Twitter Location Analysis

Team Members:

Aden Ramirez :
Abdur-Rehman Naveed:
Luis Dominguez:
Amit Nankar:

Arizona State University

CSE 467

Rakibul Hasan

5/2/2023

Table of Contents

Introduction

The widespread use of social media platforms has led to the sharing of personal information, including location data, becoming commonplace. Twitter, in particular, allows users to share their location data through coordinates, which is accurate down to 10cm of precision. Location sharing can be useful for businesses or aid in the fun of helping others find places a user finds interesting. It is unclear how many Twitter users opt-in to share their GeoJSON location and how much data they are meaning to disclose. Moreover, users may not fully understand the level of precision provided by GeoJSON, and its potential implications. We would like to take a quick look at how users share their GeoJSON data and how it might affect their personal security.

We are looking to answer the following questions:

1. **Given a randomly sampled dataset, how many tweets contain GeoJSON data?**
2. **Do users who willingly tweet their GeoJSON data habitually do so?**
3. **Can a user's personal patterns be derived from their GeoJSON tweets?**

In this project, our team aims to address these issues by analyzing Twitter data to determine how many users willingly share their location and how much data they disclose. We will also investigate whether users who disclose their location habitually do so and whether their personal patterns can be derived from their tweets.

Before beginning our deep dive we hypothesized:

1. **User supplied GeoJSON data is under 25%**
2. **Users who willingly supply GeoJSON data habitually do so.**
3. **A user's patterns can be tracked using GeoJSON data.**
4. **Users are unaware of how their geolocation data can be used.**

EDA 1 - Exploring a "Random" Dataset

To begin, we researched Twitter API access and scraping methods, we found the Python tool **Snscrape**. The selling points were that we didn't need to use Twitter's paywalled API, the code was very easy to use and it allowed us to scrape a desired amount of tweets over a given time period, by using either a username or trending string.

We wanted to do a fair job of analyzing Twitter users, so we knew we had to collect a large amount of data and wanted to collect tweets that would represent a wide swath of users from all over the world without interfering in the selection process too much. We also knew it would be a bit difficult to gather truly random data because our tool required either a username or text string to search, so we developed a method that would divide the work evenly and that would allow us to scrape data topics as they appeared in each of our "trending topic" feeds; we wouldn't be choosing the topics. A note here is this may be affected by recommendation

algorithms for each of us, which is slightly mitigated by all four of us collecting on separate accounts of varying usage.

To obtain an initial dataset, each of our group members were tasked with creating a Twitter account with our ASU email; this was an effort to mitigate any biases a personal, previously held account might bring to a feed. After establishing a Twitter account, we then gathered **500** of the most recent tweets for ten trending topics, over the course of a week. Some topics did not have 500 tweets so the scraping method was used to supplement the dataset with trending tweets until we had a large enough dataset to explore.

Our efforts scraped **21,348 tweets**, covering **44 topics**. The tweets covered topics ranging from tech, politics, entertainment, health, food, sports, religion, business and pop culture. The topics are displayed in the following table:

Tweet Topics			
Bing	Club Renaissance	Elizabeth Olsen	dementia
Harry Potter	Elden Ring	Europa League	Fox News
Judgement Day	Gen Z	Fulton County	Hogwarts Legacy
New Ceo	LakeShow	South Park	Leonardo DiCaprio
State of the Union	Mitch McConnell	The Last of Us	melanoma
Super Bowl	Most Americans	Xavi	Ohio
Tim Kelly	Russians	Alaska	Rep George Santos
Turkey	Taibbi	Nintendo Direct	Sinema
Twix	The President	Powerball	Spartans
Wrexham	UFOs	All Lies	Switch 2
Arkansas	Barcelona	Creepy Joe	The Bible

Our numerical analysis found the following:

- **21,348** tweets were posted by **17,631** users, and some of those recurring users appeared dozens of times in the dataset.
- There were **370** tweets featuring GeoJSON data, that is **1.7%** of the dataset. Out of the **370** GeoJSON tweets.
- There were at least **50 coordinate tuples** (locations) that showed up at least **twice**. Of the reoccuring 50 locations, the t**op two tuples** appeared over **ten times**.
- There were **24** users who tweeted their GeoJSON at least **twice**. The **top two users** Tweeted their GeoJSON coordinates **14** and **7** times, respectively.

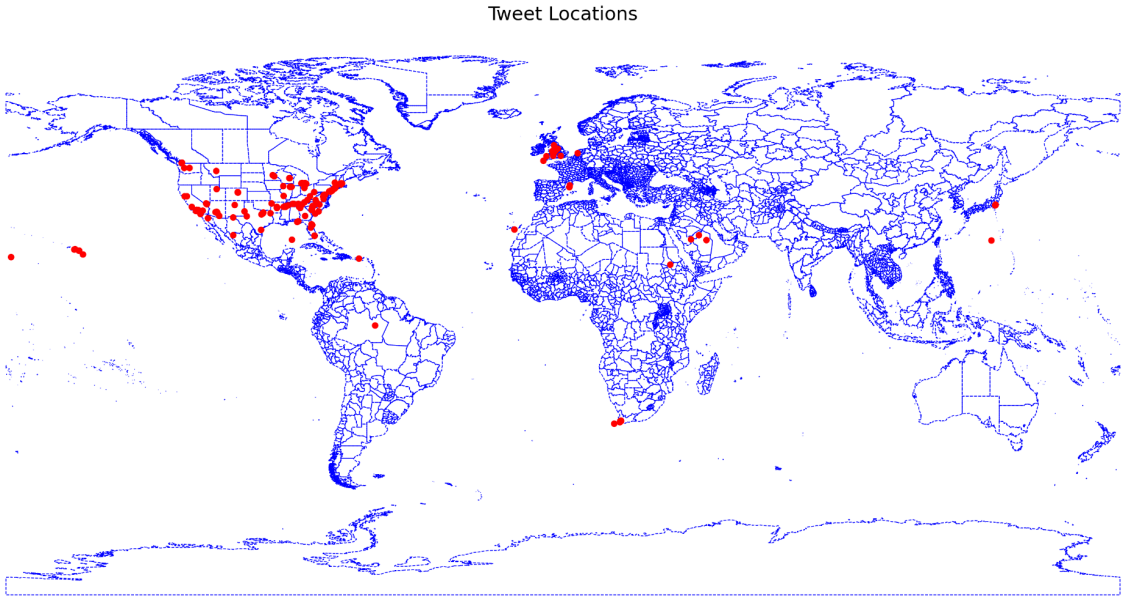
Our first exploratory data analysis allowed us to address our first research question and hypothesis:

> ***Q1: "Given a randomly sampled dataset, how many tweets contain GeoJSON data?"***
> ***H1: "User supplied GeoJSON data is under 25%."***

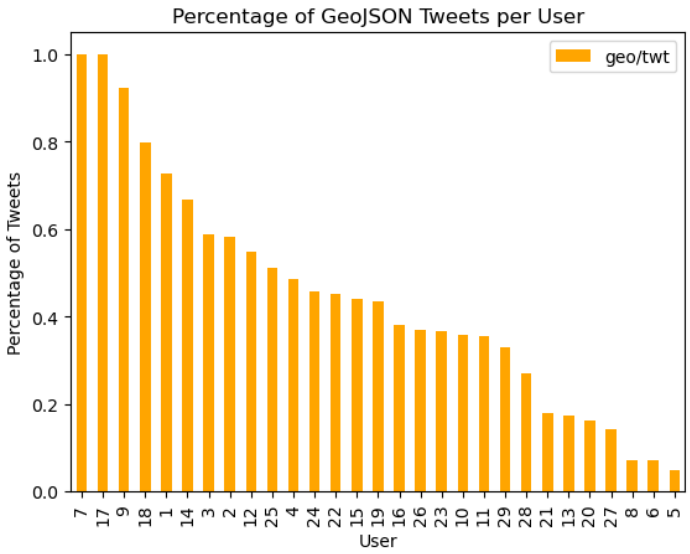
As we just discussed, in a set of over 21k tweets, 370 contained GeoJSON data. That's 1.7% of a "randomly" gathered dataset. To improve our study we would need to replicate the study several times and improve our sampling process. This would be closer to random, but for the sake of this project, we can say that we confirmed our first hypothesis; **user supplied GeoJSON data is under 25%.** At such a small percentage, one might be able to say that the security concerns associated with sharing personal location data aren't affecting a significant amount of Twitter users, but we wanted to take a closer look at the users who do share their location and how they might be putting themselves at risk. Our second and third research questions and hypotheses were dependent on a more focused study of users who willingly tweet GeoJSON tweets, so we conducted deeper research into those users from our initial scrape.

EDA 2 - Closer Look at Habitual GeoJSON Tweeters

To begin our user-focused research, we extracted the list of **29** GeoJSON tweeters from our original dataset. We then attempted to scrape **1,000** tweets per user to build a new database; we fell short of the mark on a few users, but were able to collect **25,973** tweets to analyze. Our database was filtered down to only tweets containing GeoJSON, leaving us with **11,240** tweets, or **43.28%** of the user-focused scrapes. The filtered dataset featured **234** unique locations from all over the globe, but were mostly from the **United States**, roughly shown on the following map:

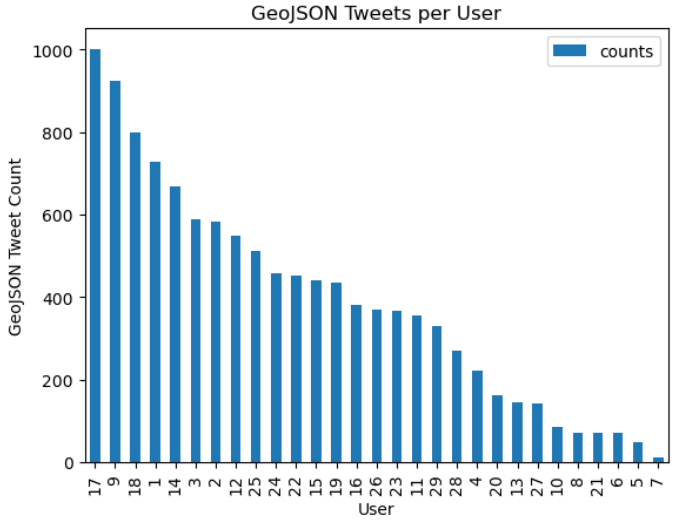


Some stats to note for the GeoJSON dataset are: Max: **1001,** Min:**10,** Mean:**387.** The mean shows that Tweeters who habitually share their location, do it over a third of the time they post. If someone with nefarious goals wanted to find their location, their tweets would be an easy way to track their patterns. This is why we were interested in exploring how often a user would willingly offer location information; if a user had to opt-out of sharing their location information, Twitter might be putting users at risk.
The following graphic helps us address our second question and hypothesis:



Q2: "Do users who willingly tweet their GeoJSON data habitually do so?"
H2: "Users who willingly supply GeoJSON data habitually do so."

Without a doubt, our second hypothesis was confirmed. When the average GeoJSON tweeter is giving their location over a third of the time, they are habitually doing so. As the bar graph demonstrates, these users tweet their location often. We can reference a similar bar graph in conjunction with the Percentage bar graph to address our third and final set of question and hypothesis:



Q3: "Can a user's personal patterns be derived from their GeoJSON tweets?"
H3: "A user's patterns can be tracked using GeoJSON data."

When we take into consideration the volume of tweets GeoJSON users produce and the staggering amount of their tweets that feature their location, we can also confirm our third hypothesis. **A user's patterns can be tracked with GeoJSON data.** Focusing only on User 1, they tweeted four different locations:

Location	Frequency
1	722
2	5
3	1
4	1

Of those four locations, one was tweeted 722 times. If a bad actor were to want to find their location, all they would have to do would be follow their Twitter feed and track the locations in their posts to map patterns and routine. We did conduct a light study on this, but have decided to keep it general and private as it is very invasive for the user that may not be aware of the usage.

Related Work

So far, our study has discussed how Twitter's geolocation data can be exploited to determine a user's location. However, Twitter is also susceptible to other location-based abuses. The paper *No Place to Hide* emphasizes the possibility of individuals inadvertently revealing their location information on Twitter, even when they do not employ conspicuous location-related phrases or geotagging. The study proposed Jasoos, a Naive Bayes-based system for determining the location of tweets. The Jasoos algorithm was tested on a large dataset of tweets and demonstrated that it can accurately guess the locations of a significant proportion of tweets despite the absence of evident location-revealing terms.

The Long Road to Computational Location Privacy analyzed various architectures and procedures for location privacy preservation methods (LPPMs) to prevent location privacy leaks. The article divided LPPMs into different categories, mix-zones, generalization-based, dummies-based, perturbation-based, obfuscation-based, or encryption-based mechanisms. In addition, the essay presents an overview of various LPPMs, as well as their architectures and methods, highlighting their strengths and shortcomings in preserving location privacy in the context of mobile users and LBSs. For instance, numerous online processes and socially conscious techniques to create mix-zones are presented. Mix-zones require the use of pseudonyms rather than real identities. The paper primarily goes over Dummies-based mechanisms which create fake users, known as dummies, to conceal the real users. Perturbation-based mechanisms are proposed to protect a user's exact location and online mechanisms based on time-to-confusion and CAP (confusion and accurate perturbation) are proposed to effectively add noise to user's location while trying to mitigate cons of Perturbation mechanisms. The paper *No Place to Hide* also proposed potential countermeasures, such as

the creation of a warning system that can alert users when they use potentially location-revealing words in their tweets.

So far we have only looked at negatives of geolocation data, however if used effectively, sharing of geolocation data can have its perks. Understanding the distinctions between social-driven and purpose-driven location sharing is crucial for the ubicomp community, as the paper *Rethinking Location Sharing* argues. The study found substantial variations between social and purpose-driven location sharing, with social-driven location sharing using semantic place names, blurring location information, and using location information to draw attention and promote self-presentation. Purpose-driven location sharing, on the other hand, was more closely related to one-to-one and one-to-few sharing, in which the requester had a specific need for the user's location.

While being used without proper education the location tracking feature on virtually all social media is dangerous as seen in our study with the GeoJSON information from a user's tweets having the capability to identify their patterns and behaviors. However, if used with less specific information, such as within a mile radius, it can be advantageous for users. Some benefits include creating localized content, increasing engagement, promoting events, and receiving personalized recommendations. By keeping location data a user might be able to connect with local audiences and businesses as well as receiving more personalized recommendations. The main benefit we found was for event organizers or businesses to reach potential customers or attendees in range of the location of their business or event. By using location-based hashtags and targeting users who are in the vicinity, they can increase the visibility of their posts and attract more people to their event or business. This can be achieved without the need for precise location data that we focused on in this study, and provide more robust privacy in comparison to being able to know precise day to day movements.

Closing

Our analysis of Twitter data revealed that user-supplied GeoJSON data accounts for a small percentage of all tweets shared on the platform, accounting for only 1.7% of the 21,348 tweets in our dataset. This finding supports our hypothesis that the percentage of user-supplied GeoJSON data is less than 25%. Furthermore, our analysis of the 29 users who willingly shared GeoJSON data in our dataset revealed that the majority of these users do not tweet their location data on a regular basis. Only 24 of the 29 users tweeted their GeoJSON data multiple times, with the top two users tweeting their GeoJSON coordinates 14 and 7 times, respectively.

While the small sample size prevented us from deriving personal patterns from these users' GeoJSON tweets, our analysis highlights the potential privacy concerns associated with sharing precise location data on social media platforms. Users may not fully comprehend GeoJSON's level of precision and its potential implications, such as the ability for others to track their movements. Overall, our research sheds light on the use of GeoJSON data on Twitter as well as the habits of users who willingly share their location information. Our findings can help to inform discussions about privacy and data protection on social media platforms, as well as the

need for users to be more conscious of the information they disclose online. Future research can expand on our findings by analyzing larger and more diverse datasets and investigating methods for deriving personal patterns from location data shared on social media platforms.

Works Cited:

Twitter Location Analysis [Computer software]. (2023). Retrieved from https://github.com/atramirez/Twitter-Location-Analysis

SnScrape [Computer software]. (2023). Retrieved from https://github.com/JustAnotherArchivist/snscrape

Bujlow, T., Huguenin, K., & Troncoso, C. (2019). No Place to Hide: Inadvertent Location Privacy Leaks on Twitter. Proceedings on Privacy Enhancing Technologies, 2019(4), 72-91. https://doi.org/10.1515/popets-2019-0064

Kalogridis, G., Kourtellis, N., Papadopoulos, S., & Rodrigues, R. (2019). The Long Road to Computational Location Privacy: A Survey. IEEE Communications Surveys & Tutorials, 22(3), 1893-1922. https://doi.org/10.1109/COMST.2019.2907818

Gross, R., Acquisti, A., & Cranor, L. F. (2010). Rethinking location sharing: exploring the implications of social-driven vs. purpose-driven location sharing. In Proceedings of the 12th ACM international conference on Ubiquitous computing (pp. 85-94). https://doi.org/10.1145/1864349.1864363