Dot, Line, Network: Helping Individuals Make Sense of "New Data"

Emilie W. Gould

Communication Department, SUNY Albany, Albany, NY,USA
egould@albany.edu

Abstract. Ubiquitous computing has led to an ever-increasing cascade of information about us, our friends, our societies, and the planet. Lima and others view this "new data" as an opportunity for individuals to develop network thinking; once people understand the whole, they can better control their contribution to global social issues like climate change. However, at present, such data is difficult to interpret by anyone, let alone by non-specialist users.

I believe that a variety of issues stand in the way of individuals understanding complex data sets. I will begin by discussing cognitive style (deductive and inductive logic). Then, after considering existing graphic principles for dealing with "visual complexity," I suggest interfaces need to provide indications of place, date, validity, probability, and privacy. Finally, I briefly discuss some of the boundaries that exist between my networks of data and yours due to the hidden algorithms of search engines and the challenge of creating common ground when visualizations are increasingly personalized.

Keywords: ubiquitous computing, visual complexity, network thinking, cognitive patterns, graphics.

1 Introduction

With ubiquitous computing, the whole world becomes data to be collected, mined, and interpreted. No longer are business decisions based on sales and profits; companies stake their futures on trends identified from aggregate Facebook pages, Google searches, and Flicker photo albums. Better and better models have been developed to identify correlations – or, at least, apparent correlations – between variables. Causality can be assigned to seemingly disconnected events. The stock market drops because of Superstorm Sandy; no, the Presidential Election; opps! the Eurozone crisis; or something…

It is one thing for companies to bet the store on all this "new data," but how can individuals navigate this new world? The traditions of European logic – deduction, classification, and problem-solving – are embedded in computer science, experimental psychology, and interface design. Such analysis is good when it leads to insights into the nature of reality and facilitates pragmatic change but, at a fundamental level, it depends on understanding one's data in terms of advance hypothesis building. By

A. Marcus (Ed.): DUXU/HCII 2013, Part IV, LNCS 8015, pp. 496–505, 2013.

contrast, non-European logics, based on non-Aristotelian philosophies like Confucianism [1], tend to emphasize induction, resonance, and circular thinking. Can interface design help people shift their cognitive styles?

Second is the issue of representation. Edward Tufte [2-5] and others [6] have proposed excellent rules for displaying quantitative information. Many of these guidelines are now conventions but, as Manuel Lima [7] shows, massive datasets and network thinking require new forms of visualization. However, his guidelines, like those of his predecessors, still tend to focus on functionality and interface aesthetics without considering the problem of variation within massive datasets. Too often visualizations gloss over missing or substandard data. Unfortunately, the more finished a graph, the more likely people will develop "swift trust" [8] in its apparent message. At a pragmatic level, designers need to develop new cues to help people recognize constraints of place, date, validity, probability, and privacy on causality.

Finally, the issue isn't whether we can escape this new world of new data – we can't – but how to better live in it. As Manuel Castells notes, "We live in a culture of not virtual reality, but real virtuality because our virtuality, meaning the internet networks, the images are a fundamental part of our reality. We cannot live outside this construction of ourselves in the networks of communication." [9] But my virtuality may not include your virtual world. Search engine algorithms deliver personalized information based on my previous queries and visits to websites. Increasing customization is a danger to establishing common ground. If individuals cannot recognize the boundaries that separate their networks of "common-sense" interpretation from the networks of others, then solving existential problems of living together will not be possible.

2 Visualizations and Cognitive Style

Most software for representing quantitative information comes from Western software companies (or Western labs in non-Western countries) and reflects the cognitive style of its designers. Despite increased testing in Asia and other parts of the world, most graphical programs reflect European norms of logic. In addition, it remains true that most users of the Internet are more like one another – young, well-educated, economically-advanced, and male – than like their national demographic. [10] They use software that has been developed for Western business and believe they understand the results of their data visualizations.

However, as Edward Tufte warned a decade ago, software embodies distinctive cognitive styles. In his rant against PowerPoint, he focused on some of the deficits: "foreshortening of evidence and thought, low spatial resolution, a deeply hierarchical single-path structure as the model for organizing every type of content, breaking up narrative and data into slides and minimal fragments, rapid temporal sequencing of thin information rather than focused spatial analysis, conspicuous decoration and Phluff ["chartjunk," logos, and clip art], a preoccupation with format not content, an attitude of commercialism that turns everything into a sales pitch." [11, p.4] He argued that presentation software was entirely speaker-oriented, ignoring both audience and content. An alternative explanation is that most software is constrained by European logic traditions.

2.1 European Logic

European logic has always been based on a linear model. Aristotle [12] defined the syllogism as consisting of a major premise, a minor premise, and a "necessary" conclusion:

- All men are mortal (major premise)
- Socrates is a man (minor premise)
- Therefore, Socrates is mortal (conclusion)

The syllogism moves smoothly from one point to the next in an ordered chain of inferences. One cannot reverse the premises or the conclusion, even though the minor premise can be dropped in an enthymeme because of shared knowledge:

- All men are mortal (major premise)
- Therefore, Socrates is mortal (conclusion)

Manuel Lima thoroughly explores the implications of modeling knowledge in linear and hierarchical terms. In particular, he explores the influence of tree diagrams – the tree of life, trees of knowledge, the tree of Porphyry, Joachim of Fiore's *Book of Figures*, tables of the liberal arts, Bacon, Descartes, Diderot and D'Alembert's *Encyclopédia* – from pre-history up through the end of the 18^{th} century. "This favored scheme, usually highlighting a hierarchical ordering in which all divisions branch out from a central foundational trunk, is ultimately a universal metaphor for the way we organize and classify ourselves and the world around us." [7, p.21]

When deduction, linear thinking, and hierarchical orders of information come together, certain types of analysis are made possible. As a result, the European intellectual tradition emphasizes cause and effect. Science relies on deductive thinking – propose a hypothesis and prove it. Our civil law is based on the idea that someone is almost always in control and responsible for events. Indeed, as we combine the two, "Acts of God" are becoming fewer and fewer and (now that 16% of the American public is atheist) less and less acceptable as an explanation for events.

2.2 Non-european Logics

Deduction represents just one cognitive style. Its reverse is induction – many observations analyzed to discover general principles. An hypothesis emerges from the data, rather than data being gathered to test a given hypothesis. One scans a wide variety of examples; if those examples do not contradict one another, one can conclude that something is true. Interestingly, truth is not a "logical necessity" as in deduction; instead, it is probabilistic – the more examples, the more likely a truth will emerge.

In addition, different classification schemes are used in non-European logics. In his introduction to Chinese categorization, Richard Nisbett contrasts the *part-whole dichotomy* of Chinese thought with the *object-class* relationship of Greek philosophy: "…[I]t is certainly the case that the ancient Chinese did not categorize the world in the same sorts of ways that the ancient Greeks did. For the Greeks, things belonged in the same category if they were describable by the same attributes. But the philosopher

Donald Monro points out that, for the Chinese, shared attributes did not establish class membership. Instead things were classed together because they were thought to influence one another through *resonance.* For example, in the Chinese system of the Five Processes, the categories spring, east, wood, wind, and green all influence one another.” [1, pp.137-138]

Influence is not the same as cause and effect; there is no necessary, linear, or hierarchical relationship between these categories. Just because “east” and “green” affect each other, east did not necessarily “cause” the change in green or vice versa. Like chaos theory, a small change in one category may ripple through others and create new relationships; another time the consequences could be different.

Condon and Yousef found that non-European argument and logic often defy formal proof. Conclusions may be grounded in anthropomorphism (the head is “higher’ than the feet), natural law (religious “first principles”), moral order (notions of “fate,” “karma,” or “justice”), or the primacy of testimony. “Facts” and physical evidence accepted in Western science or law may be ignored because of a focus on relationships and subjective interpretation. [13, pp.219-221] Mutuality leads to circular thinking. At the heart of Hindu, Buddhist, and Taoist traditions, everyone and everything are connected. Interpretation and “discovery” based on these alternative logics can reveal new insights into our world.

2.3 Best of Show?

It is not likely that people can completely escape their initial cognitive programming – from birth, infants in different countries learn to sense and think about their environments in significantly different ways. Hofstede [14, 15], in particular, points to the long-term survival of fundamentally different cultural values, attitudes, and behaviors despite globalization. But, it is important that individuals become more cognitively flexible. At the very least, they can learn the difference between hierarchical and network data and become conscious of new ways to interpret each type.

Nisbett believes that we are already changing: “A shift in characteristic social practices could therefore be expected to produce a shift in typical patterns of perception and thought.” [1, p.229] He believes we will recognize the many ways we are linked and begin to “think together” as we encounter one another more often.

Interface design should further this process and help people shift their cognitive styles by providing new modes of graphic analysis. Although experts are able to model complex processes, most individuals do not have the knowledge base, time, or resources to explore the same information. But people do have goals and want to live lives of meaning. Opportunities to explore desired outcomes and to work backwards to underlying principles could harness induction to improve sense-making. In addition, software should support additional forms of classification. Currently, one can weight different variables and look for consequences, but alterative classifications could help people discover a wider range of relationships. Finally, cause and effect take one only so far and limit discovery; resonance is a broader concept that has the potential to help individuals feel more, not less, connected to one another.

3 Guidelines for Visualization

A large corpus of rules and examples already guides graphic design. Since the 1980s, automated forms of data collection and the development of spreadsheet software have led to greater use of visualization in popular media, social science, business, and art. For me, two authors stand out: Edward Tufte for his series of four books on information display and Manuel Lima for his ground-breaking review of new modes for visualizing network data. Each author reviews the early, experimental days of data analysis to identify fundamental design rules and best practices. For Tufte, those days range from the 16^{th} to late 18^{th} century; for Lima, the innovative days of mapping new data in all its complexity are taking place right now. However each author's guidelines focus first on design, and occasionally on the data. I suggest more work can be done to help people become more data-literate.

3.1 Edward Tufte

In 1983, Edward Tufte released his landmark book, *The Visual Display of Quantitative Information*, based in part on his analysis of graphics developed in the two centuries before William Playfair, which he contrasted with late 20^{th} century graphics used in newspapers, scientific reports, and books. This book popularized the notion of "chartjunk" – "non-data-ink or redundant data-ink" [2, p.107] that fails to advance interpretation and often undermines validity.

Tufte defined graphical excellence as "complex ideas communicated with clarity, precision, and efficiency" and demanded that graphic designers focus on providing views with "the greatest number of ideas in the shortest time with the least ink in the smallest space." [2, p.51] He warned against visual distortion, scale variations, 3D representations of area that obscure numeric comparison, non-standard units of measurement, data taken out of context (lacking comparable variables such as time and place), and the creation of inappropriate user expectations. Good designs should be "data-rich" to maximize trust and increase confidence. "Low-information designs are suspect: what is left out, what is hidden, why are we shown so little?" [2, p.168]

Influential as Tufte's rules have been, they focus on presentation; underlying them is the assumption that all the data in a given visualization is equally valid and that there is one narrative to be discovered within it.

Tufte's second book, *Envisioning Information*, addressed the issue of representing multi-dimensional information on two-dimensional surfaces (computer screens as well as paper). There are references to the increasing sophistication of computer displays – movement and shifting perspectives applied to three-dimensional scatterplots; layered information; and better rendition of color (Apple was just beginning to release color monitors for its home computers). In retrospect, his arguments for micro/macro readings (dense information that reads differently at different zooms), proportion and harmony in overlapping (and separated) layers, and "small multiples" (visual sequences showing change over time) anticipate the ability of Google Maps to zoom from space to a street-side view of your home, GIS mapping software, and computer modeling.

He again railed against over-simplification of data: "If the visual task is contrast, comparison, and choice – as so often it is – then the more relevant information within eyespan, the better…. High-density display also allows viewers to select, to narrate, to recast and personalize data for their own uses. Thus control of information is given over to *viewers*, not to editors, designers, or decorators." [3, p.50]

Putting individuals in charge of their interpretation of the data sounds good. However, much of the book focuses on the need to develop "narratives of space and time," on the premise that visualizations should continue to demonstrate a master (linear/hierarchical) story.

Tufte's third book, *Visual Explanations*, "describes design strategies – the proper arrangement in space and time of images, words, and numbers – for presenting information about motion, process, mechanism, cause and effect." [4, p.9] He revisits some familiar design territory but, for the first time, discusses issues of data validity when comparing the influence of John Snow's map of the 1854 London Cholera Epidemic to the decision to launch the Space Shuttle Challenger in 1986. Snow placed his data (cholera deaths) on a map – "an appropriate context for assessing cause and effect"; noted exceptions – "quantitative comparisons" of homes that suffered mortality and a brewery and a workhouse that did not; investigated outliers in search of "alternative explanations and contrary cases'; and commented on weaknesses in his data – "possible errors in the numbers reported in graphics." [4, pp.27-37] By comparison, the data advanced by Thiokol engineers to prevent the launch of the Challenger in 29° weather failed to conclusively show the influence of temperature on O-ring erosion. The same issues showed up in charts displayed to the subsequent presidential commission. Sequencing problems (too many overheads), chartjunk (cute pairs of rockets), lack of visual clarity in handling the causal variable (temperature), and inappropriate order (time rather than temperature) all obscured the crucial link. [4, pp.39-49]

Tufte concludes that, "Visual representations of evidence should be governed by principles of reasoning about quantitative evidence. For information displays, design reasoning must correspond to scientific reasoning. Clear and precise seeing becomes as one with clear and precise thinking." [4, p.53] Elsewhere he states that the designer must develop an appropriate hypothesis and logically prove it. Is the interpretation "true? Is the visualization "accurate"? Does it avoid inaccurate readings? Is the data properly documented? Is context preserved? [4, p.70] But how?

In *Beautiful Evidence*, his most recent book from 2006, Tufte focuses on a esthetics, restating many of his previous rules for graphic design.

3.2 Manuel Lima

Like Edward Tufte, Manual Lima produced a beautiful book that analyzes the graphic principles revealed in hundreds of examples of visualization; unlike Tufte, Lima embraces the issue of ubiquitous computing and its floods of data. *Visual Complexity* (2011) focuses specifically on the visualization of networks and the tension that exists between that which is being measured, data, and representation. [7]

In the course of completing his MFA thesis on the blogosphere, Lima became entranced with networks, seeing them in all sorts of places (both natural and artificial) and recognizing them as “an inherent fabric of life.” Networks describe our brain activity, the intersecting biological and industrial cycles that produce our food, train schedules, friendship webs, citation indexing, conference attendance – practically all the relationships that physically and socially define us. At the same time, he realized foundational depictions of visual complexity were disappearing from the web. Links broke as people shifted from one institution to another, data became dated, and people neglected to update old code. Experimental visualizations might be abandoned or spawn new memes; in either case, the original image could vanish. Online visualizations are ephemeral. Lima’s book and its website, visualcomplexity.com, attempt to document and slow the process.

Many of the characteristics Lima ascribes to networks resemble characterizations of non-European logic – rejection of centralization (especially, hierarchical information), rejection of finalism (goal-direction), and rejection of essentialism (permanence). On the positive side, networks are multiple and multi-linear, interdependent, flexible, changeable, and diverse. Lima cites Warren Weaver’s claim that problems of organized complexity can no longer be solved with analytic methods developed for problems of simplicity or disorganized complexity. “The complex connectedness of modern times requires new tools of analysis and exploration, but above all, it demands a new way of thinking. It demands a pluralistic understanding of the world that is able to envision the wider structural plan and at the same time examine the intricate mesh of connections among its smallest elements. It ultimately calls for a holistic systems approach; it calls for network thinking. [7, pp.45-46]

Data is still made up of dots (nodes) and lines (links and relationships) but it can no longer be accurately represented on a two-dimensional grid; lattices, matrices, 3-dimensional space, animations, and fractals better show its complexity. One chapter identifies fifteen new forms: arc diagram, area grouping, centralized burst, centralized ring, circled globe, circular ties, elliptical implosion, flow chart, organic rhizome, radial convergence, radial implosion, ramification, scaling circles, segmented radial convergence, and sphere. The examples are breath-taking in 2-D; the notion that one could enter and manipulate these new forms of data in 3-D to discover emergent properties is tantalizing. But I, for one, lack the visual literacy to understand them …

Lima quotes Tufte on the need to represent multivariate, high-density data. Some of his other principles are conventional (start with a question; look for relevancy). Rules from graphic design, human-computer interaction, cartography, cognitive processing, visual perception, color theory, composition, typography, and spatial arrangement continue to hold true. But most of his other guidelines reflect the need to accommodate much larger datasets with vastly more nodes and connections:

1. Enable multivariate analysis (include additional variables for both causality and holism)
2. Embrace time (show the effects of historical and dynamic processes)
3. Enrich your vocabulary (use multiple coding of nodes with color; add interactivity)
4. Expose grouping (apply the Gestalt laws of similarity, proximity, and common fate/motion)

5. Maximize scaling (enable macro, micro, and relationship views)
6. Manage intricacy (disclose information gradually through adaptive zooming, overview with embedded detail, and simultaneous focus with context – for instance, a fish-eye viewer) [7, pp.81-95]

The book places network visualization within recent art traditions – abstract expressionism, generative art, fractals, and networkism (maps and hidden territories). The one area Lima misses is game theory and the notion that we may want to interact with data visualizations as we interact with the programmatic components of games.

Visual Complexity concludes with several short essays by other writers. In one of the last chapters, Nathan You discusses the need for individuals to become more involved in data collection, analysis, and interaction. People generate huge amounts of data through their mobile devices; however, to analyze and interact with it, they need to know "what is going on." [7, p.247] But You gives few guidelines for what people actually need to learn. Instead, he looks to a utopian future where computer visualizations bring us closer rather than distance us.

3.3 Some Additional Suggestions

So, "what is going on" that is not covered in Tufte's and Lima's guidelines?

Most discussions of visualization focus on the representation and assume that it accurately models the thing that is being represented. But there is no guarantee that all the information in massive data sets is accurate because much of this data comes from different sources in different places at different times. Information complexity needs a new set of visual indicators and controls to represent additional aspects of data.

At a minimum, datasets need to provide indications of place and date. Information gathered in Japan before and after the meltdown at the Fukushima nuclear plant would need to be placed in context. In a simple graphic comparing the two dates, that is easy; in a more complex network visualization comparing various power technologies and various countries over time, individually coding blocks of data could be hard.

Similarly, not all data is equally valid and not all correlations are equally probable. Medical and social scientists often assemble data from various providers and research teams to look for demographic trends. Some data may be missing or reflect different scales and collection procedures, but the aggregation still permits discoveries. However, the data density should be clear. Tell the viewer which nodes are missing due to incomplete data and show levels of validity for the links that are there. Different lines between nodes could show different probabilities of cause and effect; different hue levels could indicate problems. The graphic might look like a patchwork, but such a patchwork would remind people that a visualization is only as good as its data.

Finally, more than 89 countries have privacy laws that restrict the information that can be gathered about individuals and how it must be protected. [16] Information about children is often protected more highly than information about adults. Information about employment or health may be subject to different rules in different places. The European Union has proposed a consolidated law that covers its 27 member states; the United States often relies on industry groups developing voluntary standards. [17] Sometimes people have the ability to remove their data from collections though often that freedom is illusory – as in the "privacy forms" signed by patients in the United States that give insurance companies full disclosure. All these privacy regulations limit the data equivalence. Even procedures to "anonymize" data differ widely across different organizations and jurisdictions.

Without more transparency, new controls, and new cues, people will not be able to understand how variations within data influence causal and network visualizations.

4 Search Engines and the Problem of Finding Common Ground

Optimism about the possibility of coming together and solving global problems remains strong. But – and this is a big "but" – there are alternatives to Nisbett's belief that European and Asian cultures will learn from one another and Lima and You's hopeful views of personal empowerment through shared data.

Limited global resources are already setting countries against one another – as in the melting Arctic Ocean. Personal data is often hijacked off the Internet for private gain; someone can assume your personal identity, destroy your credit rating, and ruin your personal reputation before you realize you have a doppelgänger. In the 2012 Presidential election, both parties set up data-mining projects to ensure that individual voters in key swing states were targeted with messages crafted to their demographic. Given enough intersecting data points, anonymous data proved easy to identify.

These are serious problems but not new. Countries have always been in competition at some level; identities are surprisingly vulnerable. But, one aspect of living in a culture of real virtuality is new. Search engines increasingly filter our experience. My view of the world is no longer yours. My queries, visits to websites, gmail account, Facebook timeline, and tweets in the Twitterverse personalize the information I will see on the web. Over time, our virtual experiences diverge – we may never find one another or never find the "missing link" to some problem we need to solve together.

Much of this personalization is for commercial gain. Because of my browsing behavior, I am bombarded with ads for hotels in Montreal, Iceland, and England but not for resorts in Cancun. But what is more worrisome is that I may not see information I need to know about global warming because I am a Republican, or understand the fears of gun owners because I am a Democrat. How, then, can individuals breach their separate virtualities and find common ground?

Search engines' proprietary algorithms are not open to user control, even though companies can rise to the top of the results list by paying a fee or hiring an agency to improve their hit rate. Alternative browsers exist but many, like Firefox, use parts of more commercial browsers for search and security. Some sort of "common ground" setting, based on our existential problems rather than geolocation, is needed to reopen the web for widespread collaboration.

5 Conclusion

Helping individuals make sense of new data presents a challenge to our field. We are no longer dots in a landscape; we are nodes, lines, and networks of highly-connected world citizens. We can only understand each other if we are allowed to explore alternative perspectives, some based on unfamiliar logics, on an uncensored Internet. In addition to creating new forms of visualization, interface designers should focus on encouraging holistic thought, revealing additional properties of data, and helping individuals merge their subjective experiences of the virtual world. People need help grounding complexity in physical experience and consciously working together for good.

References

1. Nisbett, R.E.: The Geography of Thought. Free Press, NY (2003)
2. Tufte, E.R.: The Visual Display of Quantitative Information. Graphics Press, Cheshire (1983)
3. Tufte, E.R.: Envisioning Information. Graphics Press, Cheshire (1990)
4. Tufte, E.R.: Visual Explanations. Graphics Press, Cheshire (1997)
5. Tufte, E.R.: Beautiful Evidence. Graphics Press, Cheshire (2006)
6. Wainer, H.: Visual revelations. Copernicus. Springer-Verlag, NY (1997)
7. Lima, M.: Visual complexity. Princeton Architectural Press, NY (2011)
8. Meyerson, D., Weick, K.E., Kramer, R.M.: Swift Trust and Temporary Groups. In: Kramer, R.M., Tyler, T.R. (eds.) Trust in Organizations: Frontiers of Theory and Research, pp. 166–195. Sage, Thousand Oaks (1996)
9. BBC News: From Networked Protest to ‘Non-capitalism (2012), http://www.bbc.co.uk/news/world-19932562
10. Johnston, K., Johal, J.: The Internet as a Virtual Cultural Region. Internet Research 9(3), 178–186 (1999)
11. Tufte, E.R.: The Cognitive Style of PowerPoint. Graphics Press, Cheshire (2003)
12. Aristotle: Rhetoric. Modern Library, NY (1984)
13. Condon, J.C., Yousef, F.S.: An Introduction to Intercultural Communication. Macmillan, NY (1974)
14. Hofstede, G.: Cultures and Organizations, Software of the Mind. McGraw-Hill, NY (1997)
15. Hofstede, G., Hofstede, G.J., Minkov, M.: Cultures and Organizations, Software of the Mind, 3rd edn. McGraw-Hill, NY (2010)
16. Greenleaf, G.: Global Data Privacy Laws: 89 Countries, and Accelerating. Privacy Laws & Business International Report, 115, Special Supplement. Queen Mary School of Law Legal Studies Research Paper, (1998-2012), http://ssrn.com/abstract=2000034
17. Data Protection Laws, An Ocean Apart. NY Times (February 3, 2013), http://www.nytimes.com