BIG DATA

[Photo credit: Thinkstock]

NATIONAL SECURITY AGENCY · UNITED STATES OF AMERICA

RESEARCH DIRECTORATE · ADVANCING INTELLIGENCE THROUGH SCIENCE

Globe at a Glance | Pointers | Spinouts

Contents

THE Next Wave

The National Security Agency's review of emerging technologies

GUEST Editor's column

Mark E. Segal

Chief, Computer and Information Sciences
Research Group | Research Directorate | NSA

The economics of computing continues to change in ways that allow larger computational problems to be solved at lower costs. Dramatic increases in commodity computing power, high-density disks that can store vast amounts of data, and very high-speed networks capable of moving information long distances very quickly are all making it possible to analyze the contents of massive data repositories and derive new insights from them. Many people refer to this state of technological evolution, coupled with the development of sophisticated new data-analysis algorithms, as the era of "Big Data." Big Data offers the promise of being able to detect trends in large data sets in ways that were not possible with older technologies.

Big Data capabilities can be applied to a wide variety of problems in many different domains. For example, in a Big Data world, it may be possible to provide better health care by detecting how diseases propagate in large populations. Big Data capabilities may also allow companies to spot new consumer trends in order to make products that people want to buy, and manufacture them in sufficient quantities so that everyone who wants the product can buy it when they want it.

In the scientific world, Big Data capabilities are making it possible to sift through vast quantities of data from sensors, such as weather satellites and particle accelerators, to increase our understanding of the physical world. Big Data capabilities can enhance national security by allowing our military to gain better situational awareness before and during a battle. Big Data capabilities may also be used to

analyze potential actions by a country or terrorist organization hostile to the United States and prevent those actions from taking place.

In this issue of *The Next Wave (TNW),* NSA researcher Paul Burkhardt provides an overview of Big Data, some of the key technologies behind it, and some of the key innovators in the field. One technological aspect of Big Data that is relevant to a wide variety of problems is the ability to analyze very large graphs. Burkhardt's second article discusses these "Big Graphs," showing how large-graph algorithms can be applied to several kinds of Big Data problems.

For the results of Big Data analysis to be useful to humans trying to solve difficult real-world problems, they must be put into a form that humans can understand and process. In the third article in this issue, NSA researchers Randall Rohrer, Celeste Paul, and Bohdan Nebesh explore this topic and discuss the connection between data visualization and analysis.

As Big Data analytics become more ubiquitous, concerns naturally arise about how data is collected, analyzed, and used. In particular, people whose data is stored in vast data repositories, regardless of who owns the repositories, are worried about potential privacy rights violations. Although privacy issues are not discussed in detail in this issue of *TNW,* an excellent overview of the relevant issues may be found in a report titled "Big Data and privacy: A technological perspective" authored by the President's Council of Advisors on Science and Technology and delivered to President Obama in May 2014 [1]. Another useful resource on this topic and

other topics related to Big Data is the article "Big Data and its technical challenges" by H. V. Jagadish et al. published in the July 2014 issue of *Communications of the ACM* [2].

According to a 2012 study by the International Data Corporation, there will be approximately 10^{22} bytes of data stored in all of the computers on Earth by 2015 [3]. To put that number in perspective, that's more than the estimated 7.5 x 10^{18} grains of sand on all of the beaches of the Earth [4], and almost as much as the estimated 10^{22} to 10^{24} stars in the Universe [5, 6]. Let's harness the tools and algorithms currently being used to process Big Data to solve some of our planet's most critical problems. We hope you find this issue of *TNW* interesting, informative, and thought-provoking.

Mark E. Segal
Chief, Computer and Information Sciences Research Group
Research Directorate, NSA

References

[1] President's Council of Advisors on Science and Technology. "Big Data and privacy: A technological perspective." 2014 May 01. Available at: http://www. whitehouse.gov/sites/default/files/microsites/ostp/PCAST/ pcast_big_data_and_privacy_-_may_2014.pdf.

[2] Jagadish HV, Gehrke J, Labrinidis A, Papakonstantinou Y, Patel JM, Ramakrishnan R, Shahabi C. "Big Data and it's technical challenges." *Communications of the ACM.* 2014;57(7):86–94. doi: 10.1145/2611567.

[3] Gantz J, Reinsel D. "The digital universe in 2020: Big Data, bigger digital shadows, and biggest growth in the Far East." International Data Corporation. 2012 Dec. Available at: http://idcdocserv.com/1414.

[4] McAllister H. "Grains of sand on the world's beaches" [accessed 2014 Jun 16]. *The Journal of Modern Problem Solving.* Available at: http://www.hawaii.edu/suremath/ jsand.html. (Problem was solved using SureMath software.)

[5] Cain F. "How many stars are there in the Universe?" *Universe Today.* 3 Jun 2013. Available at: http://www.universetoday.com/102630/ how-many-stars-are-there-in-the-universe/.

[6] European Space Agency. "How many stars are there in the Universe?" [accessed 2014 Jun 16]. Available at: http:// www.esa.int/Our_Activities/Space_Science/Herschel/ How_many_stars_are_there_in_the_Universe.

An overview of
BIG DATA

Paul Burkhardt

What is Big Data?

Readers have probably heard or read about Big Data, but what is it exactly? According to O'Reilly Media, the term was coined in 2005 and refers to "a wide range of large data sets almost impossible to manage and process using traditional management tools—due to their size, but also their complexity" [1]. Gartner defines it as "high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making" [2]. The International Data Corporation (IDC) states "Big data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis" [3].

The three Vs—*volume, velocity,* and *variety*—are what many use to describe characteristics of Big Data. The consensus is that Big Data comes from many sources, is of any type, and the scale in both size and speed make it difficult to process and analyze. The National Institute of Standards and Technology (NIST) does not have an official definition of Big Data, but at their Joint Cloud and Big Data Workshop held in 2013, the NIST Director, Patrick Gallagher, agreed that key aspects of Big Data include notions of volume, velocity, and complexity. Gallagher went on to say, "We are really looking at a new paradigm, a place of data primacy where everything starts with consideration of the data rather than consideration of the technology" [4].

The Big Data era

If the last decade of computing could be envisioned as a journey through a landscape of rolling hills, dark forests, and winding paths, then a wayward traveler will encounter an uncertain and rough terrain, strewn with artifacts of IT efficiency and computer modernization, such as service-oriented architecture, grid computing, and web services. Below the surface are the fossils from the Big Iron era, huge mainframes that once bellowed and lumbered across the lands. At the end, the traveler will reach the Mountain of Data. It is steep, imposing, and so big that clouds shroud its peak, but yonder lies the horizon . . .

The beginning of the Big Data era is not marked by a definitive epoch, but it has been a persistent tide since the dot-com era which gave us the Amazon Elastic Compute Cloud (EC2), Wikipedia, Skype, and the mapping of the human genome. Google Trends indicate "big data" web searches were steady between 2004 and 2010 before it began increasing sharply. The September 30, 2012 issue of Research Trends found that the number of research publications on Big Data exceeded an

exponential rate of growth starting around 2008 (see figure 1).

As the web became a more important platform for both social and business needs, the amount of information began to grow dramatically. There were also significant advances in science which drove data production. The National Human Genome Research Institute reports that, in early 2008, the field of genomics developed second-generation sequencing platforms that began reducing the cost of DNA sequencing at a rate faster than Moore's Law—CPU performance doubles every 18 months—dramatically increasing the production of genomic data (see figure 2). The Large Hadron Collider spun up in September 2008, generating about 15 petabytes of data per year, leading to the 2012 discovery of a Higgs Boson particle.

The Big Data era is far from over. Smaller, smarter, web-enabled devices are becoming prevalent, from mobile phones to pacemakers. The first wireless pacemaker made by St. Jude Medical Inc. came online in 2011. Mundane appliances and systems such as your kitchen refrigerator and home security systems are also transmitting over the web. These interconnected devices and sensors make up the Internet of Things, resulting in greater information creation and consumption. The world is becoming more connected and mobile. The International Telecommunication Union predicts the number of
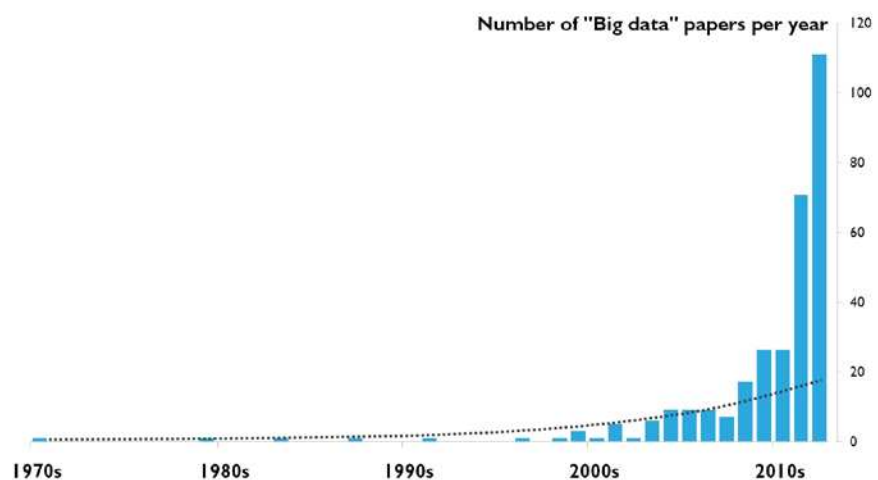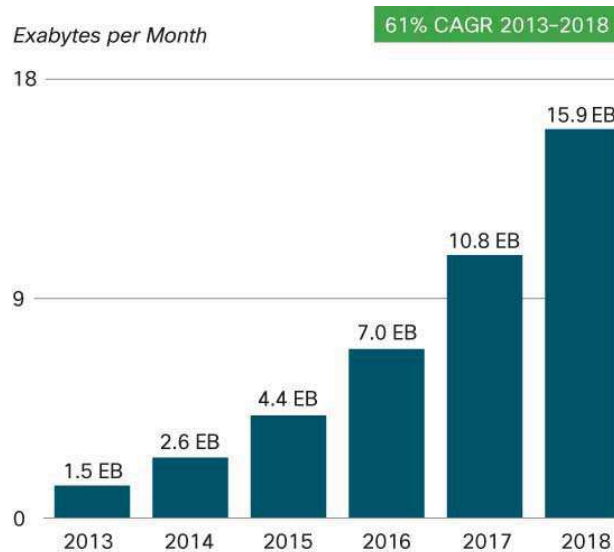

FIGURE 1. Research Trends found that the number of research publications on Big Data exceeded an exponential rate of growth starting around 2008. (Image from [5].)

mobile phone subscriptions will exceed the world population in 2014, which is over seven billion [7], and on January 2014, the IDC Worldwide Quarterly Mobile Phone Tracker reports that over one billion smartphones were shipped in a single year for the first time [8].

By 2017 the annual global Internet protocol traffic will top 1.4 zettabytes (ZB, i.e., 10^{21}) [9], and by 2018 global mobile data traffic will reach 15.9 exabytes (EB, i.e., 10^{18}) per month, according to Cisco (see figure 3) [10]. In 2014, Cisco also forecasts that 90% of all global mobile data traffic will be due to cloud applications by 2018 and that mobile cloud traffic will grow at a compound annual rate of 64% from 2013 to 2018 [10]. The forecast also predicts the number of mobile devices will reach 10 billion by 2018, with eight billion being personal devices, and over half of these mobile devices will be smart devices accounting for more than 95% of global mobile data traffic (see figure 4) [10].
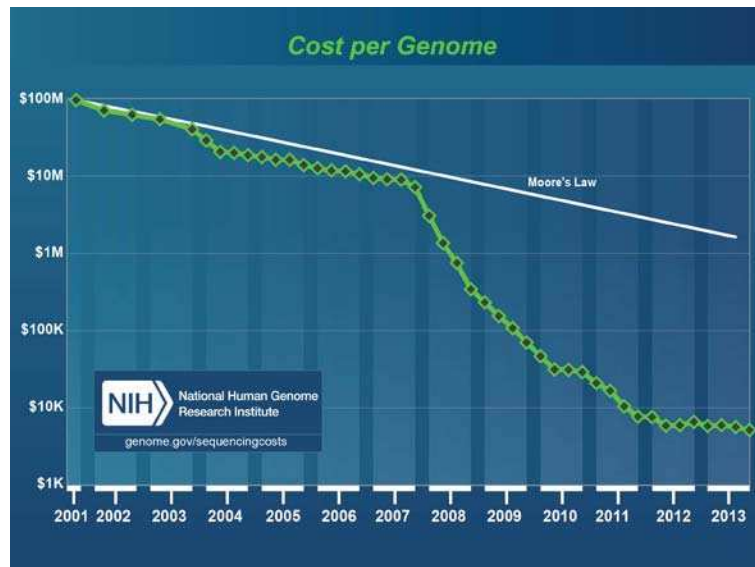

FIGURE 2. The National Human Genome Research Institute reports that, in early 2008, the field of genomics developed second-generation sequencing platforms that began reducing the cost of DNA sequencing at a rate faster than Moore's Law, dramatically increasing the production of genomic data. (Image from [6].)

FIGURE 3. Cisco predicts that by 2018 global mobile data traffic will reach 15.9 EB per month, and the annual global Internet protocol traffic will top 1.4 ZB. (Image from "Cisco Visual Networking Index: Global mobile data traffic forecast update, 2013–2018" [10].)

How "big" is Big Data?

The growth of sensors and devices, coupled with social media and scientific breakthroughs, all sharing and transmitting over the same networks, contribute to the *data deluge.* According to an IDC study, "Extracting value from chaos," information is doubling every two years while metadata, data about data, is growing two times faster than data. A 2011 study of the world information capacity [11] estimated there were 295 EB of storage, 1.9 ZB of broadcast data (i.e., TV, radio) and 65 EB of telecommunication data (i.e., fixed phone, mobile phone, Internet)—in 2007.

The 2012 acting director of the Defense Advanced Research Projects Agency (DARPA), Kaigham J. Gabriel, used this analogy, "The Atlantic Ocean is roughly 350 million cubic kilometers in volume, or nearly 100 billion, billion gallons of water. If each gallon of water represented a byte or character, the Atlantic Ocean would be able to store, just barely, all the data generated by the world in 2010. Looking for a specific message or page in a document would be the equivalent of searching the Atlantic Ocean for a single 55-gallon drum barrel" [12].
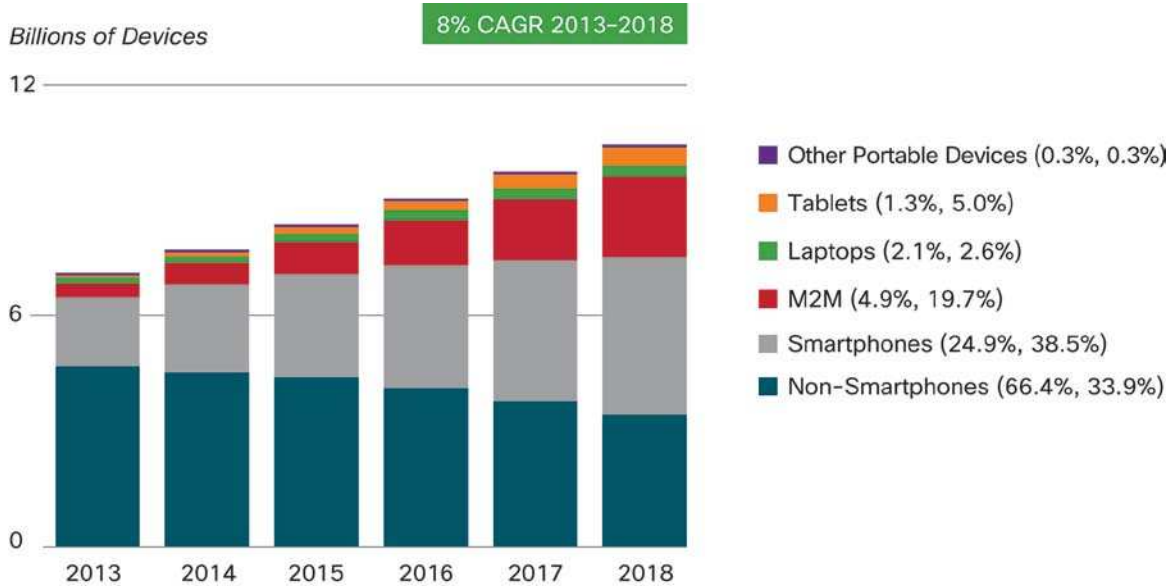
Figures in parentheses refer to device or connections share in 2013, 2018.
Source: Cisco VNI Mobile, 2014

FIGURE 4. Cisco predicts the number of mobile devices will reach 10 billion by 2018, with eight billion being personal devices, and over half of these mobile devices will be smart devices accounting for more than 95% of global mobile data traffic. (Image from "Cisco Visual Networking Index: Global mobile data traffic forecast update, 2013–2018" [10].)

The total digital information capacity, according to a 2012 IDC study, will reach 40 ZB by 2020 [13]. This projects approximately 10^{22} bytes of digital information in less than 10 years. By comparison, the number of stars in the universe is on order of 10^{22}, and the number of atoms in a mole is on order of 10^{23} (i.e., Avogadro's number). We are living in an era in which we have more data than resources to store and analyze it all.

Big Data in government

In March 2012, the White House announced the National Big Data Research and Development Initiative [14] to help address challenges facing the government, in response to the President's Council of Advisors on Science and Technology, which concluded the "Federal Government is under-investing in technologies related to Big Data." With a budget of over $200 million and support of six federal departments and agencies, this initiative was created to:

▶ Advance state-of-the-art core technologies needed to collect, store, preserve, manage, analyze, and share huge quantities of data;

▶ Harness these technologies to accelerate the pace of discovery in science and engineering, strengthen our national security, and transform teaching and learning; and

▶ Expand the workforce needed to develop and use Big Data technologies.

As part of the Big Data Initiative, the National Science Foundation (NSF) and the National Institutes of Health are funding a joint Big Data solicitation to "advance the core scientific and technological means of managing, analyzing, visualizing, and extracting useful information from large and diverse data sets." In addition, the NSF is funding the $10 million Expeditions in Computing project led by University of California at Berkeley, to turn data into knowledge and insight, and funding a $2 million award for a research training group to support training for students in techniques for analyzing and visualizing complex data.

The Department of Defense (DoD) is also investing $250 million annually to "harness and utilize massive data in new ways" and another $60 million for new research proposals. DARPA, the research arm of the DoD, will invest $25 million annually

under its XDATA program for techniques and tools to analyze large volumes of data, including

- ‣ Developing scalable algorithms for processing imperfect data in distributed data stores, and

- ‣ Creating effective human-computer interaction tools for facilitating rapidly customizable visual reasoning for diverse missions.

The Department of Energy is similarly providing $25 million in funding to establish the Scalable Data Management, Analysis and Visualization Institute to develop new tools for managing and visualizing data.

What is all of this data?

Much of the "big" in Big Data comes unsurprisingly from web data. Web browsing activities are tracked in clickstream data that can be used to market the latest fads and trends. Social media also populates web data where Facebook, Twitter, and YouTube enable users to easily post, upload, and tweet about their daily activities.

Over 90% of digital information is unstructured [5]—data which has not been converted for organized analysis. Unstructured data is raw information created by ad hoc activities and can consist of images, video, logs, and documents. Unstructured data can be rich in semantic and relational information that must be parsed, interpreted, and transformed into canonicalized formats to fit neatly into catalogs and databases. In 2012, IDC believed 23% of information in our "digital universe" has Big Data value, increasing to one third by 2020, but only if the data is tagged and analyzed.

Who is creating all of this data?

Individuals are responsible for 75% of all digital information [15], half of which is generated *by* an individual through their activities, such as e-mails, phone calls, vacation photos, and video uploads. The other half is generated *about* an individual, a person's digital shadow, such as their web-browsing habits, financial transactions, surveillance footage, medical database, and GPS tracking.

Big Data economy

Big Data is Big Business—a January market report by Transparency Market Research claims the 2012 global Big Data market was worth $6.3 billion and will grow to $48.3 billion by 2018 [16], while IDC forecasts Big Data market revenue will grow at 31.7% per year, reaching $23.8 billion in 2016 [17]. This global economy includes cloud technologies centered around Apache Hadoop software, which was inspired by Google's approaches to solving their internal Big Data dilemma. The value inherent in data can only be unlocked through analysis, which requires the right tools and infrastructure, but traditional methods are inadequate for Big Data, creating the necessity for innovation.

Can we process Big Data?

The volume and velocity of Big Data is exceeding our rate of physical storage and computing capacity, creating scalability demands that far outpace hardware innovations. Just as multicore chips were designed in response to the limits of clock speeds imposed by Moore's Law, cloud technologies have surfaced to address the impending tidal wave of information. The new cloud architectures pioneered by Google and Amazon extended distributed computing from its roots in high-performance computing and grid computing, where hardware was expensive and purpose-built, to large clusters made from low-cost commodity computers, ushering the paradigm of "warehouse" computing. These new cloud data centers containing thousands of computer cabinets are patrolled by administrators on motorized carts to pull and replace failed components.

On the software side, new parallel programming frameworks, like Hadoop MapReduce, help us crunch vast batches of data locally and independently by minimizing data dependency and the cost of transferring data between systems. Relational databases are being replaced by NoSQL and "eventually consistent" (key, value) stores such as Accumulo and MongoDB in order to scale with data demands.

Is Big Data secure?

The studies from IDC found that less than a third of the digital information was secured, and of the data requiring protection, only 20% was actually protected [3]. More than 80% of data about an individual will have passed through a commercial organization, a liability that is quietly ignored [3]. The lack of security is partly due to deference, but as data becomes more ubiquitous and available in commercial clouds, the need for security will become more imperative. Currently, the only open-source cloud database that offers cell-level security is Accumulo, a BigTable variant initially developed here at the National Security Agency (NSA).

Past the mountain of data

The DoD is leveraging the cloud infrastructure to consolidate 15,000 military networks, a potential annual savings of $680 million [18]. At NSA, the Big Data challenge is compounded by the need for security and compliance where data must be compartmentalized and audited as part of the oversight requirements to protect the Fourth Amendment.

This requires cell-level security and provenance. The challenges for NSA are not limited to infrastructure; analysis is still the primary tradecraft of the Agency. The mountain of data must be processed into intelligence products that will safeguard national security. The need for Big Data analytics will be at the core of this tradecraft.

In today's Big Data era, information is exploding as networks converge and more devices come online in the Internet of Things. The scale of Big Data poses immense challenges, and with our unrelenting consumption of information, this is likely to persist. Addressing Big Data challenges will require efforts from government, academia, and industry.

About the author

Paul Burkhardt is a computer science researcher in the Research Directorate at NSA. He received his PhD from the University of Illinois at Urbana-Champaign. His current research interests are primarily focused on graph algorithms and Big Data analytics.

References

[1] O'Reilly. "Roger Magoulas on Big Data" [YouTube video]. Uploaded 2010 Nov 16. Available at: http://youtu.be/fsT2NyM5BTI.

[2] Beyer M, Laney D. "The importance of 'Big Data': A definition." Gartner. 2012 Jun 21.

[3] Gantz J, Reinsel D. "Extracting value from chaos." IDC. 2011 Jun.

[4] Gallagher P. [Remarks from] NIST Joint Cloud and Big Data Workshop. Jan 2013, Gaithersburg, MD. Webcast available at: http://www.nist.gov/itl/cloud/nist-joint-cloud-and-big-data-workshop-webcast.cfm.

[5] Research Trends. "The evolution of Big Data as a research and scientific Topic: Special issue on Big Data." 2012 Sep 30. Available at: http://www.researchtrends.com/wp-content/uploads/2012/09/Research_Trends_Issue30.pdf.

[6] Wetterstrand KA. "DNA sequencing costs: Data from the NHGRI Genome Sequencing Program (GSP)" [accessed 2014 Mar 31]. Available at: https://www.genome.gov/sequencingcosts.

[7] International Telecommunication Union. "The world in 2013: ICT facts and figures." 2013 Feb. Available at: http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2013-e.pdf.

[8] International Data Corporation. "Worldwide smartphone shipments top one billion units for the first time, according to IDC" [Press release]. 2014 Jan 27. Available at: http://www.idc.com/getdoc.jsp?containerId=prUS24645514.

[9] Cisco. "The zettabyte era—trends and analysis." 2013 May 29. Available at: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/VNI_Hyperconnectivity_WP.html.

[10] Cisco. "Cisco Visual Networking Index: Global mobile data traffic forecast update, 2013–2018." 2014 Feb 5. Available at: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.pdf.

[11] Hilbert M, Lopez P. "The world's technology capacity to store, communicate and compute information." *Science.* 2011;332(6025):60–65. doi: 10.1126/science.1200970.

[12] Defense Advanced Research Projects Agency. "DARPA calls for advances in 'Big Data' to help the warfighter" [Press release]. 2012 Mar 29. Available at: http://www.darpa.mil/NewsEvents/Releases/2012/03/29.aspx.

[13] Gantz J, Reinsel D. "The digital universe in 2020: Big Data, bigger digital shadows, and biggest growth in the Far East." International Data Corporation. 2012 Dec. Available at: http://idcdocserv.com/1414.

[14] Office of Science and Technology Policy, Executive Office of the President. "Obama administration unveils 'Big Data' initiative: Announces $200 million in new R&D investments" [Press release]. 2012 Mar 29. Available at: http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf.

[15] Gantz JF, Chute C, Manfrediz A, Minton S, Reinsel D, Schlichting W, Toncheva A. "The diverse and exploding digital universe." 2008 Mar. International Data Corporation. Available at: http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf.

[16] Transparency Market Research. "Big Data market—Global scenario, trends, industry analysis, size, share and forecast 2012–2018." 2013 Jan 21.

[17] International Data Corporation. "New IDC Big Data technology and services forecast shows worldwide market expected to reach to $23.8 billion in 2016" [Press release]. 2013 Jan 08. Available at: http://www.idc.com/getdoc.jsp?containerid=prUS23900013.

[18] Sternstein A. "NSA chief endorses the cloud for classified military cyber program." 2012 Jun 13. *Nextgov.* Available at: http://www.nextgov.com/cybersecurity/2012/06/nsa-chief-endorses-cloud-classified-military-cyber-program/56257/.
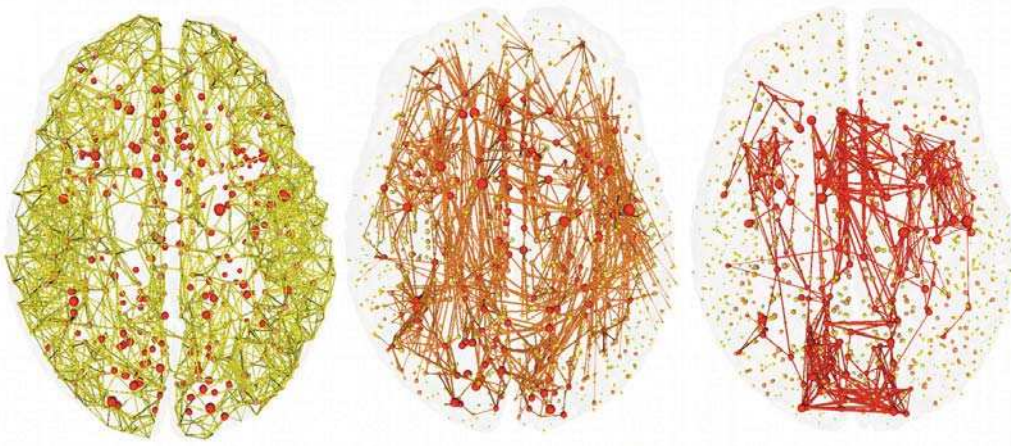
Our brain is a Big Graph, a network of trillions of neurons connected by synapses, whose topology shares common characteristics with other graphs, such as social networks. Can we unlock the secrets of our neural processing using graph theory and Big Data technologies? (Image reprinted from [1].)
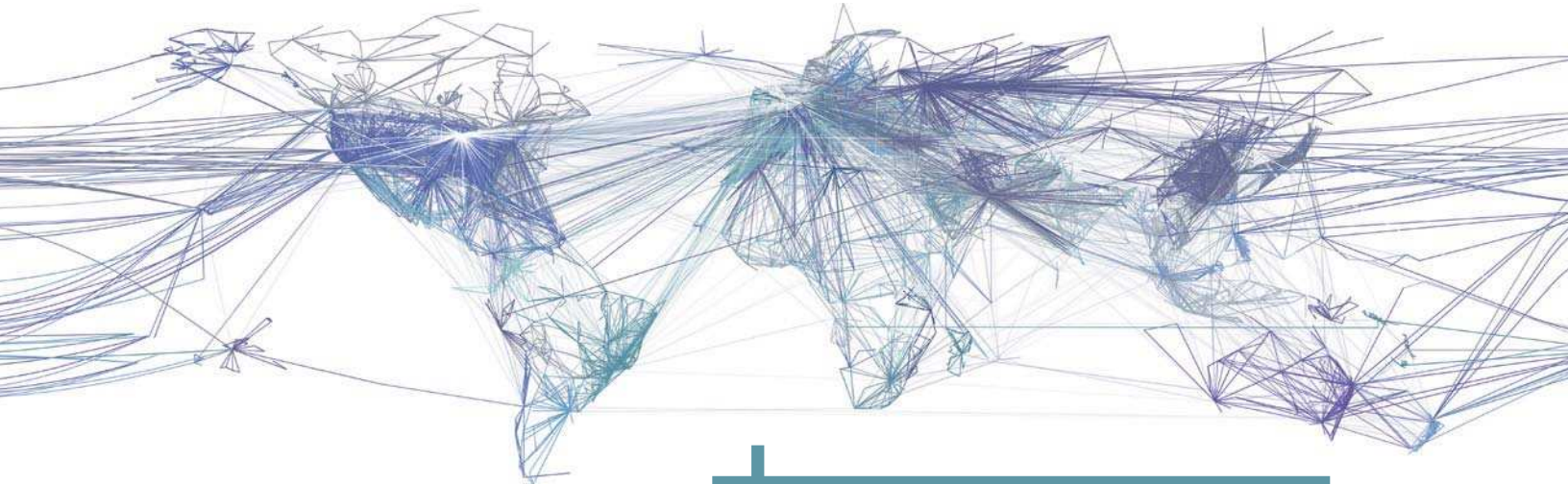
Big Graphs

Paul Burkhardt

FIGURE 1. Graphs arise naturally from physical networks, such as the flight paths between airports. (Design: Thirst. Project: O'Hare Terminal 5 Mezzanine Mural. Client: Westfield Development. Illustration built using Processing Data by http://OpenFlights.org [2].)

A graph is a group of associated objects represented by a network of vertices and edges, where a vertex is an object and an edge connects a vertex to another vertex to denote their pairwise relationship. Graphs arise naturally from physical networks, such as the roads and highways connecting our cities, the power grid that transfers electricity to our homes, and the flight paths between airports (see figure 1). Biological systems also exhibit graphs, such as the interactions between proteins (see figure 2) and the conformational topology of polymers. The neurons in our brain send signals over synapses, forming one of the largest natural networks in existence. We also engineer networks from the minute electronic circuitry in microprocessors to the massive digital network of the Internet, displayed as a graph in figure 3, facilitating communication between computers all over the world.
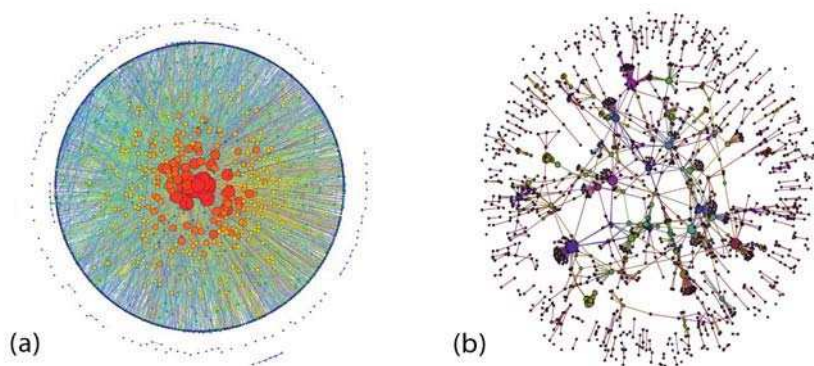
FIGURE 2. (a) Biological systems also exhibit graphs, such as the interaction between proteins. Above is a yeast protein interactome. (Graph created with Gephi, http://www.gephi.org.) **(b)** Above is a *Mycobacterium tuberculosis* interactome. (Image reprinted from [3].)

Graphs are everywhere

A graph can also be constructed from abstract and less obvious sets of relationships. For example, this article can be visualized as a graph of words. While **reading** this **sentence**, **connect** any **pair** of **words** **co-occurring** in a **span** of four **words** but **counting** only **nouns** and **verbs**. Our simple word graph in figure 4(a) reveals a number of cliques with a maximum size of four vertices. A clique is a group of vertices that are all pairwise connected, indicating the vertices are closely associated because each vertex is directly connected to any other. An interesting structure emerges where two of the largest cliques around the predicates **connect** and **counting** share the **words** vertex, thus tying any pair of vertices in this structure by two edges or less (see figure 4(b)). We can infer that connecting pairs of words in the sentence is closely associated with counting nouns and verbs, but reading is not closely associated to nouns and verbs in this context because **reading** is separated by no less than three edges to either **nouns** or **verbs**, despite the obvious grammatical relationship.

Word co-occurrence graphs are an abstract representation of written language that can help expose semantic meaning by machines. Another less obvious utilization of graphs is solving the *shortest superstring problem*—the task of creating the shortest string that contains each substring from a set of *n* substrings. If the length of the superstring did not matter, then the problem is trivially solved

by concatenating all the substrings. Constructing the shortest superstring that contains each substring exactly once is much harder but has applications in data compression and genome assembly. A brute-force method that shortens a superstring by the overlap between substrings must do so for all *n!* possible superstrings, which quickly becomes intractable (e.g., 15! is over one trillion).

The shortest superstring problem can be solved by creating a graph where vertices are the *n* substrings and all pairs of vertices are connected by edges with a weight given by the longest suffix of one vertex that is equal to the prefix of the other and a direction in that order, then finding a Hamiltonian path that visits each vertex once while maximizing the overlap (also known as the Traveling Salesman Path Problem). But finding a Hamiltonian path is in the class of NP-complete (i.e., nondeterministic polynomial
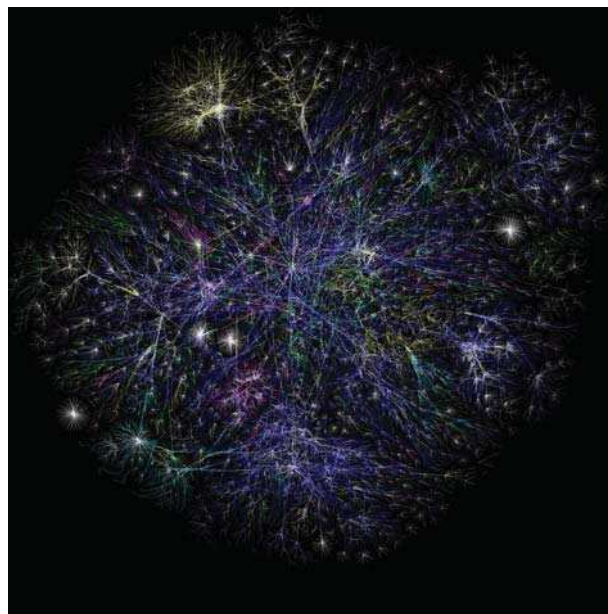

FIGURE 3. We engineer networks, such as the digital network of the Internet, displayed above as a graph. (Image from [4].)
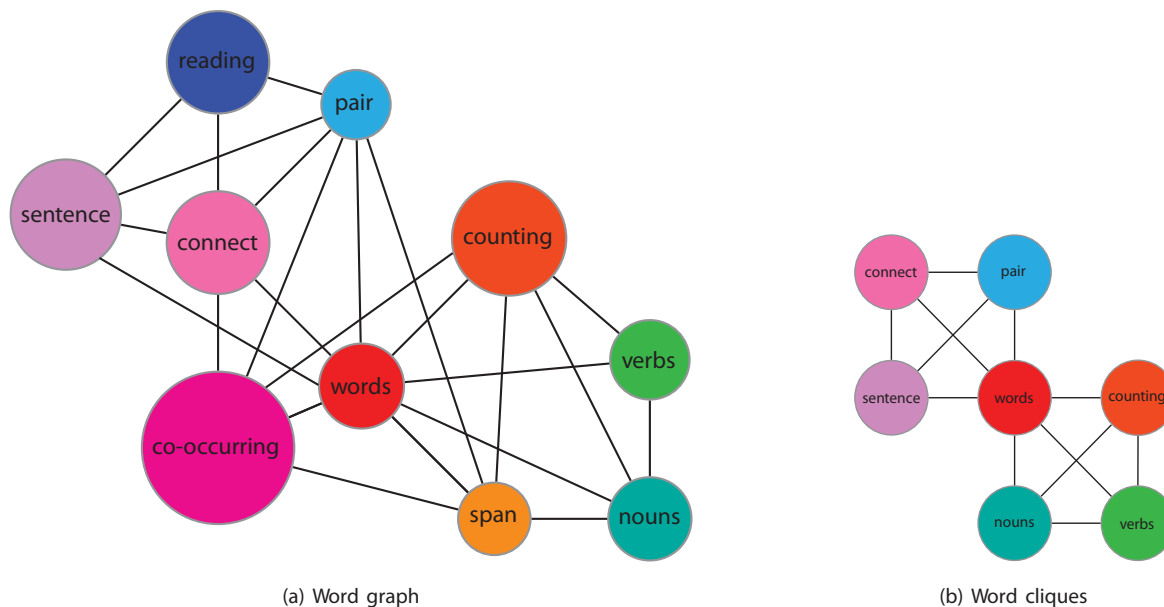
(a) Word graph

(b) Word cliques

FIGURE 4. (a) Texts can be visualized as a graph of words, such as the graph above of the co-occurrence of words in a sentence from this article. **(b)** These word cliques (a clique is a group of vertices that are all pairwise connected) from figure 4(a) reveal associations between words. Here, they reveal that connecting pairs of words in the sentence is closely associated with counting nouns and verbs.

time-complete) problems for which efficient solutions are not known.

A special case where each substring has length k over an alphabet of size n is more tractable. This problem can be solved by constructing a de Bruijn graph where each k-length substring is an edge that begins from its $(k$-1)-length prefix and ends at its $(k$-1)-length suffix, then finding a Eulerian cycle—a path that traverses each edge exactly once before returning to the origin. (Eulerian cycles are inspired by Euler's 1735 solution to crossing the Seven Bridges of Königsberg over the river Pregel which started the study of graph theory.) The de Bruijn graph in figure 5 admits a Eulerian cycle, just follow the labeled edges in order and concatenate the first symbol in each edge to construct the cyclic superstring 0000110010111101, representing all sixteen k=4 length substrings for an alphabet of 0 and 1. The graph by de Bruijn is an important method used in DNA sequencing where possibly billions of k-mers (i.e., substrings of k-length) must be assembled to construct the final genome sequence.
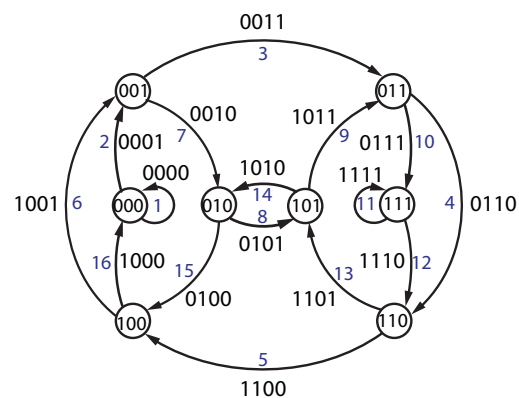
FIGURE 5. This de Bruijn graph admits a Eulerian cycle—a path that traverses each edge exactly once before returning to the origin. This type of graph can be used to solve the shortest superstring problem and is used in DNA sequencing. (Image reprinted by permission from Macmillan Publishers Ltd: *Nature Biotechnology,* available at http://www.nature.com/nbt/index.html, Compeau PE, Pevzner PA, Tesler G, "How to apply de Bruijn graphs to genome assembly," doi: 10.1038/nbt.2023, fig. 2, 2011 [5].)

What can graphs tell us?

A graph can be a beautifully complex and intriguing topology of interconnected pathways, alluding to hidden meaning and secrets available only to the intrepid willing to walk the edges. Often the graph resembles little more than a hair ball, such as the graph of the World Wide Web in figure 6, obfuscating insight by the seemingly infinite number of circuitous paths. But Google's search engine, for example, is based roughly on the concept of randomly following links from one web page to another in a gigantic web graph, ranking each page according to the popularity of the pages that link to it, and returning surprisingly accurate search results.

Social interactions can be symbolized by graphs (e.g., the Twitter graph in figure 7) and inspire colloquial phrases such as *small world* and *six degrees of separation,* indicating that we are all connected by just a few associates. The topology of our social networks was discovered to be more resistant to failure when nodes or links are removed, which can result in the dissociation of communities or the disruption of pathways, and to be better
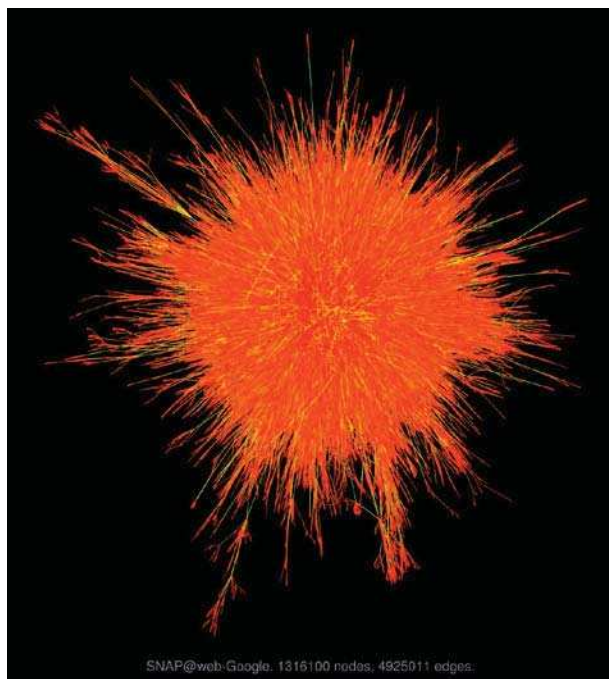

FIGURE 7. Social interactions, like those on Twitter, can be symbolized by graphs. (Graph created with Gephi, http://www.gephi.org.)

at disseminating information than other graph topologies [7, 8]. These small-world graphs have more cliques and shorter paths, but it is the severe inequities among the vertices that explain why rumors and disease quickly spread throughout these networks. Because a few vertices incur the vast majority of edges, acting as hubs, many low-degree vertices with only a few direct neighbors are able to exchange information easily [8].

Such small-world graphs can be found in many real-world networks. For example, the hub structure can be found in the network of US airports where, according to 2012 data, 80% of passengers are serviced by only 50 out of nearly 20,000 airports [9, 10]. The small-world graph properties can also be found in neural networks, such as that of the soil nematode *Caenorhabditis elegans* (shown in figure 8), implying these graph properties have an evolutionary benefit [7]. Thus, out of complex, unordered, and decentralized interactions, logic and purpose arise. Small-world graphs develop naturally without any centralized control or predefined order but rather from *preferential attachment* where popular nodes become more popular over time—just as our network of roads started as decentralized clusters localized to cities and towns, eventually connecting to other clusters creating hubs around the big cities.

Graphs are truly everywhere and can be literally constructed from any data. But graphs do

FIGURE 6. Some graphs, such as this World Wide Web graph, are so complex that they resemble little more than a hair ball. (Image from [6].)
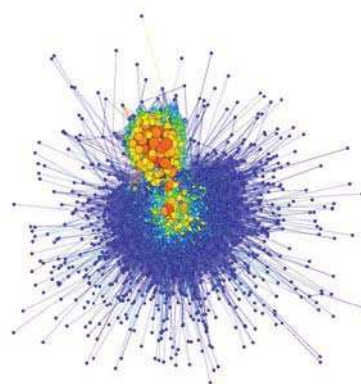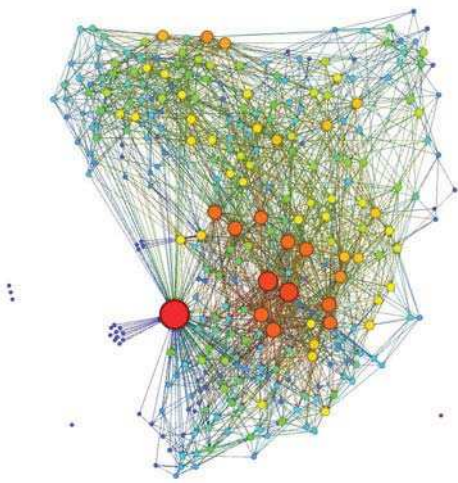
FIGURE 8. Small-world graph properties can also be found in neural networks, such as this one of the soil nematode *Caenorhabditis elegans,* implying that these graph properties have an evolutionary benefit [7]. (Graph created with Gephi, http://www.gephi.org.)

not by themselves add information; instead, they help to organize data as a collection of (key, value) pairs that effectively encode binary relationships which, when analyzed in whole, can reveal surprising insight to complex interactions. Algorithms that discover cues from navigating the graph have been studied since the days of Euler and the Seven Bridges of Königsberg (only two bridges from his day still stand) nearly three centuries ago. Searching a complex network is an exemplary application of graph algorithms and one familiar to us each time we use GPS navigation to compute the best route from one location to another. But graph algorithms are computationally challenging because of their irregular structure and combinatorial expansion.

One of the simplest data structures is a binary tree in which each node begets two more nodes. The branches of this graph expand in powers of two, so after just 16 generations, there are already 131,071 nodes, and another 16 generations later, there are over eight billion nodes. In most graphs, the branches are not regular and expand much more quickly. In many real-world graphs, the disparity in degree distribution creates significant resource contention during computation. The powerful analytic capability of graph algorithms has

motivated the design for efficient parallel processing of graphs in high-performance computing (HPC) systems. Fields such as genomics, molecular dynamics, and data science are utilizing many of these HPC graph algorithms to analyze their large and complex data sets. The rising tide of Big Data has created interest in applying graph-theoretic approaches in these fields and many others. But as data sets get larger, the challenges to graph processing increase to a point where even the most powerful HPC systems will buckle under the task of graph analysis on Big Data.

Big Graphs

The introduction to Big Data gives a sense of the massive scale of some of these data sets which would create very big graphs. On any given day the web contains about 50 billion web pages (cf. http://www.worldwidewebsize.com), and if we estimate an average of 20 URL links per page, the web graph would have one trillion edges. In 2008, Google had already claimed to have indexed a total of one trillion pages. In October of 2012, Facebook announced that their social media site had reached one billion active monthly users, connecting *one out of seven people on the planet,* and since 2004, there have been 140.3 billion friend connections. In early 2013, Facebook announced their Facebook Graph Search to harness the Big Data graph information collected in their social network which could include the more than one trillion "likes" made by their users.

The computational resources for searching the web or the Facebook network are hidden in secret data centers built by Google and Facebook. But in 2010, Google published their Pregel paper for processing large-scale graphs [11]. In this paper, Google described their distributed-memory approach, which follows the bulk synchronous parallel (BSP) model of computing rather than the parallel random access machine (PRAM) model traditionally favored for graph algorithms. A distributed-memory system is a cluster of machines, each with their own private memory, and data residing in the memory of one machine must be explicitly communicated to another machine. Increasing the memory of such a distributed-memory system only requires

connecting more machines. In contrast, a shared-memory system has a single pool of memory that is accessible to all machines, while each machine also has a small portion of private memory. Communicating data changes to all machines, therefore, simply requires updating the data in the pool. But a protocol must be enforced to ensure data remains consistent, especially when one machine has loaded data into its own private cache but, before it can process that data and replace the modifications in the pool, another may have already made changes. This cache-coherency protocol makes it much more difficult to scale shared-memory computers.

Another limiting factor is that a central processing unit (CPU) has a memory address limit. For example; the Intel Xeon E5 has a 46-bit address space [12]; therefore, a system comprised of these CPUs can have no more than 64 terabytes (TB) of globally-shared memory. It is not surprising that Google's Pregel favors the distributed-memory model. But the Big Graph challenge does not end here.

The problem with big brains

One of the largest physical networks is our own neural network, the human connectome, depicted in figure 9. If we count neurons as vertices and synapses as edges, there are approximately 10 trillion vertices and 100 trillion edges in the human brain graph. If each edge were stored in 16 bytes, our brain graph would occupy over one petabyte (PB)—that exceeds the practical memory capacity of any computing platform today. As described below, the largest memory capacity in a supercomputer is 1.5 PB.

The human brain graph stored in bytes would occupy over one petabyte. How large is that?

$(1,024)^3$ bytes	=	1 gigabyte (GB)
1,024 GB	=	1 terabyte (TB)
1,024 TB	=	1 petabyte (PB)

Leaving the memory issue aside, if we traversed edges at a pace of one every millionth of a second (microsecond) it would us take over three years to visit each neuron without ever retracing a step. This rate of one million edges per second is clearly impractical, but considering the fastest network technologies available have microsecond latency between one network interface to another, it will require careful implementation on a many-processor
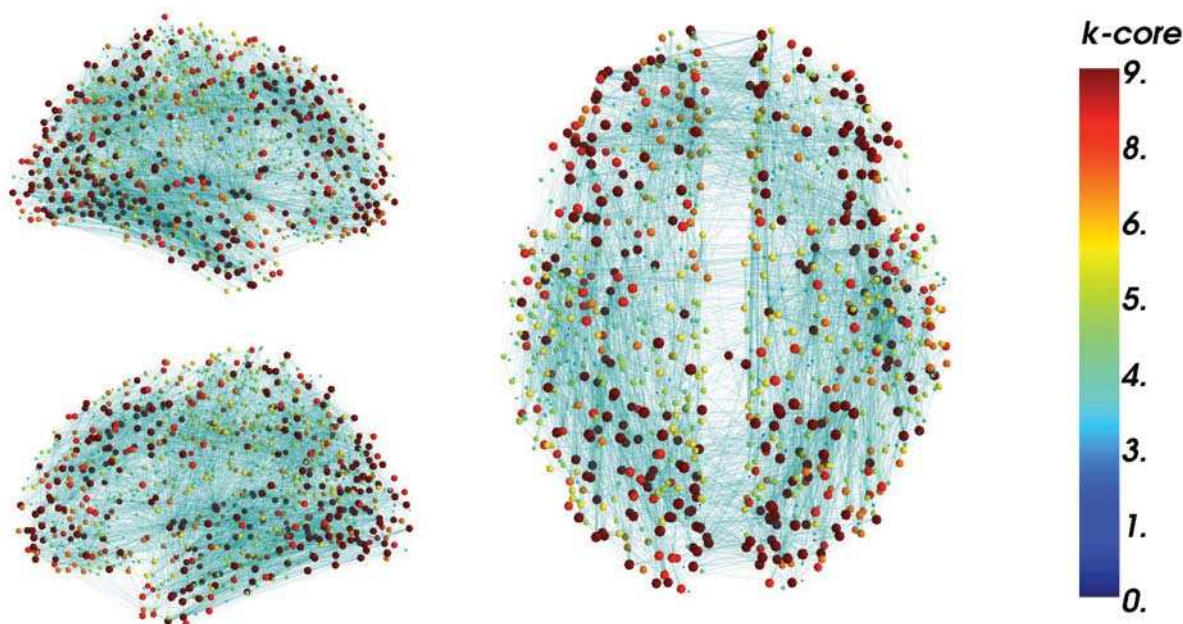

FIGURE 9. One of the largest physical networks is our own neural network, the human connectome. Copyright © 2011 Gerhard, Daducci, Lemkaddem, Meuli, Thiran, Hagmann [13].

supercomputer to overcome the latency costs incurred by traversing all the edges.

A typical CPU, the "brain" we are familiar with in our personal computers, can operate at gigahertz (GHz) frequency where the CPU can perform an instruction every nanosecond (ns) or a billionth of a second; in one microsecond, the CPU will have cycled 1,000 times. There is also a speed limitation forced upon us by physics (despite recent excitement in the now debunked "faster-than-light" neutrinos) that **light travels approximately 0.3 meters every nanosecond.** A graph with trillions of edges will necessarily be distributed across many compute racks so the distance between racks at the far ends will be a factor. This is the paradox—scaling a system to keep up with increasing data can make it more difficult to process that data!

As graphs scale with Big Data, increasing the physical memory to fit the graph may not always be practical or environmentally feasible. The Cray Titan supercomputer installed at Oak Ridge National Laboratory was the world's most powerful supercomputer in 2012 according to the Top500.org November list of that year [14]. The $97 million Titan requires 8.2 megawatts (MW) of power [14] and over 4,300 square feet of space—an NBA basketball court is 4,700 square feet—but with a total memory capacity of 710 TB, it does not have enough memory to store the human connectome. The second most powerful supercomputer on the November 2012 list, the IBM Sequoia installed at Lawrence Livermore National Laboratory, requires 7.9 MW of power [14] and over 3,000 square feet of space. The Sequoia has 1.5 PB of memory, just enough to store the human connectome, but leaves no memory for applications that could analyze the brain graph. At a hypothetical 10 cents per kilowatt-hour, it would cost about $7 million per year to power either of these supercomputers ($\frac{\$100}{MW} \times 8{,}760 \frac{h}{y}$).

Idle time on these systems is very costly, but ensuring all CPUs are performing useful work when processing a Big Graph is a daunting challenge. A single Sequoia 1.6 GHz CPU can perform 204.8 operations per nanosecond (i.e., 1.6 cycles/ns × 16 cores × 8 operations/cycle per core) [15], but if it is requesting data from another CPU that is connected 10 meters away, at least 33 ns will pass—due to the speed of light limit—before it can perform

useful work. That is a waste of 6,831 operations for just *one* CPU; there are *98,304* CPUs in Sequoia!

Graphs at Big Data scales will demand substantial system resources for processing and storing, but reality forces limitations on budget, which includes the up-front cost of an installation, lifecycle support and maintenance, and the power required to keep the lights blinking, disks whirring, and fans humming. These systems will inevitably face hardware and software failures, making fault tolerance more imperative because restarting an algorithm on a petabyte or larger graph is very costly in time and resources. We need new approaches if we are to analyze Big Graphs.

Exception! Out of memory

In addition to the limitations of power, space, and cooling, there are hardware constraints to scaling the memory capacity of a system. Data is processed by entering through the CPU pins that interface the CPU to the memory bus. The number of pins is physically limited, which results in a memory bandwidth wall. In addition, a memory controller that mediates the data between main memory and the CPU has a fixed number of memory channels for transferring data because of the electrical constraints in the circuitry. These constraints force a hard ceiling on the maximum memory capacity for a processing unit. Using the Intel Xeon E5 again as an example, it supports four channels with each channel supporting three memory slots for a total of 12 slots per CPU, and at 8 GB per slot [12], such a dual Xeon motherboard would have 192 GB of memory.

An adjacency list is a common graph data structure that uses an array for storing vertices and a doubly-linked list for storing the adjacency or neighborhood of each vertex. This adjacency list requires on order of $n + 4m$ memory locations for n vertices and m edges, and for large graphs, each location would require 8 bytes. To store the brain graph entirely in memory using the adjacency list (using 100 trillion = $2m$), a system would need over 8,000 of these Intel Xeon E5 motherboards and 204 racks to house them; there are 200 racks in the Titan supercomputer. The cost in memory alone for this system would be almost $20 million at $100 per

memory slot. The 96-rack Sequoia supercomputer supports a maximum of 64 GB of memory per CPU with 1,024 CPUs per rack, which will be useful in the event that we discover a life form with a 6 PB brain graph . . . but no bigger!

If the graph cannot fit in the aggregate memory, it cannot be processed. The conventional solution is to increase the system size (i.e., add more compute boards), but that will exacerbate the latency costs, making it harder to send data from CPU to CPU to keep them busy. Bottom line: It will be difficult to scale memory in this manner if data continues to increase at exponential rates.

We can store and process Big Graphs on modest computing clusters where the graph data itself resides on disk. If the graph gets too big, then more or bigger disks can be easily added since disks have much greater data capacity than memory modules and many more drives can be attached (possibly over 100 with port multipliers). But accessing data on disks can be one million times slower than accessing it in memory. Algorithms for graphs on disk must amortize the higher latency of disk access by increasing the throughput of data. These algorithms minimize the amount of random access to avoid a frenzy of mechanical movement from disk heads seeking for data sectors. To do this effectively, the algorithms organize data in large sequential blocks because disk heads can efficiently scan data in this manner. This external memory (i.e., out-of-core) processing was first developed in the 1980s to cope with the growing disparity in both cost and performance between disk and memory, so the problem of insufficient memory is not new [16]. Processing graphs too big to fit in memory appeared in the 1990s as streaming [17] and parallel disk model [18, 19] applications.

Big Graphs in the cloud

Open-source cloud technologies inspired by Google publications [20, 21] are being leveraged to solve Big Data problems in both industry and government. The Apache Accumulo project (http://accumulo.apache.org), originally an internal research project at the National Security Agency (NSA), can be used as a graph database that can scale with disk capacity while providing security, availability, and fault tolerance. A Big Graph can

be stored in Accumulo as a collection of sorted edges and queried using the Accumulo interfaces for scanning records. The Hadoop MapReduce (http: //hadoop.apache.org) programming framework can be combined with Accumulo for added processing power. A straightforward approach is to filter out edges from Accumulo (i.e., extract a subgraph) which can then be analyzed by MapReduce applications.

Storing a graph as edges is natural in (key, value) repositories, like Accumulo, since an edge is a vertex pair (i.e., the end points). Tables in Accumulo are distributed as a set of tablets, often many tablets on a single host in a cluster of multiple hosts. Each table is stored on disk in the Hadoop Distributed File System (HDFS), which replicates all data across the cluster to tolerate faults. Accumulo keeps track of the location of all tablets and can rebalance the distribution on demand. The tablets can migrate from one host to another depending on the load distribution or host failures. The (key, value) records are sorted in each tablet, and tablets can be grouped dynamically so scans can efficiently access only relevant subsets of data.

In real-world graphs such as the social and neural networks discussed earlier, the degree for a few vertices can be much larger than the rest, resulting in skew distribution of tablets. This skew creates a *hot spot* or bottleneck since the majority of queries will access only a few of the tablets. Additionally, adjacencies would be larger for Big Graphs, increasing the time needed to scan all entries in a tablet. In Accumulo, a large adjacency can be distributed across multiple tablets to enable greater parallel processing, and the tablet sizes can be controlled for better latency and less resource contention. The locality can be set—that is, tablets can be grouped based on types of edges (i.e., scan blue versus green edges)—to skip over data that is not relevant to the query.

Updating edges in the Accumulo edge table can be accomplished using the online ingest interface or the offline bulk load operation. The latter, as the name suggests, is reserved for large, wholesale updates that are completed in bulk. The ingest interface provides a timely, low-latency mechanism which inserts updates that are globally sorted in periodic compaction operations; deleted edges are

removed after the compaction step. In the event that a tablet fails before sorting its entries, the updates can be recovered from the write-ahead logs.

The MapReduce programming model is effective for distributed problems that can be decomposed into many independent tasks. The map step processes input data into a collection of (key, value) pairs which are then sorted and combined in the reduce step. By minimizing interdependency between processing elements, the amount of communication over the network is decreased and more time can be spent on actual processing—maximizing the computation-to-communication ratio. Increasing the number of compute resources should proportionately decrease the processing time to just about the time required to communicate data between the map and reduce steps.

The canonical example of an *embarrassingly parallel* MapReduce algorithm that minimizes communication is the simple word-count pattern described in the seminal MapReduce article by Google [21]. The algorithm counts the occurrence of every word in a large corpus of documents where each document is split into blocks of lines and distributed to many processing elements. The blocks are processed simultaneously by many independent map tasks which output (*word,* 1) pairs for each word. These pairs are collected and summed in the reduce tasks to calculate the count for each word. You could run the MapReduce word-count algorithm on this article to output how many times the words "big," "data," and "graph" were used, but the effectiveness of MapReduce is better realized on very large data sets where the latency from disk access can be amortized.

Developing effective graph algorithms in the MapReduce programming model requires "thinking in MapReduce," which may seem unnatural at first. But this recasting of conventional graph algorithms into counting (key, value) pairs in MapReduce can make it possible to analyze massive graphs residing on disk [22, 23] by exploiting locality. The complexity involved in explicitly communicating and sharing data to analyze large graphs in BSP and PRAM systems is eliminated in MapReduce because the framework manages the data movement. The result of this simpler programming interface is that it can be more difficult

to express efficient algorithms in MapReduce. But combining both Accumulo and MapReduce is a practical approach for storing, extracting, and analyzing Big Graphs. Here in the Computer and Information Sciences Research Group at NSA, we used this approach to demonstrate a breadth-first search at brain scale, traversing more than 70 trillion edges on a 1 PB graph [24]. This brain-size graph was nearly 20 times larger than the memory capacity in our moderate-size cluster, yet the rate of processing at this scale was the same at the scale of just one trillion edges, which fit entirely in memory.

About the author

Paul Burkhardt is a computer science researcher in the Research Directorate at NSA. He received his PhD from the University of Illinois at Urbana-Champaign. His current research interests are primarily focused on graph algorithms and Big Data analytics.

References

[1] van den Heuvel MP, Kahn RS, Goñi J, Sporns O. "High-cost, high-capacity backbone for global brain communication." *Proccedings of the National Academy of Sciences of the United Sates of America.* 2012. doi: 10.1073/pnas.1203593109 (figure 1(b)).

[2] Thirst. Project: O'Hare Terminal 5 Mezzanine Mural. Client: Westfield Development. Available at: http://www.3st.com/work/terminal-5-murals. (Illustration built using Processing Data by http://OpenFlights.org.)

[3] Vashisht R, Mondal AK, Jain A, Shah A, Vishnoi P, Priyadarshini P, Bhattacharyya K, Rohira H, Ghat AG, Passi A, et al. "Crowd sourcing a new paradigm for interactome driven drug target identification in *Mycobacterium tuberculosis.*" *PLoS ONE.* 2013;7(7):1–11. doi: 10.1371/journal.pone.0039808 (figure 2).

[4] Lyon B. The Opte Project. Map 1 [accessed 2014 Mar]. 2005 Jan 16. Available at: http://www.opte.org/maps/.

[5] Compeau PE, Pevzner PA, Tesler G. "How to apply de Bruijn graphs to genome assembly." *Nature Biotechnology.* 2011;29(11):987–991. doi: 10.1038/nbt.2023 (figure 2).

[6] Hu Y. Matrix: SNAP/web-Google (bipartite graph drawing) [updated 2014 Mar 12]. Available at: http://www.cise.ufl.edu/research/sparse/matrices/SNAP/web-Google.html.

[7] Watts DJ, Strogatz SH. "Collective dynamics of 'small-world' networks." *Nature.* 1998;393(6684):440–442. doi: 10.1038/30918.

[8] Doerr B, Fouz M, and Friedrich T. "Why rumors spread so quickly in social networks." *Communications of the ACM.* 2012;55(6):70–75. doi: 10.1145/2184319.2184338.

[9] Research and Innovative Technology Administration, Bureau of Transportation Statistics, U.S. Department of Transportation. "Table 1-44: Passengers boarded at the top 50 U.S. airports (a)." *National Transportation Statistics.* 2012. Available at: http://www.rita.dot.gov/bts/sites/rita.dot.gov.bts/files/publications/national_transportation_statistics/html/table_01_44.html.

[10] Federal Aviation Administration, U.S. Department of Transportation. *Administrator's Fact Book.* 2012 Jun. Available at: http://www.faa.gov/about/office_org?headquarters_offices/aba/admin_factbook/media/201206.pdf.

[11] Malewiz G, Austern M, Bik AJC, Dehnert J, Horn I, Leiser N, Czajkowski G. "Pregel: A system for large-scale graph processing." In: *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data,* SIGMOD '10; 2010; Indianapolis, IN. p. 135–146. doi: 10.1145/1807167.1807184.

[12] Intel Corporation. "Intel Xeon processor E5-1600/E5-2600/E5-4600 product families datasheet-volume 1." 2012 May. Available at: http://www.intel.com/content/dam/www/public/us/en/documents/datasheets/xeon-e5-1600-2600-vol-1-datasheet.pdf.

[13] Gerhard S, Daducci A, Lemkaddem A, Meuli R, Thiran J, Hagmann P. "The connectome viewer toolkit: An open source framework to manage, analyze, and visualize connectomes." *Frontiers in Neuroinformatics.* 2011;5(3). doi: 10.3389/fninf.2011.00003.

[14] TOP500.org. November 2012. Available at: http://www.top500.org/list/2012/11.

[15] Haring RA, Ohmacht M, Fox TW, Gschwind MK, Satterfield DL, Sugavanam K, Coteus PW, Heidelberger P, Blumrich MA, Wisniewski RW, et al. "The IBM Blue Gene/Q compute chip." *IEEE Micro.* 2012;32(2):48–60. doi: 10.1109/MM.2011.108.

[16] Munro JI, Paterson MS. "Selection and sorting with limited storage." *Theoretical Computer Science.* 1980;12(3):315–323. doi: 10.1016/0304-3975(80)90061-4.

[17] Henzinger MR, Raghavan P, and Rajagopalan S. "Computing on data streams." 1998. DEC Systems Research Center. Technical Report No. 1998-011.

[18] Chiang YJ, Goodrich MT, Grove EF, Tamassia R, Vengroff DE, and Vitter JS. "External-memory graph algorithms." In: *Proceedings of the Sixth Annual ACM-SIAM Symposium on Discrete Algorithms,* SODA '95; 1995; San Francisco, CA. p. 139–149.

[19] Vitter JS, Shriver E. "Algorithms for parallel memory, I: Two-level memories." *Algorithmica.* 1994;12(2–3):110–147. doi: 10.1007/BF01185207.

[20] Chang F, Dean J, Ghemawat S, Hsieh WC, Wallach DA, Burrows M, Chandra T, Fikes A, and Gruber RE. "Bigtable: A distributed storage system for structured data." In: *Proceedings of the Seventh USENIX Symposium on Operating System Design and Implementation,* OSDI '06; 2006; Seattle, WA. p. 205–218. Available at: http://static.usenix.org/event/osdi06/tech/change/chang_html/index.html.

[21] Dean J, Ghemawat S. "MapReduce: Simplified data processing on large clusters." In: *Proceedings of the Sixth Conference on Symposium on Operating Systems Design and Implementation,* OSDI '04; 2004; San Francisco, CA. p. 137–150. Available at: http://research.google.com/archive/mapreduce.html.

[22] Burkhardt P. "Asking hard graph questions." 2014 Feb. US National Security Agency. Technical report No.: NSA-RD-2014-050001v1.

[23] Cohen J. "Graph twiddling in a MapReduce world." *Computing in Science and Engineering.* 2009;11(4):29–41. doi: 10.1109/MCSE.2009.120.

[24] Burkhardt P, Waring C. "An NSA big graph experiment." 2013 May. US National Security Agency. Technical Report No. NSA-RD-2013-056001v1.
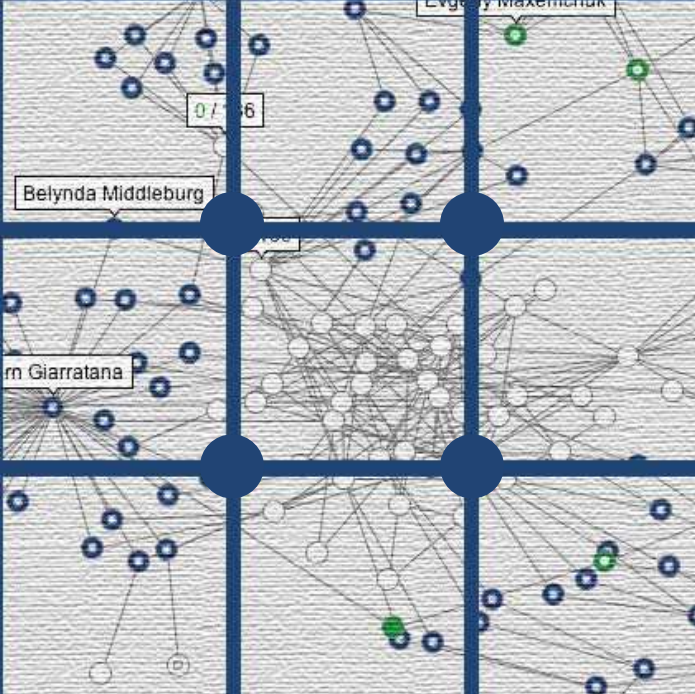
Visual analytics for Big Data

Randall Rohrer
Celeste Lyn Paul
Bohdan Nebesh

Introduction

Randomly browse nearly any publication from nearly any domain in today's information-saturated world and you'll read about Big Data. The term permeates scientific journals, technical magazines, newspapers, social sciences, and to some extent pop culture[a]. Many areas of science and government have been wrestling with Big Data for years. However, as the information age fully radiates into nearly all areas of society and application domains, Big Data problems have pushed into these areas as well. It is common to hear people simultaneously propose that they have a Big Data *problem* as well as a Big Data *opportunity*.

Certainly, new possibilities exist for better solutions and better understanding by successfully analyzing more complete, cross-correlated data sources. Big Data potentially provides an opportunity for better and more complete analysis. However, the growing size and complexity of data also obfuscates and complicates these desired outcomes on many levels. There are obvious complicating technical issues related to data management, information systems, and algorithms. Much of the Big Data publicity and splash focuses on these technical issues of management, access, and computation. There are many efforts aimed at devising new, improved, and more complete analytic capabilities

---

a. Big Data has been the focus of a number of comics such as Dilbert and XKCD.

by taking advantage of these technical advances. Big Data technology and Big Data analytics are rightfully important research focuses.

The problems and opportunities of Big Data affect many domains and applications. An important but under-examined repercussion of the Big Data surge is its effect on analysis. Does (or should) our approach to analysis change when entering this "brave new world" of Big Data? Can we use tried and true analytic approaches on an exponentially increasing scale? Or do we need to approach Big Data analytics in new or different ways to effectively deal with such size and complexity? This is not just a technical question but also a human performance question. Does (or should) the shift to Big Data significantly alter the human cognitive and sense-making approach to analysis?

These are big, hard, long-term research questions that are not easily answered in the immediate future. Each domain, application, and task dealing with Big Data will likely need to examine such questions. Perhaps some general design guidelines and principles can be teased out of the generic problem. The point is that Big Data is not just a technologic issue but also a human cognition and sense-making issue. The human factor is equally important whether you view Big Data as a problem, an opportunity, or both. How can people best analyze or cope with Big Data?

We must address our ability to understand and analyze Big Data and not just our ability to "get" and store more of it[b]. With this larger context in mind, this article narrows its focus to visual analytics for Big Data. How can visual analytics be effectively adapted and applied to Big Data problems? We want to "see" the overall shape, context, and details of our data and be able to analyze and understand the embedded detailed relationships. After first discussing visual analytics and the characteristics of Big Data, we examine issues that Big Data imposes on visualization. There are many interesting research questions and challenges introduced by Big Data that affect visual analytics. We then discuss a number of approaches and strategies

for Big Data visualization that offer promise for addressing these problems.

Unfortunately, this discussion will mostly pose open research questions with suggestions for approach rather than validated findings and solutions. These research questions, suggested approaches, and analytic needs have inspired us to establish a rather broad research agenda to examine these issues. We will close this article with an overview of our visual analytics research program at the National Security Agency (NSA) that aims to address both the problems and opportunities of visual analytics for Big Data.

Visual analytics

Visual analytics is a relatively new field of study that focuses on the tight integration of visualization and analytics. The name was coined by the noteworthy research report *Illuminating the Path: The Research and Development Agenda for Visual Analytics,* published by the National Visualization and Analytics Center (NVAC) in 2005 [1]. NVAC was a Department of Homeland Security (DHS) sponsored research program that aimed to define a long-term research agenda in visual analytics with the intent of improving analysis capabilities. To accomplish and guide its mission, NVAC convened a panel of experts to define a research and development agenda for visual analytics. The result was *Illuminating the Path,* which continues to motivate this field.

As defined in this report, "visual analytics is the science of analytical reasoning facilitated by interactive visual interfaces" [1]. The panel chose its words carefully to focus on analytics that are supported by interactive visualization; analytics is the focus. One very important factor that differentiates visual analytics from its supporting fields of information visualization and human-computer interaction (HCI) is its emphasis on analytical reasoning. Traditional visualization focuses on visual representations and visual mappings of data. HCI focuses on effective human interaction in terms of usability and utility.

---

b. One might imagine the Ark of the Covenant tucked away in some inaccessible government Big Data warehouse as depicted in the film *Raiders of the Lost Ark.*

However, the way interactive visualizations best fit into the analytic cognitive process is often overlooked. The act of using visualization may disrupt an analyst's cognitive work model and interfere with their analytic "flow" if not properly integrated. Visual analytics focuses on analytical reasoning and attempts to integrate visualization throughout the analytic process without violating the analyst's cognitive workflow. Visualization is not just used for presentation or viewing at the end of analysis but rather throughout the entire analytic process.

Humans and computers have inherent strengths and weaknesses. Computers are good at algorithmic calculations at scale but lag behind humans' ability in perceptual intelligent understanding. Some aspects of problems are best suited for fully automated algorithms while others are best accomplished with human-in-the-loop solutions. Humans are superb at visual recognition of subtle patterns, correlations, and differences [2]. Visualization takes advantage of these human abilities and presents data in ways that are optimized for effective perception and understanding. Visual analytics tries to combine and optimize algorithmic analytics with human visual perception skills to our analytic advantage.

Since analysis is very much a multidisciplinary endeavor, visual analytics is also multidisciplinary in nature. It draws heavily from a wide range of fields including information visualization, interaction, cognitive science, knowledge discovery, computer science, mathematics, statistics, perception, and data management, as well as specific problem domains [1].

As a result of widespread, strong interest in visual analytics, a dedicated international academic conference known as the Institute of Electrical and Electronic Engineers (IEEE) Conference on Visual Analytics Science and Technology (VAST) was formed in 2006. This annual conference is part of IEEE VIS (formerly VisWeek), an IEEE-sponsored suite of unified visualization conferences [3]. VAST is the premiere international conference dedicated to visual analytics and an excellent resource to gain insight and inspiration in the field.

Big Data

What is Big Data? How "big" is Big Data? Elsewhere in this issue, Paul Burkhardt provides a comprehensive overview of Big Data and nicely describes the sobering sizes, characteristics, and complexities of this growing beast [4]. Big Data is sometimes characterized in terms of the three Vs: *volume, velocity,* and *variety.* (Here, *variety* relates to complexity.) The volume or size of Big Data is an obvious issue. Scientists in some domains now discuss problems and data set sizes in terms of petabytes (10^{15}), exabytes (10^{18}), and zettabytes (10^{21}). These are truly staggering sizes.

The velocity of data is also increasing. Streaming data is being produced at increasingly fast rates, resulting in the need to dynamically process and analyze such data as it flows. Storing streaming data is not always possible or desirable, and specialized streaming analytics may need to process data flows on the fly. Bigger, faster streaming rates and volumes of data impose challenging requirements on streaming analytics. Finally, variety, or complexity, is another issue of Big Data. As data-producing and -gathering processes become more sophisticated and problem domains become more complex, we are producing and collecting more complex, detailed, multidimensional data sets.

Ironically, some of the added complexity stems from our ability to combine or "mash-up"[c] disparate data sets and dimensions together in new ways in our attempt to perform more sophisticated and complete analytics. Because of newer innovations in flexible data management (i.e., cloud technologies, NoSQL schemaless databases, Hadoop Distributed File System) as well as distributed processing (i.e., MapReduce), there is a tendency and desire to dice data sets into smaller, flexible pieces that can be recombined and cross-correlated in new ways. The result is that these new storage and processing technologies can produce new, fused data sets by mashing up parts of other data sets. This complexity is useful for solving new problems but also complicates analytics. It is ironic that the technologies used to address Big Data storage and processing can also add complexities in the process.

---

c. A data "mash-up" is fused data from disparate data sources.

In the past, the problems of Big Data were restricted to science and government. They were the main entities who had Big Data and needed to analyze it. Scientific fields such as astronomy, physics, meteorology, and medicine produce huge data sets from sampling, experiments, and simulations. Government agencies also produce and acquire large volumes of data. However, with the pervasive explosion of the information age and the Internet into all aspects of society, Big Data issues began appearing in many new application domains.

For example, social sciences, business, and communications now have Big Data issues. The digital footprint of individuals actually accounts for 75% of all digital information [4]. Pervasive digital activities such as e-mail, phone calls, photos, videos, web browsing, social media, and financial transactions account for much of this explosion in data. Because of widespread influence of the digital age, Big Data has now become part of many, if not most, areas of human endeavor. Big Data is everywhere.

Challenges and research questions for Big Data visual analytics

Visual analytics is one of many approaches to analysis. With increasingly complex problems, multiple analysis approaches are often required for successful analysis and understanding of a single problem. For example, machine-learning algorithms might be applied to train analytics to automatically detect and find patterns in data, intelligence-value-estimation algorithms might be applied to rank or score these results, and visual analytics might be inserted throughout this process to steer these algorithms or to present the results for interpretation. Visual presentation may allow one to see latent (i.e., hidden) relationships not detected in algorithmic processes. The emergence of Big Data affects all of these analytic approaches and pieces of the analytic process. Adapting and applying visual analytics to Big Data problems presents new challenges and opens new research questions.

The challenges presented by Big Data for large-scale visual analytics are difficult and numerous. Some are technological (e.g., computation, storage, algorithms, rendering) and some are related to human cognition and perception (e.g., visual

representation, data abstraction and summarization, complexity, scale). Like other human-computer interactions, visual analytics is task-specific. The specific visualizations, analytics, and interactions depend on the intended task. With the extreme scales and new data mash-ups introduced by Big Data, we now have the opportunity to ask different questions of the data. We will likely need to continue to perform previous tasks, but we now have the opportunity and need to perform new tasks.

In years past, we might have explored data relationships in a narrow time period within one data set or perhaps across a few correlated data sets or dimensions. Now with the availability of more complete data and the ability to access all dimensions of data, we can ask the same question across larger and more complete time periods and across all fused dimensions. Perhaps we should now be asking different questions of Big Data? Do we need to formulate and ask questions differently or under the guise of different tasks? Can or should our analytic questions be expressed as higher-level, big-picture questions that are not confined to past restrictions of limited or incomplete data sets? The advent of Big Data presents a new space of analysis that bears further study for optimizing our analytic opportunities and analytic successes.

The challenges that Big Data brings to visual analytics have been carefully examined by a number of leading experts. In fact, we point to several prominent discussions that provide very good summaries of these issues [5, 6, 7, 8]. In 2012, a special theme issue of *IEEE Computer Graphics & Applications* was devoted to extreme-scale visual analytics. This issue included a discussion of the top challenges that Big Data brings to visual analytics in the article, "The top 10 challenges in extreme-scale visual analytics" [5]. The challenges are listed as follows:

1. In situ analysis (in-memory analysis);
2. Interaction and user interfaces;
3. Large-scale data visualization (visual representation);
4. Databases and storage;
5. Algorithms;
6. Data movement, data transport, and network infrastructure;

7. Uncertainty quantification;

8. Parallelism;

9. Domain and development libraries, and tools; and

10. Social, community, and government engagements.

Many of the challenges suggested above are rather general and applicable to most areas of Big Data management, computation, and analytics. The challenges that seem most relevant to visual analytics are visual representation and uncertainty. Visualization of Big Data typically requires constructing abstract visual representations at multiple levels of abstraction and scale. In addition, highly scalable data-projection and dimension-reduction techniques are needed to deal with extreme data scales.

We must be careful that extreme projection and dimension reduction does not hinder the fidelity or the interpretation of the transformed data. In addition, these data transformations will likely lead to more abstract visual representations. As pointed out by Wong et al., "More data projection and dimension reduction in visualization also means more abstract representations. Such representations require additional insight and interpretation for those performing visual reasoning and information foraging" [5]. Such error-prone interpretation can easily become a vicious cycle leading to ineffective analytic thrashing and cognitive overload.

Uncertainty quantification also poses an important challenge. In order to adapt to Big Data, many analytic tasks rely on data subsampling, which introduces even greater uncertainty. Again Wong et al. states, "We must develop analytic techniques that can cope with incomplete data. Many algorithms must be redesigned to consider data as distributions" [5]. So, instead of treating data as discrete samples, we might need to treat sampled data as an aggregated distribution in order to cope with extreme scale.

The cumulative effects of projection, dimension reduction, and distribution representation may introduce new errors or uncertainty into data that likely already contained uncertainty prior to transformation. For this reason, it will be even more important for visualization to accurately convey uncertainty to help users understand risks and to minimize misleading results. Large high-resolution visual displays (e.g., power walls) can be used to aid in large-scale visualization for some tasks but are limiting and do not directly address all issues of Big Data visualization. Interaction and user-interface issues are an inseparable, intertwined problem with visualization. A key question is: How can users effectively interact with uncertain-laden abstract visual representations at multiple scales?

Interaction and user-interface challenges are critical aspects of visual analysis. Within the same publication and also expanded elsewhere, experts discussed interaction challenges in the article, "The top 10 interaction and user interface (UI) challenges in extreme-scale visual analytics" [6, 7]. The challenges are listed as follows:

1. In situ interactive analysis,

2. User-driven data reduction,

3. Scalability and multilevel hierarchy,

4. Representing evidence and uncertainty,

5. Heterogeneous-data fusion,

6. Data summarization and triage for interactive query,

7. Analysis of temporarily evolved features,

8. The human bottleneck,

9. Design and engineering development, and

10. The renaissance of conventional wisdom.

Several of these interaction challenges are of particular interest. One suggested approach is to allow users to steer or control data-reduction steps based on their own practices or analytic needs. This places an added burden on the user, but it does provide flexible control over how the data is transformed for different tasks. Analysis of Big Data often requires the data to be organized into multilevel and multiscale hierarchies. As data scale and complexity grows, the depth and complexity of resulting hierarchies also grow. This makes navigation of these hierarchies even more difficult. If we can improve the mapping between the user task semantics and the fused data semantics, we could greatly improve analysis and make user-interface issues less problematic.

In 1996, Ben Shneiderman proposed the Visual Information-Seeking Mantra (also called the Information Visualization Mantra), which offers a summary of visual design guidelines and a high-level framework for designing information visualization applications [9]. His experience showed that if one follows these simple visual design guidelines, chances are good that the resulting application will be an effective visualization for exploratory analysis. These guidelines embody the basic requirements for crafting a good exploratory visualization:

- Overview first,
- Zoom and filter, and
- Details-on-demand.

In implementing these guidelines, "Overview first" implies that the entire data set should first be displayed to provide a high-level view. This overview is likely abstracted and not every detail is explicitly visible. However, every detail should be represented in some way within the overview. Hence, you are providing the user with a global view of the entire data set. The user can then inspect and interactively explore by zooming into subregions and filtering on attributes of the data ("zoom and filter"). At any point in time, the user should be able to inspect details of abstracted data or zoom to reveal more detail ("details-on-demand"). In the ensuing years, the Information Visualization Mantra seemed to serve designers quite well. If you followed these guidelines, chances were good that you were on the right path in designing a good visualization. You were at least guaranteed a certain level of exploratory functionality.

In the Visual Analytics Research Group at NSA, we have tried to follow these design guidelines. However, with the new challenges introduced by Big Data, the Information Visualization Mantra often falls short in providing effective guidelines for visualizing Big Data. Some researchers now suggest that we need to re-examine these principles and consider new guidelines and conventional wisdom in designing visual analytics for Big Data. Can we find or adapt new guidelines that define a new *visual analytics mantra* that is effective for Big Data?

In addition to the aforementioned resources, we also recommend the 2012 book *Expanding the Frontiers of Visual Analytics and Visualization,*

which lays out future directions, needs, and ideas for research and development in visual analytics [8]. This book, compiled by leading researchers, directs much of its attention specifically to Big Data issues in visual analytics. It is a good resource for ideas and future directions and provides a nice summary of related issues. It is clear that many difficult challenges and open research questions remain.

Approaches and strategies for Big Data visual analytics

In discussing research challenges, we have hinted at some approaches and strategies for developing effective visual analytics for Big Data. Certainly this is an ongoing research issue. Here, we offer a number of ideas that show promise toward meeting this goal. Note that visualization is a cognitive process that happens in the human brain with support of the perceptual system. External visual cues that are embodied through graphics and display technologies help humans track and see abstracted visual representations of data.

Good semantic mappings from external visual cues to internal cognitive processes support human understanding. The better this semantic mapping is, the more likely humans are to benefit from visual analysis. A primary goal is to help people detect and understand both explicit and latent relationships in data as well as to interpret how these relationships inform their analytic task. Visualization designers must use appropriate visual representations, interactions, analytics, and task semantics to construct a visual analytics solution that directly supports the user's intended task and problem semantics. A big part of this process is choosing appropriate visual representations at appropriate scales that match task semantics.

One of the main benefits of a well-constructed visualization is that it provides *context* for data and relationships within the semantic problem space. One can examine a focused subset of data within the full context of surrounding data. It is often this boundary where interesting useful relationships are discovered. As an example, it is often fruitful to see the results of a query or algorithm displayed within the context of surrounding data (i.e., data that does not directly satisfy the algorithmic criteria). Hidden

relationships are often revealed at the boundary of algorithmic analysis and visual analysis. Providing semantic context of data within the entire semantic problem space is a powerful analytic tool. In visualizing Big Data, context is even more important and more difficult to achieve. The approaches we take in developing visual analytics for Big Data should attempt to convey context of data within the whole semantic space.

Abstraction and aggregation

There are a number of strategies that are important or promising for visualizing Big Data. Perhaps the most relevant are abstraction and aggregation.

Abstraction is the process by which data is defined with a representation similar in form to its meaning (i.e., semantics) while hiding away details. Abstraction attempts to reduce and factor out details without losing the semantic concept. We often refer to abstraction as graduated layers of detail with the lowest level containing full details and the highest level containing few details. For example, a subgraph within a full graph (e.g., node-link diagram) might be "rolled up" or collapsed into a single node that denotes the entire subgraph. We would say that the subgraph has been abstracted into a single representative node (see figure 1). Unlike this simple example, abstraction may include complete changes in visual representation. For example, a collection of discrete points might transform into an abstraction of an approximating surface. Abstraction may apply to both data representation as well a corresponding visual representation.
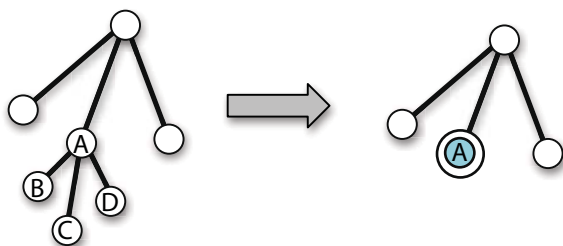

FIGURE 1. In this visual abstraction, subgraph A, B, C, D is collapsed into abstraction A.

Because of the size and complexity of Big Data, many visual abstractions might be constructed across many levels and scales. During visual analysis, we need to arbitrarily traverse and navigate these abstractions at any level or scale. However, it can be tricky to visually show this multilevel navigation without disrupting the user's analytic context or causing the user to lose context. The key is to find a mapping that preserves semantics as faithfully as possible across all levels and scales of abstraction.

Aggregation is a similar concept in that data within a certain bounded region is summarized. Multiple levels of aggregation may be applied into an organized complex structure. Again, as we traverse this structure, we must be careful to preserve context and semantics across this multilevel, multiscale structure. Analytics must be designed to interpret aggregated data at any level within this structure with minimum loss of fidelity and within acceptable error tolerance. Aggregation is a form of data reduction by summarizing subelements or subregions of data. Aggregation is a data summarization process. One could think of aggregation as a form of abstraction.

Alternative approaches

Constructing visual analytics for Big Data requires smart use of abstraction and aggregation for addressing size and complexity issues. In examining this problem, we also note several other approaches that appear promising. For example, Danyel Fisher et al. devised an approach using incremental approximate database queries or queries that operate on progressively larger samples from a database [10]. We can use approximate queries to drive approximate visualizations. By interacting with approximate visualizations, we can steer exploration, successive queries, and underlying analytics toward our analytic goals. This approach uses incremental, interactive steering by the user to explore and refine approximate solutions toward acceptable ones.

Another interesting approach is to transform data into a procedural or functional model. A mathematical procedural model is calculated to approximate a data set. Once the data is encoded in this functional form, the function can be evaluated

at any point in space to produce an approximation of the original values. Function evaluation can be performed to regenerate data at any arbitrary scale or aggregation. Procedural modeling has been used as a compression method for transmission and storage. In some cases, transformations and operations in function space are easier and faster than in the original data space.

We can construct analytics that directly operate on functional representations (i.e., function space) and evaluate the transformed functions or analytic results for visualization. Rendering can be done directly from the functional representation. For example, Yun Jang et al. demonstrated the use of procedural modeling for time-varying data visualization of volumetric data [11]. Their research uses radial basis functions (RBF) and ellipsoidal basis functions (EBF) to encode volumetric data from fluid dynamics simulations into a functional representation. The resulting function is then evaluated and rendered for selected scales of detail. Rendering performance is significantly improved with the additional benefit of selectable scale and detail. It is true that the nature of this data (i.e., time-varying, spatially coherent) is well suited for this approach. For other types of data, this approach may not be as beneficial. However, we believe that functional modeling of Big Data poses an interesting approach that is worth exploring.

In summary, we believe that smart use of abstraction and aggregation is required for effective Big Data visualization. Analytics should be designed to work on data distributions as well as discrete data. Visualization design should consider local and global context and semantics at multiple scales. Several alternative approaches like approximate queries and functional modeling show promise and are worth exploring.

Big Data visual analytics research agenda

Analysis of Big Data is a critical problem for many institutions, including NSA. We believe that visual analytics is an important and necessary part of Big Data analysis. In order to address analytic needs

and answer relevant research questions, we have established a broad research program in visual analytics. We address analysis of Big Data from three fundamental perspectives: the ability to scale **cognitive, visual,** and **computational** components. These three components are critical for visual analytics at scale.

From a cognitive perspective, how does sensemaking differ between traditional-sized data and Big Data? How can visual representations and user interactions scale to maintain effective visual metaphors and semantics? Finally, how can we leverage high-performance computing to enable large-scale analytics and visualization? We use this research framework as an overarching guide for our work in visual analytics. To explore research issues and test hypotheses and ideas, we build prototype systems and evaluate their effectiveness for analysis. Evaluation includes both formal and informal surveys, experiments, and studies. Promising techniques and prototypes developed in the lab migrate to early-deployed versions of visual analytic solutions.

Our visual analytics research program is designed to support analytic discovery, exploration, and situation awareness. It specifically includes ongoing research in graph visualization, text visualization, situation awareness, and mental models of analysis. We are using this research to examine and address the larger research questions and challenges in Big Data visualization as well as improving analysis at NSA. Here, we highlight a selection of this work.

Green Hornet: Large scale graph visualization

Graphs are very useful for modeling and solving many analysis problems in many domains. Elsewhere in this issue, Paul Burkhardt describes Big Graphs[d] and their applications [12]. Current graph visualization software is rather limiting in its ability to scale. Many of today's systems struggle to interactively display graphs comprised of 100,000 nodes/links. With the advent of Big Data, it is important to scale graph visualization capabilities to much larger sizes to meet analytic need.

---

d. Big Graphs are simply graphs that are based on Big Data.

Recall that context is one of the main goals of visualization. We do not necessarily need to see every detail of the graph at once; much of the graph may be visually abstracted. However, the entire graph should be available for interactive exploration. It is important to see focused parts of the graph within the context of the entire graph. To address this specific problem of large-scale graph visualization and to address larger issues of Big Data analysis, we are collaborating with visual analytics researchers at the Pacific Northwest National Laboratory, a Department of Energy lab [13].

Green Hornet is a prototype system that provides interactive visualization and navigation for graphs of size 10–15 million nodes/links (see figures 2, 3, and 4). We recognize that this scale is not quite up to the true scale of Big Data, but it is more than an order of magnitude improvement over current capabilities, and it provides new analysis opportunities that were not previously possible. Interactive

display and navigation is achieved by organizing the underlying data in a hierarchical structure with multiple levels of scale. Traversing this hierarchy at different levels yields varying coarseness of scale.

At high levels in the structure, we find aggregated abstracted data of the subgraph at that location, while at the lowest levels of the structure, we encounter the full detailed graph structure. This structure allows us to show multiple and arbitrary focus areas of interest within the entire graph. That is, based on user navigation and interest, we show detail in some parts of the graph but abstract other parts of the graph by collapsing subgraphs into super nodes (i.e., aggregate nodes). In addition to these basic visualization features, we have added neighborhood chaining, metadata filtering, and several useful graph analytics. Figure 4 shows a shortest path analytic computed within a fictitious paper coauthorship data set. We are also investigating graphics processing unit (GPU) acceleration
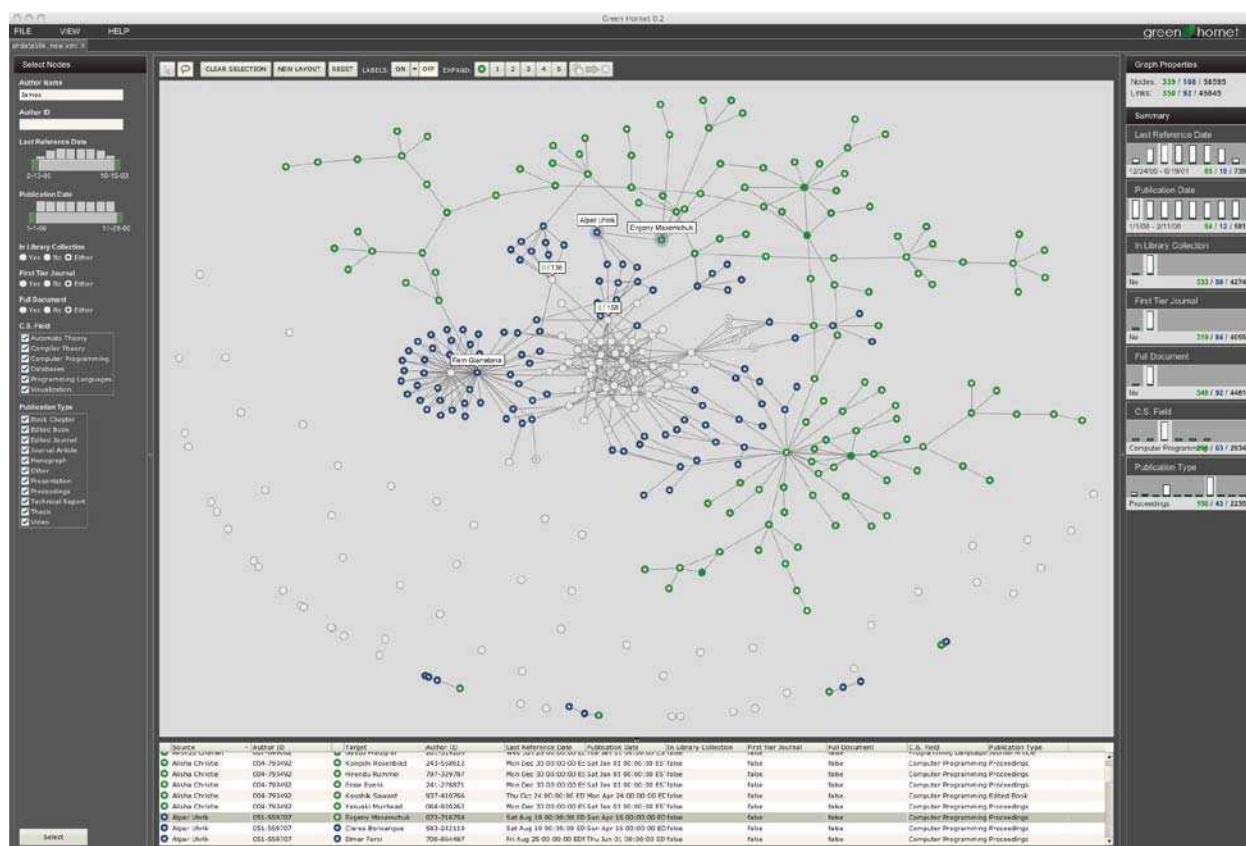

FIGURE 2. Green Hornet is a prototype system that provides interactive visualization and navigation for graphs of size 10–15 million nodes/links. Here, it is visualizing fictional paper coauthorship data.
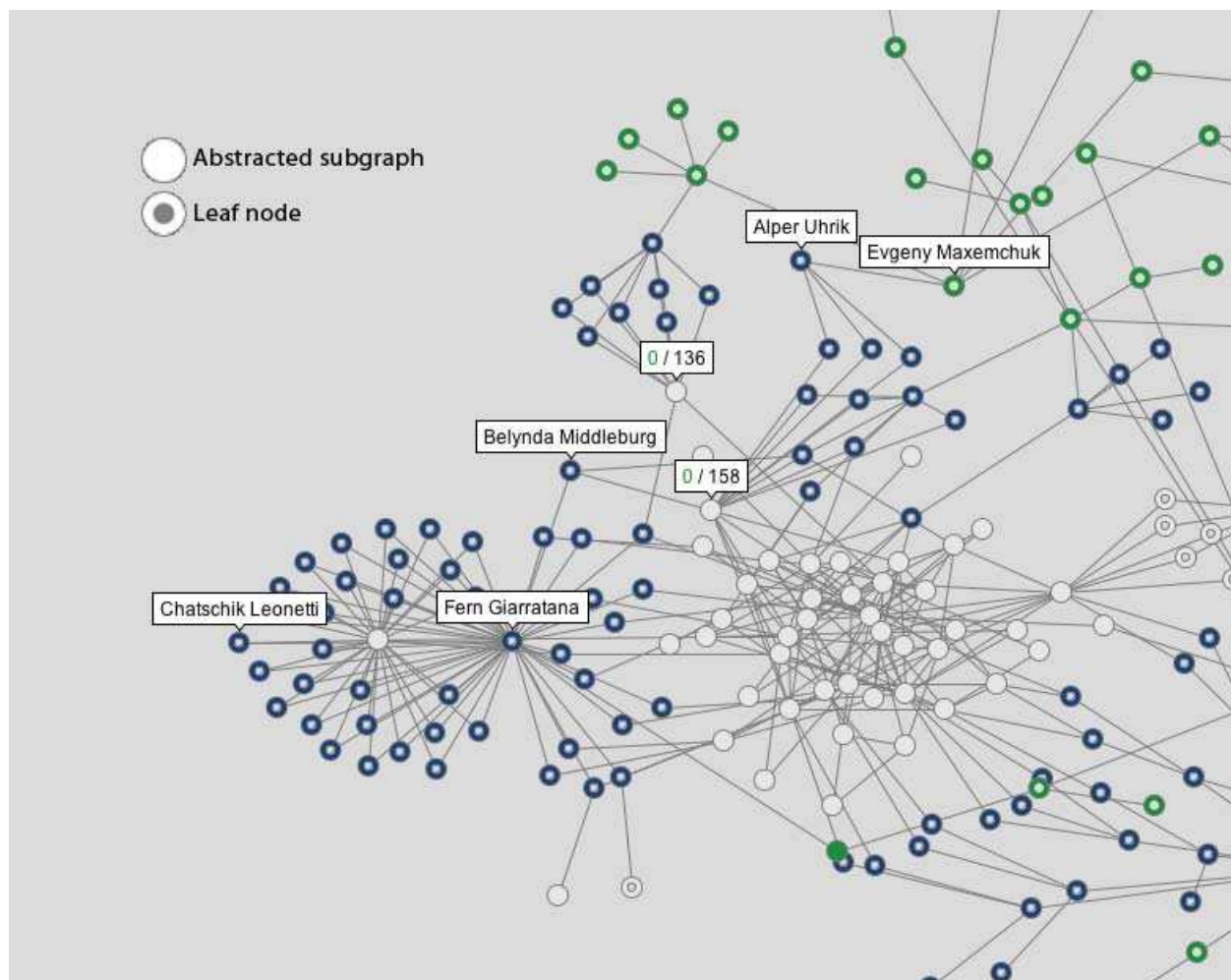
FIGURE 3. Although Green Hornet is not quite up to the true scale of Big Data, it is more than an order of magnitude improvement over current capabilities. Here is a close-up of figure 2, wherein Green Hornet is visualizing fictional paper coauthorship data.

and other means to improve scaling. An early but more thorough description of Green Hornet may be found in [14].

In order to further explore graph visualization of Big Data, we have connected Green Hornet to back-end cloud data sources such as Apache Accumulo [15] as well remote analytic services based on Hadoop MapReduce [16]. Integrating these cloud services provides direct access and analytics for Big Data. Green Hornet provides a good test bed for exploring graph analytics and visual analysis at scale.

Typograph: Visualizing large text repositories

In another research thrust, we are exploring the visualization of very large text repositories. Text is pervasive and an important basis for a lot of analysis. Text may be unstructured (i.e., free-form), structured, or a combination of both. To fully explore and evaluate our visual analytics, we use a test data source that is large (over 14 million articles), multilingual (over 200 languages), evolving, and
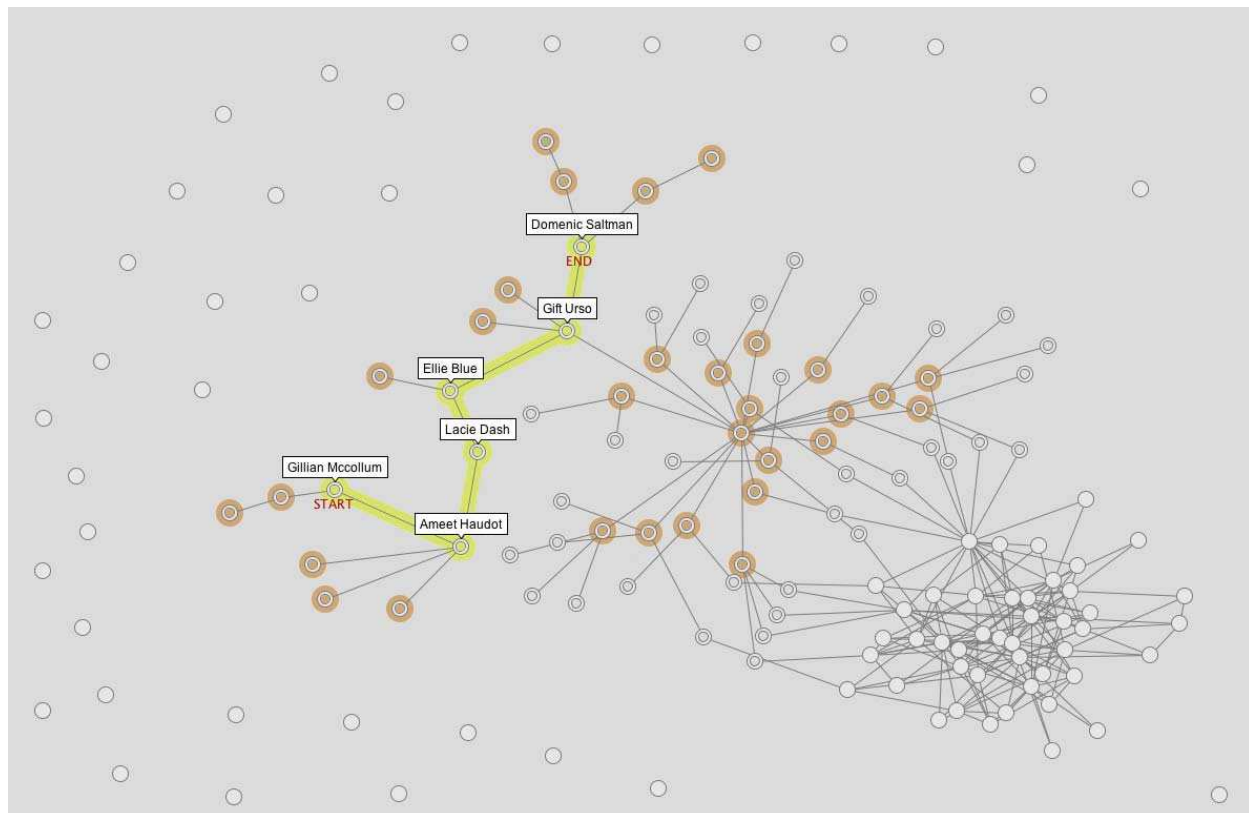
FIGURE 4. Here, Green Hornet is visualizing the shortest path between two nodes (highlighted in yellow) within a fictional paper coauthorship data set.

complex (structured and unstructured text, images, numeric tables). This realistically complex and large test bed serves a good evaluation platform for our research.

Typograph is a prototype for visualizing large text repositories that uses a spatial semantic map metaphor. One can think of it as a "geographic" map of the semantic space of the text collection. It allows global views of an entire text corpus and highlights important terms and regions of interest for further exploration. For our test-bed example data set, we first scan and parse text articles and store the immediate results for subsequent analysis. Text analytics compute the most significant terms and organize results into a multilevel, multiscale cluster hierarchy.

We then visualize the resulting term clusters, exposing related clusters and levels. Users may explore and navigate this semantic space by roaming and zooming in and out of the term space. Zooming

into cluster regions exposes more information including text snippets until detailed document content is revealed at the lowest level. Novel interaction features include semantic interaction and steering by users, landmark navigation based on important terms, and topic queries. Semantic interaction allows a user to indirectly steer the underlying analytics and clustering algorithms with simple direct manipulation actions in the visualization [17].

We are pleased with the current results for interactive exploration of large text repositories, based on our test data set, and have begun using Typograph for exploring other text collections as well. In addition to its direct analytic use, Typograph is a good way to examine how users interact with and use Big Data in text analysis. Figure 5 shows an early design of Typograph prior to our current implementation. Figure 6 shows the current Typograph prototype with an overview of the entire test data set.

FIGURE 5. This is an early conceptual design of Typograph, a prototype for visualizing large text repositories that uses a spatial semantic map metaphor.

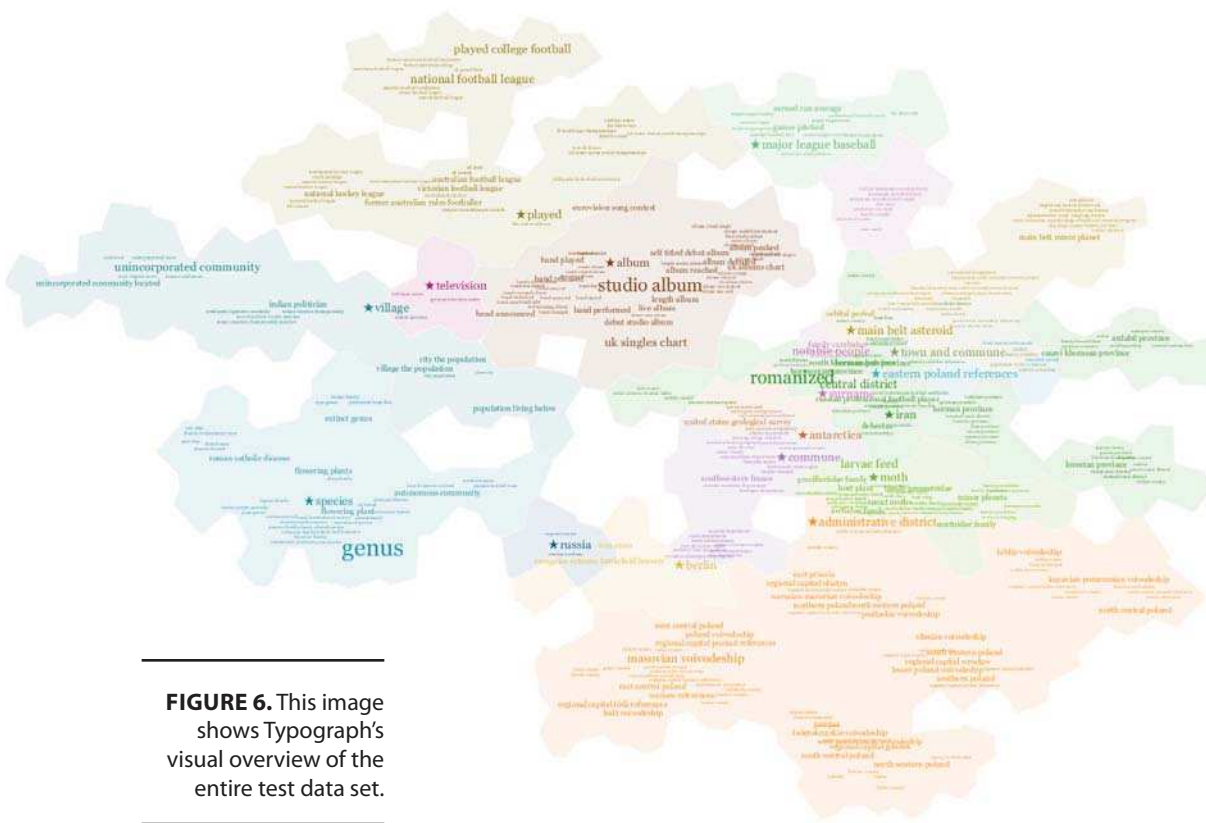**FIGURE 6.** This image shows Typograph's visual overview of the entire test data set.
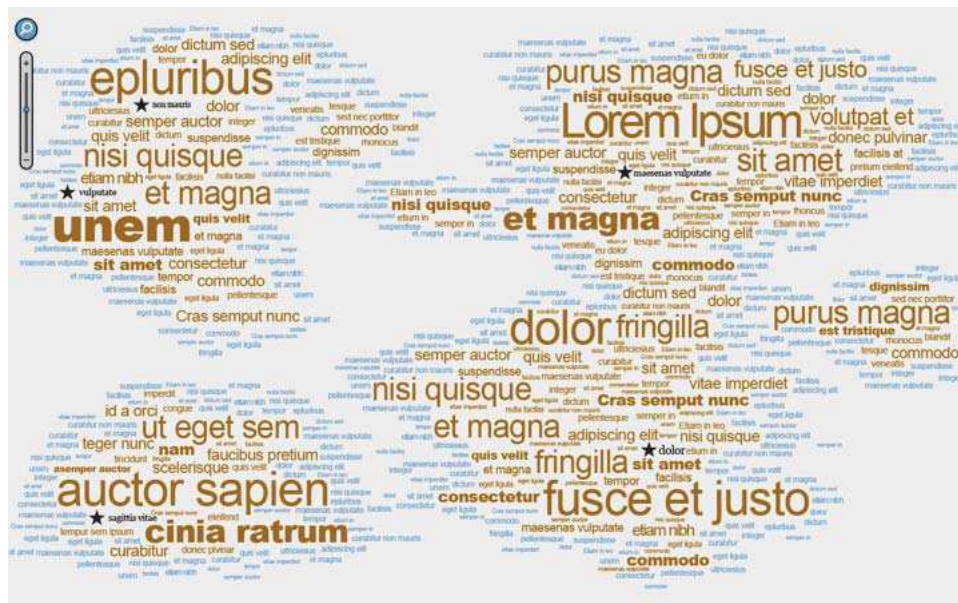
Figure 7 shows a more detailed view as one navigates and zooms into more detailed regions of the information space (in this case the "football" region). Finally, in figure 8, underlying details of the "Australian Rules football" region emerge as text snippets of associated articles appear. We continue to develop Typograph and use it to study visual analytics of Big Text Data.

VAST Challenge: Visualizing large computer networks for cyber situation awareness

The IEEE Visual Analytics Science and Technology (VAST) Challenge is an annual contest hosted by the IEEE VIS conference [18]. The VAST Challenge is designed to provide realistic data scenarios for researchers, academics, students, and industry to develop and test cutting-edge visual analytics tools. For the past three years (i.e., 2011, 2012, 2013), NSA has provided problems that were focused on the challenges of situation awareness for computer networks. Each year, we designed larger and more complex data sets that pushed the limits of data processing, data analytics, and visual display technology.

Additionally, in 2013 we introduced a new design-focused challenge to encourage creative problem solving, good visual design practices, and participation from the art community. Contest participants were asked to design a visualization to support situation awareness of a large network without a complete or sample data set. Instead, they were provided a short story that described a typical day in a network operations center. Participants were free to imagine a network as large and complex as they wanted and were rewarded for creative approaches grounded in reason. This design-first approach forced participants to focus on solving the human analysis problem rather than the technology problem.

The VAST Challenge has resulted in a number of significant contributions to the visual analytics community. Many submissions become the early prototypes for long-term research projects. Several techniques developed for the VAST Challenge have been integrated into longer-term visualization projects [19]. We have also used the results to inform our own research in the Visual Analytics Research Group. Figure 9 depicts a visualization design for situation awareness of large computer networks that is based on results from a related design challenge.

FIGURE 7. This image shows Typograph's detailed view of the "football" region of the data set after interactive navigation.
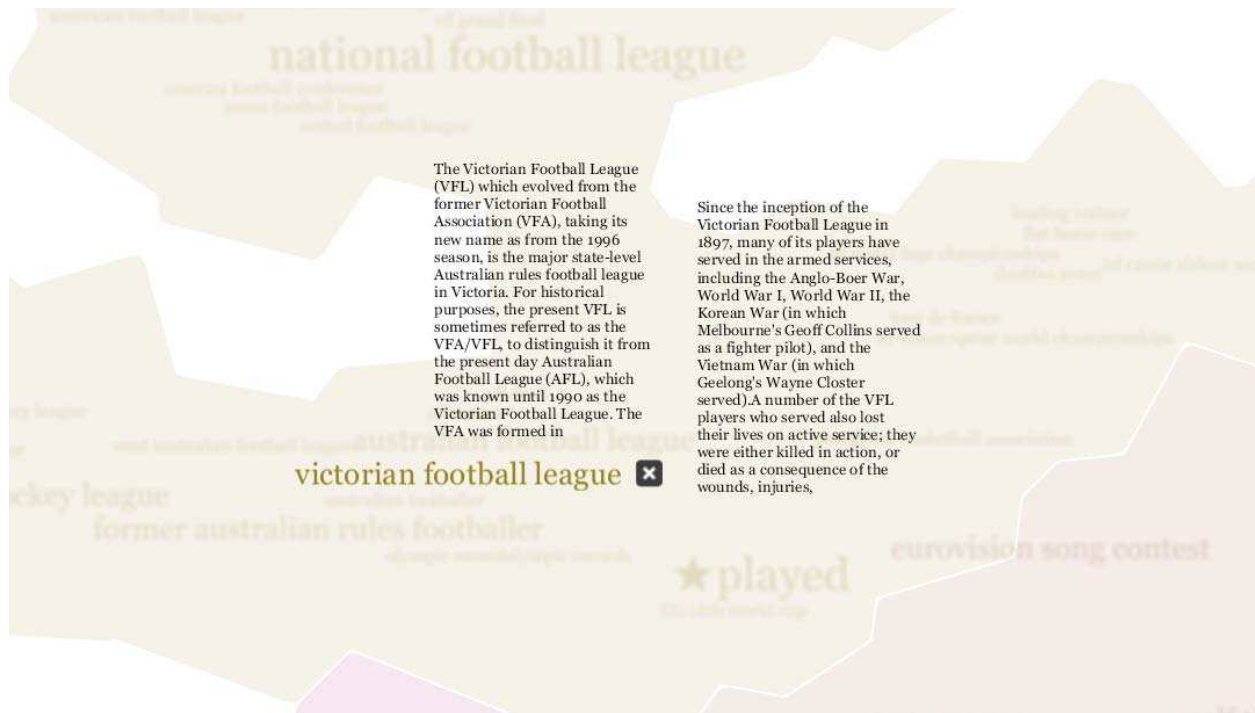
FIGURE 8. This image shows Typograph's view of the "Australian rules football" region with article details (i.e., text snippets).
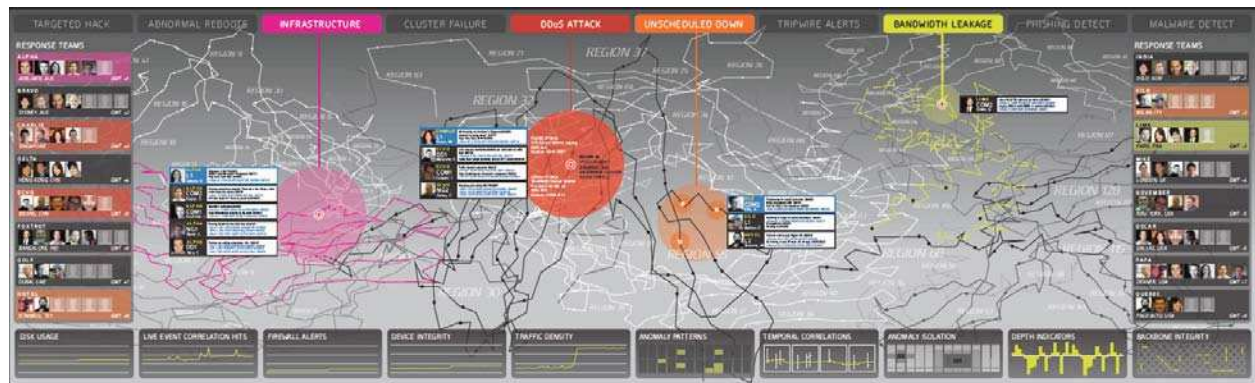

FIGURE 9. This visualization design for situation awareness of large computer networks was submitted by Gong Szeto Design Office.

Using this initial contest design as a basis, we then developed a prototype application for cyber situation awareness of a large computer network. The evolving prototype is shown in Figure 10.

Big Data mental models

Current approaches to visual analytics work well for data that is of reasonable volume and complexity. However, our data challenges go beyond reasonable data volume and complexity into the realm of Really Big Data. As we have discussed, there are a number of challenges associated with analyzing Big Data, even when using supporting visual analytics. How an analyst develops a theory and explores data is dependent on his mental model—an abstract understanding and representation—of what he thinks is in the data. This model becomes skewed or even useless as the data becomes so big and complex that it is beyond imagining. Simply scaling visual
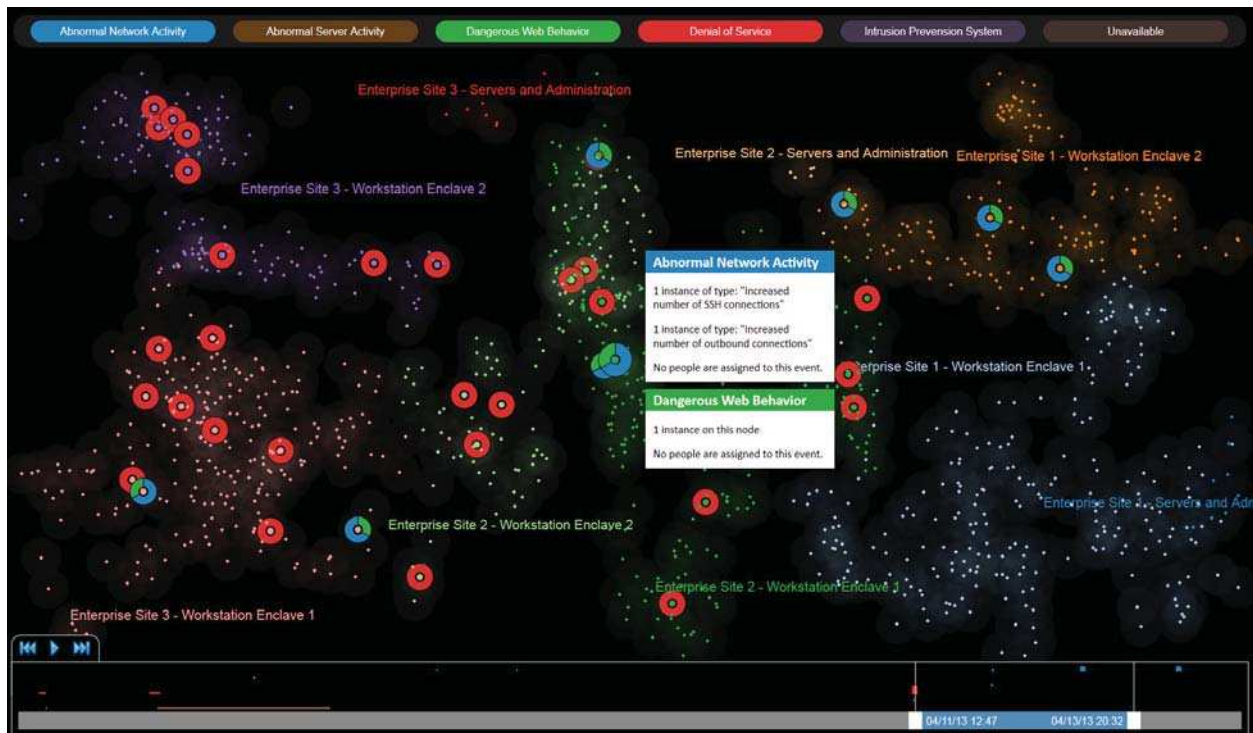
FIGURE 10. Using the visualization design from figure 9, we created this visualization prototype for cyber-situation awareness of large computer networks.

analytics does not address the added visual and cognitive complexity that more data introduces. Big Data is not simply a *more* data problem, but a *different* data problem. Does the *size of data* affect the way we *think about data?*

We are exploring how sense-making in Big Data analysis changes as the volume and complexity of data increases. For example, for those of you old enough, think back to what the world was like before the Internet. Your mental model of the world's knowledge might have been metaphorically the size of a 20-volume encyclopedia or the size of the Library of Congress. Now think about the size of the world's knowledge after the Internet. The Library of Congress is insignificant compared to the vast amounts of information available by a click and keystroke. The way we think about data has changed in less than a generation, and Moore's Law [20] suggests it will change again within our lifetimes. Understanding how we form mental models of Big Data will be essential for developing new visual paradigms to support future Big Data visual analytics. We will then be able to develop new visual analytics that take advantage of understanding

how people think about different volumes, velocities, or varieties of data.

Conclusions

Big Data is pervasive. It presents numerous problems for analysis but also opens new opportunities. We contend that visual analytics is important for analysis of Big Data. In this article, we examined the issues and challenges of visualizing Big Data as well as offered a number of strategies and approaches for effective analysis. Finally, we highlighted our current research program in visual analytics. It is an interesting, if not exciting, time to be living in the age of Big Data. Technical issues of data management, storage, and computation are certainly important. However, supporting analysis of Big Data is most important. We encourage you to share your ideas and your experiences in exploiting visual analytics for understanding Big Data.

Big data. Big data is everywhere. Big Data is good. Let's use visual analytics to understand Big Data.

About the authors

The **Visual Analytics Research** group is a small diverse team within the Computer and Information Sciences Research organization within the NSA Research Directorate. They conduct human-centered research in visual analytics and human-computer interaction (HCI) with primary focus on visual analytics for Big Data. As part of this research, the team develops visual analytics that support discovery, exploration, and situation awareness for intelligence analysis and cybersecurity.

Randall Rohrer is a computer scientist in the Visual Analytics Research group. His main research interests are visual analytics, information visualization, HCI, and computer graphics. He has over 30 years experience in both research and applications of visualization, computer graphics, and other areas of computer science. Mr. Rohrer has a BS degree in computer science from Penn State University and an MS degree in computer science from the Johns Hopkins University. He has also completed significant post-graduate work and published research while attending the George Washington University. He is a member of the Association for Computing Machinery (ACM), the ACM Special Interest Group on Computer Graphics and Interactive Techniques (SIGGRAPH), the ACM Special Interest Group on Computer-Human Interaction (SIGCHI), and the IEEE Computer Society.

Dr. Celeste Lyn Paul is a computer scientist in the Visual Analytics Research group. Her main research interests include HCI, information visualization, Big Data, and cyber. Dr. Paul received her BA degree in multimedia from Duquesne University, her MS degree in interaction design and information architecture from the University of Baltimore, and her PhD in human-centered computing from the University of Maryland Baltimore County. She is a member of the ACM and the ACM SIGCHI.

Dr. Bohdan Nebesh is a computer scientist who leads the Visual Analytics Research group. His main research interests include visual analytics, information visualization, HCI, and software agents. He has over 25 years experience in both research and applications of visualization, software agents, and other areas of computer science. He received his BS degree in computer engineering from Case Western Reserve University, an MS degree in computer science from the Johns Hopkins University, and a PhD in computer science from the George Washington University.

References

[1] Thomas JJ, Cook K, editors. *Illuminating the Path: The Research and Development Agenda for Visual Analytics.* IEEE Computer Society Press; 2005. ISBN: 0-7695-2323-4.

[2] Ware C. *Information Visualization, Third Edition: Perception for Design.* Waltham (MA): Morgan Kaufman; 2012. ISBN-13: 978-0-123-81464-7.

[3] For more information about the IEEE visualization conferences, visit http://ieeevis.org.

[4] Burkhardt P. "An overview of Big Data." *The Next Wave.* 2014;20(4).

[5] Wong PC, Shen H, Johnson CR, Chen C, Ross RB. "The top 10 challenges in extreme-scale visual analytics." *IEEE Computer Graphics and Applications.* 2012;32(4):63–67. doi: 10.1109/MCG.2012.87.

[6] Wong PC, Shen H, Pascucci V. "Extreme-scale visual analytics." *IEEE Computer Graphics and Applications.* 2012;32(4):23–25. doi: 10.1109/MCG.2012.73.

[7] Wong PC, Shen H, Chen C. "Top ten interaction challenges in extreme-scale visual analytics." In: Dill J, Earnshaw R, Kasik D, Vince J, Wong PC, editors. *Expanding the Frontiers of Visual Analytics and Visualization.* Springer-Verlag; 2012. p. 197–207. doi: 10.1007/978-1-4471-2804-5_12.

[8] Dill J, Earnshaw R, Kasik D, Vince J, Wong PC, editors. *Expanding the Frontiers of Visual Analytics and Visualization.* Springer-Verlag; 2012. ISBN-13: 978-1-4471-2803-8.

[9] Shneiderman B. "The eyes have it: A task by data type taxonomy for information visualizations." In: *Proceedings of the IEEE Symposium on Visual Languages.* Washington, DC: IEEE Computer Society Press; 1996. p. 336–343. doi: 10.1109/VL.1996.545307.

[10] Fisher D, Drucker SM, König AC. "Exploratory visualization involving incremental, approximate database queries and uncertainty." *IEEE Computer Graphics and Applications.* 2012;32(4). p. 55–62. doi: 10.1109/MCG.2012.48.

[11] Jang Y, Ebert DS, Gaither K. "Time-varying data visualization using functional representations." *IEEE Transactions on Visualization and Computer Graphics.* 2012;18(3):421–433. doi: 10.1109/TVCG.2011.54.

[12] Burkhardt P. "Big Graphs." *The Next Wave.* 2014;20(4).

[13] For more information about the Pacific Northwest National Laboratory, visit http://www.pnnl.gov.

[14] Wong PC, Mackey P, Cook KA, Rohrer RM, Foote H, Whiting MA. "A multi-level middle-out cross-zooming approach for large graph analytics." In: *Proceedings of IEEE Symposium on Visual Analytics Science and Technology (VAST 2009).* Washington, DC; IEEE Computer Society Press; 2009. doi: 10.1109/VAST.2009.5333880.

[15] For more information about Apache Accumulo, visit http://accumulo.apache.org.

[16] For more information about Hadoop MapReduce, visit http://hadoop.apache.org.

[17] Endert A, Flaux P, North C. "Semantic interaction for visual text analytics." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'12).* New York, NY: ACM; 2012. doi: 10.1145/2207676.2207741.

[18] Cook K, Grinstein G, Whiting M, Cooper M, Havig P, Ligget K, Nebesh B, and Paul CL. "VAST Challenge 2012: Visual analytics for Big Data." In: *Proceedings of IEEE Visual Analytics Science and Technology (VAST 2012).* Washington, DC: IEEE Computer Society Press; 2012. doi: 10.1109/VAST.2012.6400529.

[19] J. Scholtz MA, Whiting CP, Grinstein G. "A reflection on seven years of the VAST Challenge." In: *Proceedings of BELIV 2012, Beyond Time and Errors: Novel Evaluation Methods for Information Visualization.* New York, NY: ACM; 2012. doi: 10.1145/2442576.2442589.

[20] Moore GE. "Cramming more components onto integrated circuits." *Electronics.* 1965;38(8):114–117. doi: 10.1109/JPROC.1998.658762.

3 Mira

Specs:	IBM BlueGene/Q, Power BQC 16C 1.6 GHz
Country:	US
Site:	Argonne National Laboratory
No. of Nodes:	49,152
No. of Cores:	786,432
Problem Scale:	40
GTEPS:	14,328

7 Blue Joule

Specs:	IBM BlueGene/Q, Power BQC 16C 1.6 GHz
Country:	United Kingdom
Site:	Daresbury Laboratory, Science and Technology Facilities Council
No. of Nodes:	4,096
No. of Cores:	65,536
Problem Scale:	36
GTEPS:	1,427

7 DIRAC

Specs:	IBM BlueGene/Q, Power BQC 16C 1.6 GHz
Country:	United Kingdom
Site:	University of Edinburgh
No. of Nodes:	4,096
No. of Cores:	65,536
Problem Scale:	36
GTEPS:	1,427

7 Zumbrota

Specs:	IBM BlueGene/Q, Power BQC 16C 1.6 GHz
Country:	France
Site:	EDF R&D
No. of Nodes:	4,096
No. of Cores:	65,536
Problem Scale:	36
GTEPS:	1,427

2 Sequoia

Specs:	IBM BlueGene/Q, Power BQC 16C 1.6 GHz
Country:	US
Site:	Lawrence Livermore National Laboratory
No. of Nodes:	65,536
No. of Cores:	1,048,576
Problem Scale:	40
GTEPS:	16,599

7 Turing

Specs:	IBM BlueGene/Q, Power BQC 16C 1.6 GHz
Country:	France
Site:	CNRS/IDRIS-GENCI
No. of Nodes:	4,096
No. of Cores:	65,536
Problem Scale:	36
GTEPS:	1,427

GLOBE AT A GLANCE

The Graph500 top 11 supercomputers

The Graph500 list ranks supercomputers based on their ability to handle data-intensive loads, also known as Big Data. While the TOP500 list also ranks supercomputers, it ranks them by measuring how fast they can solve linear equations—a good metric for evaluating how a computer system will perform traditional modeling and simulation tasks. But computer scientists are increasingly using supercomputers to analyze massive data sets, and the Graph500 list ranks computer systems using a benchmark that addresses this specific task.

6 Tianhe-2

Specs:	National University of Defense Technology - MPP
Country:	China
Site:	Changsha, China
No. of Nodes:	8,192
No. of Cores:	196,608
Problem Scale:	36
GTEPS:	2,061.48

4 JUQUEEN

Specs:	IBM BlueGene/Q, Power BQC 16C 1.6 GHz
Country:	Germany
Site:	Forschungszentrum Juelich
No. of Nodes:	16,384
No. of Cores:	262,144
Problem Scale:	38
GTEPS:	5,848

5 Fermi

Specs:	IBM BlueGene/Q, Power BQC 16C 1.6 GHz
Country:	Italy
Site:	Cineca
No. of Nodes:	8,192
No. of Cores:	131,072
Problem Scale:	37
GTEPS:	2,567

1 K computer

Specs:	Fujitsu custom supercomputer
Country:	Japan
Site:	RIKEN Advanced Institute for Computational Science
No. of Nodes:	65,536
No. of Cores:	524,288
Problem Scale:	40
GTEPS:	17,977.1

7 Avoca

Specs:	IBM BlueGene/Q, Power BQC 16C 1.6 GHz
Country:	Australia
Site:	Victorian Life Sciences Computation Initiative
No. of Nodes:	4,096
No. of Cores:	65,536
Problem Scale:	36
GTEPS:	1,427

For the benchmark, each supercomputer is given a massive set of data to crunch, called a graph. A graph consists of several interconnected sets of data, with vertices and edges, similar to what a map of your Facebook network might look like. A user would represent a vertex, while the connection between two users would represent an edge. Starting with one vertex, a supercomputer is charged with discovering all other vertices in the graph by following each edge. The speed with which a supercomputer accomplishes this task, measured in gigateps (GTEPS), or billions of traversed edges per second, determines how high it ranks on the Graph500 list. The following ranking is from June 2014; the list in its entirety is available at www.graph500.org.

POINTERS

NSA nurtures growth of a science of security community

The National Security Agency (NSA) actively supports research to develop a scientific approach to cybersecurity. Currently, NSA has several initiatives to stimulate and encourage advanced work on the emerging field termed a science of security (SoS).

First, the Research Directorate announced the winner of the 2013 Science of Security (SoS) Best Scientific Cybersecurity Paper Competition in August 2014. The competition reflects the Agency's desire to increase scientific rigor in cybersecurity.

This year's winner, "Memory Trace Oblivious Program Execution," was a research paper presented at the 2013 IEEE Computer Security Foundation written by Chang Liu, Dr. Michael Hicks, and Dr. Elaine Shi. Their research centered on a scientific foundation for the use of oblivious random-access memory, or ORAM, in programs. Two aspects of the paper were compelling to the reviewers: First, it builds a bridge between cryptographic research and information flow research, and shows how the latter can help one apply cryptographic advances in a principled and secure manner. Second, it establishes a scientific foundation for the use of ORAM in programs. It provides a valuable and exciting direction toward making ORAM practical.

Of the 35 papers nominated, one paper received honorable mention, "Rethinking SSL Development in an Appified World" by Sascha Fahl, Marian Harbach, Henning Perl, Markus Koetter, and Dr. Matthew Smith from the Distributed Computing and Security Group at Leibniz University in Hannover, Germany. Their paper was presented at the 2013 ACM Conference on Computer and Communications Security. The paper studies the possible causes of SSL problems on appified platforms. The results show that the root causes are not simply careless developers, but also the limitations and issues of the

NSA is funding SoS lablets at Carnegie Mellon University, North Carolina State University, the University of Illinois at Urbana-Champaign, and the University of Maryland to advance cybersecurity science. In this photo (from left to right): Mr. Gil Nolte, chief of NSA Trusted Systems Research; Dr. Jonathan Katz, principal investigator for the UMD SoS lablet; Dr. Laurie Williams, principal investigator for the NC State SoS lablet; Dr. David Nichol, principal investigator for the UIUC SoS lablet; and Dr. William Scherlis, principal investigator for the CMU SoS lablet.

current SSL development paradigm. The authors took an unusual step which was highly important—they systematically contacted developers who had produced insecure code.

Additional details about this year's competition can be found at the SoS Virtual Organization website (http://cps-vo.org/group/sos/papercompetition).

Second, NSA has become more involved in academic collaborations, such as the Symposium and Bootcamp on the Science of Security (Hot SoS), hosted in April 2014 by the North Carolina State University SoS lablet, and held in Raleigh, North Carolina. Hot SoS is a research event centered on developing an SoS that addresses the fundamental problems of cybersecurity. Cybersecurity has been intensively studied, but previous research often emphasizes the engineering of specific solutions without first developing a scientific understanding of the problem. All too often, cybersecurity research conveys the flavor of identifying specific threats and removing them one at a time. The motivation behind the nascent SoS is to develop basic cybersecurity properties using scientific rigor to understand how to determine trust in systems. At Hot SoS, researchers from all over the country presented 12 papers and 23 posters. These presentations can be found at http://www.hotsos.org/2014/proceedings.html.

Third, NSA provided $8.2 million in direct support to Carnegie Mellon University, North Carolina State University, the University of Illinois at Urbana-Champaign, and the University of Maryland for SoS research lablets. Within this program, each lablet (i.e., a small lab) will be conducting basic foundational research, building a growing community of researchers from multidisciplines and various universities, and championing the need for an SoS. They will be identifying scientific principles upon which to base trust in cybersecurity. The overarching goal is to bring scientific rigor to research in the cybersecurity domain. The research will focus on five hard problem areas: 1) scalability and composability, 2) policy-governed secure collaboration, 3) security metrics, 4) resilient architectures, and 5) understanding and accounting for human behavior.

Carnegie Mellon University SoS lablet

The Carnegie Mellon University (CMU) SoS lablet addresses cybersecurity challenges related to all five hard problems with particular emphasis on scale and composability of modeling and reasoning, and human behavior and usability for developers, evaluators, operators, and end users. One anticipated result is progress in identifying and sharing the most effective theoretical and experimental approaches to address the scientific challenges within the five

hard problems. A second anticipated result is a better understanding of how to design and choose appropriate modeling abstractions in cybersecurity research. A third anticipated result is the identification of patterns and best practices in the way we carry out cybersecurity research, including approaches to data gathering, analysis, nomenclature, and means to promote reproducibility, enabling more rapid advances in the scientific field. Dr. William Scherlis is the principal investigator for the CMU SoS lablet.

The CMU SoS lablet projects include

- Safe programming languages,
- Binary and source code analysis,
- Data-intensive systems analysis,
- Self-healing and resilient architecture,
- Assured application programming interface (API) and framework compliance,
- Sociotechnical ecosystems,
- Development environments,
- Trusted computing,
- Specification and verification,
- Concurrent and distributed systems,
- Requirements and policy,
- Usable security and privacy,
- Intrusion and malware detection,
- Dynamic network analysis,
- Model checking,
- Secure coding practice,
- Secure process separation, and
- Verification of cyberphysical systems.

North Carolina State University SoS lablet

The North Carolina State University (NC State) SoS lablet is housed in the Institute for Next Generation IT Systems and will contribute broadly to the development of an SoS while leveraging NC State's expertise and experience in analytics, including the extensive expertise available in the NC State Institute of Advanced Analytics. The lablet's work draws on several fundamental areas of computing

research and on the related analytics. Some ideas from fault-tolerant computing will be adapted to the context of cybersecurity. Strategies from control theory will be extended to account for the high variation and uncertainty that may be present in systems when they are under attack. Game theory and decision theory principles will be used to explore the interplay between attack and defense. Formal methods will be applied to develop formal notions of cybersecurity resiliency. End-to-end system analysis will be employed to investigate resiliency of large systems against cyberattack. The lablet's work will draw upon ideas from other areas of mathematics, statistics, and engineering. Dr. Laurie Williams is the principal investigator for the NC State SoS lablet.

The NC State SoS lablet projects include

- Understanding attack surface vulnerabilities,
- Policy complexity and norms,
- Resilience requirements,
- Human information processing, and
- Metrics for security models.

University of Illinois at Urbana-Champaign SoS lablet

The University of Illinois at Urbana-Champaign (UIUC) SoS lablet, housed in the Information Trust Institute, addresses each of the five hard problems. Dr. David M. Nicol, the lablet's principle investigator, explains, "We have a portfolio of projects that will advance the science of security in both experimental and theoretical methodologies, and includes explicit consideration of both mechanized and human elements in SoS models."

The UIUC SoS lablet projects include

- Models and analysis of resiliency to intrusion in cyberphysical systems;
- Models of system and attacker behavior based on data analytics, with application to detecting the presence of intrusion prior to full-scale attack;
- Methodologies for supporting experimental evaluation of network security properties across network layers;

‣ Models and analysis of system and human behavior to support decision making in security contexts; and

‣ A science of human circumvention of security.

University of Maryland SoS lablet

The University of Maryland (UMD) SoS lablet leverages the resources of the Maryland Cybersecurity Center and brings together 15 UMD faculty from five departments, in collaboration with six external faculty from other universities, to focus on developing scientific foundations of cybersecurity. Principal investigator Dr. Jonathan Katz, professor of computer science and director of the Maryland Cybersecurity Center, says the lablet will "establish mathematical models that can be used to address cybersecurity threats broadly, carry out empirical studies to help validate those models, and develop formal techniques for reasoning about the security of large systems built from multiple components."

Particular research strengths of the lablet include using mathematical and formal tools for studying the verification and composition of security properties; conducting empirical studies based on real-world data about vulnerabilities, exploits, and end-host configurations; and understanding the role of human behavior in cybersecurity, both from the perspectives of honest users as well as attackers. Beyond the research, the lablet will also work to grow the SoS community by sharing its results with the broader public and holding workshops and tutorials.

The UMD SoS lablet projects include

‣ Verification of hyperproperties;

‣ Trustworthy and composable software systems with contracts;

‣ Empirical models for vulnerability exploits;

‣ Human behavior and cyber vulnerabilities;

‣ Whether the presence of honest users affects intruders' behavior;

‣ User-centered design for security;

‣ Understanding developers' reasoning about privacy and security;

‣ Trust, recommendation systems, and collaboration; and

‣ Reasoning about protocols with human participants.

SPIN⚙UTS

News from the Technology Transfer Program

Finding and correcting errors in Big Data

"Linear interpolative coding" invention helps find and fix errors in garbled, missing, or repeated data

[Photo credit: alphaspirit/iStock/Thinkstock]

A recent NSA invention will improve the quality of large data sets by detecting and correcting errors even in a datum or value that lacks explicit error-correcting code.

Even under the best of circumstances, it can be difficult to find errors in data sets—especially when those data sets contain millions or billions of records. The problem becomes worse when the errors are not overt, and even "look right." (Think of all the times, for example, when spell-check software passes over words used incorrectly—such as *their* for *there* or *principal* for *principle*—because each of those items appears in the dictionary.)

An NSA researcher has developed a way to resolve this issue by estimating a data point's correct value based on surrounding data. US Patent 8,539,307 was recently granted for a "Device for

and method of linear interpolative coding." This involves finding a set of numbers that can be multiplied by the data points to the right and left of the targeted sample. The system can then use this contextual data to "interpolate" or infer the target's true value. This makes it easier to detect and correct garbled data, missing or deleted data, and added or repeated data. (For examples, see table 1.)

For this approach to work properly, the data cannot be totally random—that is, they must contain natural "context" information that can be modeled in linear fashion. Data must also be statistically stationary or vary only slowly over time.

For example, the sentence in table 1 should read, "There are lots of possibilities." The kinds of errors that the invention could detect and correct are summarized in table 1.

The technique described in the patent has been used in real situations involving all the scenarios described in the table. The invention would benefit data acquisition and forwarding, as well as telecommunications equipment that interfaces between different timing and communications systems. To arrange a demonstration, please contact the Technology Transfer Program at tech_transfer@nsa.gov or 1-866-680-4539.

TABLE 1. "Linear interpolative coding" error detection

Error Type Detected	The Processing Problem	Example
Garbled Data	Datum is present, but its value is incorrect (e.g., transmission error due to noise in transmission channel).	"There **ade** lots of possibilities."
Missing or Deleted Data	Datum is supposed to be present, but is not (e.g., synchronization error between transmitter and receiver).	"There **ae** lots of possibilities."
Added or Repeated Data	Datum is present, but is not supposed to be (e.g., perceived synchronization error between transmitter and receiver).	"There **arre** lots of possibilities."