In this current “information age” that we live in, acquiring various types of data is easier than ever before. However, making sense of data and visualizing it is another, much more difficult task entirely. Network data is inherently voluminous and difficult to interpret, derive information from, and present as knowledge. Many examples of network visualizations are massive webs of interconnecting nodes that show the relationships between them, a somewhat haphazard way of organizing the information that causes difficulty when attempting to make sense of the image given the large amount of data that is being shown. What our research examines is how the ability to interact with a network visualization affects the user’s comprehension and understanding of the information contained within it. More specifically the ability to manipulate a time selection bar within a dynamic node-link diagram. In order to more thoroughly understand our field of study, first we will cover two areas that will provide a foundation of information visualization research knowledge on which to better understand our study; the first being in the field of cognitive science, and the second being a basic overview of network visualizations themselves.

Cognitive science is one of the fundamental areas of research in regards to perception of information visualizations; examining the limitations and areas of exploitations of the human body in order to realize what is and isn’t possible at the most basic level of user studies. The examination of these cognitive qualities is crucial to our research because it allows for us to account for unforeseen variables that we don’t have
control over, limiting factors that result from the subjects being studied instead of from the study itself. Existing research in this field is plentiful, a fact that’s unsurprising given the level of importance and the numerous insights contained within it. “Understanding and exploiting the abilities of the human visual system is an important part of the design of usable user interfaces and information visualizations. Good design enables quick, easy, and veridical perception of key components of that design. An important facet of human vision is its ability to seemingly effortlessly perform ‘perceptual organization’” (Rosenholtz, R., Twarog, N. R., Schinkel-Bielefeld, N., & Wattenberg, M. 2009), a facet that we want to be able to monitor and control in our experiment in order to prevent participants from creating adverse results that stem from incorrectly perceived patterns within the experimental visualizations they are shown.

The second piece of fundamental knowledge is answering the question of what exactly a network diagram is, what it represents, and the problems that can be created by them. Essentially, network visualizations are comprised of two elements; nodes and edges, with the nodes being connected with edges to illustrate relationships between the data they represent. In a basic sense, differentiation of relationships and the ability to provide a simple, easy to comprehend visualization is easy to do, accomplished by simply altering node or edge colors and shapes just to name a few. Once a user can perceive key relationships, information is extracted, information that can be hugely valuable to the world of business, science, or even society as a whole.

It is no wonder then that network visualizations have become increasingly important within the field of information visualization. They are being used to graphically represent data on a massive scale that can then be analyzed to uncover trends,
patterns, and crucial insights about the information contained within it. As stated earlier, network visualizations are, in essence, simply a visual display of elements, nodes, and the connections between them, edges. “InfoVis often involves complex and sophisticated data transformation, representation, and interaction techniques…to help identify trends, patterns, and unusual occurrences in datasets….and communicate the value of InfoVis.” (Yi, Kang, Stasko, Jacko 4). Now, while this concept seems rather basic, it can also quickly create numerous problems that are often without a clear solution.

The ability to analyze connections between pieces of data that can be from any number of datasets, which can be from the same or different sources, and then represent this information visually, is a quality that is immensely valuable. But this beneficial quality is also one of it’s fundamental flaws; that in order to gain insights on a macro level, larger and larger sets of information must be analyzed. However as more and more data is added to the visualization it quickly becomes so chaotic and busy that no clear, useful information can be extrapolated from it. Figure 1, shown on the next page, is an excellent example of this highly problematic limitation, effectively illustrating the inability to infer anything useful from the information, mainly due to the pandemonium that exists as a result of containing too much information.

We know that this image contains information and relationships, and that different pieces relate to one another in some way, but we have no way of looking at this
This image shows just how chaotic and problematic network visualizations can become when large amounts of information are represented within them.

visualization and extract any real, significant pieces of knowledge from it. This inability is the result of both the cognitive limitation faced by users (i.e. their inability to perceive any coherent regions, or patterns in the visualization) and the inability to intuitively interact with the visualization, (Rosenholtz, R., Twarog, N. R., Schinkel-Bielefeld, N., & Wattenberg, M. 2009). What makes the image even more concerning is that this visualized chaos results from a dataset that contains less than one million elements. If this results from a set of information that small, then how could a meaningful visualization be created from sets containing tens of millions, hundreds of millions, or even billions of individual elements that are interconnected?

Datasets of massive size are more common than one might think in the modern world, and the insights that are contained within them could be enormously beneficial. The social networking phenomena is one of the primary examples of a dataset containing billions, or perhaps even trillions, of individual elements that are all interconnected in a remarkably complex way. A visualization of only a small part of this network would still number in the millions, and given figure 1, it’s clear that it would be practically
impossible for average individuals to understand anything contained within a basic static, 2D network visualization.

Network visualizations come in all shapes, sizes, and styles and are unique not just in regards to the information being displayed, but in the ways they try to make understanding them as easy as possible. However, despite the fact that a good amount of these proposed solutions do indeed increase a user’s comprehension, few manage to create a significant enough increase to be deemed worthy of examination. In fact, the study relating to the “effective visualization of dynamic graphs remains an open research topic.. despite the fact that “many state-of-the-art tools use animated node-link diagrams for this purpose.” And while much research exists, the findings are often contradictory and “despite [their] intuitiveness, the effectiveness of animation in node-link diagrams has been questioned” (Ghani, S., Elmqvist, N., & Yi, J. S. 2012), which gave our team a clear problem-space to work within.

![Network visualization example](image)

**Figure 2:**

This example shows a simpler approach to how users are able to interact with a network visualization which depicts the similarity of different languages to one another. Here, users are able to control things such as the minimum amount of similarity required for the relationship between nodes to be represented as well as the ability to choose between either a network or chord visualization.
After some initial exploration and experimentation with the different types, we found that the visualizations which allow users to sort by color, arrange by quantity, zoom, and to customize what information is displayed were easiest to use because they offered users the ability to manipulate the information by enabling them to interact with it. (For examples of different ways different network visualizations enable user interaction, see figures 2 [above] and 3 [below]). This interactivity seems like a viable answer to the question of how to represent massive amounts of information in a way that individuals can not only understand, but can also perceive key connections and patterns from it. Unsurprisingly, our team is not the first to delve into researching the network visualization interaction problem. In fact, it has been researched and analyzed by multiple researchers, all of whom have offered solutions to various pieces of the puzzle.

**Figure 3:**

This network visualization, created using a dataset that was provided by Ryland Bogart, offers multiple interactive aspects including zoom, word limit, specific selection of words, and allows the user to actively click on and explore the connections between individual nodes on a micro or macro level.
One such example of this preexisting research was done by a research team comprised of Nathalie Henry, Howard Goodell, Niklas Elmqvist, and Jean-Daniel Fekete. They conducted an in-depth analysis of a few network diagrams on a massive scale delivered via direct-manipulation interaction, which resulted in a compilation of over twenty years worth of information used to answer questions, confirm existing patterns as well as discover new ones. (Henry, Goodell, Elmqvist, and Jean-Daniel 2007). While the research done by these four is highly similar in both problem-space and focus to our project, their research does differ from ours in two key ways: what the research was meant to accomplish, and who interacted with the network visualizations. The first key difference is the fact that they did not start their research to either prove or disprove a predetermined hypothesis, nor did they attempt to answer any specific question. They instead used an exploratory method to sift through the data to “generate and evaluate hypotheses—about global and local trends and outliers—interactively during the exploration” (Henry, Goodell, Elmqvist, and Jean-Daniel 2007) and the second difference is the individuals that had interact with the network visualizations are not research professionals, but our subjects were regular, average people instead, the majority of which were college students.

The goal of our research is similar to existing studies, so similar in fact that we basically share the same overall research goal as Riche and Plaisant, which is “to provide empirical evidence to guide visualization researchers in the design of effective interaction tools” (Riche, N.H., B., Plaisant, C. 2010) in order to hopefully change the way future research is conducted and how information is visualized.
In order to control the information displayed at any given time, the data that comprises the network visualization itself was sourced based on highly controlled criteria. By limiting variables outside of our control, we believe that we will be able to more effectively utilize the dynamic qualities and more easily measure individuals’ ability to “gather facts about the activities of individuals interest over a period of time by representing their networks as dynamic graphs…through animated node-link diagrams.” (Ghani, et. al. 2012). Figures 4 & 5 are both examples of the raw form of the network visualizations used in our experiment. They will be shown to experiment subjects, who will then be given tasks that will require them to navigate through the dynamic, time-varying interactive aspects that are available to them, in order to gather data as to whether or not their perception and comprehension of information increases by a measurably significant amount.

The metrics behind the creation of the visualizations were kept constant between static and dynamic tests, and were original creations sourced from raw browsing data of researcher Ryland Bogart. The development of the demonstrator visualization was a surprisingly complex process, necessary to produce quality results; it required use of five separate programs, serving as either a normalization platform or a filtering process, that applied multiple metrics in order to ensure visualizations contained significant data. In order to qualify as significant data, we attempted to limit the visual to contain only relationships that fit within certain criteria to limit existence of edges, nodes, and interactionary behavior between all elements.

Extraction of the raw data was the first step in this process; and was sourced using a program called 'ChromeHistoryView' as well as from a local 'SQLite3' database file that
contained a more comprehensive archive of browsing history. Values from both of these sources were then merged into a single text file where the values were converted into a CSV formatted file. This allowed the information to maintain universal readability between all programs being used. This produced a resulting file of node-values that were assigned a unique identifier (VisitID), a time-stamp from the time of visit (Date), and the website name (URL) for each of the 1855 entries it contained.

An analysis of this new data-file illustrated that the data gathered up to this point was unable to be accurately imported into the main visualization tool used in this experiment, Gephi; this resulting error caused any saved files to become corrupted. Eventually, it was realized that the coded time-stamp values were unable to be correctly read inside Gephi. To mitigate this error, the data was converted into a 'DD/MM/YYYY' format using Microsoft Excel 2013 (assistance provided by Michael P. Zozulia). Once the correct time-stamp values were imported into Gephi, the entire data-set was merged, creating a dynamic set of data that enabled an accurate display of time and node existence throughout the visualization.

Within Gephi, additional metrics, such as size and color, were applied to the visualization; these metrics were based on dynamic size integer values and static group IDs, respectively. The size integer metric was the dynamic value chosen to show change over time in the experiment, with each successive instance of the same URL accumulating a count value from start to end. Filters were also applied to edges, which resulted in their inclusion in the visualization to be based on whether they existed more than ten times, or existed five times within less than thirty seconds of each other, showing
big picture type behavioral aspects. The final visualization contained a total of 169 nodes, and 1855 unique instances of them spanning 70 days from start to finish.

The static visualization consists of 10 discretely rendered images, containing successively incremented 1 week time-windows, that are nonoverlapping. The dynamic animation visualization is 10 discretely rendered images, formed from the movement of a dual bound 3-day render window, being successively incremented by 1 week time-windows that are partially overlapping. (Moody, McFarland, Bender-deMoll. 2010).

Figure 4 (left):
Dynamic Animation of 10 discretely rendered images formed from the movement of a dual bound 3-day render window being successively incremented by 1 week time-windows that are partially overlapping.

Figure 5 (right):
Static Visualization of 10 discretely rendered images containing successively incremented 1 week time-windows that are nonoverlapping.

To reach a definitive conclusion as to whether or not interactivity truly increases overall comprehension and understanding of a network visualization, we created a single test (see Appendix A) for two different groups. One group of individuals were given the test along with the interactive visualization containing the information to answer the questions. The other group were given a series of 10 static images of the same
visualization at various stages the same test. Minimal information on the visualization was provided prior to the survey, in order to measure the participant’s ability to interpret and understand what the visualization was about, and any insights they had about the visualization.

For the static visualization, questions 2 and 3 both were answered correctly by all participants, but the time that it took to answer question 2, which required the participants to compare two different time-windows, was twice that of question 3, which was a false positive true or false question. However, for the dynamic visualization, question 2 was missed 20% more than in the static, but this was due to lack of answers instead of being incorrect. There was a minimal increase (+1.4) in regards to total data size perceived by users.

One key finding is that participants shown the dynamic visualization reported that readability of the visuals were 25% easier to perceive. This is supported further by the fact that participants shown the interactive visualization were 67% less likely than those shown the static visualization to have seen a network visualization prior to the experiment. However, when scores for the remaining questions were averaged, minimal effect (positive or negative) was measured from either group.

The answers given during the open ended qualitative response questions showed significantly higher understanding and comprehension of the interactive visualization. Understanding scores were found by explaining the research project, the process behind how it was carried out, and what was shown during the experiment to 10 outside individuals. These 10 individuals were then asked to rate answers from all participants on a scale of 1 to 10, in order to provide an objective measure of how well participants were
able to perceive what was being shown and why. The average "understanding" score was 4.04 for static, and 7.8 for the interactive.

We have concluded that our findings agree with current research into interactivity, supporting statements and findings that claim “interactive visualizations help users gather insights from the data by allowing them to explore it visually from different perspectives.” (Riche, N.H., B., Plaisant, C. 2010). We believe that interactivity is a positive quality, and should be used to benefit the field of network data visualization.

Future research is planned to be carried out by Ryland Bogart, with the assistance of Michael P. Zozulia, but will be carried out using an automated generation tool within Gephi. The tool is called HTTP, allowing for live visualization of browsing to be generated as it occurs, opening a world of possibilities for research to be carried out. Specifically, both researchers hope to analyze the relations and connections between Internet behavior of different individuals by giving participants a finite period of time to browse as they normally do, and then comparing the visualizations of each individual to that of other participants. This idea of using the HTTP graphing tool was introduced to Ryland Bogart after consulting Denis Parra-Santander, assistant professor at the Computer Science department at PUC Chile, about one of the issues during the development of the initial experiment visualization. Both researchers, Michael and Ryland, believe that by comparing behavioral patterns of participants’ typical Internet activity, they will be able to more easily draw insights regarding what patterns are shared commonly amongst participants and hopefully uncover why they exist.

**Course Website:**

http://mypage.iu.edu/~rbogart/i399/visual/main.html
Video:

https://www.youtube.com/watch?v=9mMs8Jtpcmg&feature=youtu.be

Citations


### Appendix A

Age: ________

Sex: ________

Occupation:________

1. Between 8/24/2013 and 9/6/2013, which site was used most frequently to visit other pages?

2. Between 8/31 and 9/13, did the frequency of visits to Reddit increase or decrease?

3. True or false: In the week of 9/14, Ubuntu was used to link to other sites more frequently than zollata.

4. During the week of 10/5/2013, which site was visited most often?

5. From beginning to end, which website (node) was visited the most?
6. In which week were the most YouTube videos viewed?

7. How many unique nodes would you estimate are represented throughout the whole 10 weeks?

8. What does the loss of a connection (branch/edge) from one week to the next represent?

9. On a scale of 1-5, how difficult was this visualization to read? Why?

10. On a scale of 1-5, how useful do you find visualizations like this? Why?

Generally speaking, what is displayed in these visualizations?

What insights, if any, did you gain from this visualization?

Have you ever seen a visualization like this before?