Do We Care About Surveillance?

Edward Snowden's Impact And Policy Implications

Arya Azma, Louis DeScioli, and Evan Marshall

Table of Contents

Introduction

In the years since September of 2001, the federal government of the United States has put boots on the ground in search of terrorists in the Philippines, the Horn of Africa, Liberia, Iraq, Afghanistan, Pakistan, the Sahara and Sahel, Yemen, Libya, and central Africa. Government officials have thwarted a number of attacks[1] on innocent United States civilians. Phone and internet surveillance have played a role in some of these investigations, including the investigation of Iyman Faris, who was tasked by Al-Qaeda leadership with the initial steps of severing suspension cables on the Brooklyn Bridge,[2] and the investigation of Dhiren Barot, who was convicted in the UK for terrorist conspiracies including a plot to destroy American financial buildings.[3] [4] However, some groups argue that the federal government of the United States has gone too far in its surveillance programs. Public outcry was especially strong in June of 2013, after the Guardian and the Washington Post simultaneously released materials leaked by NSA contractor Edward Snowden that describe unprecedented mass surveillance programs. A day after the release, President Barack Obama addressed the issue during a visit to Silicon Valley, saying, "it's important to recognize that you can't have 100% security and also then have 100% privacy and zero inconvenience. You know, we're going to have to make some choices as a society."[5] Whatever those choices may be, the policies that result will undoubtedly be influential to the development of the Internet and other communication technologies, consequential to the government's ability to maintain a secure nation, and momentous for American public life and civil liberty.

In the months since, the American people have responded in various ways to the surprising revelations that their internet data and phone activity is not necessarily safe from unreasonable scrutiny by intelligence authorities. Some have taken to social media and social news sites to discuss the Snowden revelations and familiarize themselves with the differences between PRISM, XKeyscore, Pinwale, Tempora, Boundless Informant, and other computer and phone surveillance systems. Others have gone a step further, not only discussing the topics of the Snowden leaks, but also the underlying concepts of network security, encryption, metadata, and internet privacy projects like Tor and PGP. A number have even implemented programs like Tor to protect themselves from government surveillance, and conversely, many have done none of the above. Knowing how these cohorts behave is a powerful step to making the policy choices about privacy and security to which President Obama referred in Silicon Valley.

Policymakers have competing responsibilities regarding privacy and security, and there appears to be no easy solution that satisfies all of those responsibilities. Absolutist belief in the inviolability of the rights enumerated in the Constitution may preserve the liberties that have made the United States conspicuous and admirable in its development as a free society, but it fails to

[1] http://www.washingtonpost.com/wp-dyn/content/article/2005/10/06/AR2005100600455_pf.html
[2] http://www.globalsecurity.org/security/profiles/iyman_faris.htm
[3] http://www.bloomberg.com/apps/news?pid=newsarchive&sid=av1E.SWFj7Mc
[4] http://www.theguardian.com/uk/2006/nov/08/topstories3.terrorism
[5] http://www.reuters.com/article/2013/06/07/us-usa-security-records-idUSBRE9560VA20130607

consider and counteract the very real and decentralized national security threat that has been thrust upon the United States in the first few years of the twenty-first century. Likewise, security activity that intrusively constrains the freedom of Americans to associate with one another and express opinions candidly, without fear of reprisal by the government, would be so destructive to liberty that it would replace the fear of foreign terrorists with a new fear of domestic, governmental ones. Arguably, those who wish to maximize only liberty, as well as those who wish to maximize only security, would leave little worth defending of our society. Rather than taking an absolute position, we seek to identify precisely how internet users have responded to mass surveillance revelations, and use the resulting knowledge to determine how lines can be drawn and balances struck between liberty and security in accordance with the preferences of the American public.

These responses can be characterized in a myriad of ways. Our work identifies three different stages of engagement with the topics of mass surveillance. The first stage is discovery and simple discussion of these topics. For example, a person may read an article about NSA surveillance and write a tweet saying, "Edward #Snowden on the run in Hong Kong." This kind of activity indicates the reach and impact of a news headline, press release, or other catalyst for discussion. Twitter is a good platform on which to examine this discussion stage, because the brief nature of its messages and the low familiarity needed for users to follow one another (versus becoming friends on Facebook, contacting each other via email, etc.) allows simple messages and ideas to spread quickly and broadly.[6]

The second stage is aggregation of information and discussion about the underlying concepts, indicating more than a passing interest in a news headline. For example, a person may write a comment on the social news site Reddit saying, "I'd be interested to know what you guys think is the best way for me to encrypt my web traffic." Encryption is related to aspects of the Snowden leaks, but not superficially, and a person who is curious about it is likely to have thought about internet security on a deeper level than simply reading a few online news articles or facebook posts by friends. Reddit lends itself to examination of this aggregation stage because it is a public, anonymous, easily accessible forum. Given the right search terms, Twitter is also useful for identifying whether discussions about news headlines are accompanied by discussions about related and underlying concepts.

The third stage is action based on these underlying concepts. For example, a person may query Google with the search terms "how to use Tor to be anonymous online," and navigate from there to the project website where Tor is available for free download.[7] This hit to the Tor project website, which we can observe via Alexa, reflects something deeper than discussion and aggregation of information. It suggests that the user may want to install Tor, or at the very least learn how it works. If many people are acting to improve their internet privacy, or awareness thereof, in the wake of a breaking story about surveillance programs, it is an indication of two

[6] http://www.soc.ucsb.edu/faculty/friedkin/Syllabi/Soc148/Granovetter%201983.pdf
[7] https://www.torproject.org/

things. First, there is a revealed preference for greater internet security than the status quo provides. Second, users that were previously satisfied with their internet security have found it necessary or desirable to further protect their data, which is perhaps an indication that their expectations of privacy for unencrypted data have shifted.

Before we characterize and discuss these three stages of response to revelations of mass surveillance activity, we first frame the policy context and legal significance that makes them important. Then, we review existing literature within the body of academic knowledge regarding the perceived impacts of internet surveillance on user behavior, including methods for gathering and analyzing datasets based on social media, forum comments, and web traffic. After that, we explain how we have chosen to collect (public and/or anonymized) data as the material for our analysis, as well as how we have chosen to analyze this data. Once the methods are clear, we indicate the results of our analysis, while also noting the limitations of our methods and the strengths and weaknesses of any causal arguments that may naturally flow from those findings. We then draw conclusions about how internet users respond when they find out that the government may very well be collecting data on their browsing activity and online communications. Finally, we indicate the next steps that would help to shed more light on this question, and check the strength of our conclusions.

Legal Context

We begin by framing our analysis with a discussion of the policy context and specific legal questions that make it relevant to the discourse about internet freedom and bulk collection of data by the federal Government of the United States. Even without a policy context or legal frame of reference, a wide and diverse audience is hungry for factual information about how internet users respond to news about national security surveillance activity. Not only is the internet novel and growing, but it is also so large that personal observation and anecdotal evidence alone are poor tools for really discerning its character and dynamics. However, there is a larger significance to internet usage facts, in that they shed light on the answers to various legal questions. Diligent legislators examine the behavior of individuals and cohorts affected by their policies at many stages in the policymaking process both beforehand, to determine the gaps and norms of the status quo, and retrospectively, to understand the strengths and shortcomings of their policies as revealed by implementation in the real world. Additionally, the Supreme Court has a history of using behavior-based notions to determine where to draw the line in questions of Constitutional importance. We are motivated by this second, judicial question.

Analysis of internet usage data can be used to apply a number of First and Fourth Amendment standards employed by the Supreme Court of the United States. The Court's body of precedent includes standards that are based not only on strict definitions, but also on behavior and expectations. For example, the belief that a particular policy gives rise to a chilling effect on free speech has led courts to strike it down under the First Amendment, even in the absence of an explicit restriction on speech. Similarly, the belief that a particular search or seizure violates privacy expectations has led the Supreme Court to deem it unreasonable under the Fourth Amendment, and to thereby exclude the evidence found in that search or seizure from consideration at trial. Analyzing Internet usage data can characterize the behavior of internet users and shed light on their expectations in regard to speech and privacy. Furthermore, one of the most contentious modern Fourth Amendment issues is defining the boundary between metadata and data. Today in the United States, network routing information is less protected than the contents of a data packet, but this standard is based on archaic telephone technology. With telephones, there is a clear separation between the electrical signals of a conversation and the electrical signals by which the call is routed. However, the internet does not have such clear boundaries between routing data and content data.

Are there chilling effects? Did there exist subjective and objective expectations of privacy that were violated without warrant? Where does the metadata evidentiary standard apply, and where does the content data evidentiary standard apply? With regard to the internet, these three questions are difficult to answer with a clear and consistent standard in the absence of data and analysis on the behavior of internet users. Before building and analyzing such a dataset, it is helpful to clarify these three questions one by one, as well as the standards that are used by the courts to answer them.

The First Amendment prohibits Congress from making laws that abridge the freedom of speech. With few exceptions, this language disallows explicit restraints on speech, but it also has been deemed to disallow laws that "chill," or implicitly restrain, speech. The chilling effect was first referred to by the Supreme Court in *Wieman v. Updegraff*, a 1952 case against mandatory oaths of abjurement that, in the words of Justice Felix Frankfurter, had "an unmistakable tendency to chill that free play of the spirit which all teachers ought especially to cultivate and practice."[8] A decade and a half later, in his dissent of the Supreme Court's decision in *Walker v. City of Birmingham*, Justice William Brennan pithily summarized the concept by describing the Court's "overriding duty to insulate all individuals from the 'chilling effect' upon exercise of First Amendment freedoms generated by vagueness, overbreadth and unbridled discretion to limit their exercise."[9] In the years since, the chilling effect has been used in a number of Supreme Court cases, generally to protect First Amendment freedoms that would be discouraged implicitly by a given law or policy.

It may well be the case that mass internet surveillance has a chilling effect on the free transaction of information on the internet. Innocent individuals may be so afraid of being watched or being falsely identified as criminals that they cease to exchange emails with people who may be scrutinized by the government, or to discuss or search for information about sensitive topics such as leaked cables, pressure cookers, or terrorist organizations. Alternatively, it may be the case that internet surveillance by a distant and hidden entity has no effect on what the typical citizen does with his web browser and email client. It may even be that awareness of surveillance encourages a surge in potentially precarious activity. Without a data-driven empirical assessment of internet behavior, judges can do little more than make educated guesses and philosophical arguments as to whether mass surveillance and bulk collection of data deter free speech and association on the internet.

The Fourth Amendment affirms the right against unreasonable searches and seizures. An intricate network of ideas and precedent that define this Amendment has been woven together by the courts in the years since the Bill of Rights was enumerated, and the applicability to the internet on any number of those ideas and cases could fairly be addressed. We choose to focus on two: expectations of privacy, and the line drawn between content data and metadata.

In *Katz v. United States*, Justice John Harlan wrote a concurring opinion indicating a standard for reasonable expectations of privacy, by which to determine if a warrantless search is unreasonable. "My understanding of the rule that has emerged from prior decisions is that there is a twofold requirement, first that a person have exhibited an actual (subjective) expectation of privacy and, second, that the expectation be one that society is prepared to recognize as 'reasonable.'"[10] This standard was referenced and applied by Justice Harry Blackmun in the

[8] Wieman v. Updegraff, 344 US 183, 195 (1952) (Frankfurter, J., dissenting).
[9] Walker v. City of Birmingham, 388 US 307, 344-45 (1967) (Brennan, J., dissenting).
[10] Katz v. United States, 389 U.S. 347 (1967) (Harlan, J., concurring).

majority opinion of *Smith v. Maryland*,[11] and remains a part of Fourth Amendment doctrine. To apply this doctrine to web traffic one has to ask whether an individual expects privacy in a particular communication, and whether society as a whole considers that expectation of privacy to be reasonable.

Data analysis can help to answer both prongs of this question. Aggregate data about web traffic patterns and search terms can show whether people generally changed their online behavior in a more risk-averse or encryption-seeking way after learning of web surveillance programs being conducted by the government. Sharp shifts may indicate that they had subjectively expected a more private internet prior to knowing about the availability of their information to government analysts, but became more cautious quickly afterward. Such shifts in behavior, displayed across a broad swath of internet users, would also indicate the reasonableness of those expectations by society. One may also argue that an expectation of privacy is objectively reasonable due to the secrecy of NSA surveillance operations, existence of relatively client-friendly privacy policies, protections in place for content data, and court cases like *United States v. Warshak*, where warrantless searches of internet activity were ruled unconstitutional by the Sixth Circuit Court of Appeals.

Another concern under the Fourth Amendment is the distinction between data and metadata. The Court in *Smith v. Maryland* opined that individuals do not have an expectation that the numbers that they call using the telephone will remain private because they are being handed to a third party, the phone company, which will certainly log some or all of those calls for billing and other internal purposes. This evidentiary standard for telephone communications has been applied to internet communications. Data about the traffic of a single machine may be accessed without warrant, but the content of that traffic may be considered protected. This is problematic because whereas telephone communication has a clear and operant difference between routing data and content data, internet communication affords no such natural distinction. A URL is certainly a piece of routing information, but it also may include search terms, usernames, and other data that goes beyond simple routing. This particular example is common practice for Twitter, Google, and many other major websites. In such cases, the bright line between metadata and content data is as narrow as a forward slash.

Web traffic analysis can help to determine the correctness of this distinction between types of data with different evidentiary standards. One can determine whether internet users see their metadata in the same light as their content data by seeing whether they seek to encrypt it. Multiple methods for encryption exist, and not all of them provide protection for metadata. By examining the difference in percentage increase in traffic to sites where users can acquire metadata-protecting encryption protocols such as Tor and I2P, versus sites for non-metadata-protecting encryption protocols such as PGP, one can see if internet users reveal a preference for better protections for their metadata. If so, then it may be the case that the

[11] Smith v. Maryland, 442 U.S. 735 (1979) (Blackmun, J. opinion).

metadata standard inherited from the pen register question of *Smith v. Maryland* (1979) is the wrong fit for the routing data of today's internet. If people flock to metadata encryption, it is a natural indication that they view this data as content.

The internet is so widespread and diverse that without data analysis any of these three standards would be difficult or impossible to apply by reasoning alone. However, data analysis is impactful in this legal context. It can clarify whether there is a chilling effect among internet users as a result of surveillance; it can determine whether internet users exhibited a subjective expectation of privacy that changed due to the revelation of surveillance policies; it can even identify whether internet users have flocked to metadata-protecting encryption mechanisms, revealing a preference to protect their metadata as though it were content data. These are questions that cannot be answered by deep thought alone, no matter the wisdom and judicial experience of the thinker.

Literature Review

It is important to support each phase of the process independently, analyzing how social media supports and enables the discovery, aggregation, and action phase of a response.

Discovery

At this point in time, in the year 2013, it is clear that people use social media to talk about news and current events. It's commonplace to read a headline in the newspaper, or more often, on another website, and share it on Facebook, Twitter, Reddit, or any other number of social media services. The mass sharing that happens through social media enables original opportunity for research that were not possible before, such as analyzing sentiment of a response to a news story[12] or event. Although Twitter is known for being home to a very vocal minority, it is still possible to get an accurate sample size from top tweets[13]. The speed of the technology and the breadth of the usage of social media across the general public provides original and powerful possibilities for analyzing the nature of sharing, and has been used time and time again the past few years.

Discovery Alternative Discussion

Analyzing social media to learn about the popularity or discontent about a story or event is better than the alternatives used in the past. Strictly looking at the number of news stories written about a subject, or measuring the airtime that a story gets on television or radio are top-down approaches that don't accurately capture the viewers feelings about the story. Widespread polling and monitoring, such as Nielsen ratings for television or Gallup for national polls, do a better job of a user or viewer-centric approach to analyzing response and sentiment, but are slow and still limited in their scope.

Research & Aggregation

Social media also provides a mechanism for people to aggregate information and learn from one another. Like-minded people posting about their activities, hobbies, or interests naturally attracts others to learn about those pursuits as well. As people work on those activities and come across challenges or breakthroughs, they are likely to make posts about those proceedings for others to assist in or benefit from. This is the essence of any online forum or mailing list, and has existed in some form for 30 years.

---

[12] Mitchell, Amy; Hitlin, Paul. "Twitter Reaction to Events Often at Odds with Overall Public Opinion," Pew Researcher Center, March 2013.

[13] Himelboim, Itai; McCreery, Stephen; Smith, Marc. "Birds of a Feather Tweet Together: Integrating Network and Content Analyses to Examine Cross-Ideology Exposure on Twitter." Journal of Computer-Mediated Communication, January 2013, Vol. 18, No. 2, pp. 40-60. doi: 10.1111/jcc4.12001.

Modern social media consists of large communities with standardized-yet-complex structures, making it easier to analyze than the mailing lists, bulletin boards, or forums of the past. Studies have leveraged this to evaluate the movement of information aggregation through a network of people. One such study showed that weak social ties among online users is primarily responsible for the large scale dissemination of new information[14]. This is highly applicable to studying a community such as Reddit, where users are mostly anonymous and avoid tying their online identity and friends to their offline lives. It is also highly applicable to Twitter due to the public nature of the community. Since users can follow or be followed from any other user, weak ties are very common. The study shows that while most interaction in a social network occurs between users with strong ties, weak ties are what drive the propagation of information throughout a large network.

One other study shows that while users involved with movements propagated through social media mobilization often overestimate the efficacy of the group, large information gathering helps smaller subgroups succeed[15]. This is another item of support for using Reddit, where communities are capable of easily fragmenting into smaller subreddits, or groups, and using the information provided by the larger subreddits to continue to learn and make progress while avoiding their slow and cumbersome nature.

Action

Having shown that social media enables people to share information, research and aggregate new information, it logically follows that is also moves people to take action. When a certain threshold of excitement and interest is reached, online social media moves people to take offline actions. This has been documented by analyzing campaigns to increase voting[16], analyzing the adoption rate of an application or plugin[17], and to fight unpopular laws, such as SOPA and PIPA [18]. The social media presence in conflicts such as the Arab Spring and revolutions in Egypt in 2012 are also a testament to the powerful nature of social media to bring people to act.

[14] Eytan Bakshy , Itamar Rosenn , Cameron Marlow , Lada Adamic, The role of social networks in information diffusion, Proceedings of the 21st international conference on World Wide Web, April 16-20, 2012, Lyon, France [doi>10.1145/2187836.2187907]
[15] A. Rutherford et al., Limits of social mobilization,Proceedings of the National Academy of Sciences.110(16) 6281-6286 (2013).
[16] Bond, Robert M., et al. "A 61-million-person experiment in social influence and political mobilization." *Nature* 489.7415 (2012): 295-298.
[17] J.-P. Onnela and F. Reed-Tsochas. Spontaneous emergence of social influence in online systems. Proceedings of the National Academy of Sciences, 107(43):18375--18380, 2010
[18] http://www.journalism.org/2012/01/26/social-media-win-big-one-washington/

Methods

With the goal of identifying quantitative effects in social networks and various privacy protection related applications, we decided to collect numerical data to quantify social media responses and the traffic data associated with the appropriate applications. Analyzing social media reveals how information spreads and collects, such as informing users about the Edward Snowden leaks, or identifying the spread of related privacy applications and tools like Tor. We planned to use the traffic data to represent the individual behavior changes as individuals learn about and potentially adopt new tools to protect their privacy.

We first began by collecting social media responses to the Edward Snowden leaks. Our goal was to quantify and distinguish the response and general informing about the leaks, and information sharing about potentially useful privacy related tools. We specifically chose to analyze Twitter and Reddit as our social media responses.

Twitter and Reddit

We used social media data to inform our analysis because of the large number of users and the relatively wide availability of data. The highly structured nature of the data also facilitates the possibility for in-depth network and sentiment analysis that would be impossible with other tools. Also, social media analysis has garnered much attention for ability to produce useful insights for academic studies[19] and large brands like Apple[20]. Social media analysis continues to grow in legitimacy and in use as the number of social media users swell.

Despite these benefits, social media often may the disadvantage of not truly reflecting the demographics of society. Although an important consideration, social media analysis still provides useful insight into the ways that real world events engage particular communities. The Pew Research Center conducted a survey in March 2013 about the differences between Twitter reactions and the general public opinion and found some significant discrepancies[21]. The study went further to analyze the particular communities engaged on social media and found that despite overall differences, the sentiment of these online communities largely reflected the opinion of their offline counterparts. By analyzing social media, we can thus essentially get a reasonable indication of what the online community thinks even though real world discrepancies will exist.

With the decision to use social media made, there are only four social media outlets with a large enough user base to consider worth analyzing: Facebook, Twitter, Tumblr, and Reddit. Other

[19] A collection of Twitter Studies can be found at:
http://journalistsresource.org/studies/politics/campaign-media/us-government-twitter-research
[20] Apple purchased the Twitter Analysis Firm Topsy for $200 million
http://online.wsj.com/news/articles/SB10001424052702304854804579234450633315742
[21] The Study can be found at:
http://www.pewresearch.org/2013/03/04/twitter-reaction-to-events-often-at-odds-with-overall-public-opinion/

social media sites with similar page ranks like Pinterest and Instagram were not considered because of they focus on types of social interaction that would not be useful for the purpose of this study. Tumblr also turned out to be a similarly inappropriate media to discuss politics and encryption. Although Facebook is used to communicate political information, limitations on data collection including scope and amount made Facebook data relatively difficult to collect and even more difficult to verify the accuracy. Since Facebook does not have a feature to search across all posts that mention Edward Snowden or Tor, we would have had to randomly collect information from individuals and hope that it was representative. Twitter and Reddit avoided these particularly large disadvantages even though Twitter provided challenges in the data collection.

Twitter is a microblogging platform that is used to share short messages (140 characters or less) across potentially vast networks of users. About 15% of American adults actively use twitter regularly with about 8% of Americans using Twitter on a daily basis[22]. When significant news develops occur, the information is soon after spread through the Twitter network, quickly informing large numbers of users. Historically, Twitter has also facilitated the gathering and planning of special interest groups for everything from winning Internet competitions to protesting in Iran. Overall, Twitter was chosen for its wide usage, relatively representative demographics, and its established use in the literature.

Reddit is a social news and entertainment website where users submit either link or text based posts. About 6% of Americans use Reddit on a regular basis, making Reddit the 5th most popular American social networking tool[23]. Reddit's demographics are highly skewed to younger males with interests in humor, technology, and politics. The close similarity between the demographics of Reddit and the likely adopters of encryption or other security applications makes Reddit a potential source of vital insights into information aggregation through social media. The ability of users on Reddit to gather related comments of arbitrarily long lengths enables users to form coherent, in-depth conversations. We chose to analyze Reddit because it is a likely source of information gathering for privacy related applications.

[22] http://pewinternet.org/Media-Mentions/2012/Study-Number-of-Daily-Twitter-Users-Have-Doubled.aspx
[23] Duggan, Maeve, and Aaron Smith. *Six Percent of Online Adults Use Reddit*. Rep. Pew Internet, 3 July 2013. Web. 1 Nov. 2013.

Search Keywords and Methods

In order to gather information from Twitter and Reddit, we had to be able to find the appropriately related tweets and Reddit posts. We selected a list of primary terms that would characterize and differentiate related and unrelated content. Here is the list of terms we decided to search for:

nsa, *prism*, *snowden*, *encryption*, *privacy*, *whistleblower*, *tor*, *vpn*

Using the search features built into Twitter and Reddit, we filtered appropriate tweets and Reddit posts, checking for these keywords to include any particular content in our dataset. These keywords were chosen as they closely relate to the topic at hand. More importantly, content mentioning any of these terms serves to shed light the social response to the Edward Snowden leaks.

After defining our list of keywords to distinguish relevant content from noisy chatter, we implemented code in python to scrape the specified tweets from Twitter and posts from Reddit. We specifically built a python system on top of the WebKit technology to allow our scripts to emulate a browser to continuously load content from Twitter and Reddit. It is important to note that we choose to avoid using the application programming interfaces (api) for both Reddit and Twitter.

In the Twitter case, we avoided using both the Search API and the Firehose API. The Search API sets very strict limits on the number of requests and provides almost no guarantees on the availability of historical data. The Firehose API would provide many more tweets than we could collect otherwise, but it lacks the ability to collect any tweet before the date the collection begins. The lack of useful features for both APIs proved very important reason for us to collect tweets directly from the pages, scrolling down to load more tweets as if a user was requesting them.

In the Reddit case, the API provides similar restrictions, prohibiting more than 100 posts for each search. With a user emulated search, we were able to collect around 1000 posts per search, greatly expanding our dataset. We also found that we could add in additional secondary terms to our primary terms to further increase the number of relevant searches we could perform. By creating longer search queries, we were able to identify relevant posts that otherwise might have gone missed due to the 1000 posts per search limit.

Avoiding these APIs and employing our own script allowed us to collect about 15,000 tweets from Twitter and about 44,000 posts from Reddit. We filtered the two datasets for the primary terms and for duplicates, producing the stated numbers. For each tweet we collected the number of retweets, number of favorites, the username of the tweeter, the content of the tweet, and the tweet's timestamp. We focused most on using the number of retweets and favorites as a measure of the tweet's popularity and reach. For each Reddit post, we collected the number of

upvotes, the number of downvotes, the overall post score, the poster's username, the content of the post and the timestamp of the post. We chose to use the sum of upvotes and downvotes of the post to determine its popularity and reach.

Twitter and Reddit Data Limitations

Despite a good sample size and filtration of data, there are some inherent limitations to the insights provided from the data. Most importantly, we rely on the assumption that the search features on Twitter and Reddit return the most popular, relevant tweets/posts. It is possible that the searches from either social media contain gaps in time by leaving out older posts or even some sort of down scaling by recency. Gaps in content are also possible as a result of relying heavily on very specific keywords to search for content. It is possible that many relevant posts were left out of the dataset as a result of not containing a specific keyword. For example, our data collection process would have prevented a tweet mentioning Edward, Russia, Plane Travel, asylum from entering our dataset even though it seems at least moderately, if not highly related to the topic at hand.

Another important limitation to note includes the inability of our data to accurately represent the real magnitude of a response. Instead the data can only provide relative insights because of the millions of possibly relevant tweets not contained in our dataset. Thus our dataset may provide useful analytics for the relative magnitude of an event, but we cannot use the data to accurately predict the number of informed users. Preemptive use of the Firehose API for Twitter might provide such estimations in future studies.

Privacy Applications Data Collection

After collecting social media data from Twitter and Reddit, we gathered the traffic information for specific applications likely to be used by people seeking to protect their anonymity. Traffic data provided a more reliable measure than the actual application use which could be artificially inflated by botnet piggybacking as was the case with the huge increase in Tor usage starting in mid-August (0). We used the Alexa Web Information Service hosted on the Amazon Cloud to gather the traffic data. Alexa is the best source of traffic information on the Internet with rankings for more than 30 million websites. Alexa collects traffic information by receiving data from Internet users who install the Alexa toolbar or one of many other toolbars that tracks browser usage and reports the data to Alexa. Alexa collects usage data from tens of millions of internet users around the globe. More recently, Alexa has introduced a javascript plugin that web developers can install to a website to provide even more accurate analytics. As of 2013, about 9 million unique visitors go to the Alexa website per month, helping to establish it as the leading authority on Internet traffic estimation and rankings. Research through Alexa helped determine the most relevant and widely used applications where newly informed privacy seekers might attempt to find tools to protect their identities and data. We first had to identify the websites that we wanted to measure. We choose to measure the traffic incoming to the Tor Project , TrueCrypt,

GNUPG, and Silent Circle websites.

Tor

The Tor Project website (www.torproject.org) was chosen because it was a tool directly connected with Edward Snowden (Snowden used Tor to communicate with Wikileaks Lawyers) and because it remains one of the most widely used privacy protection tools. The Tor protocol works by using onion routing which adds layers of encryption at each server that relays the TCP packet. The Tor protocol effectively anonymizes the identity of the computers on the network, making it difficult for spying techniques to identify anyone. The Tor Project website specifically manages the release and download of the Tor Browser as well as insights and data related to the Tor network usage.

When choosing how to measure the increase of Tor usage, we first investigated the possibility of using the actual network data to estimate usage increases. We eventually discarded this data as it became widely known that a botnet network moved on to the Tor network two months after the Snowden leaks, rapidly increasing the usage of Tor. Although Tor usage might still have increased, it remains practically infeasible to separate the botnet usage from authentic users. Instead we decided to measure the traffic to the Tor Project website. Such traffic information provides a solid measure of general Tor interest and knowledge. This traffic data also avoids the problems associated with the Tor usage data because a botnet would have no reason to visit the website, risking exposure and identification of the source controller.

TrueCrypt

With almost 29 million downloads, TrueCrypt is one of the most popular freeware applications for on-the-fly encryption. TrueCrypt can be used to encrypt a file before emailing it or even encrypting a partition of a hard drive for continuous protection. Along with many other performance features, TrueCrypt consistently ranks as one of the highest quality and most used encryption softwares. Once again, we measured the traffic going to the TrueCrypt website (www.truecrypt.org) to understand if there was any increase in usage during the months following the Edward Snowden leaks.

PGP

The GNU Privacy Guard (GNUPG) is a GPL-licensed alternative and open sourced implementation of PGP. With a combination of public key encryption and symmetric key cryptography, GNUPG allows mostly Linux users to encrypt their communication with another user. The encryption program can also be used to encrypt emails and other file sending. Since the GNUPG only works from the command line and doesn't support an API, the program usage can indicate a higher level of learning of encryption techniques. There are also dozens of programs that act as an interface or wrapper for GNUPG allows multiple operating systems to support GNUPG.

Although another good indicator of privacy program adoption, the specific skill sets to use GNUPG make it a more difficult tool to use and perhaps more indicative of behavior switches by sophisticated internet users.

Silent Circle

The last website we chose to measure the traffic information for is Silent Circle. Silent Circle provides an easy to use commercialized encryption service. With the use an application available on the app store for every major mobile phone operating system, Silent Circle is one of the most popular paid-for encryption services. Unlike the other open source programs, Silent Circle was started by Navy Seals with expertise in encrypted communication. These former Seals then started the company to provide special ops levels of security to normal business operations and various government projects. Silent Circle provides an interesting point of analysis because of its close connections with the US Government. Any suspicion of governmental data sharing becomes ever more important with the shutdown of similar encrypted email services like LavaBit who closed down because of an unwillingness to cooperate with the US Government data collection programs.

Privacy Application Traffic Data Limitations

Our choices for potential applications sought out newly informed internet users trying to protect their data fairly represents popular tools and applications that users would pursue. However, the traffic data does impose some limitations. For example, there are numerous other possible explanations for any rise in traffic. Although a significant correlated response for all chosen websites might help eliminate uncertainty, other unknown causal relationships may still exist. For example, the botnet internet story probably caused an increase in traffic to the Tor Project website and we want to carefully construct our interpretations of the data to avoid misattributions.

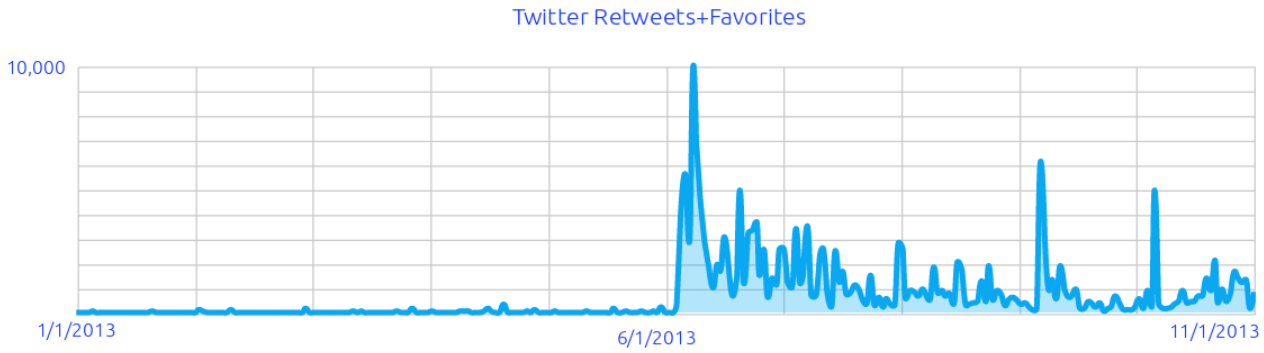
Other factors may still cloud the data by hiding the increases generated by the Snowden Scandal. Such factors may include a news break about how a particular service experienced a flaw and currently remains vulnerable to attack. With all the uncertainties about the causes of potential traffic data changes, it is important to aggregate the traffic information as a comprehensive picture of the Snowden Scandal responses instead of a single application gaining traction. Depending on the data, it may also end up being necessary to analyze other factors of website usage including time spent per visit, the number of unique visitors and the alike to accurately determine traffic change triggers. Despite these limitations on the insights we can generate from traffic data alone, the context of the social media data can provide useful insights for creating a broad but insightful interpretation of the reactions of internet users after hearing about the Edward Snowden leaks.

Results and Discussion

As described in the preceding section, we collected a set of posts and messages from Reddit and Twitter based on this small set of keywords:

nsa, *prism*, *snowden*, *encryption*, *privacy*, *whistleblower*, *tor*, *vpn*

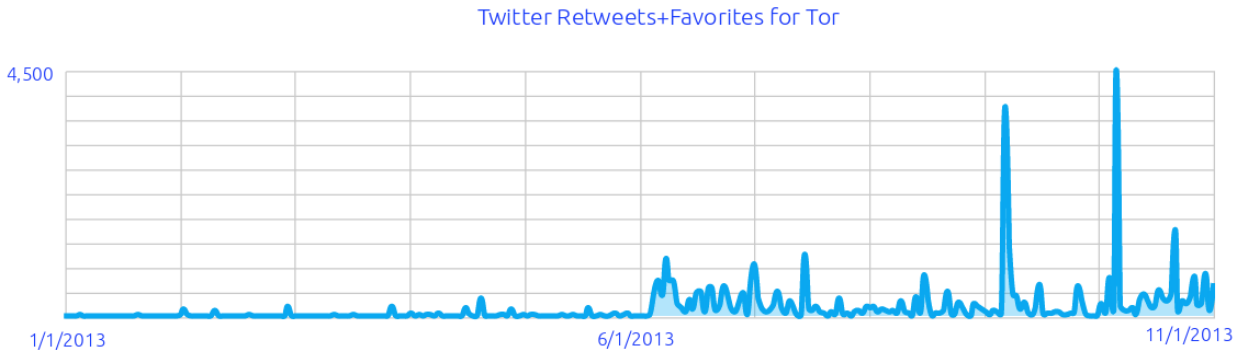
Overall, aggregation of tweets using this set of keywords resulted in the following time series, where the vertical axis shows reach for all tweets in the dataset:

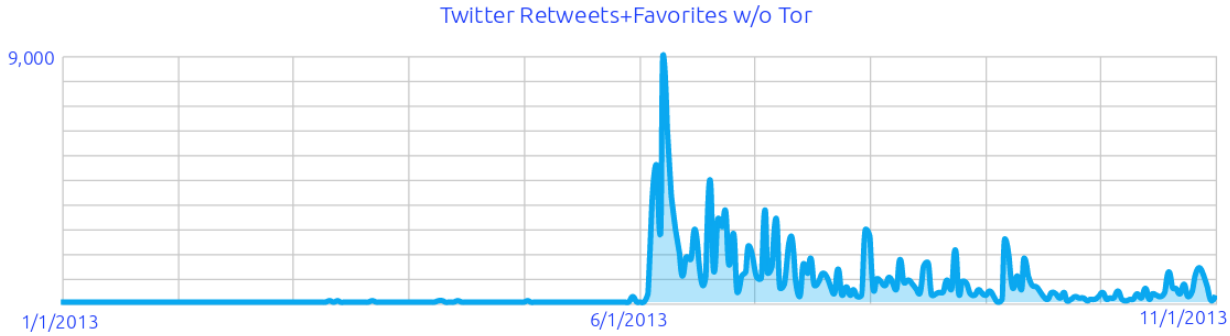


There is an extremely significant double-spike in Twitter activity across all keywords in the first few days of June 2013. Note that the first media disclosures about PRISM were published in the Guardian and the Washington Post on June 6, 2013, and that Edward Snowden's identity was published two days later, on June 8, 2013.[24] This accounts for the initial double spike. Twitter activity carries on with a decay that appears bi-exponential in form, generally attenuating through June and July, during which time Snowden traveled to Hong Kong and Moscow, and began filing asylum applications to a number of countries.[25] Another spike in activity occurs near August 1, which is the date upon which Russia granted Snowden a year of asylum.[26] These can all be considered as activity in the first stage of our three-stage model for internet activity, where news stories propagate and catalyze discussion of the most apparent topics.

Two more sharp spikes follow, one in early September and the other in early October. Separating the keywords, we find that these spikes are very likely due to discussion of Tor. Of the following two graphs, the first shows Twitter activity for the Tor keyword, and the second shows Twitter activity for all keywords except Tor.

[24] http://www.cnn.com/2013/09/11/us/edward-snowden-fast-facts/
[25] http://wikileaks.org/Edward-Snowden-submits-asylum.html
[26] http://www.nytimes.com/2013/08/02/world/europe/edward-snowden-russia.html?_r=0

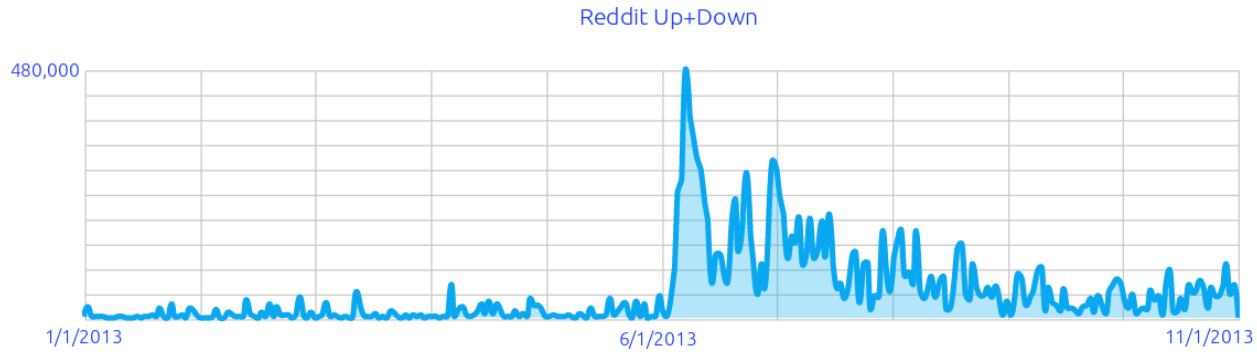




The September spike in Tor is accompanied by a soft spike in the other keywords, indicating a potential relation to the other keywords. Tweets from this time period include multiple messages exclaiming that the NSA can actually crack Tor communications. The October spike in Tor is very interesting due to its magnitude, but the lack of a covariant spike in the other keywords is peculiar. As it turns out, a Tor-based black market called Silk Road was shut down by the FBI on October 2, 2013,[27] which accounts for this spike. In general, comparison across the June 6 boundary indicates that there was certainly some interest in Tor by the Twitter community after the Snowden revelations. We categorize this interest in the second stage of our three-stage model for internet activity. However, the magnitude of this interest was slight in comparison to the discussion generated by the Silk Road shutdown. The only comparable spike attenuated quickly.

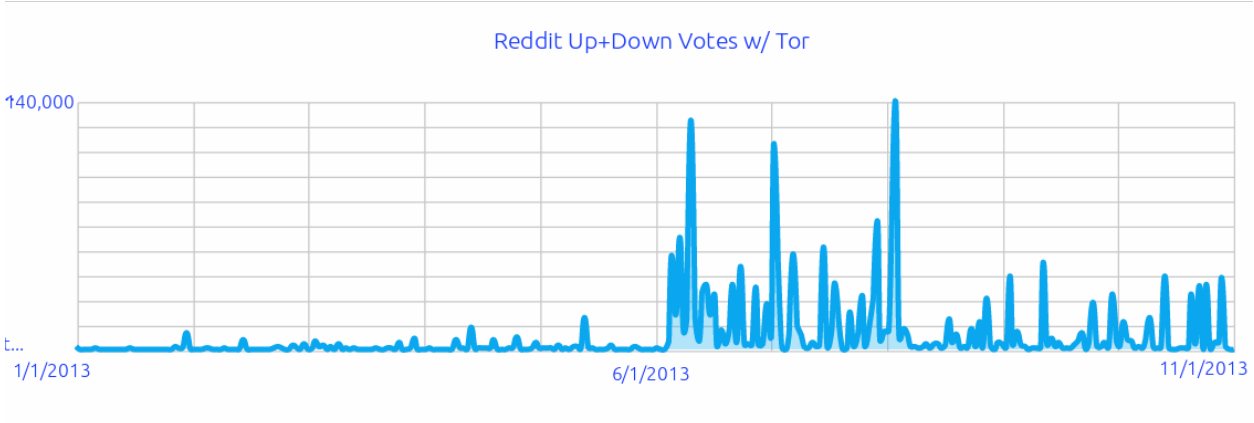
We continue our exploration of the first two stages, discussion and aggregation, with a look at Reddit. For all keywords, if we take all posts in the dataset and aggregate their upvotes and downvotes as a proxy for reach, we generate the following time series:

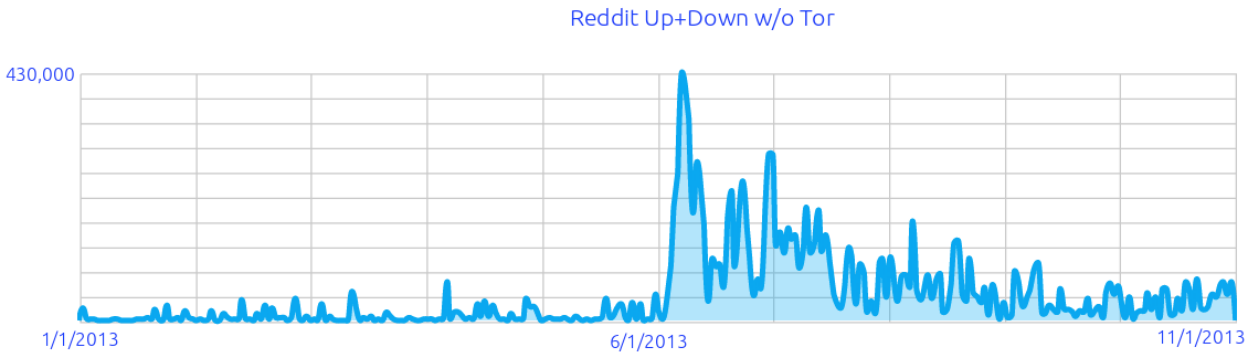
[27] http://www1.icsi.berkeley.edu/~nweaver/UlbrichtCriminalComplaint.pdf



Note that, as with the Twitter data for all keywords, there is a massive spike in activity about June 6 and June 8. Activity falls off for a short time near the end of June, then surges again by July 1. It is unclear what caused this surge. Some posts in the dataset encourage Reddit users to protest in various cities on Independence Day (July 4), especially in New York City and Washington, D.C. Many other posts around July 1 are in reference to press commentary and interviews, and tend to be more political than technical. A surprisingly large number of posts are self-referential discussions of Reddit's reaction to the NSA leaks. This indicates the same sort of first-stage (discussion) activity as shown on Twitter, as well as some political activism by Reddit users, which is arguably third-stage (action) activity.

In the following two graphs, we separate Tor from the other keywords, as we did with the Twitter data. The first graph shows Reddit activity for the Tor keyword, and the second graph shows Reddit activity for all keywords except Tor.

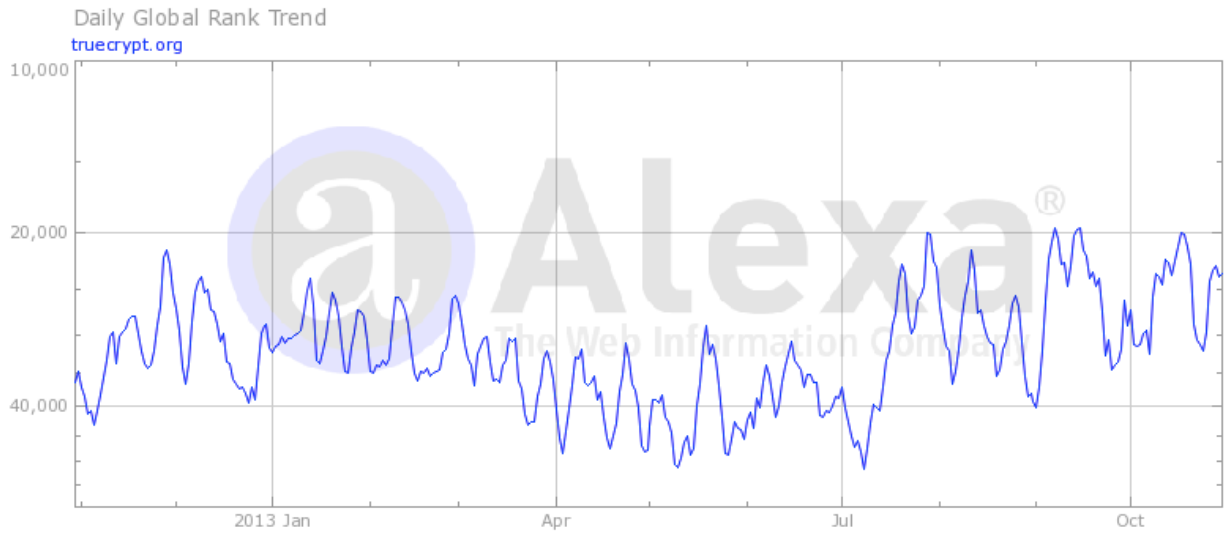




Both graphs show a stark difference in activity across June 6, 2013. Note, however, that activity for all keywords except Tor changes in a smoother fashion than activity for the Tor keyword, which shows three major spikes that all fall between early June and early August. The first two of these spikes are mirrored in the non-Tor keywords, and the first spike appears to be a product of the initial Snowden revelations around June 8. The second spike does not have an analogue in the Twitter data, and is potentially just a period of intense attention on Tor, including some of its technical aspects, by small clusters of Reddit users.[28] [29] [30] [31] [32] [33] The third spike, around the beginning of August, is perhaps attributable to the takedown of Freedom Hosting, a Tor hidden site, on August 4th.[34] [35] Regardless of these three spikes, it is clear that subreddits (topic-specific subforums of Reddit) for Tor saw a sustained increase in activity after the Snowden release, and that a non-negligible portion of this activity was technical in nature. This suggests that some people were curious enough about encryption to go beyond the content of news headlines, searching online and asking questions about encryption mechanisms. This led to the aggregation of a body of information on Tor on Reddit.

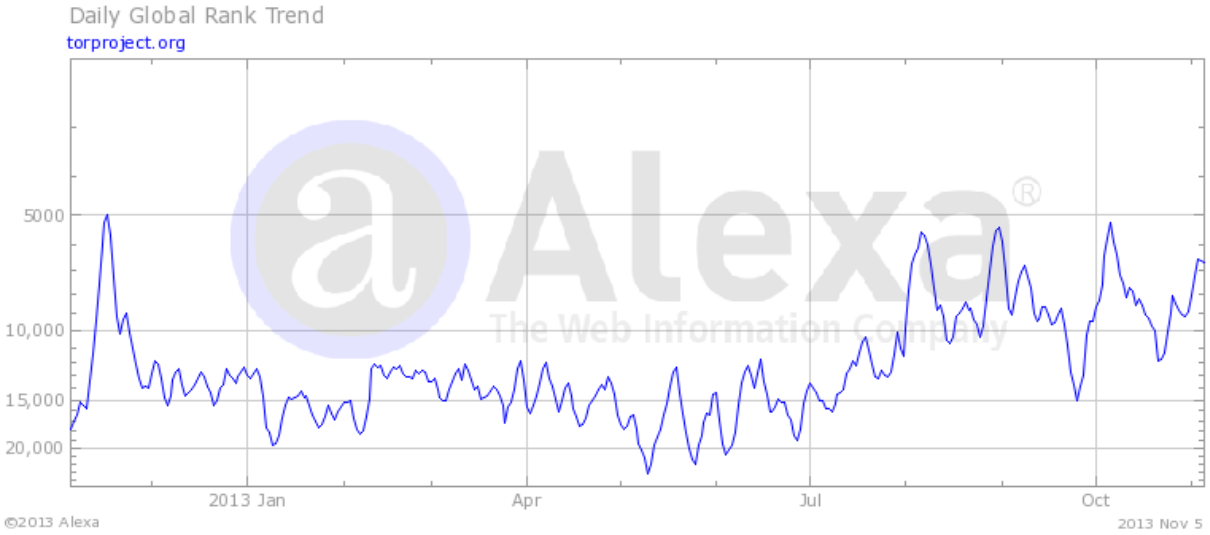
In order to examine the third stage (action) of our three-stage model, we shift our attention from Reddit and Twitter to Alexa, a service for tracking web traffic. First, we look at the page ranking of TrueCrypt, an on-the-fly disk encryption freeware program, across time:

[28] http://www.reddit.com/r/TOR/comments/1he4b3/do_you_run_a_tor_relay_if_so_since_when_where/
[29] http://www.reddit.com/r/TOR/comments/1hbbs3/can_i_use_the_tor_as_a_gateway_for_all_my_lan/
[30] http://www.reddit.com/r/onions/comments/1hft2m/how_can_i_host_a_website_on_the_tor_network/
[31] http://www.reddit.com/r/onions/comments/1h9se8/can_you_use_tor_on_your_mobile_device/
[32] http://www.reddit.com/r/onions/comments/1h6g3v/ive_always_like_the_idea_of_anonymous_browsers/
[33] http://www.reddit.com/r/onions/comments/1hkafa/easy_guide_for_a_linux_nonexit_relay/
[34] http://www.theguardian.com/technology/2013/aug/05/tor-deep-web-servers-offline-freedom-hosting
[35] https://blog.torproject.org/blog/hidden-services-current-events-and-freedom-hosting



The daily rank does not appear to fluctuate much outside of its drifting variance. There is a sharp increase in rank, from about 50,000 to about 20,000 in the month of July 2013. Whatever the case, it is clear that there is a general increase in pagerank from July 2013 onward. Next, we take a look at GNU Privacy Guard, an alternative to PGP:

Here, too, we see a sharp increase in pagerank. In this case, the jump is from an unmeasured rank over 100,000, to a peak rank just under 40,000. This is a substantial jump, and as with TrueCrypt, it takes place shortly after the first few days of July 2013. Finally, we study the pagerank time series for the Tor project website:



The Tor project enjoys a high pagerank throughout the time series, but like the other two encryption websites, it sees a marked increase in pagerank from July 2013 onward. Indeed, the pagerank improves from around 15,000th at the beginning of July, to a high of almost 5,000th by October 2013. In fact, this is not a fluke in the data. According to Roger Dingledine, the project leader for Tor, their usership -- that is, the number of Tor clients running -- doubled from August 19 to August 27.[36]

Across the board for encryption websites, it seems that July was the beginning of an increase in interest from internet users. It is certainly possible that this increase is exogenous to the Snowden leaks — although the likelihood that page ranks would improve by a factor of two or three by unnoticed factors seems small, it is not necessarily safe to assume a causal relationship whereby this increase is due entirely to traffic from concerned netizens. However, if a significant component of the increase in pagerank for encryption sites in July 2013 is, in fact, due to interest stirred up by news about NSA programs and globetrotting leakers of classified information, then this is a very exciting finding.

If the increase in encryption site traffic from July onward is due in part to an increase in awareness and curiosity resulting from the Snowden leaks, it means two things. First, it means that the third stage of our three-stage model, in which internet users take action to improve the privacy of their data, is strong. Second, it means that there is a lead-lag effect of about one month between the dissemination of information (first stage in our model) about NSA programs, and the actions that internet users took to realign their use habits in an equilibrium where they increasingly protect themselves from government intrusion by using encryption. This indicates that there is a time in which internet users researched and learned about encryption programs like Tor. Indeed, our dataset shows that the aggregation of information about Tor on Reddit was highest in June and July of 2013. The increase in pagerank for Tor and other encryption sites began in the middle of that time period.

[36] https://lists.torproject.org/pipermail/tor-talk/2013-August/029582.html

Further, there may be direct implications about user preferences on metadata, as well. Of the three encryption mechanisms that we examined, the Tor project enjoys the most popularity and the largest increase in pagerank from July onward. Tor is the only one of these three software programs that encrypts internet traffic metadata to a useful degree. TrueCrypt is fundamentally a mechanism for encrypting storage devices. GNU Privacy Guard encrypts email content, but does not preclude surveillance experts from eavesdropping metadata.[37] This is due to the way that SMTP works.[38] Together, this points toward the likelihood that Tor is preferred because it not only protects the content of messages, but it also obscures the path that a particular packet takes along the network. In other words, Tor protects the routing data by which surveillance experts could otherwise piece together networks between machines, and thus presumably between individual people. At a high level, interest in protecting routing data reveals that the internet-at-large is willing to treat what the courts deem to be "metadata" as though it were content data, too sacred to be observed by interloping third parties.

* * *

This analysis is motivated by three specific legal questions. First, does internet usage data indicate a chilling effect due to surveillance programs conducted by the federal government of the United States? Second, do people change their online habits as a result of surveillance revelations, in such a way that indicates a subjective expectation of privacy prior to realizing that organizations like the NSA conduct scrutinous surveillance of internet traffic? Third, does internet usage reveal a preference to treat metadata more similarly to content data than the precedents applied from telephony would require, or vice versa?

Indeed, the results of this analysis contribute to the process of answering these questions. An affirmation of the first question, as to whether a chilling effect exists due to surveillance, would suggest by First Amendment doctrine that surveillance curtails free speech. One can make the argument that increases in encryption activity reflect a need for more advanced technology in order to maintain the same level of free speech as would be available in a surveillance-free Internet. However, for a chilling effect to be shown empirically, an observable decrease in discussion must take place. If anything, discussion of NSA surveillance and Edward Snowden's actions increased due to the leaks of June 6, 2013. It may be the case that other types of speech were chilled, but political speech about surveillance activity was definitively emboldened by the leaks. Thus, a chilling effect does not appear to exist.

The second question asks whether a subjective expectation of privacy is implied by changes subsequent to surveillance. If people do not encrypt in the status quo, and then you introduce an

---

[37] http://www.theguardian.com/technology/2013/sep/30/email-surveillance-could-reveal-journalists-sources-expert-claims

[38] https://tails.boum.org/doc/about/warning/

intrusion into their privacy, the magnitude of their reaction to that intrusion indicates an expectation that it is unjust and unexpected. Here, we definitively see that, in at least the final two stages of our three-stage model -- aggregation and action -- internet users tend toward a rejection of intrusions into their privacy. Thus, one can reasonably argue that the surveillance programs violated a subjective expectation that those intrusions would not occur. By Fourth Amendment doctrine, this suggests that NSA surveillance programs, taken as a whole, violate the civil liberties of the U.S. persons whom they are used to monitor.

The third question asks whether the boundary between metadata and data, as inherited from telephony and applied to internet communications, is a bright line properly placed. Given the higher take-up of Tor versus two other encryption programs, our analysis suggests that internet users have a revealed preference for more protections on their metadata than the courts presently afford. Thus, based on this analysis, the line between internet metadata and content data may be improperly placed, and ought to be shifted in favor of warrant protections for what is currently considered metadata. Realistically, the boundaries between metadata and content data on the internet are unclear wherever they are drawn. A URL, for example, can include both a domain name which may be considered routing data, as well as search terms or usernames, which are components of the conversation between client and server machines. Additionally, the subject line of an email may contain non-routing information, and if unencrypted, the content of that email may be as accessible for surveillance as its origin and recipient. Not only are the lines between metadata and content data on the internet hazily drawn, but they also appear to fall in the wrong place based on the revealed preferences of internet users. Thus, it may be the case that warrantless searches of what the courts deem internet metadata, actually result in the warrantless disclosure of the content of communications between two machines or two individuals.

Next Steps and Conclusion

In addition to the research we have done, many related inquires may be addressed with a few follow up studies. Three such issues include: the specific types of events for which these steps and variants occur, the quantification of results, and the ability to predict events in real time. By answering some of the related questions, real policy and other impacts may be realized. The subject remains young in its study and requires further inquiry.

By determining the existence of separate phases in online response to the Edward Snowden leaks, it becomes important to understand the types of events for which these phases occur and any other variants. Many specific questions may rise independently such as did Bradley Manning provoke a similar internet effect and does this process apply to other subjects like online corporations? Many of these questions may have answers with political or monetary value. If it becomes realized that many internet users adopt or leave online applications in a similar process, the value of quantification rises significantly.

After understanding the various subject matters that we can study though this news disseminating , information aggregating, and behaviour shifting phases, it quickly becomes important to quantify the effects. Comprehending the numerical implications of real world events may help researchers develop political and legal models for understanding the interface between social media and the real world. The key to these models relies in large data collection as well as statistical and computational techniques. Utilizing big data algorithms may assist future study with greater depth and accuracy. For example, if a group captured the entirety of Twitter for an extended enough period, machine learning techniques may be able to predict changes in application usage given a Tweet and the number of retweets and favorites. With such prediction abilities, a group may quickly learn how to foretell the stage of progress in real time.

As the online experience becomes an ever more important part of the average American's life, the cyberspace and real world may more significantly affect each other. Even new financial companies such as SNTMNT thrive by using sentiment analysis to buy and sell stocks as well as provide informational services to much larger firms. Timing is a very important aspect for many of these applications. Such ability to quantify data, especially timing and geographic, enables real-time predictive models. Political leaders may utilize sophisticated services in the future to create stories that propagate to the internet, shifting the tides of public opinion. Social media has already shown its power to significantly change the course of bills in the US Legislature as evidenced by the failure of SOPA and PIPA.

As companies in the tech industry demonstrated by leveraging social media against the entertainment industry, proper understanding of the timing and dynamics may help avoid public apathy and exhaustion for a cause. Real world events do seem to bounce into the online world and then propagate back out again. The structure of the internet and its protocols make social media a

system that can be manipulated, hacked, and gamed. We may immediately conclude that understanding the steps of social media response, its quantification and its progression in real time may aid both system designers and the people that hack them.