Gina Stalica

Data Driven Societies

Final Paper

#Tor: Visualizing the Anonymous Web

Introduction

Tor – originally an acronym for "The Onion Router" – is a government-created project that uses layers of encryption to allow users to browse the Internet anonymously (Grossman 2013). Though its original uses were for the security benefits of government activity, it is now used much more widely as a portal to what is known was the "Deep Web," the portion of the Internet that cannot be reached by common search engines (Grossman 2013).

For some time, I have been interested in taking a look at the Silk Road – an underground, anonymous, online drug market – I had never even heard of "Tor" until I began to actually research the Silk Road. After a quick Twitter search and a Google scan, I came across a Forbes article by Andy Greenberg, explaining the current status of the Silk Road 2.0. I quickly learned that Tor is the name of the anonymity tool that allows the Silk Road to function and, thus, allows criminals to conduct illegal transactions (Greenberg 2013). Needless to say, I was excited to learn more about Tor's implications and thus decided to take on an exploration of the use of the hash tag "#Tor" on Twitter in order to analyze the discussion of the anonymous Web that is taking place online.

Research Question

As I started this project, I did not believe that I had a complete understanding of the Deep Web's underground community. While I understood its basis, I was eager to learn more about the Silk Road and its other functions. Overall, I was confident that #Tor would yield interesting ideas and information about the Deep Web. For this reason, I was able to postulate my research question: Is there a clear discussion or Tor that is occurring online, specifically on Twitter?

Early Hypotheses

Before having extracted any tweets, I formulated a few hypotheses. I expected to see tweets from all sorts of sources – from media figures to everyday people giving their input on Tor – and hopefully some from users themselves. Ultimately, I did expect that I would see mostly non-participatory input from computer- and Web-interested tweeters, simply commenting on the wonder of the anonymity power of the Silk Road and the "Deep Web" as a whole. I expected them to focus on the latest updates of the status of the Silk Road and the controversy that anonymous Web use inevitably sparks within the United States.

Literature Review

Today's academic literature regarding Tor and the anonymous Web largely focus on the current weaknesses in Tor's anonymity and the future of the network as a whole. Adversaries observe other users' activity before entering the Tor network in order to eventually locate said users (Johnson 2013). According to the analysis of Johnson, et al., 80% of Tor users could be made not anonymous over the time span

of six months (2013). This number was derived from simulations run on the actual Tor network using realistic models that allow for the evaluation of the security of the network. Johnson essentially invites Tor users to reconsider how useful Tor may actually be (2013).

A seemingly hot topic among the academics of online security is the idea that the future of Tor requires more relays, or more volunteers to serve as nodes for the network's layers of encryption. Westermann et al. cite motivating users to volunteer to serve as relays – components of the "layers" of Tor's routing encryption - as one of the largest issues regarding Tor that currently exits (2011). Their paper suggests splitting Tor's entire network into subnetworks (Westermann, et al. 2011). Unfortunately, through analysis of this proposal, it is ultimately concluded that this splitting would likely lead to great imbalances in the size and speed of the operating subnetworks, leaving the authors with no real solution in the end (Westermann, et al. 2011).

Alternatively, Androulaki et al. suggest another method: providing users compensation for volunteering to serve as relays in the Tor network (2008). This method calls for a payment method that utilizes hybridization of payment using Peppercoin Micropayment and a new type of electronic currency (Androulaki et al. 2008). Though a bit far-fetched, this idea could plausibly function within the Tor network, potentially boosting relay participation, improving the network's functioning overall (Androulaki et al. 2008)

A much more basic – and much more realistic – way to increase participation is increased publicity (Lawrence 2014). It can certainly be argued that Tor only receives its most press following negative events, including criminal activity with use

of the Silk Road (Lawrence 2014). Along these lines, "hacktivist" groups like Anonymous – groups that use code and the Web to perform actions from pranks to human rights activism – use Tor as a means of protecting their identities (Coleman). Even these groups carry a slightly negative connotation with them, as most people who have heard of their activity jump right to the thought of their malicious acts. If some of this press could be turned positive, showing the positive light of Tor's ability to shield users' location from other dangerous Web-users, perhaps Tor could draw more relay participation (D. Lawrence 2014).

Methods of Data Collection

In order to collect Twitter data, I utilized ScraperWiki's Twitter tools to extract public tweets including "#Tor" from January 28, 2014 to February 25, 2104. Once compiled, this provided me with 5,553 tweets – a number much lower than some of my other classmates' data sets, but a number much larger than any data set with which I had previously worked. When seeking another data set, I turned to Tor's website, where I found an array of different Tor metrics that are free and available for use and analysis (Tor 2014). I decided to download the file of all Tor client requests from 2011 to the present day. At this point, I was able to conduct statistical, graphical, spatial, text, and network analyses on both data sets in order to learn more about Tor. I was able to use both R and Microsoft Excel to perform graphical analyses; Wordle to perform text analysis; Social Explorer, CartoDB, and BatchGeo to perform spatial analyses; and, finally, Gephi to perform a network analysis of my Twitter data set. Some of these visualizations proved exciting and useful, while others seemed to bear little revelation.

Analysis

Analyzing my Twitter data set in terms of representation turned out to be a bit more complicated than I had anticipated. The key sentiments in my data set are related to issues of privacy. The key words and phrases that kept showing up along my hash tag include "malware," "deep Web," "hack," "NSA," "privacy," and "anonymity." Most of these tweets express opinions on the issues of Web privacy that are currently in question. Though I was excited to keep exploring the progression of my data set to seek some answers to these questions, I soon realized that these answers would not come straight from my Twitter data set.

Unfortunately, Twitter limits my ability to represent what interests me about my hash tag. For example, plenty of my tweets are unrelated to the Tor Internet platform. The hash tag has also been used to replace "Toronto," the name "Tori," and several alternative topics. This brings into question the issue of "data cleaning." According to boyd and Crawford, the "data cleaning process" is defined by "making decisions about what attributes and variables will be counted, and which will be ignored." (2012) Because I could not locate the point at which I should filter through my data set, I decided to refrain from doing so. For this project, this would have felt as though I had gone against proper research methods.

Furthermore, it is argued that Twitter cannot provide truly representative samples, especially seeing as only sixteen percent of online adults in the United States make use of Twitter. (Crawford 2013). Speaking more specifically to my hash tag, it could be of concern that the most interesting conversations about anonymous Web occurring on Twitter might very well be taking place in interactions between users of private accounts. This is certainly plausible, as Tor users must value privacy

and anonymity at some level. Ultimately, I was able to approach the issue of proper representation by acknowledging and accepting that my data set cannot possibly be a complete, unbiased, heterogeneous sample of conversations about Tor that are taking place.

Findings

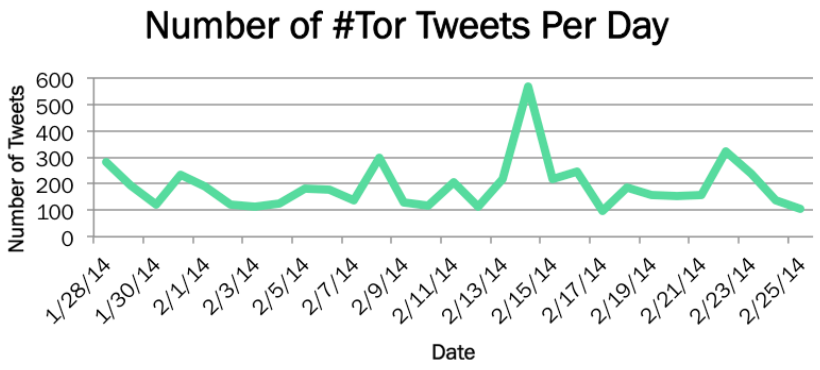

Figure 1. Number of #Tor Tweets Per Day. (Source: Stalica, 2014.)

When I first began to analyze Figure 1, I was sure that the variations in the number each day could be attributed to day-to-day fluctuations or mass retweeting on a given day. After doing some research, I have concluded that this is true, but the fluctuations also work alongside some notable events on certain days. I found that on February 13, the day that my number of tweets jumps to about 550, the Tor project released its newest version (Tor 2007). I then continued to search the dates for events that may have occurred on days with noticeable spikes in number of tweets. It was frustrating to find that on February 8 and February 23, two other days with a considerable spikes in tweets, the spike in tweets were likely due to Toronto sports events – what appears to be a hockey game on the 8th and a basketball game

on the 23rd. These were certainly events that were far less exciting to discover, however, they brought me back to the question of data cleaning. Perhaps it would have been more beneficial for me to somehow have "cleaned out" my scraped tweets, but I am ultimately not sorry that I was unable to do so. My data has revealed that Tor and Internet anonymity are being discussed online all across the world, whether Toronto wins its next hockey game or not.

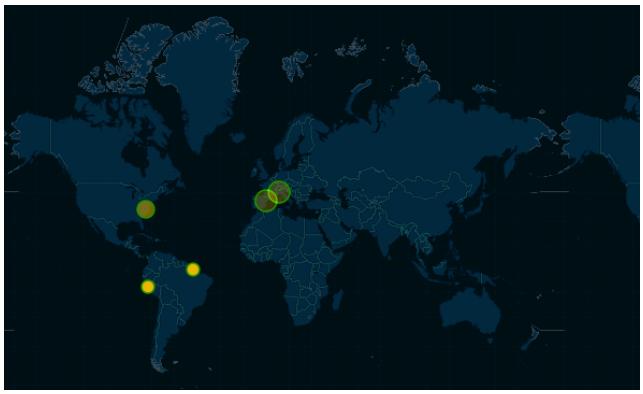

Figure 2. #Tor Tweets Over Time. (Source: Stalica, 2014.)

Next I began to spatially analyze my data, first using the CartoDB platform. Figure 2 visualizes the georeferenced #Tor tweets I extracted over the course of my data collection period. While I did not extract quite as many georeferenced tweets as I would have hoped, I found the timing of the tweets to be quite interesting. For some dates, tweets seem to come from areas across the globe – for example, the northeastern United States, Europe, and the Middle East– simultaneously. This felt like a breakthrough; I had extracted tweets that could be related to a political event involving Tor or the release of a Tor update. At the same time, my randomly dispersed tweets reminded me that my research could certainly be strengthened with

more data.

Figure 3. Georeferenced #Tor Tweets. (Source: Stalica, 2014.)

The map I created using BatchGeo proved to be most interesting to me due to the fact that it further proved one of my initial assumptions to be wrong. Figure 3 clearly shows that discussions of Tor on Twitter are occurring globally, not only in the United States. While a high count of my georeferenced tweets were concentrated in the northeastern United States, I also found concentrations of tweets in Europe, South America, and the Middle East. After noticing the spread of languages from my dataset of all of my scraped tweets, I am no longer surprised by this. I am, however, surprised that the tweets containing #Tor that were georeferenced in the United States came exclusively from the northeast. While New York and Boston are both certainly home to such hubs, California is as well, and I expected to scrape tweets from Silicon Valley and its other nearby hubs. This concentration can perhaps be attributed to Lessig's idea of "east coast code" and "west coast code" (2006). "East coast code" describes government-created code that regulates online behavior as a method of control; whereas "west coast code" describes coder-created material that essentially creates the cyberspace we know and use today (Lessig 2006). The

falsification of my hypothesis is supported by Lessig's claim that "east coast code" is growing in power over "west coast code," which is to say that government control seems to be overtaking coder freedom (2006). Whether or not this is exactly what my Twitter locations represent or not may still be in question; however, I was excited to find that Lessig's work seemed to support my findings.

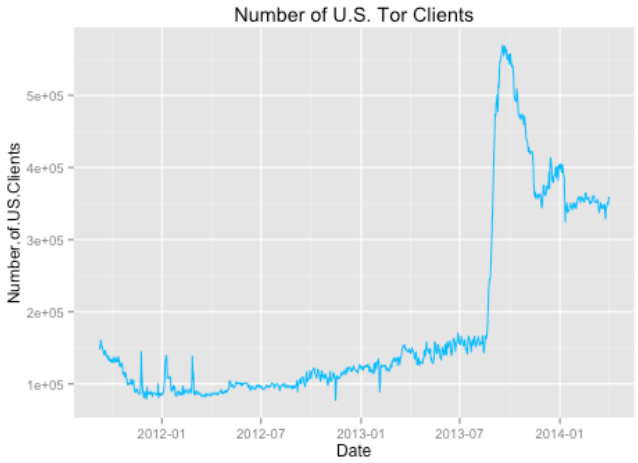

Figure 4. Number of U.S. Tor Clients. (Source: Stalica, 2014.)

When I first began analyzing my data set of all global Tor requests since 2011, I expected to see a fairly steady but small increase in Tor usage over time. What I found was something far more interesting. Figure WHAT FIGURE displays the dips and dives in Tor user requests from January 1, 2012 to the present day. The dramatic increase in user requests – peaking on September 15, 2013 - shocked me. I was certain that a major event must have inspired such a quick spike.

Interestingly enough, after doing some research, I simply could not find one single event that inspired the spike. It appeared that even Tor's creators were baffled. After getting in contact with a Tor employee, I was finally introduced to their

current explanation: a botnet (arma 2013). Their proposal is that millions of malware-infected computers had Tor installed, making for an increase in Tor requests that could not even be explained as having a human user behind each computer client (arma 2013). This revelation was particularly alarming because I had not considered such an extreme and unusual event to be the reason for the spike in client requests. Though this explanation helped me to understand my data, it also helped me to see that the Tor network is somewhat mysterious due to its anonymity and its usage is, too.

At this point, I was able to conduct a T-test to compare the number of client requests for 198 days before the spike on September 15, 2013 and the number of requests for 198 days after the spike using R. The p-value I extracted – a number so small that R could not even print its value – was far enough below five percent that I was able to conclude that there is, in fact, a statistically significant difference in the number of Tor client requests before and after the observed spike. Needless to say, this statistical analysis provided further support for the ability of bots and other

external noise to interfere with the analysis of Tor data.

Figure 5. Network Analysis of the #Tor Twitter Data Set. (Source: Stalica, 2014.)

The network created using all of the #Tor tweets I have scraped has proven to be drastically different from what I expected. My initial expectation was that my network would be clearly divided into subgroups either composed of tweets that were about Tor and privacy or tweets that were about other topics – whether that meant Toronto hockey game scores or shout-outs to friends named Tori. Instead, I have found that my tweets are actually more divided into either subgroups based on highly retweeted users or singletons. Almost all of the connected components greater than size two are based around users such as "@YourAnonNews" or other Internet news sources or forums. This proved to tell a much more interesting story than I had expected. It now seems clear to me that there is, in fact, a clear online discussion of Tor and, more broadly, issues of privacy and anonymity. Had I not visualized my

tweets using a network analysis, I am not sure that I would have been able to draw such a clear conclusion. While my various graphical analyses showed me a story in the changes in Tor use, my network analysis showed me much more about the actual dynamics of the discussion of Tor that is taking place on Twitter.

Discussion

Tor is being discussed online and offline, but perhaps not most productively on Twitter. My graphical analyses show that "#Tor" is being used on Twitter, but with a certain amount of variability in frequency that sometimes, but not always, aligns with relevant events. My hopes were that I would find this Tor press – whether good or bad – cultivating within the Twitterverse. I did find that Tor is being discussed, but not with as much coherence as I might have hoped. After conducting some research outside of Twitter, it was easy to see that, as word spreads about the anonymous Web, academics, hackers, and everyday people are discussing its capabilities and risks. Whether these conversations occur on the Tor site's blog or in any number of online news sources, a quick Google search yields an array of material regarding Tor.

My graphical analysis of the number of Tor client requests tells a story far more interesting than I had originally anticipated. I would never have known the true driving force of the statistically significant increase in Tor clients observed in September in Figure 2, had I not reached out and contacted Tor developers via email. This pushed me to think more abstractly about my research process. It forces me to consider seemingly out-of-the-ordinary events that might inspire changes in both my Twitter data and Tor usage, speed, and other data.

Along the line of interference, some of my data visualizations mask actual stories regarding Tor due to interference of external content. For example, in Figure 1, we can see spikes in tweets per day due to basketball games. Ultimately, I feel that my network analysis of my Twitter data set tells the clearest story of Tor in relationship to Twitter, making my most compelling data visualization technique. My network shows that there are conversations, however concentrated, about Tor occurring on Twitter, specifically. While many tweets including the hash tag Tor may not be connected, there certainly are connected components that cultivate relevant conversation.

At the same time, my research has shown that Tor's creators and employees are working hard to positively publicize the network (Heffernan 2010). Jacob Appelbaum, one of Tor's developers, explains that he is trying to spread the network to users by helping people to see it as easy to use (Heffernan 2010). This seems particularly relevant to my data set, specifically my network visualization, in which Appelbaum's Twitter handle, "@ioerror," is a centrally located node for one of my network's smaller connected components. This provided me with a direct connection between my background research on Tor and my data findings. Twitter is certainly being used as a platform for – at the very least – causal discussion of Tor, even involving its head operators at times.

Conclusion

Overall, my findings matter because Tor and the anonymous Web is a relatively new topic and even data analysts had not been exposed to such information until very recently. When studying Tor and the Deep Web, it is absolutely

imperative to remember that there are limitations to trying to uncover the stories of anonymity. For this reason, I suppose that it is not completely surprising that my research has required a lot of outside exploration and the analysis of a data set other than my Twitter extraction.

In the end, my attempt was truly oxymoronic – I have been trying to study the public social media discussion of anonymous activity. Through the analysis of my Twitter data set and the exploration of other material regarding Tor and the anonymous Web, I have finally come to conclude that there is a discussion of Tor that is taking place both on Twitter and elsewhere online. These discussions, however, can be hard to locate and understand, as there is certainly an air of mystery surrounding the Deep Web. Whichever way its developers choose to motivate relay volunteers, I predict that Tor is a tool whose use and discussion appears to have no end in sight.

Works Cited

Anrdoulaki, E, Raykova, M, Srivatsan, S, Stavrou, A, & Bellovin, S 2008, 'PAR: Payment for Anonyous Routing', *Lecture Notes in Computer Science,* vol. 5134, pp 219-236. Available fro: Springer Link [10 May 2014].

arma 2013, How to handle millions of new Tor clients. 15 September 2013. *The Tor Blog.* Available from <https://blog.torproject.org/blog/how-to-handle-millions-new-tor-clients>. [10 May 2014].

boyd, d & Crawford, K 2012, 'Critical Questions for Big Data", *Information, Communication & Society,* 15:5, pp. 662-679. [8 May 2014].

Coleman, G 2012, 'Am I Anonymous?', *Limn,* vol. 2. Available from < http://limn.it/am-i-anonymous/>.

Crawford, K 2013, 'Think Again: Big Data', *Foreign Policy* 9 May. Available from <http://www.foreignpolicy.com/articles/2013/05/09/think_again_big_data>. [May 8, 2014].

Greenberg, A 2013, '"Silk Road 2.0'" Launches, Promising A Resurrected Black Market For The Dark Web', *Forbes* 6 November. Available from < http://www.forbes.com/sites/andygreenberg/2013/11/06/silk-road-2-0-launches-promising-a-resurrected-black-market-for-the-dark-web/>. [1 May 2014].

Grossman, L, Newton-Small, J, Roy, J & Stampler, L 2013, 'The Deep Web', *Time*, vol. 182, no. 20, pp. 26-34.

Heffernan, V 2010, 'Granting Anonymity', *New York Times* 17 December. Available from <http://www.nytimes.com/2010/12/19/magazine/19FOB-Medium-t.html?pagewanted=all&_r=0>. [8 May 2014].

Johnson, A, Wacek, C, Jansen, R, Sherr, M & Syverson, P 2013, 'Users Get Routed: Traffic Correlation on Tor by Realistic Adversaries', *Proceedings of the ACM Conference on Computer and Communications Security*, pp. 337. Available from: Scopus [10 May 2014].

Lawrence, D 2014, 'The Inside Story of Tor, the Best Internet Anonymity Tool the Government Ever Built', Bloomberg Businessweek 23 January. Available from <http://www.businessweek.com/articles/2014-01-23/tor-anonymity-software-vs-dot-the-national-security-agency>. [7 May 2014].

Lessig, L 2006, 'Regulating Code', In *Code: Version 2.0*, New York, Basic Books, New York, pp. 61-80. [9 May 2014].

Tor, 2014. Available from: <https://www.torproject.org/>. [7 May 2014].

Westermann, B, Chia, PH & Kesdogan, D 2012, 'Analyzing the Gold Star Scheme in a Split Tor Network', *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, vol. 96, pp 77-95. Available from: Springer Link [7 May 2014].