Privacy Protection

Social Construction of Privacy: Reddit Case Study

A Thesis Prospectus Submitted to the

Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements of the Degree
Bachelor of Science, School of Engineering

,

Technical Project Team Members

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Signature Rajiv Sarvepalli ______________________________ Date 12/4/20 _____

Approved ____ Date __________
, Department of

Approved ____ Date __________
, Department of Engineering and Society

Introduction

The issue of privacy is an ever-growing concern in our information-driven society. Despite this growing concern, there is a lack of change from large technology companies and corporations to orient themselves around privacy, and therefore, most likely also a failure of individuals to economically pressure these corporations. Every year a new scandal is in the news about large social media companies inappropriately using data, yet all these companies continue to exist and gain large economic profits.   Therefore, there is not only a technological deficit for protecting privacy but more importantly a social one. In this paper, I will investigate the social construction of privacy using Reddit as a case study to understand the changing definition and perspective of privacy. The technical aspect will introduce a new way for individuals to protect their privacy by understanding the total information they give away. This technical solution to privacy does not counteract the greater social construction of privacy but provides individuals with the means of understanding the technical impact of their posts. In other words, this software will indicate to a user how much information a specific post gives away through analyzing how much location information is given away and how confidently can we identify who this person is. The STS related-topic and technical topic for Dept. of Computer Science are directly connected and will help enhance the understanding of privacy through different perspectives.

Technical Topic
(Project assigned by your home department)

Within anonymous social media, many users have no understanding of the potential impact of their words. Normally, people think about their posts or content put for everyone to see as only viewed by a human. As such, they interpret the amount of information given away in how much a human could interpret. However, with modern-day computing techniques, the information typical users put out can be used to identify due to a large amount of information on the web. While on most social media, people may be aware of this, anonymous social media users expect their accounts are unique and non-identifiable. By providing a software package that allows users to determine how much information they are giving away in a post, they can be better informed as to the impact of whatever content they post to the whole world. It is unfair for people to be unaware of how much personal information they are giving away, and therefore, this tool will help provide the means to protect their identity and information.

This problem is not simple, and therefore, does not have a simple and readily available solution. There is not data that easily lends itself to this task. As such this problem is best broken down into subtasks. There are several subtasks such as geolocation (locating a post based on text or images), identifying personal names in a post's text, and, broadly, identifying individuals based on anonymous post data. Using publicly available datasets for image geolocation-based purely on pixels is simple enough (but a relatively hard problem). In fact, several publications have some degree of solutions to this problem. Name entity recognition can enable the identification of names, locations, and other personal information. This sort of natural language processing datasets is readily available. Besides, the problem space for geolocation can be limited to a smaller region than the whole world to at least demonstrate a proof of concept. Finally, identifying individuals based on anonymous post data is quite a hard task. There is no readily available public data for this, and it seems the best thing is to take multiple public social media accounts of individuals across different platforms and train a model to match their social media together. Essentially, the concept is to identify a person based on an anonymous social media post, if you can link that post to a Facebook/Twitter account, you can identify who they are.

All these models can be created in PyTorch. In a past class project, with group members Ramya Bhaskara and Nick Mohammad, we worked to geolocate images and text for the sake of privacy. I am looking to extend this project to encompass a more complete concept of privacy, a first-pass attempt to understand what information truly compromises a person's identity in this data-driven internet age.

STS Prospectus

Introduction

In the technological age of today, privacy becomes a more and more valuable commodity. With so many companies that live off the idea that information is money, it becomes increasingly concerning the amount of an individual's information that is public. It is public in every sense of the word, not just to a group of people, but to the whole world. Consider the constant data scandals that plague our technological world. Whether it is Facebook, Google, or governments, someone is always getting caught selling, collecting, or losing data that many consider infringes on their privacy. Therefore, as stewards of these technologies, we must develop preemptive ways of protecting the privacy of the individual in an information-based world focused on the collective. The heterogeneous nature of society, especially with respect to privacy, makes the perspective vary greatly from person to person. This study shall focus on Reddit, an anonymous social media since individuals within anonymous social media communities tend to view anonymity as some form of privacy and therefore tend to care about in some manner about privacy. In order to understand the perspective and definitions of privacy, privacy needs to be analyzed in the context of a society.

The conundrum of our information society is the constant complaints of lack of privacy, yet no privacy-invasive companies change their conduct. Every year, new scandals from large technology companies are reported in the news. The same indignant response from the public happens every time, yet those companies only continue to grow. This paradox illustrates the privacy of modern society is too complex to simply view as an ethical issue but must be viewed as a social product. The ambiguous and changing definitions of privacy are complicated to be considered anything less than a sociotechnical establishment.

Research Question

The idea of privacy itself needs to be defined within this context. In order to do this, a survey can analyze the connection between interest groups (subreddit) and perception of privacy. These surveys will be reactionary based, asking questions such as How do you feel about governments analyzing your Reddit posts?

As we address the concerns of privacy in social media, this study hopes to approach the questions: how do text and imagery in social media expose locality that may be considered private by the individuals sharing that information? What constitutes privacy in terms of locality in a public social media setting? What constitutes a breach of privacy in anonymous social media? Does Reddit's perspective on privacy shift the communities' perspective in one direction or another? How do the economic players that have influence over Reddit influence the communities' privacy concerns? What are the connection between government powers, economic players, and the actions of Reddit's administrators? By limiting the groups studied to only anonymous social media and specifically Reddit, the definition of privacy and its relation to other factors (socioeconomic, interest groups, etc.) can be explored in more depth.

Reddit is chosen as the case-study since it is one of the most prominent anonymous social media websites in the world. Anonymous social media in general should have more concern for privacy and thus more involvement in a study and analysis of privacy itself. Additionally, Reddit's community tends to be concerned with privacy and worried about data being used in inappropriate manners. Finally, the style and composition of Reddit make it accessible and simple to attain information about previous posts and previous timelines (that occurred on Reddit).

Literature Review

The problem of privacy in anonymous social media is touched upon by several publications, but few explore in-depth the societal communities and their perspectives on privacy within anonymous social media. Only one article seemed to explore privacy in the same social media but from a completely different perspective. In that publication, they analyze the ability of machine learning techniques to classify users based on their information. The results in this paper indicate the ability to classify gender and citizenship relatively well purely based on the user's text and therefore indicate the possibility of privacy infringement in ways the user may not have perceived as possible.

However, analyzing why users perceive this as a privacy violation is not something analyzed in this paper, and something our proposition is interested in (Fabian et al. 2015).

Socioeconomic Construction of Privacy

Privacy negotiation is discussed in an STS-related journal with heuristics designed to enable better methodologies of resolving individuals conflicting opinions and perspectives (Such and Rovatsos 2016). Another STS-related journal discusses how anonymity is perceived by the users in anonymous social media, and the impossible nature of anonymous networks (to be connected in some manner, one cannot be completely anonymous). This paradox affects how well this model for communication will survive (Sharon and John 2018). Additionally, an STS-related journal analyzes the socioeconomic factors' connections to online privacy literacy, perceptions, and definitions (Epstein and Quinn 2020). An STS-related journal analyzes the demographics with connections to social media behavior, specifically the connection with demographics to privacy settings on social media accounts (Madden 2012). Finally, an STS-related journal identifies the demographic of teens and analyzes the understanding of privacy teens have. Additionally, it studies the ways privacy is achieved and attempts to determine how privacy can be achieved in public networks (Marwick and Boyd 2014). Overall, STS-related journals explore the multitude of ways demographics, societies, and individuals interact with privacy in technology, analyzing the different perspectives, characterizations, and literacy of privacy itself.

Technical Shaping of Privacy

One publication indicates the inability of private users to remain private in social media with a mix of public and private users. The public user's data of social media can be effectively used to predict the private user's behavior indicating an inability to remain truly private in most of the world's current social media (Zheleva and Getoor 2009). One paper studies the effect of Big Data on the average user's privacy and what parts they should be concerned about. They propose a concept that will enable users to be effectively informed about the relevant privacy concerns within Big Data (Smith et al. 2012). Whether anonymous social media is truly anonymous is analyzed by finding out if the information on anonymous social media is sufficient to track or identify users (Chatzistefanou and Limniotis 2019). Another publication surveys the work done so far in user privacy protection and analyzes different techniques and algorithms for privacy prevention and anonymization to determine future research directions
and issues. (Beigi and Liu 2018).

Overall, the literature indicates some investigations into the problems of privacy inside social media, the failure of anonymous social media to be truly anonymous, and the different demographics perspectives on privacy. However, the current literature fails to investigate the different relationships with interest groups (communities created around a shared interest) and their definitions, perspectives, and literacy on internet privacy. The current literature investigates the technical or social impacts of privacy but taking a sociotechnical lens to privacy is missing. Privacy is mutually shaped by both technological advancements and social changes, so to analyze only one aspect you lose the interconnectedness between the dimensions. This paper will take into account both aspects, and hope to truly explore the dynamic product that privacy is. By analyzing Reddit, a dominant anonymous social media/news platform, these relationships can begin to be explored.

STS Framework and Method

Pinch & Bijker's SCOT Framework

To find consider the relevant social groups concerning privacy in Reddit, first, we have to consider the heterogeneous opinions on the matter of privacy. The obvious groups are the separation of the community and the company. However, there are more social groups that should be investigated here. The investors and governmental powers that influence those investors are important players in defining privacy. Additionally, the entire community is not just one individual and there is a wide variety of opinions regarding privacy. Since Reddit is anonymous social media, define interest groups to be the defining categories of the community which has many connections to a multitude of socioeconomic groups. The economic players influence Reddit's policies concerning privacy and governmental entities have control over these players. For example, Tencent, a technological investor based in China, caused much controversy by investing in Reddit since the community, which is an American majority, feared Chinese governmental restrictions and influence, especially concerning privacy.

All of these social groups designate their definitions for privacy and, in doing so, attempt to enforce these definitions through whatever means available to them. Reddit as a company decides what policies to allow the entire platform to have: what types of speech are allowed and what they do with the collected user data. Users enforce their concepts through being moderators which set community rules within their specific community (subreddit), and through market economics. Reddit attempts to please users, so indirectly, users can influence the overall companies' policies and privacy perspectives. The economic investors of Reddit influence the company's decision behind the scenes as do the governmental players. Market economics play a role here, but government concerns, laws, or desires also heavily influence the decisions of these players, but most of this is not done openly. As a result of this, the community's reactions also feedback into this loop refining and reconstruction of the rules and privacy definitions that govern Reddit as a whole.

The economic investors and Reddit as a company (employees) work together internally to resolve disagreements over the direction of the company. The governmental players have indirect influence over the economic players through laws and monetary influence. The community of Reddit itself communicates directly to the company with complaints, but its main form of influence comes through market economics. Reddit desires to please the community since they are their user base. Their monetary model requires customers (users) to be financially successful. Therefore, they work to ensure that their users do not leave for other social media sites in most cases. It is important to note the distinction between purely pleasing and pleasing to ensure users stay. Reddit does have a monetary incentive in the short term to do anything more than keeping users on their websites. As such, this can lead to pervasive incentives such as working to ensure a relative monopoly of the popular anonymous social media websites.

The loop of interactions that involves all the social groups does not have a clear resolver, but there are sort of checks and balances in place that disable any player complete control. Reddit as a company can make its own decisions for some time, but economic investors will use their control to ensure that profit is being made. Just as the community cannot make direct decisions but can influence both the economic players and the company to make decisions that align with their beliefs. Finally, governmental powers have to play all their parts concerning their citizens, laws, and leaders. No group or individual has complete resolve power, but the checks and balances tend towards the common goals, aggregating when there are disagreements. The concerns of privacy follow this pattern as do most decisions concerning the use of the technology of Reddit.

Latour's ANT Framework

In an anonymous social media, specifically Reddit, actors such as the Reddit community, Reddit employees, and economic investors are important social groups that play a role in indicating what privacy means to them. These definitions are not just limited in analysis to their social groups but also actants who play a role in supporting, defining, or altering definitions of privacy. Adopting the Actor (Actant) Network Theory will enable the analysis of privacy with both human and non-human factors. In order to consider the ways in which actors and actants work with and against each other, actants in the specific context must be defined. Actants include the money gained by advertising and selling information, the design of the website to post certain material (pictures, etc.), and the technologies that are commonplace in people's homes (smartphones, computers, etc.).The Reddit community are users of the technology the company Reddit provides; they are economic buyers and they heavily influence the manner in which Reddit is used as an anonymous social media and the information available on it. They are deciders of the actual use of the website and through their economic influence and social influence, they dictate what privacy means to each other and to the company. In turn, the company makes overarching decisions about the admins and influences what privacy means to them. Additionally, they make mission statements that influence both the public perception and dictate their direction with the company rules. Finally, economic investors play roles in influencing the direction the company takes by putting pressure economically. The decisions between the actors and actants are working towards is created through the manifestations of the available technology and choice of the uses of said technologies. The most important non-human is privacy itself as a network. The collectivist nature of privacy in modern times lends itself to being within a network-like form. With the technology of today, privacy has shifted from being in the hands of an individual into the arms of the collective. For this reason, privacy is best considered a non-human actor within this network. For example, the different technologies available limit the information available inherently by what they are capable of sharing. If smartphones were not commonly available, it would be a lot harder to share pictures or imagery (some medium people might consider invasive in the manner it is used).The heterogeneous nature of the different definitions of privacy makes the Actor (Actant) Network Theory especially relevant in this study. Additionally, the combination of technology and people are what defines the ways in which privacy is viewed, defined, and executed.

Methods for Data Collection

The definition of privacy with respect to the company (Reddit) will have to also be analyzed through both interviews (if possible), and company statements. Finally, the effect of investor's and different economic players in Reddit with respect to the company's perspective on privacy will be investigated. Also, the perception of these economic players entering the Reddit community is just as important. It is important for surveys to also analyze the connection between the community members of Reddit and the company through surveys that ask a question about how the community feels about the current direction of the company, especially with respect to privacy.

Despite the many news stories of today concerning themselves of the data scandals of the last few years, confirmation that individuals within the Reddit community feel their privacy is threatened is required through more personal forms. Surveys and interviews can help substantiate these claims and ensure that it is more than just news stories. To understand the average view of these data privacy issues, this study will use social media and news coverage to find an aggregate view. Surveys and interviews will also ensure that individuals believe in the importance of protective methods, making them stakeholders. Additionally, the company of Reddit itself may find themselves, stakeholders, because they have the interest to serve the interests of their consumers. The problems of location-specific and password related data in photography and text are specifically concerning. Current research has worked on basic privacy solutions but failed to build adaptative, broad solutions that harness the power of computer vision and natural language processing. Through developing a system that recognizes inappropriate text and photo usage by users and warning them, privacy can be better maintained in the social media world of today. More importantly, current research has failed to establish a basis for privacy concerns and analyzed the multitude of perspectives that constitute what privacy means to people. Technological solutions can only slightly help the larger problem of privacy.

The primary forms of data collection will through anonymous based surveys that focus on the definition of privacy for different community members within Reddit. These surveys will focus on analyzing the differences between privacy definitions and interests since Reddit is anonymous social media categorized by interest groups. These surveys will analyze different scenarios and record community members' reactions to these scenarios concerning privacy. Using a sample of around 50 for a handpicked selection of around 20 popular interest groups(subreddits) should provide enough data to be able to draw some conclusions. However, since the surveys will be open, the more users reply the more information will be available for use. If there is much more than the expected amount of replies, similar questions with slightly different phrasing can be used to eliminate survey answers that are poorly constructed. Another form of data collection could be through documents on Reddit's privacy policy provided through their website. These policies are outlined in company statements on Reddit's websites. Finally, news articles can be used to view the community's overarching reactions to different economic investors in Reddit, and also Reddit posts that discuss these investors. These news articles can provide a basis for understanding the overarching community perspective on government and economic interests of Reddit. For example, when Tencent (a Chinese multi-national company) invested in Reddit, there was a lot of concern for the potential suppression of controversial opinions/ideas. Several news articles discussed these reactions and would provide a good starting point to analyze the overall perspective.

Additionally, national legal documents can illuminate the perspectives of governments on privacy. These documents such as the US Constitution and others can help build an understanding of what governments consider privacy to be and the manner in which legal parties get involved. This macro analysis of how different national parties perceive and protect privacy should elucidate the larger impact and role of privacy.

Timeline

First, the data collection methods explained will need to be executed, then STS frameworks can be more accurately applied to results in order to interpret and explain them. Following this, final conclusions can be drawn with statistical significance in order to indicate the accurateness of the data sample in representing the overall population (the entire Reddit community).

Conclusion

The expected outcome is that the interest groups in Reddit will have very differing opinions on the definitions of privacy within anonymous social media. Additionally, privacy literacy will most likely differ from interest group to group, but many will probably be surprised at the lack of anonymity in anonymous social media. The younger demographics will probably view their privacy as somewhat compromised already and are less concerned. These contributions will be enabling an understanding of how people view privacy in the anonymous social media world and what concepts/technologies they might consider invasive.

Bibliography

Beigi, G., & Liu, H. (2018). Privacy in social media: Identification, mitigation and applications.CoRR,abs/1808.02191. http://arxiv.org/abs/1808.02191

Chatzistefanou, V., & Limniotis, K. (2019). Anonymity in social networks: The case of anonymous social media. International Journal of Electronic Governance,11(3-4), 361–385.

Epstein, D., & Quinn, K. (2020). Markers of online privacy marginalization: Empirical examination of socioeconomic disparities in social media privacy attitudes, literacy, and behavior. Social Media + Society,6(2), 2056305120916853. https://doi.org/10.1177/2056305120916853

Fabian, B., Baumann, A., & Keil, M. (2015). Privacy on reddit? towards large-scale user classification.

Madden, M. (2012). Privacy management on social media sites. Pew Internet Report, 1–20.

Marwick, A. E., & Boyd, Danah. (2014). Networked privacy: How teenagers negotiate context in social media. New Media & Society,16(7), 1051–1067. https://doi.org/10.1177/1461444814543995

Sharon, T., & John, N. A. (2018). Unpacking (the) secret: Anonymous social media and the impossibility of networked anonymity. New Media & Society,20(11), 4177–4194. https://doi.org/10.1177/1461444818768547

Smith, M., Szongott, C., Henne, B., & von Voigt, G. (2012). Big data privacy issues in public social media.2012 6th IEEE International Conference on Digital Ecosystems and Technologies(DEST), 1–6. https://doi.org/10.1109/DEST.2012.6227909

Such, J. M., & Rovatsos, M. (2016). Privacy policy negotiation in social media. ACM Trans. Auton. Adapt. Syst.,11(1). https://doi.org/10.1145/2821512

Zheleva, E., & Getoor, L. (2009). To join or not to join: The illusion of privacy in social networks with mixed public and private user profiles. Proceedings of the 18th International Conference on World Wide Web, 531–540. https://doi.org/10.1145/1526709.1526781