Analysis of Twitter 2.8 Billion User Profile Breach

Summary

This document discusses the breach of about 400GB of information on 2.8 billion Twitter (now X[1]) users that was leaked in January, 2025. TL;DR: **Most or all of this data is real, and Twitter *does* have over 2.8 billion accounts in its database. No Emails/phones/passwords were present, but there is at least a possibility that they were present in the breach but not leaked.**

It may be helpful to completely **ignore the 200M file that I released** (the "mashup" that appended data from the 2.8B dataset to the 200M dataset). It is pretty much irrelevant to the 2.8B breach[2].

Sorry if this is bit messy/ugly/incomplete, I spent little time formatting and/or proofreading it.

PART I: Background and Initial Response

Background

On January 23, 2025, an anonymous user going by the handle "ebiuprsy" leaked a massive ~400GB dataset of 2.8 billion Twitter user profiles on a popular hacking website[3]. The subject was "[2022] Twitter account data for 2,873,410,842 accounts (103 GB)". This was by far the largest social media breach in terms of numbers of users and quantity of data[4].

The data contained the following fields:
ID,name,screen_name,location,url,description,protected,verified,followers_count,friends_count,listed_count,favourites_count,statuses_count,default_profile,default_profile_image,last_status_created_at,last_status_source,created_at,utc_offset,lang.
Notably, no emails, phones, or passwords were leaked[5]. However, it likely contained personal data for 2.8 billion users per the GDPR[6].

1 For simplicity, I will primarily use "Twitter" since the data was taken before the company was renamed.
2 To answer the obvious question: I released it as a last resort to make sure X was aware of the breach, as well as the general public, to whom I feel this should be newsworthy as the largest ever social media berach.
3 https://breachforums.st/Thread-2022-Twitter-account-data-for-2-873-410-842-accounts-103-GB
4 Facebook and LinkedIn each had scrapes of around 500-600M users around 2019-2021, of about 50GB-150GB of data when in text (CSV) form.
5 There are indications suggesting that there is at least a possibility of more data having been taken
6 The GDPR says that "personal data" includes information relating to an "indentifiable person", and an "online identifier" can be used to identify someone. "identifiable person" treats an "online identifier" as personal data, per https://gdpr-info.eu/art-4-gdpr/. says "The username is personal data if it distinguishes one individual from another regardless..." [source: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/personal-information-what-is-it/what-is-personal-data/what-are-identifiers-and-related-factors/]

Initial Response – January, 2025

The person who leaked this made one critical error that caused this dataset to remain hidden for a long time: they started the subject with "[2022]", which likely led people to assume that this was an old breach. And people likely either assumed it was 2.8 *million*, or assumed it had to be fake since Twitter claims roughly 600M users (fake data is not uncommon on Breach Forums). There was also the issue that the data was hard to download[7].

The "black hats" that *did* look at the breach and understood it didn't care about it: it did not contain passwords, emails, phones or other sensitive information. Security researchers that monitor Breach Forums seem to have all overlooked it.

Somehow, the largest social media breach ever flew under the radar.

Secondary Response – March, 2025

After I released the 200M dataset appended with data from the 2.8B dataset, word started spreading online. It made the rounds briefly on Twitter, but again there was a lot of disbelief. xAI's Grok gave varying responses, one of which said that it "*seems unlikely... far exceeds the user base, suggesting exaggeration or misinformation... a possible mix-up... evidence leans toward this being false or overstated*".

The was a lot of disbelief, however, since nobody could find any references to Twitter having anywhere near 1 billion accounts (I do not use a Twitter account, so I could not easily reach them). So, as I started writing this 3 days later, only 1 article has been written about it, and zero mainstream media attention.

PART II: VERIFICATION

Verification Step 1: Reconciling 2.8B users versus 600M users

Many people have heard that Twitter has roughly 600M users, or predictions that it would hit 1B in the future. If true, a 2.8B user dataset would obviously have to be fake. So this question needs to be answered before spending time analyzing the data. This was a big red flag for me, as it is for many others.

What few people realize is that social media companies rarely publicize the total number of users that have accounts. Instead, they have a common metric: Monthly Active Users (MAU). This is not officially defined, but a typical definition would be the number of accounts that logged in within the past month.

7 It was originally only available as a torrent. It took me over a day to download the entire dataset, and many people complained of no "seeds" (without which it is nearly impossible to download it in its entirety).

Simply knowing about Monthly Active Users reconciles the 2.8B users in the dataset with the 600M figure commonly seen. Without further details, it is impossible to say if 2.8B could be accurate... but at this point, it is at least plausible, and we cannot rule it out. So on to the next step.

Verification Step 2: Are there really 2.8B users in the dataset?

When answering this question, I ignore all the supplemental data, and focus on *whether there are 2.8 billion valid user IDs*.

To answer this question requires either looking at the data, or trusting someone who has. While most people cannot easily manipulate 400GB of data, it should be no problem for most security researchers and tons of data nuts like myself.

My method involved taking a representative sample[8] of 100 users, and looking up their screennames at https://x.com/[screenname][9]. This would let me know if an account existed, was suspended, or did not exist. This also returned the user ID, which I could then compare with the one in the dataset. For the screennames that did not exist, I then went to https://twitter.com/intent/user?user_id=[user ID] and if it returned a screenname, I then checked that new screenname[10].

Of the 100 users, 92% had both a valid Twitter user ID and a screenname that matched what the Twitter database currently has.

There were another 4% where Twitter reported the user ID as unknown, 2% where the screenname had changed, and 2 that were suspended (and Twitter did not return an ID). That means that 96% of the records in the sample from the 2.8B dataset had either a valid ID or valid screenname or both. The 4% that is unaccounted for were likely deleted.

I'm sure if you asked a statistician (and trust my numbers, or run them yourself), they would say that it is statistically impossible that Twitter has less than 1 billion accounts. Or even less than 2 billion.

Verification Step 3: Is the rest of the data accurate?

Honestly, I have not spent much time on this, as there is little need: if someone was able to enumerate 2.8 billion Twitter accounts, it would make no sense for them to fake the other data.

However, *spot-checking* quite a few accounts shows *realistic* data (e.g. that could not have easily been forged, such as 248 status updates when archive.org shows 244 status updates the week before).

8 I took every 100,000th line from each file to get a representative sample, sorted randomly via EmEditor, and took the first 100 lines.

9 I would have preferred a larger sample, at least 1,000, but manually went to each webpage to avoid any possibility that it would be considered scraping or abusing their API. The problems with not intentionally violating the law.

10 This step allowed me to find valid data in the 2.8B dataset where the screenname had been changed. For example, if ID 10 was listed as having a screenname John, but now the ID 10 has a screenname Johnny. In this case, a search for screenname “John” failed but I was able to determine that the ID 10 is a valid Twitter ID.

Further, the 200M 2023 breach merged with the 2.8B dataset makes spot-checking easy, by comparing the number of followers from the 2023 breach (taken in late 2021) with the data in the 2.8B breach (taken in November, 2022)[11], as well as comparing the created_at dates. Of the 100 representative sample, when Twitter showed a protected account, sure enough, the 2.8B dataset showed it as protected as well.

I've seen enough to be convinced that at least *much* of the data beyond user ID and screenname is accurate, and to me it seems far-fetched that any of the data would have been fabricated given how much is real. It would make no sense to do so.

Part III: Was this just a scrape of public data?

No, no, no.

A scrape either expands on known data (such as taking a list of known screennames and looking up the profiles), or expanding on guesses of data (such as looking up IDs starting at the number 1, and continually incrementing the ID). A scrape of *public* data would entail doing that without accessing Twitter's database/API.

The 2023 breach used the Twitter API[12], but could also be considered a scrape. The person behind it took billions of email addresses found in other breaches, and used a vulnerability in Twitter's API to see if they exist, and if so, get profile data.

However, this 2.8B breach is very different. Somehow, the person who obtained the data was able to *enumerate every valid Twitter user ID*[13]. Twitter does not, as far as I have seen, have any method using their API to do that. And a scrape starting at 1 and going up would take forever[14].

Part IV: GDPR

I include this simply because of the lack of media attention: I'm pretty confident that would change if it were clear that this was 2.8 billion user records meeting the GDPR definition of a data breach. If so, Twitter would at least be required to make determinations regarding the breach.

I know very little about the complex GDPR, so I cannot say whether this is considered a data breach. However, I submit the following:

GDPR Article 4 Section 12 defines a "personal data breach" as "*a breach of security leading to the accidental or unlawful destruction, loss, alteration,*

11 There is a flaw in this logic, though: in theory, the 2.8B dataset could have been faked by using the 200M dataset as "guidance". But that is an extremely unlikely possibility.

12 Proof of this would take another document similar to this one, I won't burden you with reading that, nor do I plan to burden myself in writing it. But logic strongly suggests that it would not be possible to, for example, determine that "mary1234...@gmail.com" was Twitter user "mary smith8" and not "mary j smith", without accessing the Twitter database.

13 In theory, they could have enumerated all screennames, rather than IDs... but the evidence says otherwise. Another document for another day if people don't believe it.

14 Not quite literally forever. But effectively forever, with 64-bit user IDs, to get all of the valid ones. It would take less time, but still unrealistic amount of time, to try to use a smart algorithm to guess user IDs as the format of them is known. Again, I'll save that for another document that I don't plan to take the time to write.

unauthorised disclosure of, or access to, personal data transmitted, stored or otherwise processed;". To me, this means that *if* the 2.8B dataset includes personal information that Twitter stored and did not intended to disclose, the GDPR considers it a data breach.

GDPR Article 4 Section 1 defines "personal data" as "*any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an* ***online identifier*** *or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;*" To me, this means that if user IDs are considered an "online idenifier", they are personal data.

UK's ICO (Information Commissioner's Office) states that "*An individual's social media 'handle' or username, which may seem anonymous or nonsensical, is still sufficient to identify them as it uniquely identifies that individual. The username is personal data if it distinguishes one individual from another regardless of whether it is possible to link the 'online' identity with a 'real world' named individual.*"[15] To me, this confirms that user IDs and/or screennames are considered an "online identifier" and therefore personal data.

To me, it sounds like a good case could be made that the GDPR considers this a data breach. **If emails, phone numbers, and/or passwords were taken, that could push it into the category of a risk rating that requires reporting**.

PART VI: Misc.

Who exfiltrated the data?

This is not known. An anonymous Breach Forums user (an account created the same day as the post, named "ebiuprsy") leaked the data on January 23, 2025. They said very little. The one significant piece of information provided outside the dataset was a brief README.txt file that ended with emotionally charged words "Heil Elon".

Two other clues are the timing: when the data was taken, and when it was leaked. It was taken in November, 2022[16], at the same time that employees were learning that mass layoffs were going to be a reality, and leaked on January 23, 2025, 3 days after President Donald Trump was inaugurated and it was known that there might be mass layoffs of federal employees.

Those clues strongly suggest that this came from an ex-employee that was laid off, not happy with it, and possibly was hoping this data would get people thinking that federal employees concerned about being laid off might behave the same way.

Further, while anyone with access to the Twitter API could have scraped profile

15 https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/personal-information-what-is-it/what-is-personal-data/what-are-identifiers-and-related-factors/

16 According to the person who leaked it. I have confirmed to the minute when the data was exfiltrated, but choosing not to include that.

data (albeit subject to rate limiting), nobody seems to know how someone could have enumerated *every* user account other than having access to the internal Twitter network[17].

The only 2 explanations that come to my mind are that this was done by an employee, or an extreme hack (as opposed to simply scraping the data, as was the case with previous breaches).

Statistics on 2025 2.8 Billion Breach

Total Size:	About 400GB[18]
Number of files:	288 "users" files, 94 "extra" files.
Records per file:	10,000,000 (plus a header)
Number of records:	2,873,410,842[19]
Date of exfiltration:	November, 2022[20]
Rate of exfiltration:	I am not releasing this[21].
Date of leak:	January 23, 2025[22]
Unique screennames:	2,847,814,795+[23]
Duplicate screennames:	Under 10 million[24]
Unique user IDs:	2,873,410,842[25]
Timing of CSV Files:	They were created *after* the data was exfiltrated[26].
Duplicate user IDs:	0
Data Sort Order:	The data is in order by Twitter user ID.

17 I have looked at the Twitter API, and asked on Breach Forums, and nobody has come up with an explanation. "Brute Force" (trying all screennames and/or user IDs) would not be realistic. Screennames would require 2,922,075,441,452,077,677 lookups (for just 15 characters, that wouldn't get the 16-20+ character screennames). User IDs would require at minimum around 5 quintillion lookups.

18 408GB including the "extra" data files (timezone/language), 392GB excluding them.

19 See the README.txt file, verified by counting the number of records in the CSV files (a linecount will show a higher total since some records span multiple lines)

20 According to the person who leaked the data. I was able to pinpoint the timeframe that the data was exfiltrated, down to the minute, but not including that here.

21 I have determined the amount of time from the start of data collection to the end of data collection, but am choosing not releasing it. I don't want to give away all my work.

22 See the Breach Forums post.

23 By my count, there are 2,847,814,795 screennames that have *no* duplicates. The number of unique screennames is higher as this count excludes any screennames that appear multiple times.

24 9,767,683 screennames appear twice, 173,690 appear 3 times, 5,665 appear 4 times, and another 600 or so appear as many as 13 times. These duplicates appear to be *valid*, e.g. due to screenname changes.

25 I wrote a program to go through each of the 288 CSV files, and each user ID is higher than the previous one, which means that there are no duplicates.

26 I am not releasing the method I used to determine this.

Misc. Screenname Notes

While the vast majority of screennames are limited to the letters a-z, digits 0-9, and underscore, there is a screenname with a space in it ("Knight 9999").

Screennames used to be limited to 20 characters, but most accounts created on/after June 3, 2009 were limited to 15 characters. Yet a 26-character screenname "zyxwvutsrqponmlkjihgfedcba" exists.

There are numerous accounts with screennames prefaced by "erased_" and followed by the user ID. For example, "erased_1153812924684615680".

There are numerous accounts with a name (not screenname) of "Mysteriously Unnamed" and a screenname of "Mysteri" followed by an 8-digit number (with leading zeroes if needed).

Another mystery is "get_randlet". There are thousands of accounts that end with the characters "get_randlet" and one extra character (some truncated), most/all from Brazil.

About ThinkingOne

First, I am not a hacker (at least not by the mainstream definition): I'm a data enthusiast. I live and breathe data, and I am very careful to never intentionally violate any laws.

Some people like sitting down with a long book, or an intricate 2,000 piece puzzle. Me? Give me a 400GB dataset with 2.8B records, and it's like that book and puzzle at the same time. It starts off with glancing at the data, asking questions, and paying attention as the data tells a story. Figuring out the exact time the data was collected, the order it was collected in, whether it came from a backup or API, finding test accounts, how many servers generate user IDs, what oddities appear in the data. I assume most neurotypicals know this joy, right?

My goal in releasing the data and writing this was to ensure that X is aware of what appears to be the largest social media breach in history (at least by user count), that could possibly be a reportable breach (e.g. if phones/emails/passwords were taken). Despite having tried to contact them in several ways, I have not received a response, nor any sign that they are aware of the breach. It is now 3 days after releasing this data, I still have *zero indication* that X is aware of the breach. Hence, this document.

Maybe this will nudge someone who is able to ask the question "*Um, did someone steal 2.8 billion emails, phones, and passwords*?" until an answer is provided.