

Graph Analytics—Lessons Learned and Challenges Ahead

Pak Chung Wong ■ *Pacific Northwest National Laboratory*

Chaomei Chen ■ *Drexel University*

Carsten Görg ■ *University of Colorado Denver*

Ben Shneiderman ■ *University of Maryland*

John Stasko ■ *Georgia Institute of Technology*

Jim Thomas ■ *Pacific Northwest National Laboratory*

Graph analytics is among the most interesting and important topics in the visualization and analytics community. The mission of graph analytics research isn't merely about research per se; it has the essential and enduring purpose of producing pragmatic working solutions that meet real-life challenges.¹ (For more information,

Lessons learned from developing four graph analytics applications reveal good research practices and grand challenges for future research. The application domains include electric-power-grid analytics, social-network and citation analytics, text and document analytics, and knowledge domain analytics.

see the “Graph Drawing, Visualization, Mining, and Analytics” sidebar). After delivering applied graph analytics technologies in separate groups over the past few years, we feel the need to reflect collectively on the similarities and differences in our different research projects. We also wish to explore the similarities and differences between our research and other related areas such as graph drawing and graph visualization.

When discussing graph analytics technologies, we can't neglect the applications that leverage them. A successful application can bring out the best of the underlying technology. Visual analytics researchers often use the degree of success achieved by an application to measure the corresponding technology's quality, instead of using the algorithmic or

aesthetic criteria usually found in graph-drawing and graph visualization research.

Such technology-and-application pairs, which depend on and complete each other in much visual analytics research, form the basis of our discussion. We've organized the lessons we report here by individual applications that highlight the corresponding technologies' specific contributions. These applications fall into four domains: electric-power-grid analytics, social-network and citation analytics, text and document analytics, and knowledge domain analytics.

This discussion's primary focus isn't graph drawing or theoretical graph algorithms but the application of graph analytics in successfully deployed or applied systems. The four applications use graph analytics differently and use nonstandard layout paradigms or adjust existing layout approaches to a specific domain. The successes and mistakes we describe are real and in certain ways unique to the selected applications. However, many of the lessons learned and the ongoing challenges we describe are also applicable to the broader perspective of visual analytics.

Electric-Power-Grid Analytics

A power grid is an electrical network that transmits and distributes power from generators to consum-

Graph Drawing, Visualization, Mining, and Analytics

In the world of graph studies, we often see terms such as *graph drawing*, *graph visualization*, *graph mining*, and, more recently, *graph analytics*. (In our context, “graph” refers to a set of entities and their relationships. This differs from graphics such as bar graphs or histograms, which depict the relationships among variables as adopted by the engineering and scientific communities.) These terms represent the major R&D communities that approach graph problems from different disciplines and perspectives. Although these four areas have similarities, each has distinctive objectives and research methods.

Graph drawing is the oldest of the four areas. According to a Google Timeline search, the term first appeared in the public literature in the 1970s. Graph-drawing researchers develop computational techniques and algorithms to automatically lay out (draw) graphs. The annual International Symposium on Graph Drawing (www.win.tue.nl/GD2011) is the community's flagship conference. *Graph Drawing: Algorithms for the Visualization of Graphs*¹ and *Graph Drawing and Applications*² are two of the more popular textbooks. Many of the results generated by the graph-drawing community have become the R&D foundation of the other three areas.

Graph visualization is a subarea of information visualization. A Google Timeline search indicated that this subarea entered the public literature in the 1990s. Graph visualization differs from graph drawing in that it involves humans in visually and interactively exploring graphs. The annual IEEE Information Visualization Conference (www.visweek.org/visweek/2011/info/infovis-welcome/infovis-welcome) is the premier conference. Both Stuart Card and his colleagues³ and Chaomei Chen⁴ have covered graph visualization extensively from the information visualization perspective. Graph visualization almost always involves some sort of smart graph-drawing technique as part of the exploration process.

Although the knowledge discovery and data mining community has studied different aspects of graph mining since the mid-1990s, the term “graph mining” didn't appear in the public literature (determined again on the basis

of a Google Timeline search) until the early 2000s. Graph mining employs computation to identify graphs' structural features and their interrelationships. The proceedings of the annual ACM SIGKDD Conference on Knowledge Discovery and Data Mining (www.kdd.org/kdd2011) extensively cover this topic. Deepayan Chakrabarti and Christos Faloutsos have surveyed the cutting-edge areas of graph mining.⁵ Unlike graph visualization, graph mining usually doesn't include human visual intervention in the exploration process.

Graph analytics is a major topic of visual analytics studies. Visual analytics has been called “the science of analytical reasoning facilitated by interactive visual interfaces.”⁶ Graph analytics is a transdisciplinary R&D area, involving information retrieval, data management, human-computer interaction, computer graphics, and visualization. It aims to bridge the gaps in the other areas' graph studies. The annual IEEE Conference on Visual Analytics Science and Technology (www.visweek.org/visweek/2011/info/vast-welcome/vast-welcome) is the main academic conference. The applications we describe in the main article are typical examples of graph analytics.

References

1. G. Di Battista et al., *Graph Drawing: Algorithms for the Visualization of Graphs*, Prentice Hall, 1999.
2. K. Sugiyama, *Graph Drawing and Applications*, World Scientific, 2002.
3. S.K. Card, J.D. Mackinlay, and B. Shneiderman, *Readings in Information Visualization: Using Vision to Think*, Morgan Kaufmann, 1999.
4. C. Chen, *Information Visualization beyond the Horizon*, 2nd ed., Springer, 2004.
5. D. Chakrabarti and C. Faloutsos, “Graph Mining: Law, Generators, and Algorithms,” *ACM Computing Surveys*, vol. 38, no. 1, 2006, article 2.
6. J.J. Thomas and K.A. Cook, eds., *Illuminating the Path: The Research and Development Agenda for Visual Analytics*, IEEE CS Press, 2005.

ers. The North American power grids necessitate and facilitate the sharing of resources and infrastructures across companies and territories under federal regulation in the US and Canada. It has been suggested that electric power has the highest network reachability of today's energy infrastructure because all the other energy resources depend on electricity to operate. So, maintaining the power grids' stability and interoperability is vital and directly contributes to the security, economy, and social order of the nation and beyond.

To this end, visualization actively supports day-

to-day grid operations ranging from planning to maintenance to failure recovery. Among these operations is the proactive detection of grid vulnerabilities and signs of potential crises, on which we focus here.

The Domain and Task

In the electric-power industry, a power grid visualization typically refers to an overlay of a network of buses (nodes) on top of a geographic map, with different colors indicating the voltages of the transmission lines (links). In addition, dynamic

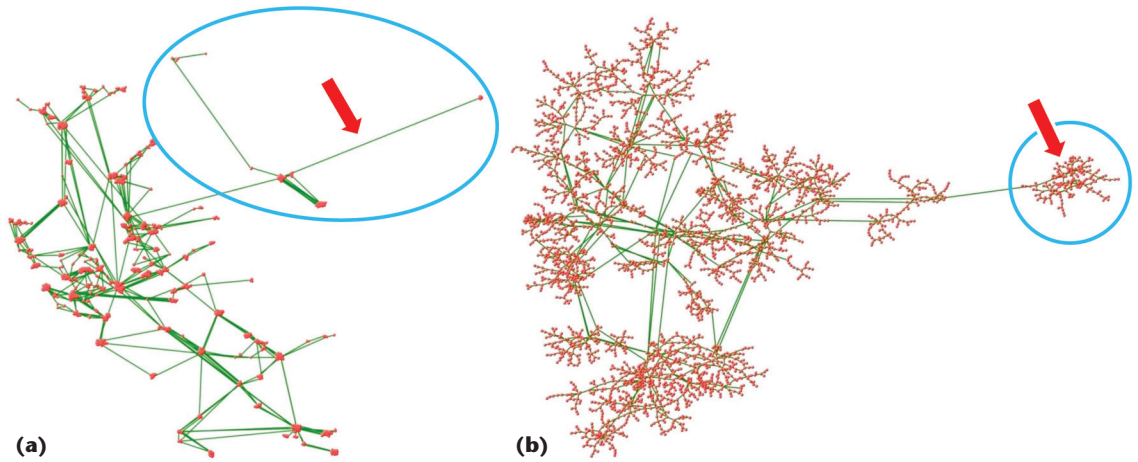


Figure 1. Visual analysis of the California power grid. (a) A traditional static geographic view. (b) A dynamic GreenGrid view, with additional information about the transmission links' inductive reactance. The highlighted area in the GreenGrid visualization provides more, and more accurate, information in terms of circuit size, link connectivity, and so on.

information can be animated on an electronic display showing the latest changes at different map locations. Icons display multivariate information.

For routine tasks such as assessing system vulnerabilities and threats (which usually involve a range of human, machine, and environmental factors), operators must use their expertise and experience to mentally decipher the integrated meaning of the colors and icons at different locations and decide when and how to respond. We call this traditional analytical practice “visualizing the geography” of power grids.

The Problem and Solution

Unfortunately, disaster can strike fast and often with little warning in electric transmission and distribution systems. For example, only six minutes elapsed between the first sign of system stress and the total breakup of the entire Western Electricity Coordinating Council (WECC; www.wecc.biz) grid during the last major US blackout on 10 August 1996. A 1.5-minute delay between an event in the grid and the communication of the relevant information to the operators worsened the situation. In other words, from the moment the problem started, the operators had less than five minutes to diagnose the situation and prevent the further deterioration that ultimately led to a total blackout.

Making a sound decision in moments calls for an intuitive visual analytics tool that goes beyond the geographical divide and focuses on the electronics of a geographically integrated power network. We call this analytical approach “visualizing the physics” of the power grid, which reflects what electrons look like in an electric circuit. In other words, we want to analyze the electron movement and the corresponding behavior in an electric cir-

cuit instead of the geographic locations of generators, transformers, and so on.

Working with researchers and engineers from the electric-power industry, we developed GreenGrid, a graph analytics tool that bridges the power grid's geography and the circuit board's electronics.² GreenGrid uses a weighted force-directed graph layout to model a power circuit's repulsive forces (such as voltage angle) and attractive forces (such as impedance). A GreenGrid graph layout is not just a geographic template that ties icons together; it's a visualization that physically reflects the grid's contents and structurally characterizes its behavior. For example, Figure 1a shows a static view of the California power grid based on its geographic information. Figure 1b presents a dynamic GreenGrid view of the same network, with additional information on the transmission lines' inductive reactance.

From an information visualization viewpoint, the GreenGrid visualization has many advantages over the traditional one. For example, although the cyan circles in Figures 1a and 1b highlight the same power grid area (the same set of nodes and links), the highlighted area in Figure 1b clearly provides more, and more accurate, information (in terms of circuit size, link connectivity, and so on).

Also, from a power-grid-visualization viewpoint, the visualization in Figure 1a can be misleading if you're analyzing an electric circuit. Because electrons move quickly (near the speed of light) through a circuit, the geographic distance between two network nodes, such as the one highlighted by the red arrow in Figure 1a, is mostly irrelevant. GreenGrid, on the other hand, presents the power circuit's underlying physics (inductive reactance in this case) in Figure 1b.

Perhaps the example in Figure 1 is too simplistic to claim any meaningful success. We put GreenGrid to the test using a simulation environment that models the 1996 WECC blackout. By combining node and link mappings, we used GreenGrid to detect the power grid's potential vulnerabilities.² Six of the vulnerabilities that GreenGrid identified were ones that eventually broke the WECC grid into four isolated islands, which eventually brought down the entire system.

Lessons Learned

We learned to not overemphasize the importance of spatial details and accuracy of a dataset involving geography. Including spatial information almost instantly eliminates one of the major data-mapping options, which include color, shape, icon, and texture, to visually analyze the underlying data.

In populous areas, the small-world³ characteristics of many geographic datasets also challenge human cognitive and sensory limits. For example, the physical size of Los Angeles on a map is usually very small, yet the information related to the city is almost always too rich to be properly visualized within the city boundaries on the map. Attempts to balance the two extremes usually create new artifacts and meaningless distortions. Although this lesson seems fairly obvious and is probably well-known to some, many visualization designers, including those who develop analytical tools for the power grid industry, have repeatedly forgotten it.

We also learned to not directly visualize the data but instead visualize the theory governing or underlying the data. There's a fine difference between

- visualizing the data and using the visualization to discover or corroborate the theory and
- visualizing the theory directly and eliminating the extra step to substantiate the theory.

The former approach requires little or no knowledge of the data. This fundamental visualization approach largely involves data exploration, which has been the hallmark of the modern data visualization era. The latter, analytical approach requires deeper understanding of the theory (either empirical or mathematical) behind the data and largely involves data exploitation. In our case, visualizing power grid geography clearly belongs to the former approach, whereas visualizing power grid physics belongs to the latter.

Our experience with power grid analytics also reaffirms the mission of visual analytics, which challenges the conventional data exploration wisdom of information visualization. Instead, visual analytics

aims to integrate visual and analytical methods for broader knowledge and deeper understanding.

Turning Lessons Learned into Lessons Applied

The Electricity Infrastructure Operations Center (<http://eioc.pnl.gov>), a fully functioning backup power grid control room for the Pacific Northwest, is testing and evaluating GreenGrid. We've also delivered copies of the system to our industrial partners for transitioning our technologies. Additionally, we're incorporating the lessons learned from developing GreenGrid into a new visual analytics tool for analyzing critical-infrastructure protection, which includes power grids as part of the energy infrastructure.

Social-Network and Citation Analytics

The growing production of Web-based documents, including email, scientific papers, legal briefs, and Wikipedia, offers opportunities to improve collaboration and resolve conflicts. In some cases, the challenge is due to the scale.

The Domain and Task

Social-network and bibliographic-citation analysts have been studying human relationships as manifested in documents for almost a century. The growing availability of online data has dramatically increased the opportunities, and the parallel development of advanced analytic tools is having a profound effect. Analysts can now trace patterns of influence by coauthorship and co-citation analysis, as well as more subtle use of key phrases, arguments, and examples.

The Problem and Solution

Once the linkages among thousands or millions of documents have been extracted, analysis can begin. Standard graph-theoretic algorithms, such as those involving prestige or betweenness centrality, have been extended with thousands of specialized metrics that might highlight key documents that influence or bridge disciplines. In addition, visualization tools are increasingly effective in presenting the relationships in ways that support exploration and discovery.

The most common approach for network visualization is the force-directed layout. Ideally, strongly related nodes are clustered together, revealing communities of shared interest. Isolated communities and nodes are placed far from the clusters.

Unfortunately, this approach draws richly connected clusters in a tightly packed area, often making it impossible for users to follow links from the source to the destination, count the degree of

a node, or even read node labels. Some strategies for improving legibility are emerging. However, other problems, such as layout instability as nodes or links are added or deleted, undermine force-directed layouts' utility.

Our research takes a completely different approach to node layout: *semantic substrates*.^{4,5} We've found that in many applications, nodes have rich attributes that are important to users, such as the publication year or venue. If you lay out nodes on a meaningful substrate, users can see patterns related to these important and familiar attributes.

A common attribute is time, with users typically expecting older documents to appear on the left and newer documents on the right (on the *x*-axis). Then, users can see the growth of document production over time as well as quieter or more intense production periods. If you use the *y*-axis for other meaningful attributes, such as the 1st through 11th District Courts in US legal documents or low- to high-ranked journals (for example, by impact factor), users can quickly see further patterns, clusters, trends, outliers, and gaps. A further benefit of meaningful substrates is that you can have several regions—for example, separating district court cases from circuit court cases or journals from conferences. Users can then quickly get further useful information.

Once users specify the node layout, they can review the links, revealing which sets of nodes are strongly or weakly connected. With more than a few dozen nodes and links, the display can become cluttered, so user control over link visibility becomes valuable. Selective display of links will reveal whether the connections are local or distant, scattered or concentrated. An often-cited key document stands out, owing to the large number of incoming links. Similarly, an unusual or rarely cited document also stands out because there are just a few incoming links in a generally empty area.

Lessons Learned

We found that for problems in which nodes have multiple attributes, semantic substrates can be very helpful. It's natural to group male and female nodes in separate regions so that relationships between men and women stand out, and relationships between men and men as well as women and women are easy to follow. Another lesson is that such groupings readily reveal the number of nodes in each category. Any interesting distributions, such as age differences between groups, will also stand out.

Another lesson is that choosing the right region

layouts is extremely important, thereby raising the importance of a flexible design tool that lets users specify and adjust regions. Two principles were

- appropriate alignment of regions—for example, with a common timeline—and
- placement of regions to minimize long edges that might cause clutter.

Certainly, other principles will emerge with further applications.

Turning Lessons Learned into Lessons Applied

Semantic substrates are most effective when the nodes have many attributes that are familiar to users. We've found good applications for other analytic domains such as predator-prey relationships and US Senate voting patterns. Our software tool, Network Visualization by Semantic Substrates (NVSS; www.cs.umd.edu/hcil/nvss), has a powerful specification component that lets users specify up to five regions, each of which can have a different layout. Users can tie node sizes to attribute values, and we've added novel link visibility controls.

To support scaling to thousands of nodes and links, we added a metanode feature that allows aggregation of nodes with common attribute values. Replacing many nodes with a single metanode dramatically reduces the number of links and makes some patterns more apparent.

We continue to refine our software as we expand the range of our case studies. Although semantic substrates don't solve all network and citation analysis problems, they're proving valuable when knowledgeable users work on applications with multiple attributes for nodes. Figure 2 shows an application of NVSS.

We've taken the lessons learned from NVSS and embedded some of them in the more ambitious open source software tool called NodeXL (www.codeplex.com/nodexl). It includes graph importing from multiple social-media software sources (email, Twitter, Facebook, YouTube, Flickr, and so on), a rich set of analysis tools, and elaborate network layout possibilities. Some of the layouts use traditional force-directed approaches, but users can also specify *x*, *y* locations for nodes based on attribute values. A major step forward is the group-in-a-box feature, which uses a treemap algorithm to create rectangular regions whose size is based on the number of nodes in each group. In each box node, any algorithm can determine the layouts. NodeXL, like NVSS, lets users turn off links that cross regions so as to highlight each region's links.

Text and Document Analytics

Investigative analysts such as news reporters and law enforcement and intelligence professionals routinely work with large collections of documents. They might have a particular incident or individual to investigate; the task here is to gather as much related knowledge about that incident or person as possible. This situation is a targeted knowledge-gathering or learning scenario. Alternatively, investigators simply might be exploring collections, browsing for some interesting story, narrative, or theme to explore further. In this case, they might be developing a news story or identifying a threat that should be communicated to other officials for further action.

The Domain and Task

In investigative analysis of document collections, the documents and their attributes are the primary objects of interest. Our research focuses on the entities in the documents' text, such as the named people, places, dates, and organizations that likely will be vital during investigations.

To identify the narratives embedded throughout document collections, investigators must understand the relationships between entities. For instance, a reporter researching a money-laundering story will need to identify the people involved, where the acts occurred, what banks were used, how much money was involved, and so on. The reporter might find such connections only through long chains of subtly connected documents. Each new relevant document will provide a new set of potentially related entities to explore.

The Problem and Solution

As long as the number of documents is small, investigations such as those we just described are likely manageable. As the number of documents and entities grows, reading all possible related documents becomes less realistic, and investigators will have more difficulty finding pertinent documents and developing a coherent schema for the information being uncovered. Consequently, mechanisms that help investigators focus on the most related and pertinent documents can reduce investigation time and lead them to the most interesting findings.

In the Jigsaw system, we model two graphs.⁶ In the *document-entity graph*, documents and their entities define the set of nodes, and their contained relationship defines the set of edges. That is, entities are linked to all documents in which they occur. In the *entity-entity graph*, the entities define the set of nodes; two entities are connected if they co-occur in at least one document.

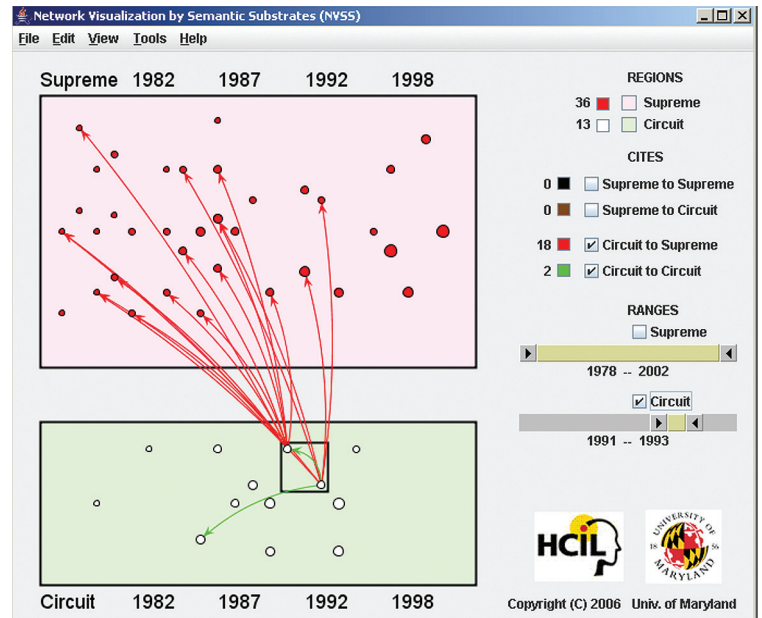


Figure 2. A Network Visualization by Semantic Substrates visualization that shows 18 red Supreme Court citations and two green circuit court citations in 1990 and 1991. The alignment by common timeline makes the links understandable, distinguishing the citations of recent cases from those of older cases.

Jigsaw provides multiple visualizations of these two types of relationships. For example, Figure 3a shows the *graph view* of a document-entity graph. Small white rectangles represent documents in the collection. Small circles represent entities, drawn in a color unique to the entity type (for example, person, place, and date). Thin white lines connect documents and their entities.

In the graph view, when the investigator expands documents or entities, Jigsaw draws new items in circular clusters around the expanded element. The view contains a special circular-layout operation to redraw the view. Jigsaw draws the documents on a circle and places highly connected entities inside it. The more connections entities have, the closer they are to the center.

Figure 3b shows the *list view* of an entity-entity graph. This visualization doesn't use a traditional graph layout. It shows the nodes (entities) in lists organized by their type. It displays connections to one or more selected entities in two ways: drawing lines to connected entities in neighboring lists and using color to show connections across all lists. Shades of orange indicate a connection's strength: entities co-occurring in multiple documents are more strongly connected and thus are darker.

In large document collections, many documents and entities are just noise and don't contribute to the sought-after thread. So, investigators care more about individual documents and entities and the set of connected items that provide context

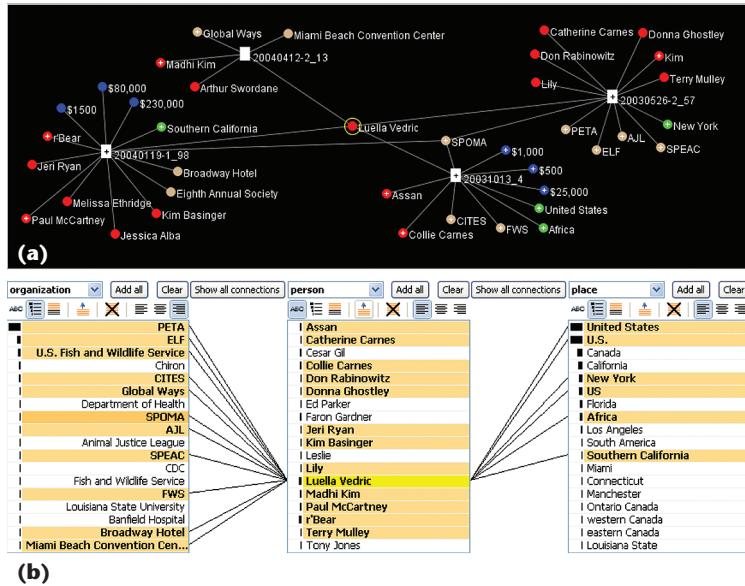


Figure 3. Views of graphs in the Jigsaw system. (a) A graph view of a document-entity graph. (b) A list view of an entity-entity graph. The views in Jigsaw are connected, and user interactions in one view are reflected in other views.

to them than they do about the document set’s global structure. Hence, Jigsaw uses an interactive, incremental-explorative approach. Nodes appear or disappear only through user directives, so exploring connections is much more dynamic. The investigator chooses which documents or entities to explore further.

The views in Jigsaw are connected, and user interactions in one view are reflected in other views. This lets investigators examine different aspects of entities and documents simultaneously under different perspectives.

Lessons Learned

Much graph visualization R&D focuses on drawing an entire (large) graph and uses sophisticated layout algorithms. Frequently, such research produces an initial complete static layout, and analysis proceeds from there. Jigsaw takes a very different approach. The views start initially empty; as we mentioned before, nodes are added only through user direction. Additionally, the graph view employs relatively simple layout heuristics.

Combining the list and graph views to visually explore document-entity and entity-entity connections is powerful. The list view provides a high-level overview of connections between entities, and its sorting capabilities help users quickly find the most frequent or highest-connected entities in a set. The graph view then reveals the connecting documents and provides more details of the connections.

The analysts we worked with liked Jigsaw’s user-

focused approach because Jigsaw is easy to use and puts them in control. Jigsaw shows items only when users want to see them. Each new node expansion is generally of the form, “Show me what’s related.” Investigators can manually change the layout according to their semantic model of the entities and additional information extracted from the documents.

The Jigsaw views also benefit from the straightforward user interface and interactions. Single and double mouse clicks map to natural operations—selecting and expanding items. A right click exposes a small but powerful set of auxiliary operations to assist browsing. Because the graph view doesn’t provide sophisticated automated layout, it’s crucial to have a simple but powerful user interface to manually adjust the layout. Our technique works well as long as the graph of interest is relatively small. Larger graphs require more sophisticated approaches, or manual node repositioning becomes too tedious.

Turning Lessons Learned into Lessons Applied

Using Jigsaw as an analytic aid, we won the university category of the 2007 IEEE Visual Analytics Science and Technology Symposium contest. This contest presented a collection of approximately 1,500 documents and asked participants to identify a synthesized threat embedded across the documents.⁷ The list and graph views were likely the most valuable system views we used. During a special conference session, a professional analyst working with us particularly liked the circular-layout function.

Since then, we’ve demonstrated Jigsaw to analysts across different domains (law, law enforcement, and biology) and started several collaborations. Law enforcement professionals at the Washington Joint Analytical Center in Seattle have used Jigsaw, providing positive feedback and many ideas for further enhancements.

Knowledge Domain Analytics

Integrating graph-theoretic and information-theoretic approaches is a challenge for graph analytics.⁸ On one hand, each approach has unique strengths, well-established metrics, and expectations of plausible patterns. On the other hand, each draws input from distinct data sources, uses different data transformations, and is optimized to identify different types of patterns of interest. The conceptual gap between the two will likely hinder an analytic pursuit that must draw on clues and follow leads across the boundaries of information patterns—for example, between structural proper-

ties identified by graph theory and uncertainties presented by information theory.

Juxtaposing results obtained from the two approaches on the same raw data might not be considered a genuine integration because analysts would still have to establish any possible connections between two sets of patterns. Graph-theoretic representations of an underlying system are an abstraction. Graph analysis focuses essentially on structural components and corresponding properties. Typical tasks include

- identifying structural components on the basis of connectivity,
- determining the shortest path between two nodes,
- estimating a graph's resilience to removal of a node, and
- identifying the generic type of a graph, as in the case of scale-free or small-world networks.

In contrast, information-theoretic approaches

- measure textual information's uncertainties or ambiguities,
- estimate how additional information reduces or increases uncertainty, and
- provide a generic class of metrics called *information metrics* to measure the overall distance between different information sources.

One active research area involves devising effective measures of interestingness such that human users can efficiently direct their attention to potentially significant targets.

The challenge for integrating the two approaches is due largely to the mismatch between the abstraction levels and the lack of theoretically or operationally defined connections between structural patterns and numerical metrics. To illustrate this challenge and such integration's potential benefits, we look at analysis of the domain of scientific discovery.

The Domain and Task

We selected scientific discovery for four reasons. First, it shares fundamental goals of visual analytics in terms of making unforeseen connections or linking previously unconnected threads. Second, the study of scientific discovery has a long history in several disciplines, including psychology, cognitive science, artificial intelligence, sociology, history, and philosophy. The related literature can provide valuable insights into discovery. Third, the scientific literature is a readily available source of

continuously evolving data with which we can explore and validate various theories, models, and metrics. Finally, graph analytics advances will make immediate contributions to emerging fields such as cyber-enabled discovery and e-science.

A central question in the study of scientific discovery is how to explain a discovery in terms of the relevant scientific knowledge before and after it. There's a wide variety of scientific discoveries; one type closely related to visual analytics involves revealing hidden links or creating new links. To some extent, this involves knowing where to look or what to ask. The challenge arises because identifying the most plausible places to look involves cognition at two levels:

- a macroscopic level in terms of structural patterns and
- a microscopic level in terms of information at the sentence or even the concept level.

Generating graphs is in the direction of aggregation, whereas breaking down documents into linguistic and statistic patterns is in the direction of decomposition.

The Problem and Solution

We illustrate the challenge with a typical study of the structure and evolution of a scientific domain—for example, terrorism research, string theory, or climate change. There are two motivating questions:

- What does the scientific community know collectively about the subject?
- What evolutionary paths have led to the most influential insights in the subject matter?

To address these questions, we tapped scientific literature as a publicly available source.^{9,10} In essence, we created a network of co-cited references.

The traditional way to make sense of the network is to focus on its structural groupings. A typical method is to apply a clustering algorithm, factor analysis, or principle-component analysis to the network's corresponding similarity matrix and decompose the network into clusters or components. The next task is to make sense of each cluster on the basis of analysts' knowledge or an additional search for further information about the cluster members. Knowing that a set of items belongs to a group often is insufficient in itself to determine the nature of the group. Automatic feature selection and text summarization can be useful.

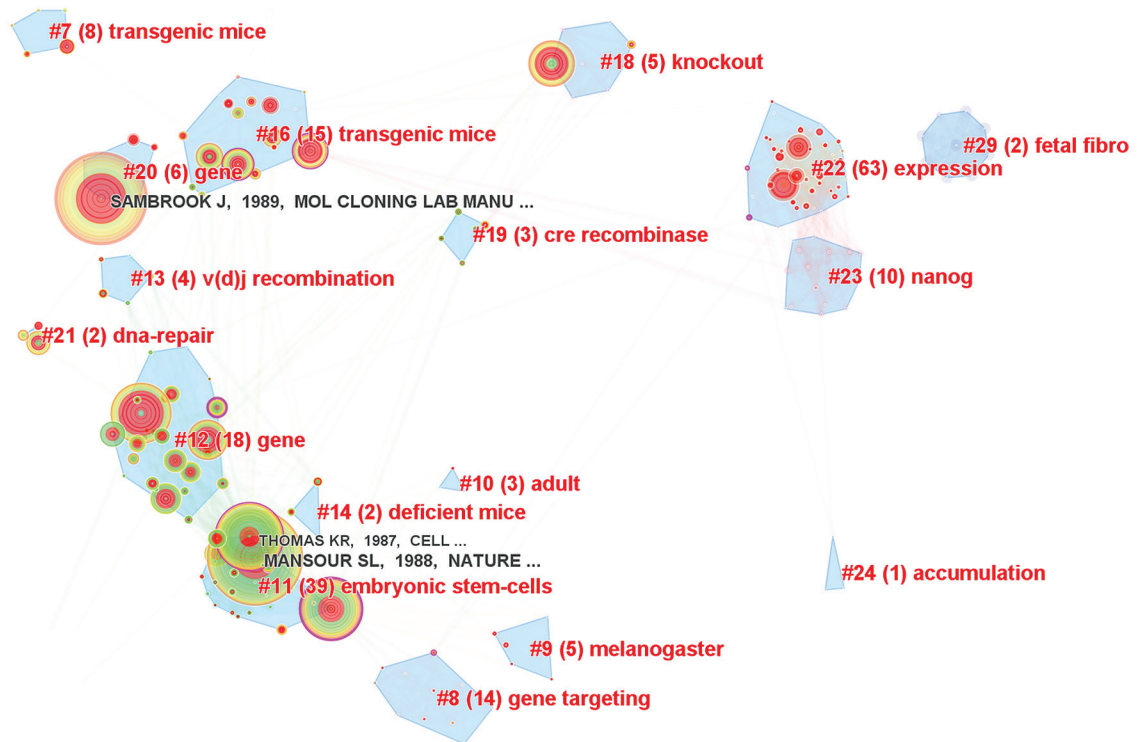


Figure 4. A multilayered network visualization of co-cited references on gene targeting. CiteSpace automatically identifies and labels each cluster with the most discriminant term found in text that contributes to the cluster.

We've found two interesting connections between graph-theoretic and information-theoretic approaches. First, we've proposed an explanatory and computational theory of scientific discovery based on integrating structural and temporal properties.¹⁰ Second, we've incorporated a clustering step that relies on structural information only such that the resultant graph decomposition clearly identifies the responsible text in the source data. Then, we can link structural patterns and linguistic and statistical patterns.

Figure 4 shows a network decomposed into clusters through the unsupervised algorithms in our CiteSpace application. CiteSpace automatically characterizes each cluster's meaning at multiple levels, from individual instances of terms in the corresponding latent semantic space to cluster labels summarizing the most salient and unique aspects. It encodes the temporal and historical properties as visual attributes.

Lessons Learned

From developing our theory of discovery, we learned that structural patterns alone often aren't sufficient to identify a potential path of discovery. For example, structural patterns carry no temporal information such as recency, growth, or decay. Providing temporal information at multiple levels of abstraction or aggregation can signifi-

cantly facilitate relevant analytic tasks. Semantic indicators, similarly, are absent from typical graph representations or are often simply superimposed on the graph, which makes sensemaking at various abstraction levels difficult.

We also learned that a tighter integration of graph- and information-theoretic approaches opens up new ways of interacting with graph visualization, with the benefits of text mining, summarization, and many other potentially useful capabilities. Understanding a graph visualization is no longer limited to structural patterns; analysts now can cross-validate clues conveyed by both structural components and multiple layers of textual information.

The abstract nature of the graph-theoretic approach lets us use many generic operations. On the other hand, structural properties might have a diverse range of associated semantics. This is a particularly promising area in which information-theoretic approaches and graphic models for reasoning can significantly enhance the graph-theoretic foundations of graph analytics.

Turning Lessons Learned into Lessons Applied

We're incorporating an integrated approach combining information theory, information-foraging theory, and Bayesian search theory with graph-theoretic approaches.⁸ Such an approach could guide the search and assimilation of macroscopic

knowledge patterns on the basis of information metrics that measure information associated with particular topological and temporal properties.

Discussion

We began writing this article expecting to find considerable evidence of common lessons learned among the four graph analytics applications. After all, we all deal with information governed by the same data type and structure definitions. If one person can find a working solution, the others can go along with it. But when we actually wrote down the problems, solutions, and lessons learned, we saw instead interesting differences regarding fundamental approaches to graph analytics problems.

For example, the power-grid-analytics application applies the top-down approach extensively to proactively detect the vulnerabilities of the entire grid. The text-and-document-analytics application, on the other hand, champions the bottom-up approach to isolate interesting actors in the network and build associations among them. It also contradicts the knowledge-domain-analytics application's general principle that you can't fully understand emergent properties by examining individual components because those components might conflict with one another at the component level.

Unlike the other three graph analytics applications, the social-network-and-citation-analytics application defies the tradition of using a standard node-link layout and favors node-based visualization. The polarization of these analytical approaches suggests not just fundamental ideological differences but also practical reasoning strategies.

We now look at the graph analytics challenges we identified when developing our applications. These challenges are by no means unique to these applications. Indeed, many aspects of these difficulties are common to other application areas.

Dynamic Graph Stream Analytics

Regarding the power-grid-analytics application, we must address two major challenges before we can integrate the technology into a working environment.

First, at the technical level, we're trying to identify the most effective way to apply the visualization technology in real time. In our research, we've found that a nonstop animation of GreenGrid can cause motion sickness. Also, the screen space in a grid room is too limited to support a time series or matrix of graph visualization. The challenge is to properly present the GreenGrid visualization in real time and yet bring out the time-varying nature of the underlying data without causing operator sickness or fatigue.

The second challenge is at the operational level. Power grid companies have invested significantly in their infrastructure management systems. Introducing new visualization technology without rigorous testing and evaluation is expensive and risky. We've conducted a usability study on daily operation of GreenGrid.² A bigger challenge is to find out how GreenGrid performs during emergencies, when operators have only minutes or even seconds to respond. With the US Department of Energy's support, we've invited consultants from the power grid industry to conduct situation awareness exercises using GreenGrid, in hopes of discovering its hidden strengths and weaknesses.

Node and Link Types Plus Multiple Layout Strategies

The rich variety of networks and tasks provides many challenges. For example, existing graph analytics systems don't deal well with networks having many node or link types with different attributes. Coloring the nodes or links differently is a start, but if filtering or computed attributes could be different for each node or link type, analysts could be much more effective. A second helpful design addition to support discovery would be to allow multiple layout strategies (force-directed, semantic substrates, geographic maps, temporal histories, and so on) in coordinated windows.¹¹

Combining Automated and Interactive Layout Techniques

No one has yet successfully combined traditional automated graph layout techniques and user-centered interactive-explorative techniques. With automated techniques, users can't easily influence or constrain the layouts or adjust them to specific tasks' requirements. Users can adjust the parameters used to compute the layout, but usually these parameters only allow adjustment within the applied layout paradigm (for example, a circular or hierarchical layout). Also, these techniques are neither intuitive nor easy to understand for someone unfamiliar with the layout algorithm.

When users control layout creation, they often must perform time-consuming and tedious work to make the layout readable by repositioning nodes to reduce edge crossings and label overlapping. These techniques' difficulty only increases as the graph grows.

To effectively explore parts of large graphs, users need task-specific automated layout techniques that let them adjust the layout with simple interactions.

Integrating Information-Theoretic Approaches

A significant direction for the next generation of

graph analytics is to combine the graph-theoretic approaches' strengths with those of information-theoretic approaches. This will let users clearly identify the quality of information with reference to structural patterns and other temporal and semantic patterns. Incorporating information metrics in addition to the abstract topological properties will be essential for enhancing the ability to identify information concerning uncertainty.

Integrating information-theoretic approaches will also help address the issue concerning strong- and weak-profile patterns because information metrics measure the underlying system's macroscopic properties.

Visual analytics is a young, rapidly evolving field that demands immediate response and quick solutions to solve pressing real-world problems. The visual analytics community's educational and public-outreach activities often spark imagination and inspire proactive thought even beyond what was intended. Lessons learned such as the ones in this article could encourage knowledge sharing, promote replication of successful practices, and accelerate technology transfer within the community and across disciplinary boundaries.

We hope this discussion crystallizes the issues and helps generate new and better ideas. We've focused on our own experience and application domains in which we have extensive expertise. However, many important issues concerning graph analytics remain beyond this article's scope. One such issue is the lack of publicly available resources for formal validation and evaluation of graph-analytic features and systems with real users in realistic settings.

We envision our research as a pioneering effort to systematically organize and report lessons learned from different visual analytics areas. Eventually, we want to establish and maintain an online presence similar to Visual Analytics Digital Library (VADL; <http://vadl.cc.gatech.edu>) for a lessons-learned database, as part of our community outreach to enable and promote the reuse and spread of the knowledge. ■■

Acknowledgments

GreenGrid's development has been supported partly by the US Department of Energy (DOE) Office of Electricity Delivery and Energy Reliability and the National Visualization and Analytics Center (NVAC), a US Department of Homeland Security (DHS) program at the Pacific Northwest National Laboratory (PNNL). The

Battelle Memorial Institute manages PNNL for the DOE under contract DE-AC05-76RL01830. Jigsaw's development has been supported partly by the US National Science Foundation (NSF) via awards IIS-0414667, CCF-0808863, and IIS-0915788; by NVAC, under the auspices of the Southeast Regional Visualization and Analytics Center; and by Vaccine (Visual Analytics for Command, Control, and Interoperability Environments), a DHS Center of Excellence in Command, Control and Interoperability. CiteSpace's development has been supported partly by the NSF under grant IIS-0612129 and by DHS through NVAC. The Network Visualization by Semantic Substrates research has been supported partly by the NSF grant "Inter-court Relations in the American Legal System: Using New Technologies to Examine Communication of Precedent II."

References

1. J.J. Thomas and K.A. Cook, eds., *Illuminating the Path: The Research and Development Agenda for Visual Analytics*, IEEE CS Press, 2005.
2. P.C. Wong et al., "A Novel Visualization Technique for Electric Power Grid Analytics," *IEEE Trans. Visualization and Computer Graphics*, vol. 15, no. 3, 2009, pp. 410–423.
3. S. Milgram, "The Small-World Problem," *Psychology Today*, vol. 2, 1967, pp. 60–67.
4. A. Aris and B. Shneiderman, "Designing Semantic Substrates for Visual Network Exploration," *Information Visualization J.*, vol. 6, no. 4, 2007, pp. 1–20.
5. B. Shneiderman and A. Aris, "Network Visualization by Semantic Substrates," *IEEE Trans. Visualization and Computer Graphics*, vol. 12, no. 5, 2006, pp. 733–740.
6. J. Stasko, C. Görg, and Z. Liu, "Jigsaw: Supporting Investigative Analysis through Interactive Visualization," *Information Visualization*, vol. 7, no. 2, 2008, pp. 118–132.
7. C. Plaisant et al., "Evaluating Visual Analytics at the 2007 VAST Symposium Contest," *IEEE Computer Graphics and Applications*, vol. 28, no. 2, 2008, pp. 12–21.
8. C. Chen, "An Information-Theoretic View of Visual Analytics," *IEEE Computer Graphics and Applications*, vol. 28, no. 1, 2008, pp. 18–23.
9. C. Chen, "CiteSpace II: Detecting and Visualizing Emerging Trends and Transient Patterns in Scientific Literature," *J. Am. Soc. Information Science and Technology*, vol. 57, no. 3, 2006, pp. 359–377.
10. C. Chen et al., "Towards an Explanatory and Computational Theory of Scientific Discovery," *J. Informetrics*, vol. 3, no. 3, 2009, pp. 191–209.
11. G.M. Namata et al., "A Dual-View Approach to Interactive Network Visualization," *Proc. ACM Conf.*

Information and Knowledge Management, ACM Press, 2007, pp. 939–942.

Pak Chung Wong is a chief scientist and project manager at the Pacific Northwest National Laboratory. His research interests include visual analytics, predictive analytics, visualization, privacy and security, and social computing. Wong has a PhD in computer science from the University of New Hampshire. He's on the IEEE Computer Graphics and Applications editorial board. Contact him at pak.wong@pnl.gov.


Chaomei Chen is an associate professor at Drexel University's College of Information Science and Technology. His research interests include information visualization, visual analytics, knowledge domain visualization, mapping scientific frontiers, and theories of scientific discoveries. Chen has a PhD in computer science from the University of Liverpool. Contact him at chaomei.chen@cis.drexel.edu.

Carsten Görg is an instructor in the University of Colorado School of Medicine's Computational Bioscience Program. His research interests include the design, development, and evaluation of visual analytics tools to support biomedical discovery. Görg has a PhD in computer science from Saarland University. Contact him at carsten.goerg@ucdenver.edu.

Ben Shneiderman is a professor in the Department of Computer Science and a member of the Institute for Advanced Computer Studies at the University of Maryland, College Park. His research interests are human-computer interaction, user interface design, and information visualization. Shneiderman has a PhD in computer science from the State University of New York, Stony Brook. Contact him at ben@cs.umd.edu.

John Stasko is the associate chair of the School of Interactive Computing at Georgia Tech's College of Computing. He's also the director of Georgia Tech's Information Interfaces Research Group. His primary research area is human-computer interaction, focusing on information visualization and visual analytics. Stasko has a PhD in computer science from Brown University. Contact him at stasko@cc.gatech.edu.

Jim Thomas was a Laboratory Fellow at the Pacific Northwest National Laboratory and the founding director of the US National Visualization and Analytics Center. Thomas had an MS in computer science from Washington State University. He died in 2010.

 Selected CS articles and columns are also available for free at <http://ComputingNow.computer.org>.

Seasonal Specials Are Back!

Act now and get a free membership

CSDP Bundle: Normally \$595, now \$495

www.computer.org/certification



Distinguish Yourself From the Crowd Earn Your CSDP

Earning the Certified Software Development Professional (CSDP) credential is the best way to prove your abilities, skills, and knowledge.



www.computer.org/getcertified