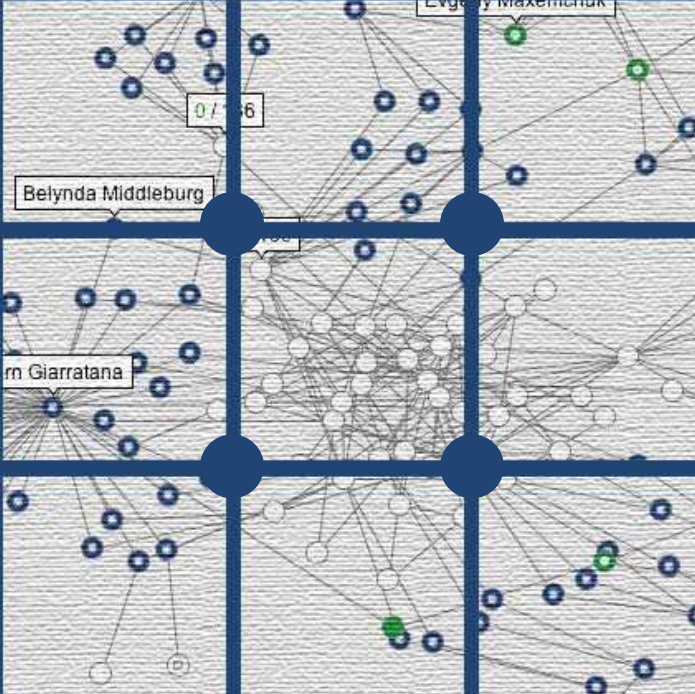




“ Big Data. Big Data is everywhere. Big Data is good. Let’s get more data. ”

~Anonymous analyst~

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 as a corresponding visual representation.

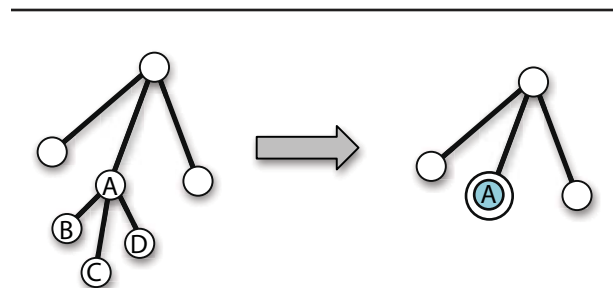


FIGURE 1. In this visual abstraction, subgraph A, B, C, D is collapsed into abstraction A.

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 sense-making 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

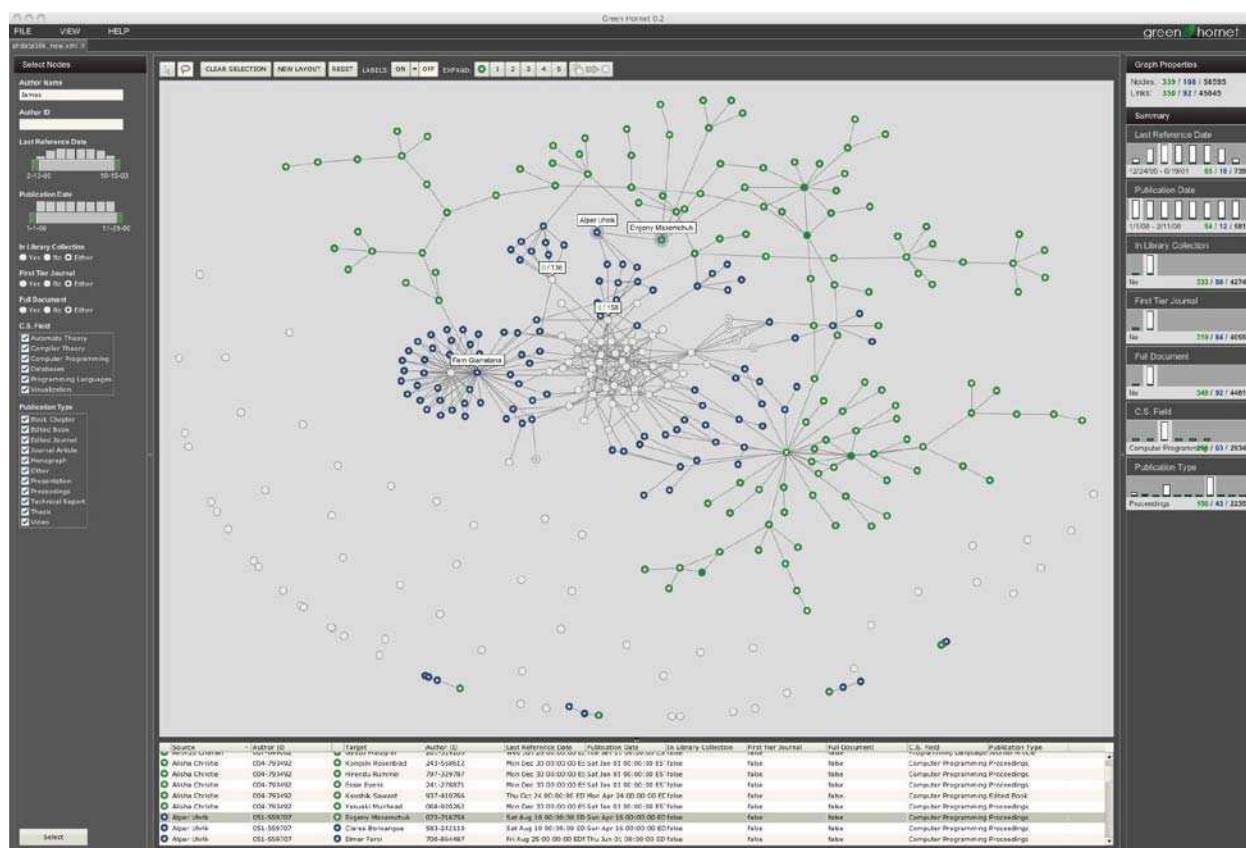


FIGURE 2. Green Hornet is a prototype system that provides interactive visualization and navigation for graphs of size 10–15 million nodes/links. Here, it is visualizing fictional paper coauthorship data.

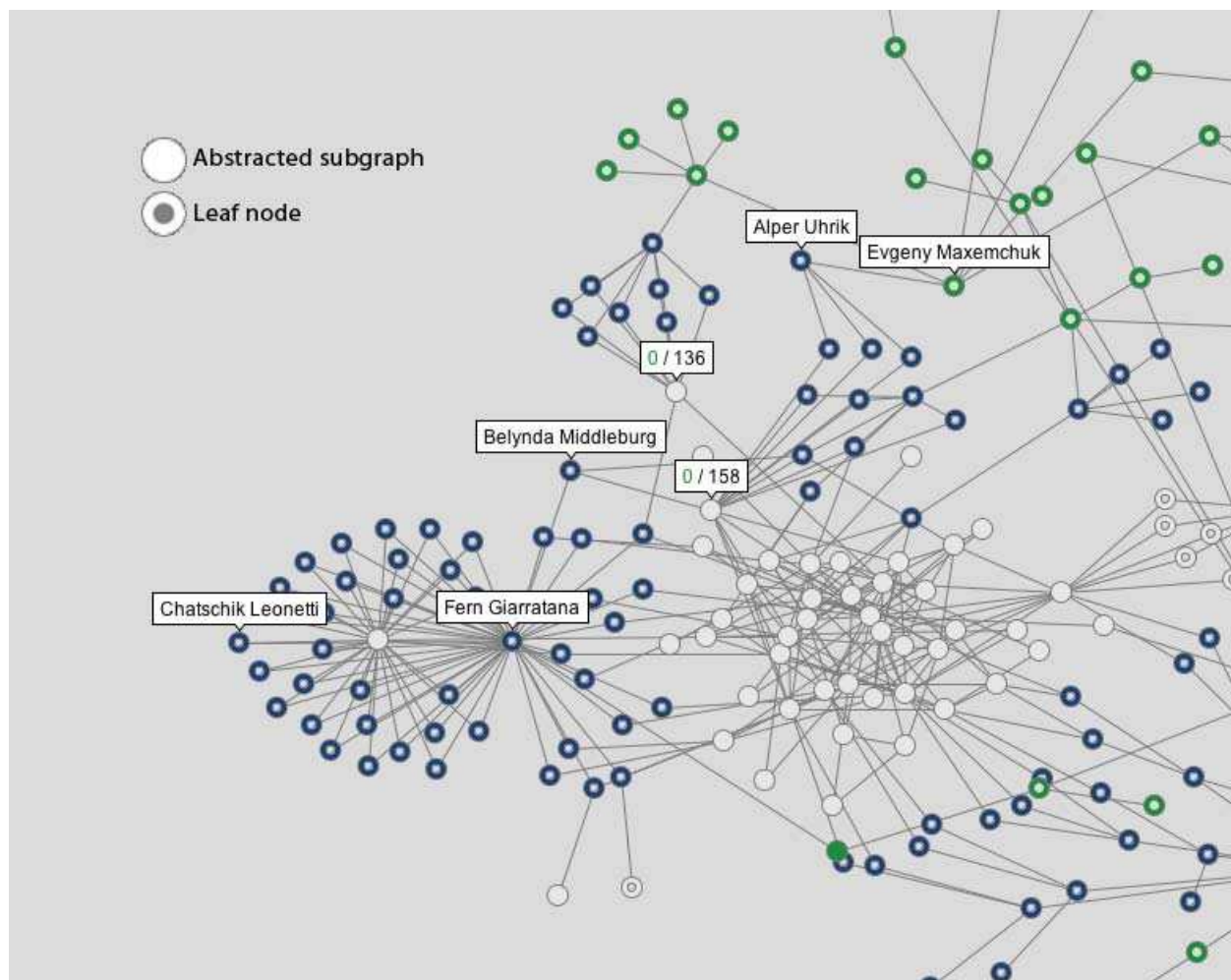


FIGURE 3. Although Green Hornet is not quite up to the true scale of Big Data, it is more than an order of magnitude improvement over current capabilities. Here is a close-up of figure 2, wherein Green Hornet is visualizing fictional paper coauthorship data.

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

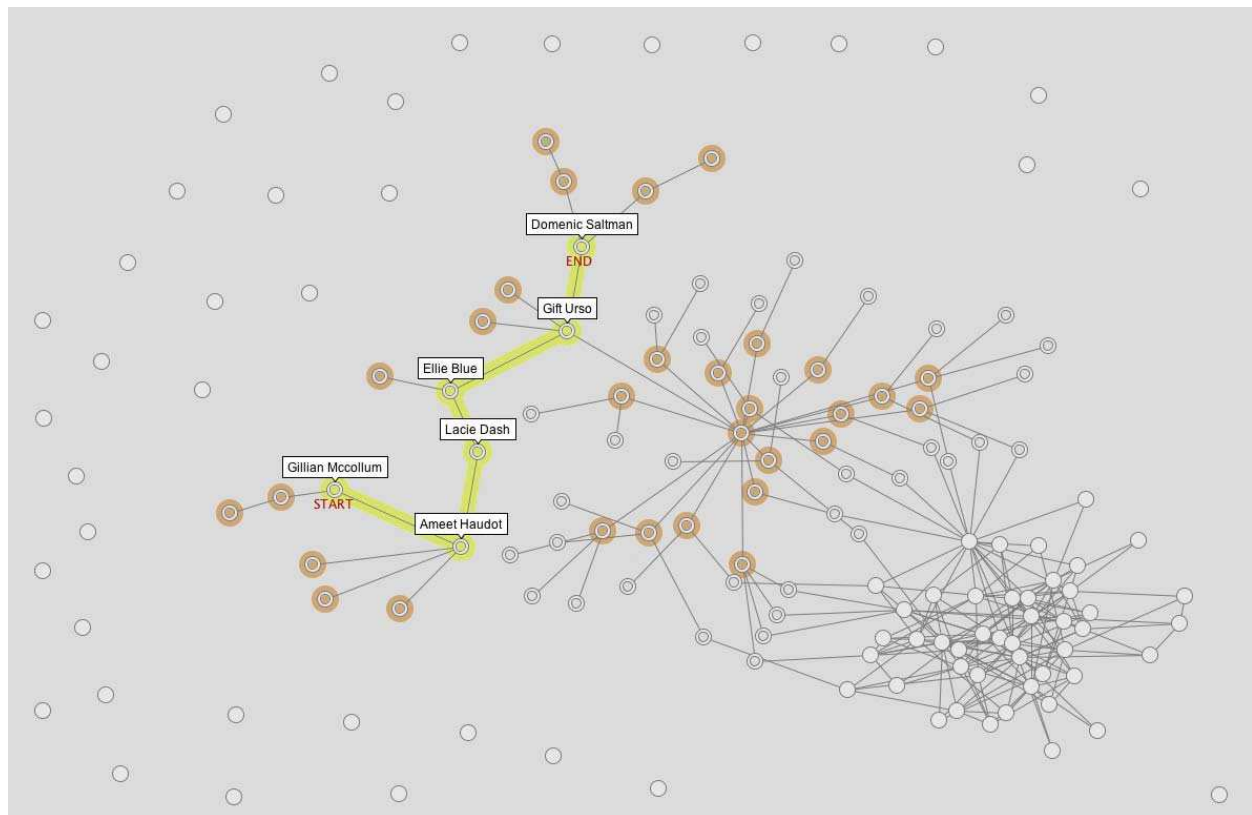


FIGURE 4. Here, Green Hornet is visualizing the shortest path between two nodes (highlighted in yellow) within a fictional paper coauthorship data set.

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.

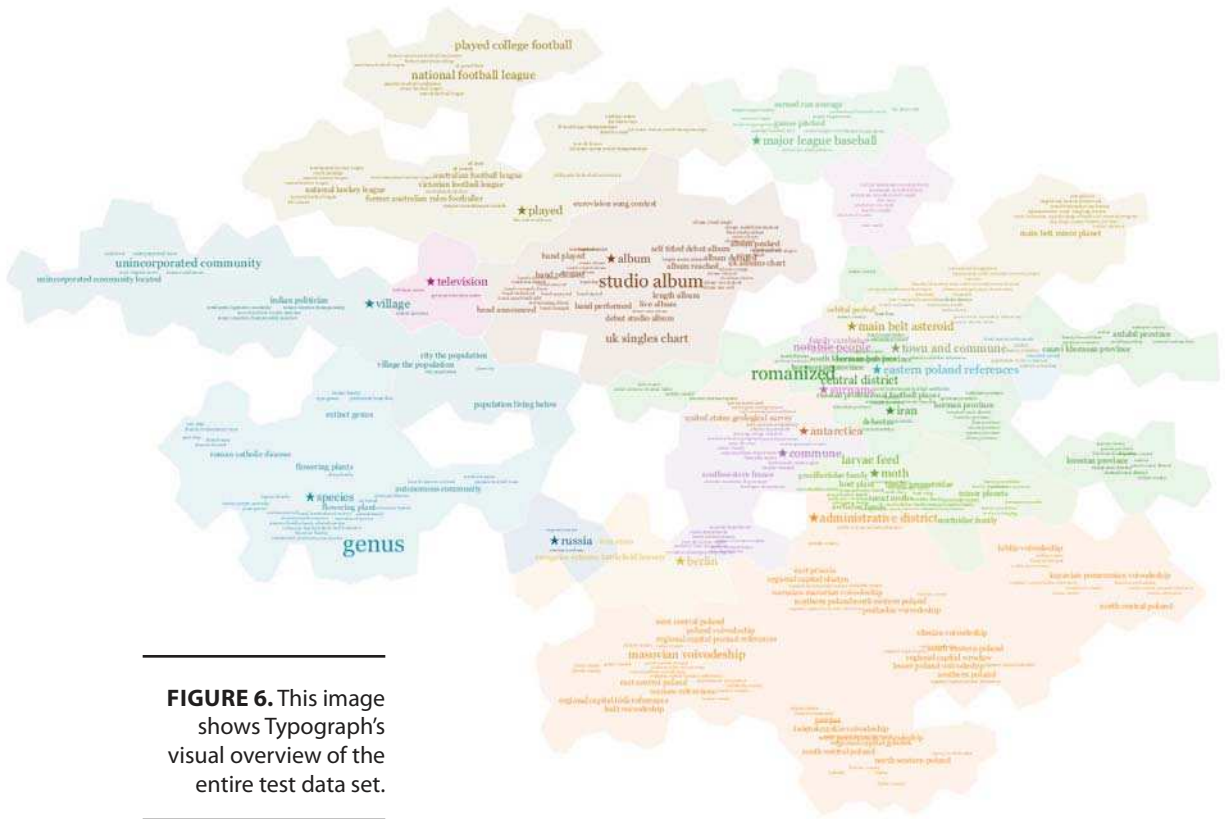
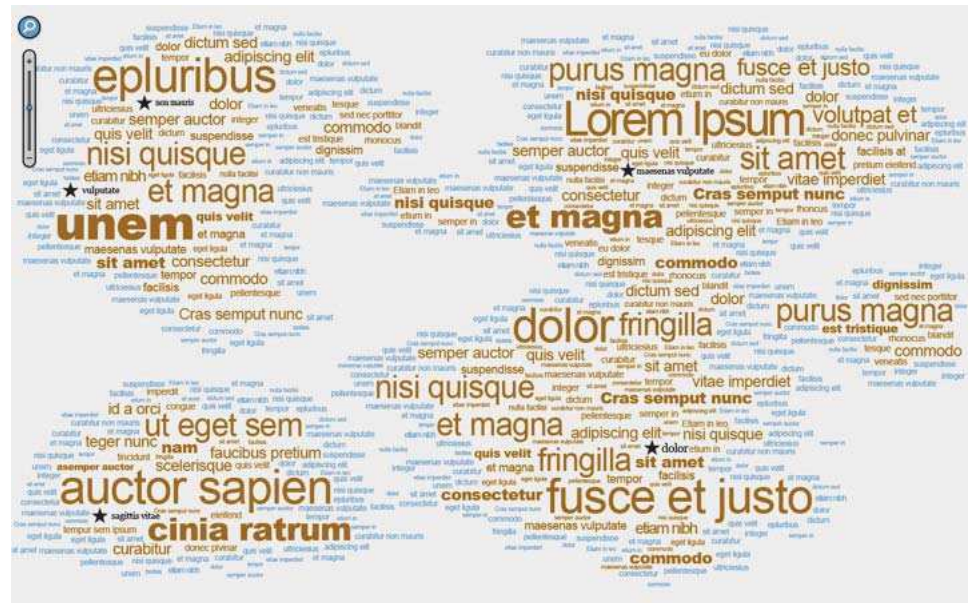


FIGURE 6. This image shows Typograph's visual overview of the entire test data set.

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.

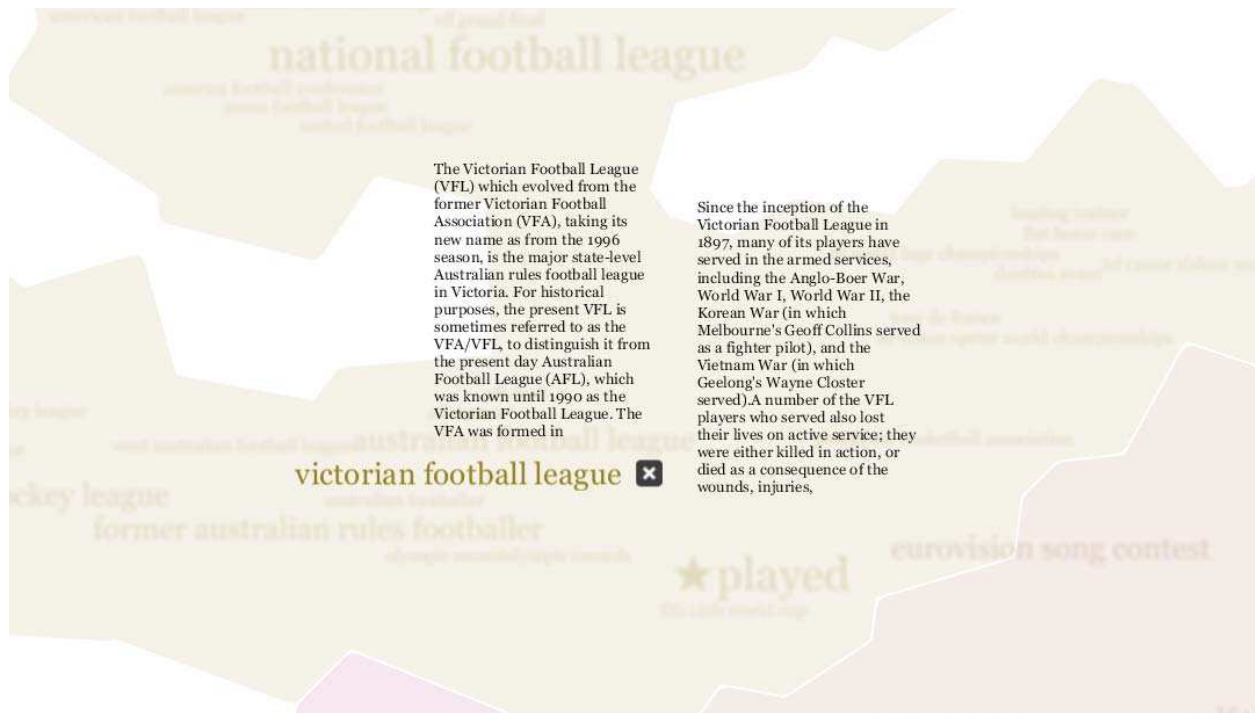


FIGURE 8. This image shows Typograph's view of the "Australian rules football" region with article details (i.e., text snippets).

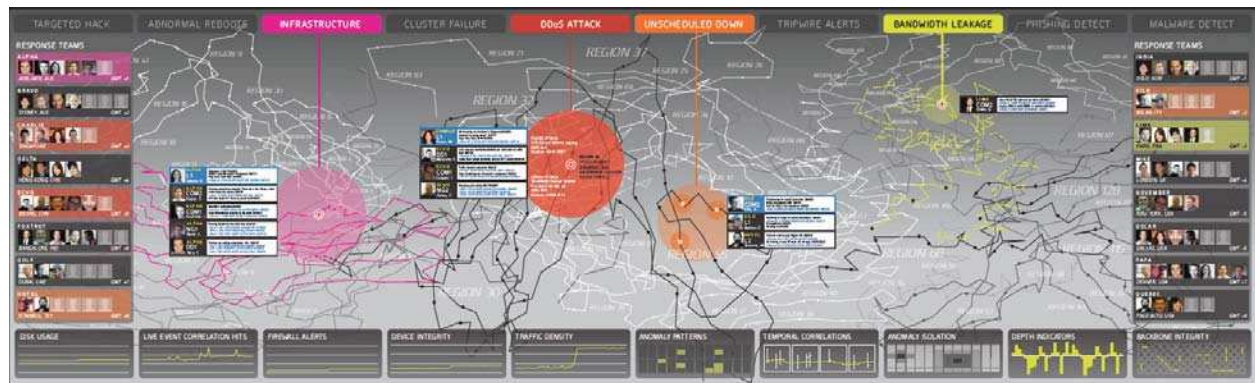


FIGURE 9. This visualization design for situation awareness of large computer networks was submitted by Gong Szeto Design Office.

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

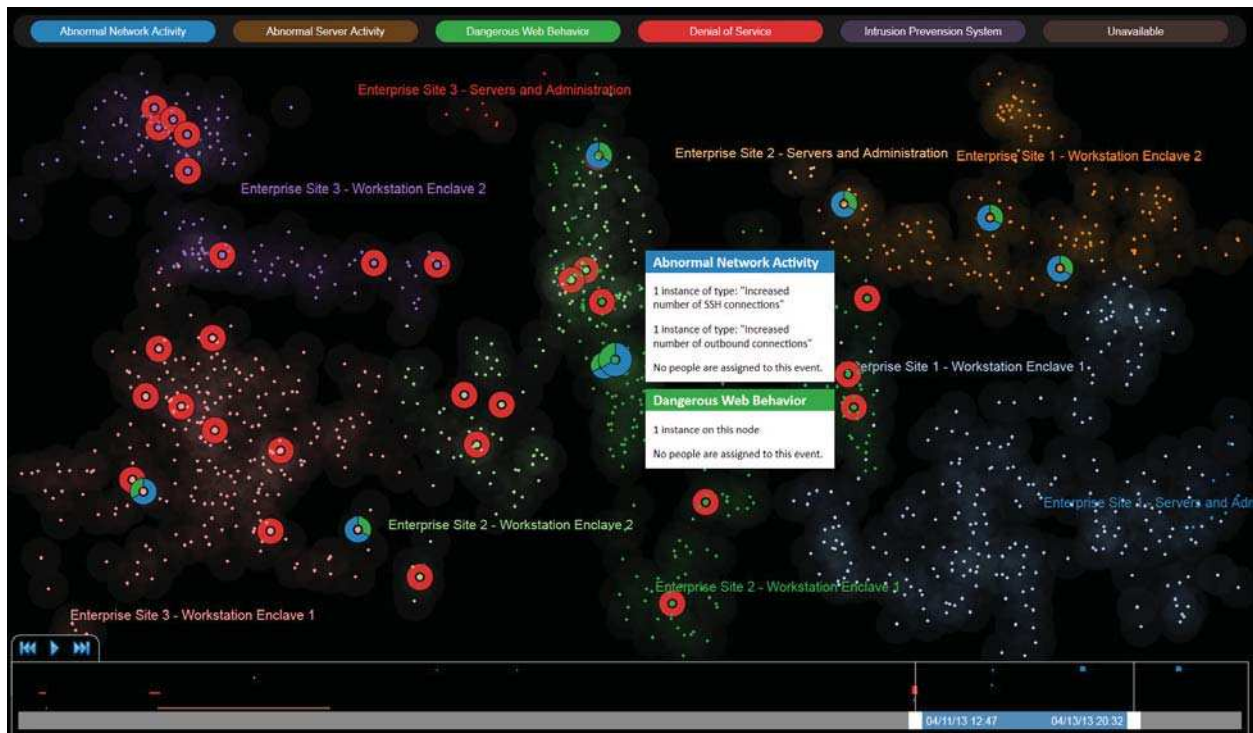


FIGURE 10. Using the visualization design from figure 9, we created this visualization prototype for cyber-situation awareness of large computer networks.

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://ieevis.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.

The Next Wave is published to disseminate technical advancements and research activities in telecommunications and information technologies. Mentions of company names or commercial products do not imply endorsement by the US Government.

This publication is available online at <http://www.nsa.gov/research/tnw/index.shtml>. For more information, please contact us at TNW@tycho.ncsc.mil.

