Study of Website Usage of Students via Data Visualizations

Data Visualization Class Csc 83060
Zoë Markovits, Cairo Thompson, Steven Alsheimer, Kei Nemoto

Abstract

The online world has exploded over the past 30 years; from its creation in 1990 to over 90% of North America being online, and around 58% of the world population. Many of today's studies focus on the country-wide or worldwide impact of internet use. Our research team steps forward to begin the exploration of a more individualistic study of internet usage. Through self-collected data of our own team's online history, we discovered new trends, patterns and correlations of graduate student's unique website usage. The data was aggregated into four distinct visualizations to better depict the outcome and to present an easily comprehensible image that reflected our conclusions. . This research produced the foundational steps to a new way of visualizing online usage using individual's website history versus a much grander scale.

Index Terms - Personal Data, Internet Usage Data, Multi-dimensional Data, Multi-coordinated Visualization

I. Introduction & Previous Work

The internet's domination over our generation has become abundantly clear. You can see from a short ride on the metro or walking down 5th Ave, that the internet has left us tethered to our devices almost continuously. This strong reliance on our laptops, mobile devices, and desktops has led to a different kind of society where information is coming from a brand new platform. This, of course, has led to a large influx in the research being done on the internet's relationship with our community. The relationship starts with the way we get our news, the way we interact with each other online (social media), to the way it impacts our world outside of the internet. Today nine in ten adults in the United States use the internet, and one hundred percent of adults 18-29 use the internet. (PEW) The field of research that can be done based on internet consumption is vast. We set out to become a part of this research sphere and add our own twist to what has been discovered so far.

The research up until this point has had a strong concentration on the widespread impact of internet use. The purpose of this exploration is to see the country-wide or global perspective of the internet, as opposed to more individualistic applications. There have been countless studies on how many people use the internet, how many hours people use the internet [5], the comparisons between age groups and their internet use, [2] and the most commonly used websites [3] One of the most recent examples of this is a visualization project by Max Roser, Hannah Ritchie and Esteban Ortiz-Ospina based on a World Bank data set. Within this study they used data starting from 1990 to 2017 from people across the globe that was collected by the World Bank. The main components were the total population, internet users, fixed (wired) broadband subscriptions and mobile cellular subscriptions. They chose 1990 as the starting point because it is the year in which Berners-Lee released the first web browser and the first website was put to use in a lab during this year. The conclusion of this study was multiple visualizations analyzing the change in internet usage over time and comparisons between different countries. They used a heat map of the world with a slider to change the year [Figure 2] and a multiple line graph with each line indicating a country, the x-axis being the year and the y-axis being the percent of internet users. [Figure 1] These visualizations can be seen below.

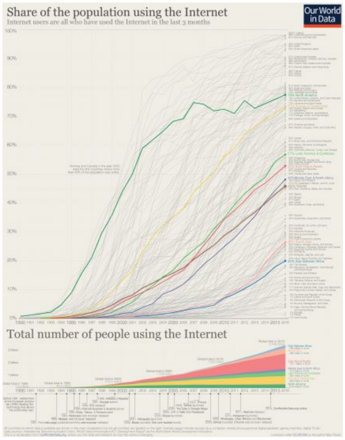


Figure 1: Multiple line graph comparing different countries' percent of internet users between 1990-2017

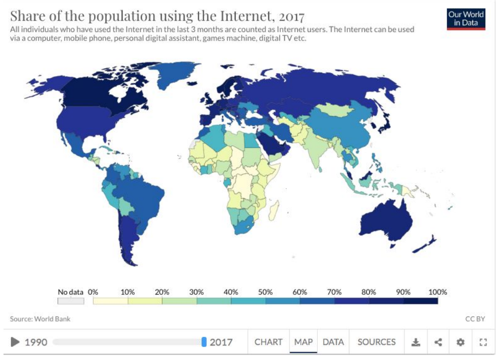

Share of the population using the Internet, 2017
All individuals who have used the Internet in the last 3 months are counted as Internet users. The Internet can be used via a computer, mobile phone, personal digital assistant, games machine, digital TV etc.
Our World in Data
No data
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Source: World Bank
CC BY
▶ 1990
2017
CHART
MAP
DATA
SOURCES

Figure 2: Heat map of the world representing the percent of internet users with a slider to change the time (year)

Although the above visualizations are uniquely created by these researchers, they were repeated with multiple variations across the past decade. Studies such as this allow the viewer to gain a comprehensive understanding of the change in world-wide internet usage compared with different attributes described above. However, the missing component is the internal comparison of individuals. This is where we wish to begin our exploration.

Another, well cited, article whose focus is more strongly correlated with our study, was *"Student Use of the Internet",* found in the Journal of Educational Technology Systems. [7] The paper was written by Dr. Gary D. Malaney at University of Massachusetts, Amherst. This paper set out to explore the way in which undergraduate students, specifically at UMASS Amherst, used the internet for both school work and personal use, and how this progressed over time. The process included sending out a survey to around 900 students in 2000 and another group of 900 students in 2003. The final results included about 600 students for each year. The questions included things such as, 'In the past seven days, have you gotten less than four hours sleep in a night because you were on the Internet?' or ' In the past seven days, have you gotten less than four hours sleep in a night because you were on the Internet related to school or work?' or 'Do you currently own a personal computer?'. It is clear through the last question about owning your own computer that this study is out of date with our latest advancements in online use. However, the conclusions of the test are applicable in the sense of finding a way in which to measure how often students are using the internet and for what use. This study concluded that internet use was exploding exponentially, even over the small gap of 2 years. Despite its strong relation to our topic, this study is outdated and fails to explore exactly which sites the students are visiting, what time, and exactly how often. The questions were relative and by asking the questions through a survey, the researchers make the assumption of the students being truthful.

We are setting out to research a more detailed account of the internet usage of college students, in our case, four graduate students of CUNY Graduate Center. We hope to bring what the other studies have not yet, a more individualistic perspective. We also look to demonstrate this view through visualizations, so to allow a more diverse audience. Through aesthetically pleasing, and self-explanatory visualizations and graphs, we will show the viewer a thorough, clear understanding of today's student internet usage.

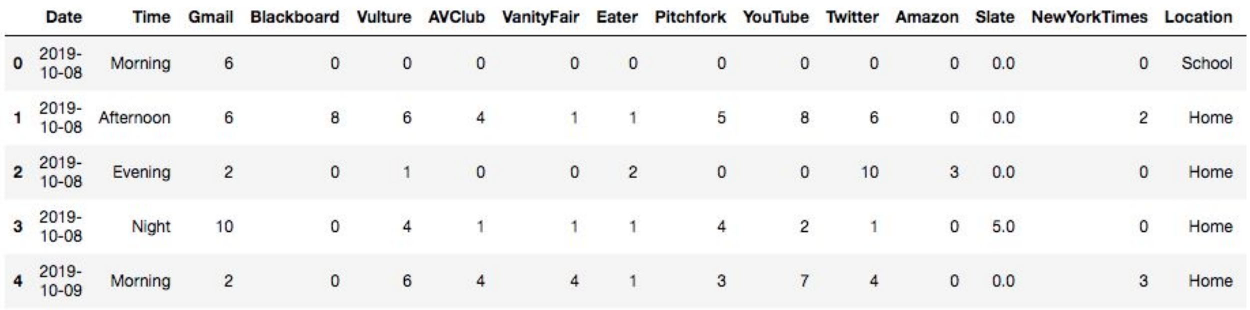
The questions we are setting out to answer focus on the change in our internet usage over a month's time split into the categories of which websites we are using. We are asking, How often do we work online versus "play"? What time of the day are we most active on the internet? How much does our location impact our internet usage? There should be patterns displayed here for each individual, but how strong are these patterns, generalized over 4 individuals? Extending the research to four different people we can include questions such as, who is the "laziest" in the group, who spends the most time working online in the morning, and how similar four students internet usage is given that they are all in the same small age range and going to the same school?

Although the scope of our visualization research is small, the impact of a study such as this on a larger scale could be used to improve a company's marketing strategy. A company could use this study to determine, for example, when the best time to send a push notification, or an email to a user. It is also insightful for a single user to optimize self productivity by recognizing subconscious patterns of their website usage and the specific times of day they are more distracted from their work.

II. Dataset

We began our research by finding a collaborative way to track our online history over a month. Our group formed a shared Google Sheet where we each manually entered our web history every day for a month (October 8 - November 8). This included, the date, the time of day (morning, afternoon, evening, or night), the website view count, and our location (home, work, or school). The time of day was split into the four categories to have a better understanding of when exactly the online usage was heavier within a day. We defined morning as 8am to 12pm, afternoon as 12pm to 4pm, evening as 4pm to 8pm and night as 8pm to 12am. We did not include 12am to 8am with the assumption that most of us spent this time sleeping.

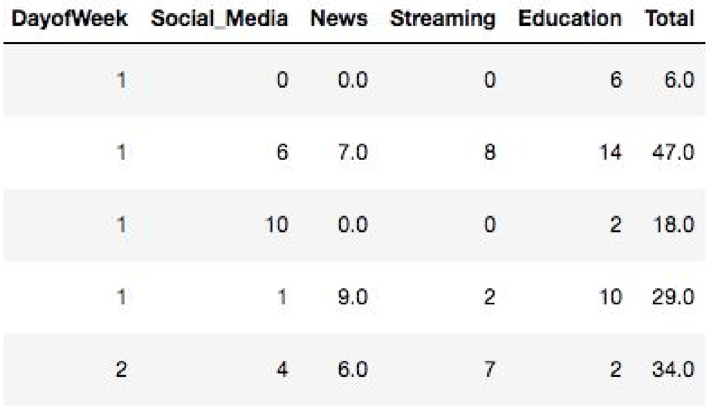
We then took the collected data, and in a Jupyter notebook we broke up our data into four dataframes, one for each of us. Each dataframe was then cleaned and pre-processed. Our data consists of a column for each of the websites, with the rows as a different date and time of day and the website view counts for that time of day. There is also one row for the location of the individual during that time. A sample of that data can be seen below.



	Date	Time	Gmail	Blackboard	Vulture	AVClub	VanityFair	Eater	Pitchfork	YouTube	Twitter	Amazon	Slate	NewYorkTimes	Location
0	2019-10-08	Morning	6	0	0	0	0	0	0	0	0	0	0.0	0	School
1	2019-10-08	Afternoon	6	8	6	4	1	1	5	8	6	0	0.0	2	Home
2	2019-10-08	Evening	2	0	1	0	0	2	0	0	10	3	0.0	0	Home
3	2019-10-08	Night	10	0	4	1	1	1	4	2	1	0	5.0	0	Home
4	2019-10-09	Morning	2	0	6	4	4	1	3	7	4	0	0.0	3	Home

Table 1: Sample raw dataset

During the preprocessing, we added a day of the week column from our initial date, and then decided to break up each person's websites into four categories: education, news, social media, and streaming. This was done because the sites we frequent the most vary person to person. In order to be able to compare across each individual, the categories need to be the same. Thus we added a column to each of our datasets for the four categories and an additional column "Total" that sums all of the individual website columns. These manufactured columns can be seen below.



DayofWeek	Social_Media	News	Streaming	Education	Total
1	0	0.0	0	6	6.0
1	6	7.0	8	14	47.0
1	10	0.0	0	2	18.0
1	1	9.0	2	10	29.0
2	4	6.0	7	2	34.0

Table 2: Sample pre-processed dataset

III. Methodology

The process of our research can be broken down into four main parts. First, we collected our own data. Second, we drew out a rough-rough draft on paper to get a better understanding of the visualizations we thought would work the best. Third, we took the drawn out visualizations and created a D3.js html rough-draft version. Within the final step, we pulled together all of the previously created graphs to design a fully connected html page of the entire research visualizations. The first section of data collection is described above in the dataset component.

The most important part of our research was conducted before the second step, deciding which visualization best represented our dataset and answered our questions. The ideal visualization will efficiently display the salient parts of a dataset in an effective and efficient way. It will answer the questions clearly and succinctly in the most comprehensive way possible. This is why we began this process with a simple drawing, to be able to quickly physically see how the viewer would be perceiving our dataset. Below you can see the results of the second part of our research.

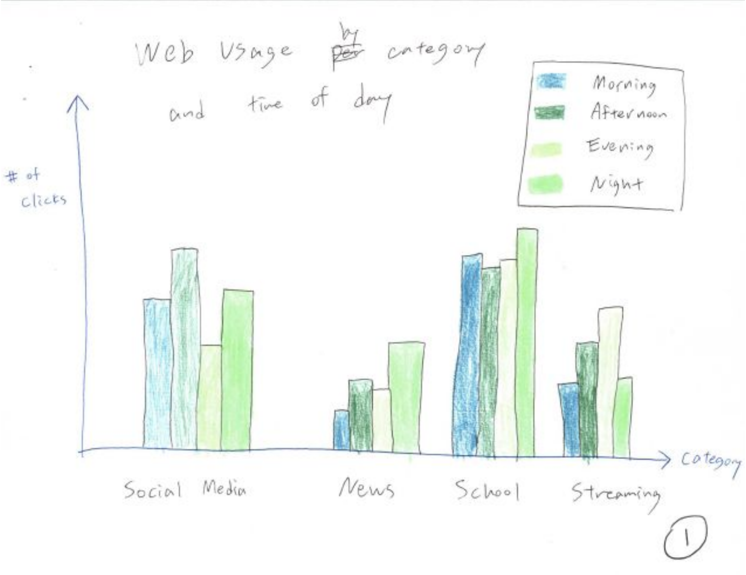


Figure 3: Sketch of Plot One

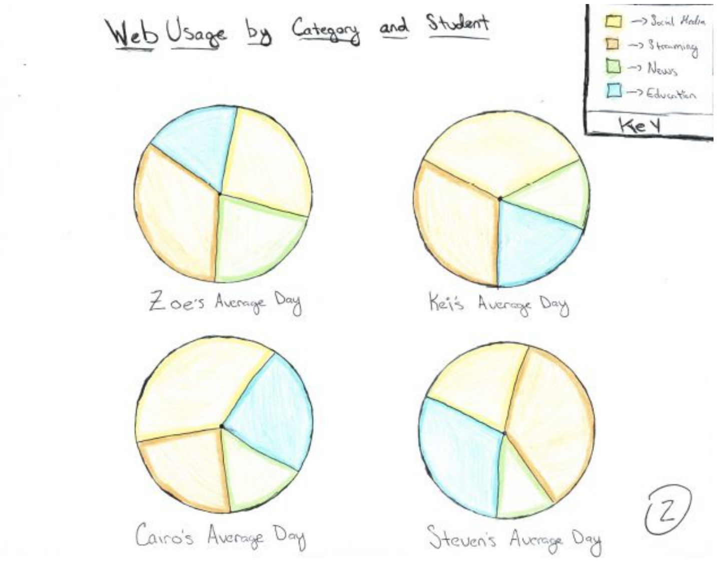


Figure 4: Sketch of Plot Two

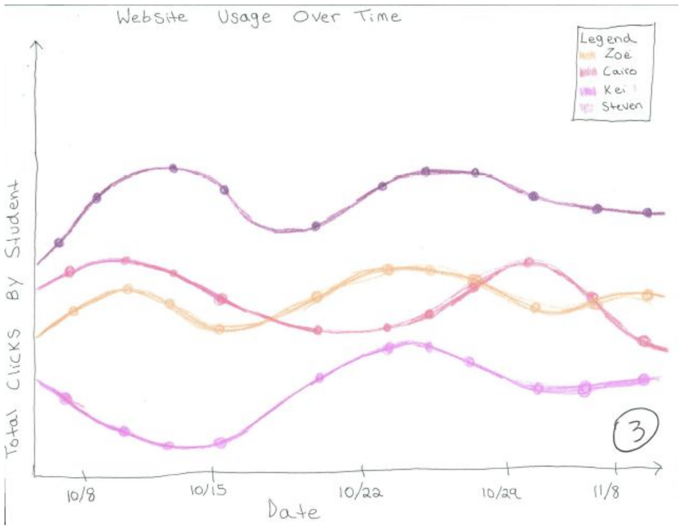


Figure 5: Sketch of Plot Three

First Visualization: Grouped Bar Charts

The first visualization we decided upon was a grouped bar chart. In general, a grouped bar chart is used when you are trying to express two different categories of data over a quantitative value within one visual representation. A grouped bar chart is, as the name implies, an extension of a simple bar chart. So, like a bar chart, it does well at presenting the details or the numerical values of the data and comparing these values across categories. The key point of choosing a grouped bar chart was this advantage of easily read values, not just visual size comparisons. The benefit of using a grouped bar chart over multiple bar charts is that it allows the comparison of “within the group” and “between the groups” In our case, the two sets of categories were defined as the category of which website visited and the category of the time of day. We also created four different group bar charts for each team member so that the comparisons were more detailed and focused on each team member separately. So in the end we had four different grouped bar charts of each member where there were four large chunks that marked the time of day(morning, afternoon, evening, night) and within each chunk was a bar representing one of the categories of websites visited (education, news, social media, and streaming).

The other option that comes up when we were discussing a rendition of a classic bar chart, is a stacked bar chart. We attempted to draw out a version of a stacked bar chart, but upon seeing the result, decided against it. This is because of the lack of a zero baseline in a stacked bar chart. Without a zero baseline, the comparison “within the group” is near impossible to understand quickly. If we used a stacked bar chart, the comparison of each category of websites visited within each time of day category would be very difficult. That is why we chose to use a grouped bar chart, that has a zero baseline so that the bar’s length is equal to its value.

The other consideration, when creating a grouped bar chart, is whether to place each of the bars within each time of day section in consistent order of each category of websites visited, or to place the bars in order of length. In other words, to have the bars go Education, Social Media, News and Streaming for each time of day block or to put each bar in order of the length. In a classic bar chart example it is often useful to put the bars in order of length for a more efficient reading of the comparison of values.

However, for our example, placing the category of websites visited in order of length would make the comparison between each time of day much harder to read.

The final decision was what color pallete to use to represent each of the “with-in the group” pieces (the categories of websites visited). We chose the classical qualitative color range, sharp contrasting different hues. This makes it very clear to the viewer upfront the difference in categories and by not using a gradient of saturation or lumiance, we do not leave any confusion of a connection between categories.

Second Visualization: Donut Charts

The second visualization we chose to use were donut pie charts. The pie charts represent the overall month’s average web usage for each group member split into 4 different categories, streaming, education, social media, and news. The general use of this visualization being the comparison, first between each group member, and second between each category within each group member’s individual graph. A pie chart is the best visualization tool to express these comparisons because it is a part-to-whole analysis, each category of internet use compared to the overall amount of internet use. It clearly expresses the difference in each piece quickly, and is a straightforward graph that can be understood by an audience without any statistical background. The benefit of using a donut pie chart versus a classic pie chart is the white space in the middle. The white space draws the eye of the audience or viewer, bringing attention to the comparison we are attempting to express.

The other important aspect of our donut pie charts is the chosen color palette. The sharp contrasting hues are always a good tool to use when demonstrating separate, non-continuous, and unconnected categories. The colors make it blatantly obvious to the viewer the difference between each category. We did not use a progression of luminance or saturation to avoid the insinuation of continuity or dependence between each category.

The biggest limitation to a pie chart is that it expresses a quick, basic understanding of the difference between categories but does not provide numerical, or exact differences. It allows only for a rudimentary visual comprehension. In order to compensate for this limitation, we added a component to give the percentage of each category that shows up when the cursor is above said category. This not only helps accommodate for the constrictions of a pie chart, but gives the visualization another dimension to draw the audience in. We also already have made a grouped bar chart that represents the numeric values of the category of websites visited for each group member very clearly. The pie chart just allows for a more general, overall view of this same data but in a more widespread view of the entire month.

Third Visualization: Stacked Line Chart

The third visualization we attempted was a line chart with the y-axis being each day of the month and the x-axis being the total amount of clicks online containing multiple lines for each team member. The classic line chart is used to show the relationship between independent and dependent values, the x-axis being independent and the y-axis being dependent. The most common y-axis being time, so that the line chart can demonstrate trends over time. We originally chose the line chart for this exact reason, to have a visualization that better demonstrated the trends over the entire month. The line chart we made is displayed below.

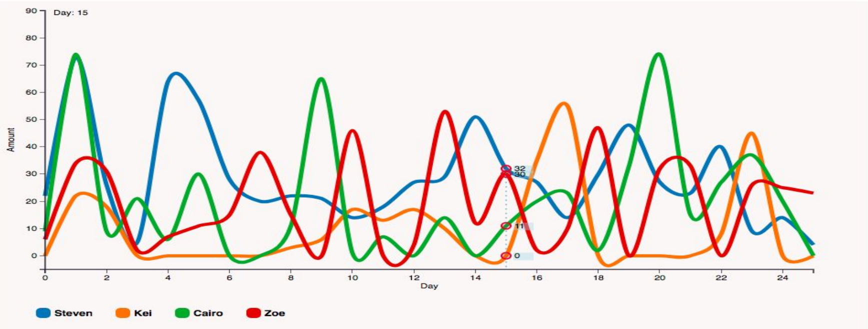


Figure 6: Stacked line chart of total website clicks over time (October 8th - November 8th)

As you can see, the graph looks very busy, each line varies greatly over time and overlaps the other line many times. It creates a graph that is hard to read and to understand. Even with the inclusion of a tooltip to distinguish the distinct values, it still does not provide a clear and instant understanding of the data being presented. The graph also has smoothed lines going from point to point, instead of using dots to mark each point along the x-axis. The data set being presented has one value for each day, in total 30 points. However, by using curves instead of dots with connecting lines, it insinuates continuity. For instance, on the green line between day 0 and 2 the line jumps sharply from around 10 on day 0 , up to a value of around 75, and then drops sharply back to a value of around 9 on day 2. This jump gives the impression of a consistent exponential-type increase in online use from day 0 to 1, when in reality it is a single value on day 0 and a single value on day 1. In order to account for the messy graph and the insinuation of continuity, we attempted another visualization graph that works well on demonstrating trends over time, a streamgraph.

Fourth Visualization: Streamgraph

The final visualization we drew out and transformed into a D3.js representation, was a streamgraph. A typical streamgraph displays change of data over time by category through the use of flowing, organic shapes that resemble a river. We chose to have a streamgraph whose x-axis represents time over the month, October 8th through November 8th. This visualization was instantly much more visually appealing. The graph provided an easily comprehensible idea of the way in which each team member's web usage changed throughout the month and also how it changed compared to the other team members.

The downfall is that you cannot easily deduct the numeric values. This was accounted for by including a tooltip to read each numeric value of websites visited depending on where the mouse arrow is placed. We also added a component to decrease the lumination of all surrounding category's (group members) "stream" when the mouse arrow was over a different "stream". This allowed for a very clear conception of what the numeric value was in response to. The other downfall that is common in a streamgraph is that it gives the impression of larger contrasts in categories than it actually is. This is common when you have a very large data set of highly varied data points. In our case, the values are close enough that it does not give off this impression and also has a much smaller dataset which helps with the issue of outliers and extremes.

The streamgraph result is aesthetically pleasing and engaging. It gives the viewer the opportunity to spot a more general view of changes over time and to catch things like periodic patterns or days in which certain group members are more likely to be using online.

Final Product

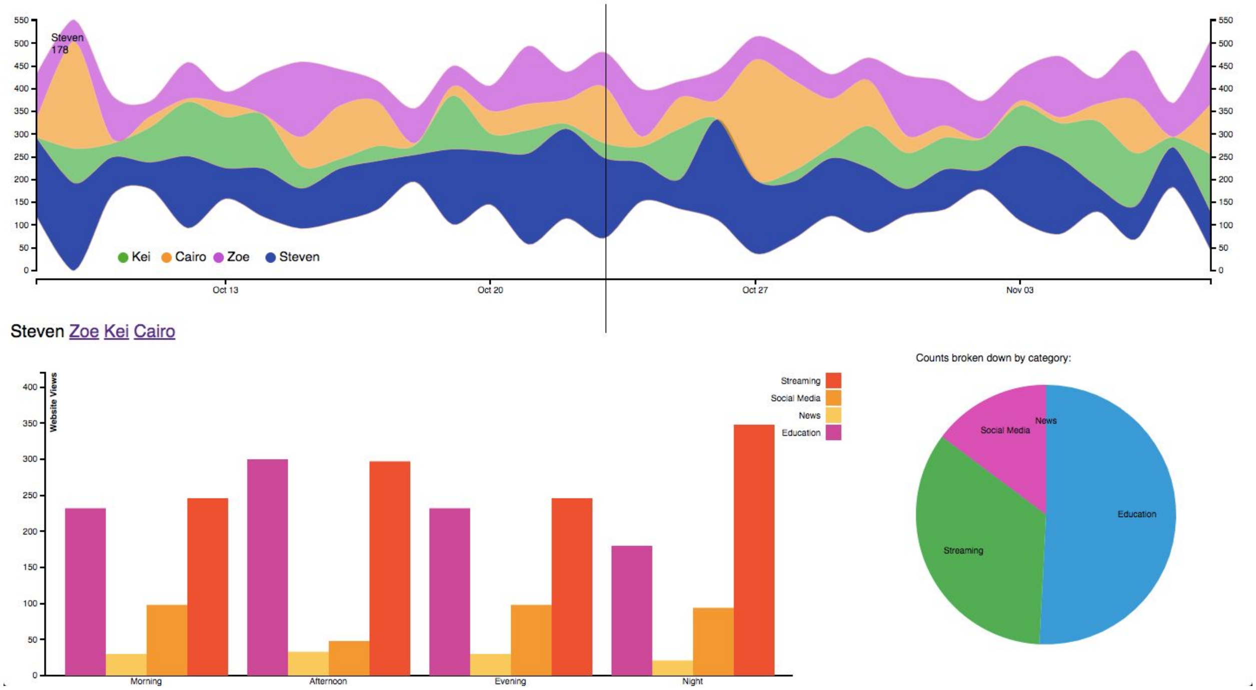


Figure 7: Multi-coordinated view of all visualizations

At this point, we have now completed the collection of data, drawing rough-rough draft visualizations, and creating D3.js rough draft visualizations. The final step is creating a collective representation of all three final visualizations (donut charts, grouped bar charts and the streamgraph). The goal in creating a multi-coordinated view is to be able to visual all of the information at once and provide a much broader and complete understanding of the dataset, the questions about it, and the answers drawn from the visualization. The ultimate objective being to allow the user to have dialogue with the data. This is done using interactions that help to illuminate the most important parts of the data, the connections between different visualizations and an interface where a story can be told. [Our multi-coordinated visualization is seen above in figure 7]

Our multi-coordinated view involves a main html page with a link to each group member (and their associated html pages), so that each member has their own page of visualizations. Within each page is a grouped bar chart with each group being the time of day(morning, afternoon, evening, night) and each bar within the group being a category of websites visited (education, streaming, social media, news) whose length is equivalent to the amount of times that category was visited during that time of day. The grouped bar chart includes a tooltip that marks the exact value of the given bar you have your arrow on top of. This allows for further interact with the viewer.

There is also a stream graph above the grouped bar chart. The stream graph is consistent across each group member's individual page (the same graph). It demonstrates a stream for each group member over the entire month. On the upper left corner of the streamgraph a html tooltip shows the user and total counts that day for the user. The added pie chart tooltip on the streamgraph will produce a pie chart of the given date and team member, that represents a piece of the pie for each category of websites visited (i.e. education, streaming, social media, news). This pie chart gives further information about the stream graph and demonstrates to the viewer the connection of data across all visualizations. By putting all graphs on the same page it is possible to get numeric values, a part-to-whole relationship, the day to day and intra - day trends and trends over time all within the same visualization.

IV. Results

Plot One - Grouped Bar Chart

The first visualization we focused on was a grouped bar chart that investigates how time of day effects category. When looking at the category feature, we can see that all four of us visit educational sites more than the other three categories, Kei visit news sites less than the three of us, and Steven visits streaming sites more than the three of us. Likewise when examining the time of day feature, it appears that Cairo, Kei, and Zoe are less active at night than the other times of day, and Steven's website usage remains steady throughout all times of the day. Finally, when looking at the visualization as a whole it is evident that Steven's website usage on average is the largest while Kei's average usage is the smallest.

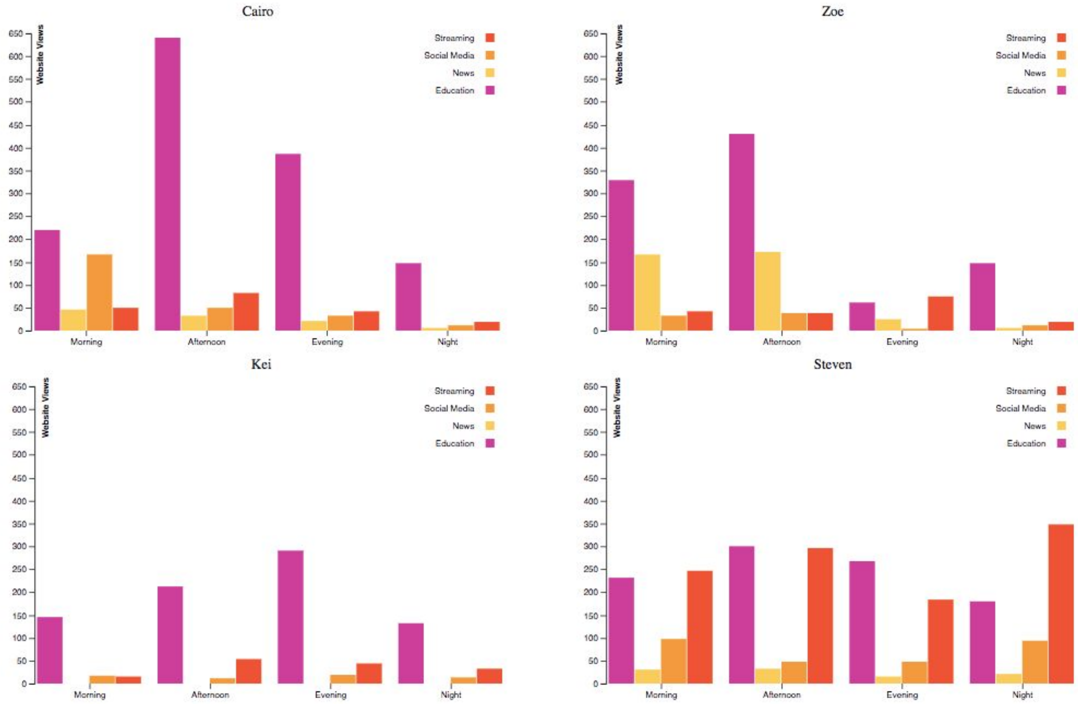


Figure 8: Grouped bar chart of website views by time of day and category of site per person

Plot Two - Donut Chart

The second plot shows a more generalized overview of each users data. It simply shows the total counts per category and displays donut based pie chart for each user. This adds to the first visualization. It is even more clear that all of the users are on “education” based websites more often than not. It is even more salient that Steven uses “Streaming” profusely and that all the users are almost even on their social media use. Zoe is evidently the newsophile of the group.

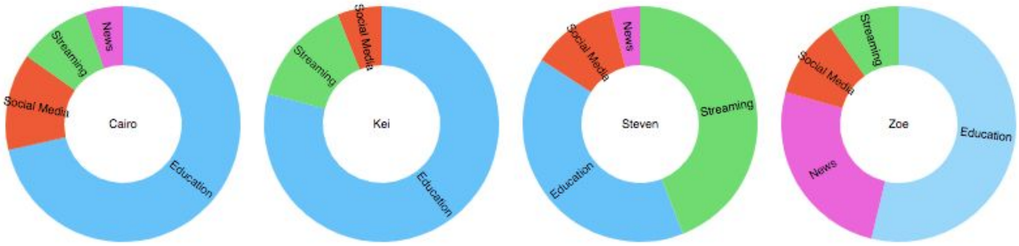


Figure 9: Donut chart of individual group members website usage by category

Plot Three - Stream Graph

The primary purpose of the third visualization is to directly compare each member's website usage by days. As you can see from the graph below, Steven constantly uses the web on a daily basis. It’s interesting to note that Zoe and Kei seem to show a similar trend, meaning that when Zoe visits more websites than the previous day, Kei also tends to visit more websites than the last day. If we focus on the difference between the web usage in the weekdays and weekends, almost all of us visit fewer websites in every weekend than we do on weekdays.

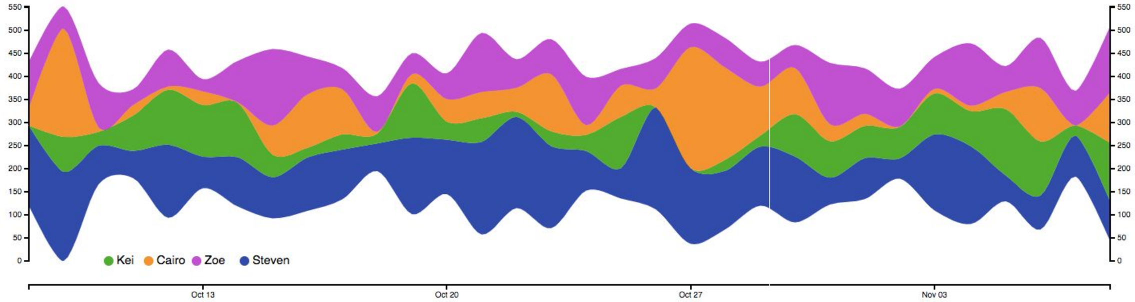


Figure 10: Streamgraph of website counts over time. The stream is an additive sum to maintain its symmetry. The total counts of all the users data is shown here, note that not all of the counts per day were binned into one of the four main categories. Total counts may be higher than the sum of the categories

V. Conclusion

As we discussed in the introduction, all four of our visualizations are done in an effort to dig deep into the website usage of individuals, as opposed to larger demographics such as an entire country or the globe. Through our visualizations we worked to find some general patterns and trends for each person and the similarities between different team members. In the previous sections, we confirm that appropriate visuazalitions allow us to extract meaningful information such as the difference in the web usage purpose per person (i.e.) some people visit more websites related to their education or some people spend more time on streaming services than other categories. Even, those four visualizations above tell us that the website usage patterns are totally different among students who are in the same program at the same school and are in the same age range. For example, the donut chart shows that each group member uses the web for totally different reasons, such as one third of websites visited by Steven falls into Streaming category, whereas Kei visits most of websites for education purpose. Also, we can see from the multiple line chart that Cairo visits much more websites than others in certain days but her average usage is lower than Zoe or Steven, who consistently use the web on a daily basis. From the observations we gained so far, we conclude that data visualization is a useful way to extract interesting insights on the website usage of individuals, even though the number of participants is limited in this project.

VI. Future Work

In possible future endeavours, we would love to expand on both the number of participants in our study and the data sources. Although the four of us as students spend much of our day on our computers, we can’t count out the importance of cellphones when looking at internet usage. It would be interesting to expand our studies to tracking the time spent on our phones, especially if we’re keeping our four categories from this study as many of us use social media apps on our phones more than our computers. As mentioned above, investigating website usage, especially the volume during different times of day, has many practical implications for a company’s strategic decisions. For instance many e-commerce sites send out promotional emails that drive sales and traffic to their business. If these companies knew what time of day their customers were most active on their computers, especially their email, they would be able to send out their emails at targeted times. Likewise with apps sending out push notifications. Furthermore, if we’re interested in expanding and digging deeper into our research questions it would be interesting to increase the number of students’ website usage we look at to see larger trends in the student community.

VII. References

[1] Apuke, O. D., & Iyendo, T. O. (2018). University students' usage of the internet resources for research and learning: forms of access and perceptions of utility. *Heliyon*, *4*(12), e01052. doi:10.1016/j.heliyon.2018.e01052

[2] Roser, M., Ritchie, H., & Ortiz-Ospina, E. (2015, July 14). Internet. Retrieved from https://ourworldindata.org/internet.

[3] Dougherty, J. (2019, May 1). Internet growth usage stats 2019: Time online, devices, users. Retrieved from https://www.clickz.com/internet-growth-usage-stats-2019-time-online-devices-users/235102/.

[4] Waddell, K. (2017, February 6). Your Browsing History Alone Can Give Away Your Identity. Retrieved from https://www.theatlantic.com/technology/archive/2017/02/browsing-history-identity/515763/.

[5] Kemp, S. (2019, January 30). Digital 2019: Global Internet Use Accelerates. Retrieved from https://wearesocial.com/blog/2019/01/digital-2019-global-internet-use-accelerates.

[6] Demographics of Internet and Home Broadband Usage in the United States. (2019, June 12). Retrieved from https://www.pewresearch.org/internet/fact-sheet/internet-broadband/.

[7] Malaney, G. D. (2004). Student Use of the Internet. Journal of Educational Technology Systems, 33(1), 53–66. https://doi.org/10.2190/VQRQ-YQX6-ARKL-7D2T