Using Data Visualizations to Study Digital Public Spaces

Yee Man Margaret Ng, Ph.D., Assistant Professor
Department of Journalism & Institute of Communications Research
University of Illinois Urbana-Champaign
Direct comments to ymn@illinois.edu

Accepted to First Monday (2022)

Abstract

This article reviews the history and current data visualizations in studying digital public spaces. I will discuss the recent development of visualizing raw data numerically, relationally, spatially, and textually. Each method involves different visual representations to integrate data collection with analysis and presentation of results. Through a case study of global web use, this article also demonstrates a thinking process and analytical workflow to incorporate data visualizations when studying digital public spaces, particularly in the midst of a global crisis.

Keywords

Data visualization, digital public spaces, social scientists, graphical representation

Introduction

Since the World Health Organization declared COVID-19 to be a pandemic, we have been inundated with daily news on the spread of the coronavirus filled with rates and percentages, charts and graphs, projections and probabilities. Many compelling analyses of the virus in news media have taken the form of visualizations—for example, the illustration and concept of “flattening the curve” (“Flattening the curve for COVID-19: What does it mean and how can you help?,” 2020)—to educate the public on the need to practice social distancing to reduce the spread of the virus. These visualizations do not simply present public health data but also possible future scenarios, guiding our behavior in a time of crisis.

“A picture is worth a thousand words.” Graphs and diagrams not only help readers grasp the essential content of a study, but also provide insight that traditional, descriptive statistics cannot. Research has long shown that individuals understand and better remember information communicated via pictures rather than through single words or short sentences (Carney and Levin, 2002; Few, 2009). Indeed, the ability to read and construct data visualizations is as critical as the ability to read and write text. Traditional data visualization methods—such as scatter plots, bar charts, histograms, line charts, and pie charts—have been widely used in social science research. Too often, however, graphs and diagrams that accompany most scientific research are created as afterthoughts and not given the attention they deserve.

Emerging Web 2.0-enabled technologies impact human interaction and participation. Social media, particularly, is a novel avenue for disseminating content and forming communities, providing a massive volume of data for social scientists to understand the underlying user behavior in digital public spaces. However, ingesting, visualizing, and analyzing such massive amounts of data is a substantial challenge. Therefore, new techniques that visualize both

quantitative and qualitative data are more critical than ever. If researchers seek to examine how ideas spread or how virtual communities form, it is important to understand the strengths and limitations of methodologies that analyze and visualize online activities.

This paper reviews the history and current data visualizations in studying digital public spaces. I will discuss the recent development of visualizing raw data numerically, relationally, spatially, and textually. Each method involves different visual representations to integrate data collection with analysis and presentation of results. Through a case study of global web use (Ng and Taneja, 2019), this article also demonstrates a thinking process and analytical workflow to incorporate data visualizations when studying digital public spaces.

Background

A Brief History of Data Visualization

Data visualization is not a new subject area. It has deep roots that stretch from early map-making and visual depictions to modern cartography, statistics, and other fields. Understanding its historical background to help us properly apply and execute visualization concepts that we still use today.

The use of data visualization dates back to 2500 BC, when the Babylonians used columns and rows to display transactions (Few, 2009). During the 10th century, tables and graphical depictions were used to display star and celestial body positions (Friendly, 2006). Several historical examples continuously arouse attention even to this day. One of the classics is French civil engineer Charles Joseph Minard's figurative map "Napoleon's March" (Figure 1). The graphic depicts the horrific loss of life that Napoleon's army suffered in 1812—of the 422,000

soldiers set off from Poland, 100,000 reached Moscow, but only 10,000 returned. The graphic rose to its prominent position in the data visualization world primarily thanks to Edward Tufte, one of the field's modern giants. In his classic 1983 text *The Visual Display of Quantitative Information*, Tufte declared that Napoleon's March "may well be the best statistical graphic ever produced" for its clarity and data density (Tufte, 1983, p. 40). The graphic is able to condense six different numeric and geographic facts into one graphic to illustrate the downfall of Napoleon's army:

1. Line orientation shows the direction of invasion and subsequent retreat
2. Line thickness indicates the number of troops who survived the hunger and cold
3. Line scale shows distance traveled
4. Labels depict notable rivers and cities
5. Dates indicate progress
6. The line chart below tracks the freezing temperature.

[INSERT FIGURE 1 ABOUT HERE]

Today, the graphic is not only a wonderful example of how visualizations can turn raw numbers into engaging stories about human events but also a powerful anti-war statement, conspicuously presenting the loss of human life.

Besides Minard, a few other key figures revered for their data visualizations continue to be influential in documenting humanitarian crises. For example, rather than plotting cases over time, physician John Snow mapped each cholera patient to their home during the 1854 London cholera epidemic (Figure 2). By visualizing the data in this way, Snow was able to interpret the virus was spread through contaminated public wells and discounted the miasma theory of foul air. His statistical mapping brought fundamental changes in London's water and waste systems.

The mapping was also recognized as a breakthrough in using geographical analysis to understand and solve a complex health problem (Tufte, 2004). The method is widely used today. Since the early stage of the COVID-19 pandemic, health institutions and universities have created many innovative trackers and maps. The one of John Hopkins University (“COVID-19 Dashboard,” 2020) is the most prominent.

[INSERT FIGURE 2 ABOUT HERE]

Florence Nightingale—often called “the Lady with the Lamp”—is most remembered as a pioneer of modern nursing, but her medical report also revolutionized the field of visual representation. Nightingale noticed that the main cause of death among the soldiers was not related to the war itself, but to infectious diseases that spread through British military hospitals during the Crimean War (1854–56). To alert the British government to these conditions, she marshaled data and presented the evidence as a set of polar area diagrams (Figure 3).

Nightingale’s diagrams resemble pie charts but segmented into twelve slices, each representing a month. Each slice has three sections: one for deaths from wounds in battle, one for disease (e.g., preventable illnesses such as typhus and dysentery), and one for “other causes.” The area of each colored section, measured from the center, is proportional to the represented statistics. The diagram evidenced that many soldiers were dying of infectious diseases. The diagram underscores how critical health reforms were in battlefield hospitals. These later became standard practice worldwide, and eventually helped save the lives of countless soldiers throughout history.

[INSERT FIGURE 3 ABOUT HERE]

In the latter half of the 20th century, publications about good statistical visualization practices abound, setting up exceptional visualizations that have the power to effect widespread social and political changes. These legendary publications include:

- Mathematician John Tukey's (1977) *Exploratory Data Analysis*, highlighting visualization as a critical step in understanding data sets
- Computer scientist William Cleveland's *The Elements of Graphing Data* (1985) and *Visualizing Data* (1993), stressing the use of visualization to thoroughly study the structure of data and to check the validity of statistical models fitted to data
- Statistician Edward Tufte's (1990, 1997, 2001, 2006) set of illuminating books, understanding the best way to display quantitative information
- Statistician Leland Wilkinson's (2005) *The Grammar of Graphics*, which shuns the notion of a fixed "chart typology" and instead encourages building up a graphic from multiple layers of data

These works, as well as the rapid progress in computing power and advancements of statistical software, led the way to a resurgence in scientific visualization.

Compelling data visuals and their power to relay complex information extend to recent times. In 2006, Swedish health expert Hans Rosling gave an inspiring TED talk about social and economic developments in the world over the past 50 years ("The Best Stats You've Ever Seen," 2006). In the talk, Rosling presented a series of bubble charts showing the relationship between global income and life expectancy across decades. He used statistics and visualizations to debunk myths of the developing world, revealing how world health and living standards are improving each day. In particular, the data demonstrated the tremendous social change in Asia, and how these developing countries were pulling themselves out of poverty—news that was under-

reported and overlooked. Enjoyable animations accompanied his energetic presentation: visualizations that added a sense of excitement to the data. Rosling's TED talk was an incredible and classic demonstration of the power of animated visual communication.

Growing Need for Data Visualization

Epidemiologists have long used visual methods to communicate scientific findings. The need for social scientists to bring data to life via visualizations is also growing. Data visualization is not the icing on the cake but serves to explore data patterns, enhance reader comprehension and memorization, and facilitate trust.

Better Exploratory Data Analysis

Researchers tend to perceive visualization as the end product of analysis—an afterthought. For social scientists, however, visualization could be an immediate exploratory tool that provides initial “clues,” leading to deeper analysis and greater insight. Exploratory data analysis primarily transpired through making charts and other visualizations of a dataset. For instance, when working with continuous variables, histograms help examine if the data exhibits a normal or long-tailed distribution. If the latter is found, researchers would consider taking a logarithmic transformation before the analysis; when working with categorical data, bar charts help identify the most/less frequent category, presenting whether there are abnormalities in the dataset. Exploratory analysis can also use scatterplots to highlight relationships between variables. Overviewing of those associations may help uncover previous “blind spots” and stimulate a fresh scientific perspective. Therefore, visualizations could allow much higher transparency than summarizing results through a descriptive or regression table (Healy and Moody, 2014).

Better Comprehension and Memorization

Traditional data analysis usually presents information in a numerical table, which depends heavily on cognition. In contrast, graphs and diagrams are graphical, making greater perception use. As Few (2009) explains, seeing (i.e., perception)—work of the visual cortex—is fast and efficient, whereas thinking (i.e., cognition)—work of the cerebral cortex—is slower and less efficient. Data visualization shifts the balance between perception and cognition, allowing our eyes to discern patterns to engage and amplify cognition (Few, 2009). Further explained by Tufte in his well-cited book, *The Visual Display of Quantitative Information* (1983), designers can further achieve this cognitive goal by maximizing *data-ink ratio* (reducing information to the most important points), avoiding *chartjunk* (redundant display such as distracting background colors and irrelevant visual decorations), and leveraging labeling and graphical formats that decrease cognitive processing by readers. Studies littered with poor data visualizations can mislead researchers, impede the progress of scientific research, and confound readers. Tufte posed five *graphical integrity* principles for efficient graph design (1983, p. 51). Those rules are:

- *Graphical excellence is the well-designed presentation of interesting data—a matter of substance, statistics, and design.*
- *Graphical excellence consists of complex ideas communicated with clarity, precision, and efficiency.*
- *Graphical excellence is that which gives the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space.*
- *Graphical excellence is nearly always multivariate.*
- *Graphical excellence requires telling the truth about the data.*

Therefore, when data visualization follows graphical integrity principles and comes with thoughtful color, space, and alignment, it can quickly communicate stories associated with scientific research and improve readers' comprehension.

Better Trust

The Artificial Intelligence revolution has been reshaping academic research. AI research tools present several advantages over traditional research methods: They support analyses of large datasets and identify patterns that would be imperceptible to human analysts. However, the wonders of AI research are not without perils. Because of their complexity, the inner workings of algorithms—such as topic modeling and multiclass classification algorithms—often remain obscure. The “black box” effect, due to a lack of explainability and transparency, has fueled distrust and suspicion toward researchers.

Some researchers are hoping data visualization could act as a bridging solution between unraveling the “black box” and could bring forward the behind-the-scenes in an understandable way. Researchers Viégas and Wattenberg made a keynote speech “Visualization: Secret Weapon of Machine Learning” in 2017 to advocate the use of visualizations that expose how those black-box algorithms make decisions. Elijah Meeks, a data visualization engineer at Netflix, also emphasized the growing importance of designing visualization to identify anomalies and generate trust in algorithms. Visualization techniques, such as parallel coordinate plots, scatterplot matrices, and scagnostics, are developed to help understand relationships within high-dimensional datasets. By transforming and integrating data sets into visual representations, researchers can then search visually for patterns or trends in varied data configurations. This offers transparency about how the algorithm comes to its conclusions.

Types of Visualization

Data visualization is useful for multivariate data, numeric data with a broad range, geographic data, as well as texts (Tufte, 2004). Remarkably, social scientists usually involve in interdisciplinary research, which comes with complex and unstructured data. There is no one-size-fits-all approach to create a visualization as every dataset is unique. With different emphases and for different purposes, there could be multiple ways to depict the same dataset.

Structured Data: Uni- and Multivariate

Structured data is data that can be represented as rows and columns. Each row is a single data record, and each column is a specific attribute of the dataset. Continuous, numeric data and discrete, categorical data are common forms of structured data types.

Univariate analysis is the simplest form of visualization where researchers analyze the distribution of one data attribute. The type of variable—whether it is numerical or categorical—will influence the type of chart: Histograms or density plots are suitable for visualizing numeric data and their distributions; boxplots are correct for emphasizing outliers; bar charts or pie charts are appropriate for categorical data attributes—with pie charts being helpful to display shares for a small number of categories (usually less than five).

Multivariate analysis involves at least two attributes at play. Besides distributions, multivariate analysis also concerns potential relationships amongst attributes. One common example of bivariate data visualization is scatter plots, which are frequently used for visualizing bivariate correlations and linear regressions. An aggregated matrix version of scatter plots is usually represented as a heat map. Heat maps use color hue to help visualize the degree of correlation among attributes: darker color means higher correlations while lighter shading

represents lower correlations. Scatterplots and heat maps can quickly unravel key insights, trends, and correlations between either categorical or numerical variables.

Structured Data: Geospatial Data

Geography adds another pivotal dimension to how humans interact with their environments. Social reality is heavily dependent on spatial features. As a popular way of sharing users' statuses, social media content is often geotagged, either as precise coordinates of their posting location or as toponyms of these locations. Those geotagged posts become a valuable proxy for understanding people's mobility and allow researchers to explore their physical presence together with their online activities on a massive scale. For example, geography researchers Tsou and Leitner (2013) were among the first to popularize the emerging field of *cyber geography*, which studies the interconnected spatial patterns and relationships between cyberspace and the real world. Examining Twitter during Hurricane Sandy, Tsou, *et al.* (2013) use a series of cartographic visualizations to highlight the complex nature of sociospatial relations. This innovative method can facilitate the tracking of the dissemination of ideas and social events in cyberspace from a spatial-temporal perspective.

Flow Maps. By placing stroked lines on top of a geographic map, a *flow map* can depict the movement of a quantity in space and in time. Charles Minard's Napoleon's March visualization we reviewed earlier is an example of a flow map. Flow lines typically encode a large amount of multivariate information: path points, direction, line thickness, and color can all present dimensions of information to the viewer. Cartographer Eric Fischer explores virtual communication trends by mapping language communities on Twitter ("Language communities of Twitter," 2011). During a five-month period in 2011, he tracked people who were @replying using geotag data to locate the users on both ends of the conversation, mapping their virtual

communities. For instance, Fischer showed that the United States is heavily connected to various parts of the world, indicating a significant presence of global virtual communities. The powerful imagery shows what is inherent in the data and accurately depicts the world's connectedness in the digital era. However, areas of spatial visualization do not necessarily reflect the relative importance of regions (e.g., Montana has fewer people than New York City, but is much bigger) and that spatial distance is not directly associated with nearness (e.g., countries divided by natural features like mountain ranges). There is substantial literature in geography regarding these display issues, such as using color schemes to show values (i.e., choropleth maps).

Unstructured Data: Network Data

Visualizations are essential to exploring numerical and geospatial data, as well as relational data that are often used for network research. Social network analysts see the social world as structured by a web of connected agents tied together by specific relationships (Wasserman and Faust, 1994). Particularly, nowadays, social media rely heavily on well-defined relationships (e.g., Twitter's following/followers). The ability to demonstrate these relationships using visual network data gives researchers an edge over long-winded explanatory text.

Network research has made extensive use of visualization since psychiatrist Jacob Moreno—the father of network analysis (Burt, *et al.*, 2013)—developed *sociograms* (geometric shapes and lines) to depict friendship patterns among elementary school students and identify children “at-risk” (Figure 4, Moreno, 1934, p. 38). Since then, network visualizations have grown ubiquitous, illustrating every topic from networks of corruption (Chang, 2018) and clusters of political conversations on Twitter (Smith, *et al.*, 2014) to the spread of epidemics (Brockmann and Helbling, 2013).

[INSERT FIGURE 4 ABOUT HERE]

Traditionally, network visualizations were hand-drawn node-link diagrams that served for descriptive purposes, while more advanced analytical results were in verbal or tabular form (Brandes, *et al.*, 2001). Recent work, however, involves a progressive shift to computational software. Graph layout algorithms, such as force-based or tree-based layouts (Bender-deMoll and McFarland 2006), optimize the spatial layout of nodes and edges (e.g., organized nodes with respect to the number of edges they are connected to or with respect to their importance to the network's structure). Graph theory examines key network structural properties, including clustering and connectivity (Correa and Ma 2011). These developments facilitate inductive identification of the underlying structure of narrative data, and reveal the complexities of the links between differently positioned actors in a structure that a personal attribute-based analytical method might overlook (Contandriopoulos, *et al.*, 2018).

Community Detection and Formation. Examples of network visualizations are numerous, but to study the digital public spaces, many researchers have used network visualizations to show aggregate patterns of sharing/retweeting and friending/following to estimate the formation of virtual communities. Communities are of particular importance in social media analysis as they convey the underlying organization and structure of social media users, which often leads to a better understanding of the role groups of users have in the social space, as well as to insights into how information propagates between user groups. For example, the study by Usher and Ng (2020) examined Washington DC political journalists as Communities of Practice to better understand the sense-making and knowledge-producing contexts of the journalists' work. The researchers used an inductive computational approach that combined a social network analysis of journalists' Twitter interactions with a qualitative, thematic analysis of the journalists' work histories, organizational affiliations, and self-

descriptions. Findings showed that journalists' peer-to-peer engagement facilitated a diversity of knowledge-producing communities within political journalism, neglected in previous research. Another study by Weltevrede and Helmond (2012) mapped and analyzed the historical changes in the Dutch blogosphere and networks of connections between blogs by using the Wayback Machine to trace and map transitions in technologies and major platforms and practices in the blogosphere. The authors developed a series of yearly visualizations that show the changing structure of the Dutch blogosphere from different perspectives.

Diffusion of Information or Influence. Network visualizations are also used to examine the diffusion of information or influence in digital public spaces. One of Google's early innovations was analyzing the network structure of the Internet—i.e., determining which pages link to/from other pages—in order to rank web pages by relevance. Network theory algorithms that weigh connections among entities to gauge their importance have proven useful to help navigate millions of pages in document dumps such as WikiLeaks and the Panama Papers. Network analysis and visualization can help make these large sets of data navigable and give researchers and the public a starting point toward understanding connections between parties. For instance, the Carter Center utilizes network analysis to estimate the chains of command and track emerging and shifting alliances in Syria among the government, the opposition, the Kurds and their allies, and ISIS by analyzing social media postings and YouTube videos by approximately 5,600 armed groups (“Syria Conflict Mapping Project,” 2016). Those analyses and visualizations provide mediators and humanitarian responders with up-to-date information on developments throughout Syria.

While there is a long tradition in studying and visualizing static networks (e.g., roads and railway lines, Haggett and Chorley, 1969), social media—and the social networks derived from

them—tend to be much more dynamic. This dynamic nature originates from the rapid creation and change of content, users, and links over time. The diffusion process of rumors and misinformation during a global pandemic on social media is an example showing the dynamic nature of social media. Discerning the anomalous information behaviors on Twitter, Zhao, *et al.* (2014) developed *FluxFlow*, an analytic dashboard with interactive visualizations that visually summarize important information such as keywords, temporal dynamics, and the relationships and connections among threads and authors of anomalous information. In particular, *FluxFlow* introduces an aggregated temporal circle packing design that demonstrates how an original message is disseminated and propagated among people over time. Each circle denotes a user who retweeted the original tweet: the circle's size denotes users' importance as defined by the number of their followers; and the circle's color indicates its anomaly score as computed by the analysis model.

Unstructured Data: Textual Data

Moving beyond the relational aspect of social media, visualizing content derived from social media also poses unique challenges. Thematic and contextual information of social media messages derives a valuable understanding of public opinion and collective action. However, unlike numerical data, textual data is one form of unstructured data: its rich structure, syntax, and semantics are hard to identify and handle. As a case, textual visualization is a solution to improve textual analysis—in terms of speed and clarity—by providing researchers a top-down view of the topics in a corpus and identifying the relationships between topics and other attributes (e.g., political ideologies, gender, etc.). Text visualization is now used in a wide variety of domains, from communicative (Viégas, et al., 2009) to exploratory analysis of topic models (Sukhija, et al., 2016) and single document visualizations.

Word Frequencies. One commonly used method to visualize thematic information is word clouds (or *tag clouds*). Popularized by sites such as del.icio.us and Flickr, word clouds have become widely used tools for web content exploration and navigation (Heimerl, *et al.*, 2014; Viégas, *et al.*, 2009). It visualizes words *that* appear *more* frequently with greater prominence through font size or color (McNaught and Lam, 2010). Despite being subjected to usability critiques, word clouds are frequently used for their ability to effectively summarize large amounts of data and present it qualitatively (Jung, 2015; Wu, *et al.*, 2011).

Word Context. While word clouds summarize keywords in a corpus, they do not explain word choice in context, limiting the degree to which a user can engage with themes or commentary across the documents. To address this, Wattenberg and Viégas (2008) introduced *WordTree* (Figure 5) (<https://www.jasondavies.com/wordtree>), which shows the relationship of phrases in a dataset. A word tree places a tree structure onto the words or phrases that follow a particular word or phrase and then uses that structure to arrange those words or phrases spatially. The tree structure makes it easy to spot repetition in the contextual words that follow a word or phrase. For example, Mitra and Gilbert (2014) examined whether the language used in the crowdfunding site Kickstarter predicted campaigns' successes. Their study found that phrases used in successful campaigns exhibited the general persuasion principle. For instance, the phrase "pledgers will" was often followed by positive words such as "receive," which conveyed that one would receive gifts or other benefits after funding the project. In contrast, the phrase "even a dollar" was often followed by negative words such as "short," "will," and "helps," which might be interpreted as desperation for money and, therefore, less appealing.

[INSERT FIGURE 5 ABOUT HERE]

Case Study of Global Web Use

In this section, I will present a co-authored study (Ng and Taneja, 2019) that illustrates different applications of data visualizations. My aim here is not to report the study results, but to demonstrate the use of visualization for organizing, analyzing, and integrating multidimensional data to study public digital space. Additionally, I will assess their merit, utility, and ability to derive insights from the visualization tools used.

Research Description

The World Wide Web turned 30 years old in 2019, with half the world online. However, it is far from being a global platform with a universal language as early visionaries had imagined. While political and Silicon Valley elites continue to suggest that the Internet's growth is making distances, languages, and geographies somewhat irrelevant, my colleague Harsh Taneja and I (2019) did an empirical reality check against such normatively optimistic prescriptions. Drawing on the literature of media globalization, as well as Internet geographies, we examined how and why countries are (dis)similar in their web use patterns.

Data Types and Forms

We considered nations, rather than individuals, to be their principal units of analysis. We first obtained the ranked lists of the 100 most-visited websites for 174 different countries from Alexa, a web analytics company, in July 2018 and February 2019. Alexa ranks web traffic based on its global panel, which consisted of millions of Internet users who used one of Alexa's toolbar browser extensions (Alexa, 2018).

To determine the extent that online consumption is similar across countries, we computed pairwise similarities between countries using the Rank-Biased Overlap algorithm (Webber, *et al.*,

2010). With those pairwise similarity scores, we constructed a symmetric country-by-country matrix (174 x 174), treating it as a network graph.

Visualization and Analytic Lines

To perform an exploratory analysis, we first plotted a histogram (Figure 6) to examine the distribution of weighted degrees for each country. Caribbean nations had the highest weighted degrees, beginning with Barbados (72.2), followed by Belize (70.76) and Trinidad and Tobago (68.98). The United States ranked tenth (66.97). On the opposite side, Turkmenistan (29.25) was among the lower rank, along with China (13.82), scoring the lowest weighted degree among all 174 countries. No country stood out as exceedingly similar to most others in terms of website usage (Gini coefficient = 0.08).

[INSERT FIGURE 6 ABOUT HERE]

Next, we performed an agglomerative hierarchical cluster analysis on the similarity matrix. The cluster forms through a bottom-up process: each country starts in its own cluster, and pairs of clusters merge as one moves up the hierarchy. The dendrogram (Figure 7) identifies clusters of countries with similar web use patterns.

[INSERT FIGURE 7 ABOUT HERE]

However, as is often the case with cluster analysis, setting a cut-off point to separate cohesive subgroups required qualitative judgment. Thus, we further created 29 choropleth world maps (from 2 to 30 clusters) to interpret the relationship between clusters and spatial patterns. By shading in the choropleth map based on membership, spatial patterns between communities become noticeable. We found that large clusters split into smaller groups of countries of geographically contiguous or linguistically similar regions. For example, when global web use manifested as five clusters, major countries composed the largest cluster from South and

Southeast Asia, the Middle East, Africa, and Western Europe. It also included most Caribbean and Latin American countries (e.g., Mexico and Brazil), as well as the United States. For the 18-cluster solution, this large cluster split into seven smaller groups. Latin American countries remained as a cluster; but the United States, Singapore, and a few Western European countries clustered into one group; and regions of France (i.e., France, Réunion, French Guiana, Guadeloupe, and Martinique) formed their own cluster. Thus, the choropleth world maps (Figure 8) illustrate that global web use manifests as a mosaic of regional cultures, composed of geographically adjacent and linguistically similar countries.

[INSERT FIGURE 8 ABOUT HERE]

The Alexa web traffic data was a two-mode network, with 174 countries and 6,252 unique websites as the two sets of nodes. To evaluate the robustness of our finding and reassure that the country-by-country projection did not result in a loss of valuable structural information, we projected Alexa traffic data to its other network projection: a website-by-website similarity matrix. We conducted cluster analysis on the website-by-website matrix using the Louvain-clustering method, a popular community detection algorithm appropriate for large weighted but undirected networks (Blondel, *et al.*, 2008). This analysis revealed 17 clusters (Figure 9, modularity = 0.256). In general, websites with the same language, especially when their content focused on countries that share a border tended to belong to the same cluster. Therefore, both country-to-country projection and website-by-website projections led to similar inferences.

[INSERT FIGURE 9 ABOUT HERE]

In summary, we created *network graphs* of countries that are connected based on their web use similarities. We analyzed the network properties via *histograms*, and identified each country's weight degree—the higher the score, the more similar a country's news consumption is

to other countries. We further applied hierarchical clustering and used a *dendrogram* to find communities of comparable nations. We employed *choropleth world maps* as a visual solution to interpret the relationship between spatial patterns and web use. Those visualizations were created via software *Gephi* (network), *matplotlib* (histogram), *geopandas* (choropleth map), and *scipy* (dendrogram) libraries of the Python environment, which included a large inventory of visualization approaches. These libraries not only added power and flexibility, but also allowed animation of the visualizations.

Ethical Considerations and Implications

Visualizations have an enormous impact on how data influences decisions across all areas of human endeavor. Visualizations, however, are not immune to prejudice and misrepresentation. All visualizations, not only future-looking models, are sensitive to bias and underlying assumptions during data collection and processing; presentation and design are susceptible to distortion and misinterpretation. Jason Moore of the US Air Force Research Laboratory, as quoted during the 2011 VisWeek Conference, suggested a Hippocratic oath for visualization, which contains the essence of responsible visualization:

“I shall not use visualization to intentionally hide or confuse the truth which it is intended to portray. I will respect the great power visualization has in garnering wisdom and misleading the uninformed. I accept this responsibility willfully and without reservation, and promise to defend this oath against all enemies, both domestic and foreign.” (cited in A Reader on Data Visualization, 2019, The Data Visualization Hippocratic Oath section, para. 3)

Misleading visualizations can affect a message's clarity and damage research efforts and credibility (Pandey, *et al.*, 2015). To prevent this, one should follow specific standards to generate meaningful and accurate visuals. The process breaks down into three steps, each with its own guiding rules.

Data Collection and Storage

The first step is data gathering. Data is not a naturally occurring phenomenon. Instead, data is always collected or processed by someone, for certain aims. Since data is the foundation and pillar of a project, it must be trustworthy and verifiable. Besides collecting data from reliable sources, information designer and journalist Alberto Cairo (2014) also suggests four reminders for information gathering:

1. Beware of selection bias while using an existing dataset or creating a new one.
2. False or irrelevant information does not improve anyone's decision-making capacity.
3. Even if the information is both accurate and relevant, moral pitfalls may remain.
4. To avoid the unethical trap of inscrutable or misleading graphics, take an evidence-based approach when possible. The purpose of the graphic dictates the form it takes; aesthetic preferences should never override clarity.

Traditional ethical principles—such as consent, anonymity, and avoiding undue harm—should always be applied to social media research (Beninger, *et al.*, 2014). Specifically, for the issue of anonymity, one might consider removing the name and sensitive information of participants would be enough to protect individuals' rights. However, even after deleting all identifying information, random bits of social media data that alone seem anonymous can often be pieced together, possibly exposing clues to subjects' identity (Zimmer, 2008). For example, the fact that a dataset includes each subject's gender, hometown state, and a social media post can be far

enough to identify the individual. Visualization may make these issues more prominent as network graphs disclose nodes' names. Therefore, researchers must take extra care to further anonymize before dissemination.

Representation of Visualization

“A poor chart is worse than no chart at all” (Wallgren, *et al.*, 1996, p. 98). Without consideration of how visualizations will be interpreted (or possibly misinterpreted), researchers run the risk of confusing audiences rather than enhancing their understanding. Graphical excellence requires telling the truth about the data (Tufte, 2004, p. 51). In a series of experiments performed by the Center for Human Rights and Global Justice, the empirical analysis shows how common distortion techniques can affect the way information in the graph is perceived and how it potentially could mislead viewers (Emerson, *et al.*, 2018). Distortion techniques include improper extraction, tactical omission of data, using a truncated y-axis (starting at a number greater than zero when illustrating percentages), and using area to represent quantity (such as comparing areas of circles) (Cleveland and McGill, 1984). Transparency is thus essential, not only as a pre-condition for scientific rigor and replicability but also to increase the participatory potential of data visualizations.

Readability of Visualization

Researchers should think carefully about the technical and substantive choices underlying graphical representation and their readability for non-specialist audiences. What content is to be displayed and how? Are dynamic formats preferable to static ones? What should labels show? If readers, especially laypersons, are not aware of the basic principles underpinning these choices, they will have limited capacity to appraise visualizations critically. Therefore, researchers should use labels, reference lines, and annotations wisely to increase the readability of

visualizations. Researchers should be mindful of making certain design choices, such as consistent use of color answering individual questions rather than attempting to serve all needs.

Interactivity elements are also suitable for analyzing high-dimensional data (Weber and Hauser, 2014). Interesting applications of interactive visualizations abound in the literature. For example, sociologists Abramson and Dohan (2015) illustrate the use of an ethnoarray—loosely adapted from the graphical heatmap approach—for analyzing, representing, and sharing ethnographic data. However, if the graphical representation is confusing to readers, researchers should use an analogy or connect the implications to the person’s value system. It is ideal to seek input and feedback from other experts and laypersons and iterate over time. More generally, public education about data—particularly how to interpret data visualizations—is likely to become a pressing need if the use of visualization in digital space (and other) research is to bloom.

Conclusion

This article outlines various visual methods that can be utilized to make sense of the numeral, relational, spatial, and textual patterns in datasets related to digital public spaces. It highlights the historical and current state of the art along with some future directions with a discussion on the accompanying challenges and pitfalls. However, limitations remain, and this article has no pretension to review exhaustively what can and cannot be done with data visualizations. The examples in this article may create bias as they may be seen as US-centric, but the overarching purpose of this article is to understand some effective strategies for visualizing data, especially when dealing with various data types.

Charts and graphs are powerful, and they appeal to our natural visual processing power. When we take a more holistic approach to quantitative research, the ability to comprehend and construct charts and graphs critically is pivotal. It seems more timely than ever due to the COVID-19 pandemic. These images of the pandemic produce a social imaginary expressed as curves, distributions, and maps. The global crisis has forced our society to rethink the value of data visualization in convincing people of a drastic shift in behavior. We have become very disciplined in a very short time, partly through data visualization. Besides its educational role, data visualizations became indispensable tools for governments to take the right decisions at the right time. They helped to flatten the curve and saved lives while limiting economic damage.

Visualizations should be simple and easy to understand, but at the same time, it is critical to consider the ways to make sure data visualizations are “responsible artifacts.” Researchers must practice ethical procedures throughout the steps of visualization. Collaboration, iteration, and feedback are important steps of the visualizing process of data related to digital public spaces at any time, but particularly when visualizing sensitive data in the midst of a global crisis. I hope the work will spark further conversations around visualizations and encourage researchers to leverage these snippets for visualizing their own datasets in the future.

About the author

Yee Man Margaret Ng (Ph.D., University of Texas) is an Assistant Professor in the Department of Journalism and Department of Computer Science (faculty affiliate) at The University of Illinois Urbana–Champaign. Her research interests include computational social science, journalism, and communication technology.

References

- Kelsey Beninger, Alexandra Fry, Natalie Jago, Hayley Lepps, Laura Nass, and Hannah Silvester, 2014. "Research using social media; users' views." *NatCen Social Research*, pp. 1–40.
- Ulrik Brandes, Jörg Raab, and Dorothea Wagner, 2001. "Exploratory network visualization: Simultaneous display of actor status and connections." *Journal of Social Structure*, volume 2, number 4, pp. 1–28.
- Dirk Brockmann and Dirk Helbing, 2013. "The hidden geometry of complex, network-driven contagion phenomena." *Science*, volume 342, number 6164, pp. 1337–1342.
- Ronald Burt., Martin Kilduff, and Stefano Tasselli, 2013. "Social network analysis: Foundations and frontiers on advantage." *Annual Review of Psychology*, volume 64, number 1, pp. 527–547.
- Russell Carney and Joel Levin, 2002. "Pictorial illustrations still improve students' learning from text." *Educational Psychology Review*, volume 14, number 1, pp. 5–26.
- The Carter Center, 2020. "Syria conflict mapping project reports," at https://www.cartercenter.org/peace/conflict_resolution/syria-conflict-resolution.html#reports, accessed 15 January 2022.
- Zheng Chang, 2018. "Understanding the corruption networks revealed in the current Chinese anti-corruption campaign: A social network approach." *Journal of Contemporary China*, volume 27, number 113, pp. 735–747.
- William Cleveland, 1993. *Visualizing data*. AT&T Bell Laboratories: Hobart Press.

William Cleveland, 1994. *The elements of graphing data*. AT&T Bell Laboratories: Hobart Press.

William Cleveland and Robert McGill. “Graphical perception: Theory, experimentation, and application to the development of graphical methods.” *Journal of the American Statistical Association*, volume 79, number 387, pp. 531–554.

Damien Contandriopoulos, Catherine Larouche, Mylaine Breton, and Astrid Brousselle, 2018. “A sociogram is worth a thousand words: Proposing a method for the visual analysis of narrative data.” *Qualitative Research*, volume 18, number 1, pp. 70–87.

John Emerson, Margaret L. Satterthwaite, and Anshul Vikram Pandey, 2018. “The challenging power of data visualization for human rights advocacy.” In M. K. Land & J. D. Aronson (editors.). *New technologies for human rights law and practice*. Cambridge University Press, pp. 162–187.

Eric Fischer, 2011. “*Language communities of Twitter*.” At <https://flowingdata.com/2011/10/27/language-communities-of-twitter/>

Kara Gavin, 2020. “Flattening the curve for COVID-19: What does it mean and how can you help?,” at <https://healthblog.uofmhealth.org/wellness-prevention/curve-fattening-not-flattening-what-can-we-do>, accessed 26 December 2021.

Kieran Healy and James Moody, 2014. “Data visualization in sociology.” *Annual Review of Sociology*, volume 40, number 1, pp. 105–128.

- Florian Heimerl, Steffen Lohmann, Simon Lange, and Thomas Ertl, 2014. “Word cloud explorer: Text analytics based on word clouds.” In *2014 47th Hawaii International Conference on System Sciences*, pp. 1833–1842.
- Jin-Kyu Jung, 2015. “Code clouds: Qualitative geovisualization of geotweets.” *The Canadian Geographer*, volume 59, number 1, pp. 52–68.
- Johns Hopkins University, 2020. “COVID-19 Dashboard” at <https://coronavirus.jhu.edu/map.html>, accessed 15 January 2022.
- Carmel McNaught and Paul Lam, 2010. “Using Wordle as a supplementary research tool.” *Qualitative Report*, volume 15, number 3, pp. 630–643.
- Elijah Meeks, 2018. “What is a senior data visualization engineer?” at https://medium.com/@Elijah_Meeks/what-is-a-senior-data-visualization-engineer-eb032996297, accessed 26 December 2021.
- Friendly Michael, 2006. “A brief history of data visualization.” In C. Chen, W. Hardle and A. Unwin (editors.). *Handbook of computational statistics: Data visualization*. pp. 16–48. Heidelberg, Germany: Springer-Verlag.
- Vincent Miller, 2008. “New media, networking and phatic Culture.” *Convergence: The International Journal of Research into New Media Technologies*, volume 14, number 4, pp. 387–400.

Tanushree Mitra and Eric Gilbert, 2014. “The language that gets people to give: Phrases that predict success on kickstarter.” *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing*, pp. 49–61.

Jacob Levy Moreno, 1934. *Who Shall Survive? A new approach to the problem of human interrelations*. Washington, D. C.: Nervous and Mental Disease Publishing Co.

Fabian Neuhaus and Timothy Webmoor, 2012. “Agile ethics for massified research and visualization.” *Information, Communication and Society*, volume 15, number 1, pp. 43–65.

Anshul Vikram Pandey, Katharina Rall, Margaret L. Satterthwaite, Oded Nov, and Enrico Bertini, 2015. “How deceptive are deceptive visualizations? An empirical analysis of common distortion techniques.” *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 1469–1478.

Responsible Data, “Responsible Data Visualization New York,” January 15 2016 at <https://responsibledata.io/event/responsible-data-visualization/>, accessed 15 January 2022.

Hans Rosling, 2006. “*The best stats you’ve ever seen.*” at https://www.ted.com/talks/hans_rosling_the_best_stats_you_ve_ever_seen

Hans Rosling, 2007. “*The seemingly impossible is possible.*” at <http://www.gapminder.org/videos/hans-rosling-ted-talk-2007-seemingly-impossible-is-possible/>, accessed 15 January 2022.

Michael Schermann, 2019. "A Reader on Data Visualization." at https://mschermann.github.io/data_viz_reader/ethics.html#the-data-visualization-hippocratic-oath, accessed 15 January 2022.

Ian Spence, 2006. "William Playfair and the psychology of graphs." *Proceedings of the American Statistical Association Section on Statistical Graphics*, pp. 2426–2436.

Nitin Sukhija, Mahidhar Tatineni, Nicole Brown, Mark Van Moer, Paul Rodriguez, and Spencer Callicott, 2016. "Topic modeling and visualization for big data in social sciences." *International IEEE Conferences on Ubiquitous Intelligence and Computing*, pp. 1198–1205.

Harsh Taneja and James G. Webster, 2016. "How do global audiences take shape? The role of institutions and culture in patterns of web use." *Journal of Communication*, volume 66, number 1, pp. 161–182.

Ming-Hsiang Tsou and Michael Leitner, 2013. "Visualization of social media: seeing a mirage or a message?" *Cartography and Geographic Information Science*, volume 40, number 2, pp. 55–60.

Ming-Hsiang Tsou, Jiue-An Yang, Daniel Lusher, Su Han, Brian Spitzberg, Jean Mark Gawron, Dipak Gupta, and Li An, 2013. "Mapping social activities and concepts with social media (Twitter) and web search engines (Yahoo and Bing): a case study in 2012 US Presidential Election." *Cartography and Geographic Information Science*, volume 40, number 4, pp. 337–348.

Edward Tufte, 2001. *The visual display of quantitative information*. Cheshire, CT: Graphic Press.

Fernanda Viegas and Martin Wattenberg, 2017. “Visualization: The secret weapon of machine learning.” *Keynote at the Visualization in Data Science Symposium at EuroVis 2017*. Barcelona, Spain at <https://www.youtube.com/watch?v=E70lG9-HGEM>, accessed 15 January 2022.

Fernanda Viégas, Martin Wattenberg, and Kushal Dave, 2004. “Studying cooperation and conflict between authors with history flow visualizations.” *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 575–582.

Fernanda Viegas, Martin Wattenberg, and Jonathan Feinberg, 2009. “Participatory visualization with wordle.” *IEEE Transactions on Visualization and Computer Graphics*, volume 15, number 6, pp. 1137–1144. <https://doi.org/10.1109/TVCG.2009.171>

Anders Wallgren, Britt Wallgren, Rolf Persson, Ulf Jorner, and Jan-Aage Haaland, 1996. *Graphing statistics & data: Creating better charts*. London, UK: Sage Publications, Inc.

Martin Wattenberg and Fernanda B. Viégas, 2008. “The word tree, an interactive visual concordance.” *IEEE Transactions on Visualization and Computer Graphics*, volume 14, number 6, pp. 12217–1228.

Esther Weltevrede and Anne Helmond, 2012. “Where do bloggers blog? Platform transitions within the historical Dutch blogosphere.” *First Monday*, volume 17, number 2, pp.1–34.

Leland Wilkinson, 1999. *The Grammar of Graphics*. New York, NY: Springer.

Angela Xiao Wu and Harsh Taneja, 2016. “Reimagining internet geographies: A user-centric ethnological mapping of the world wide web.” *Journal of Computer-Mediated Communication*, volume 21, number 3, pp. 230–246.

Yingcai Wu, Thomas Provan, Furu Wei, Shixia Liu, and Kwan-Liu Ma, 2011. “Semantic-preserving word clouds by seam carving.” *Computer Graphics Forum*, volume 30, number 3, pp. 741–750.

Jian Zhao, Nan Cao, Zhen Wen, Yale Song, Yu-Ru Lin, and Christopher Collins, 2014. “#FluxFlow: Visual analysis of anomalous information spreading on social media.” *IEEE Transactions on Visualization and Computer Graphics*, volume 20, number 12, pp. 1773–1782.

Michael Zimmer, 2008. “On the “anonymity” of the Facebook dataset.” at <http://www.michaelzimmer.org/2008/09/30/on-the-anonymity-of-the-facebook-dataset/>, accessed 15 January 2022.

“Napoleon’s March” (1812), Charles Joseph Minard. Wikimedia Commons [Public Domain]. https://en.wikipedia.org/wiki/Charles_Joseph_Minard#/media/File:Minard.png

“Snow Cholera map” (1854), John Snow. Wikimedia Commons [Public Domain].
<https://en.wikipedia.org/wiki/File:Snow-cholera-map-1.jpg>

Figure 3

“Diagram of the causes of mortality in the army in the East” (1858), Florence Nightingale.
 Wikimedia Commons [Public Domain]. <https://commons.wikimedia.org/wiki/File:Nightingale-mortality.jpg>

Figure 4

“Jacob Moreno’s Sociogram of a fourth-grade class (1934).” Redesigned by Grandjean, Martin (2015). *Wikimedia Commons [Public Domain]*.

https://commons.wikimedia.org/wiki/File:Moreno_Sociogram_4th_Grade.png

Figure 5

“Word Tree showing how languages used in crowdfunding site Kickstarter predict the success of campaigns,” (Mitra and Gilbert, 2014, p.55). Redesigned for this article.

Figure 6

*“Distribution of weight degree of 174 countries’ web use similarity,” (Ng and Taneja, 2019).
Redesigned for this article.*

Figure 7

“The dendrogram of global web use similarity.” (Ng and Taneja, 2019). Redesigned for this article.

Figure 8

“Choropleth world maps: Web use similarity in 5 clusters and 19 clusters,” (Ng and Taneja, 2019). Redesigned for this article.

Figure 9

“Bipartite graph of countries and websites,” (Ng and Taneja, 2019). Redesigned for this article.