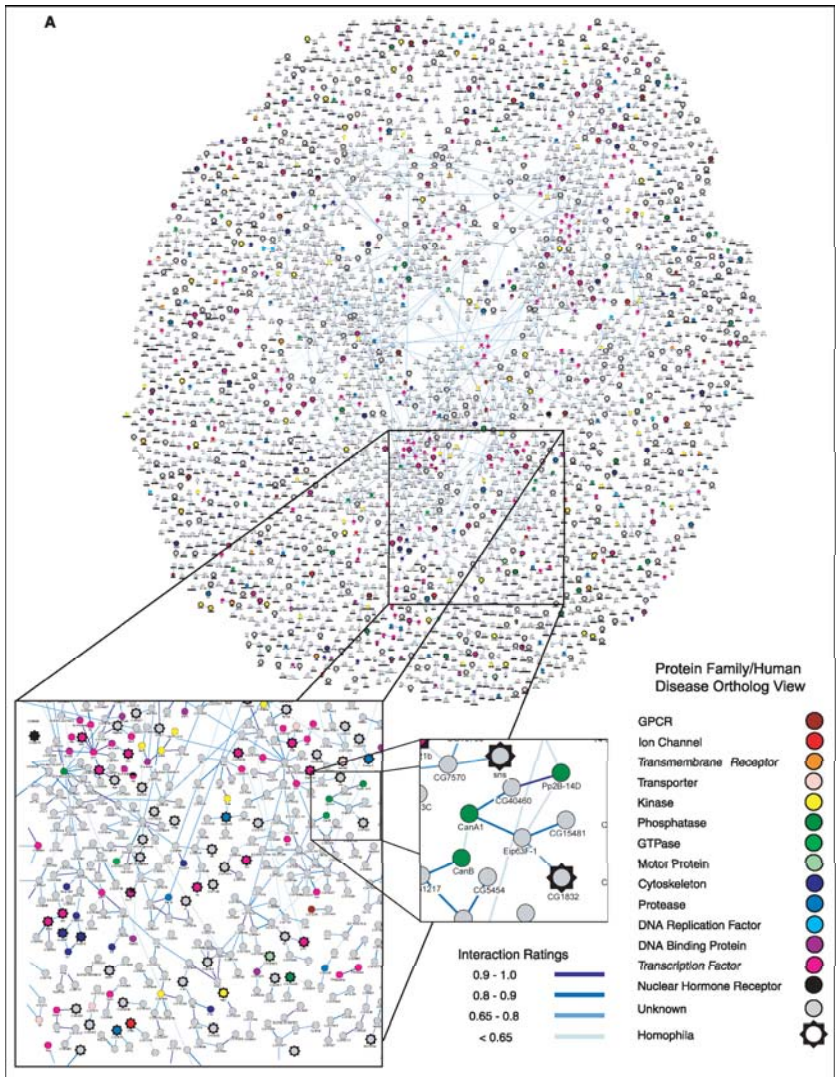
Tree of Life

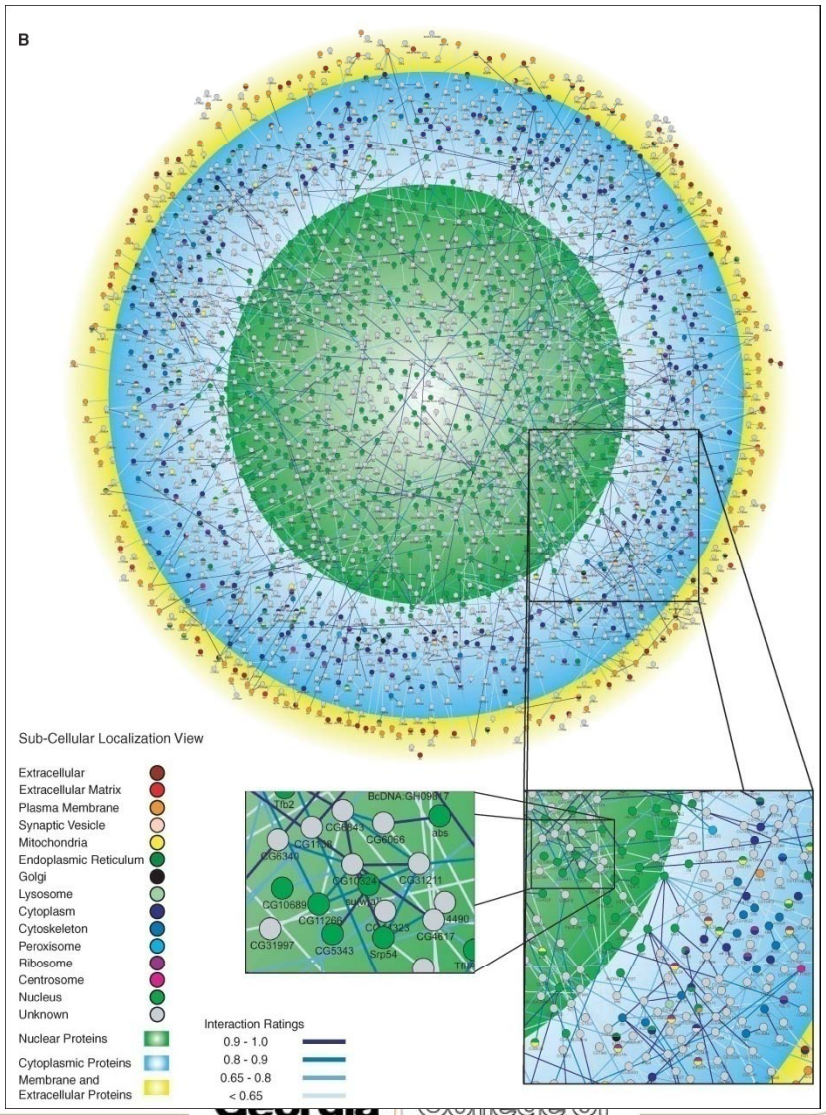


Protein-interaction networks

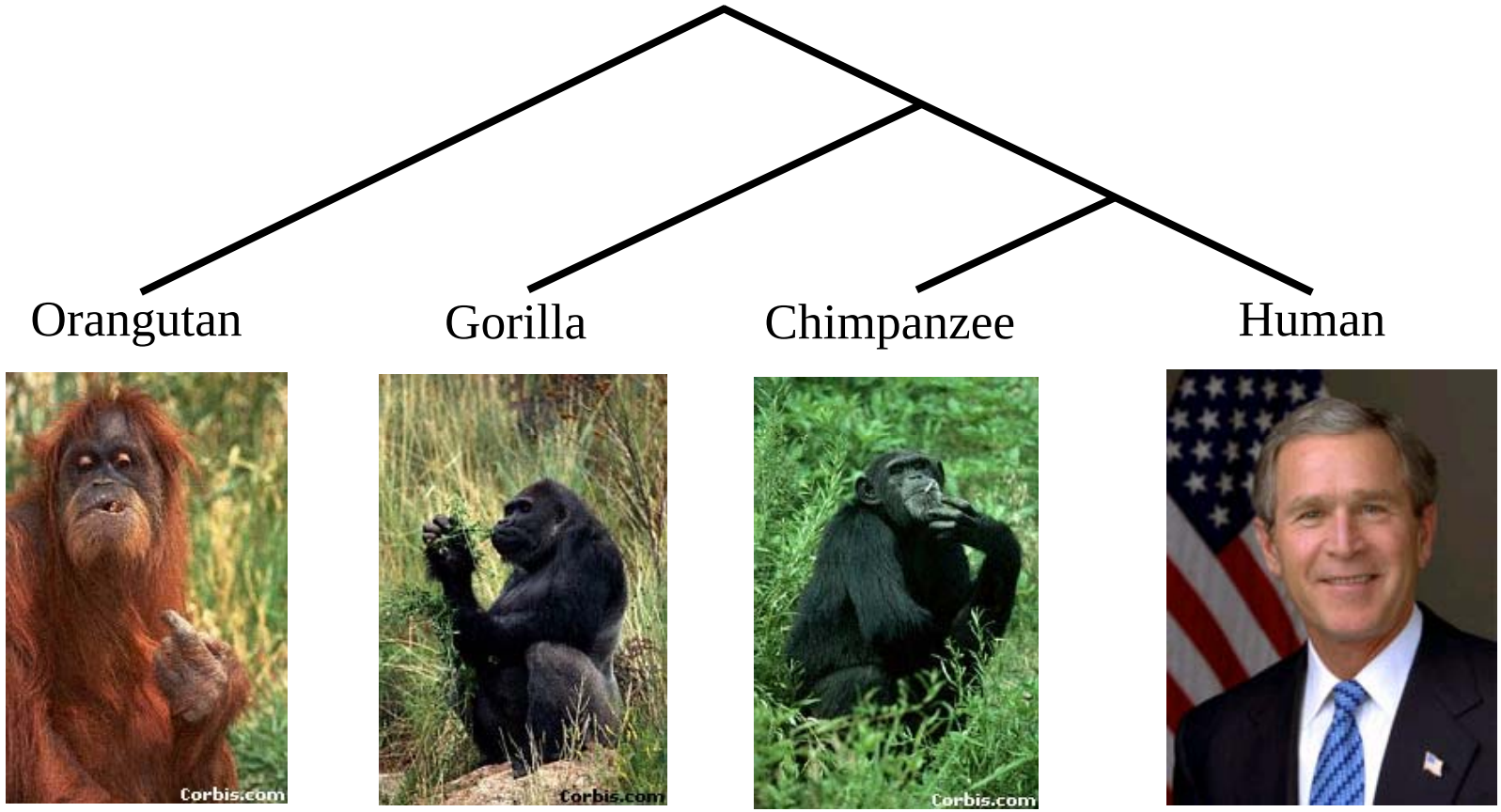
Sources: C. Faloutsos talk [IPAM 05]

Giot L, Bader JS, …, Rothberg JM,
A protein interaction map of *Drosophila melanogaster*
Science 302: 1727-1736, 2003.

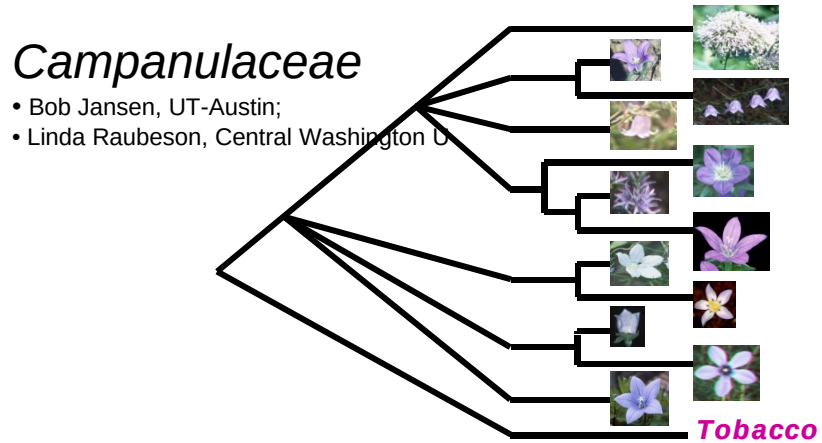


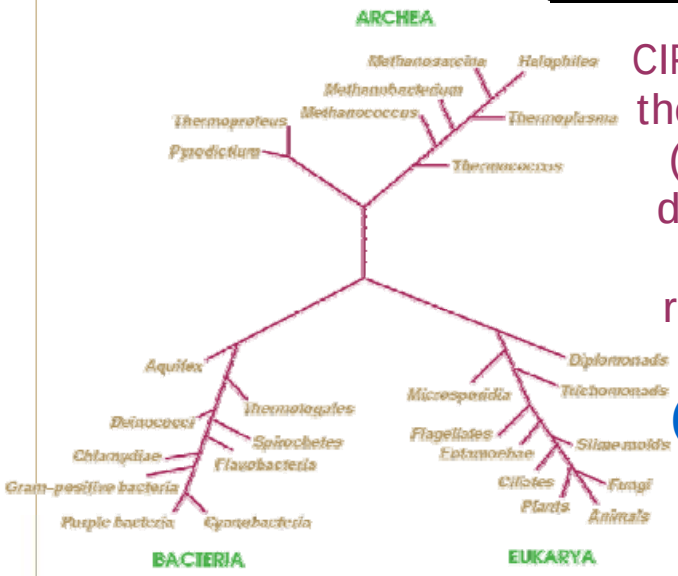


Phylogeny


Orangutan
Gorilla
Chimpanzee
Human
Corbis.com
Corbis.com
Corbis.com

Computational Phylogeny
GRAPPA





The Tree of Life

CIPRES aims to establish the cyber infrastructure (platform, software, database) required to attempt a reconstruction of the Tree of Life **(10-100M organisms)**

- 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

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.

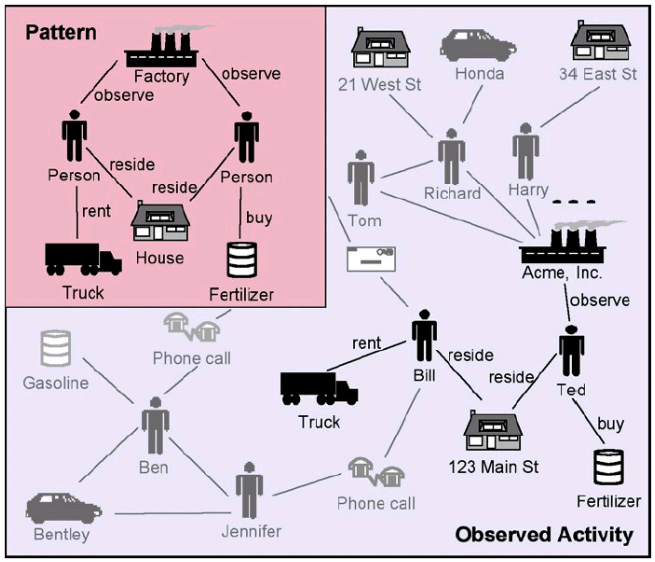


Figure Credit: *Graph-based technologies for intelligence analysis,*T. Coffman, S. Greenblatt, S. Marcus, Commun. ACM, 47(3):45-47, 2004.

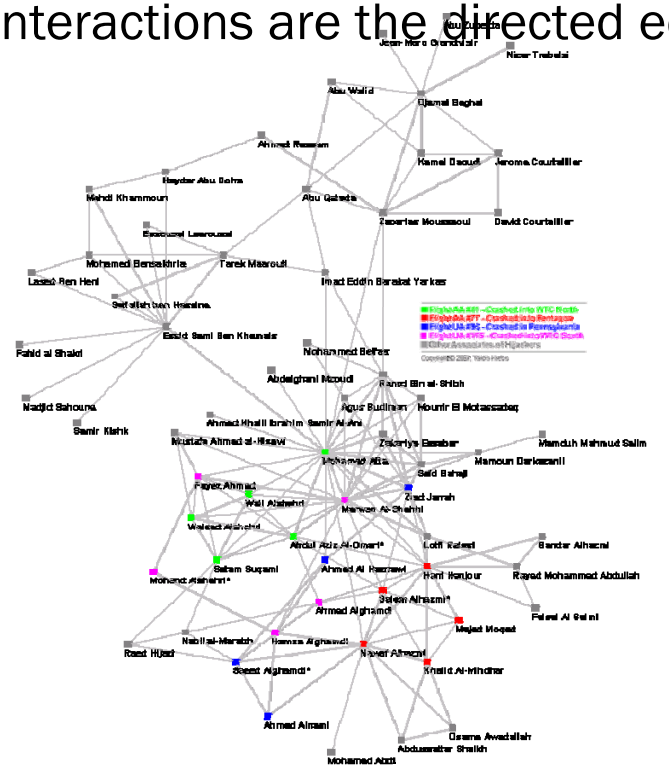
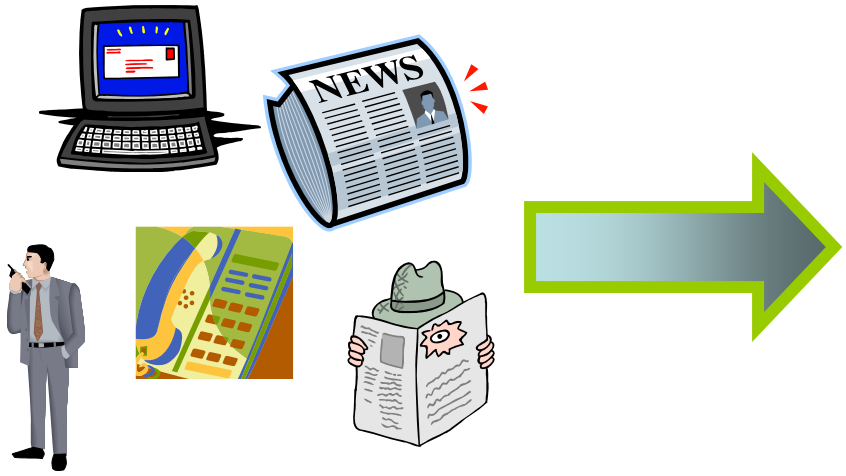
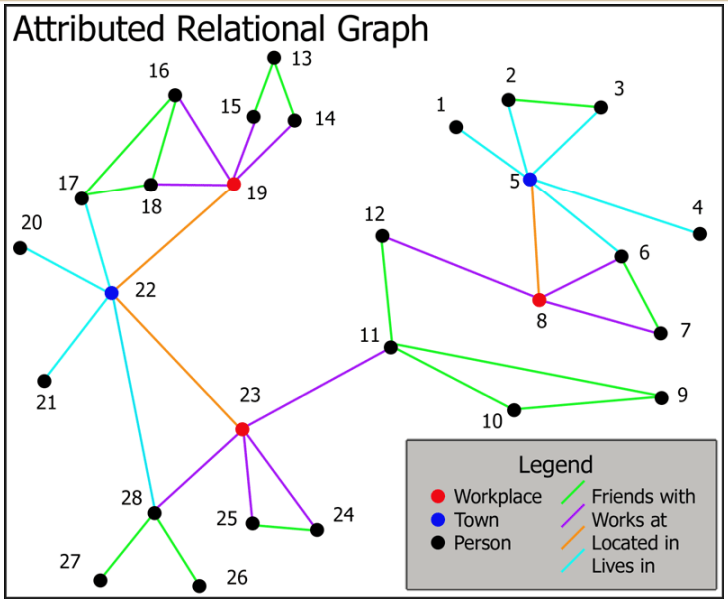


Figure Credit: *Uncloaking Terrorist Networks*, V.E. Krebs, First Monday, 7(4), April 2002.

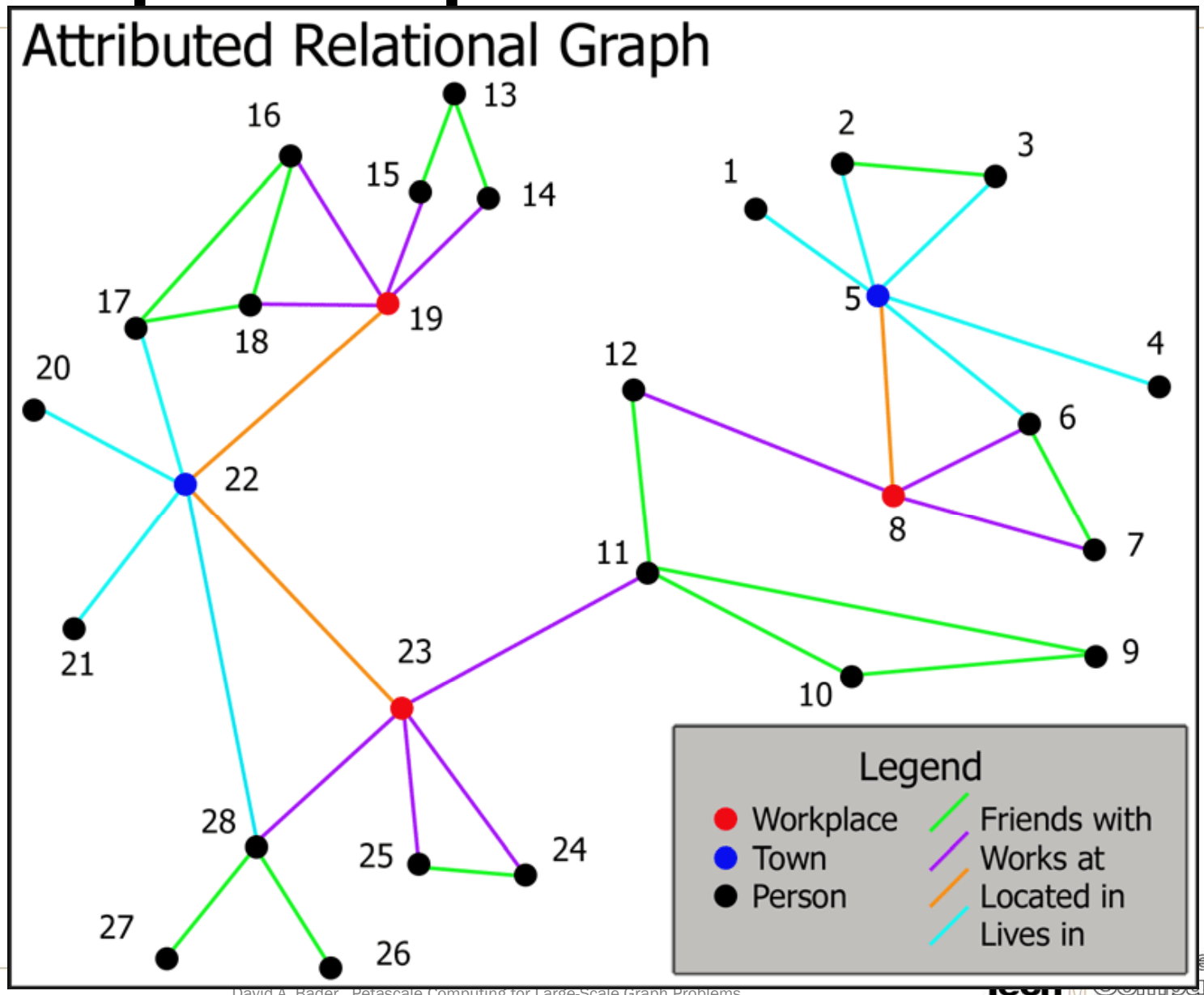
Information Overload

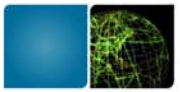




- *Challenge*: Piecing the data together and extracting critical, relevant information in a timely manner
- Semantic Graphs (or Attributed Relational Graphs) are one way to integrate data from disparate sources
  - Vertices represent people, places, locations, events, etc.
  - Edges represent the relationships between the vertices
  - Semantic graph encodes web of relationships

Simple Example

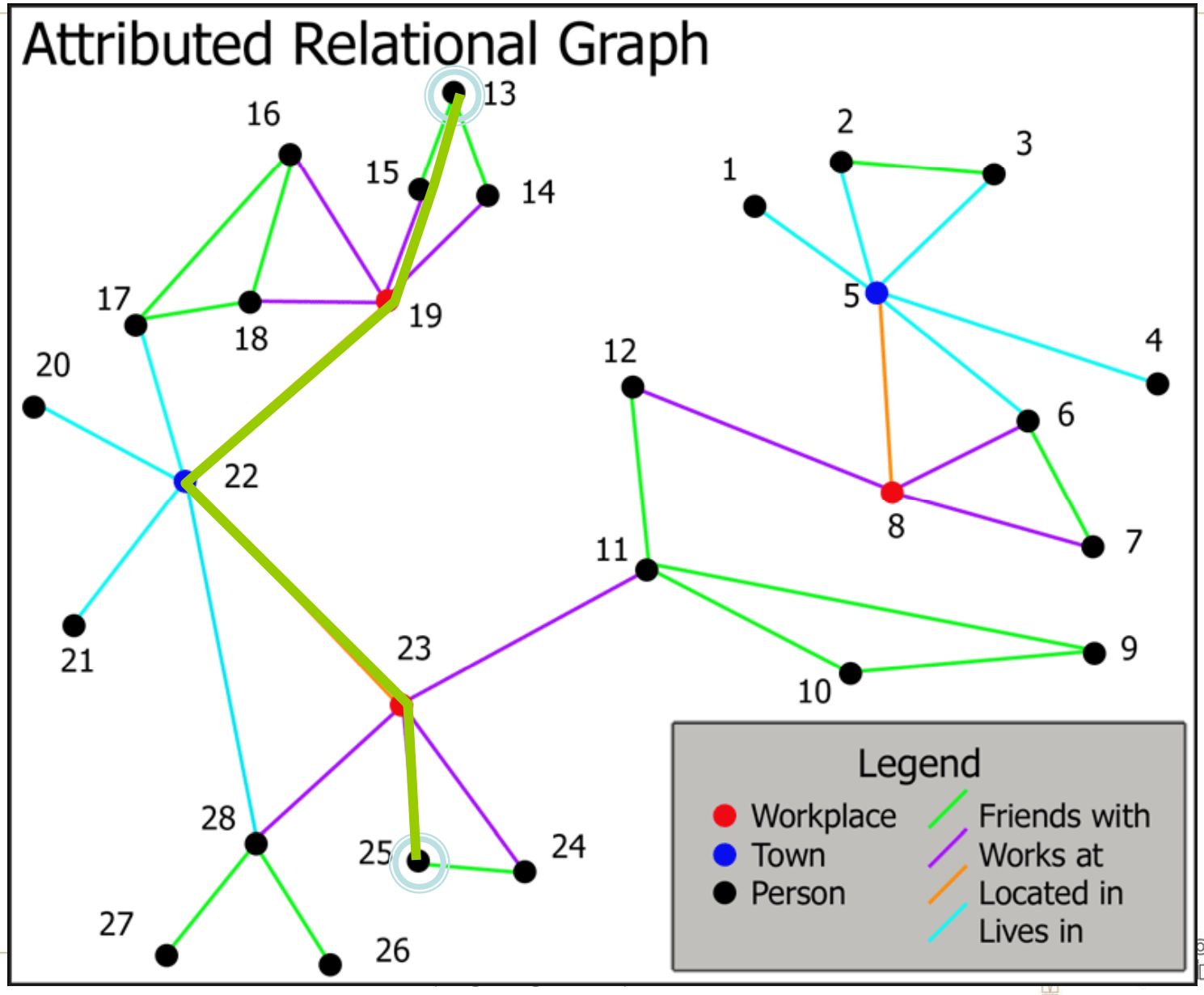



Advantages of Semantic Graphs

- Much smaller than raw data. Can fit in memory of large computer
  - Fast response to queries
  - Pre-join of database
- Combine data from different sources and of different types
- Some common intelligence and law enforcement queries are naturally posed on graphs
  - Particularly for the terrorist threat

Query Example I: Short Paths



Query Example II: Motif Finding

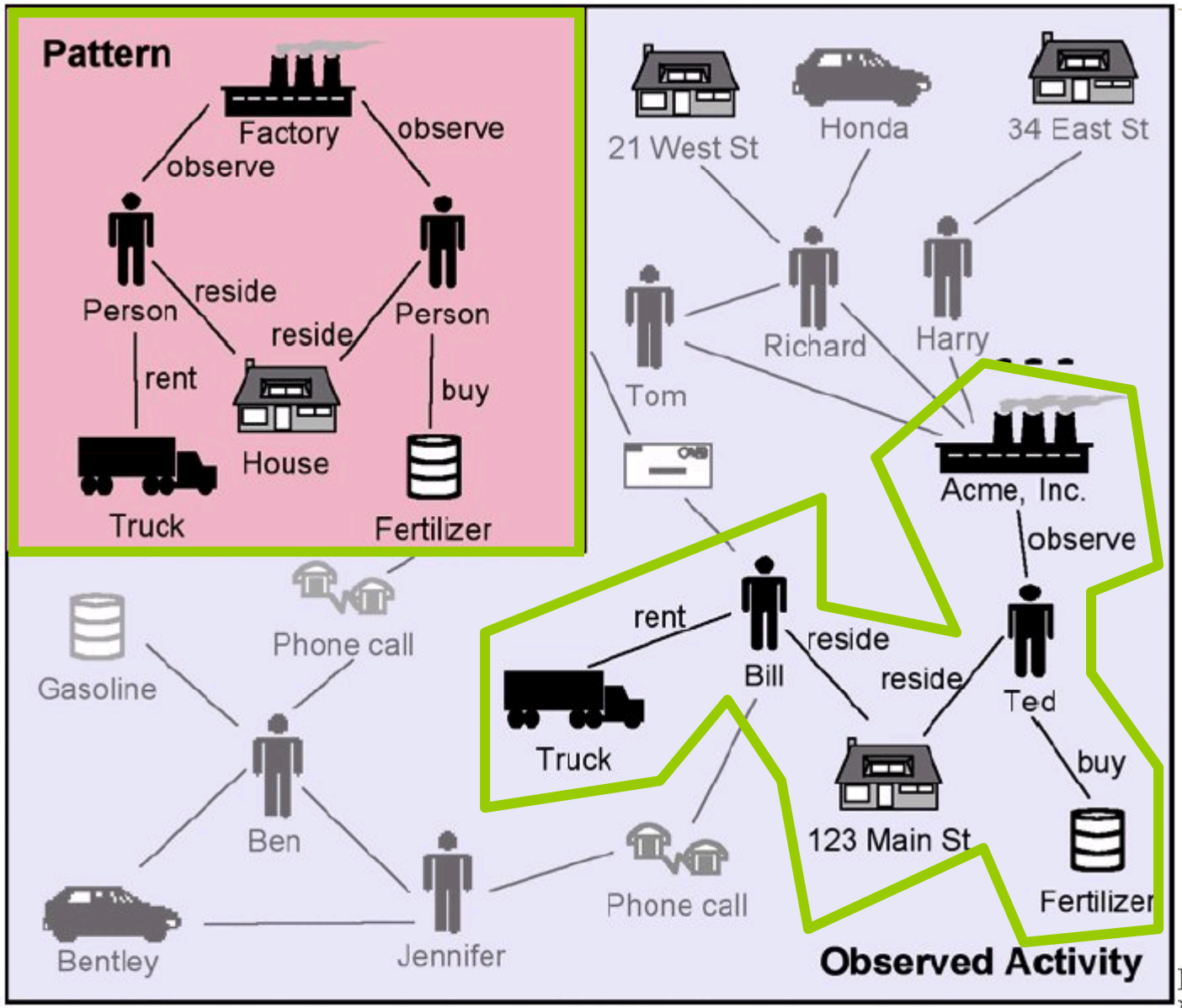
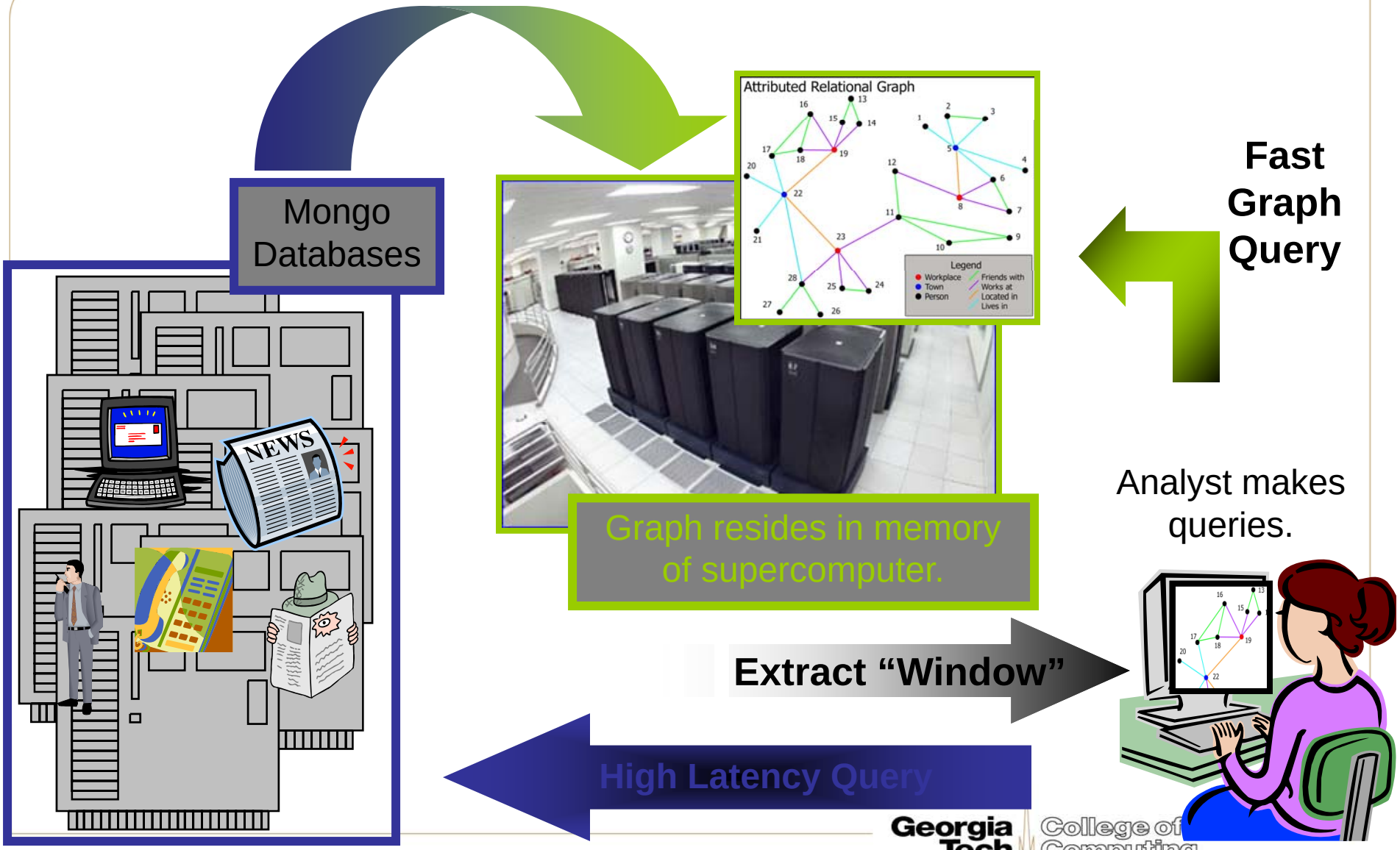
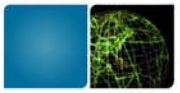


Image Source:
T. Coffman,
S. Greenblatt,
S. Marcus,
Graph-based technologies for intelligence analysis,
CACM, 47
(3, March 2004):
pp 45-47

The Big Picture



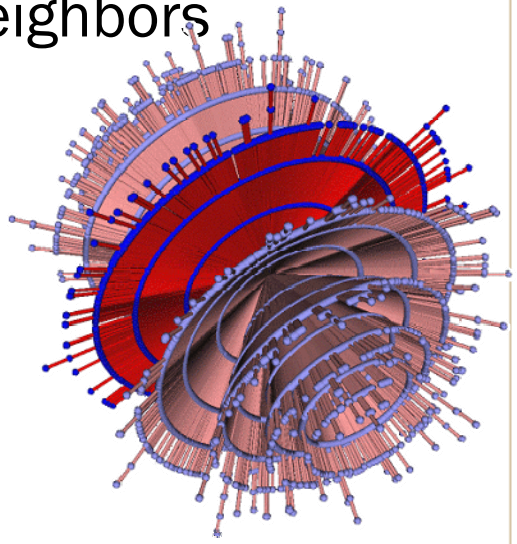




Graph algorithms

- Driving applications are not traditional HPC:
  - health care, proteomics, security, informatics, …
- Fundamental abstraction
  - Standard introductory material covered in a computer science course on data structures and algorithms, but…
- There are few (if any) efficient distributed-memory parallel implementations of even the simplest algorithm for sparse, arbitrary graphs!

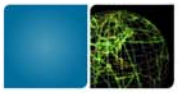
Informatics Graphs are Tough

- Very different from graphs in scientific computing!
  - Graphs can be enormous
  - Power-law distribution of the number of neighbors
  - Small world property – no long paths
  - Very limited locality, not partitionable
  - Highly unstructured
  - Edges and vertices have types



Six degrees of Kevin Bacon
Source: Seokhee Hong

- Experience in scientific computing applications provides only limited insight.

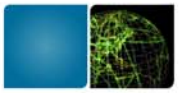

Architecture

- 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

- Desired Features:
  - 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
  - One machine with these properties is the Cray MTA-2
    - And its successor, the Cray XMT ("Eldorado")

How Does the MTA Work?

- Latency tolerance via massive multi-threading
  - Each processor has hardware support for 128 threads
  - Context switch in a single tick
  - Global address space, hashed to reduce hot-spots
  - No cache or local memory. Context switch on memory request.
  - Multiple outstanding loads
- Remote memory request does not stall processor
  - Other streams work while your request gets fulfilled
- Light-weight, word-level synchronization
  - Minimizes access conflicts
- Flexibly supports dynamic load balancing
- Notes:
  - MTA-2 is 7 years old
  - Clock rate is 220 MHz
  - Largest machine is 40 processors

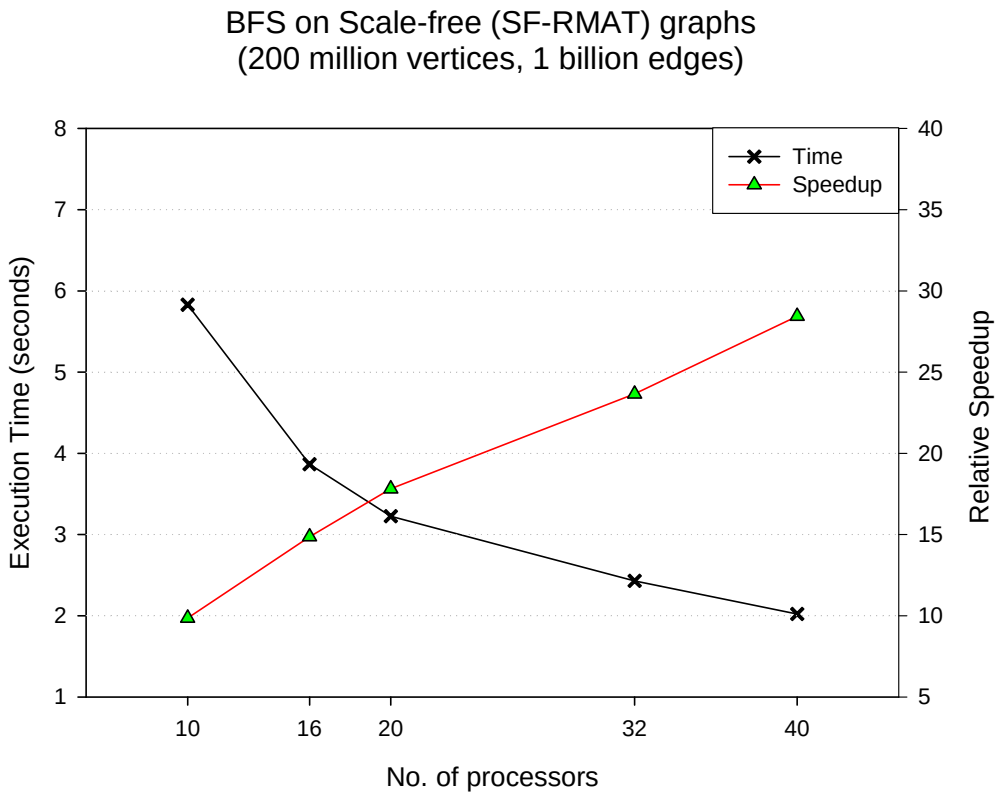

Our recent work on Multithreaded Algs.

- List ranking and connected components.
  - List ranking runs 40 times faster
  - Connected components runs 6 times faster
  - on 220MHz Cray MTA-2 processors compared with a commodity 400MHz Sun SMP.
  - [Bader, Cong, Feo; ICPP 2005]

- Graph theory applications
  - Parallel breadth-first search; approximate clique extraction; DARPA SSCA2 [Bader, Madduri, Feo, in progress]
  - st-connectivity [Bader, Madduri; ICPP 2006]
  - Betweenness Centrality [Bader, Madduri; ICPP 2006]
  - Advanced Shortest Paths [Croback, Berry, Madduri, Bader; MTAAP 2007]

Case Study 1: Breadth-First Search (BFS)

- Sequential BFS: $O(m+n)$ using a FIFO queue
- Recent algorithms and implementations for handling large-scale graphs:
  - graph partitioning [Yoo et. al. 2005]
  - external memory [Meyer et. al. 2006]
- Our design is a fine-grained algorithm, suited for multithreaded architectures
  - All vertices at a given *level* in the graph can be processed simultaneously, instead of just picking the vertex at the head of the queue
  - The adjacencies of each vertex can be inspected in parallel

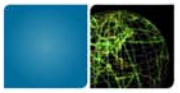


Large-Scale Graph Results: Breadth-First Search



Problem	Graph Instance	Result	Comments
Multithreaded (OUR RESULT)	Random graph $n=2^{28}$ vertices $m=2^{30}$ edges	2.3 s (p=40) 73.9 s (p=1) Cray MTA-2	Works well for sparse real-world graphs
External Memory [Ajwani et al., 2006]	Random graph $n=2^{28}$ vertices $m=2^{30}$ edges	8.9 HOURS (MM_BFS_R)	State-of-the-art external memory BFS
Multithreaded (OUR RESULT)	Random graph Scale-free graph n=400 M vertices m=2 B edges	4.53 s 5.2 s (p=40) Cray MTA-2	Largest arbitrary BFS known results
Distributed Memory [Yoo et al., 2005]	Random graphs n=3B vertices m=32B edges	4.7 sec on p=32K IBM BG/L	Works only for Erdos-Renyi random graphs.

WORKS ONLY FOR SYNTHETIC GRAPHS

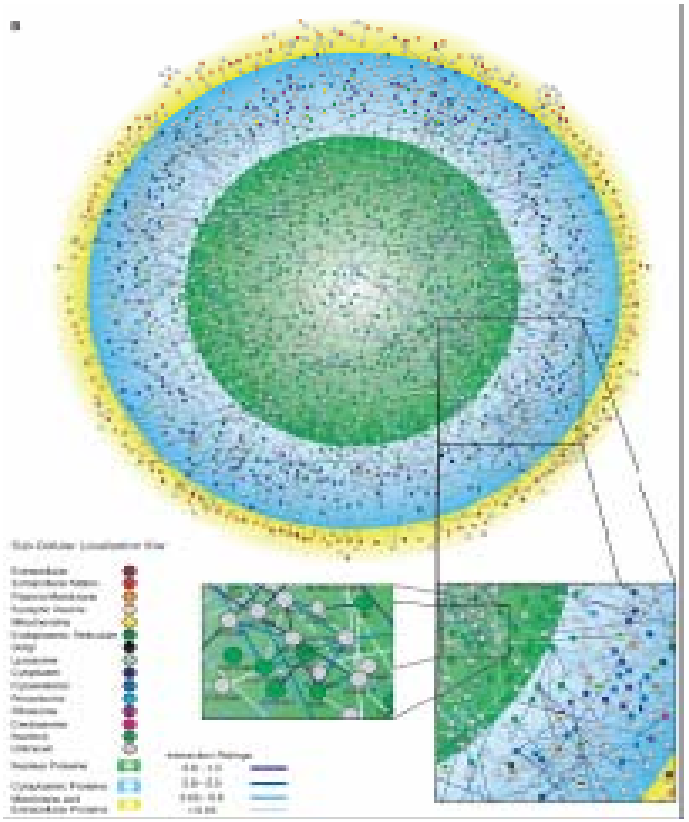




Case Study 2: Social Network Analysis

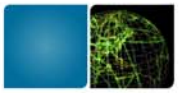
- Centrality metrics: Quantitative measures to capture the importance of a node/vertex/actor in a graph
  - Degree, Closeness, Stress, Betweenness
- Identifying *central* nodes in large complex networks is the key metric in a number of applications:
  - Biological networks, protein-protein interactions
  - Sexual networks and AIDS
  - Identifying key actors in terrorist networks
  - Organizational behavior
  - Supply chain management
  - Transportation networks

Biological Complex Networks

- Protein-interaction networks (PINs), signal transduction networks, biological pathways, food-webs
- PIN analysis: Novel protein function prediction, identification of critical nodes
- High-throughput experimental techniques → lot of biological data
- Protein-interaction datasets are available for yeast (high-confidence), human, fly



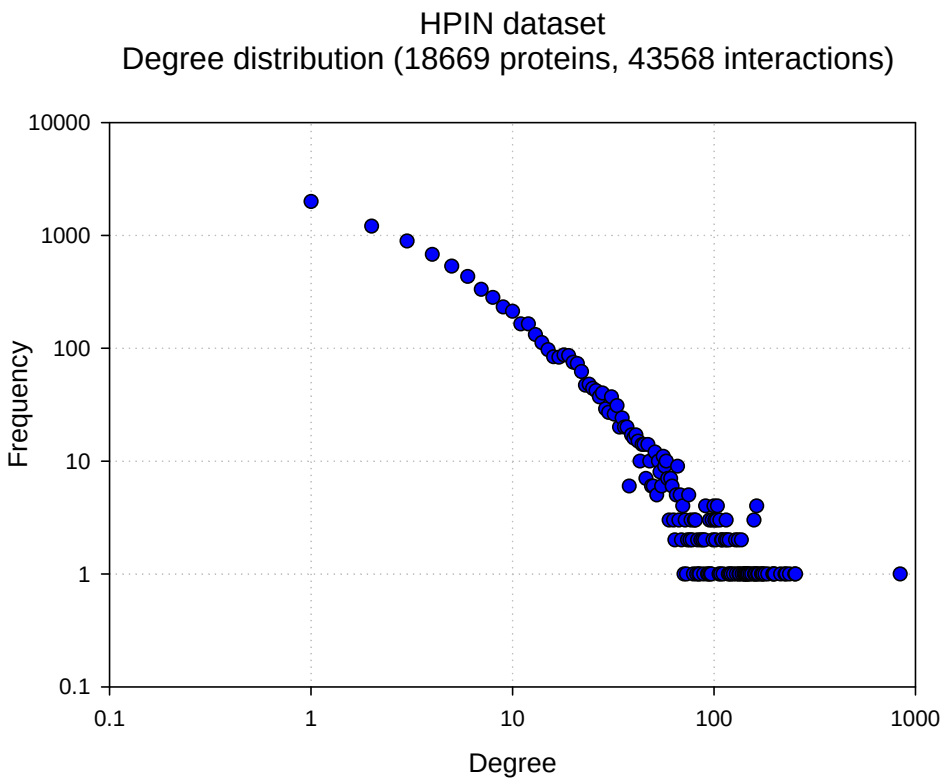
Giot L, **Bader JS**, …, Rothberg JM,
A protein interaction map of *Drosophila melanogaster*
Science 302: 1727-1736, 2003.


Our Contributions

- Graph-theoretic analysis of the Human protein interaction network, comprising nearly *18,000 proteins and 44,000 interactions*.
  - [Bader,Madduri; HiCOMB 2007]
- Parallel algorithms for analysis of large-scale interaction networks
  - [Bader, Madduri; ICPP 2006]
- Comparison of the yeast and human protein interaction networks, over time.

Human PIN: Degree Distribution

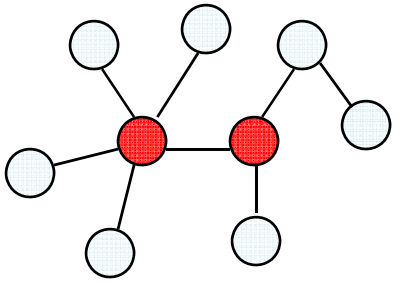


- The degree distribution is similar to the yeast protein interaction network (unbalanced, few high-degree proteins)
- Power-law graph models can mimic the degree distribution of PINs
- Protein with the highest degree: *Solute carrier family 2 member 4*, Gene Symbol SLC2A4, HPRD ID 00688, biological function: transport

Betweenness Centrality (BC)

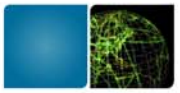
- Key metric in social network analysis
  [Freeman '77, Goh '02, Newman '03, Brandes '03]

$$BC(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$



- σ_{st} : Number of shortest paths between vertices s and *t*
- $\sigma_{st}(v)$: Number of shortest paths between vertices s and *t* passing through *v*

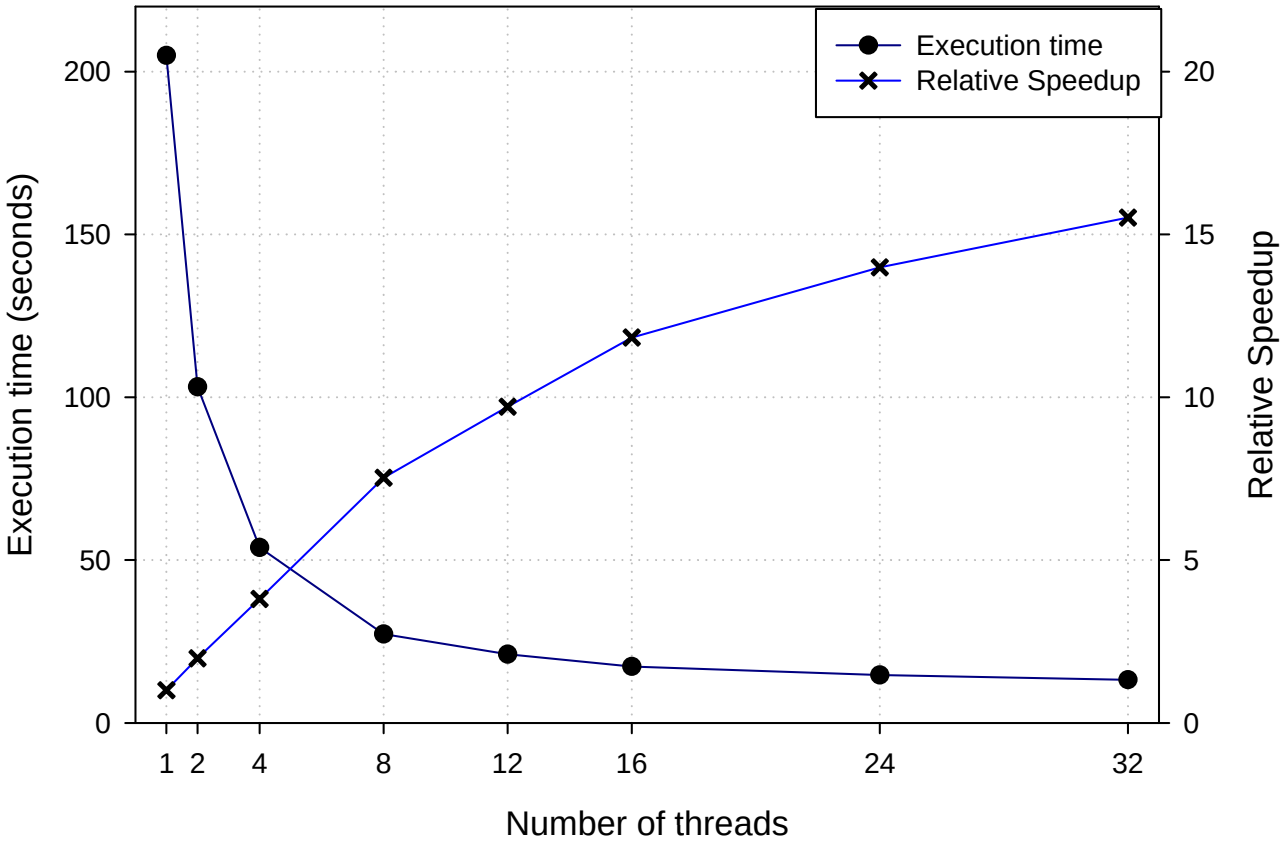
- Exact BC is compute-intensive


BC Algorithms

- Brandes [2003] proposed a faster sequential algorithm for BC on sparse graphs
  - $O(mn + n^2 \log n)$ time and $O(n)$ space for weighted graphs
  - $O(mn)$ time for unweighted graphs
- We designed and implemented the first parallel algorithm:
  - [Bader, Madduri; ICPP 2006]

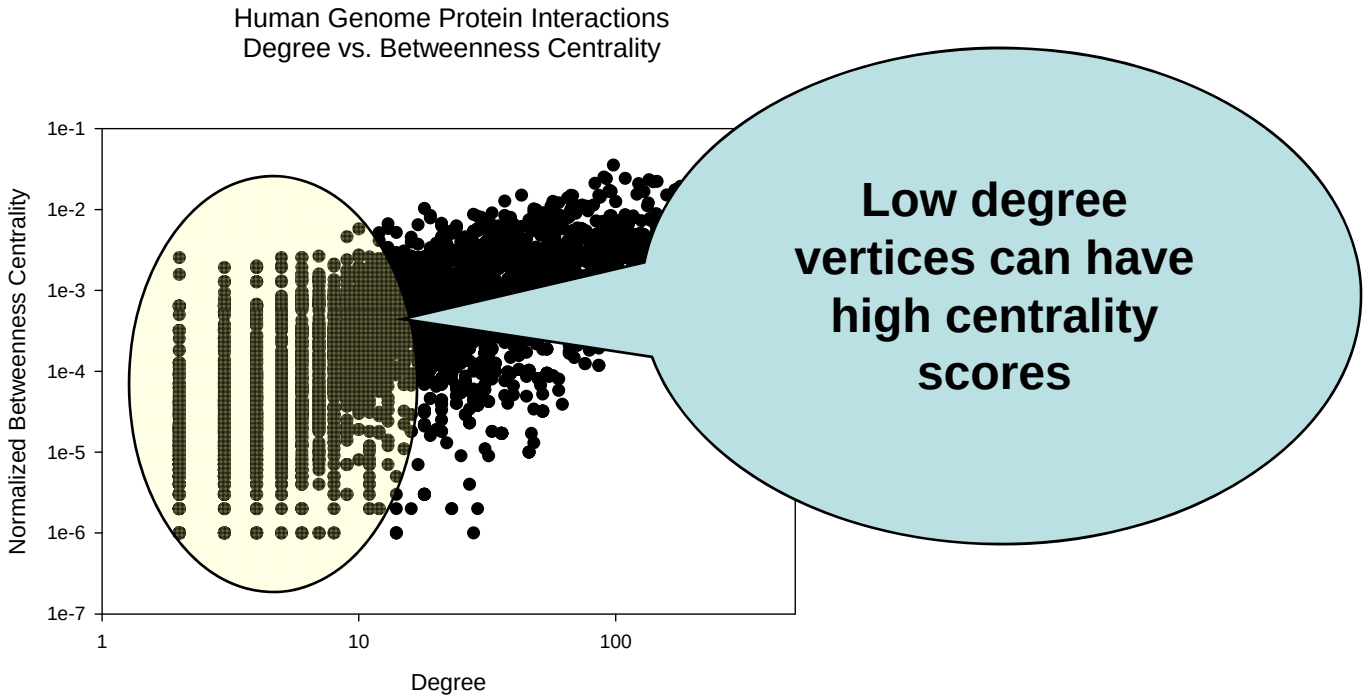
BC Computation: Parallel Performance



BC Analysis: Protein-protein interactions

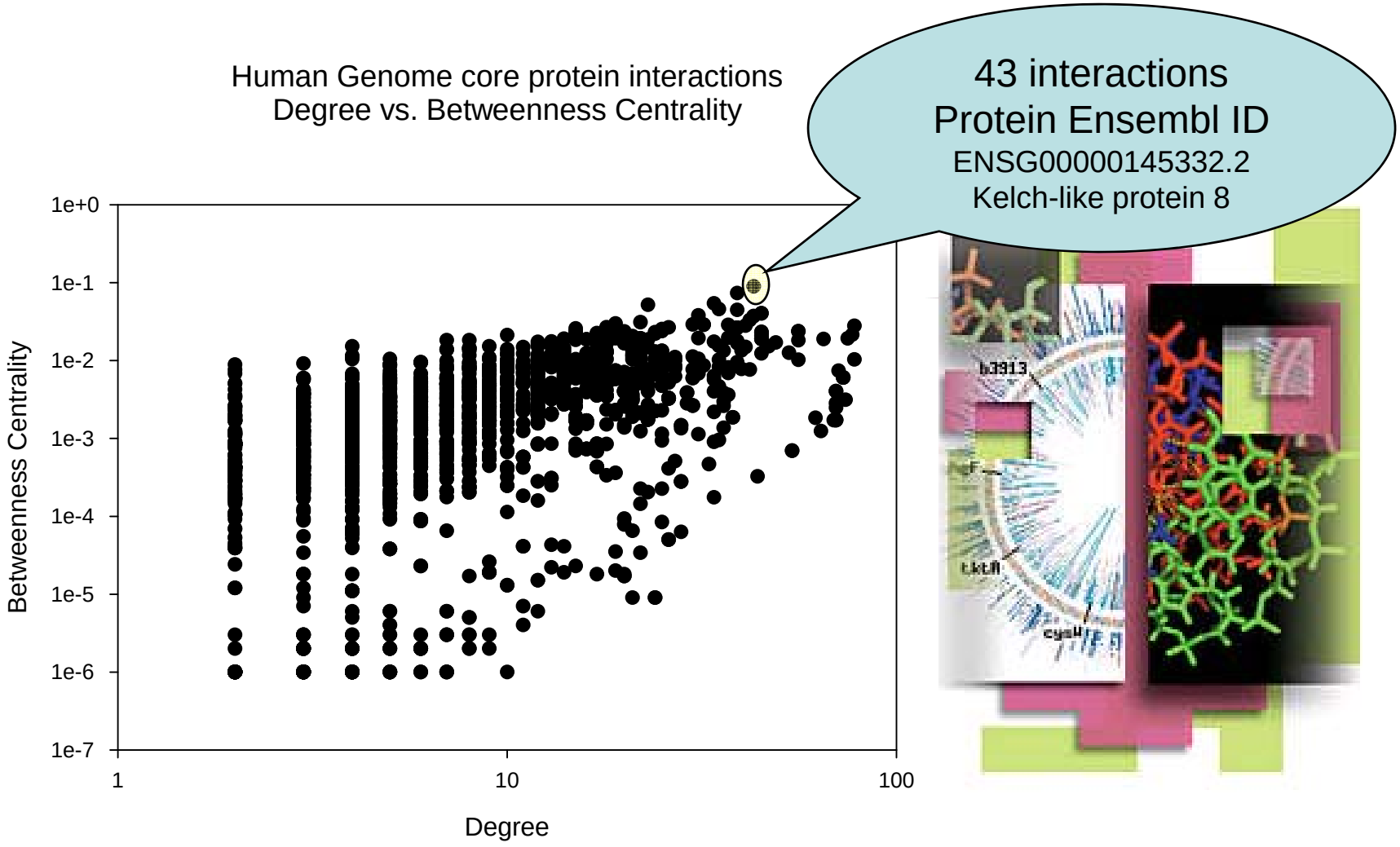


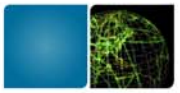
- We recently computed betweenness centrality scores for the human genome[1] protein interaction network
  - [Bader, Madduri; HiCOMB 2007]



[1] Lehner, Fraser. A first draft human protein interaction map, http://genomebiology.com/2004/5/9/R63

BC Analysis: Protein-protein interactions




BC Implementation Details

- We have designed and implemented parallel betweenness centrality for two shared memory platforms:
  - Symmetrical multiprocessors (SMPs)
    - Modest number of processors
    - Coarse-grained implementation, BFS/SSSP computations are done concurrently
    - Implemented on IBM p570
  - multithreaded architectures
    - Thousands of hardware threads
    - Individual BFS/SSSP computation is parallelized
    - Implemented on Cray MTA-2

IBM p5 570

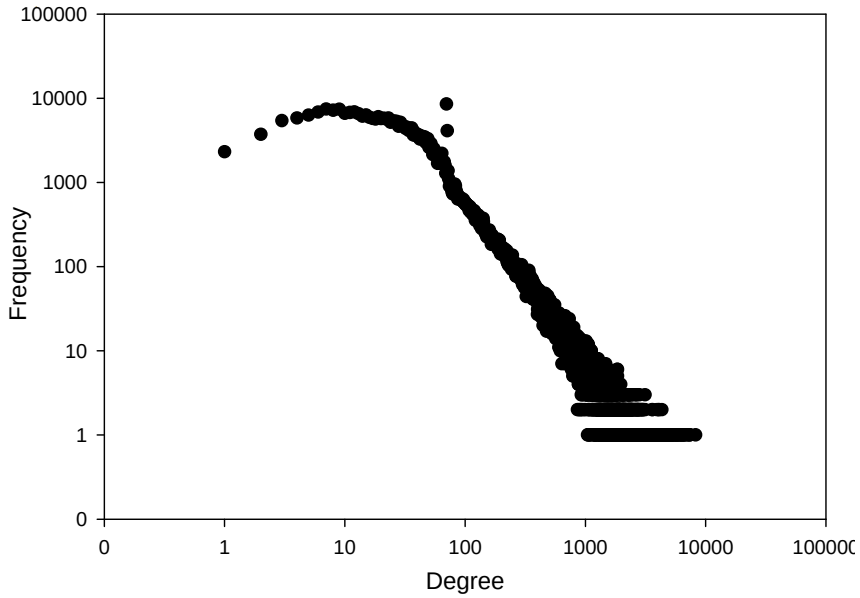
- 16-way Power5 symmetric multiprocessor
- 1.9 GHz processor
- 256 GB physical memory
- 32KB L1D, 1.9MB L2, 32MB L3
- 8-way superscalar
- SMT on each core

- Supports a C and POSIX threads parallel implementation

BC for IMDB movie actor network

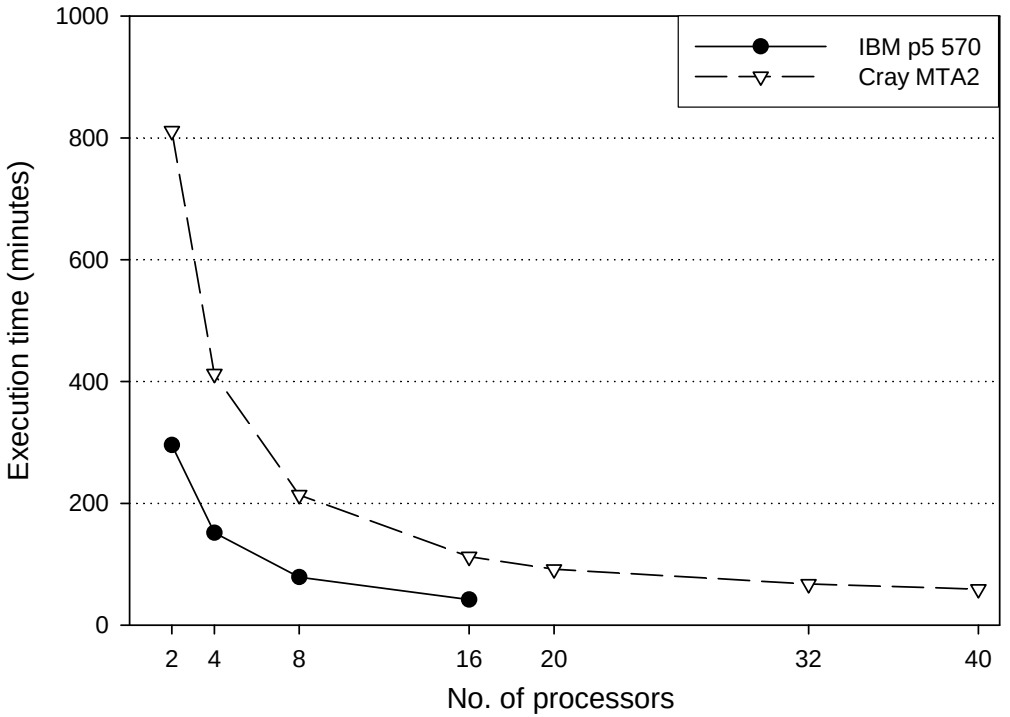
Real-world instance: an undirected graph of 392,400 vertices (movie actors) and 31,788,592 edges. An edge corresponds to a link between two actors, if they have acted together in a movie. The dataset includes actor listings from 127,823 movies.

ND-actor : IMDB movie-actor network
(392,400 vertices and 31,788,592 edges)

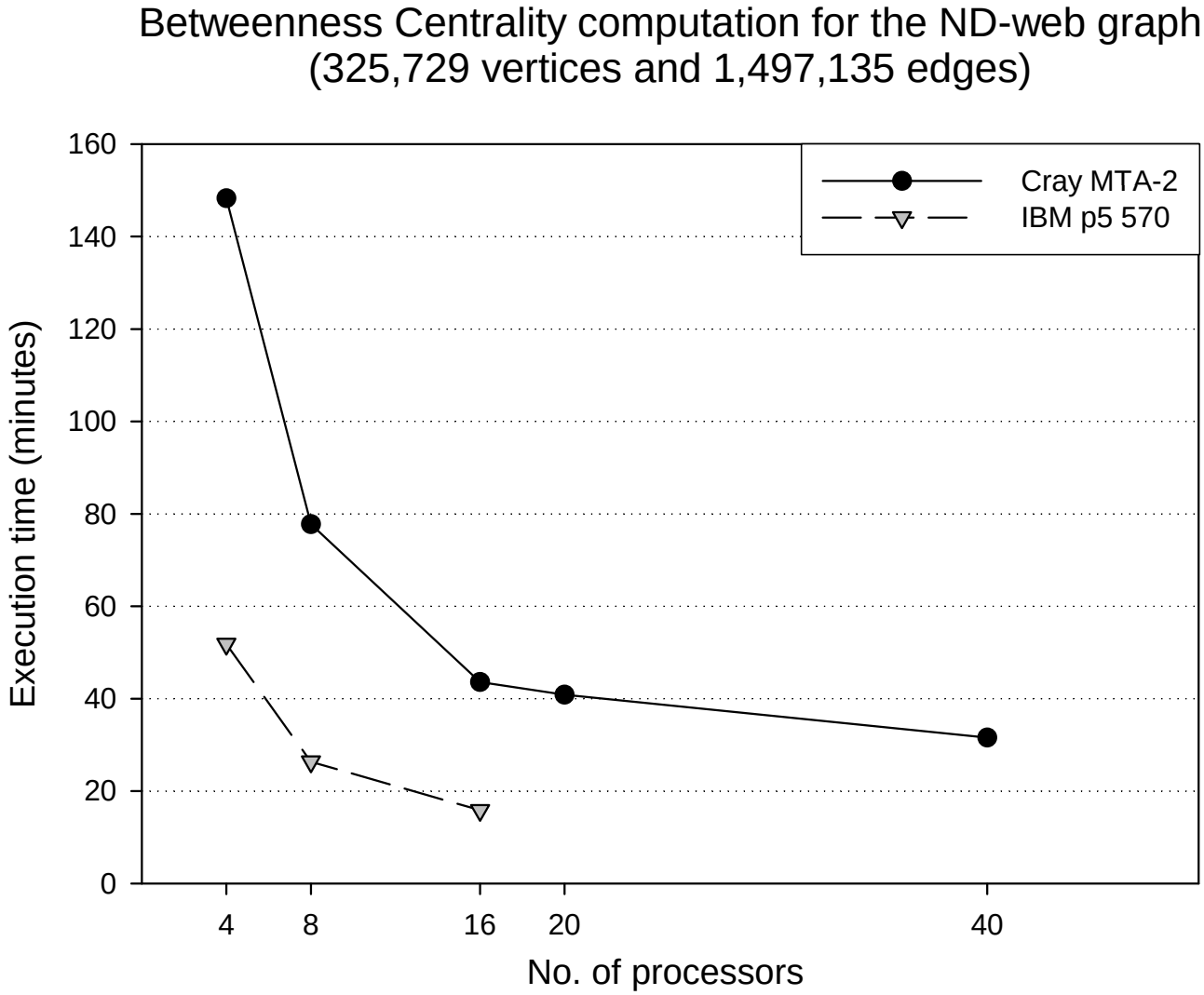


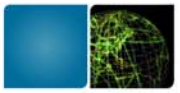
Degree Distribution: Scale-free

Betweenness Centrality computation for the ND-actor graph
(392,400 vertices and 31,788,592 edges)



BC for web graph



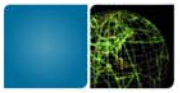




Collaborators

- Kamesh Madduri (Georgia Tech)
- Bruce Hendrickson (Sandia National Laboratories)
- Jon Berry (Sandia National Laboratories)
- Vipin Sachdeva (IBM Austin Research Lab)
- Guojing Cong (IBM TJ Watson Research Center)
- John Feo (Microsoft)

Acknowledgment of Support

- National Science Foundation
  - **CSR:** A Framework for Optimizing Scientific Applications (06-14915)
  - **CAREER**: High-Performance Algorithms for Scientific Applications (06-11589; 00-93039)
  - **ITR**: Building the Tree of Life -- A National Resource for Phyloinformatics and Computational Phylogenetics (EF/BIO 03-31654)
  - **ITR/AP:** Reconstructing Complex Evolutionary Histories (01-21377)
  - **DEB** Comparative Chloroplast Genomics: Integrating Computational Methods, Molecular Evolution, and Phylogeny (01-20709)
  - **ITR/AP(DEB):** Computing Optimal Phylogenetic Trees under Genome Rearrangement Metrics (01-13095)
  - **DBI:** Acquisition of a High Performance Shared-Memory Computer for Computational Science and Engineering (04-20513).
- IBM PERCS / DARPA High Productivity Computing Systems (HPCS)
  - DARPA Contract NBCH30390004
- IBM Shared University Research (SUR) Grant
- Sony-Toshiba-IBM (STI)
- Microsoft Research
- Sun Academic Excellence Grant


Conclusions

- "High-Performance Computing" needs to move from FP-centric to data-centric computing
  - Impact to emerging areas such as life sciences and informatics
- Several architectural features reduce the programmer's burden and enable high-performance large-scale applications with irregular data structures
- How will we program multicore processors, especially for these applications?
- Will Microsoft and Intel reach this before the traditional HPC community? ☺