

A Sociotechnical Analysis of Social Media Data Usage by Tech Companies

A Research Paper submitted to the Department of Engineering and Society


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

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

Ethan Choo
Spring, 2021

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 _____  _____ Date 5/4/2021 _____
Ethan Choo

Approved _____  _____ Date 5/4/2021 _____
Hannah Rogers, Department of Engineering and Society

Abstract

The use of social media has been dramatically increasing over the last two decades and in today's internet, for the everyday user, it is impossible to remain truly anonymous because of all the data that is tracked and collected pertaining to internet activity. Excessive data collection often can lead to unintended consequences and so in this paper, how social media companies including Facebook, Google, and Twitter have affected the quality and diversity of information that users see on social media platforms is examined. To help explain how this came to be, the Social Construction of Technology framework will be used in combination with historical case studies to analyze the different groups that interact with social media and to examine different scenarios in which the social media data is utilized. From there, it is found that because of the differences in interpretations, a lack of diversity in thought is observed for individuals utilizing social media apps from Facebook, Google, and Twitter. Furthermore, the SCOT framework helps show that these negative impacts will influence how users perceive and interpret the use of social media data and create a view that conflicts with the developers interpretation of social media and data. The SCOT framework will then show how changes will have to be made to social media and data handling such that the interpretations of social media by the user and the developers will converge in a way that is beneficial to both parties.

A Sociotechnical Analysis of Social Media Data Usage by Tech Companies

Introduction

“On the internet nobody knows you’re a dog.” This quote described how the internet was able to allow users to become almost entirely anonymous on the web (Steiner, 1993). This came at a time when the internet was just starting to be used by everyday people for enjoyment and expression rather than by computer scientists for academia. However, since the rise of ecommerce in the mid 2000s, this quote hasn’t been entirely true. Now every time you visit a website, search something up, or browse social media, different bits of information about you and your internet activity are being tracked and stored. Slowly, as the internet evolved and became a place for ecommerce, businesses wanted to know the people who were looking at their online businesses (Schwartz, 2001). Thus more and more user information started to be tracked. No longer is it the case that the internet is anonymous. Instead, now almost every action that is taken online is being monitored. This tracking of user data has many purposes and many implications, however, in this paper, the use of user data by social media companies for data filtering will be examined. In particular, Facebook, Google, Twitter, and ByteDance, and their respective social media applications will be discussed.

Social media is being examined here because the use of social media platforms has risen dramatically over the last two decades. As shown in a 2019 survey, 72% of the American population used some form of social media, and that number will likely continue to rise in the future (“Demographics of Social Media”, 2019). From this survey, it can be seen that these platforms have a large influence on the general public. Part of the reason these platforms have succeeded so far is because of something called curation algorithms. Facebook noted that they store over 300 petabytes of data and generate 4 new petabytes of data every day (Wiener &

Bronson, 2014). For comparison, 1 petabyte of data is approximately equivalent to 3.4 years of video content (Fisher, 2020). Users are thus unable to filter through the amount of content available on a platform like Facebook manually, and thus require these curation algorithms to make it possible to access the content that they want to see.

Although the actual way in which the algorithms run or are implemented in various social media platforms is unknown, certain platforms such as TikTok have explained some of the high level factors that go into determining what content is shown to a user. As it is relatively easy for companies to simply store and display data, the largest differentiating factor is how these companies personalize data for users. Thus, these algorithms have one of the largest impacts on a social media platform's success and would not want others to be able to easily replicate this technology. In TikTok's press statement, it was noted that how the user interacts with content, the specifics of the content, and the device and account settings of the user all play a part in determining what other content will be suggested to the user ("How TikTok recommends", 2020). Additionally, it is generally believed that companies like TikTok use some sort of machine learning component, which uses user data to learn about the user's taste, to provide even greater personalization in terms of content delivered (Narayan, 2020).

Curation algorithms have played a large role in leading social media platforms to where they are now and how user data is used. However, the way in which social media companies handle the data used by curation algorithms has led to a loss of diversity of thought through both direct and indirect means. This is evidenced by the way in which curation algorithms develop filter bubbles, the lack of knowledge people have about data collection, and the numerous data breaches that affect people's trust in these companies. Since technology related to the usage of user data developed in response to how individuals used the internet, the Social Construction of

Technology (SCOT) theory will be able to effectively frame the analysis since the underlying idea of SCOT is that society influences how technology is created. The SCOT theory will first be introduced and explained in relation to social media technology. Following this, the SCOT theory will be used to analyze specific scenarios in which users have been unsatisfied with how data is handled by social media companies. These examples include the Facebook Cambridge Analytica scandal, Twitter's reliance on cookies, filter bubbles, and inherent AI bias. Finally, the SCOT framework will be used to analyze how a hypothetical social media company would develop a profitable business model and why this leads to differing interpretations of their technology. Through the lens of SCOT, these examples will show how a change in how data is collected and managed will be seen in the near future in order to satisfy both developers and users.

Social Construction of Technology Framework

In order to analyze this topic and various examples while keeping society and technology's relationship in mind, this paper will utilize the Social Construction of Technology framework to draw meaning from different scenarios involving data in social media. Our understanding of the SCOT theory comes in large part from Trevor Pinch and Wiebe E. Bijker's "The Social Construction of Facts and Artefacts: Or How the Sociology of Science and the Sociology of Technology Might Benefit Each Other" as Pinch and Bijker are two of the originators of this theory. The basis of this theory is that society and human activity is what determines technology instead of the opposing belief that technology determines how society is run. Thus, a large emphasis is placed on examining the social environment that is present during the development and adoption of technology (Pinch & Bijker, 1984). As mentioned before, the curation algorithms seemed to have been created in response to the need to filter through

petabytes of data that would otherwise be impossible for a single human to do. Thus it seems appropriate to utilize the SCOT theory as it allows us to identify the relevant social factors that led to the creation of these algorithms and how data came to be used by these algorithms. Additionally, this framework would allow for social factors to be examined in a way that shows the shortcomings of this technology as it is now for various social groups. The key tenants of the SCOT framework include relevant social groups, interpretive flexibility, and closure (Pinch & Bijker, 1984).

The first tenant of SCOT, relevant social groups, refers to the need to identify the many different social groups that are affected or have an interest in the technology that is being analyzed (Pinch & Bijker, 1984). By identifying these groups and examining their interpretations of the technology, the social context that the technology developed in becomes clearer. This makes it easier to see what the potential problems were at the time, and what specifically the technology being analyzed was trying to solve. In the case of social media data usage by curation algorithms, the relevant social groups include the developers, social media users, advertisers, and the government. Each of these groups are in some way interested in the algorithm, but find value and use from different parts of it. The developers need to create something that will allow their company to generate revenue. Social media users want to be able to find the content they are interested in. Advertisers want to be able to know what other users are interested in so they can advertise to them. Finally, the government needs to make sure basic privacy rights are not being infringed on and that issues of censorship are being addressed. Additionally, some of these groups can be divided even further into subgroups. For example, the social media users group can be divided up by age or socioeconomic status. This would affect

what specifically they want to see as it might be the case that younger users use social media for entertainment while older users may use it for keeping up with the news or with friends.

The second tenant of SCOT, interpretive flexibility, refers to how the technology is interpreted. More specifically, there is flexibility in how people interpret or view technology and how technology can be designed (Pinch & Bijker, 1984). Importantly, it is noted that “There is not just one possible way, or one best way, of designing an artefact” (Pinch & Bijker, 1984). Essentially, there is usually more than one specific problem being solved as multiple groups have vested interest, and so decisions have to be made to determine which problems are the most important. In the case of curation algorithms, it can be seen that users saw the technology as a way for them to get content, whereas developers and advertisers saw the technology as a way to make a profit through the usage and sale of consumer data. Finally, for the government, it may have brought up questions about how this may impact political information and messages. Through this, it is clear that society places a high importance on data, money, and entertainment.

Finally, closure relates to the convergence of the different interpretations of the technology when there are conflicting interpretations (Pinch & Bijker, 1984). In order for closure to be achieved, the social groups need to view the problem as being solved and as a result, new designs for the technology will decline. In our case, if the users view the algorithm as an invasion of their privacy with regards to how their data is being handled, this interpretation would be in direct conflict with the advertisers interpretation. Since social media and the internet as a whole is such a profitable business, companies will continue to try to develop new designs for social media in order to appeal to the most users and advertisers. In doing so, new technology that achieves closure between conflicting interpretations will slowly emerge. Thus, new designs

for social media and the handling of data will continue to develop until the problem is solved without conflicting interpretations.

Impacts of Social Media Data Usage

One point that is argued is that a lack of trust in how companies handle data will affect what people say online. If users are not sure that companies will handle data safely and securely, it is likely that they will avoid searching for, reading, or saying certain things that may not be popular or that they would not like to have traced back to them. If the SCOT framework is used to analyze the users as a group, it can be seen that they interpret the technology differently than the developers do. The developers believe that users always want to see what their mind likes and thus need curation algorithms to know as much about the user as possible. However, users do not always want to see exactly what they are thinking as sometimes it goes against society's norms or they are not particularly proud of their thoughts or may be ashamed of their interests. Thus this technology will have to continue to develop in order to reach closure between these two interpretations.

One example that illustrates the lack of trust users have in companies to appropriately manage their data is the Cambridge Analytica scandal. In 2018, it was revealed that Cambridge Analytica, a political consulting firm had improperly used the personal data of 200,000 Facebook users to create psychological profiles of 87 million Facebook users. (Heawood, 2018) This was done by accessing all of the connections each of the 200,000 original users. The resulting profiles that were created were then used by multiple political campaigns around the world to target and send very specific messages and ads to the individuals whose data had been used. The result of this incident was a very harsh response from the public towards Facebook and Cambridge Analytica and government investigations of Facebook. A survey from the Ponemon

Institute found that after this scandal, only 28% of Facebook users believed Facebook was committed to privacy, which was a drop of 66% from the previous year (Weisbaum, 2018).

It could be argued that companies give users options to manage how much data is collected and used by their applications. If the user has the ability to choose how much data is used, then in the eyes of the developers, there wouldn't be a problem in however they choose to utilize any collected data since they explain how data is handled. When using a SCOT viewpoint, this becomes a possible explanation of how the developers of social media technologies came to an interpretation that does not focus much on data usage or privacy since in their eyes it seems to be voluntary. However, there are many reasons why this choice to not have data collected is not really a choice. Twitter technically has integrated different ways of managing what data is collected ("Personalized Content Based on Your Third-Party Web Activity", n.d.). Twitter allows users to disable cookies, which is one of the main ways browsing activity is tracked. However, this will often result in several features not working. Twitter also claims they don't associate web browsing history with names or email addresses, however, this does not solve the problem as cookies are the only thing needed to track a user's browsing activity across the web. Since many people are not knowledgeable about how the internet works or how data is tracked, it is easy for companies to explain data usage policies in ways that are not clear to the user. Thus since many companies such as Twitter rely on cookies and do a poor job of explaining how data is used, the collection of user data isn't truly voluntary.

As more users begin to lose trust in various social media platforms, a new subgroup of social media users will develop. This social group will consist of those who have been negatively affected by improper usage of social media data and those who are aware of how social media companies handle data and have a negative view of this data usage. In a 2019 study, it was found

that 79% of Americans said they were concerned with how companies use collected data (Auxier et al., 2019). Through SCOT it is clear that this new group will interpret the curation algorithms and social media as a whole differently than the previous social media user group who saw these algorithms as a way to simply find content they enjoy. This new interpretation will lead to a push to change the way in how data is managed and used.

Another reason why these curation algorithms have created a lack of diversity in online thought is because of the phenomenon known as the filter bubble. In a study of 1.2 million Facebook users that interacted with two distinct and conflicting narratives, it was found that the confirmation bias led to social interaction that further formed polarized subgroups. Individuals in these polarized groups were found to be confined to only a few like-minded pages and largely ignored conflicting viewpoints on Facebook. (Brugnoli et al., 2019) Companies such as Facebook, Google, Twitter, and ByteDance all keep the details of how the curation algorithms work a secret, so it has been difficult to prove whether it is specifically the algorithms and its usage of data has been the sole cause of filter bubbles and polarization or if it's a combination of multiple factors. However, a study on YouTube recommendations concluded that on the whole, the YouTube landscape generally showed signs of confinement which leads to filter bubbles (Roth et al., 2020). Thus, it can be seen that the algorithm has at least some form of contribution to the polarization problem. However, it is not a fact that curation algorithms have to cause filter bubbles. Research on the curation algorithm and filter bubble problem has shown that the design choices are what drives the possible creation of filter bubbles. Certain algorithms can actually decrease polarization in applications and can encourage senders of information to increase the quality of information that is being shared (Berman & Katona, 2020).

Artificial intelligence has also been seen to reflect biases that the people developing the algorithm may have. It is believed that Facebook and TikTok utilize some sort of machine learning component for their curation algorithm (Narayan, 2020). Machine learning is a troubling subset of artificial intelligence as they essentially learn how to make decisions based on large datasets that are given to them (Howard & Borenstein, 2018). Thus the biases that exist in the dataset will be present in the decision making process that the algorithms use. The data that is being used to “teach” the algorithm will be inherently biased due to simply how it's being collected. This issue has been shown when a Twitter bot had to be disabled when its AI algorithm started promoting hate speech and when Facebook wasn't able to have its algorithm differentiate between hate speech and opinion (Berman & Katona, 2020). For our instance dealing with curation algorithms, web roaming data will not include poor underprivileged individuals who do not have the ability to leisurely use the web. Thus, another group will not receive the same benefits as everyone else from this piece of technology and thus view the technology differently. Additionally, when applied to curation algorithms, this problem will result in certain topics being recommended more than others which will affect the diversity of opinions and quality of opinions available on social media.

Social Media Business Models

In order to understand why social media companies make the decisions they make and why they have similar business models and data usage practices, the way in which a hypothetical social media company grows is examined. The most important aspect of a social media app that determines its success is its users. Without users, there's no content and no incentive for more users to join or use the app. Thus, a new social media application would need to be very easy for new users to pick up and start using, especially considering how easy it is to use all the other

existing social media applications available now. To achieve this, it is almost necessary to make the service free, at least initially in order to attract users and develop a user base that will allow the app to perform like a social media app. Simply charging users for the app would drive users away no matter how cheap it is, and this would result in a failing social media application and business. However, since the service is free, the company would need to do something in order to make a profit.

There are two obvious solutions. The first solution is to switch to a form of a paid service. This would either be in the form of charging all new users after the social media app has achieved some desired number of users, or to charge both existing and new users in a form of subscriptions that only allows them access to all features of the app through recurring payments. However, with these methods, there will be harsh criticism as people may feel conned when they are suddenly charged for a service that was initially free and the amount of new and recurring users will likely decline and result in a less successful or even failing app. The second obvious solution would be to collect and sell user data, which is what most successful social media companies do now. This form of “payment” is less likely to be criticized as harshly as it is not as noticeable and is a more passive form of payment in that users don’t have to get out a credit card like they would have in the previous solution. Thus, in the eyes of many, this does not seem as costly to them. Additionally, because of how privacy policies are set up, it is very hard for users and developers to be on the same page about data usage. This is shown from a survey conducted which found that only 22% of American adults always or often read privacy policies before agreeing to them while 36% of American adults said they never read privacy policies (Auxier et al., 2019). The same study found that only 22% of American adults read the entire privacy policy

before agreeing to them and that only 13% of American adults actually understand a great deal of the information that is in the privacy policies (Auxier et al., 2019).

This clearly shows where the social media users group starts to have a different interpretation of data usage and social media than the developers. Even if it is not intentional for developers to make privacy policies confusing, the results is that these two groups are understanding and interpreting the technology differently. Initially, this may not be a conflicting view, as users are unaware of what is really happening with their data, however, as more information about this topic is spread, users may begin shifting their views and interpretations to one that is in direct conflict with the developers' interpretation. By using the SCOT framework, it can be predicted that companies will continue to make changes to their technologies and business models. An important note is that the two solutions for generating a profit that are explained above are not necessarily the only solutions or business models that social media companies have to follow. They are simply the most common or obvious. As more conflicting interpretations of social media data usage arise, it is likely that a new, less obvious solution will be developed that allows for many different interpretations to converge and satisfy the needs of the most social groups.

Counter Arguments

The main counterargument that can be made is that the closure aspect of SCOT does not hold true for companies of this magnitude and power. This section will examine this counterargument and respond to the points made. Essentially the counterargument is that these companies hold an oligopoly over the social media industry and have a lot more power than what is assumed in this paper. Specifically, this would mean that they would not have to respond to every customer need or want since there isn't much competition that would force them to make

drastic changes. This would be especially true if the change that is being asked of them would cause them a significant decline in revenue which would be the case when dealing with data and advertisers. However, one way to get around this issue would be for legislative changes to be made that force companies to comply with the needs of the customer. If other parts of the world are examined, different privacy bills that cover different aspects of social media data usage are seen. In Europe, the General Data Protection Regulation sets guidelines for collecting and processing personal data in the EU (Frankenfield, 2020). Furthermore, several bills in the U.S. are under consideration that would change how data is handled by social media companies. Specifically, the DASHBOARD act would require platforms that collect data to be more transparent about what is being collected and how the data is being used (Chakravorti, 2020).

The other way this issue could be resolved given the current social media companies' power is through changes in hardware companies. Recently, Apple has announced a privacy change for iOS 14 that will make it so users have to actively enable device ID tracking that allows for targeted ads. This will make it so users are not automatically tracked as the default option and as such, more users will be aware of what apps are actually tracking their browsing activity and will be able to more easily disable some of the tracking that is unwanted (Rodriguez, 2021). Using the SCOT method, it can be seen that this is a different subgroup of the developer social group, however, these developers do not make profit off the user's data and so they can respond to the needs of the consumer and force social media companies to adapt to hardware changes. In response to this privacy change, Mark Zuckerberg, the CEO of Facebook, stated that this change could be beneficial to Facebook if more companies decide to do business on their app in response to these changes (Rodriguez, 2021). This shows that there are feasible ways for social media companies to remain competitive and profitable without the high degree of data

collection and processing that is present now. The SCOT framework suggests that because data usage is becoming such a large issue for a large relevant social group of users, a new social media structure in which extreme data collection is not needed will develop to solve this problem and result in closure.

Conclusion

By examining specific aspects of social media data usage and the specific technologies that deal the most with data collection and processing, it can be seen that there is a lack of trust in social media companies and a disjoint in views between the people providing social media services and those who are using social media. This is seen from surveys that show that over half of Americans are concerned with how data is handled by companies (Auxier et al., 2019). This has led to a loss of diversity of thought and general wariness and unrest in how people should operate on the internet and throughout the web. However, as this feeling is echoed by more and more people, shown through the various survey studies, the SCOT framework shows why companies and the government will continue to develop new ideas that will help create a convergence of interpretations that allow for users and social media companies to both benefit from how data is used to help improve the social media experience. Thus, it is evident that the social media that is used now will go through drastic changes in the future. As a result of this research on social media, it is likely that a combination of increasing conflicting interpretations as well as changing legislation will force companies to take action and innovate in a way that allows users to find what they want on social media while also not feeling concerned about the amount of data that is being collected. Especially when considering Apple's recent changes that make it so users are not tracked by default, it is likely that these companies will not be able to hold on to their particular business model for much longer. Furthermore, understanding the

negative viewpoints that many people have on data usage may also lead to changes in how other technologies that rely heavily on data, such as machine learning algorithms, are utilized and could suggest that the machine learning field as a whole will be subject to change.

References

- Auxier, B., Rainie, L., Anderson, M., Perrin, A., Kumar, M., & Turner, E. (2019, November 15). Americans and Privacy: Concerned, Confused and Feeling Lack of Control Over Their Personal Information. Pew Research Center: Internet, Science & Tech. <https://www.pewresearch.org/internet/2019/11/15/americans-and-privacy-concerned-confused-and-feeling-lack-of-control-over-their-personal-information/>
- Auxier, B., Rainie, L., Anderson, M., Perrin, A., Kumar, M., & Turner, E. (2019, November 15). 4. Americans' attitudes and experiences with privacy policies and laws. Pew Research Center: Internet, Science & Tech. <https://www.pewresearch.org/internet/2019/11/15/americans-attitudes-and-experiences-with-privacy-policies-and-laws/>
- Berman, R., & Katona, Z. (2020). Curation Algorithms and Filter Bubbles in Social Networks. *Marketing Science*, 39(2), 285-457. <https://doi.org/10.1287/mksc.2019.1208>
- Berman, R., & Katona, Z. (2016). The Impact of Curation Algorithms on Social Network Content Quality and Structure. *SSRN Electronic Journal*. doi: 10.2139/ssrn.2848526
- Brugnoli, E., Cinelli, M., Quattrociocchi, W., & Scala, A. (2019). Recursive patterns in online echo chambers. *Scientific Reports*, 9(1), 1–18. <https://doi.org/10.1038/s41598-019-56191-7>
- Chakravorti, B. (2020, January 30). Why It's So Hard for Users to Control Their Data. Harvard Business Review. <https://hbr.org/2020/01/why-companies-make-it-so-hard-for-users-to-control-their-data>
- Demographics of Social Media Users and Adoption in the United States. (2019, June 12). Retrieved October 18, 2020, from <https://www.pewresearch.org/internet/fact-sheet/social-media/>

- Fisher, T. (2020, March 25). Terabytes, Gigabytes, & Petabytes: How Big Are They? Retrieved October 18, 2020, from <https://www.lifewire.com/terabytes-gigabytes-amp-petabytes-how-big-are-they-4125169>
- Frankenfield, J. (2020, November 11). General Data Protection Regulation (GDPR). Investopedia. <https://www.investopedia.com/terms/g/general-data-protection-regulation-gdpr.asp>
- Heawood, J. (2018). Pseudo-public political speech: Democratic implications of the Cambridge Analytica scandal. *Information Polity: The International Journal of Government & Democracy in the Information Age*, 23(4), 429–434. <https://doi.org/10.3233/IP-180009>
- How TikTok recommends videos #ForYou. (2020, June 18). Retrieved October 18, 2020, from <https://newsroom.tiktok.com/en-us/how-tiktok-recommends-videos-for-you/>
- Howard, A., & Borenstein, J. (2018). The Ugly Truth About Ourselves and Our Robot Creations: The Problem of Bias and Social Inequity. *Science & Engineering Ethics*, 24(5), 1521–1536. <https://doi.org/10.1007/s11948-017-9975-2>
- Narayan, A. (2020, May 13). How ByteDance's TikTok is using Machine Learning to learn and build a huge dataset of human... Retrieved October 18, 2020, from <https://medium.com/@aniketnarayan/how-bytedances-tiktok-is-using-machine-learning-to-learn-and-build-a-huge-dataset-of-human-52458dfaf5c4>
- Personalized content based on your third-party web activity. (n.d.). Retrieved April 5, 2021, from <https://help.twitter.com/en/using-twitter/tailored-suggestions>
- Pinch, T. J., & Bijker, W. E. (1984). The Social Construction of Facts and Artefacts: or How the Sociology of Science and the Sociology of Technology might Benefit Each Other. *Social Studies of Science*, 14(3), 399–441. doi: 10.1177/030631284014003004

- Rodriguez, S. (2021, March 18). Zuckerberg: Facebook may actually be in a “stronger position” after Apple’s iOS 14 privacy changes. CNBC. <https://www.cnbc.com/2021/03/18/zuckerberg-facebook-may-be-in-stronger-position-after-apple-ios-14.html>
- Roth, C., Mazières, A., & Menezes, T. (2020). Tubes and bubbles topological confinement of YouTube recommendations. *PLoS ONE*, 15(4), 1–17.
<https://doi.org/10.1371/journal.pone.0231703>
- Schwartz, J. (2001, September 4). Giving Web a Memory Cost Its Users Privacy. *The New York Times*. <https://www.nytimes.com/2001/09/04/business/giving-web-a-memory-cost-its-users-privacy.html>
- Steiner, Peter. “On the Internet, nobody knows you're a dog”. Cartoon. *The New Yorker*, 5 July 1993.
- Weisbaum, H. (2018, April 18). Trust in Facebook has dropped by 66 percent since the Cambridge Analytica scandal. NBC News. <https://www.nbcnews.com/business/consumer/trust-facebook-has-dropped-51-percent-cambridge-analytica-scandal-n867011>
- Wiener, J., & Bronson, N. (2014, October 22). Facebook's Top Open Data Problems. Retrieved October 18, 2020, from <https://research.fb.com/blog/2014/10/facebook-s-top-open-data-problems/>