TikTok's Societal Impact of Data Collection

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

Spencer Portuese

Spring 2024

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

Advisor

Pedro A. P. Francisco, Department of Engineering and Society

STS Research Paper

Introduction

TikTok is one of the largest social media platforms and is continuously growing, hosting more than 500 million users. Its main focus is short-form videos that can be watched in succession on a user's "For You" page, which algorithmically becomes tailored for each individual user based on the other videos they have watched and their interests (Herrman, 2019). This algorithm requires a very large amount of data in order to determine what to show its users, and therefore TikTok collects a large amount of data to make it work, which is often a worry to its users.

Within the app, users can request their data in one of two formats, either an easy-to-read .txt format or a more machine-readable .json format, which is a file structured as nested arrays. The app claims to include data on a user's profile and app settings, but also user activity on the app, which contains a significant amount of data on what videos are watched when, and what was done with each video. When users download this data, it is overwhelming to process it all alone, and certain critical pieces of data are hidden away in files, so even when users try to understand what is within this data they do not get a firm enough grasp of it. To help users understand this large swatch of data, I created a tool that looks through the files and pulls out the pertinent information. This tool is further explained in the background section.

That being said, this data is only meaningful within certain contexts. If the application can tell users that the application knows their interests, it may be unclear what it can do with that information and how that can impact them. Even on a larger scale, TikTok became put in the eyes of the US government over concerns about this data becoming negative for national security (Fung, 2023). This concern over national security centers around the concern that China's

security laws seem to force any company under their jurisdiction (such as ByteDance, the owners of TikTok) to provide data if asked under the guise of national security (Fung, 2023). The US government claims that data can be used against them in two major ways. First, China could use this data as intelligence, such as if they suspect someone is a spy against them or there is someone they want to blackmail, the data within the app could showcase potential interests or things to look into to learn more about that individual (Fung, 2023). Furthermore, there is concern that the Chinese government could influence ByteDance to censor or promote certain content on the app, which is less related to data but still showcases the impact the app could have over the population, and why America could be worried about it (Fung, 2023). Most individuals will be worrying about what this data can say about themselves individually, especially from parents due to the app being targeted at children (Roth, 2021).

To determine the societal impact of TikTok's data collection, this paper will investigate what personal information can be discovered from this data file utilizing the tool created and other public information to determine how dangerous this data collection is to users of TikTok, as well as the effects of data collection in the first place.

Background/Significance

As mentioned earlier, the TikTok data file is very large and unclear what it contains for users who decide to download their data. If they choose to select the .txt option, they are given a minimum of 19 files spread among 4 folders, the longest of which in the data for this experiment were almost 150,000 lines long, or even longer depending on app use for other users.

The data is much easier to parse in its .json format, which can be processed using Python. After the data is analyzed, a Python library can also be used to show this data in an easy-to-read .pdf format. The overwhelming amount of data was split into two main categories for analysis: biographical information and app use data.

The biographical data is about the users themselves, not how they use the app. This includes information about what platform they view the app on, operating system, email, birthday, and most recent IP, amongst other things. This type of data is very simple to access within the data, it just might be a bit more hidden in it. Generally, this data may not be as large of a surprise that it is known, but certain users may have this data hidden to the app which would be an actionable thing users could do if they want to protect their privacy.

The second type of data is more complicated and requires some analysis. The “app use” category includes when the user opens the app, what videos were watched or liked, and when they were watched or liked. Combining these with queried details of the videos can show off hashtags of videos that users tend to watch and like. While there are many possible options for analysis here, the most critical ones would be the times that users use the app, and what their interests would be, both of which could be determined by this data. Regarding the time that users use the app, after accumulating the times that each videos were watched, a heatmap could be created based on short “buckets” of time throughout the week. This can be compared day to day to see which days users are more active and at what time. This data is valuable considering that while the security of users’ data on the app seems to be secure, video requests are unencrypted, meaning that sniffers (programs that analyze network traffic) could determine this data themselves about a user given enough surveillance (Neyaz et. al, 2020).

The main flaw with this approach is that the data being analyzed is only the data collected on the app, and this is only data that TikTok chooses to give out to its users. TikTok also collects data from other apps or websites, which could tell TikTok analytics about the user but would not appear in this report (Germain, 2022). In the past, TikTok seemed to also collect data about user's clipboards before Apple exposed this and the feature was removed, which is not contained in the data file (Doffman, 2019).

In short, my tool allows users to upload their TikTok data and see what specifically the data knows about them, and give a .pdf report of these findings and what the potential consequences could be. This tool then can be used to analyze the societal impacts of the data.

Going into previous research, a meta-analysis determined that most studies on the topic of TikTok analyze the content side of the platform: what content is generated, how the algorithm pushes different types of content to users, and how it affects cultures, with much less focus on privacy and ethics of data collection (Kanthawala et. al., 2022). This focus is likely on the format of the platform, which due to its innate design is incredibly user-oriented and leads to the desire of people to be "micro-influencers," which has had a considerable impact on the types of content delivered to users on the app (Jaffar et. al., 2019).

Most of the app's users tend to be young users, and therefore much of the content of the app is tailored around them (Shutsko, 2020). This raises concerns over the data security of these younger users, especially as they may be less aware of what they would be providing to the application, and especially to the public. However, most parents of tween children are more worried about children interacting with strangers than revealing personal information, although many parents still monitor their children regarding that as well (De Leyn et. al., 2022).

All of this goes to show that the specific question of what is being gathered and what can be done with it has not been a major focus when researching TikTok and its impacts, as how the app impacts people, especially younger people, are more important to most users or their parents.

Methodology

In analyzing the impact of TikTok's data collection, we will analyze data collection through the latent and manifest functions and dysfunctions STS framework. Clearly, this data is being collected for some purposeful reasons (manifest), which either can have positive consequences (functions) or negative functions (dysfunctions). However, some effects would not be intended (latent), which similarly can result in functions or dysfunctions. To understand the societal impact of data collection, all of these elements should be analyzed.

To determine what exactly these latent and manifest functions and dysfunctions are, and how they can impact the users of the app, two main things will be done. First off, research will be conducted to further understand how the app and its data collection impact users using it via reading recent research. Secondly, a small experiment will be conducted to determine potential consequences that could occur in the event of a data breach of user TikTok data.

To determine the potential consequences of a data leak, trials could be conducted of technologically adept participants being given the completed report from the tool created earlier and then, playing the role of an adversary, attempting to determine private information about that individual. This part however would have a significant breach of privacy, so participants who submit their data for these trials would have to fully consent to this, which may prove difficult. The amount of information gathered from users would be separated into categories such as public profile knowledge (such as username, profile picture, public videos, etc.), private profile

knowledge (birthday, phone number, etc.), personal information (hometown, location, schedule, etc.), and intimate knowledge (interests, relationships, career, etc). Based on how much information could be gathered from each profile in each category, an estimate of how much information could be gathered from this data would showcase to what extent this TikTok data would be considered a breach of privacy.

The results of this experiment would showcase the scale of personal impact the data getting out could have on user. Furthermore, this would showcase purely the latent dysfunctions of TikTok data collection, which while important does not explain the complete impact of the data collection. To fill these other categories, research and survey will be conducted to examine their impact.

The manifest functions are clear and can be determined straightforwardly from the data file in the first place. Content like IP address and user operating system are simply data that the app needs in order to operate, and further details such as watch history would be used in the FYP algorithm. However, there likely could be additional reasons that need to be considered that can be found in research.

The manifest dysfunctions are less explicitly written and require thoughts about how the app was designed and how the data collection fits into this. While there is much out there about the dysfunctions of the app as a whole, manifest dysfunctions of the data collection itself is a lot more of a gray area, so viewing the design decisions from the perspective of ByteDance can clear up what dysfunctions were considered and why the decisions were made to fit their goals.

Finally, the latent functions require analysis of benefits that were not factored into the design of the data collection. While some could appear in the adversary experiment, this is by far

the most specific category to look into. The collection of data for the app innately is good for the app and bad for the users, and this balance tends to be known by developers and therefore is limited, as the data tends to be collected as needed to run the app. This means that each piece of data tends to have a specific purpose and limits the opportunity for it to have effects other than is intended.

To understand the total impact of the data file all of the latent and manifest functions and dysfunctions need to be understood to evaluate the total societal impact of the collection.

Literature Review

As mentioned earlier, it is important to look at each subset of latent/manifest functions/dysfunctions, and looking at previous research can help showcase the content of most of these categories. As mentioned earlier, the core of the additional research method of the analysis of the data mainly will focus on latent dysfunctions, although it will also be discussed in this section.

Before jumping into the research, it is important to show the current state of TikTok research. A meta-analysis of the research done on TikTok in 2022 looked at 58 journals on the topic and investigated what type of information has been collected about TikTok, specifically looking at it through an ethical lens. In this review, they discovered that the main topic on TikTok that was being researched was the cultural impact of the app and trends, while there was a significantly smaller focus on the ethics of data collection, which is noted as concerning considering the younger age of its users (Kanthawala et. al., 2022). By far, it appears that researchers are more interested in the social impact TikTok has on the world and its implications, which still are helpful for determining the latent/manifest functions/dysfunctions of the app. The

hard part here is connecting these specifically to the collection of data, which will be addressed in the following sections.

Starting with the manifest functions, the clearest category. These are the positive benefits that were intended to come from the decision to collect a large amount of data. The core reason why this data is obtained is for the core algorithm of TikTok to get to know its users and present its content to them in a way that they would enjoy, as stated by Klug et. al. (2021), the algorithm is "based on previous and continuous user engagement with presented video content through video viewing time, liking, commenting, and sharing." All of these parameters are held within the downloadable data file. This interaction is purely unique to TikTok, but Bhandari and Bimo (2022) state, "TikTok is the only one to position its algorithm at the center of the social experience it engenders; the algorithm determines the type of video content the user is exposed to, and viewing this content makes up the majority of the experience on the platform." In the eyes of TikTok, this is a very intentional decision to engage users in the application and have them enjoy their experience, which is a clear manifest function.

Latent dysfunctions also contain some fairly straightforward contents, especially looking at the social ramifications of this data collection, especially due to how it relates to "the algorithm" of TikTok. While using all of this data to tailor users towards content they like, it also provides a change in the opposite direction: content creators now have to tailor their content to reach more people, or at least, they think they do. Most content creators just felt confused about the algorithm, as Klug et. al. (2021) determined that they "had rather confusing experiences and primarily understood the TikTok algorithm as erratic without observable patterns of distributing videos." This confusion for something many creators rely on for their livelihood is very concerning.

Manifest dysfunctions regarding users of the app seem to mostly fall outside of the distinct data file users can collect. This likely would be due to TikTok wanting to hide negative experiences for the users. Germain (2022) goes over some of these uses of data, specifically things called “pixels” which are embedded in other websites to send user advertising data to TikTok, which increases the specificity of ads users receive, and potentially worse. As much of Congress is concerned, TikTok’s relationship with China means that theoretically, this data could be forced to be given to the Chinese government for espionage or targeting government officials for their own benefit, although it remains purely hypothetical, not something that would be likely to happen (Fung, 2023). While this interaction with China cannot be known as intentional or not, many people seem to think it is, and it is surely a consequence of the explicit decision to collect significant advertising data on each user.

Latent functions describe how this data collection helps people outside of its express purpose. Its express purpose is to tailor content for the algorithm that users like and to distribute appropriate advertisements to them. The positive benefits of the algorithm are very present, for example as De Leyn et. al. (2021) discuss how the function of the app allows for beneficial tween self-expression, development, and connection, while balanced with family values. They state “tweens’ networked participations are managed within the structural context of the household, and informed by socio-cultural norms, values and assumptions on what it means to be a child, teenager and adult” (De Leyn et. al., 2021). While the app was generally targeted at kids, this significant interaction likely was not and would not have been accomplished without the data collection required for the app.

Discussion/Results

It becomes apparent that throughout the multiple consequences of TikTok's data collection, there are certainly benefits and detriments. The benefits are tailored around users' experiences within the app and the consequences of using the app, while the detriments tend to focus on how the data could be used outside the content of the apps or users attempting to outplay how the data is used. I believe that the lower risk of something negative happening regarding the data being sold out means that for most users of the platform, the societal impact of TikTok's data collection would likely be positive. However, there are a significant amount of forces at work here, specifically the powers of the United States and China, and all the people who work for them and interact with the app. The US has its own interests to protect, which in a worst-case scenario could be exploited by the data collection of TikTok, and they would prefer not to take that risk despite the benefits other people may experience from using the app and having their data used in an algorithm to accommodate them.

Looking at the latent and manifest functions and dysfunctions of collecting this data shows off the overall scope of impact the data collection has. While neither side seems to overpower the other, and there is no denying that this data collection does have positive consequences, it is clear that this data collection is impacting everyone who uses the app. There is no TikTok in its current state and societal impact without the data collection it is dependent on, and therefore the question of whether or not this data collection is good for society is a much broader question on whether or not TikTok is good for society, which is a much larger topic. Regarding the dysfunctions, it also is clear that in worst-case scenarios it could be really detrimental, but in the average case, it appears to be minor.

Regarding the small experiment, I was unable to obtain consenting participants to be used as data points, so a smaller version was run using just my own data. Using my tool program, a

few things were clearly available revealing private things about each user. Firstly an IP address could be used to determine the general location of the user, showing Charlottesville but not as specific as an address. Secondly, some sleep patterns could be determined from the app use, showing that on weekends I was more likely to be up late using the app than on Saturdays and on Wednesdays, as I tend to more frequently watch videos on the app at those times. Additionally, an email address was provided, which could in the event of a leak open up to additional attacks or spam. Lastly, the iPhone type and operating system were provided, which if it has particular vulnerabilities could be used for an additional attack. The data file also contains every single comment and direct message sent, which was not analyzed in my version of the tool to not overwhelm the document, but is something that if the data were to be leaked would be visible. That being said, this was the extent of what was easily accessible. There were fields for birthday and phone number, which would be bad to have, but neither of them had any content so they were left unknown. Similarly, there were large sections for advertisement and third-party data, which were similarly blank in this data file.

There was very little significantly of note in this data file, more things for the app to run and track the user's activity to recommend videos. While the tool is fairly limited and further analysis can be done on the content of the videos watched, what videos were liked, and what videos were favorited, they are also core necessities for how the app works. For most people, this data ending up in the wrong hands would have very minor consequences outside of an email potentially getting spam or private information being disclosed in direct messages or comments. That being said, as mentioned earlier if this data became accessible to potential blackmail targets or people with sensitive positions things such as personal interests could be used by an adversary to manipulate the target.

Conclusion

Using social media gives significant amounts of information to the app for it to function. It is a common worry that this data could be detrimental for the users and the data collection is inherently dangerous and a breach of privacy. Through this research, it has become more clear, by looking at the consequences of the data collection, that while there can be negatives of this data if somehow it gets into the wrong hands or if TikTok wants to take advantage of it, the data itself tends not to be the concerning part. In fact, the consequences of this application being so based on an algorithm seem to more frequently lead to a negative impact on its users.

The main thing learned from creating and using the tool was how while there is a ton of data stored for each user on TikTok, mainly revolving around the videos watched, most of this is very specific to what the app requires to run and feed for its algorithm. That being said there is no denying that the dysfunctions of the app cannot be ignored, but generally, these dysfunctions tend to be something a user should be aware of using the app, and going forward TikTok should make these issues clear so users can be informed in how they use the app. Data leaks are a concern for every form of media used, so hopefully users can protect themselves on what they do on the app, but this is something widespread across all social media platforms and is not specific to TikTok. However, TikTok being focused on children also heightens this issue, so internet safety needs to be heavily taught to kids using the app by their parents and the app itself.

References

Bhandari, A., & Bimo, S. (2022). Why's everyone on TikTok now? The algorithmized self and the future of self-making on social media. Social media + society, 8(1), 20563051221086241.

De Leyn, T., De Wolf, R., Vanden Abeele, M., & De Marez, L. (2022). In-between child's play and teenage pop culture: tweens, TikTok & privacy. Journal of Youth Studies, 25(8), 1108–1125. https://doi.org/10.1080/13676261.2021.1939286

Doffman, Z. (2019, March 9). Warning—Apple Suddenly Catches TikTok Secretly Spying On Millions Of iPhone Users. Forbes. Retrieved October 20, 2023, from https://www.forbes.com/sites/zakdoffman/2020/06/26/warning-apple-suddenly-catches-tiktok-secretly-spying-on-millions-of-iphone-users/?sh=42911ddf34ef

Fung, B. (2023, March 24). *TikTok collects a lot of data. But that's not the main reason officials say it's a security risk*. CNN. Retrieved April 14, 2024, from https://www.cnn.com/2023/03/24/tech/tiktok-ban-national-security-hearing/index.html

Germain, T. (2022, September 29). *TikTok Tracks You Across the Web, Even If You Don't Use App*. Consumer Reports. Retrieved April 14, 2024, from https://www.consumerreports.org/electronics-computers/privacy/tiktok-tracks-you-across-the-web-even-if-you-dont-use-app-a4383537813/

Herrman, J. (2019). How TikTok is rewriting the world. The New York Times, 10, 412586765-1586369711.

Jaffar, B. A., Riaz, S., & Mushtaq, A. (2019). Living in a moment: Impact of TicTok on influencing younger generation into micro-fame. Journal of Content, Community and Communication, 10(5), 187-194.

Kanthawala, S., Cotter, K., Foyle, K., & DeCook, J. R. (2022). It's the methodology for me: a systematic review of early approaches to studying TikTok.

Neyaz, A., Kumar, A., Krishnan, S., Placker, J., & Liu, Q. (2020). Security, privacy and steganographic analysis of FaceApp and TikTok. International journal of computer science and security, 14(2), 38-59.

Roth, S. M. (2021). Data Snatchers: Analyzing TikTok's Collection of Children's Data and Its Compliance with Modern Data Privacy Regulations. J. High Tech. L., 22, 1.

Shutsko, A. (2020). User-generated short video content in social media. A case study of TikTok. In Social Computing and Social Media. Participation, User Experience, Consumer Experience, and Applications of Social Computing: 12th International Conference, SCSM 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part II 22 (pp. 108-125). Springer International Publishing.