Science Communication
1–25

DOI: 10.1177/1075547014556540
scx.sagepub.com

Visualizing Computational Social Science: The Multiple Lives of a Complex Image

Brooke Foucault Welles[1] and Isabel Meirelles[1]

Abstract

Parallel advances in communication and visualization technologies have enabled the study and visualization of human behavior at a scale and level of detail never before possible. Nowhere are these advances more evident than within the emerging field of computational social science. Using Adamic and Glance's image of the political blogosphere as an example and social representations theory as a guiding framework, we explore how computational social science visualizations may aid and complicate public understanding of this new science. We conclude with a discussion of best practices for the production and reuse of computational social science images for public consumption.

Keywords

computational social science, scientific visualizations, big data, network science, network graphs, complexity

Thanks to the increasing prevalence of digitally mediated communication along with the trace records it generates, social scientists from a number of

[1]Northeastern University, Boston, MA, USA

Corresponding Author:
Brooke Foucault Welles, Department of Communication Studies, Northeastern University, 204 Lake Hall, 360 Huntington Avenue, Boston, MA 02115, USA.
Email: b.welles@neu.edu

fields are able to observe and analyze human behavior at an unprecedented scale and level of detail. Lazer et al. (2009) urged researchers to embrace big data and computational analytic techniques to usher in a new era of social science dubbed *computational social science*. Broadly, computational social science is an interdisciplinary approach to social science that applies advanced computing techniques to study human social systems. It includes five core methodological techniques: automated information extraction, geospatial analysis (geographical information systems or GIS), complexity modeling, social simulation models, and social network analysis; and it frequently leverages electronic data to draw conclusions about human behavior (Cioffi-Revilla, 2010). In just a few short years since Lazer and colleagues' call, researchers have used computational social science to advance theory and knowledge across the social sciences, unlocking long-standing puzzles about topics such as the mechanics of political mobilization (Bond et al., 2012), the motivations for assembling into teams (Zhu, Huang, & Contractor, 2013), and the processes involved in the diffusion and adoption of new technology (Aral, Muchnik, & Sundararajan, 2009), to name just a few.

The outcomes of computational social science research are compelling; however, the specific methods and analytic techniques remain shrouded in mystery for many. Referring to Nate Silver's application of computational social science techniques to accurately predict the outcome of the 2012 U.S. presidential election (Silver, 2012), the satirical website *Is Nate Silver A Witch?*[1] declared, "Probably. While we on the Is Nate Silver a Witch editorial board are strict rationalists, Mr. Silver's performance has been uncanny enough to raise small but significant doubts as to whether his methodology is entirely of this world."

Of course, neither is Nate Silver a witch nor are his methods other-worldly; they are simply unfamiliar. Traditional social science research, such as survey or experimental research, has had the benefit of decades of public exposure to both the methods and results of those techniques. This is not yet the case for computational social science research. Moving forward, it will be important for computational social scientists to develop strategies for effectively communicating the complexities of the process and output of their work in order to affect social science understandings within the scientific and lay communities.

We believe that this challenge can be partially addressed through the use of effective scientific visualizations. In this article, we discuss the strengths and weaknesses of computational social scientific visualizations in general, and social network visualizations in particular, using a visual representation of the political blogosphere by Adamic and Glance (2005) as an example. Our goal is to examine the communicative power of that particular

visualization, including the reasons behind its widespread dissemination within the scientific and among the broader nonscientific community. Using social representations theory as a theoretical lens, we argue that complex computational social science images that visually confirm public expectations about human behavior may spread more readily, advancing general acceptance of computational social science approaches, but potentially obscuring its full value.

The Purposes of Data Representations

In the past 20 years, we have experienced a boom in data visualization practices fostered by the availability of large volumes of data. In 1985, a National Science Foundation initiative on scientific visualization helped catalyze the use of computer-supported visual representations by diverse communities, from earth scientists analyzing satellite data to computer graphics and artificial intelligence communities researching automatic visual presentations of data. Although visualizations can serve a number of purposes, in this article, we focus on those visualizations designed to enhance human understanding of data, and we adopt the definition by Card Mackinlay, & and Shneiderman (1999) that *information visualization* is "the use of computer-supported, interactive, visual representations of abstract data to amplify cognition" (p. 7).

Independent of the data type, information visualizations serve two distinct, although sometimes related, purposes. First, information visualizations are frequently used by scientists for exploration and analysis of data, such as helping reveal underlying structures and hidden patterns and helping with hypothesis generation. In this case, the visualizations are typically part of the scientific process, and they are neither shared with the general public nor designed with public consumption in mind.

Perhaps more often, information visualizations are used for the communication of information (e.g., scientific discoveries) to specific audiences, whether informing experts (e.g., in journals or proceedings) or communicating findings to general audiences (e.g., in news media). Some evidence suggests that the general public tends to perceive visual representations of science as "the truth," regardless of whether or not the visualizations were intended to convey one single truth, or indeed, any truth at all (Pauwels, 2006).

This issue may be further complicated in the case of computational social science visualizations, when a single visualization cannot represent the whole "truth" because of the size and complexity of the data. As information complexity increases, it becomes challenging to effectively represent data in a single image. All visualizations entail compromises, and in many fields, it is common practice to use multiple visualizations to accurately convey information. Take

for example, architectural representations, which are usually a compound of plans, sections, facades, electrical diagrams, and so on. A single architectural drawing cannot communicate the whole complexity of an architectural structure. Instead, architects commonly use a series of drawings to understand a building; they would never select just one drawing to stand for an architectural structure, as none could be considered the "most representative."

As computational social science gains traction within the scientific and lay communities, we argue for a similar approach to computational social science visualizations. Rather than selecting a single image to represent a complex system, we argue in favor of using multiple visualizations to represent various dimensions of a single system. We believe that such an approach could preserve the complexity that is a central asset of computational social science research and also aid in the general public understanding of computational social science visualizations as multifaceted and "multitruthed." In what follows, we describe one computational social science image in detail, using social representations theory as a guiding framework to suggest why this one image was selected to represent a complex idea that might be better illustrated with multiple images. We then offer suggestions for best practices to aid in public understanding of complex images and computational social science more generally.

Fostering Public Understanding of a New Science

Computational social science is a new scientific approach, enabled by the availability of detailed online data about human social behavior (Lazer et al., 2009). It has gained significant momentum within the scientific, professional, and lay communities in the past several years, spawning academic conferences, special issues of journals, popular press books, and the emerging professional field of "data science." However, relative to other modes of social scientific inquiry (e.g., surveys, interviews), the precise methods and analytic techniques of computational social science remain obscure for many, even many trained social scientists. In order for computational social science to gain widespread understanding (and, some would argue, acceptance), it may be important for computational social scientists to frame their work in terms of patterns and scientific methods that are more familiar.

Social representations theory offers a useful theoretical lens for explaining how this might happen. Originally described by Moscovici (1961) and subsequently described in English elsewhere (Moscovici, 1988), social representations theory explains how objects become collectively recognized, such that they can be understood, communicated, and reasoned about by a social group. The process involves two central steps: objectification, where a relatively

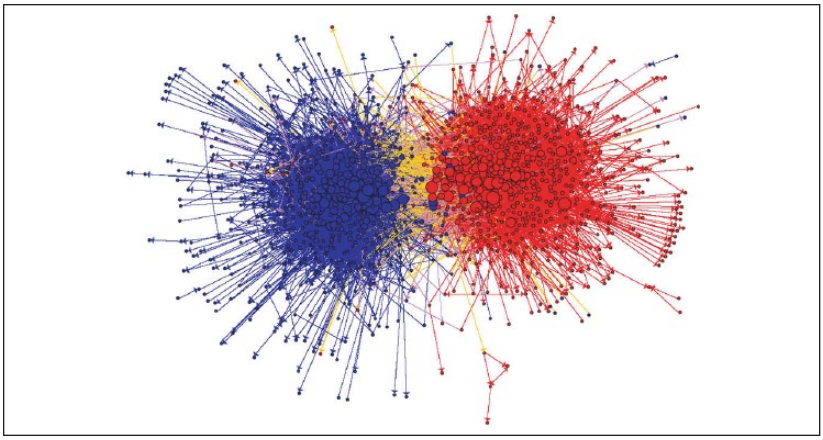

Figure 1. Original blogosphere visualization from Adamic and Glance (2005).
Note. The original caption describes it as follows: "Community structure of political blogs (expanded set), shown using the GUESS visualization and analysis tool. The colors reflect political orientation, red for conservative, and blue for liberal. Orange links go from liberal to conservative, and purple ones from conservative to liberal. The size of each blog reflects the number of other blogs that link to it," (Adamic & Glance, 2005, p. 37).

complex concept becomes simplified to the point where common understanding is possible, and anchoring, where the simplified object is interpreted through a lens of preexisting understandings about related objects (Moscovici, 1988; Rateau, Moliner, Guimelli, & Abric, 2012).

Following the logic of social representations theory, as computational social science gains momentum within the scientific and lay communities, we might expect that as the output (in this case, visualizations) spreads, it will have a tendency to become simplified and interpreted in terms of existing notions about the social world. Helpful in terms of cultivating widespread acceptance, this tendency risks obscuring one of the central assets of computational social science—the ability to study human behavioral complexity at a scale and level of detail never before possible. In what follows, we examine one notable example of a computational social science visualization and identify how its widespread dissemination both helps and hinders public understanding of this new scientific approach.

"Divided They Blog"

The visualization under scrutiny (Figure 1), which we will refer to as the "blogosphere visualization," first appeared in a paper coauthored by Adamic

and Glance (2005) and presented at LinkKDD, a satellite workshop at the Conference on Knowledge Discovery and Data Mining. The blogosphere visualization was arguably among the first to represent computational social science findings, combining a large data set with computational analytics to draw conclusions about the patterns of interaction among political blogs (and, by extension, the human bloggers). For that reason, it serves as an excellent case study for examining how it was used and reused over a 9-year period of circulation and reproduction by multiple communities.

The blogosphere visualization is a network image, representing the hyperlink connections (the links) between a set of political blogs (the nodes). Broadly, the image served the purpose of communicating scientific findings to an audience of experts, or more specifically, members of a subfield of computer science that focuses on automatically extracting useful information from large data sets. The stated scientific goal of the paper was to describe the hyperlinking patterns among a sample of political blogs analyzed in the lead-up to the 2004 U.S. presidential election between incumbent president George W. Bush and Democratic challenger John Kerry. Adamic and Glance chiefly focused on identifying the discussion patterns among and between 40 "A-list" blogs, defined by the proportion of total web traffic they received, to determine whether or not there was a single "noise machine" influencing online conversations about political candidates. In order to accurately identify those "A-list" blogs, they first gathered a list of all active political blogs at the time. The authors included the now-iconic visualization of the hyperlinks between all of the active political blogs they identified, as a way of documenting their process.

Interestingly, given its subsequent widespread reuse, for Adamic and Glance, the blogosphere visualization was somewhat orthogonal to the central analyses and conclusions of their paper, which featured six figures total and focused primarily on a small subset of 40 "A-list" blogs (20 conservative, 20 liberal) that drew the most web traffic at the time. The main findings of the paper involved analyses of the content of the posts on the A-list blogs, including the tendency for certain words and phrases to be reused within partisan communities, and for partisan blogs to link to partisan mainstream media sources. Although the authors found evidence of partisan linking behavior (both to other blogs and to mainstream media sources), they found no evidence of language reuse within political parties, discounting the idea of a "noise machine" setting the talking points for political blogs (Adamic & Glance, 2005).

Notably, the blogosphere visualization, although consistent with the findings regarding linking behavior, was used to illustrate the process of data exploration and justify the selection of the 40 A-list blogs that were examined

in more detail. The blogosphere visualization could arguably have been omitted from the paper entirely with no great loss to the central claims. Yet, for reasons we will discuss below, it went on to be reproduced much more frequently than other images in the same paper. The blogosphere visualization was subsequently reproduced hundreds of times in blog posts and scientific articles about computational social science and political communication, including in the widely read journal *Science* (Lazer et al., 2009, cited 849 times[2]), and in the popular book *Connected* (Christakis & Fowler, 2009, cited 584 times).

Organizing Visual Principles

Information Representation

The blogosphere visualization represents abstract information. Hyperlinks between political blogs have no literal or physical form; one could not go online and observe the structure being represented (Gershon, Eick, & Card, 1998). Generally, visualizations of abstract data help viewers—experts and nonexperts—understand concepts by providing spatial models and concrete ways of grasping knowledge (Larkin & Simon, 1987; Pinker, 1990). The blogosphere visualization is a node-link graph, a particular type of social network visualization that uses physical space to represent the interconnections between a set of related objects. Like any projection, a node-link graph entails both distortion and simplification. Think, for example, about geographical maps: We often take for granted that geographical map projections result in distortions of one or more of the geometric properties of the angles, areas, shapes, distances, and directions of physical space. Similarly, node-link graphs involve distortions and simplifications of relational information. As Newman (2010) explains,

> A network is a simplified representation that reduces a system to an abstract structure capturing only the basics of connection patterns and little else. Vertices and edges in a network can be labeled with additional information, such as names or strengths, to capture more details of the system, but even a lot of information is usually lost in the process of reducing a full system to a network representation (p. 141).

In the blogosphere visualization, 1,494 individual U.S. political blogs (the nodes) are represented by circles of various sizes scaled by popularity and colored by partisanship[3]—blue for liberal blogs and red for conservative blogs. The circles are connected in a network by lines (the links) representing over 20,000 hyperlinks between the blogs. Like the circles, the lines are also

colored by partisanship—blue lines for links between liberal blogs, red lines for links between conservative blogs, and orange and purple lines for liberal-to-conservative and conservative-to-liberal links, respectively. The spatial arrangement of circles and lines was defined computationally by an algorithm that positioned blogs that are linked close to one another, and blogs that are not linked far from one another. The result is that graphical proximity in the display encodes conceptual ideological distance in the source domain.

To be clear, representing the hyperlink relationships between political blogs in this way was an active choice made by the researchers; there is no single "best" way to represent relational data, and researchers often switch between different spatial layouts while searching for the one that depicts what they perceive to be the most meaningful properties of the network (Bender-deMoll & McFarland, 2006). Adamic and Glance did not specify which layout algorithm they used in this particular image, just that they rendered it using the visualization and analysis tool GUESS. It is likely that they chose one among the number of layout algorithms provided in GUESS, which includes a number of standard network visualization algorithms, such as Fruchterman-Reingold, Kamada-Kawai, Sugiyama, GEM, ISOM, and radial layouts (Adar, 2006), although it is impossible to tell through visual inspection alone how the graph was generated.

Partisan Divide

Before we examine how and why Adamic and Glance's blogosphere was reused so often, let us first step back and look at how vision and cognition allow us to derive meaning from the blogosphere visualization. If we were to describe the image from the point of view of its basic visual appearance, we would say that it depicts two sets of connected nodes (the blogs), one mostly composed of blue circles (to the left of the display) and the other mostly of red circles (to the right). Lines, mostly orange in color, connect blue and red sets horizontally. Although it obscures some finer detail, a blurred version of the image (Figure 2) well represents the first impression that we just described, that is of two clusters—blue and red—separated horizontally by an orange zone, which is as distinct as the other two colors.

Two Gestalt perceptual principles play a major role in organizing how we perceive information in the blogosphere visualization: similarity and proximity (Ware, 2012). Given that most circles to the left are colored blue and most to the right are red, one might infer that the representation depicts two different groups that are color-coded based on their similarity (similarity principle). The suggestion of a two-group system is amplified by the spatial position of the circles: Circles with the same color are located closer to each

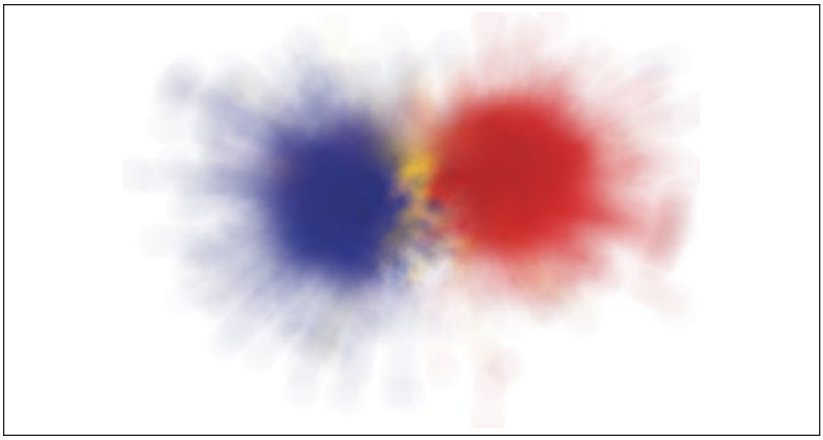

Figure 2. Adamic and Glance's (2005) blogosphere visualization presented here as a blurred image with the purpose of providing a sense of what one might perceive at first glance.

other and further away from circles with a different color, which indicates association among them (proximity principle). The two groups are separated horizontally, while connected by lines tinted in four colors: two familiar colors—blue and red— and two new colors—orange and purple. Without knowing anything about the information being visualized, the graphical features of the image invite three impressions: (1) an unmistakable association among like elements, given the color and spatial proximity of circular elements; (2) a clear distinction between the two main groups, given the spatial arrangement of similar elements—the blue and red clusters; and (3) a clear connection between the two groups, provided by orange and purple connecting lines.

Given the additional information that the image represents partisan affiliations of political blogs in the United States, viewers might easily infer that the colors are those of the bipartisan political system in the U.S., where blue stands for the Democratic Party and red for the Republican Party. The spatial configuration reinforces the underlying political ideologies, in that traditionally the U.S. Democratic/liberal party is referred to as the "left," and the U.S. conservative/ Republican party as the "right." To that end, the image confirms common conceptions about the U.S. political system in general, including a partisan divide, with the expected depiction of a larger number of connections happening within either political party than between them (Dimock, Doherty, Kiley, & Oates, 2014).

We would like to suggest that one of the factors that might have contributed to the wide dissemination of this image is that it visually "confirmed" the common intuition of a politically divided nation. This made the social representations processes of both objectification and anchoring easier, allowing the image to make a relatively easy transition as "visual evidence" of what many people already thought was the case, but had not been until then "scientifically proven" at this scale. Given its familiarity, one needs to understand relatively little about the science that produced this graph in order to integrate it into existing notions of what many believe to be "true" about U.S. politics.

We are not suggesting that Adamic and Glance claimed the image depicted "the truth." On the contrary, the image was one among several and perhaps the least central to the arguments and findings described in the original paper, as previously discussed. However, the familiarity of the patterns depicted in this image in particular may have especially enabled it to transcend its original scientific audience and become part of the mainstream political discourse, leaving it open for conclusions other than those originally claimed. Without knowing how the image was constructed, such as the metrics and the algorithm used to spatially encode the blogosphere into this particular graph, the image is easily simplified for public consumption.

Partisan (Dis)connection

While the general visual impression of a divided political system is communicated clearly, more ambiguous in the blogosphere visualization are the orange and purple lines connecting the two clusters, which represent about 9% of the total linking behavior. Although it is relatively easy to infer that these lines represent interconnections between the blogs of different partisan ideologies, the extent and nature of this interconnectedness is difficult to determine using the image alone. Adamic and Glance make no official claims about the degree of political disconnect represented by the visual, although the subtitle of the paper, "Divided They Blog," offers a clue about how they think the image ought to be interpreted.

There is also communicative ambiguity around the colors of the links between blogs of different political ideologies. Adamic and Glance (2005) provided the following caption for their visual to aid understanding:

> Community structure of political blogs (expanded set), shown using the GUESS visualization and analysis tool. The colors reflect political orientation, red for conservative, and blue for liberal. Orange links go from liberal to

> conservative, and purple ones from conservative to liberal. The size of each blog reflects the number of other blogs that link to it. (p. 37)

From this, readers can understand that the different colors represent different directions in the linking patterns. However, even the most careful reader could easily make a false inference about this directionality. Orange and purple are not standard colors in the U.S. political system, and in the absence of a mental model that can serve as a basis for interpreting what they mean, readers must rely on perceptual visual clues in the image. If anything, one might intuitively associate orange with red, and purple with blue, but that is the *opposite* encoding that we see in this system, where orange lines stand for links originating in blue circles (Democratic nodes).

Moreover, the orange lines were rendered after the purple ones so that they appear on top, and orange is the more salient of the two colors, the general impression is that of mostly orange connections, with purple lines barely visible at first sight (Figure 3).

This exemplifies how the complexity enabled by computational social science techniques may be partially or wholly obscured by the visualization. Based on the latter visual impression, readers might mistakenly believe that cross-party links most commonly originated from liberal blogs. But, mathematically, the paper indicates the opposite is true—conservative blogs link to liberal blogs slightly more often than liberal blogs link to conservative blogs (Adamic & Glance, 2005, p. 40). Although Adamic and Glance were careful to base their claims on mathematical calculations, a reader relying on the image alone would very likely reach the wrong conclusion. We point this out, not as a critique of Adamic and Glance who were meticulous in documenting their claims, but to underscore how computational social science visualizations can simultaneously depict many types of information that might invite different inferences—some more visually apparent than others. This challenge might be further intensified by the fact that many reproductions of this visualization did not include description of the encoding system, the accompanying mathematical results, or any of the other visualizations included in the original paper.

Broader Dissemination

Following publication in 2005, Adamic and Glance's paper received over 1,000 citations in academic publications,[4] a figure that is high, although not unheard of in the information science community. More atypically, the blogosphere visualization has been reproduced hundreds of times on its own, mostly with reference to the original article but often without any additional context. These reproductions serve as useful examples to demonstrate how a single

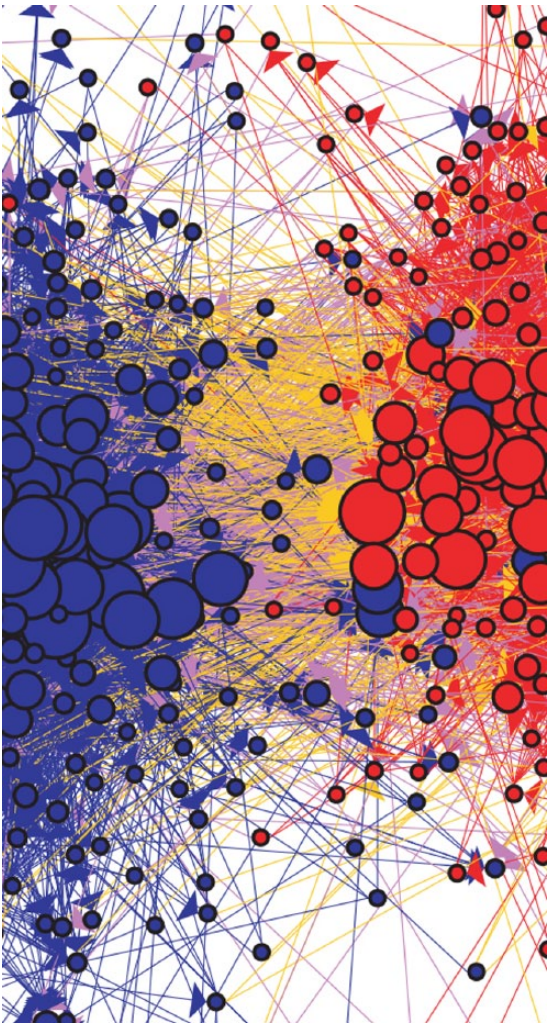

Figure 3. Detail of the central part of Adamic and Glance's (2005) blogosphere visualization to show the general impressions caused by the orange and purple connecting links between opposing political blogs.

complex image can transcend its original scientific purpose, simultaneously promoting computational social science as a mode of inquiry and obscuring the complexity that is one of the hallmark values of the approach.

Connected

Connected is a popular-press book about network science, aimed at informing the general public about how social networks influence human behavior in a variety of contexts (Christakis & Fowler, 2009).[5] Christakis and Fowler reproduced the blogosphere visualization in a chapter of the book examining American politics, titled "Politically Connected" (pp. 172–209), as Plate 6. The original caption by Adamic and Glance was replaced with one describing the encoding—what circles, lines, and color stand for—and it concluded, "This network map shows that the political blogosphere is *highly polarized* [italics added]" (Christakis & Fowler, 2009, Plate 6).

For Christakis and Fowler, Adamic and Glance's visualization is used to discredit the assumption that the Internet would bring people of opposing political positions closer:

> The hope was that we would discuss issues of the day in a nearly ideal Jeffersonian form of democratic exchange. But Lada Adamic . . . has produced some stunning images of these exchanges that show nothing could be further from the truth

They continue,

> What immediately stands out is the extreme separation between liberals and conservatives. If the hope of the Internet was that these two groups would talk to each other, the blog network reveals that these hopes have been utterly dashed. Just like the real-world political networks studied by Lazarsfeld and Berelson and later by Huckfeldt and Sprague, the online social network appears to be strongly homophilous and polarized. This suggests that political information is used more to reinforce preexisting opinions than to exchange differing points of view. (Christakis & Fowler, 2009, p. 206)

Christakis and Fowler present the blogosphere visualization as emblematic of a "strongly homophilous and polarized network," something of a departure from and simplification of the original text. Furthermore, they suggest the visualization depicts truthfulness, a statement they claim to be self-evident from the image. Identification of a "highly polarized network" is demonstrated by means of comparison with a set of models of hypothetical networks depicting four potential states: complete, high, medium and low polarization (Christakis & Fowler, 2009, p. 197). Given that the orange zone is as perceptually evident as the two opposing blue and red clusters in the blogosphere visualization, we posit that it is a simplification to suggest that the image self-evidently suggests political polarization. Interpreting the

meaning of the orange and purple lines in the visualization is central to understanding the segregation between the two political parties being represented, for reasons discussed above. We contend that the extent of that connection or segregation is not immediately clear from the visual alone.

Interestingly, in Adamic and Glance's original paper, there is another figure that we consider to be more emblematic in depicting polarization. The figure illustrates the subset of 40 A-list political blogs that were examined in more detail, and it shows a less dense graph in which the lines depicting fewer than 25 hyperlinks were removed. The result is a completely fragmented graph of liberal and conservative blogs, with no links across political affiliations (Adamic & Glance, p. 41). However, we believe that graph was not chosen because it was not so easily objectified for public consumption. The 40-blog figure is less familiar, carrying less visual weight with only 40 nodes and a few dozen links, such viewers are unlikely to understand that they are looking at a depiction of a two-party political system. In the blogosphere visualization, more data not only introduces noise (incomplete separation) but also increases the visual salience of a system of two (mostly) disconnected parties. If readers simplify the image by ignoring the links connecting the parties, then the full blogosphere visualization more easily communicates the desired message of political polarization.

The use of the blogosphere visualization to represent political polarization underscores one of the challenges in network representations, and in computational social science visualizations more broadly. Complex systems are often described by more properties than one can either visually perceive or visually represent in the constrained dimensions of the paper or the monitor screen. For example, one of the largest problems of network visualizations results from the occlusion of nodes and link crossings that tend to obliterate the structure they are supposed to reveal. This is not a trivial problem, given the large data sets used in these graphs, and is perhaps one of the reasons we so often see "hairball" network displays—a visual mass of circles and lines that are hard for even trained experts to read and extract meaning from. As Hansen, Shneiderman,and Smith (2010) contend,

> Network visualizations are often as frustrating as they are appealing. Network graphs can rapidly get too dense and large to make out any meaningful patterns. Many obstacles like vertex occlusions and edge crossings make creating well-organized and readable network graphs challenging. (p. 47)

Christakis and Fowler rely on the spatial schema in the blogosphere visualization to simplify the interpretation of the graph. However, this can be a risky interpretative practice as the partiular spatial arrangement in that image

is only one of several possible layouts. Selection of the spatial layout often happens on a trial-error process toward finding the algorithm that bests displays and, thus, conveys the important aspects of the research being undertaken. There are few standards that guide the spatial arrangement of network visualizations.

Furthermore, each algorithm imposes a set of constraints and generates a set of problems. For example, algorithms might address how to minimize edge crossing and node overlap, or how to maximize discovery of clusters, and so on. Not only does the algorithm affect the graph topology but, more importantly, how the graph might be perceived and interpreted. In a seminal study examining how the manipulation of a graph layout changed viewers' perceptions of grouping, McGrath, Blythe, and Krackhardt (1996) concluded,

> The results of our analysis suggest that spatial clustering has a significant effect on viewers' perceptions of the existence of groups in networks. . . . These preliminary results emphasize how fragile conclusions drawn from layouts of networks can be, and how important it can be to seek a clear depiction of a network. (p. 28)

To illustrate this point, we used GUESS, the same tool as Adamic and Glance, to represent the political blogosphere data set but using different algorithms as can be viewed in Figure 4.[6] We would argue that without the network metrics, each of these graphs would be interpreted differently, when in fact the differences are an artifact of the layout construction and not of the phenomena being depicted.

Furthermore, rendering graphs with an algorithm will result in a slightly different spatial layout at each iteration (see Figure 5). This has two consequences that are worth pointing out: First, there is not a single image that will result from automated methods of rendering graphs, rather this will depend on the algorithm and the metrics as well as the number of iterations (Figures 4 and 5); and second, interpretation of the graph might be affected by unintentional spatial groupings and other visual residues (or noise) generated by the algorithm.

On the other hand, presenting only the metrics might also lead to false conclusions. This point was demonstrated 40 years ago by Anscombe (1973), who showed that four data sets with the exact same statistical measures behaved differently when represented graphically (Figure 6).

Given the complexity of the phenomena under examination in the blogosphere visualization, or in any other computational social science research, there is no easy way to simplify an image for public consumption, while retaining the complexity and detail in the data. We contend that one

Figure 4. We used the GUESS online demo applet (http://www.hpl.hp.com/re-search/idl/demos/politicalblogdemo.html) to render these six layouts of the citation pattern between the 40 top political blogs in the months preceding the 2004 election. Note that the data set in the applet is the same as the one used by Adamic and Glance (2005, Figure No. 3). Top row, from left to right: GEM, Circular, Physics; and in the bottom row: Kamada-Kawai, Fruchterman-Rheingold, Random. This comparison helps demonstrating the fallacies of looking only at network layouts without inclusion of the metrics used for their construction.

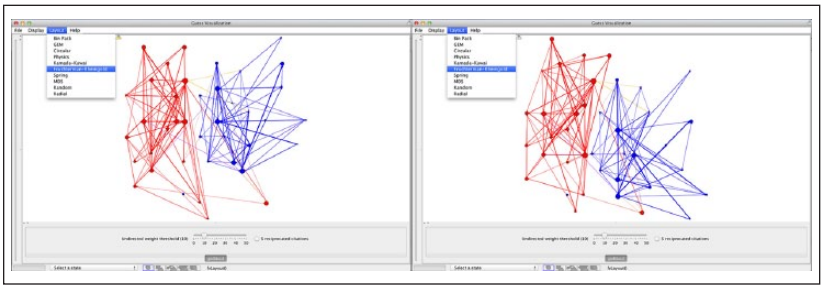

Figure 5. These two images were generated in the GUESS online demo applet (http://www.hpl.hp.com/research/idl/demos/politicalblogdemo.html) with the same data set used by Adamic and Glance (2005): the citation pattern between the 40 top political blogs in the months preceding the 2004 election. Note the slight differences in the positioning of nodes and links between the two graphs, both rendered with the Fruchterman-Rheingold algorithm.

visualization is just not enough to convey the whole story. Furthermore, we suggest reproducing multiple images as well as accompanying those with the metrics used for their construction, so as to provide a fuller understanding of

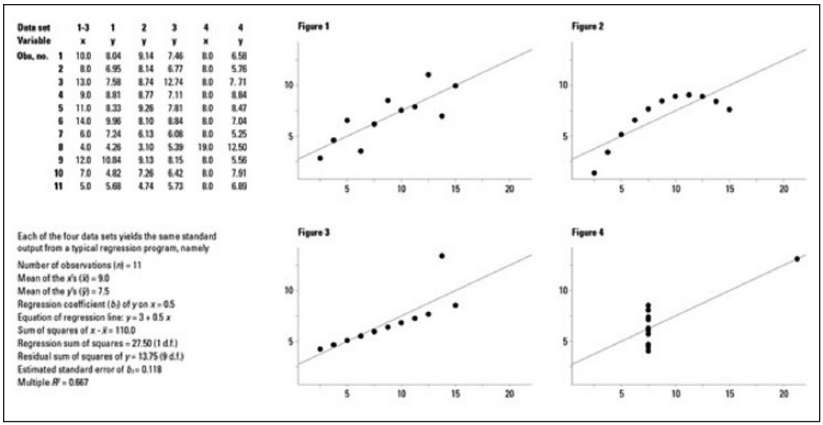

Figure 6. Table and graphs redrawn from Anscombe (1973) demonstrating the fallacies of relying only on numeric metrics without also considering graphical layouts when interpreting statistical data.

the scientific findings. In this case, Adamic and Glance provided a suite of visualizations and network metrics in the original paper that could have been included in those later reproductions, but were not.

From Evidence to Metaphor

In 2009, David Lazer and 14 colleagues (including Adamic) from across the social and information sciences published a call to action for researchers to embrace digital data to understand human behavior using a new approach they dubbed *computational social science* (Lazer et al., 2009). Alongside their call, the authors reproduced the blogosphere visualization as an example of the sort of work computational social scientists might pursue. In this case, rather than representing a particular scientific finding, the blogosphere visualization was used as an illustration in the traditional sense of the word, visually representing what the authors meant by "computational social science."

Just as an image of the brain can be used as a visual stand-in for the field of neuroscience, in *Science*, the blogosphere visualization was used as a visual stand-in for the emerging field of computational social science.[7] The blogosphere visualization itself is not referred to in the text of the article, although the possibility of tracking online communication about politics is among a dozen examples of the promise of computational social science. However, the selection of this particular image was not arbitrary. The work of

Adamic, one of the *Science* coauthors, was an early representative of the types of research the authors wanted to encounter more often. The blogosphere visualization represented a suitably large and complex data set and yet had the benefit of being anchored in the general public understanding of the U.S. political system, making the image understandable with only a very short caption to explain. Particularly apropos was the fact that the blogosphere visualization represented insights about the democratic process and not about capitalist and/or private enterprise. One of the central goals of the *Science* piece was to encourage scientists to take back data-driven social science before it becomes "the exclusive domain of private companies and government agencies" (Lazer et al., 2009, p. 721). To that end, this particular visualization represented exactly the type of work the authors hoped to see more of.

We argue that the reproduction of the blogosphere visualization in *Science* helped with its current status as an iconic visualization within computational social science, to the point of becoming somewhat synonymous with the term *computational social science*. The reproduction of the blogosphere visualization in *Science* as well as in the book *Connected* also likely contributed to its wide dissemination, as evidenced by reproductions in hundreds of articles and blog posts written since 2009, including articles in several languages.[8] Not surprisingly, since it was anchored in existing public understandings of U.S. politics, as the image transcended scientific publications into more popular sources, it was frequently used in the context of political discussions. We found several examples of reproductions that were used to illustrate polarization in American politics (Benkler & Shaw, 2010; Berkowitz, 2011; Freeberg, 2012; Zhou, 2012) and occasionally in discussions of politics in other countries as well (Hopkins, 2011; Lee, 2011).

Strikingly, many of the reproductions omitted a caption describing the image. It is not simply that the original caption was not reproduced, rather, in several instances in popular writing, the reproduction of the blogosphere visualization lacked any caption at all. Without a caption, the image fully realized its transformation from an illustration of scientific process, to evidence of scientific results, and finally to an icon emblematic of the whole topic of American politics, or more specifically, to the political division in the United States. To that end, for the general public, the blogosphere visualization may serve as a conceptual metaphor (Lakoff & Johnson, 2003), confirming the apparent "truth" of political division in the United States, evidenced by a complex representation of that divide.

Lakoff and Johnson (2003) contend that metaphors play a central role in the construction of social and political reality, especially when imposed on us by those in power:

> They do this through a coherent network of entailments that highlight some features of reality and hide others. The acceptance of the metaphor, which forces us to focus only on those aspects of our experience that it highlights, leads us to view the entailments of the metaphor as being true. Such "truths" may be true, of course, only to the reality defined by the metaphor. (pp. 157-158).

Given that "truth" is relative to a conceptual system defined by the metaphor, they suggest we examine the perceptions and inferences that follow from the metaphor as these are more relevant than examining whether a metaphor is true or false, in that we define our reality and how we proceed to act in the world based on those metaphors.

Discussion

While the communicative value of information visualizations to help convey findings in computational social science research is unquestionable, in this article, we have highlighted some concerns about how these images are disseminated and repurposed. Specifically, we highlighted how complex computational social scientific visualizations might spread, and through the processes involved with social representation, become simplified and situated within existing understandings of human behavior, to the extent that their original purposes and insights are obscured. The blogosphere visualization underscores both the opportunity and risk of visualizations that illustrate complex social processes. Because this image is relatively easy to objectify and anchor, it is not especially surprising that this image, and not others from that same paper, was the one that gained so much popularity. The image confirms many of our preexisting suspicions about American politics as bi-partisan, divided and oppositional. The image *makes sense* for many U.S. Americans (and, indeed, anyone with passing familiarity with the U.S. political system) and could be easily taken at face value as truthful. One can only speculate about whether a similar image showing a deeply connected political blogosphere, or one showing blogosphere with more than two constituent groups, would have garnered as much attention.

Familiarity affords several opportunities for an emerging and unfamiliar new science, such as computational social science. First, and most obviously, if we accept that images that are consistent with social expectations about the world are more likely to be taken as valid than those that are not, a familiar image lends credibility to computational social science as scientifically useful and accurate. Such images may also be easier to spread, as objectification and anchoring also afford the benefits of facilitating public communication

about otherwise unfamiliar objects. Finally, to the extent that they are used as visual metaphors, familiar images are easy to interpret, perhaps increasing the likelihood of reproduction because they do not require lengthy explanations of all they represent.

There are, however, also risks. First and foremost, one of the central risks of an image that conforms to our social expectations is that the image, and the new science it represents, will artificially reify existing social boundaries and biases (Howarth, 2006). Although the blogosphere visualization is intuitively easy for many to accept, there is no particular reason why the hyperlinking patterns among political blogs have to resemble the U.S. political system more generally. There are a number of social—and even physical—structures that enforce a two-party oppositional system in the offline world that simply are not present online. To be clear, we are not arguing that the blogosphere visualization is wrong, or that the researchers were somehow biased in their analysis; we simply mean to underscore how visualizations that confirm our conceptions of the world may be seen as inherently better than those that do not, especially when the world of study is a new one that was not and could not be studied before. However, when the blogosphere visualization was repurposed as a metaphor for the U.S. political system, which happened mainly outside the scientific community, most scientific references were removed, leaving readers with few cues to accurately interpret its scientific meaning.

In theory, this could have been fine—after all, many sciences have visual metaphors that are not intended to convey actual scientific meaning, but rather to signal the sort of science that went into producing the results. For example, it is not uncommon to see images of DNA sequences on popular press articles about genetic disease. We do not suppose that readers are expected to understand or interpret the gene sequences themselves; instead these images serve to signal that genetic science is happening behind the scenes. It is possible that images like the blogosphere visualization might come to be used as metaphors to signal computational social science (indeed, this is arguably how it was used in the *Science* article); however, in many cases the image is presented as evidence of a specific scientific result and not of a mode of scientific inquiry in general. Our concern, therefore, is that through the natural processes of objectification and anchoring, complex images without their underlying metrics could easily invite false conclusions.

To avoid this possibility, we have a number of suggestions. First, one recommendation is to include the caption and/or metrics within the image itself, such that when reused, it would always contain that minimum scientific information. This practice is commonplace in other forms of scientific data visualizations, such as histograms and pie charts that often include

details such as the margins of error on the measurements within the body of the graph. As computational social science visualizations become more popular, we propose that similar norms on reporting metrics within visualizations should become standard. Of note, leveraging new media and interactive publication formats, this added layer of information could also take the form of a hyperlink that would point back to the original research paper, helping the viewer further inspect and understand the visualization within its original context. Establishing standards and linking practices is an important area for future research, where we could examine ways to incorporate additional layers of meaning to enhance general public understanding of complex images.

Second, we would encourage both the authors and reproducers of computational social science images to embed several images together to ensure that a complex graph is not too easily simplified. Similar to the architectural drawings discussed above, it could become common practice for computational social scientists to rely on several images to fully illustrate their research processes and results, and for readers to *expect* several illustrations on the same premises. Imagine, for example, if Christakis and Fowler had used both the blogosphere visualization and the secondary visualization of 40 blogs that illustrated complete partisan separation of the blogs. This would have complicated their message somewhat, but we believe that the two images would ultimately have worked together to more accurately support the argument. The smaller graph could be used to demonstrate the central point about polarization at some scales, while the blogosphere visualization would have shown the full complexity of the idea space, allowing readers to understand the source domain, while also recognizing the limits of the polarization argument.

Finally, and most critically, we call on computational social scientists, especially network scientists, to interrogate their own visualization practices. As discussed above, constructing network graphs remains as much an art as a science, with few conventions regarding the "right" way to represent node-link data. Subject to the same norms of social representations as anyone else, there is a temptation for computational social scientists to choose a graph layout that seems to conform to general social understandings of the system(s) being represented. Indeed, if there were an unspoken convention to representing network graphs, it might be parsimony—where, of course, parsimony is influenced by subjective expectations about what the simplest explanation might be. Instead of relying on intuition about which graph layout best conforms to expectations, we would challenge computational social scientists to intentionally seek out graph layouts that challenge expectations about the world. Seeking counterevidence is common practice in qualitative research

and provides an excellent critical lens for understanding the range of claims that can (and cannot) be supported by the data at hand. Such counterexamples should not detract from the explanatory power of computational social science. Instead, coupled with the other techniques suggested here, they can become one of the central assets for driving new understandings about the complexities of human social behavior.

We hope critical evaluation of the use and reuse of computational social science visualizations, such as the one attempted in this article, might help determine when and how to reproduce them in the future. There are, of course, still many questions that need to be examined, in particular, conducting empirical research on how diverse groups interpret the blogosphere visualization given the different reuses and repurposes.

Nevertheless, our hope is that this article serves as a useful reference for how the public dissemination of computational social science visualizations, exemplified by the blogosphere visualization, comes with both opportunities and risks for advancing the public understanding of science. These displays of complexity can be compelling artifacts in the scientific discourse, advancing both scientific and lay understandings of computational social science. By adopting some of the techniques proposed here, we believe computational social science visualizations can retain their intuitive appeal, while still highlighting the central value of complexity in the study of human social behavior.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. http://www.isnatesilverawitch.com. Mark Coddington, a PhD student in the School of Journalism at the University of Texas at Austin, created the site and updated it throughout the 2012 election season. The quoted declaration appeared on the site the day following the 2012 presidential election and was still posted as of April 1, 2014.
2. Citation figures from Google Scholar, as of April 1, 2014.
3. Readers are directed to the online version of this article for full-color graphics to support this discussion.
4. Citation figures from Google Scholar, as of April 1, 2014.

5. *Connected* was written for the general audience. Although we have no way to learn the total number of copies sold, we imagine it is quite large given that it has been translated into 18 languages at the time of this writing and since its original publication in 2010.
6. Readers can explore the political blog data using GUESS online at: http://www.hpl.hp.com/research/idl/demos/politicalblogdemo.html. Images for this article were created on August 10, 2014.
7. *Science* is a peer-reviewed general-science weekly journal published by the American Association for the Advancement of Science in the United States. According to the 2014 Media Kit-Product Advertising document, the magazine has 129,551 worldwide print subscribers, 570,400 readers of the print weekly publication, and 3.4 million unique visitors to the *Science* website each month.
8. As evidenced by a Google image search, as of April 1, 2014.

References

Adamic, L. A., & Glance, N. (2005). The political blogosphere and the 2004 US election: Divided they blog. In *Proceedings of the 3rd International Workshop on Link Discovery* (pp. 36-43). New York, NY: ACM.

Adar, E. (2006). GUESS: A language and interface for graph exploration. In *ACM CHI '06 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 791-800). New York, NY: ACM.

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, *27*(1), 17-21.

Aral, S., Muchnik, L., & Sundararajan, A. (2009). Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences of the USA*, *106*, 21544-21549.

Bender-deMoll, S., & McFarland, D. A. (2006). The art and science of dynamic network visualization. *Journal of Social Structure*, *7*(2), 1-38.

Benkler, Y., & Shaw, A. (2010). *A tale of two blogospheres: Discursive practices on the left and right* (Berkman Center for Internet and Society Working Paper Series). Retrieved from http://cyber.law.harvard.edu/publications/2010/Tale_Two_Blogo-spheres_Discursive_Practices_Left_Right

Berkowitz, R. (2011). *The occupy movement: Visualizing change*. Retrieved from http://www.hannaharendtcenter.org/?p=3561

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature*, *489*, 295-298.

Card, S. K., Mackinlay, J. D., & Shneiderman, B. (1999). *Readings in information visualization: Using vision to think*. San Francisco, CA: Morgan Kaufmann.

Christakis, N. A., & Fowler, J. H. (2009). *Connected: The surprising power of our social networks and how they shape our lives*. New York, NY: Little, Brown.

Cioffi-Revilla, C. (2010). Computational social science. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*, 259-271.

Dimock, M., Doherty, C., Kiley, J., & Oates, R. (2014). *Political polarization in the American public*. Washington, DC: Pew Research Center.

Freeberg, J. (2012). *Hey presidential candidates, welcome to the age of social media*. Retrieved from http://blogs.cornell.edu/info2040/2012/09/05/5291/

Gershon, N., Eick, S. G., & Card, S. (1998). Information visualization. *Interactions*, *5*(2), 9-15.

Hansen, D., Shneiderman, B., & Smith, M. A. (2010). *Analyzing social media networks with NodeXL: Insights from a connected world*. San Francisco, CA: Morgan Kaufmann.

Hopkins, J. (2011). *How many blogs are there in the Malaysian SoPo blogosphere?* Retrieved from http://julianhopkins.net/index.php?/archives/308-How-many-blogs-are-there-in-the-Malaysian-SoPo-blogosphere.html

Howarth, C. (2006). A social representation is not a quiet thing: Exploring the critical potential of social representations theory. *British Journal of Social Psychology*, *45*, 65-86.

Lakoff, G., & Johnson, M. (2003). *Metaphors we live by*. Chicago, IL: University of Chicago Press.

Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, *11*, 65-100.

Lazer, D., Pentland, A. S., Adamic, L. A., Aral, S., Barabasi, A. L., Brewer, D., . . .van Alstyne, M. (2009). Life in the network: The coming age of computational social science. *Science*, *323*, 721-723.

Lee, C. (2011). *The Irish blogosphere*. Retrieved from http://sociograph.blogspot.com/2011/02/irish-blogosphere.html

McGrath, C., Blythe, J., & Krackhardt, D. (1996). Seeing groups in graph layouts. *Connections*, *19*(2), 22-29.

Moscovici, S. (1961). *La psychanalyse, son image et son public*. Paris: Presses Universitaires de France.

Moscovici, S. (1988). Notes towards a description of social representations. *European Journal of Social Psychology*, *18*, 211-250.

Newman, M. (2010). *Networks: An introduction*. New York, NY: Oxford University Press.

Pauwels, L. (2006). *Visual cultures of science: Rethinking representational practices in knowledge building and science communication*. Lebanon, NH: University Press of New England.

Pinker, S. (1990). A theory of graph comprehension. In R. O. Freedle (Ed.), *Artificial intelligence and the future of testing* (pp. 73-126). Hillsdale, NJ: Lawrence Erlbaum.

Rateau, P., Moliner, P., Guimelli, C., & Abric, J.-C. (2012). Social representation theory. In Van P. A. M. Van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *The handbook of theories of social psychology* (Vol. 2, pp. 477-497). Thousand Oaks, CA: Sage.

Silver, N. (2012). *The signal and the noise: Why so many predictions fail-but some don't*. New York, NY: Penguin.

Ware, C. (2012). *Information visualization: Perception for design (interactive technologies)*. San Francisco, CA: Morgan Kaufmann.

Zhou, J. (2012). *The mental models of politics*. Retrieved from http://systemsandus.com/2012/07/19/the-mental-models-of-politics/

Zhu, M., Huang, Y., & Contractor, N. S. (2013). Motivations for self-assembling into project teams. *Social Networks*, *35*, 251-264.

Author Biographies

Brooke Foucault Welles is an assistant professor of communication studies at Northeastern University. Her research focuses on the relationships between social networks and human communication, with particular emphasis on how new media influence the role that networks play in the pursuit of personal and organizational goals.

Isabel Meirelles is an associate professor of graphic design at Northeastern University. Her intellectual curiosity lies in the relationships between visual thinking and visual representation, with a research focus on the theoretical and experimental examination of the fundamentals underlying how information is structured, represented, and communicated in different media.