Does geographical location have an impact on data samples extracted from Twitter?

Rositsa V Ivanova*, Stefan Sobernig*, Mark Strembeck*†‡
*Vienna University of Economics and Business (WU), Vienna, Austria
†Secure Business Austria (SBA), Vienna, Austria
‡Complexity Science Hub (CSH), Vienna, Austria

rositsa.ivanova@wu.ac.at, stefan.sobernig@wu.ac.at, mark.strembeck@wu.ac.at

***Abstract*—We report on an experiment that used ten different machines running on a standardized cloud platform in five different geographical locations around the globe (Frankfurt/Germany, Mumbai/India, Sydney/Australia, Seoul/South Korea, Virginia/USA) to collect datasets using Twitter's public free-of-charge API. Each of the ten machines extracted the tweets at the exact same time and using the exact same Twitter API parameters. We found that the characteristics of the datasets collected in different locations vary considerably, potentially affecting any analysis performed on such location-biased data. For example, the number of exactly identical tweets (i.e. all 90 metadata attributes of the tweets are the same for all ten machines) lays only between 0.15% and 20%. Based on these findings, we derive recommendations on how to mitigate the location-bias in practice.**

***Index Terms*—Data analysis, Data collection, Data quality, Network Science, Social Networks, Twitter**

I. Introduction

In recent years, datasets consisting of Twitter messages (so called tweets) have been extensively used for studies in various scientific fields ranging from sociology and psychology to network science, see, e.g., [1]–[6].

Twitter's popularity among researchers as a data source results from clear terms of use and a well-maintained application programming interface (API) that allows for an automated collection of large datasets. Twitter's API is publicly available and offers various API endpoints to access different data, e.g. one can collect tweets that include user-defined combinations of keywords and/or hashtags. The downside of using Twitter's open (free-of-charge) API as a data source is that researchers are confined to samples of data. This is, on the one hand, due to technical restrictions of the API (varying service levels and rate limits) as well as different commercial API options offered by Twitter.

While Twitter provides access to commercial and special-purpose academic accounts without or with lowered restrictions (e.g., increased rate limits or transfer volumes), the costs of obtaining unlimited commercial access are prohibitively high for most researchers. Moreover, an academic account is only granted on request provided that the respective research topic and the researcher's current job position meet certain criteria. Hence, Twitter's free API endpoints remain very popular and widely being used for collecting datasets.

A second reason for researchers usually being limited to data samples is the dynamic, ever-changing nature of the whole population of messages and users over time. Twitter users can modify or remove content as well as their relationships to other users. Moreover, Twitter's content moderation remains a blackbox for the most part. These conditions lead to varying samples depending on when a dataset has been collected. The short- and long-term decay of Twitter data has been actively investigated by prior research [7].

Therefore, the question must be asked whether different samples are identical for researchers who are extracting tweets regarding the same topics (i.e., the same keywords and/or hashtags) but at a slightly different time or from a different location (for example, another Internet node, country, or continent).

Previous research investigated whether certain types of samples provided by Twitter are representative for all existing tweets including a particular hashtag. To this end, Morstatter et al. [8] explored Twitter's Streaming API[1][2] which promises a 1% sample of all tweets for a particular hashtag. They compared certain characteristics of the retrieved dataset to the full collection of tweets collected via the commercial Twitter Firehose API. Our study complements the work of Morstatter et al. by investigating if the datasets that researchers collect via Twitter's search API [3] are comparable or differ when considering the researcher's network-topological (Internet node) and geographical locations (country, continent).

Thus, in contrast to previous studies [7]–[11] which compare datasets collected via different Twitter APIs, our work compares datasets retrieved via same Twitter API but from different geographical locations. Our study aims at making recommendations for researchers who collect datasets from Twitter in practice.

Section II gives a brief introduction to important concepts used in our work. Section III then documents our setup for orchestrated data collections of Twitter data from different locations and gives an overview of the resulting data collections. In Section IV, the key findings are presented.

[1]https://developer.twitter.com/en/docs/twitter-api (v.1.1)
[2]All links have been last accessed 28.07.2022
[3]https://developer.twitter.com/en/docs/twitter-api

Section V iterates over the known limitations, followed by the recommendations for other researchers VI. An overview of related work is given in Section VII. Section VIII concludes the paper.

II. Background

a) Twitter API: At the time of writing, Twitter offers two versions of its API in parallel, version 1.1 (v1.1) and version 2 (v2). Tweets can either be collected as a real-time data stream or as a historical collection. Both of these data extraction methods can be used with Twitter's Standard (free-of-charge) API v1.1 and promise to provide the user with a sample of all tweets on the requested hashtag. In contrast, Twitter's commercial Premium API v1.1 enables users to extract all tweets that have been sent over the last 30 days.

Twitter's API v2 has been first introduced in 2020. Like v1.1, API v2 provides free access via the Essential and Elevated access types which enforce different rate limits for tweet extraction. In addition, API v2 offers a new access type for Academic Research[4] which is available for researchers at academic institutions or universities. An Academic account also enforces monthly rate limits (10 million tweets per month) but allows users to access all tweets that have been published since 2006.

For many scientific purposes, the analysis of historical tweets is preferable as compared to a real-time analysis, because a tweet's attributes may change over time (e.g. the favorite/like-count or the retweet count might change). Previous studies found that for digital messages most responses happen within the first day and that barely any response occurs three days after a message has been sent (see, e.g., [12]–[14]). Thus, in order to consider dynamic message attributes such as retweet-count or favorite-count in a data analysis, tweets should be collected three or more days after they have been published and the values of dynamic attributes have settled. In this context, it has to be mentioned that Twitter's free API only provides access to historical tweets if they are not older than 6-9 days[5], leaving a window of 3-6 days for the collection of tweets related to a particular topic of interest.

b) Twitter data model: Each tweet collected via one of Twitter's APIs contains 90 different attributes (such as status id, tweet text, retweet count, favorite count, or attached media). These 90 attributes can be separated into categories based on the consistency of their respective attribute values. The first category includes attributes that are static over time such as a tweet's status id, the sender's user id, and the tweet's text. The second category includes attributes whose values may change over time (i.e. dynamic attributes). Examples of dynamic attributes are the retweet-count or the favorite-count of a tweet. The third category includes attributes which are dynamic but typically do not change in the short-term. Examples for such attributes are the user name or the status of a user account (e.g. a user may be a "verified" user or not).

c) Network derivation: The tweets, including the 90 attributes mentioned above, can be used to derive various types of networks (e.g., a retweet network or a messaging network based on @-mentions). Such network structures are then used for applying different network analysis techniques (see, e.g., [15]–[18]). Moreover, the hashtags and keywords included in a tweet can be used to derive topic models. The co-occurrence of the same hashtag(s) in different tweets can also be used to identify relations between users who are discussing similar topics (see, e.g., [19]).

III. Multi-site Twitter mining

a) Infrastructure: For our study, we simultaneously collected multiple datasets based on the same hashtags but from different geographical locations. The data collection was performed using a distributed system of ten virtual machines (VMs) running in five different geographical locations around the globe (Frankfurt/Germany, Mumbai/India, Sydney/Australia, Seoul/South Korea, Virginia/USA). For each of these five locations we rented two independent VMs via Amazon Web Services (AWS)[6].

In order to coordinate the data collections procedures running in the 10 VMs, an additional device, running in Vienna/Austria, acted as an orchestrator. Moreover, a storage server, also running in Vienna/Austria, was used for permanently saving the data collected from the 10 VMS (see Figure 1). First, the orchestrator dispatches the data collection scripts to the VMs and the storage server (orange arrows in Figure 1). By executing these scripts, the VMs perform a scheduled data collection. Moreover, in 15-minutes intervals the data is fetched from the individual VMs and stored on the central storage server (blue arrows in Figure 1). In order to ensure that all VMs begin collecting Twitter data at the exact same time, we automatically synchronized the corresponding execution times.

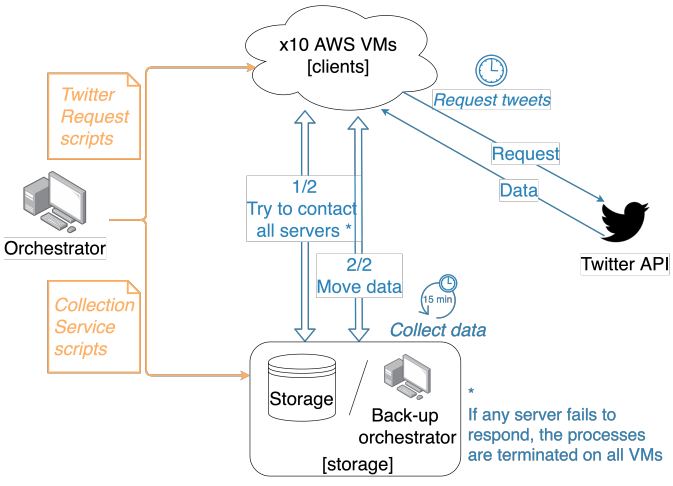

Fig. 1. Architecture for multi-site Twitter mining

[4]https://developer.twitter.com/en/products/twitter-api/academic-research

[5]https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/guides/standard-operators

[6]For our study, we selected the lightest VM option that was available at the time of writing (512 MB RAM, 1 CPU core and 20 GB of hard disk storage).

Our data collection scripts accessed Twitter's API via the "rtweet" library[7]. Moreover, each of the 10 VMs used a separate Twitter developer account for data collection. The physical locations of the corresponding Twitter accounts correspond to the respective AWS regions.

b) Data sets: Table I presents a list of the collected hashtags, the maximum number of tweets collected, the collection time window (from, until), as well as the day of data collection. For the first group of hashtags, our aim was to cover a wide variety of topics (e.g. major events, pop culture) and collection sizes. The second group of hashtags is related to the currently ongoing war in Ukraine.

Moreover, for two of the hashtags related to the Ukraine war, we performed the tweet collection procedure twice in order to explore, whether the differences in the number of collected tweets per VM will persist during a second collection. Each of the 17 data collection procedures are denoted by the prefix "C" and are numbered chronologically. Each of the collections consists of ten individual datasets, collected by the ten different VMs respectively. Each collection of tweets was extracted once and the extraction was not restarted if errors occurred during the extraction procedure (e.g. the Twitter API closed the connection; this extraction procedure was chosen on purpose, for further details see below).

TABLE I
OVERVIEW OF TWEET COLLECTIONS FOR A GIVEN HASHTAG INCLUDING THE NUMBER OF MESSAGES (#T: MAXIMUM NUMBER OF TWEETS OBTAINED FROM ANY OF THE 10 LOCATIONS), TIME WINDOW FOR DATA COLLECTION, AS WELL AS COLLECTION DATE. "REDO" DENOTES A REPEATED COLLECTION PERFORMED ON A SUBSEQUENT DAY.

Coll.	Hashtag	#T (max)	From	Until	Coll. on
C01	covid19	310.879	18.11.21	21.11.21	25.11.21
C02	BlackFriday	250.570	23.11.21	26.11.21	29.11.21
C03	BlackFriday	550.361	26.11.21	28.11.21	01.12.21
C04	Omicron	526.927	29.11.21	03.12.21	06.12.21
C05	HongKong	18.957	19.12.21	24.12.21	29.12.21
C06	Happy-Birthday-Taehyung	2.577.930	29.12.21	01.01.22	04.01.22
C07	Djokovic	218.966	04.01.22	08.01.22	11.01.22
C08	tsunami	125.793	14.01.22	20.01.22	24.01.22
C09	Ukraine	85.641	18.01.22	22.01.22	25.01.22
C10	SuperBowl	1.826.490	13.02.22	15.02.22	20.02.22
C11a	Putin	197.941	21.02.22	23.02.22	27.02.22
C11b	Putin (redo)	197.904	21.02.22	23.02.22	28.02.22
C12	Ukraine	436.420	21.02.22	23.02.22	25.02.22
C13a	Putin	590.680	23.02.22	25.02.22	01.03.22
C13b	Putin (redo)	585.561	23.02.22	25.02.22	02.03.22
C14	Ukraine	1.144.923	10.03.22	13.03.22	17.03.22
C15	Ukraine	1.474.915	15.03.22	20.03.22	23.03.22

IV. FINDINGS

We analyzed the 17 collected datasets (see Section III) from three different perspectives. First, we analyzed the data based on the geographical location of the respective VM (node-level data). Second, we performed a content analysis to compare the exact content of the collected tweets (i.e. attribute-level data). Third, we analyze and compare the retweet-networks that can be derived from the datasets.

[7]https://cran.r-project.org/web/packages/rtweet/rtweet.pdf

A. Node-level data

In the first step, we look at the number of unique tweets collected per VM and hashtag. A unique tweet is identified by its status id. Therefore, duplicates can easily be detected (in contrast to retweets, duplicates are not counted as additional tweets). The total number of tweets per location and hashtag are shown in Table II. For 10 of our 17 collections, we observe an apparent size difference of at least one of the datasets. In these 10 cases, the API connection was closed by Twitter (see also the names of logged messages in Table II).

Note that for collections C05 and C12, the differences in the sizes are less than 5%, therefore one might still consider them as regular collections. However, the variations in the remaining 8 collections are considerable (see Table II).

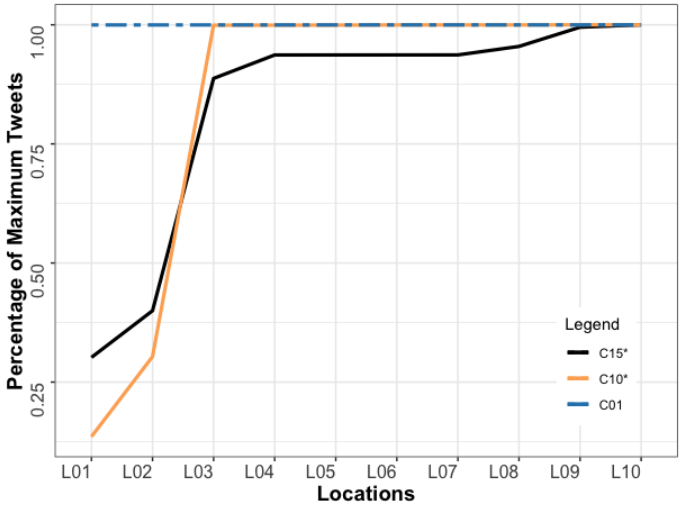

Fig. 2. Cummulative fraction of tweets (unique status ids) collected per location, for collections C01 (blue), C10* (orange), C15* (black)

In a basic research scenario, one may be tempted to collect a single dataset, from a single location. The numbers documented in Table II already indicate that such a single-location collection might be biased. Therefore, we now explore how many additional unique tweets can be found by collecting another dataset (including the same keywords/hashtags) from a different geographical location.

In this context, Figure 2 depicts a worst-case scenario for three collections (here C01, C10*, C15*)[8]. Here, the x-axis depicts the ten VMs/locations (labeled from L01 to L10). The y-axis depicts the respective fraction of unique tweets (in %) that a particular VM/location adds to the overall dataset.

For example, for C10* the worst-case location (here Mumbai 1, see also Table II) provides only 13.5% of the maximum number of unique tweets. If a researcher decided to additionally send the same request from the second-worst location (L02, here Frankfurt 1), this would increase the percentage of unique tweets to just 30.3%. Only after including the dataset

[8]We define this worst case as one in which a researcher collects tweets from a single location (here L01) while the resulting dataset consists of the smallest number of tweets (as compared to the overall dataset resulting from a union of all datasets collected by the ten VMs).

TABLE II
NUMBER OF TWEETS COLLECTED PER DATASET PER HASHTAG AND THEIR RESPECTIVE LOGGED MESSAGES

Collection	Frankfurt 1	Frankfurt 2	Mumbai 1	Mumbai 2	Sydney 1	Sydney 2	Seoul 1	Seoul 2	Virginia 1	Virginia 2
C01	310.802	310.803	310.811	310.810	310.808	310.806	310.805	310.879	310.814	310.808
C02	250.557	250.559	250.560	250.560	250.561	250.559	250.567	250.570	250.563	250.560
C03	550.321	550.312	550.361	550.361	550.332	550.339	550.330	550.313	550.310	550.304
C04	526.804	526.811	526.803	526.804	526.813	526.815	526.927	526.830	526.809	526.812
C05	18.886	18.868	18.366	18.366	18.371	18.371	**17.883**[1]	18.532	18.935	18.957
C06*	2.351.611	2.577.831	2.574.730	2.577.930	2.575.355	2.574.470	38.985[2]	2.577.219	2.575.290	2.576.477
C07	218.831	218.827	218.841	218.842	218.841	218.842	218.965	218.966	218.858	218.836
C08*	125.789	125.793	**78.711**[1]	**78.711**[1]	**78.710**[1]	**78.712**[1]	124.518	**78.714**[1]	125.779	125.788
C09	85.640	85.637	84.406	84.406	84.409	84.412	85.622	85.622	85.640	85.641
C10*	554.498[4]	1.825.922	247.731[3]	1.825.615	1.825.616	1.825.577	1.826.056	1.826.146	1.826.476	1.826.490
C11a*	197.909	197.910	197.937	197.941	197.939	197.932	197.939	63.356[2]	197.907	197.909
C11b	197.876	197.873	197.902	197.902	197.904	197.914	197.891	197.897	197.862	197.862
C12	436.391	436.392	436.379	432.209[2]	436.401	436.409	436.420	436.399	436.409	436.409
C13a*	590.543	590.548	34.577[2]	590.600	590.646	590.674	590.653	590.680	590.607	590.601
C13b*	585.428	585.434	585.518	477.115[2]	585.588	585.561	585.547	585.549	585.430	585.436
C14*	1.068.768	1.100.893	1.100.851	1.100.832	1.144.885	1.144.892	1.144.923	150.033[2]	1.100.894	1.100.907
C15*	1.309.517[5]	1.408.257	590.046[3]	445.348[3]	1.381.884	1.381.923	1.382.012	1.382.025	1.467.636	1.474.915

Logged messages:
[1] Two confirmations that the script exited correctly
[2] Error in curl::curl_fetch_memory(url, handle = handle) : OpenSSL SSL_read: SSL_ERROR_SYSCALL, errno 104
[3] Error in curl::curl_fetch_memory(url, handle = handle) : transfer closed with outstanding read data remaining
[4] Killed
[5] Over capacity - 130

obtained from a third location (e.g. L03, here Sydney 2), one arrives at the maximum number of tweets (100%) for C10*.

Under the worst-case assumption, two of our collections would require three or more locations to obtain a dataset including at least 90% of all tweets. For five other collections it would requires two or more locations to collect a corresponding dataset (see also Table II). Nevertheless, in the best case a single location may already offer more than 90% of all tweets available (see, e.g., C01 in Figure 2).

However, note that, based on our results, we cannot give any recommendation for the "best" data extraction location. Yet, there is clear evidence that collecting a second dataset (for the same keywords/hashtags) from the same or a different location may considerably increase the number of unique tweets. A dataset that merges the tweets collected from three different notations will most often include the majority (at least 90%) or even all tweets (100%).

a) Hashtag Cooccurrence: For the hashtag cooccurrence analysis, we parsed the corresponding message texts (i.e. the publicly available text of a tweet). The text of a tweet is static and therefore always identical for each occurrence of a particular status_id. In contrast, other attributes, including the "hashtags" attribute of a tweet, might include varying values, even for the same status_id.

For the collections with a similar size across all locations (such as C01) we found that the hashtag cooccurrence does not show noticeable differences between the individual locations. However, as soon as a collection includes at least one outlier dataset (i.e. a location that extracted a smaller dataset) we observe clear differences in the resulting hashtag cooccurrence and, specifically, in the order of cooccurrence (i.e., from most to least frequent).

Table III depicts one such example for C10 in which the extracted hashtags for the two smaller datasets in this collection, extracted at Frankfurt 1 and Mumbai 1, are different. In the case of C15 (see Table IV), the differences in the results of the hashtag cooccurrence are not only found in the datasets that suffered from an error during data extraction. Here, the list of hashtags for one of the affected datasets (here Frankfurt 1) matches the list of the majority of the rest of the collected datasets. However, the two datasets from Virginia offer different hashtags, despite them receiving all available tweets from the API.

TABLE III
COMPARISON OF HASHTAG COOCCURRENCE (C10*)

Frankfurt 1	Mumbai 1	Rest
HalfTimeShow	HalfTimeShow	HalfTimeShow
PepsiHalftime	dogecoin	PepsiHalftime
SuperBowlLVI	PepsiHalftime	SuperBowlLVI
dogecoin	SuperBowlLVI	NFL
RamsHouse	NFL	RamsHouse
NFL	CryptoInu	SBLVI
SBLVI	metaverse	Bengals
metaverse	Eminem	RuleItAll
Eminem	bnb	Rams
NFTs	RamsHouse	NFTs

TABLE IV
COMPARISON OF HASHTAG COOCCURRENCE (C15*)

Frankfurt 1,2 Sydney 1,2 Seoul 1,2	Mumbai 1	Virginia 1,2	Mumbai 2
Russia	Russia	Russia	Russia
StandWithUkraine	StandWithUkraine	StandWithUkraine	StandWithUkraine
Putin	Putin	Putin	UkraineUnderAttack
UkraineRussiaWar	UkraineUnderAttack	UkraineRussiaWar	Putin
Russian	UkraineRussiaWar	Russian	RussiaInvadedUkraine
UkraineUnderAttack	Russian	Kyiv	UkraineRussiaWar
Kyiv	RussiaInvadedUkraine	UkraineUnderAttack	RussianUkrainianWar
UkraineWar	RussianUkrainianWar	UkraineWar	Russian
Mariupol	UkraineWar	Mariupol	StopPutin
Ukrainian	Kyiv	Ukrainian	Kyiv

B. Attribute-level data

For the attribute-level analysis, we first examined the number of exactly overlapping tweets per collection (i.e. all 90 attributes of a particular tweet are the same for all 10 datasets/locations). The results are summarized in Table V.

Figure 3 shows the maximum number of tweets per hashtag, the partial overlap, and the exact overlap between all 10 locations. Note that the exact overlap between the different locations is always relatively low. In cases where all locations within a particular collection have a comparable number of tweets the overlap lays between 8.4% (for C04) and 20% (for C02) only. For the cases where considerable differences in the dataset sizes were observed the overlapping percentage of tweets drops to numbers between 0.15% (for C10) and 14% (for C13b).

TABLE V
NUMBER OF SAME STATUS_IDS COLLECTED ACROSS ALL DATASETS AND NUMBER OF TWEETS WITH ALL EXACTLY MATCHING ATTRIBUTES ACROSS ALL DATASETS PER COLLECTION

Collection	Same Tweet	Exact overlap
C01	310.757	28.837
C02	250.523	51.771
C03	550.172	80.513
C04	526.672	44.169
C05	17.395	3.076
C06*	38.947	9.878
C07	218.785	32.611
C08*	78.707	12.802
C09	84.393	13.540
C10*	247.679	28.182
C11a*	63.318	12.488
C11b	197.807	35.451
C12	431.996	32.886
C13a*	34.550	6.923
C13b*	476.819	81.310
C14*	149.967	22.094
C15*	445.210	46.322

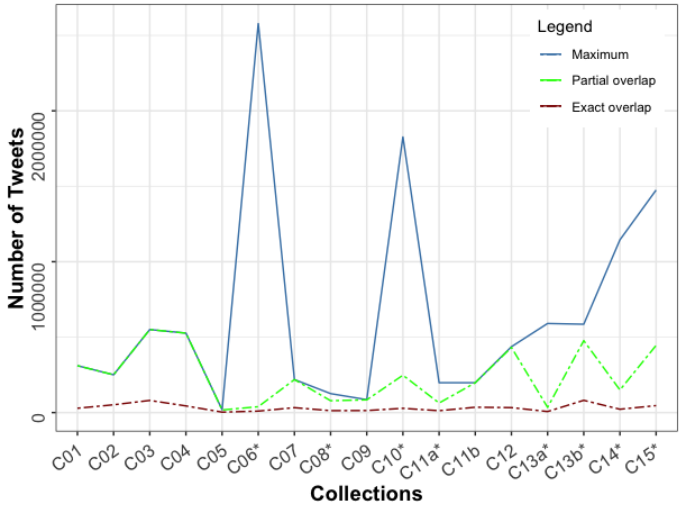

Fig. 3. Maximum number of tweets, Partial overlap, and Exact overlap between all 10 datasets per hashtags

Next, we analyzed tweets that have been extracted at all ten locations (i.e. a tweet with the same status_id has been extracted at each location) but show at least one difference across the 90 attribute values. We found that the relative number of partial overlaps (i.e. same status_id while other attribute values differ) is significantly higher than the number of exactly overlapping tweets (i.e. all attribute values are the same for all ten locations). Figure 3 show the maximum number of tweets (blue) and partially overlapping tweets (green).

In order to further analyze the partial differences between tweets with the same status_id, we use the classification of static and dynamic attributes mentioned in Section II.

As a little surprise, we also found variations in static attributes (e.g. retweet_location, mentions_user_id, hashtags). Such variations are difficult to explain, as the corresponding attributes are static and therefore are expected to remain stable over time. However, it should also be mentioned that the fraction of tweets which showed varying values for at least one static attribute was lower than 1%.

Moreover, we found unexpected mismatches between timestamps (e.g. for the quoted_created_at attribute) with an exact difference of 1 hour. Such mismatches, however, may simply be explained by different time zones. Yet, such time differences are still unusual since all timestamps are recorded in UTC[9].

Furthermore, we also found a few variations in the ”hashtags” attribute. Such variations are also unusual, since a tweet’s hashtag attribute is based on the hashtags in the corresponding text, and the text of a tweet is not supposed to change. Upon further investigation, we extracted the hashtags from tweet texts ourselves and discovered that sometimes not all hashtags found in the text are listed in the ”hashtags” attribute field.

Nevertheless, aside from such smaller and unusual differences between tweets with the same status_id, most differences can actually be tracked back to dynamic attributes such as the retweet_count. As a matter fact, the majority of such differences can be attributed to the 19 different counter attributes included with each tweet (e.g. 1.605.046 variations in approximately 526.000 tweets for C04 result from differences in one or more counter attributes, such as retweet_count). Figure 4 depicts the maximum differences in the retweet count values between all 10 locations per tweet per hashtag.

While in many cases the tweet counter attributes from at least one location show differences of the corresponding counter (e.g. retweet count) by 1 or 2, in some extreme cases the values may be up to 20 times bigger in certain locations (e.g. 23.585 retweets in four locations and 458.438 retweets in the remaining six locations for C03). On the one hand, this indicates that for the majority of the cases, a tweet’s attribute values are comparable, yet one should be cautious when using the exact count values.

Similarly, values of the favorite count attribute (i.e. likes) per tweet typically vary by 1 as shown in Figure 5. The biggest observed differences concern the values describing the number

[9]https://developer.twitter.com/en/docs/twitter-ads-api/timezones

of tweets that a user has liked (i.e., favorites_count) as seen in Figure 6.

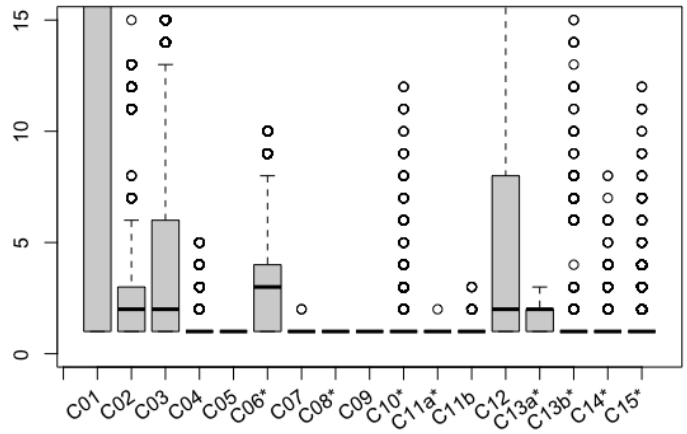

Fig. 4. Maximum difference in retweet count values per tweet per hashtag

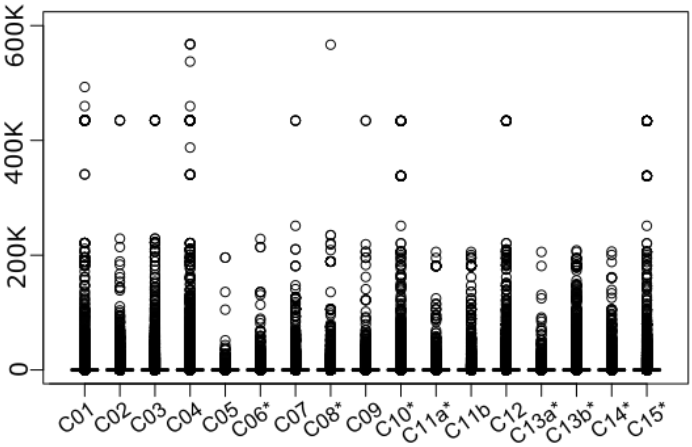

Fig. 5. Maximum difference in favorite count values (i.e., likes) per tweet per hashtag

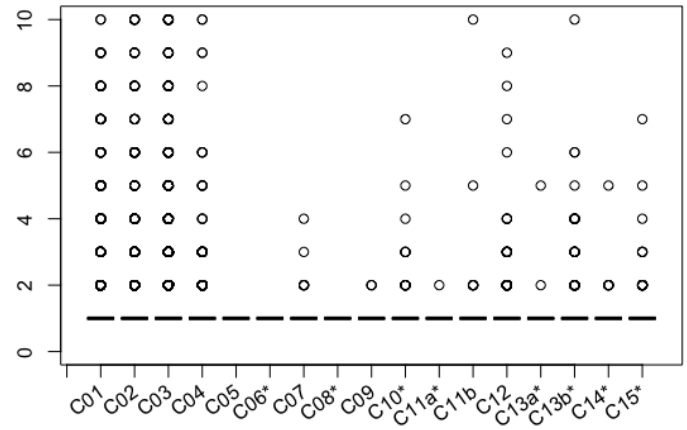

Fig. 6. Maximum difference in user favorites count values per hashtag

C. Network-level data

In order to evaluate the differences in the datasets with regard to network structures that can be derived from the respective tweets, we construct and analyze retweet networks for each of the locations. In this step, the objective is to investigate whether characteristic and substantial differences exist in the network structures derived from different datasets.

In order to derive the retweet networks, we first subset all retweets for all locations. Each tweet is then modeled as a node and edges connect an original tweet to the corresponding retweets. Next, we compute basic network measures (number of nodes, number of edges, number of connected components, maximum degree) to enable a basic comparison of the resulting networks.

In the collections where all locations provided a similar number of tweets, we found small differences in the basic characteristics of the derived networks. One such example for C03 is depicted in Table VII. However, as soon as an outlier dataset/location exists in a collection, we observe considerable variations between the retweet networks. Table VI displays one such example based on C15*, which contains three outlier datasets. Here, the datasets with the biggest variations ("Mumbai 1" and "Mumbai 2") lead to networks containing less than half of the vertices and edges of the remainder of the datasets. The differences of these two networks are also obvious when comparing the number of connected components and the maximum degree. However, we also see that in the case of the "Mumbai 1" dataset, the max degree is the same as for the non-outlier networks. Depending on the type of network analysis a particular researcher is planning to conduct, the results may thus differ significantly, depending on the location the respective dataset has been extracted from.

For example, the network resulting from the dataset extracted from "Seoul 1" includes 76.656 (or 7%) additional edges and 6.076 (or 5%) additional connected components as compared to "Virginia 2".

TABLE VI
NETWORK ANALYSIS MEASUREMENTS FOR THE RETWEET NETWORKS CONSTRUCTED USING EACH OF THE INDIVIDUAL DATASETS AND A MERGED DATASET, CONTAINING ALL TWEETS FROM THE INDIVIDUAL DATASETS FROM C15*

Dataset	$\|V\|$	$\|E\|$	Connected Components	Max Degree
Frankfurt 1	1.151.689	1.048.508	103.181	15.223
Frankfurt 2	1.239.081	1.128.180	110.901	15.223
Mumbai 1	531.705	481.243	50.462	15.223
Mumbai 2	404.316	364.838	39.478	9.473
Sydney 1	1.215.285	1.106.309	108.976	15.223
Sydney 2	1.215.290	1.106.322	108.968	15.223
Seoul 1	1.215.341	1.106.359	108.982	15.223
Seoul 2	1.215.336	1.106.350	108.986	15.223
Virginia 1	1.291.652	1.177.045	114.607	15.222
Virginia 2	1.298.073	1.183.015	115.058	15.222
All	1.298.647	1.183.473	115.174	15.223

V. LIMITATIONS

To the best of our knowledge the experiment described in this paper is the first to address the comparability of Twitter datasets collected from various geographical locations at the exact same time and using the exact same Twitter API parameters. Even though we aimed to conduct our experiment in a reproducible manner, our approach does have certain limitations.

TABLE VII
NETWORK ANALYSIS MEASUREMENTS FOR THE RETWEET NETWORKS DERIVED FROM EACH OF THE INDIVIDUAL DATASETS AS WELL AS THE MERGED DATASET (CONTAINING ALL TWEETS FROM THE INDIVIDUAL DATASETS IN C03)

Dataset	$\|V\|$	$\|E\|$	Connected Components	Max Degree
Frankfurt 1	441.065	395.160	45.911	31.391
Frankfurt 2	441.064	395.160	45.910	31.391
Mumbai 1	441.042	395.138	45.910	31.390
Mumbai 2	441.043	395.139	45.910	31.390
Sydney 1	441.042	395.137	45.911	31.390
Sydney 2	441.040	395.135	45.911	31.390
Seoul 1	441.055	395.151	45.910	31.390
Seoul 2	441.056	395.151	45.911	31.389
Virginia 1	441.045	395.143	45.908	31.389
Virginia 2	441.043	395.141	45.908	31.389
All	441.106	395.197	45.915	31.393

In order to exclude any influence resulting from differences in the technical environment that is used for executing the data collection scripts we chose to use a standardized cloud platform; in our case we used AWS VMs. While to the best of our knowledge all VMs in each of the locations are set up in the same manner, we cannot exclude that the results might have been slightly different if we used another cloud platform (e.g. Microsoft Azure).

Moreover, the iterative collection method used in our study (i.e. collecting 200 tweets every 10 seconds) originated from the limitations of the VMs that we rented for our experiment. Alternatively, all Twitter data could also be collected via a smaller number of requests (e.g. a batch of 50.000 tweets at once).

VI. RECOMMENDATIONS FOR TWITTER-BASED RESEARCH

Based on our findings, we can recommend the following measures:

R1. Use three or more locations: We recommend collecting Twitter data from a minimum of three independent locations and machines in parallel (while all extraction parameters other than location remain unchanged). Here, a unique location is defined as a distinct machine using a distinct Twitter account. The actual network-topological and geographical locations (e.g., using a second node within the same or in another AWS availability zone) did not show a significant difference in our setting.

This way, even in a worst case scenario, approx. 90% of tweets that are retrievable via the free search API can be collected (see Fig. 2). If unfeasible (i.e. if it is not possible to extract data from three or more locations/machines), aim for replicating the data-extraction procedure within a narrow time window from the same location, in order to extract a more representative dataset. Make sure to eliminate partial and exact (inter-source) duplicates from the combined dataset before analyzing the respective data. Additionally, based on the collection setup (i.e. programming language, library used) the output messages and logs of the collection process may be a good indication if the extraction procedure has been interrupted (i.e. an error occurred). API errors usually result in a dataset only representing a subset of all of the possible tweets which are available via the free version of the API.

R2. Use a three-day delay: Because some of the 90 attributes associated with each tweet are dynamic, the corresponding attribute values (such as retweet count or favorite/like count) may differ depending on the exact extraction time as well as depending on the location the data is extracted from. Previous research has shown the vast majority of all reactions happen within the first day and almost no reactions happen three days after a message has been sent [12]–[14]. Therefore, researchers who are not performing real-time data analysis tasks, should collect Twitter datasets with a three-day delay, after the dynamic attribute values have settled.

R3. Use consistent/stable attributes: Network structures derived from Twitter messages (such as retweet-networks or @-messaging networks) should either be derived from datasets that have been extracted with a three-day delay (R2) or should be based on static attributes only (such as @-mentions included in the text of a tweet). Table VIII shows a list of static and dynamic attributes which have been identical in our experiment for each unique tweet over all locations/datasets.

Moreover, remember that certain attributes of a tweet are directly connected to others. For example, a Twitter user name is also included as part of the URL referring the respective profile picture. Thus, if a user name is changed the corresponding URLs are adapted accordingly, and the previous URLs are no longer valid.

TABLE VIII
CONSISTENT ATTRIBUTES AND THEIR EXPECTED CHANGE OVER TIME

Attribute name	Expected attribute type
account_created_at	static
account_lang	dynamic
ext_media_type	static
is_retweet	static
lang	static
profile_background_url	dynamic
protected	dynamic
quote_count	dynamic
quoted_location	static
reply_count	dynamic
retweet_created_at	static
retweet_source	static
retweet_status_id	static
retweet_text	static
retweet_user_id	static
source	static
text	static
user_id	static
verified	dynamic

Note: For this evaluation, we used the status_id as a unique identifier. It is assumed to be a reliable attribute.

VII. RELATED WORK

Previous research has shown that factors such as API endpoint and access level used (see Section IIa) as well as the time of a (repeated) tweet collection influence the data in terms of it's representativeness as a *sample population*.

Morstatter et al. [8] compared the Streaming API to APIs offering bigger or full samples (e.g., academic API, Firehose). Data obtained from two different priority types of the

Streaming API (i.e., Spritzer and Gardenhose) were found to vary based on activity patterns of the Twitter users and their sentiments [10]. Another experiment by Pfeffer et al. [9] revealed using a different access level to the Sample API (academic, free) can lead to over- or under-representing certain user accounts in the corresponding sample. Furthermore, [20] investigated Twitter datasets when the filtering option is used with the Streaming API.

Pfeffer et al. [7] analyze and compare data collected via the various v1.1 and v2 APIs in terms of their coverage over time, showing an approximate drop of 10% in the number of tweets over time. This decay increases over longer periods of time up to 30% after four years. In addition, tweet metadata has been found to change over time [21]. Timondea [22] focused on the removal of tweets with strong political content over time. He found that 20-30% of tweets deemed potentially sensitive could not be recovered using the Search API, and 2-5% were not retrieved via the Streaming API. Kim et al. [11] compared the Stream, Search, and Firehose API endpoints by collecting data over a certain period of time using the same search parameters, finding that the covered user accounts, the quantity of the dataset, and the tweet content vary heavily between the different endpoints.

VIII. Conclusion

In this paper, we report our findings regarding the impact of geographical location when extracting data using Twitter's free-of-charge search API. Our work complements previous contributions focusing on different API endpoints as well as on different extraction times. We applied a multi-site Twitter mining approach for orchestrated data retrieval, deployable on a conventional, standardized, low-cost Cloud infrastructure. The resulting data collections are analyzed at three abstraction levels: node level, attribute level, and derived network structures.

Our key findings include: Datasets from single locations are frequently incomplete due to unexpected API-level and connection-level failures. When comparing all datasets collected in parallel from 10 different machines for otherwise identical searches, we found different variations in terms of the data records (tweets) retrieved. While comparatively small in absolute numbers (# tweets), we show that even small data inconsistencies between samples may severely affect analyses at all three levels. At the attribute level, only 7.4% of the retrievable data records (tweets) carry identical metadata (e.g., retweet count, likes) when comparing the individual search results. Even attributes deemed fixed or unchangeable are found to vary between geographical locations.

We also derived retweet networks for each location and compared the resulting network structures (vertices, edges, connected components, and max degree). The derived networks vary substantially depending on the analysis task at hand. We compiled these findings into three recommendations to improve Twitter data mining in practice. In our future work, we will incorporate additional access levels to Twitter's API and we will extend the network-level analysis.

References

[1] E. Kušen and M. Strembeck, "Emotional Communication During Crisis Events: Mining Structural OSN Patterns," *IEEE Internet Computing*, vol. 25, no. 02, pp. 58–65, March 2021.

[2] N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, and D. Lazer, "Fake news on Twitter during the 2016 US presidential election," *Science*, vol. 363, no. 6425, pp. 374–378, 2019.

[3] F. Morone and H. A. Makse, "Influence maximization in complex networks through optimal percolation," *Nature*, vol. 524, no. 7563, pp. 65–68, 2015.

[4] D. A. Broniatowski, A. M. Jamison, S. Qi, L. AlKulaib, T. Chen, A. Benton, S. C. Quinn, and M. Dredze, "Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate," *American Journal of Public Health*, vol. 108, no. 10, pp. 1378–1384, 2018.

[5] A. Boutyline and R. Willer, "The social structure of political echo chambers: Variation in ideological homophily in online networks," *Political Psychology*, vol. 38, no. 3, pp. 551–569, 2017.

[6] E. Kušen and M. Strembeck, "Building blocks of communication networks in times of crises: Emotion-exchange motifs," *Computers in Human Behavior*, vol. 123, 2021.

[7] J. Pfeffer, A. Mooseder, L. Hammer, O. Stritzel, and D. Garcia, "This sample seems to be good enough! assessing coverage and temporal reliability of Twitter's academic API," *CoRR*, vol. abs/2204.02290, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2204.02290

[8] F. Morstatter, J. Pfeffer, H. Liu, and K. Carley, "Is the sample good enough? comparing data from Twitter's streaming API with Twitter's Firehose," in *Proc. International AAAI Conf. on Web and Social Media*, vol. 7, no. 1, 2013, pp. 400–408.

[9] J. Pfeffer, K. Mayer, and F. Morstatter, "Tampering with Twitter's sample API," *EPJ Data Science*, vol. 7, no. 1, p. 50, 2018.

[10] Y. Wang, J. Callan, and B. Zheng, "Should we use the sample? analyzing datasets sampled from twitter's stream api," *ACM Transactions on the Web (TWEB)*, vol. 9, no. 3, pp. 1–23, 2015.

[11] Y. Kim, R. Nordgren, and S. Emery, "The story of Goldilocks and three Twitter's APIs: A pilot study on Twitter data sources and disclosure," *International Journal of Environmental Research and Public Health*, vol. 17, no. 3, p. 864, 2020.

[12] J. Mahmud, J. Chen, and J. Nichols, "When will you answer this? Estimating response time in Twitter," in *Proc. International AAAI Conf. on Web and Social Media*, vol. 7, 2013, pp. 697–700.

[13] N. O. Hodas and K. Lerman, "How visibility and divided attention constrain social contagion," in *Proc. International Conf. on Privacy, Security, Risk and Trust and 2012 International Conf. on Social Computing*. IEEE, 2012, pp. 249–257.

[14] F. Kooti, L. M. Aiello, M. Grbovic, K. Lerman, and A. Mantrach, "Evolution of conversations in the age of email overload," in *Proc. 24th International Conf. on World Wide Web*, 2015, pp. 603–613.

[15] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?" in *Proceedings of the 19th International Conf. on World wide web*, 2010, pp. 591–600.

[16] Y. Xiong, M. Cho, and B. Boatwright, "Hashtag activism and message frames among social movement organizations: Semantic network analysis and thematic analysis of Twitter during the #MeToo movement," *Public Relations Review*, vol. 45, no. 1, pp. 10–23, 2019.

[17] A. Gruzd and J. Roy, "Investigating political polarization on Twitter: A Canadian perspective," *Policy & Internet*, vol. 6, no. 1, pp. 28–45, 2014. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/1944-2866.POI354

[18] E. Kušen and M. Strembeck, "Evacuate everyone south of that line: Analyzing structural communication patterns during natural disasters," *Journal of Computational Social Science*, vol. 4, no. 2, November 2021.

[19] J. Ince, F. Rojas, and C. A. Davis, "The social media response to Black Lives Matter: How Twitter users interact with Black Lives Matter through hashtag use," *Ethnic and racial studies*, vol. 40, no. 11, pp. 1814–1830, 2017.

[20] A. Campan, T. Atnafu, T. M. Truta, and J. Nolan, "Is data collection through twitter streaming api useful for academic research?" in *Proc. IEEE International Conf. on Big Data*. IEEE, 2018, pp. 3638–3643.

[21] A. Zubiaga, "A longitudinal assessment of the persistence of Twitter datasets," *Journal of the Association for Information Science and Technology*, vol. 69, no. 8, pp. 974–984, 2018.

[22] J. C. Timoneda, "Where in the world is my tweet: Detecting irregular removal patterns on Twitter," *PloS one*, vol. 13, no. 9, 2018.