Deciphering TikTok's User Data File

TikTok's Data Collection and Potential Impacts

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By
Spencer Portuese

November 3, 2023

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.

ADVISORS

Prof. Pedro Augusto P. Francisco, Department of Engineering and Society

Prof. David Evans, Department of Computer Science

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.

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. 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. This is where the technical project comes in, creating an application that when this data file is input, it will output the most critical information within it and give users a better grasp of what the app knows about them.

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).

The technical part of this research will focus on analyzing the data file TikTok provides users and making a tool to decipher it, and the sociotechnical part will focus on the data itself and how this data being collected could impact individuals.

Investigating the TikTok Data File

As mentioned earlier, the TikTok data file is very large and unclear about what it contains to users that 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 user 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 that 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 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, this project provides a tool that users could upload their TikTok data into and see what specifically the data knows about them, and give a pdf report of these findings and what the potential consequences could be.

TikTok's Data and Impacts

The data collected from the data file in the technical section and general research for data not included in the data file can help answer the question of what exactly TikTok is gathering about people, and what can be done with that information.

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 it impacts people, especially younger people, are more important to most users or their parents.

To determine the full extent of what data can be collected by TikTok and what can be done with it, the results of the technical component will considered as research is conducted on what specifically can be done with it. Ideally, trials could be conducted of technologically adept participants being given the completed report from the technical section and attempting to determine details 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. If this is feasible, 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.

Conclusion

TikTok has had multiple concerns over its privacy, notably how, by design, it collects a lot of data about each user. However, the exact extent of information gathered on each user is not formally known. The technical part of this research will determine what this extent might be, based on the data that TikTok provides on a user, but more noticeably the technosocial aspect approached in the STS section would approach how other sorts of data could reveal private information about users. TikTok is an immensely popular app right now, and it is important to make sure that its data collection does not cross any boundaries for users without their knowledge. Based on what is known so far, it seems that the data would not be above and beyond any other social media application, but it is possible that more is collected that could impact individuals in various ways.

References:

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

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 October 20, 2023, from https://www.cnn.com/2023/03/24/tech/tiktok-ban-national-security-hearing/index.html

Germain, T. (2022, September). How tiktok tracks you across the web, even if you don’t use the App. Consumer Reports. 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.