Visualization of Online Information Space

Justin Ko and Jerry Lei

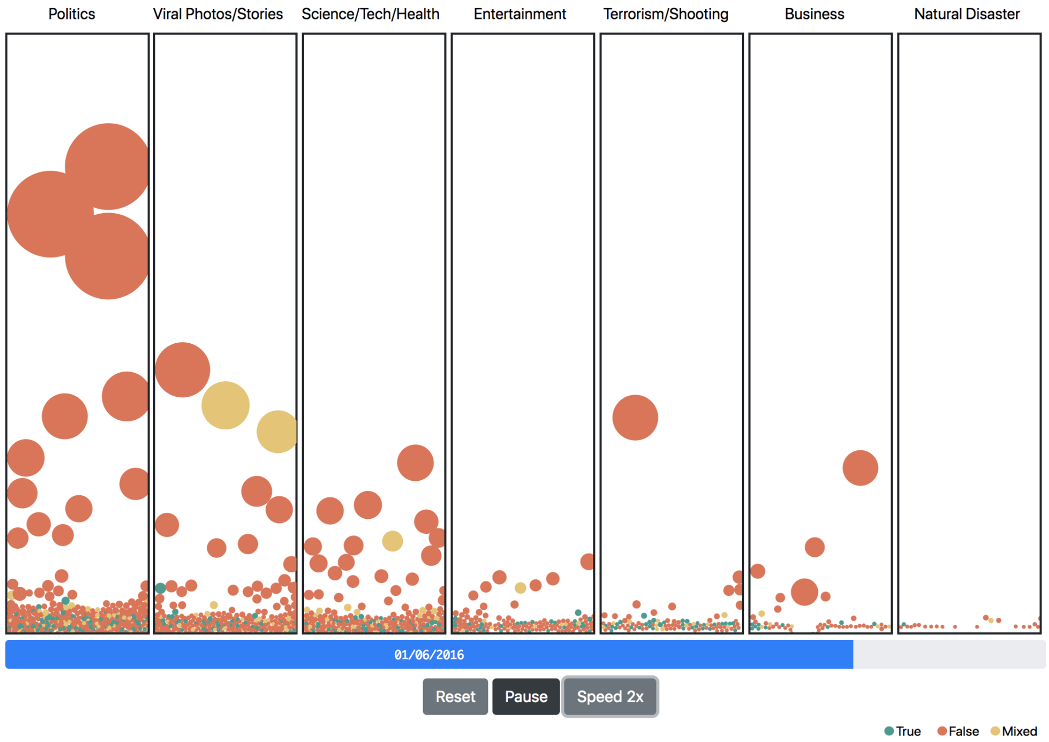

Fig. 1. Final Visualization of the online information space of Twitter.

Abstract— The internet has made the distribution and consumption of content easier than ever before. The inventors of the internet thought that this would open the floodgates to a new enlightenment era, where truth is universal and accessible. In reality what we have become painfully aware of is that the internet has made all kinds of information spread faster, not just truth. In this project we created a dynamic visualization to explore this online information space to investigate what kind of information it is made out of and what kind of information is popular in it. We focused on the Twitter online information space in particular. Our visualization represents each piece of information or factoid as a bubble. The color of the bubble represents its veracity (green = true, red = false, yellow = mixed). The size of the bubble represents its magnitude (magnitude = sum between the total number of retweets and independent tweets). We can see that the Twitter online information space is made out of mostly false information (way more red bubbles) and that false information is more popular (red bubbles are larger on average).

Index Terms—Social media, information, news, visualization

1 Introduction

What was once a discussion among only a relatively small group of internet savvy media consumers has expanded into a global discourse in the light of serious consequences that the online world has had on the offline world. Online channels, especially social media, have become a dominant part of people's daily routine around the world For many people social media is their primary source of information and news. Just like consumers have been made aware of what is inside their food, we believe that media consumers should be made aware of what kind of information is on these social media platforms. It is important, then, that we understand what kind of information is shared online and what kind of information is dominant online.

We were motivated to create a visualization that would be informative to people in understanding what kind of information is shared online. We coin the term online information space to refer to the general "shape" of the information shared online (e.g. a question we might ask would be is the online information space made out of mostly true or false information). Our goal is to create a dynamic visualization of the online information space for the general public. Our hope is that people will learn what kind of information is shared online and practice caution when they come across suspicious information while browsing on their social media feeds. We also hope that people will reflect on their own online behavior (e.g. retweeting a tweet) as a result of interacting with our dynamic visualization.

Our research question is two folds:

1. What does the online information space look like?
2. What kind of information is dominant online?

We hypothesize that there is an equal amount of true and false information, but that false information is dominant. We expect to see an even split of true and false information in our dataset. We also expect to see that false information would have greater magnitude, which is a number that represents the sum of retweets and independent tweets.

2 Data Source & Related Work

Our data came from a paper, *The Spread of True and False News Online*, which analyzed tweets since Twitters inception. They found 126,000 rumor cascades, sources of false information using the following fact-checking sites: snopes.com, politifact.com, factcheck.org, truthorfiction.com, hoax-slayer.com, and urbanlegends.about.com The authors gathered their tweets by scraping the data from these sites and searched for tweets that referenced those sites to address a rumor. They would then trace that tweet to generate a chain with other tweets that have some relationship to the original tweet. We were originally concerned about the methodology that the authors chose to use because we were not sure that this data was representative of the actual overall Twitter space. We were also worried that the authors had filtered out too much data, because we received a data set with only 4 million tweets, and approximately 500 million tweets get sent out every day. We learned that the authors of the paper addressed the concerns by running an additional study. They took an independent sampling of 3,000,000 tweets and found comparable results in terms of veracity. They also ran their analysis with and without bots, and found that bots did not significantly influence Twitter's online space. The authors found that false news usually spread through users that had significantly fewer followers, followed significantly fewer people, were verified significantly less often, and had been on Twitter for significantly less time. It turns out that false information is 70% more likely to be retweeted than true information, which makes it diffuse faster and farther. This is due to false rumors being more novel compared to true information which tends to cause more shock and surprise. The authors also did an interesting study on the word choice of tweets and how people reacted to true and false information They found that false rumors tended to invoke emotions of surprise and disgust, while true rumors resulted in replies with more sadness, anticipation, joy, and trust [1].

Our initial design for the visualization represented tweets as nodes on a board. Each node had a color representing the veracity of the tweet, and the category the tweet fell under. The color palette used for the set of categories was retrieved from a qualitative color scheme on ColorBrewer, and it appeared to be effective for differentiating the categories [3]. During our peer feedback session, we were told that the colors we chose to represent veracity clashed with the colors we chose to use for category. We suspect that the greens and reds used may have fallen under too few just noticeable differences [2, 6]. We also came across an issue where the nodes frequently overlapped, and we planned to use an efficient algorithm to remove overlapping nodes [5]. We also planned to use trees to represent the online space and ideally we would utilize edge uncertainty techniques to indicate when a tweet was far away from a source tweet in a graph [4].

3 Visualization Design Evolution

Our initial idea for the visualization was to have a canvas in the center of the page where graphs of nodes and edges would appear. We wanted to have some statistics on the left side and controls/buttons on the right side. Our idea was to represent each piece of information with a graph, where each edge represents a retweet or an independent tweet relating to that information.

An iteration on this idea was to have all the graphs representing true information start from the top and have all the graphs representing false information start from the bottom. We thought that this would create a novel visualization where our audience could see how information spreads online and the ratio between true and false information.

We planned to simultaneously work on the user interface and explore the dataset, and come back to discuss whether we should continue on the path with the original storyboard or come up with a different plan. Jerry wrote the scripts to collect statistics from the dataset and Justin created the user interface the controls for the visualization We discovered that our dataset was too vast and varied for us to be able to draw all the nodes and edges on the screen.

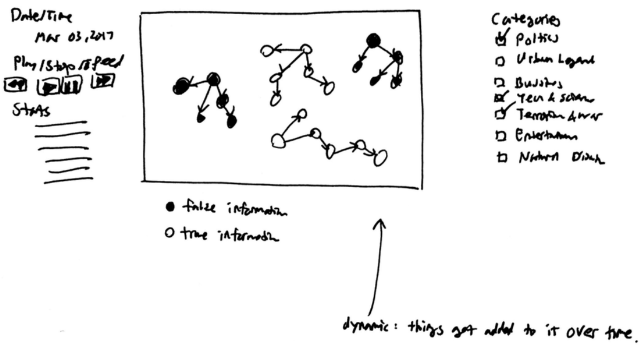

Fig. 2. Storyboard.

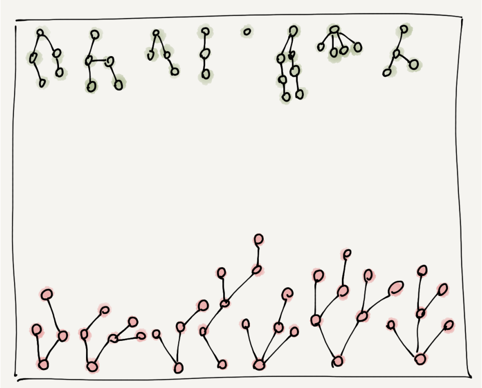

Fig. 3. Graph representation where true and false information grow from opposite ends.

We brainstormed for alternative visualization ideas and worked on preprocessing scripts and techniques to reduce the size of our raw dataset. We came up with a new idea of representing each information as a circle, where the radius of the circle represents the magnitude. Unlike the original idea where the size of the graph implicitly represented the magnitude of the information (since the more edges and nodes it had the more space it took), the new idea explicitly recorded the magnitude and did away with granular information such as edges between individual tweets. One of the feedback we received was that this kind of visualization is not dynamic enough and boring to watch. So we brainstormed on new visualization ideas that kept the concept of the circles but added motion. We experimented with a number of

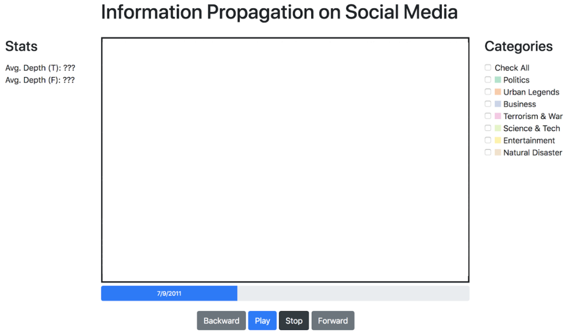

Fig. 4. User interface.

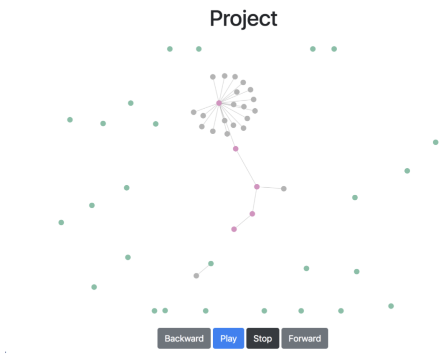

Fig. 5. Graph representation.

concepts and landed on a concept of bubbles.

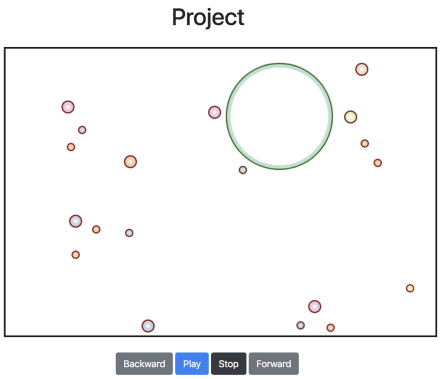

Fig. 6. Circle representation.

The idea was to think of each circle as a bubble. It would pop into existence and as its magnitude increased, it would go higher. We saw that this approach had two main advantages. First, it was fun to watch the bubbles rise. Second, it organically separated the large bubbles from the small ones, reducing visual clutter and improving readability of the visualization. Once we landed on this idea, we focused on improving it, reiterating on small details such as the rate at which bubbles grow and time increment step. The final big improvement we made was eliminating the radio buttons for choosing which categories to show on the visualization. Instead, we separated the large canvas into seven distinct boxes with each box representing a different category. In the final visualization, which is a web application, the viewer can control the speed, play/pause, and reset of the visualization. Bubbles of different color pop into existence when first mentioned on Twitter, and grow as more people tweet about them. Green bubbles represent true information, red bubbles false information, and yellow ones mixed. It is easy to compare online information space across different categories, too.

4 Peer Feedback

Our target audience being the general public, the feedback we got from our peers in class was not a perfect match to our intended audience since there is low variance of age and education gap within the group. In order to deal with this problem, we made our questions simple and easy to understand. We did not ask any technical questions as we wanted to ask questions that anyone could answer without prior knowledge or experience.

We asked two questions during the peer feedback session:

1. Of the two different visualizations (circle version and graph version) which do you prefer more and why?

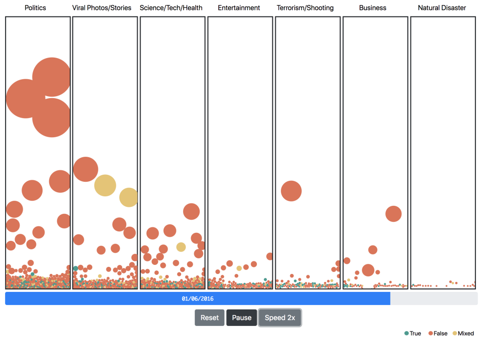

Fig. 7. Bubble representation.

2. What is your first impression of the visualization? Was there anything unclear or confusing?

For the first question, we prepared two different visualizations. Version (a) [Fig. 6] represents each piece of information as a circle, and it grows in size as it is talked about more or retweeted. Version (b) [Fig. 5] represents each tweet as a node, branching out of a central node which represents a piece of information. The latter version also allows users to expand/contract each node. We wanted to understand which visualization people would prefer more and why, and also whether interactivity was important.

We received a lot of good feedback from our peers. The preference for each visualization was even. People who preferred the circular representation liked the fact that the visualization was simple to understand and effective at communicating the overall information landscape on Twitter. They thought that the graphical representation made the visualization needlessly complicated. Many also expressed that they did not find the interactivity important to our visualization as they were more interested in seeing the general trend play out. On the other hand, the people who preferred the graphical representation thought that the structure that this view presents provided key insight to our research question. They felt that having the branching nodes showed exactly how information spread, instead of just showing its magnitude However, they also agreed with the first half of the people on that the particular interactivity we had was not something that they thought was important to our visualization Some people suggested that it would be useful to have a hover behavior for each node so that it would display more detailed information only if the viewer wanted to dive deep.

For the second question, we wanted to get a general sense of people's reaction to our visualization. We wanted to see what confused people about our design, what they didn't like about it, and how they thought that it might be improved.

We got a lot of good feedback for this question as well. We had forgotten to label our True/False/Mixed colors, because we thought it was intuitive, but we realized that people were initially confused by lack of context in which they appeared. They told us that it was unclear which color was which and were able to put things together after our initial brief on what our visualization was. Another feedback we got was that the design of each circle/node was too confusing (we had used two colors on each circle/node: one color for veracity and another color for category). Lastly, one concern that many people shared was that it could get boring to watch the visualization. For both versions, the viewer is looking at either circles growing or nodes popping in and out. This was particularly useful and we focused our efforts on dynamism and fluidity with force behavior to make our visualization not just informative but also fun to look at.

Some future work remains. First, we want to give the viewer more control over the visualization. Currently we have made a design decision on how much time should increase on each animation step (0.5 week/frame). We want to create an input field where the viewer can specify the time-step. Another mode of control that would be useful is having a slider for controlling the time. Currently there is no way of going back in time except for clicking the reset button or refreshing the browser A slider that allows the viewer to freely move back and forth in time would be a useful improvement. Second, as useful as our dataset is, it has its limitations. One such limitation is that it is fully anonymized. If we get to collect the data ourselves by scraping Twitter along with fact checking websites, we would still anonymized the data for viewer privacy reasons, but we could keep what each information id represents (i.e. what each bubble represents). Currently, there is no way to say for certain what each bubble represents. We can only make educated guesses. We think our visualization can provide more insight with this additional layer of information. Third, we think this visualization is suited for live data. Because of the sheer volume of the number of tweets generated every second, we would need to implement some kind of sampling algorithm and live updating mechanism that pipes the sampled and anonymized data to the visualization. Fourth, we want to add an option to let the viewer choose from alternative color schemes. We were careful to choose the colors that we chose in order to not create any political links. We did not use red and blue because of their political symbolism. We want to create other color schemes for two reasons One is to improve accessibility for the visually challenged. Another is to have a color combination that brings out the true bubbles more, because in current the version the green bubles can be hard to see on small screens. We believe these four additional features are a good next step for us to take beyond the scope of this class project.

5 User Study

5.1 Target Participants of the User Study

We want a group of participants from the general public. Ideally, we would want a distribution of demographics similar to ones of people that read news articles regularly. This could maybe be handed through a survey sent to individuals that hold physical or online subscriptions to a news source. Older people are more likely to have a print subscription, and younger people are more likely to have digital subscriptions. However, we want to sample from both to ensure that our visualization is easy to understand. There is a relatively even split between digital and print for the age bracket of 35-49 year olds, and it would be a nice demographic to ask questions about our visualization. We think that getting participants from just off the street could be very effective as well. To get our target demographic from off the street, we could stop someone with a paper on the train, or someone picking up a paper at a coffee shop.

5.2 Data collectors

We think responses to a more open ended user study will be more effective to make our visualization easier to understand. These sort of questions dont necessarily require an expert (familiar with UI design concepts) in the field. We would prefer the person collecting the data be conversational, and friendly to allow for more free flowing and genuine responses. This should also help a participant with sharing their feedback, rather than just answering a few predetermined yes/no questions.

5.3 Where & When

Because we think our visualization fits well in a news setting, we want our user study to take place in an environment similar to where people spend time reading news. We suspect that our few sample questions should not take more than a few minutes to answer, and because we would sample random people from off the streets, we probably do not want to exceed that.

5.4 Experiment Details

We do not think that a comprehensive user study on our visualization would require too much. The content does not require much in depth knowledge, and is directed towards the general public to answer the question of "how different rumors spread". We think our visualization would fit really well in an article that explores the spreading of rumors online and because of this we are directing our user study towards the general population.

We have three specific questions that we will ask participants to gauge their initial impressions and understanding of the content without the full information. It is important that the questions are asked in this order.

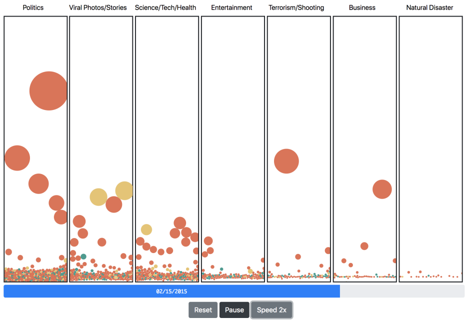

Fig. 8. Image to be part of the user study.

1. Given the above snapshot, with a title "Rumors", what would you assume this snapshot is trying to show you?

2. If I also told you that a each circle represents a rumor that was spread online, what conclusions can you draw?

3. The source of our data set concluded that false rumors spread a lot quicker than true rumors, and often spread out to a lot more people. Does this finding surprise you given what you have seen from the visualization?

Furthermore, we would like to have a more relaxed portion where we can get honest feedback about what a participant did not really understand without more explanation. We would not place any time restrictions to answer any of the questions. Our goal is to find out how clear our visualization is.

5.5 Data Collected

We want our responses to be as open ended as possible, so we want to be able to acquire written/verbal responses. If the data were to be collected in person, an interviewer with a voice recorder asking about the visualization would suffice. If we wanted more responses from more people, we could get responses from a mailed response (physical & digital), through a survey including the snapshot with the first question and a suggestions box to make our visualization more clear (after informing the participant with the answer).

On top of the open ended user study, we would use web analytics tool to evaluate how the participants interact with our visualization. Because our visualization is web-based, we are able to taken advantage of many existing software that can do a range of tracking: click behavior, mouse movement, gaze tracking. Of course, we would make our participants aware that we are tracking their behavior while they are participating in our user study.

5.6 Cost & Time Concerns

We think that the cost of the user study can be conservative. But because of the conversational and open-ended questions that we want to ask, it might take a lot of time to conduct the study and gain quantifiable feedback from them. On the other end of the spectrum, automatically collected data will be easier to work with so we expect that aspect to take significantly less amount of time.

5.7 Goals and Desired Outcome

Our goal in this study is to find out how the general public might interact with our visualization. We want to find out whether we have successfully communicated what we wanted to communicate, and also find out what aspects of our visualization that might be confusing.

6 Features & Implementation

We think our visualization has three core features. First, we think we were able to create a visualization that is intuitive and dynamic. We included labels where necessary and kept it simple. For example, we made use of basic shapes and only used three colors. Dynamism adds visual intrigue without being ostentatious and distracting. This is important because we want to maintain the simplicity of our visualization to keep it easy to understand, but also not too blend and boring for the viewers. Second, we think our visualization is a good way to provide a birds eye view of online information space. It only takes a glance to get a gist of the shape of the online information space being visualized. There are only two dimensions color for veracity and size for magnitude. Third, we think our visualization is good for observing surge events. Surge events are events where there is an explosion of information. For example, in our visualization one can observe a rise of few but large bubbles during election seasons and terrorist attacks.

We used vanilla Python without any other libraries for analysis and data preprocessing One challenge with our visualization was the size of our dataset. Therefore data preprocessing was necessary to reduce the size of our dataset. We incorporated three key techniques to reduce the size of our data from 500mb to 50mb, reducing the loading time from minutes to seconds and making our visualization accessible to more devices. First technique was to encode data into a more memory efficient format Some examples of this is changing strings to ints and dates to unix timestamps. Second technique was to remove whitespaces. This is a very simple thing to do, but it saved a lot of memory space. CSV file format divides each row by a whitespace, so we switched to JSON with no whitespace. Third technique proved to be the most effective. We split the raw dataset into two different files. One file contains the sequential data with an id, and the other file serves as a lookup table. Because the original dataset contained repeating information, we were able to save a lot of space by removing these repeated information and replacing them with a new short unique id.

For implementing the user interface, we made use of the Bootstrap CSS library. This library was helpful in making the user interface nice to look at and flexible to screen size. For implementing the visualization algorithm and animation we used the d3 library, specifically d3.forceSimulate(), d3.interval(), and d3.circle(). They were used to simulate bubbles colliding and floating, animate specific frames over time, and draw circles respectively.

7 Division of Work

After the brainstorming session where we decided on roughly what we wanted the visualization to look like, Jerry worked on python scripts for analyzing the data we had gathered together and Justin worked on the user interface using HTML/CSS/Javascript. In order to present two different versions for the peer review session, we worked on each in parallel: Jerry worked on the graph version and Justin worked on the circle version When it became clear that we should go with the circle version, Jerry and Justin brainstormed on how to make it better and add animation. Both worked on the final version of the visualization, building on top of our previous works. The presentation and the write-ups were a joint effort.

References

[1] S. Vosoughi, D. Roy, and S. Aral, The spread of true and false news online, Science, vol. 359, no. 6380, pp. 1146-1151, 2018.

[2] D. A. Szafir, Modeling Color Difference for Visualization Design, in Proceedings of the 2017 IEEE VIS Conf.: IEEE Transactions on Visualization and Computer Graphics, 2018. Available: http://cmci.colorado.edu/visualab/VisColors/. [Accessed: March 23, 2018].

[3] M. Harrower, and C. A. Brewer, ColorBrewer.org: An Online Tool for Selecting Color Schemes and Maps, The Cartographic Journal, vol. 40, no. 1, pp. 27-37, June 2003.

[4] H. Guo, J. Huang, and D. Laidlaw, Representing Uncertainty in Graph Edges: An Evaluation of Paired Visual Variables, IEEE Transactions on Visualizations and Computer Graphics, vol. 21, no. 10, pp. 1173-1186, 2015.

[5] E. R. Gansner, and Y. Hu, Efficient, Proximety-Preserving Node Overlap Removal, in Journal of Graph Algorithms and Applications, vol. 14, no. 1, pp. 53-74, 2010.

[6] C. C. Gramazio, D. H. Laidlaw, and K. B. Schloss, Colorgorical: Creating discriminable and preferable color palettes for information visualization, in IEEE Transactions on Visualization and Computer Graphics, vol. 23, no. 1, pp. 521-530, 2016.

Acknowledgments

We wish to thank Soroush Vosoughi et al., the authors of the paper of the data this visualization is based on, for generously sharing their data with us.