Carnegie Mellon University
Department of Statistics and Data Science

Visualization Analysis

How can we read data visualizations?

Jacky Lao

April 10, 2019

Contents

List of Figures

1 Introduction

This paper presents a question-based framework for analyzing the quality of a visualization in a general manner. Visualizations are tools that can convey complex ideas and encapsulate large amounts of data in a relatively small format. These representations can present coherent narratives and lead readers to conclusions supported by data. At the same time, the intent of the designer of a visualization will affect the form that the visualization takes, and in doing so, may lead readers to specific conclusions that the designer desires, regardless of what the data actually says. Propaganda, for example, has been used to great effect in convincing the general public of various countries of different narratives. For this reason, I consider the designers' intent in addition to the visual elements of a graphic, to form a more holistic picture of what the graphic is saying.

The designer communicates ideas through the medium of a graphic to various readers. In the process of creating the graphic, the designer picks and chooses what parts of the data to show, and what visual elements to display them. These actions are based heavily on a designer's expertise, intent, and background knowledge, which comprise a knowledge base of understanding. Similarly, readers come to view visualizations with their own particular background knowledge, with various degrees of understanding of what visual elements mean, and of the conventions and usages of these elements. This communication process has two main weak points, creating what are known as pre-visual and post-visual errors.[1] Pre-visual errors are committed by designers during the process of creating a graphic, by misrepresentation of the data. Post-visual errors are committed by readers when the figure introduces properties or relationships of the data that are not actually relevant. These types of errors are addressed in various case studies within this paper, primarily in the context of problem solving, where committing these errors may compromise a solution.

In addition to displaying large amounts of data, visual representations have been used in problem solving processes. Drawing a diagram is often a good strategy, as diagrams can orient readers in the state of knowledge of the problem, and lead to new lines of thought. The use of visualizations in problem solving involve four actions: inferring additional consequences, elaborating on new information, stating a new goal, and monitoring the process.[2]

This paper is comprised of three parts. First, I present a series of case studies of visualizations, specifically ones with some intention to inform and depict data for the benefit of the reader of a

visualization. For each data graphic, an analysis of the visual elements are performed, and the relevance of components are discussed in relation to the overall message and author intent. Second, a series of problems are presented, along with diagrams drawn to aid in solving the problem. Last, I discuss the question-based framework that has been developed from the considerations of the case studies.

2 Data Graphics

2.1 A Pie Chart

Figure 1: 3D Pie Chart Infographic

I begin a discussion about the quality of visualizations by first looking at images of questionable quality as examples of "what not to do". This picture was retrieved from a digital marketing agency, and the title, placed in the bottom of the image, is "Top 6 Multi-platform Social Network-

ing Websites".[3, 4] The content of the image appears to be a 3D pie chart, with 6 logos from the top 6 social networking sites connected to each of the 6 slices in the pie chart. Each label has the logo of the website, and a percentage label. Visually the image is well-designed, with six distinctive colors to represent six social networking sites, and a light blue background. There is no excessive verbiage and it's clear which company is associated with which pie slice.

Before getting into details of the visualization itself, my first inclination is to understand what the data that is being depicted is. The dataset depicted in this graphic is quite literally a total of 6 percentages, printed by each company name. The visualization presents absolutely no further information. We have no source for where, when, how the data was collected, or even what the percentage represents. This visualization commits what I consider to be the primary failure of any data visualization - it doesn't present tidy, quality data to begin with, but fluffs it up in what Edward Tufte would consider "chartjunk".[5] Without a known data source, a skeptical reader has no reason to believe what's being shown in the visualization to begin with. Furthermore, the percentages don't even sum to 100%, so there's a missing 20.9% of the total that isn't being displayed.

A key component of a visualization is the encoding of information along dimensions of the chart. This allows comparisons between data points without explicitly writing out the data points. A 3D pie chart has 2 spatial dimensions that information could be encoded in. The height of a pie slice and the angle of the slice can be used to encoded different covariates, but in the case of this visualization, it doesn't appear any information is actually encoded in either of these dimensions. Initially the slices look like they could be encoding the percentages, but the angle of the slice for Reddit is certainly not a quarter of the angle of the slice for YouTube, or a ninth of the slice for Facebook. Similarly, the difference in height of the slices between Facebook and YouTube looks exactly the same as the difference in height between Reddit and Twitter, but the difference in percentage according to the numbers is 22.1% compared to 0.3%. This is further exacerbated by the fact that our perception of height and the area of the slices is distorted by the 3D perspective of the image. A pie chart generally is used to display percentages of a whole, but since the data depicted doesn't sum to 100%, the angle and size of the slices are doubly misleading. Ultimately, none of the spatial dimensions are used to encode information whatsoever, so the entire pie chart of the visualization is essentially fluff.

The text within a visualization is meant to guide a reader to understanding the content and message of the graphic. First, the title of this visualization is located at the bottom, making it

somewhat unobtrusive. The title "Top 6 Multi-Platform Social Networking Websites" is clear and readable in a sans-serif font, but also has a shadow cast over part of it, which is mildly distracting and serves to redirect attention back to the pie chart graphic. This title tells only part of the story - "what" is being displayed is the social networking sites, but details of the data, such as what the percentages represent, is absent. The labels for each of the slices are the names of the social networking sites, and are in the same consistent font, although colored according to the color scheme of the slices. This presents a slight dissonance, as the color of the text doesn't match the color scheme of the companies, and the yellow color for text makes Reddit clash with the background and difficult to read. The percentages are printed clearly by each company, in a light gray, which serves to make them fade into the background.

While visually appealing, there is no substance to the graphic. The visual elements of the image aren't actually being use to display the data, and it's almost as if the intention is to hide the data. The 3D pie chart in the center is pretty, but no element of the chart itself serves a purpose in visually displaying the percentages, and the fact that it casts a shadow over the title draws attention away from actual informative elements, such as the names of the social networking sites. The base of the 3D pie chart has a nice grey arc with 6 colors in it, but serves no purpose other than looking good. The center colored portion of the slices also serve no purpose related to the data, since the radius of these colored slices are all the same. Putting the logos next to the company names is a good visual cue, but its usefulness is dependent on whether the intended audience will recognize them.

Given that the dataset being presented is so small, the visualization is best reformulated as a table. I also add an "Other" row to clarify the missing percentage points. However, issues of data quality can't be rectified by revising the graphic - those need to be dealt with before a good graphic can be made. Even with this table, the definition of the percentage being displayed isn't clear. I've removed most of the visual elements from the graphic, essentially leaving behind the pure data, primarily because the data are so sparse in and of itself. I could also plot this in a bar chart, which would encode the percentage on the y-axis, and accurately portray the relative differences in the percentages by height of the bars.

2.2 NYT Fast Food and Wealth

This figure is from the New York Times' "What's Going On In This Graph?" column, which posts a new curated graphic each month and brings in a statistician from the American Statistical

Top 6 Multi-Platform Social Networking Websites	
Facebook	43.9%
YouTube	21.8%
Reddit	5.3%
Twitter	5.0%
Pinterest	1.6%
LinkedIn	1.5%
Other	20.9%

Figure 2: Table Revision of the 3D Pie Chart

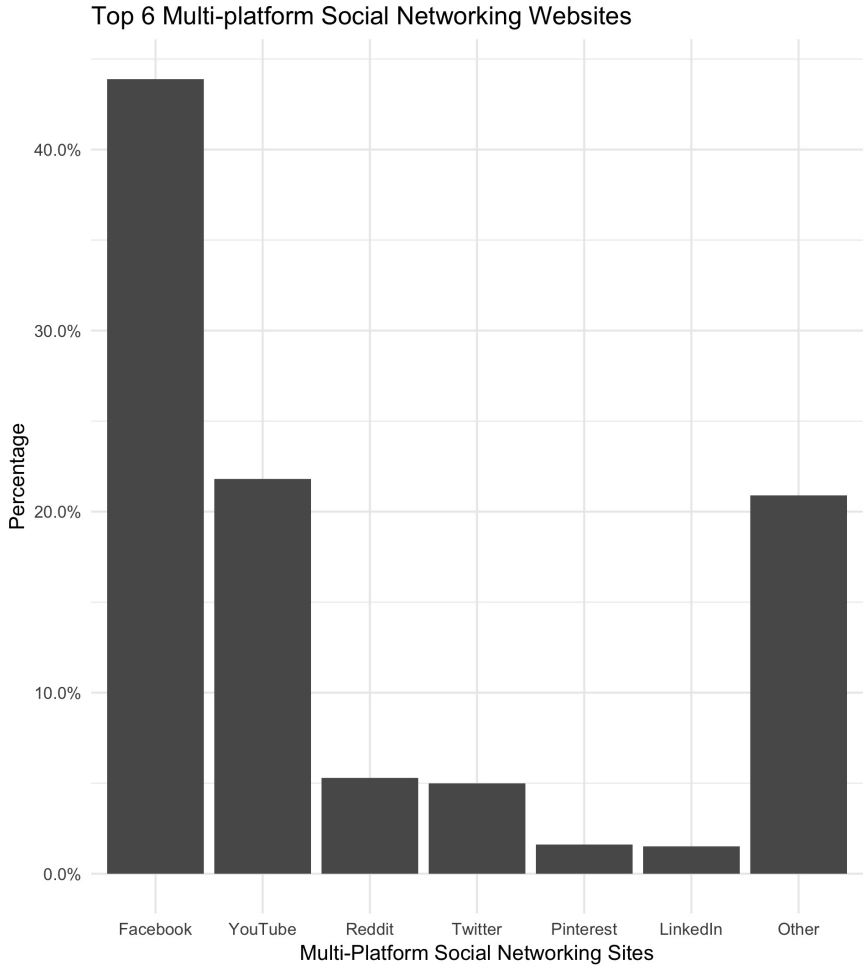


Figure 3: Bar Graph Revision of the 3D Pie Chart Infographic

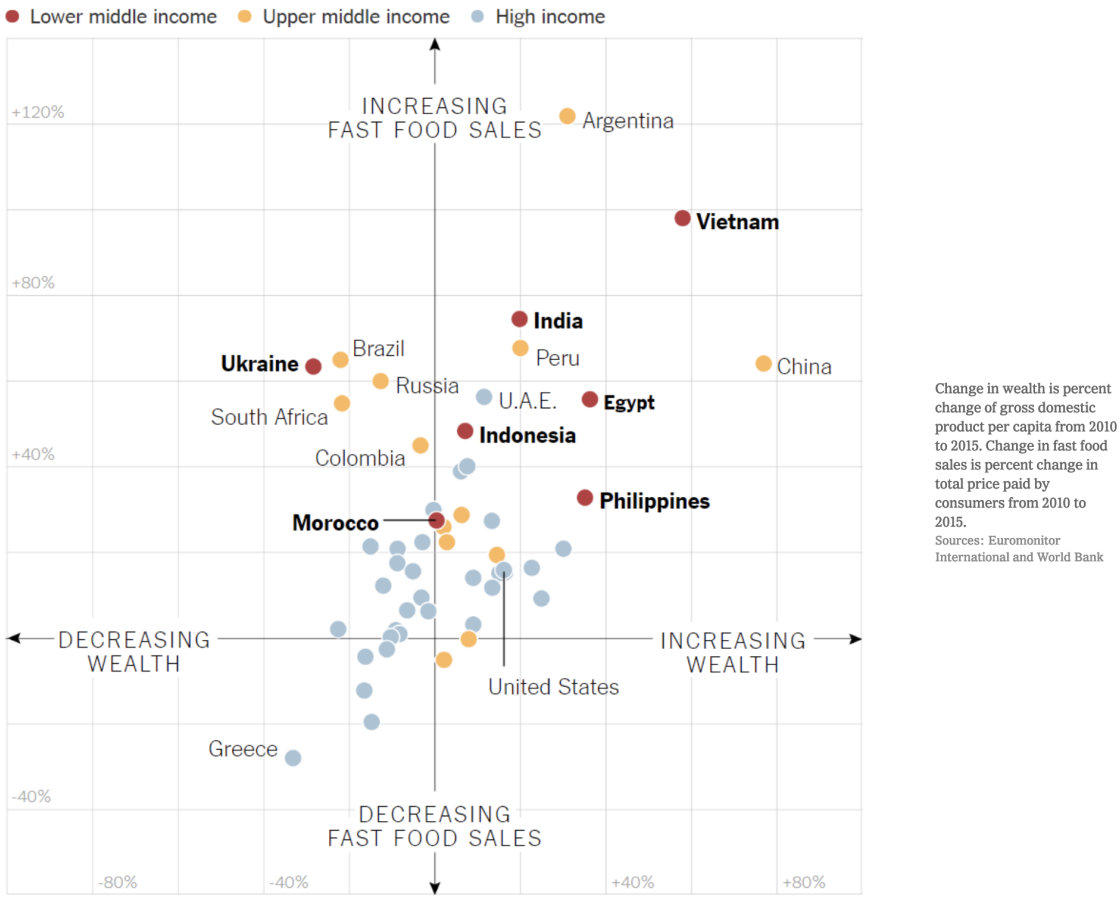


Figure 4: Change in Country Wealth against Change in Fast Food Sales

Association to moderate a discussion.[6] The image is a scatter plot, with 3 different colors for the data. Interestingly, instead of tickmarks on the x and y axes to mark the scale of the variables, the scale is written on the grey grid lines on the left and bottom of the plot. No title is provided within graphic itself. A caption on the right presents information about the graphic, explaining the variables being compared, the time range of the data, and the units of the variables. A handful of the data points in the plot are labeled with the specific country names, and the labels for red data points are bolded. The grey grid lines create a rather aesthetically pleasing 10x10 grid to frame the graphic, with the origin of the x and y axes placed 3 rows from the bottom.

The source provided by the caption on the right is Euromonitor International and the World Bank (assuming that the caption is correct). A quick search shows that Euromonitor International is a market research agency that runs global surveys, socioeconomic research, and industry research.[7] The World Bank is an international financial institution that provides loans to coun-

tries for projects.[8] Presumably the World Bank contracted out the data collection process to Euromonitor International. The question to consider now is whether the data is trustworthy. It's quite likely that the data collection procedure and actual dataset is unavailable to the public and can't be vetted. There are two potential sources of bias, the first being biases stemming from the World Bank that have been transmitted to Euromonitor International (such as through oversight or interference in Euromonitor's procedure in data collection), and the second being biases from Euromonitor International itself (such as through choice of data collection procedure). However, both organizations have a large presence and good reputations, so it seems likely that the dataset is of good quality.

Each data point in a scatter plot encodes information from its position in space relative to the axes. The distance from the x-axis and the distance from the y-axis can encode different covariates, and the distance from a particular data point to another also a visual display of relative differences in the covariates. From the caption, the graphic encodes the percent change of gross domestic product per capita from 2010 to 2015 on the x-axis, and the percent change in total price paid by consumers for fast food from 2010 to 2015 on the y-axis. In addition, color of the data points is used to encode "Lower middle", "Upper middle", and "High" income, as seen in the legend at the top of the graphic. However, it is not explained what this means in the context of a countries. Based on the text, each data point is an individual country, and the baseline for what constitutes a "Lower middle income" country is unspecified. From the data points themselves, a total of 7 countries fall in the "Lower middle income" category, 12 in the "Upper middle income" category, and approximately 32 in the "High income" category. In general, the use of the spatial dimensions is quite good, but I would prefer having smaller dots to avoid the overlap in clusters of data points, such as around the United States. In addition, I would add some specification, either in the figure's caption or in the color legend on the top, of what thresholds define the categories.

The text within the graphic is very well done, with bolding and fonts used for emphasis and with good effect. The placement of the all-caps "INCREASING" and "DECREASING" labels are substituted for the tickmarks you'd find on a standard scatter plot, and provide a capsule summary of what the axes encode and their directionality. This qualitative measure gives a new reader a quick way to understand what's being plotted, but is also misleading, in combination with the fact that the variables here are percentage changes from 2010 to 2015. Increasing and decreasing wealth could be operationalized in a different way from how a reader would expect, such as in absolute terms rather than the percentage change. A reader that considers the graphic alone may

wonder why the United States is to the left of the Phillipines, and also how the color encoding for "Income" is connected to "Increasing Wealth". In addition, "Increasing Fast Food Sales" could be interpreted differently from the actual variable encoded based on the caption, which states that the y-axis is actually the change in total price paid by consumers. What is meant by change in total price paid by consumers is ambiguous - it could be interpreted as change in the amount consumers spent on fast food from 2010 to 2015, or a change in the price of some standard fast food item (like the Big Mac index). In either case, a correlation with increased fast food sales numbers isn't clear, since increased amount spent on fast food could occur with the same number of sales if the price of fast food increased, and we would expect that an increase in fast food prices to have the effect of decreasing fast food sales. In addition to being ambiguous, the operationalization of the covariate as a change in percentage price hides factors such as the influence of inflation between 2010 and 2015 in the various countries, and whether currency exchange rates are considered. The data may appear different if all of the country prices were normalized to 2010 US dollars versus kept as percentage changes within each country's respective currencies.

A number of the data points are labeled with the country they represent, with lines connecting labels to points when appropriate due to clustering of the data. Notably, all of the "Lower middle income" countries have bolded text, directing reader attention specifically to them. In addition, the labels for the non "Lower middle income" countries are specifically for the data points at the entremes of their category - Greece for being so far to the bottom left, UAE for having the highest percentage change in fast food prices for "High income" countries. The remaining labels for "Upper middle income" countries are ones that are further out from the cluster near the origin of the plot. Overall the text does a good job of drawing attention to particular data points with text, and giving enough context for understanding what is being shown. However, the simplifying labels on the x and y axes sacrifice precision to aid in reading the graphic quickly. I would have preferred they write precisely what the covariate actually is than a simplified "Increasing Wealth" label.

The chart is quite well designed, with no large items that I would consider non-essential to understanding the plot. The grid lines are greyed out and visually placed in the background of the chart along with the tickmarks that show the actual quantitative scale of percentage change. The colors used to encode the income category stand out, and present something of an ordinal visual scale, which is appropriate when the covariate encoded is lower, middle, and high income. Bolding is used to draw attention to the lower income data points, and the caption is placed on

the right side of the graphic in a readable fashion. While this graphic doesn't have the same immediate visual "pop" as the 3D pie chart, no visual element stands out as unnecessary, which I consider more important.

The intent of the authors and the conclusions readers are expected to draw are unknown. Based on the visual cues that I have outlined previously, the graphic shows that lower middle income countries tend to have higher percentage changes in "total price paid by consumers" from 2010 to 2015, relative to other countries, and about half of the upper middle income countries had percentage change in fast food spending above 40%. It's difficult to say whether there is a distinct relationship between fast food spending and change in GDP per capita, but few countries had increased GDP per capita with a decrease in fast food spending. The variability in the percentage changes of both variables also appears to increase with decreased income of the countries. The majority of the high income countries are clustered around the origin, between -20% to 20% changes GDP per capita and -20% to 30% changes in fast food spending, compared to the range of -30% to 60% changees in GDP per capita and 20% to 100% changes in fast food spending for lower middle income countries. There is a small drift ot the upper right when looking at countries from high income to lower middle income. I believe that the use of percentages on both of the x and y axes is slightly misleading since percentage increases all depend on the base value (GDP per capita, fast food sale price in 2010), which will be different for each country. This may be the explanation for the clustering and visual drift of countries. However, it's very likely that the scale of the changes in GDP per capita and fast food spending will be orders of magnitude different in absolute terms, so the designers may have chosen to use the percentage change measure rather than a logarithmic transformation for readability.

The only changes I would make to the graph would be textual and for clarity. "Total Price paid by Consumers" needs to be explained, and "FAST FOOD SALES" and "WEALTH" labels on the axes need to be more informative. In addition, the caption on the left should explain the differences between the color-coded categories of "Income," including the thresholds between the three.

2.3 NYT Industry Oligarchies

Figure 5 also comes from the New York Times' "What's Going On In This Graph?" column. The image could be considered a form of bar graph or a form of line graph, but certainly doesn't fit squarely in either category. Each arrow shows a change in market share from the "Early 2000s" to

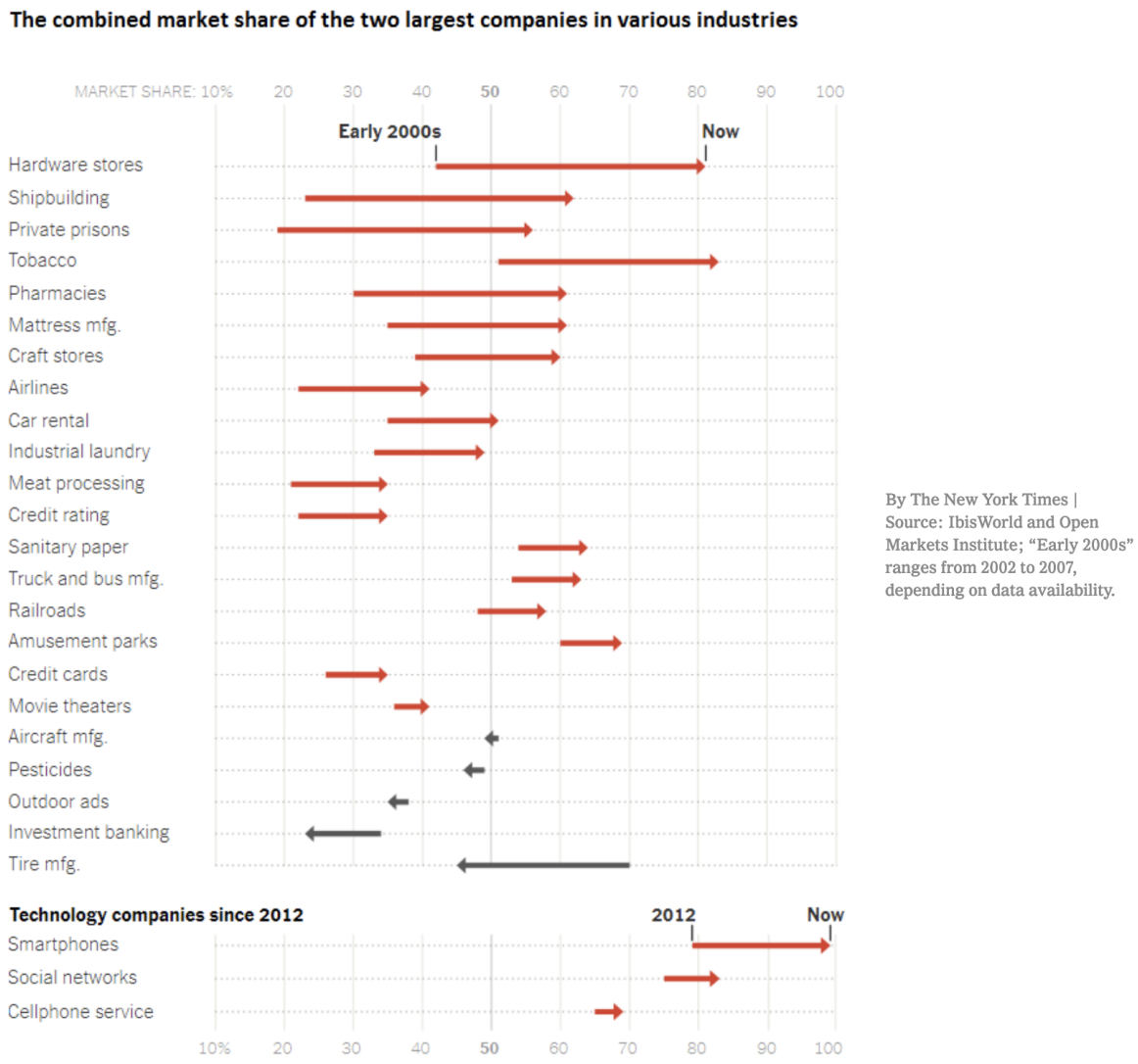


Figure 5: "Dominance of Corporate Behemoths"

2018, the year this graph was published. The caption on the right explains the definition of "Early 2000s" ranges from 2002 to 2007 based data availability, and the title within the graphic states that the data is based on the combined market share of the two largeset companies in these industries. The x-axis label on the top displays the variable encoded: market share, ranging from 10% to 100% by the tickmarks, while the y-axis is a list of various industries. A portion of the chart is sectioned off, denoting specifically technology companies, with market share data ranging from 2012 to "Now", 2018. A slightly darker grey line is used to denote the 50% market share line for each industry. Red and gray arrows are used to distinguish industries in which companies increased

their market share over time from those that decreased over time. The industries themselves appear to be ordered by the magnitude of the difference in market share, from largest change at the top to largest decrease at the bottom.

Some of the data presented in this plot can be found at the Open Market Institute's website, at "https://concentrationcrisis.openmarketsinstitute.org." Open Markets Institute purchased "extensive, up-to-date industry intelligence from IBISWorld, a team of analysts who collect economic and market data, with the intention of releasing the information regarding industry concentration to the public." [9] A number of the industries presented here are not found at that website. With an organization with the express intent to raise public awareness of the concentration of market share in a few companies, it is reasonable to expect that the graphic will best express the intent of the authors. The industries picked here are by no means all of the industries with data available, and it's likely that they were selected to prove a point. It would be interesting to see whether the two largest companies would remain the two largest companies over the time span of the dataset, or if one company holds a large market share in multiple industries, and whether this would support the organization's intent.

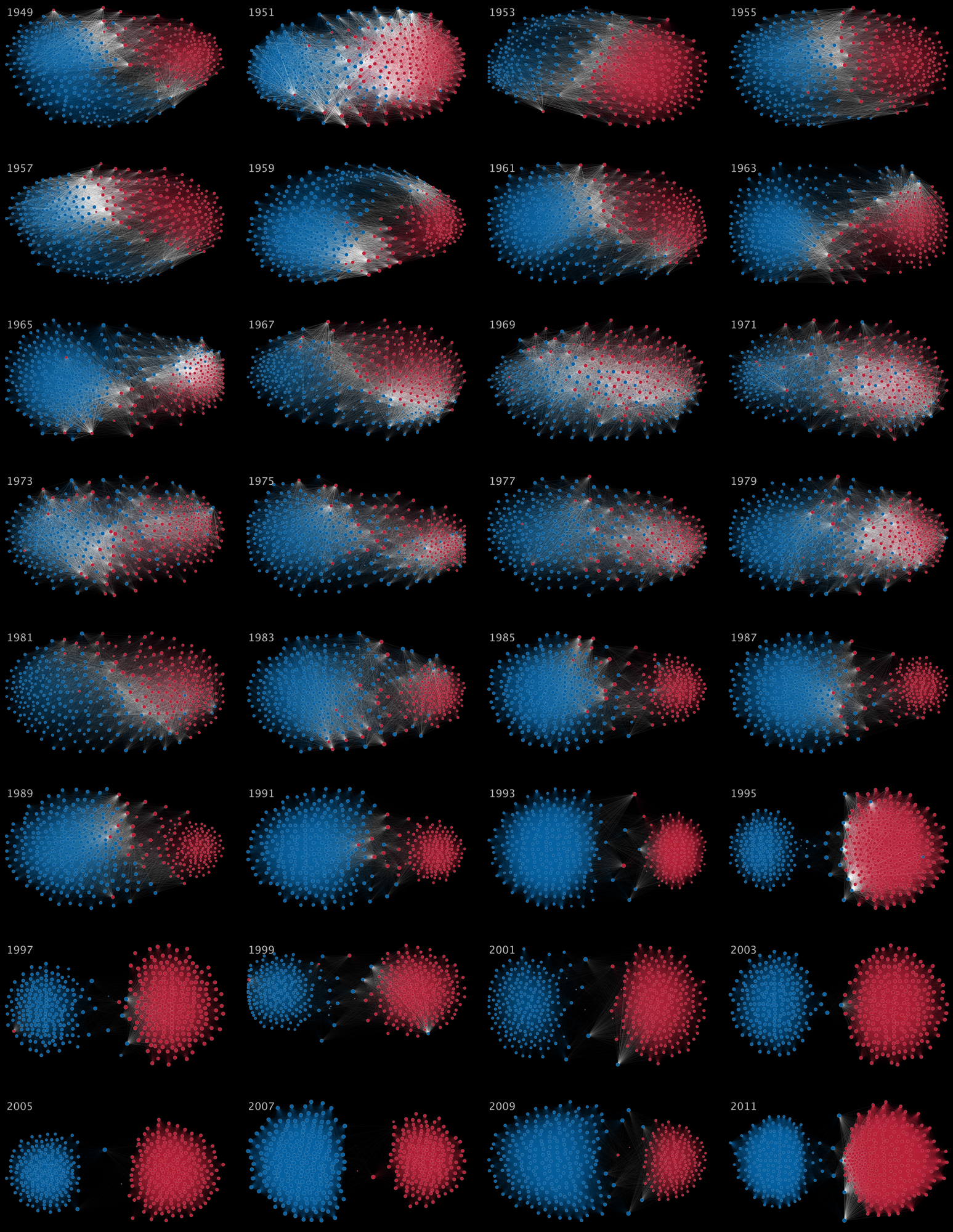
The chart uses arrows, with a tail starting at the market share value of the two largest companies in the Early 2000s, based on data availability, and the head pointed to the market share of the two largest companies in 2018. The chart does a good job of allowing resolution of the values, since each 10% slice of market share is marked by 10 small dots, allowing curious readers to obtain an actual percentage change down to the percent. The use of arrows is appropriate for showing a change over time, and the use of color is very good for distinguishing between industries that have increased in market share and those that have decreased in market share. I also appreciate the separation of the graphic into industries with data from the Early 2000s and technology companies with data from 2012, which is quite clear and avoids misinterpretation of the time range for that particular set of industries. The vertical axes allows some amount of comparison of change in market share between industries, but since they have no common baseline, it's rather more difficult to get a good comparison. However, this seems to touch upon the purpose of the graphic - not to compare market dominance in different industries, but to get a general sense of how many industries are now dominated by their two largest companies. In fact, 15 industries shown have greater than 50% market share in the industry, which would potentially indicate oligopoly within those industries. The positioning of the industries in decreasing order of change serves to again emphasize market dominance of these companies - only 5 industries are displayed with decreases

in market share of their two largest companies, and they're set at the bottom.

There is a limited amount of text within the graphic, primarily text labels for the industries along the y-axis, and informative titles. The caption displays source of data and defines "Early 2000s." The title within this graphic defines what is being shown at each head or tail of the arrow. Further emphasis on this is provided by 10% to 100% tick labels at the top and bottom of the graphic, and an unobstrusively grayed "Market Share" for an x-axis label. The y-axis is not labeled specifically except through the title label, but the various industries are clearly marked down the left side of the plot. The label "Technology Companies Since 2012" interrupts this flow with three technology industries on the bottom portion of the plot. to aid in understanding the heads and tails of the arrows, the first of each section of the graphic has the head and tail annotated with the year, although the use of "Now" for 2018 makes the graphic slightly out of date at the time of writing of this document. All of the text labels in the chart serve a purpose, and while occasionally redundant, they serve to emphasize important parts of the graph without being particularly intrusive.

The graph as a whole is not visually stunning, but instead is informative and clear in conveying its point. Every element of the plot has a point and contributes to the understanding. Color is used to distinguish two categories - industries that showed a decrease in market share over the time interval, and industries that showed an increase. The title labels define what the data is, the caption defines the term "Early 2000s," and the arrow labels define the beginning and end of the interval for the arrows. Even the grid lines, despite being clearly in the background, convey additional information about the arrows, namely the specific percentage of that data point. The only thing I would be concerned about is the source, which potentially has some variable omission bias, as some industries may not be displayed because they don't contribute to the overall message.

2.4 Political Polarization


1949
1951
1953
1955
1957
1959
1961
1963
1965
1967
1969
1971
1973
1975
1977
1979
1981
1983
1985
1987
1989
1991
1993
1995
1997
1999
2001
2003
2005
2007
2009
2011

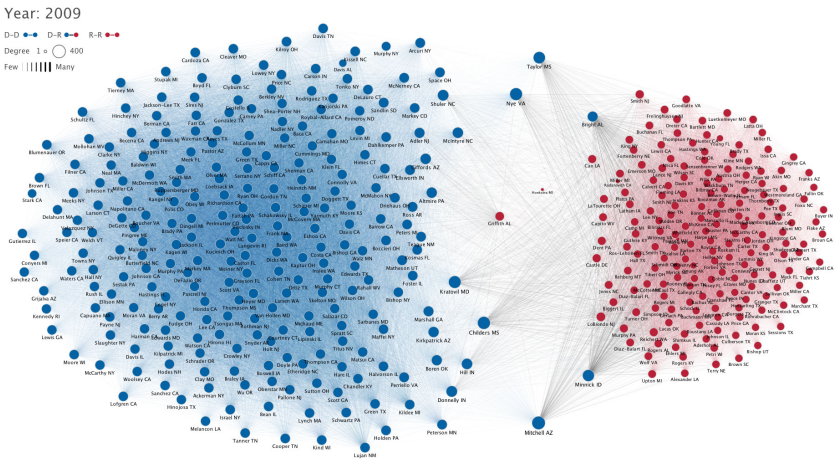


Fig 2. Division of Democrat and Republican Party members over time. Each member of the U.S. House of Representatives from 1949–2012 is drawn as a single node. Republican (R) representatives are in red and Democrat (D) representatives are in blue, party affiliation changes are not reflected. Edges are drawn between members who agree above the Congress' *threshold value* of votes. The *threshold value* is the number of agreements where any pair exhibiting this number of agreements is equally likely to comprised of two members of the same party (e.g. D-D or R-R), or a cross-party pair (e.g. D-R). Each node is sized relative to its total number of connections; edges are thicker if the pair agrees on more votes. The starting year of each 2-year Congress is written above the network. The network is drawn using a linear-attraction linear-repulsion model with Barnes Hut optimization [33].

Figure 6: Political Polarization of US House of Representatives

The above figures and the associated page are from the PLOS One article "The Rise of Partisanship and Super-Cooperators in the US House of Representatives." [10] The first part of it acts as a legend for the network diagrams that the viewer sees on the following page. The visualization is embedded in a paper that provides a detailed explanation and context for what is being seen. On the top left, a visual explanation of what elements correspond to what types of relationships. Blue circles appear to represent Democrats, and red circles indicate Republicans. It is unknown where Independents fall in, if they exist at all within this year. Democrat-Democrat cooperator edges are colored blue, cross-party edges are colored grey, and Republican-Republican edges are colored red. The degree of each node is represented by the size of the circle, and the thickness of the edges between nodes qualitatively represents how much cooperation occurs between the two nodes. As an example, the network for the year 2009 is shown, with a large mass of blue nodes on the left and a potentially smaller mass of closely knit red nodes on the right. Only a small number of nodes lie in the middle between the masses, particularly "Griffith AL." Every node has an associated last name and 2-letter state ID for identification. However this identification is lost when looking at multiples of these network diagrams.

PLOS One helpfully requires authors to acknowledge the data sources. The data for these visu-

alizations originally come from Congressional Roll Call Vote Data, provided by the Office of the Clerk of the US House of Representatives, and is accessible at "https://www.govtrack.us/congress/votes." Additional data on Congressional productivity and approval rate come from a different paper. [11] With any political dataset, the risk of bias, intentional or otherwise, should be expected. GovTrack is an independent website that retrieves official governmeent data as well as community data repositories, and their "About Me" notes that they take no donations whatsoever from outside organizations and rely exclusively on advertising revenue and crowdfunding. Overall the public face of the website and data source appear genuine and relatively verifiable.

The main visualization here is the 4 by 8 set of network diagrams, of a different year of House of Representatives. The years range from 1949 to 2011, in intervals of 2 years. The diagrams are in chronological order, reading from left to right and top to bottom. Each network diagram has the same elements of the "legend" figure of the year 2009. Network visualizations have a number of dimensions to encode information. Nodes can be treated as data points as in scatter plots. Color, size, and shape can each encode different pieces of information. Edges are used to indicate a relationship between nodes, and their color and thickness can be used to encode information as well. In the case of this graph, nodes represent representatives. The paper also creates a threshold number of agreement votes for each Congress, which is the value where each pair of representatives is equally likely to be either the same party or different party. Edges represent dyads of representatives that are above this threshold of votes. Node color encodes party affiliation of the representative, while node size indicates the degree - how many other members does this particular node have above threshold value of agreed votes. With regards to edges, edge color indicates whether the relationship is same-party or cross-party, and thickness qualitatively encodes how many agreements there are between the two representatives. It should be noted that perception of the colors is based on thickness - a reader may perceive higher rates of cross-party interactions with a few thick relationships while all of the thicker same-party relationships are masked by the many red and blue nodes.

This graphic uses the technique of small multiples to encode time. By arranging the network diagrams in a grid, a progression of the network over time can be qualitatively determined by a reader, simply by reading from left to right and top to down. Since each network diagram is structured the same way, with the left hand mass of Democrats, and the right hand mass of Republicans, comparison between years is simple. However, allowing this easy comparisons comes at the cost of resolutions of individual elements of the network diagrams. Thickness of lines and

size of nodes become harder to distinguish when scaled down in small multiples. On the other hand, the progression of colors over time become easier to track, which ultimately is the intent of the visualizations. Tracking the amount of white in the network diagrams is equivalent to qualitatively tracking the amount of cooperation occurring across parties. We can see that the amount of white decreases over time, which is the takeaway of the diagrams according to the paper. There are also a few oddities within the graphic which could be due to the force-directed algorithm used to create these networks, such as the lone red node in the mass of blue in 1997. We can also see that certain years had only a small number of cross-party dyads above the threshold value.

Text in this graphic is limited to year labels, which is all that is necessary to define the scale of the small multiples as time. The figure label provides all of the information about the plot itself, defining what variables the node color, node size, edge color, and edge thickness encode. All understanding of the plot can be found in the figure label, which is particularly beneficial, since the plot can be understood without needing the paper to provide context.

The visualization has a pretty clear singular message wrapped up in a rather complex graphic. It makes for a quite appealing poster, to watch party polarization over time, but since no quantitative details are possible to accurately perceive, what's left is a qualitative message. The use of color is the visualization's strongest point, and while it attempts to encode other information like node degree and a relative measure of how strong the cooperator pairs are, the effect is masked by resolution issues with the use of small multiples. The use of small multiples was a great choice to compare different House of Representatives over time, but showing the decrease in cooperation across parties could be demonstrated by a different plot, such as a bar plot. Using a network diagram as a unit graph provides a striking visual of two amorphous masses intermixing and separating in recent years, while using a bar plot would sacrifice this visual effect for a more information visualization. This is a trade off that the authors have made, and given the fact that this graphic is embedded in a paper providing full context and analysis of the dataset, the network diagram presents a more memorable visual for the reader.

2.5 Sobolev et. al

The above figure is from Sobolev et al. "An olivine-free mantle source of Hawaiian shield Basalts," an article published in *Nature* in 2005.[12] The graphic is primarily a scatter plot, with the addition of multiple colored fields demarcating boundaries of composition of certain specific igneous rocks. Data points from a total of 12 locations for igneous rocks are plotted by forsterite

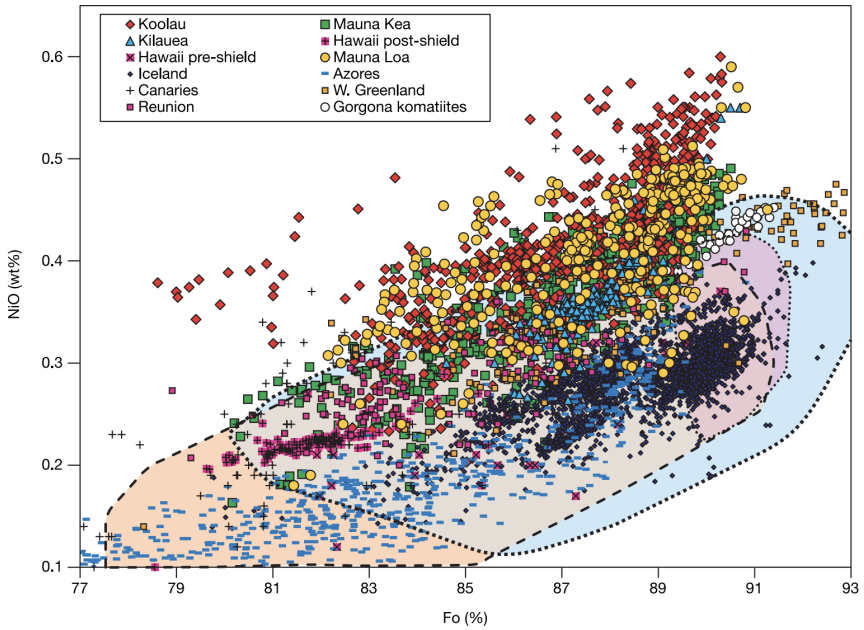


Figure 1 Compositions of olivines from mantle-derived rocks. Blue field, peridotites from mantle xenoliths, orogenic massifs and ophiolites; purple field, oceanic abyssal peridotites; beige field, phenocrysts from mid-ocean-ridge basalts; light green field, overlap between peridotite and phenocryst fields; pink field, overlap between oceanic abyssal peridotites and phenocrysts from mid-ocean-ridge basalts. Most data are from our unpublished database (data of A.V.S. on Hawaii, D. Kuzmin on Iceland, V. Kamenetsky on Gorgona, I. Nikogosian and T. Elliott on the Azores, I. Nikogosian on the Canaries and Reunion and V. Batanova for olivines from mantle peridotites). Olivines of Archaean komatiites from Belingwe show NiO contents only 0.02 wt% higher than Gorgona komatiites (L. Danyushevsky, personal communication) and follow the upper boundary of the mantle peridotite field (blue). Additional data are from the GEOROC and PETDB databases[46] (see Supplementary Information for major references) and from ref. 47. Olivines from shield-stage Hawaiian basalts vary significantly in Ni content at constant Fo, with the majority systematically enriched in Ni compared with olivine from mantle peridotites, komatiites and common basalts. Olivines from post-shield and pre-shield Hawaiian basalts are similar to peridotites and common basalts.

Figure 7: Figure 1 in Sobolev et al.

content and nickel oxide content, distinguished by a variety of colors and point shapes. The figure label primarily shows the multiple sources the data is aggregated from, including multiple different authors and petrology databases, published and unpublished. Explanation for the intent of the figure is again in the paper. Based on the paper, forsterite content is a proxy for magnesium content and nickel oxide is the proxy for nickel content, and the paper contests existing hypotheses for mantle melting, which make assumptions that constrain nickel concentrations. The reason for the figure is to make two points: Olivines from shield-stage Hawaiian basalts have large ranges of NiO content for a given Fo content relative to the ranges of olivines formed from melts at shallow depths, such as those from Iceland, Azores, Reunion, West Greenland, Gorgona, and the Canaries, and adiabatic cooling of melts generated at high temperature and pressure do not explain the high NiO content as one would similarly expect to see high NiO content in high pressure-temperature melts such as Gorgona komatiites and West Greenland picrites.

The source of the data is outlined in the figure label, naming authors, organizations, and databases

that this plot drew from. These happen to be primarily published papers in various petrology journals as well as from the GEOROC and PETDB databases. GEOROC is maintained by the Max Planck Institute for Chemistry, and PETDB is maintained by Columbia University and funded by the US National Science Foundation. It would be difficult to see where there is any bias or any reason to insert fake or low-quality data. The remaining unpublished data are from authors of other petrology papers from a variety of research institutions.

The chart is a scatter plot, encoding Forsterite content on the x-axis and Nickel oxide content on the y-axis. Rocks from 12 different locations are plotted, each with a different point shape and color, explained in the legend on the top left. Both dimensions of shape and color are used to distinguish location categories, which opens more possibilities for visually distinct categories. The legend is visually noisy to the point of hiding certain categories of rocks. The bulk of the data at 83 to 90 Fo and 0.3 to 0.5 NiO appears to be yellow circles, red diamonds and green squares, on first glance, but it's possible to miss the purple squares in those colors, or the blue triangles that look like they're hidden under the mass of yellow circles at approximately 89 Fo and 0.4 to 0.5 NiO. Similarities in the point shapes for Hawaiian pre-shield and post-shield also make it difficult to distinguish the two. In the mass at 81 to 83 Fo and 0.2 NiO, it primarily appears to be Hawaiian post shield rock, but we can see a few pre-shield rock near the bottom of the mass. At the same time, the data points for Hawaii pre-shield rock are scattered across the plot over a large range, often hiding under other points. Rock from the Canaries are shown using a plus sign symbol, but Hawaiian post-shield rock are shown using a plus sign with a purple square - and unfortunately, both types are rather close to each other, making careful consideration of the plot necessary to distinguish the two in all of the noise.

In addition to the points, there are also 3 lightly colored fields to distinguish peridotites from 2 different origins and phenocryst formations from a third origin. These are distinguished by different boundary patterns, namely dashed lines, dashed circles, and dashed squares. While 3 fields are used to distinguish categories, 2 additional colored fields are generated by their overlap. These fields appear to have been plotted after certain categories of rocks but not others - their dashed boundary lines are above the green squares and white circles, but below the yellow circles.

The primary criticism I have with this graph is just the sheer amount of visual noise that is particularly unnecessary for the intent of the plot. The paper uses the figure to illustrate the observation that Hawaiian shield-stage rocks (Mauna Kea, Mauna Loa, Koolau, and Kilauea) have higher variance in nickel content for a given forsterite value, and a higher than expected nickel

content compared to mantle peridotites, komatiites, and common basalts, and that pre-shield and post-shield Hawaiian rocks are similar to peridotites and common basalts. The multiple categories - Icelandic versus Canaries versus Reunion versus Azores should be recolored into a single category relevant for comparison. West Greenland picrites and Gorgona komatiites could be their own category since they are used to compare the low variance in nickel content compared to the Hawaiian shield stage rocks. Rather than going by location, going by pre-shield, shield-stage, and post-shield stages of Hawaiian rock would simplify the plot dramatically. The colored fields aren't referred to in the paper at all, and while they arguably contribute by giving general values for the categories of rocks relevant to the authors' model, using a light colored pattern such as stripes would avoid complicating and mixing with the colors used in the data points. All of the points made in the using the plot would be much easier to see if the categories were condensed, with the same point shape and different colors - the higher variance in shield-stage Hawaiian rocks would be more obvious if the 4 locations had the same shape and color, and you could compare them to a combined category of high forsterite West Greenland picrites and Gorgona komatiites. Post and Pre shield Hawaiian basalts could be their own individual colors, and could be compared to their own category of "common basalts" composed of the rocks from Iceland, Canaries, Reunion, Azores. The fields themselves in this particular figure aren't discussed in the paper, except for an aside comment in the figure label about a type of rock from a location not even shown in the figure, so they can be removed without major informational loss.

The only text in the figure are the x and y-axis labels, which helpfully show the units and the elemental compositions being compared, and the legend, which explains each of the 12 locations or type of rocks being displayed by the data points. The x-axis label abbreviates Forsterite as Fo, which I consider just a little bit unclear - the figure label doesn't explicitly say forsterite, and the y-axis label is a chemical compound, so I was primed to expect a chemical compound here as well. The figure label is a full paragraph detailing what the fields are, the multiple data sources that go into the plot, and the two points that the figure is intended to convey. I personally believe that the data sources are useful, but the figure should begin with the conclusions being made first, followed by the data sources. The definition of the fields and the categories they represent could be displayed in a legend in the plot, but in the case they are redundant or unnecessary, could be removed. In particular, the overlap in colors being defined appears to be a limitation of the software used to construct the plot - the authors don't use the overlap in categories for anything relevant. Lastly, I do not believe the absence of a figure title to be particularly harmful in this case,

since the figure label provides the intent of the figure, more than a short figure title could.

The figure presents a lot more information that is necessary for the intent of the figure, judging from the content of the paper it is embedded in. In addition, the presentation of the data is confusing and complex, combining two visual elements (point color and point shape) to make more categories for the same covariate (origin). While the fields are a nice visual touch, they add additional colors, and colored overlapping fields that contribute no useful information, even according to the figure label. The figure label is probably the most informative part of the figure, as it clearly explains where the data is from and states the observations the authors drew from the figure. In my opinion, the origin of the rocks is not as important as what categories are being compared by the authors. Four of the origins are treated as one category - Hawaiian shield-stage rocks, and should be condensed as one in the plot. To avoid overuse of color, the boundary fields, if the authors actually use them or refer to them, should be patterns rather than colors, which when overlapped, create differently colored fields with no more significance than "this is the overlap between these two categories." Alternatively, the authors could split up the large figure such that each figure only has one major point to emphasize. This may or may not be feasible due to page constraints, but one figure to compare the range of NiO content of shield-stage Hawaiian basalts to one category of "other olivines at shallow depths" and a separate figure for comparing Hawaiian basalts to Gorgona komatiites and West Greenland picrites would serve the authors point more effectively.

3 Visualizing Problems

Visualizations are widely used in mathematics and statistics textbooks to provide a centering framework for understanding a concept, often in the form of a diagram. Information is be encoded in diagrammatic representations to help the reasoning and problem solving process. The representation can help to infer new conclusions, elaborate on new information, and monitor the process for consistency [2, 13]. However, visualizations also have the risk of leading to incorrect conclusions, if the representation relies on extra or untrue assumptions, or if extra properties are erroneously attributed to parts of a problem due to their presence in a representation, but don't actually apply to the problem [1]. In the following case studies, I consider visualizations of mathematical problems and the quality of their solutions.

3.1 The Triangular Duel

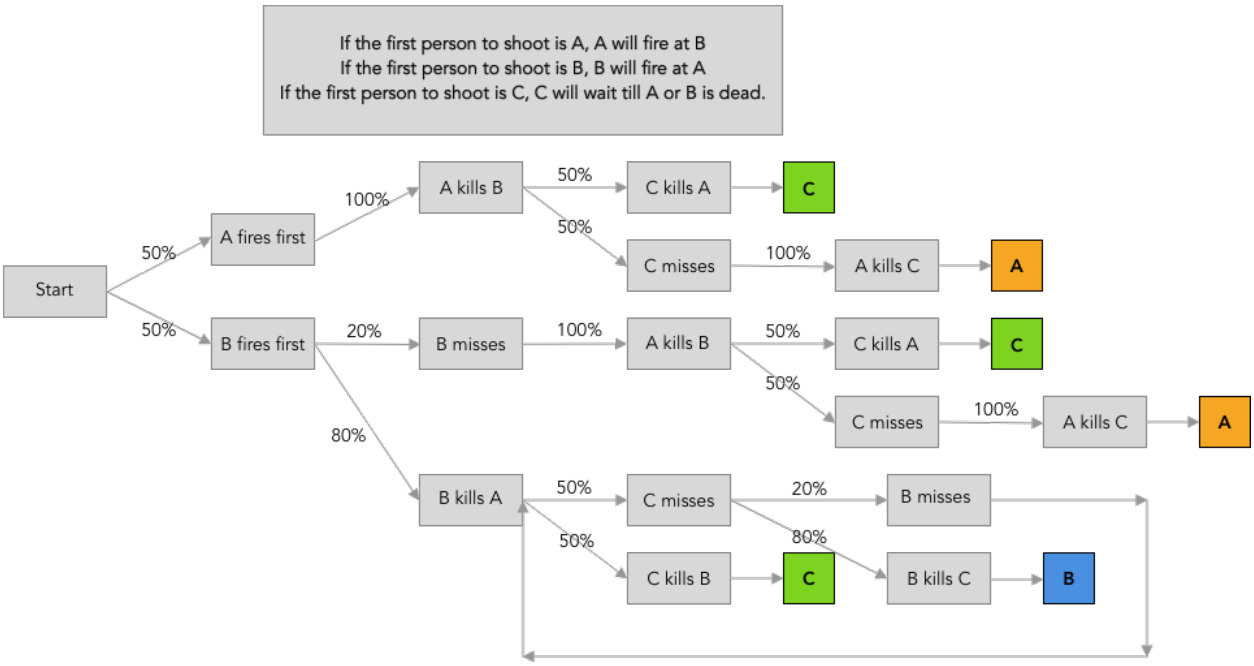


Figure 8: Tree Diagram

3.1.1 Description

Three gunmen A, B, C at equilateral triangle, will fire single shots in turn and continue in the same cyclic order until two are dead. A always hits target, B is 80% accurate, C is 50% accurate. What is survival probability of each?

3.1.2 Initial Solution

A number of prior assumptions was made for this problem. First is that no gunman necessarily must aim at someone, and second, each gunman has perfect information of the others' accuracy and follows the optimal strategy at each decision. These assumptions have the benefit of limiting the scope of the tree.

The solution to this problem can be found by propagating the probability of each decision down to the survivor. For A, note that only two leaves end with A surviving.

$$\frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{5} \times \frac{1}{2} = \frac{3}{10}$$

For B, B is caught in a cycle, where C and B aim at each other and potentially both can miss. The first iteration has a $\frac{1}{2} \times \frac{4}{5} \times \frac{1}{2} \times \frac{4}{5} = \frac{4}{100}$, the second iteration $\frac{1}{2} \times \frac{4}{5} \times \frac{1}{2} \times \frac{1}{5} \times \frac{1}{2} \times \frac{4}{5} = \frac{4}{1000}$, and so on. This is a geometric sum with $\frac{4}{100}$ being the first term and successive terms being $\frac{1}{10}$ of the previous. The sum of this geometric series is given by $\frac{a}{1-r}$ where a is the first term and r is the ratio.

$$\frac{\frac{4}{100}}{\frac{9}{100}} = \frac{4}{9}$$

For C, we could propage and add up the geometric series on top of the two branches at the top, but we could simply look at the survival probabilities of the B and A, and see that since only one gunman can survive at the end of the duel, the survival probability of C is one minus the survival probabilities of B and A.

$$1 - \frac{3}{10} - \frac{4}{9} = \frac{23}{90}$$

The availability of the tree above shows not only the sequence of events and survivors, but allows calculation of the survival probability for each sequence of decisions. However, it's possible to see that this approach may not be as scalable as one might like. Based on the assumption that C can fire and miss, and that this is the optimal strategy for C until either B or A is dead, then A and B have the initial nodes branching from start with a probability of 50%. A more accurate tree may include turn order, with 3 nodes branching from the start with $\frac{1}{3}$ probability to each A, B, and C, only to have C miss with a 100% and turn order going to B and A subtrees with 50% probability. Survival probabilities would not change, but it would add more edges, nodes, and two cycles.

3.1.3 Generalization

Aspects of the problem can be further generalized. The assumptions we've made might not hold. The players may have imperfect information and fire at random. They might have non-100% accuracy. Rather than going in order, they may roll a die to decide who shoots next, and said die may or may not be weighted. Each of these generalizations adds more and more branches, and more and more decision nodes, until visualization may be intractable. The process of problem solving may be aided by the visualization of the tree on paper. It allows a person to follow the path of decisions semantically and match it to the probability of reaching that decision node. However, with some generalizations, such as imperfect information, cycling within the tree, and

combinatorial explosion of decisions, visualization of the tree may be non-trivial.

An aspect of the tree itself should be considered. The tree as drawn is non-unique. Another person may draw it with different edges, or ignore edges with 100% probability to save space. You could also notice that there is a identical subtree and link "B misses" to "A kills B" in Fig. 8, rather than copying the subtree.

3.1.4 Visualization Quality

This graphic was drawn in the process of figuring out the problem statement. To solve for the survival probabilities, I considered the outcomes that would occur and their probabilities. Certain assumptions were made in the process of building the tree, and these were noted in the initial solution as well as in the textbox at the top of the figure.

The graphic is a tree diagram of the outcomes. The key assumption is that there is one optimal strategy for each gunman at every stage of the duel, meaning there is an optimal decision of choosing a target, and as a result, there are no decision nodes. Outcomes are encoded in the gray nodes of the diagram, and probabilities are encoded in the directed arrows. Three separate colors and associated letter labels are used to denote the winners of a particular set of outcomes. A total of 17 nodes are sufficient to fully show the possible decisions under the assumptions made for the problem. However this could be condensed further by merging nodes that are connected by 100% directed arrows. I limit the degree of pre-visual errors by explicitly stating the assumptions in the textbox above the tree, which outlines the optimal strategy the gunmen are assumed to follow. In the process of making a graphic to address this problem, I ensured I covered the possibilities by checking that a full 100% exited each node.

The diagram does not solve the problem outright and provide survival probabilities for each path. I could possibly have done some of the work in calculating the probabilities by having a running total for each nodes and the probability of reaching the node, and then have arrows to sum up the colored nodes. However, drawing the running total is complicated by the cycle in the bottom. The intent of the diagram was not to solve the problem outright, but to facilitate calculation of survival probabilities. In this I consider it useful. Calculation of the survival probabilities can be done by pinpointing which paths end in a particular color, representing one of the three gunmen, and tracking the probability of getting to that endpoint, which was done in the solution above. In terms of post-visual errors, I make the assumption that probabilities can be multiplied as one proceeds down a particular path, which is a property of decision trees rather than tree

diagrams. However, the diagram was generated with this in mind.

In terms of improvement, the first is the lack of guiding text. I made the diagram for myself, with little consideration for other readers. As a result, I have assumptions and prior knowledge built into the diagram that aren't explicitly labeled. There is no label for what problem this diagram is for, although I've placed the figure within a paper that provides that context. The text box at the top outlines the optimal strategy, but is not labeled as such. I've defined the nodes as actions taken by one gunman, with the exception of the start node. It's also not immediately clear how to arrive at the final survival probabilities. I've purposefully drawn the diagram to look like a decision tree, but the knowledge to how to calculate survival probabilities from one is expected prior knowledge for a reader.

3.2 Displaying Informational Entropy

3.2.1 Description

How should the relationship between entropy and information gain be presented?

3.2.2 Initial Solution

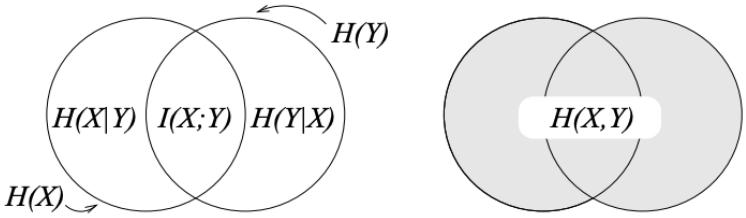


Figure 9: Traditional Venn Diagram of Entropy and Information Gain

Figures 9, 10 are both from MacKay's "Information Theory, Inference, and Learning Algorithms." [14] In the field of information theory, mutual information, or information gain, is a quantity representing the "amount of information" in bits, obtained about one random variable X, from observing a different random variable Y. This is naturally linked to the concept of entropy, which quantifies the expected value of the information content of a random variable.

The entropies (H) and conditional entropies of random variables X, Y are related to each other

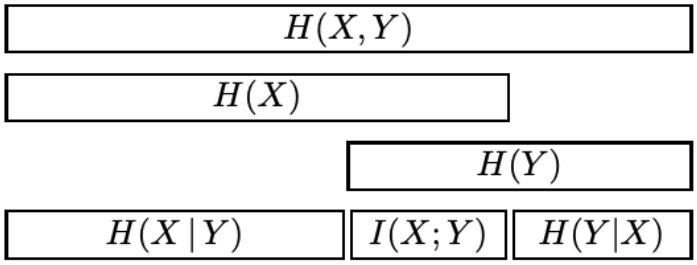


Figure 10: MacKay's Version of Entropy and Information Gain

and information gain (I) in multiple additive and subtractive ways. A common way to visualize these relationships is through a Venn Diagram as in the upper figure. MacKay provides the alternative bar representation in the lower figure.

3.2.3 Generalization

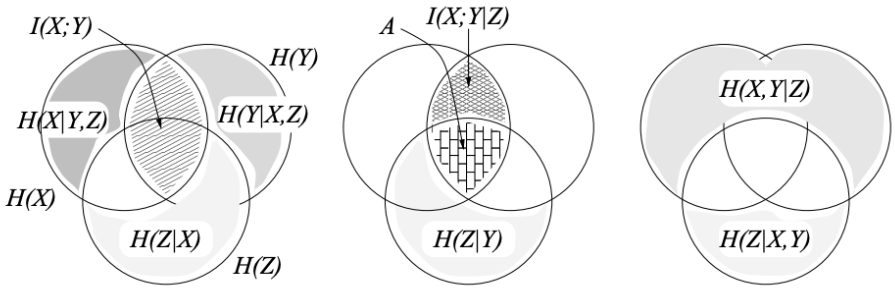


Figure 11: Traditional Information Gain Diagram

In generalizing to multiple random variables, the number of additive relationships and different conditional quantities to consider exponentially grows. The traditional information diagram uses a Venn Diagram to illustrate the relationships, and are a useful pedagogical tool. However, their primary caveat is that multivariate mutual information is a signed measure, and can be negative. As long as the Venn diagram is interpreted in the sense of relationship between sets, with set unions, intersections, and differences, then the understanding of the relationships between the multiple conditional entropies and multivariate mutual information quantities is maintained. However, the quantities themselves are not sets - there is no member of the 'set' $H(X)$, and think-

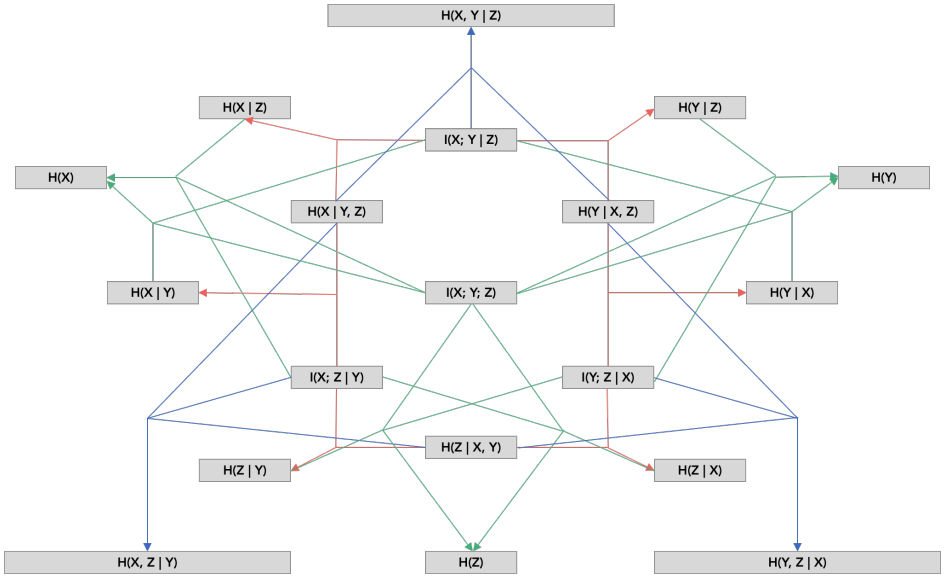


Figure 12: Information Gain Diagram Using a Network

ing of a random outcome (x, y) as a point within the diagram would be confusing entropies and probabilities. In addition the relative areas of the Venn diagram may not correspond to the actual relationships between conditional mutual information. For example, $I(X;Y|Z)$ can be greater than $I(X;Y)$, when the corresponding $I(X;Y;Z)$, at the center of the Venn diagram, is a negative quantity.

Overall, the purpose of the information diagram is to illustrate relationships between quantities. Venn Diagrams are useful in this manner because they show all possible logical relationships between the quantities. Unfortunately, generalizing MacKay's stacked bar representation of these relationships isn't quite possible, since any particular quantity will be involved in more than 4 additive relationships at a time, so there isn't any "base" line with component quantities, as there is in Figure 10. More standard methods of purely depicting relationships, such as the network diagram in Figure 12, aren't scalable. Even with only the additive relationships shown, is a rather complicated mess of lines.

3.2.4 Visualization Quality

Figure9 is a Venn Diagram, $H(X), H(X|Y), I(X;Y), H(Y|X), H(Y)$ are shown in the left, at their associated areas, and $H(X,Y)$ on the right, with the diagram grayed out. Without context, it's possible that a reader would not understand that the labels represent quantities and not sets, as

noted before. Additive and subtractive relationships are identified in a similar manner to the set union, set intersection, and set difference operations. For example, $H(X|Y) + I(X;Y) = H(X)$, and this is observed by noting that the left circle of the Venn diagram is labeled $H(X)$, and is composed of two areas, labeled $H(X|Y)$ and $I(X;Y)$. As long as the quantities $H(X)$, $H(X|Y)$, $I(X;Y)$, $H(Y|X)$, $H(Y)$, $H(X,Y)$ are understood to be individual quantities and not sets, and that relationships between the quantities are given by standard set operations in the context of the diagram, then the diagram is useful as a mental image or pedagogical tool for understanding.

Figure 10 is MacKay's version for two random variables. A rectangle is used to represent each quantity, and these are stacked on top of each other to show the relationship between the quantities. The relationships are encoded by looking at which bars add up in length to other bars. "Size" of a quantity is encoded in the length, and relationships are read by considering both the length of the rectangle and whether the rectangle is above or below other related quantities. The relationship $H(X|Y) + I(X;Y) = H(X)$ is encoded by having the rectangle labeled $H(X)$ above the rectangles $H(X|Y), I(X;Y)$, which occupy the same row, with a total length that of the rectangle for $H(X)$. Similarly, $H(Y) = H(Y|X) + I(X;Y)$ has the rectangle $H(Y)$ above the rectangles $I(X;Y)$ and $H(Y|X)$, with the same total length. Unfortunately, it's still possible to draw inappropriate analogies from this diagram, as the lengths themselves are used to encode some concept of size or magnitude. It would not be unexpected for a reader to compare the length of $I(X;Y)$ to $H(Y|X)$ and expect them to be equal in magnitude, or compare $H(Y|X)$ to $H(X|Y)$ and expect $H(Y|X)$ to be less than $H(X|Y)$, when those comparisons are inappropriate.

Figure 11 is the generalization of the two random variable Venn Diagram. It shows all of the logical relationships among the 7 component areas of the diagram. On the left, various conditional and marginal entropy quantities are colored in various shades of gray, and mutual information quantities shaded in different patterns. Only certain quantities are labeled, and other conditional quantities can be deduced by symmetry. The relationships between the quantities are valid as long as one remembers that the "sets" are signed quantities, and a point within the "sets" doesn't have any useful meaning in the context of information theory. Three Venn diagrams are used to fully show at least one example of marginal and conditional entropy distributions - $H(X,Y|Z)$ is shown on the right, which is composed of, and read as equal to $H(X|Y,Z) + I(X;Y|Z) + H(Y|X,Z)$, which are shaded various shades of gray on the left and middle diagrams. The shading is somewhat useful, but less for showing relationships and more for understanding which component areas of the diagram correspond to what is labeled. The use of patterns for mutual information

quantities and shades of gray for entropy quantities is a helpful visual cue, and ultimately necessary to distinguish the many labels in the diagram. The figure also uses text labels to establish where the quantities are represented, and this is absolutely necessary for understanding. A lot of information about relationships between these quantities is encoded in the Venn diagram, in a very concise way. However, in taking advantage of readers' presumed understanding of Venn diagrams, post-visual errors occur, with inappropriate analogies possibly being drawn from them.

Improvement on the Venn diagram to avoid the post visual errors is difficult to do. Standard visualizations for illustrating relationships is not as scalable or as concise as in the Venn diagram. The three random variables require 7 component entropy quantities, of which 12 more conditional entropy quantities and 7 component information quantities can be generated through additive relationships, and every component quantity is part of at least 4 additive relationships. In the network diagram in Figure 12, the nodes encode quantities, and edges are used to shown directional additive relationships. This makes for a rather complicated figure, even when particular relationships are distinguished by color. In the above figure, the 7 component quantities are placed in the middle, with red lines depicting additive relationships between two random variable quantities, green lines for the marginal quantities, and blue lines for the conditional distributions of two variables conditioned on a third. The network diagram avoids the assumptions made with Venn Diagrams, but this comes at the cost of readability. Even without all of the mutual information quantities such as $I(X;Y)$ displayed, the figure requires many crossing lines due to the many relationships any one component is part of, and is hardly as concise as the Venn Diagram.

3.3 Mutilated Chessboard Problem

3.3.1 Description

Consider a standard 8 by 8 chessboard, and domino pieces of size 2×1. The problem question is the following: "Is it possible to cover the entire board with 2×1 pieces, without overlapping pieces, if we remove opposite corners of the board?"

3.3.2 Initial Solution

The answer, for a standard 8 by 8 chess board, is no. One way to discover this answer is a parity argument - Note that a 2×1 piece must cover a black square and a white square. In Fig 13, it's clear that the opposite corners of the board are the same color. The remaining board after opposite

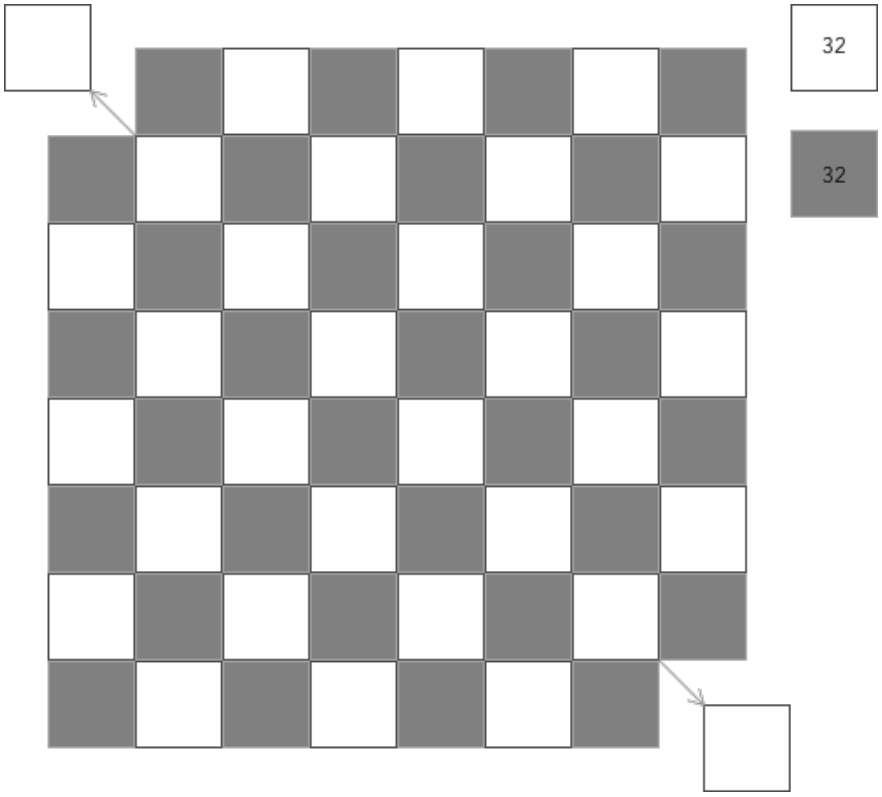


Figure 13: Mutilated Chessboard Problem

corners have been removed will have 32 black squares, and 30 white squares.

3.3.3 Generalization

The problem can be further generalized by extending the board, and considering an $n \times n$ board. However, the parity argument still stands, as opposite corners of a square remain the same color. Furthermore, for odd n, there are $n^2 - 2$ tiles remaining after removing the corners, and a simple proof shows that $n^2 - 2$ has an odd number of squares, so it wouldn't be possible to fill the remaining board with 2×1 pieces regardless.

The second extension to the problem is whether we could tile dominos on the board if we removed 2 squares from arbitrary places in the board, given that we remove squares of different colors. Now the parity argument no longer works. If the board is still a square with an even side length, then it's possible for dominos to tile the board. It's also possible to decompose the problem and construct a visualization that allows us to take advantage of techniques from other fields, specifically graph theory. Consider Fig 14. Decomposition for this problem involved converting

each cell of the board into a node, and connecting each node to adjacent nodes in a grid pattern, exactly how dominos would overlay on the board.

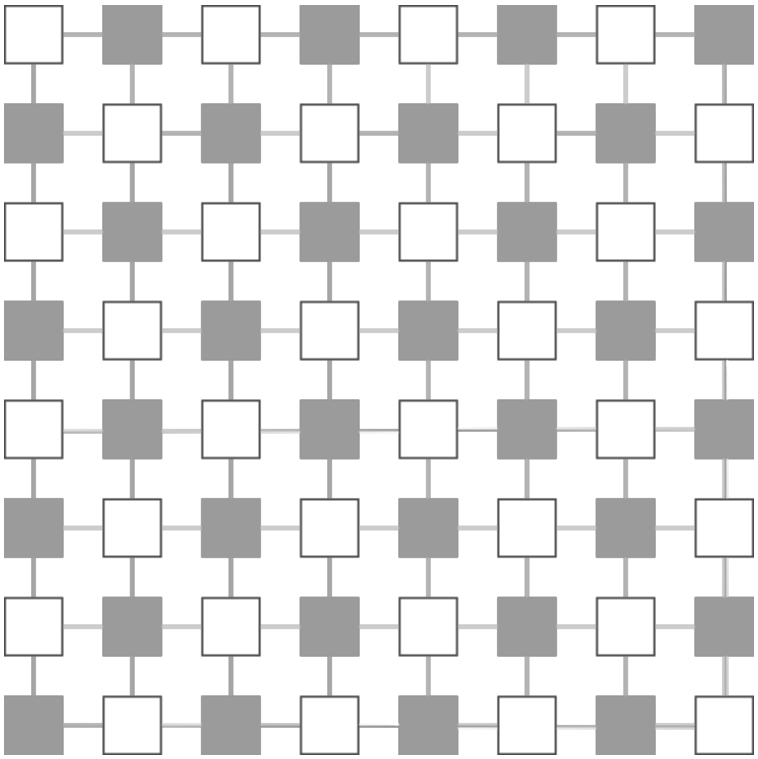


Figure 14: Network Decomposition

Note that each cell of the board can only be covered once. In graph theory, a Hamiltonian cycle is a path that visits each node exactly once. Once such cycle is shown in Fig 15. If I remove two arbitrary nodes from this cycle, the cycle is split into at most two paths. If the two nodes are of different color, it can be shown that these paths are of even length. Since these paths are even length, not only does a solution exist for tiling the board after removing the two squares, but I have a specific arrangement of tiles where the 2×1 pieces should be placed. If the nodes are of the same color, (setting the parity argument aside) then the two remaining paths have odd length. As a result, the decomposition remains consistent in showing that there is no solution for removing two squares of the same color.

The next generalization I consider is the extention of the board from a $n \times n$ square to a $n \times m$ rectangle. In this case, I can still use the same network decomposition, and this is where I can draw on graph theory. If a Hamiltonian cycle exists, then a tiling can be found. Luckily, there is an existence theorem for Hamiltonian cycles in rectangular grid graphs, and it can be shown that

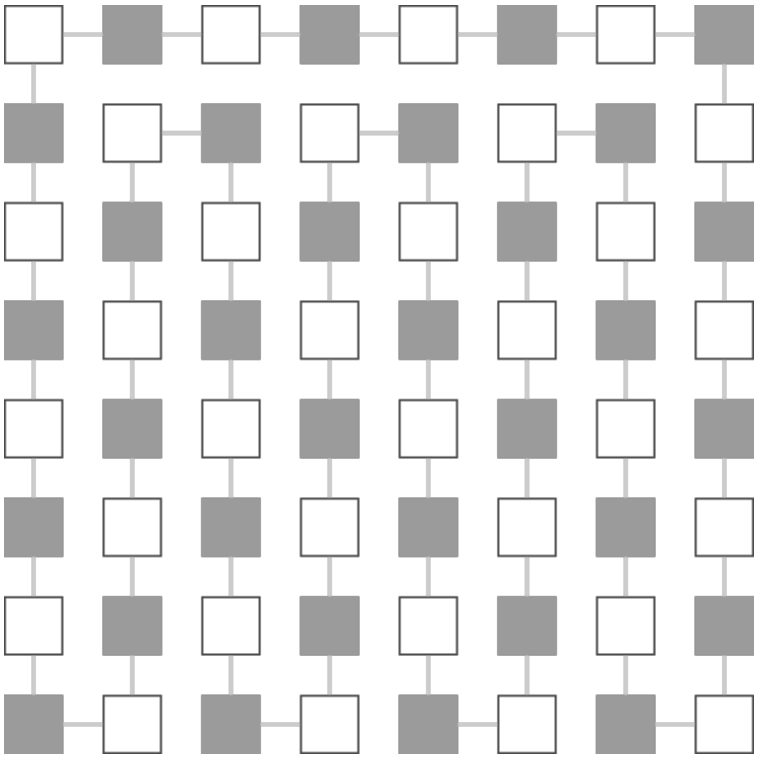


Figure 15: Cycle Decomposition

if mn is even or $mn = 1$, then there exists a Hamiltonian cycle. [15] However, these cycles are not unique, and it can't be shown that if a Hamiltonian cycle does not exist, then a tiling doesn't exist. That statement is equivalent to the contrapositive, that a tiling existing implies a Hamiltonian cycle. However, the statement is false, a 2x3 rectangle with opposing corners removed does not have such a cycle, while a tiling does exist. Since a tiling can exist without a Hamiltonian cycle, the network decomposition cannot show that a tiling exists. While the network decomposition may be useful under certain conditions, the visualization may not generalize and new lines of thought for solving the problem may need to be considered.

3.3.4 Visualization Quality

The visualizations in the above figures are primarily intended to suggest different thought processes to solve the problem and generalizations of the problem. Fig 13 is essentially a picture of a chessboard with two white corners removed. On the top left are a white square and black square, with "32" labeled on them. To some degree, the visualization provides the beginnings of the solution, specifically within the black and white coloring of the chessboard. However, there is

no black and white domino image to complete the picture and provide all the components of the parity argument.

Fig 14 is a visualization of the network decomposition concept. By decomposing the problem, we can potentially draw new ways of solving the problem from different fields, such as graph theory. The grid graph created by the chessboard is depicted by encoding nodes as chess tiles, and the edges representing potential places for a 2×1 domino to be placed. This allows me to use concepts such as Hamiltonian cycles in Fig 15 to algorithmically solve the problem, but ultimately is not sufficient to rigorously prove whether a tiling exists or not. I would consider this a type of pre-visual error, where analogies are drawn from a visualization of a chessboard that are not actually entirely valid in the context of the problem. A Hamiltonian cycle actually allows for two possibilities for a tiling, so tilings are not unique even with a particular Hamiltonian cycle. Whether the visualization was successful in solving a generalization of the problem is dependent on whether one considers a new line of thought useful, regardless of whether it was effective in generating a proof.

4 Discussion

4.1 Data Visualization

In the first five case studies of this paper, different graphics were presented and analyzed across a number of different axes from the perspective of someone trying to draw as much information from a graphic as possible. In the New York Times' column "What's Going on in this Graph?," readers are instructed to consider three questions:

1. What do you notice?
2. What do you wonder?
3. What might be going on in this graph?

In addition to these questions, I would add the following

The goal of this thesis has been the development of what considerations a reader should have to judge the "quality" of a graphic, or "What makes a good graphic?" However, this is an intensely subjective question, and judging quality with as few assumptions as possible is what motivated the development of these questions.

1. Where does the data come from?
2. What is the designer's intent?
3. How is the data presented?
    a) Does the graphic use a standard type of visualization?
    b) What covariates are encoded in what visual element?
    c) Are there any visual elements that distort the data?
4. Why is the data presented, in this particular way?

Figure 16: Questions for Data Visualizations

A disconnect exists between the designer of a graphic and the reader of the graphic, mediated by the visual elements of the designed graphic. The designer may have particular understanding of what any visual element within the graphic is meant to be carry - that is, an intent. A reader of the graphic will have their own knowledge base or understanding of the visual elements in the graphic, and as a result, may not fully receive the message that the designer intended in the graphic. Different readers may look at the same graphic and leave with completely different conclusions. Many of us may understand how information is encoded in the spatial dimensions of a point in a scatter plot, but Pew Research conducted a study in 2015 that showed that only 63% of American adults could read and understand a scatter plot.[1] The questions posed above are tools for a reader to understand a graphic, with the assumption that a reader is capable of reading one. These questions have the goal of encouraging readers to spend more time and increase the amount of information they gain from reading graphics.

The first question is centered on the source of the data, and is important in grounding a reader in where a designer is starting from, and making a reader look more critically at data quality. Data are primarily generated through experimental or observational studies, and can be of varying levels of quality. A graphics reader is generally unable to specifically evaluate the quality of the data used in the graphic, but based on the context of the graphic, such as if it were embedded in an academic paper as in Fig. 7, it may be possible to evaluate the procedures used to collect the data. In addition, one can consider different levels of bias in the data generation process. The data for Fig. 5 was sourced by IbisWorld and Open Markets Institute - with the express mission

to expose and "reverse the stranglehold that corporate monopolies have on our country." [9] It would be expected for the story of the graphic to align with this mission. Certain covariates might be omitted from a study, or additional data that could be found in a scraped database is not used in the graphic present. A graphic is often designed with a story in mind. One should critically consider if the data is being mutated to tell a story different from what the data is, to align more closely with the goals of the data source. A skeptical reader should separate what the data says and what the source wants them to say.

The second question considers the graphic designer's intent in packaging the data into the graphic. I separate the designer's intent from that of the organization that commissioned the graphic, or that of the source of the data, as they are not necessarily aligned. I assume that designers have an intent, otherwise they would not go through the trouble of creating a graphic in the first place. In some cases, this may be persuasive in nature - as in Fig. 1, which is visually attractive but with minimal data - or informative, such as in Fig. 4, with the a small number of key takeaways. One would hope designers create graphics to inform and allow readers to draw their own conclusions, but a large number of graphics are persuasive, and misleadingly so.

The designer's intent is a "why" of the graphic. Designers transform and repackage data into a presentable form, for a purpose, and readers will get hints from elements of the graphics, but mostly from the context of the graphic. Fig. 7 is embedded in the authors' paper, and the discussion and references to the figure rather directly explains the intent of the authors in creating the graphic - to challenge existing hypotheses with empirical evidence and motivate the development of their own model for the composition of particular igneous rocks. In contrast, Fig. 1 is provided without context, and was pulled from the website of a design agency, on a public facing part of the site. Presumably it was to illustrate an example of the agency's work, but it could also have been the use of a particular slide template to "spice" up business presentations. A reverse image search shows that the image is primarily a template, with an unknown author. The intent of the author was probably to create something visually appealing that could be sold. Context in this case was only mildly helpful, and definitely not helpful in understanding the data presented.

The third question goes specifically to the visual elements of the graphic. There are many subquestion associated with this, due to the massive number of ways to encode information. I choose to start by considering the different knowledge bases of the designer and reader, and the use of standard graphics. A number of standard plots are commonly used to display data, primarily due to simplicity and wide applicability. For example, scatter plots are used for displaying two contin-

uous covariates by encoding information in the position of a data point relative to two axes. Bar graphs use bars to display one categorical covariate and one continuous covariate. A reader that can recognize when a standard plot is used can quickly observe what data are being presented and what visual element is used to display the data. Being able to recognize what covariates are being shown and what visual elements are used to encode the covariates provides a reader a guide to what relationships there are between the covariates, and the intent of the designer in showing the data. In general the goal of a visualization is to tease out the relationship (or lack of) between covariates, and a reader has to know what is being shown and how to understand a graphic.

In addition to understanding what is encoded in a graphic, a skeptical reader needs to consider how well the visual elements encode the information. If colors are used to distinguish the origins of rock as in Fig. 7, are they used appropriately? The data points within the figure overlap and hide each other such that a reader can't see all the data points at once. This creates the effect of visual noise, and hampers understanding of the graphic. Psychologists have developed what is known as the Gestalt principles of grouping - or how we distinguish different groups within an image.[16] Knowledge of these principles will lead an informed reader to how data might be distorted, and to either disregard the graphic or correct for the distortion in the readers' understanding of the graphic.

The last question ties the designer intent to the understanding that a reader has generated from parsing the visual elements and data shown within a graphic. Optimistically, the decisions a designer makes in creating a graphic are motivated by the data, but realistically, may be motivated by an intent to mislead or bend the understanding of the data readers generate, or just a basic need to make the graphic visually appealing. By this point, a reader has considered the intent of the designer in some way, and has generated an understanding of a graphic in terms of the covariates presented, and the conclusions drawn purely from the details of the graphic. I now consider the decisions the designer has made in creating the graphic, while being cognizant of their intent, and qualitatively judge how successful they were. This places the reader as sort of an outside viewer looking into the process, which is partly the intent of the development of these questions. Being in this outside position is beneficial in objectively evaluating the "quality" of a graphic.

4.2 Applications in Problems

In the last three case studies, different problems and generalizations of these problems were presented along with visualizations created to represent the problem and possibly lead to a solu-

tion of the problem. The ”data” encoded in the graphics are elements of the problem statements, and naturally, each graphic was specific to the problem. I do not believe that it's valid to judge the quality of a visualization generated from the statement of a problem by if and to what extent the visualization leads to a solution for the problem statement. The quality of a visualization of a problem should be examined by whether it accurately represents the state of the problem solving process, and whether it minimizes pre-visual and post-visual errors.

First and foremost is the difference in situation compared to the graphics discussed above. Graphics created in the problem-solving process have the designer of the graphic and the reader of the graphic being one and the same. The intent of the designer is the elucidate some understanding of the problem, and to represent the state of knowledge that the designer has of the problem. The designer intent is quite clear, and the purpose of asking half of the questions above no longer apply. In addition, the ”data” of a problem are the parts of the problem statement, limiting the use of standard plots and other data presentation methods to orient readers, and requiring customized displays of data. An outside reader who views the graphic that another reader has generated for themselves may not understand the graphic because of a difference in background knowledge. The meanings of particular symbols and arrangements of visual elements may have certain meaning in one field, and different meaning in others. As a result, whether a visualization actually leads to a solution will depends heavily on a readers' background knowledge and its alignment with the designer's.

Second, judging a visualization by whether it leads to a solution is a categorical scale for quality. With questions such as ”how can the relationship between entropy and information gain be presented,” Fig. 9, 10 are both ways to illustrate accurately the relationship with different merits to each, and both are considered ”solutions”. With the Mutilated Chessboard problem however, Fig. 13 does not solve the problem, only represents it visually. Comparisons of the depictions of entropy and information gain based on the end result of creating a ”solution” is limiting. Creating a visualization that shows a solution without knowing a priori what the solution is, or if one even exists, would be a matter of chance. If visualizations are being created to aid in the understanding and problem solving process, judging the quality of one by the end result would be counter productive.

Visualizations in the problem solving process have the capability to orient readers with the state of knowledge about the problem, and to point in multiple directions of inquiry. At the same time, they may lead to pre-visual and post-visual errors. Since the information presented in the

statement of a problem is usually quite different from the presentation of data as in the first 6 case studies, a different set of questions should be considered with graphics created in the problem solving process. I present the following as a general guide for both readers and designers of graphics.

1. Are all relevant components of the problem statements represented in the graphic, and if not, is there a reason why they were not included?

2. Are there visual elements that might have different connotations or meanings?

3. Are there relationships between components of the problem that exist in the graphic but not in the problem statement?

Figure 17: Questions for Visualizations of Problems

The first question addresses the elements of the graphic similar to the data visualizations, but in context of the problem solving process. A graphic used should include the state of knowledge of the problem. An omission can be interpreted as a pre-visual error on the part of the designer, who has missed a pertinent component of the problem in the visualization, or as a piece of information, that the component is not relevant to the solution. In either case, the reader should consider critically for their own thought process why the component is not relevant, and whether they agree with the designer that the component is irrelevant.

The second and third questions emphasize the possibilities of pre-visual and post-visual errors from visualizations for problems.[1] Interpretation of visual elements depend on a person's prior knowledge for recognition and understanding. However, visual elements can be found in multiple different settings, and the mixing of usages in interdisciplinary work can cause confusion. A misalignment between the reader and designers' interpretation of visual elements can lead to post-visual errors, as readers attempt to synthesize understanding, and draw conclusions based on visual elements that were meant to be understood in a particular way. For designers, these questions should encourage looking at their graphics from other angles, particularly in relation to individuals that haven't gone as far in the problem solving process. In addition to the visual elements, the consideration of the relationships between the components of the problems is one that creates new directions of inquiry. As with network decomposition in Fig. 14, graphics can lead to new perspectives on a problem, and point towards to solutions based on work in otherwise

unrelated fields. At the same time, this approach is exactly where post-visual errors can occur, since it is unknown which relationships that are not in the problem statement are actually valid and which ones are not. The only defense against post-visual errors, given careful consideration to the answers of the questions above, is ensuring consistency in pursuing these new relationships and the lines of inquiry they advance. If an inconsistency occurs, backtrack and discover what assumptions were made, then consider how the graphic can be redrawn to update and remove the assumption or relationship that led to the inconsistency.

5 Conclusion

Questions\Figures	3D Pie Chart	Fast Food and Wealth	Industry Oligarchies
Data Origin	Unknown	Little Agenda World Bank Euromonitor International	Definite Agenda Open Market Institute IBISWorld
Designer Intent	To sell the slide template as a product	Indicate a relationship between change in fast food sales and change in wealth in multiple countries	Persuade reader of increasing concentration of market share in few companies in many industries
Standard Plots	Pie Chart	Scatter Plot	Line Plot Variation
Covariate Encodings	Height: Nothing Angle: Nothing Color: Company Text: Percentages	X-axis: Change in GDPPC Y-axis: Change in Fast Food Sales Color: Income level	Segments: Market share (2000s-now) Y-axis: Industry Color: Increase/Decrease Text: Industry
Visual Distortion	Percentages don't add up to 100% Visual dimensions unused	Income thresholds not defined	Not a full picture of all industries
Why this format	Visually appealing product Information presentation is not the focus	Two continuous variables, one categorical variables Easy to read in scatter plot with color	Shows magnitude and scale of increase in market share over time Uses understandable symbols

Table 1: Summary of Questions Responses for Data Visualization Case Studies Part 1

Questions\Figures	Political Polarization	Hawaiian Magma Composition
Data Origin	Little Agenda Congressional Roll Call Vote Data	Little Agenda Various authors and studies in petrology
Designer Intent	Characterize cooperation between US Representatives over time	Present counterexample to existing models of composition of Hawaiian magmas
Standard Plots	Force Directed Network algorithm	Scatter Plot
Covariate Encodings	Node: US Representatives Node Size: Degree of cooperation Node Color: Political party Edge: Agreement above threshold Edge Weight: Agreed votes Edge Color: Same/Cross-party Small Multiples: Time	X-axis: Forsterite Content Y-axis: NiO Content Color: Origin Shape: Origin Fields: Expected Compositions
Visual Distortion	Text is unreadable Nodes, Edges are unidentifiable Edge weight differences are not perceptually distinguishable	Too many data points obscuring each other, making visual noise Overlapped colored fields contribute no information
Why this format	Usage of network in rest of analysis The general trend of decreasing white/ decreasing cooperation is highly visible	Comparison of composition of rocks from different origins Depict as much data as possible as currently known

Table 2: Summary of Questions Responses for Data Visualization Case Studies Part 2

In this article, I have dissected five data visualization case studies on what information is presented within the graphics themselves, and what can be inferred about the authors intent. This dissection has led to the creation of a question-based framework for rating the quality of a visualization. Quality of a visualization is determined not by visual appeal, but by the content of the visualization, the designers' intent, and the amount of information that can be conveyed. This question-based framework attempts to bring out the components of the visualization that are relevant to the communication of information through the medium, and the determination of quality rests on the reader and the subjective nature of how any particular aspect is weighed.

In the above tables I've summarized my application of this question-based framework to the five case studies presented in this paper. The origin of the dataset presented is described in terms of possible agenda along with the actual source of the data. Whether or not there is an agenda is a function of the organization gathering the data and a subjective personal belief of the objectivity of the procedure. In no cases were the data-gathering procedures public. The awareness of a potential agenda is expected to temper a reader's interpretation of the results and encourage objectivity. The evaluation of designer intent is based on the context of the visualization, if any, and is partly subjective. This particular qualitative explanation is important in assessing a visualization since it provides information on the thought process of the designer, and can help in the interpretation of the visualization. In many cases, the relationship between two covariates was the suspected primary intent of the designer. The third component of the framework is the analysis of the visual elements within the graphic. This is the medium that a designer uses to communicate information. A clear understanding of the usage of the visual elements in central to understanding the graphic, and guiding questions within the framework ask readers to capitalize on prior knowledge of standard statistical plots, if any. Assuming no prior knowledge, the reader is encouraged to see how information is stored in various dimensions and visual elements, and to consider if visual distortion exists that can mislead their interpretation of the data. The explicit connection between covariate and visual element gives a qualitative sense of how much complexity the visualization can potentially have. A reader can identify if dimensions are unused, as in the 3D Pie Chart, or if multiple dimensions are used for one covariate, as in the Hawaiian magma chart. The final question of the framework ties together the designer intent and the visual elements, and is a question of whether the chosen format of the visualization is appropriate for the designer's intent. This question is essentially checks for success in visualization, of whether the intent of the authors is actually conveyed by the visualization. All of the components that are

brought out by the framework can be weighted differently based on a reader's inclination.

Questions\Figures	Decision Tree	Entropy Venn Diagram	Cycle Decomposition
All Relevant Components	Probabilities of gunmen shown Strategies described in text All possible paths shown	All relationships shown	Chessboard represented with nodes Potential domino placements by edges
Visual Elements	Directed arrows may not be understood	Venn diagram might be interpreted as sets	Chessboards are not usually represented as networks Nodes might be seen as dominos
New relationships in figure but not in problem	None	Relative areas to positive quantities	Cycles result in tilings Tilings however, do not result in cycles

Table 3: Summary of Question Responses for Problem Case Studies

In addition to the five data visualizations, three mathematical problems were solved, with visualizations used during the problem solving process. A visualization created during the problem solving process is different from that of the visualization of a dataset. The judgment of quality depends on different components. The usage of a visualization created during the problem solving process is primarily to aid in the process, but at the same time can create pre-visual and post-visual errors. The purpose of the question-based framework is to make components that directly impact the creation of pre-visual and post-visual errors, and make them explicit. The first component is whether the visualization shows all of the information presented in the problem statement, which is the initial state of knowledge in the problem solving process. The second component considers whether the interpretation of the visual elements may differ based on different prior knowledge. This speaks to the readability of the visualization to different audiences, but also to the production of post-visual errors. Conclusions drawn from misinterpretation of visual elements will impact the problem solving process. The last component relates to the interactions of visual elements, which primarily lead to new directions of inquiry. However, these also create the potential for post-visual errors. These components are made explicit by the question-based framework, and provide a means for readers to judge quality in their own subjective manner.

References

[1] Valeria Giardino. Intuition and visualization in mathematical problem solving. *Topoi*, 29(1):29–39, 2010.

[2] Despina A. Stylianou. On the interaction of visualization and analysis: The negotiation of a visual representation in expert problem solving. *Journal of Mathematical Behavior*, 21(3):303–317, 2002.

[3] Infographic - 3d pie chart social media platforms. (n.d.). https://dabaran.com/services/design/infographic-design/infographic-3d-pie-chart-social-media-platforms/.

[4] Business infographics circle graph vector illustration. can be used for workflow layout, banner, diagram, number options, step up options, web design. https://www.shutterstock.com/image-vector/business-infographics-circle-graph-vector-illustration-172917602.

[5] Edward R Tufte. *The visual display of quantitative information*, volume 2. Graphics press Cheshire, CT, 2001.

[6] What's going on in this graph? — jan. 30, 2019. https://www.nytimes.com/2019/01/24/learning/whats-going-on-in-this-graph-jan-30-2019.html.

[7] Euromonitor international. https://www.euromonitor.com/about-us.

[8] World bank. https://www.worldbank.org/en/who-we-are.

[9] Open markets institute. https://concentrationcrisis.openmarketsinstitute.org.

[10] Clio Andris, David Lee, Marcus J. Hamilton, Mauro Martino, Christian E. Gunning, and John Armistead Selden. The rise of partisanship and super-cooperators in the U.S. House of Representatives. *PLoS ONE*, 2015.

[11] Norman J Ornstein, Thomas E Mann, Michael J Malbin, Andrew Rugg, and Raffaela Wakeman. Vital statistics on congress data on the us congress–a joint effort from brookings and the american enterprise institute. *Strengthening American Democracy*, 4, 2013.

[12] Alexander V. Sobolev, Albrecht W. Hofmann, Stephan V. Sobolev, and Igor K. Nikogosian. An olivine-free mantle source of Hawaiian shield basalts. *Nature*, 2005.

[13] J Larkin and H Simon. Why a Diagram is (Sometimes) Worth Ten Thousand Words. *Cognitive Science*, 11(1):65–100, 1987.

[14] David J.C. MacKay. *Information Theory, Inference, and Learning Algorithms David J.C. MacKay*. 2005.

[15] Christina Zamfirescu and Tudor Zamfirescu. Hamiltonian Properties of Grid Graphs. *SIAM Journal on Discrete Mathematics*, 2005.

[16] Isabel Meirelles. *Design for information: an introduction to the histories, theories, and best practices behind effective information visualizations*. Rockport publishers, 2013.