Third-party Link Shortener Usage on Twitter (extended)

Martin Lindblom, Oscar Järpehult, Mathilda Moström, Alexander Edberg, Niklas Carlsson
Linköping University, Sweden

***Abstract*—Twitter has proven a powerful tool to shape peoples' opinions and thoughts. One efficient way to spread information is with the use of links. In this paper, we characterize the link sharing usage on Twitter, placing particular focus on third-party link shortener services that hide the full URL from the user. First, we present a measurement framework that combines two Twitter APIs and the Bitly API, and allows us to collect detailed statistics about tweets, their posters, their link usage, and the retweets and clicks 24 hours after the original tweet. Second, using two one-week-long datasets, collected one year apart (April 2019 and 2020), we then characterize and analyze important difference in link usage among users, the domains that different users and shorteners (re)direct users too, and compare the click rates of such links with the corresponding retweet rates. The analysis provides insights into link sharing biases on Twitter, skews, and behavioral differences in usage, as well as reveal interesting observations capturing differences in how a tweet containing a link may be retweeted versus how the embedded link is clicked. Finally, we use click-based results for covid-19 tweets to discuss the importance of controlling the spread of (mis)information.**

I. Introduction

Increasingly many people stay connected, obtain, and share news via social media sites. Today, a single person can reach millions of worldwide users within a very short time, sharing everything from plain text messages to videos and other media.

Twitter is perhaps the most influential such service. It has 186 million daily active users (March 2020) [1] and is closely followed by the major traditional news channels (e.g., TV and newspapers), who regularly quote information shared on Twitter. This has resulted in many people and organizations, including celebrities, presidents, and other world powers, using Twitter as one of their primary dissemination platforms.

One of the reasons that Twitter has been so successful is that it limits the length of the "tweets" (i.e., messages) that its users are allowed to post. This character limit (originally 140 characters; today 280 characters) forces users to create terse messages that bring forward the key messages that they want to convey and simplifies the presentation of streams of tweets on small mobile devices. However, the original length limitation also resulted in an ecosystem of link shortener services; each of which provides users with compressed URLs that they can use in their tweets (or other terse messages), which when clicked on redirect the users to the original (typically longer) URL. These services therefore help users effectively share links, while consuming less characters. Although Twitter since then has created their own link shortener and modified their policy such that all links consume the same number of characters (regardless of URL lengths), there are still many third-party link shorteners being used on Twitter.

978-3-903176-40-9 ©2021 IFIP

In this paper, we characterize the link sharing usage on Twitter, placing particular focus on the usage of third-party link shorteners. We first present a careful data collection methodology that involves collecting data in parallel via multiple different APIs and information sources. Using two one-week-long datasets collected using the developed collection tool, each separated one year apart (April 2019 and 2020), we then characterize and analyze the link usage on Twitter, the users of link sharing services, the domains that the users direct their followers to, and compare the click rates of such links with the retweet rates of the corresponding tweets. We also provide insights into various differences that we observe across tweet categories based on whether they include a link, a link shortener, or a Bitly link, for example, as well across user groups and the domains being linked.

In contrast to prior works (§ V), which primarily focus on security perspectives arising due to linked URLs being concealed, we take a popularity-based perspective on the general link shortener usage. This includes comparing the popularity of the linked domains (e.g., based on Alexa/Majestic ranks and frequency of use) and the popularity of the users posting the links. Among the third-party shorteners, Bitly is by far the most popular service. We therefore take a closer look at its usage and incorporate its API into our collection methodology so as to also extract the clicks of each such link.

By comparing datasets, we identify several seemingly invariant properties and reflect on current trends. Similar to several other popularity-based contexts, we observe a significant skew also in link usage. This includes with regards to who posts most links, which shorteners are most often selected, and which domains are most frequently linked. Furthermore, the most tweeted, retweeted, and clicked domains often are not the most popular domains on the internet. Instead, the most linked/clicked domains (from all ranks) often provide a service that in some way is well-suited to be shared via Twitter (e.g., YouTube videos, Spotify playlists, daily horoscopes, or honesty surveys). While most observations are invariant over the datasets, we also observed some changes/trends (e.g., reduced Bitly usage and increased use of website-specific shorteners) that opens for interesting future work.

Finally, we present a click-based analysis that confirms prior work's [2] observation that users often share links before clicking the links. While we do not consider misinformation in

this paper, we find it concerning that covid-19 related tweets appear to have a relatively higher fraction of tweets with more retweets than clicks compared to other Bitly tweets. The generality of these and other findings have been validated using two additional week-long 2020 datasets.

The properties identified throughout the paper have implications on information sharing and highlight differences in how different content types are both shared and clicked. Based on our click-based analysis of covid-19 related posts and others' findings regarding the virality of fake news [3], we also make a case for the importance of policies that try to prevent news from spreading faster than users click/read the shared links.

Outline: After a brief background (§ II), we present our methodology (§ III) and characterization results (§ IV). Finally, related work (§ V) and conclusions (§ VI) are presented.

II. Background

Twitter provides users with multiple ways to interact. For example, a user can post their own tweets ("general tweets", which can contain text, photos, GIFs, and videos) that others can read, create mentions of other users (using the @ symbol), reply to tweets, and retweet other users tweets (i.e., re-post tweets), select to follow another user, resulting in that user getting one more follower (that is following it) [4].

Twitter APIs: Twitter provides application programming interfaces (APIs) that allow anyone to collect public tweets and user data via HTTP requests [5]. At the time of the collection, there were two ways to collect tweets for free. First, a search-based API allows users to search tweets among the tweets posted in the past 7 days. This interface has a rate limit of 450 request each 15 minutes, where each request can contain 0-100 tweets. Second, a real-time streaming API continually returns approximately 1% of all tweets currently being posted on Twitter, with the possibility to add custom filters to the stream. In this paper, we use both APIs.

URL shortening: There are many link shortener services that creates short URLs (e.g., https://sho.rt/3wK5) that redirect users using such a link to a corresponding long URL (e.g., https://a-very-long-url.com/some-file.html). URL shortener services are typically of one out of two types: public or internal. Public shorteners (e.g., Bitly, goo.gl and TinyURL) allow anyone to shorten any URL. In contrast, some services use their own service-specific implementations. For example, youtu.be only links YouTube videos and t.co is used by Twitter to automatically shorten URLs in tweets.

Bitly API: Bitly is the most popular public third-party link shortener among those that we observed on Twitter. Bitly offers users to easily view link statistics for all their URLs either using their API (which we used) or by adding a plus (+) sign to the URL and accessing a statistics page.

III. Data Collection Method

To understand how tweets with clicks were retweeted and clicked, we split the data collection into two phases. In the first phase, we use Twitter's streaming API to collect as many tweets as possible together with information about each such tweet and the poster of the tweet. In the second phase, which took place 24 hours later, we then collected information about retweets of these particular tweets, the URLs that link shorteners in the tweets redirected to, and in the case that it was a Bitly link, we also collected Bitly specific information (e.g., click statistics) about the embedded Bitly link.

Block-based scheduling: We use blocks-based scheduling to simplify data management during the second phase. During the first phase we group tweets into four-hour blocks and at the end of the collection of such a block, we schedule the data collection of the second phase for the tweets in that block to take place 20 hours later. This ensures that the first retweet and click data always are collected exactly 24 hours after the corresponding tweets were collected (during the first phase).

With relatively small time variation in the retweet collection, most retweet statistics were collected roughly 24 hours after the original tweet. At that time, due to the ephemeral nature of tweets, most retweets typically have occurred [2]. However, due to rate limitations (discussed next) the Bitly-related collection that we run in parallel is often slower than the retweet collection. To keep track of, and account for, such timing differences and any potential biases they cause, for each individual tweet, we therefore record the time instance when each value is recorded and use these timings in our analysis. Perhaps most importantly, the collection is designed such that the retweets always are collected ahead of (or at least no later than) the corresponding Bitly link statistics, ensuring that the measured retweet-to-click ratio always is conservative.

Rate limitations: When implementing our data collection, we had to adhere to the APIs' rate limitations. First, Twitter's streaming API (used during the first phase) restricts us to a 1% sample of the total tweet volume. While others have observed some biases in the streams provided by this API [6], the observed biases should not significantly impact the high-level conclusions reported based on our (multiple) one-week-long datasets. A preliminary sanity check, comparing the tweet volume observed for individual users during the week with the number of tweets each of these accounts generated between the first and last observed tweet during the week, suggests that the tweets by these users were sampled at a rate close to 1%.

During the second phase, we had to adhere to rate limits of both Twitter and Bitly. For Twitter, the rate limit differed between endpoints. However, for the one we used, the limit was 900 requests per 15-min window [7], which was sufficient (when combined with batch queries allowing queries about 100 tweets per request) to consistently complete the collection of all tweets of a block within a four-hour window. For Bitly, the documented rate-limit specifications for different endpoints and call types are less clear. However, with the endpoints that we used, we were able to achieve the only rate limits that we found listed: 1,000 calls per hour and 100 calls per minute [8]. (This is substantially higher than the 200 requests per hour reported by Gabielkov et al. [9].) With access to four Bitly accounts, this allowed us to make 4,000 requests per hour.

Other API limitations: The Bitly API does not distinguish if a click is due to a particular instance of a link. Instead, the

TABLE I: Categorical breakdown of observed tweets.

Category	2019		2020	
All Tweets	25,482,108	*(100%)*	33,281,088	*(100%)*
Link Tweets	4,026,101	*(15.8%)*	3,803,233	*(11.4%)*
Shortener Tweets	322,954	*(1.27%)*	310,915	*(0.93%)*
Bitly Tweets	159,143	*(0.625%)*	52,517	*(0.158%)*

finest granularity it provides is whether a click is due to a link shared on Twitter or some other website. To avoid inflated values due to older links, we limit our analysis to Bitly links for which we observe the tweet of interest within 10 minutes of the creation of the Bitly link and only count clicks via Twitter. Finally, we note that Bitly counts clicks by the same users as distinct clicks as long as they are at least a few seconds apart. In contrast, a twitter user can at most retweet a tweet once. This further contributes to the retweet-to-click ratios observed being conservative (if interpreted on a per user basis).

During the collection of retweet stats, we resubmitted failed requests one more time before moving to the next set of tweets, whereas for Bitly we did not retry failed requests. This design decision was motivated by the Bitly process already being slower than the Twitter process and the Bitly process sometimes taking considerable time (exceeding four hours). In particular, we argued that it was better to lose some data than delaying the collection of all other data. For the comparisons of retweet and click data we only perform our analysis on the tweets for which all data was successfully obtained.

Long URLs of other link shorteners: Finally, we identified all link shorteners and looked up their full URLs. Here, we followed all URLs associated with one of 284 manually identified shorteners (aggregate of public lists of known shortener services). While we did not have any problem looking up the full URLs of Bitly links (collected using their API), we had some problems with other shorteners. For example, goo.gl flagged us as a bot after just a few lookups and some others frequently redirected to an invalid page. This almost exclusively was observed by: (1) a 40X/50X status code, (2) a page timeout, or (3) the service stating that the link does not exist or has been removed. Out of the 284 known shorteners, we observed links to 37, and successfully looked up at least one URL for 27. The other shortener services were excluded from our domain popularity analysis. In total, we successfully looked up 87.9% of the links.

Implementation and datasets: Implementation details and dataset descriptions are provided in the Appendix.

IV. Results

To improve the understanding of the general third-party link shortener usage on Twitter, we present a popularity-based analysis. First, we analyze the relative popularity of the link shorteners themselves (§ IV-A) and the domains being linked (§ IV-B). Second, we look closer at the users using these services, whose usage indirectly drive the observed popularity distributions. Here, we also study the language used, the users popularity (e.g., how many followers they have), their behavior, as well as compare the success of the posts using both retweet- and click-based metrics (§ IV-C).

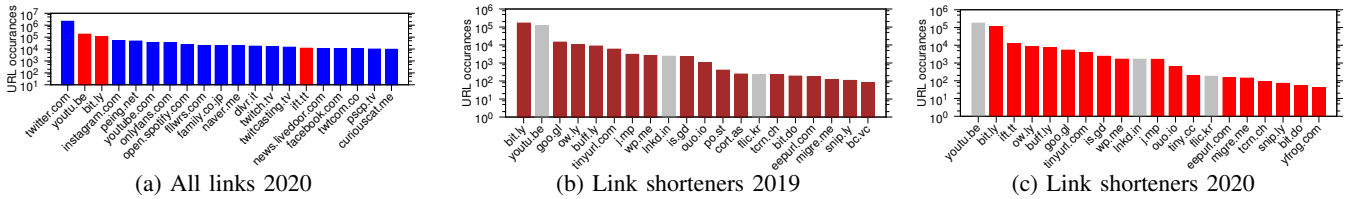
A. High-level link shortener usage

We focus on two one-week-long datasets collected one year apart: (1) April 26 - May 3, 2019, and (2) April 18-25, 2020. (Similar results have also been observed for two other week-long datasets from Mar-Apr. 2020.) Table I summarizes the fraction of the total tweets (25.5 and 33.3 million) over the two datasets (labelled "2019" and "2020" in the following) that contained links (15.8% and 11.4%), shortener links (1.27% and 0.93%), and Bitly links (0.625% and 0.158%). Naturally, the later subsets in this list are subset of the earlier ones. This is seen in Figure 1(a), where we plot the number of link occurrences, before redirects, for the top-20 domains in the 2020 dataset. Here, three out of the top-20 most frequently linked domains are link-shorteners: youtu.be (rank 2) bit.ly (rank 3), and ift.tt (rank 15). In the top-25, we also find ow.ly (rank 23) and buff.ly (rank 25). Of these, youtu.be and bit.ly are responsible for the majority of the shortener links (54.5% and 34.4%, respectively). After these two giants, there is a big drop to the third most popular link shortener (ift.tt; 3.6%). This big drop can be seen in Figures 1(b) and 1(c), which show the usage of shorteners in the two datasets. In addition to youtu.be, which is used for YouTube videos, several other popular websites also use domain specific shorteners, including LinkedIn (lnkd.in, rank 10), Flickr (flic.kr, rank 14), and Spotify (sptfy.com, rank 22). Overall, we observed a reduced Bitly usage and increased use of website-specific shorteners.

B. Domain-based analysis

Top linked domains: To understand what domains are being linked to using different methods, we extracted the long URLs that each link shortener directed clients to and analyzed the relative usage frequencies of the observed domains. Tables II-IV show the top-6 domains and their link occurrences in the 2020 dataset, as broken down for all links, shortener links, and Bitly links, respectively. We also include the 2019 rank (within brackets in the first column), the fraction of the total links each domain makes up, the percentage change in this fraction, as well as their Alexa and Majestics rankings when available (obtained for last day of each collection period); otherwise we list "-". Some shortened URLs were invalid or had been deleted - most common for goo.gl (98%), ow.ly (95%) and buff.ly (99.9%). These links are not included in the domain analysis. While this introduces some bias, their usage is much less than that of Bitly.

Twitter itself has by far the most link occurrences (56% of all links; Table II). This is due to a Twitter link being embedded in every mention or retweet. Among the domains being linked using shorteners (Tables III), youtube.com and twittascope.com stands out. Most YouTube links are from YouTube's dedicated shortener youtu.be. These URLs are generated whenever a YouTube user shares a video. twittascope.com is (by far) the most linked domain using Bitly (Table IV). Despite only observing 12 unique Bitly links for this domain (one for each zodiac sign) and the user posting most such links only having three such tweets (in our dataset), these 12 links combine for 23,173 tweets in the 2019 dataset and 9,374



(a) All links 2020 (b) Link shorteners 2019 (c) Link shorteners 2020

Fig. 1: Top-20 lists of domains directly linked in a tweet and the corresponding number of occurrences for that domain: (a) all domains in 2020, (b) link-shorteners in 2019, (c) link shorteneres in 2020.

TABLE II: Top-6 most frequently linked domains: any link.

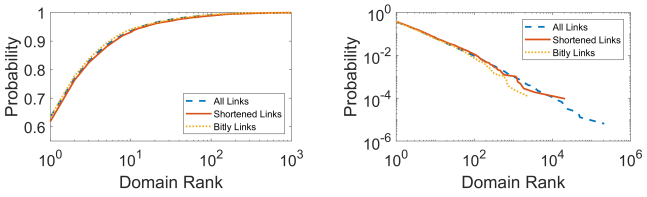
		Domain	Occurrences			Alexa	Maj.
1	(1)	twitter.com	2,134,010	(56%)	+4.2%	55	4
2	(3)	youtube.com	213,002	(5.6%)	+53%	2	3
3	(5)	instagram.com	50,572	(1.3%)	-6.3%	32	5
4	(8)	peing.net	46,834	(1.2%)	+98%	17,513	185,762
5	(-)	onlyfans.com	34,244	(0.90%)	-	1,060	25,602
6	(12)	open.spotify.com	23,339	(0.61%)	+75%	-	155

TABLE III: Top-6 most frequently linked domains: shorteners.

		Domain	Occurrences			Alexa	Maj.
1	(1)	youtube.com	177,203	(57%)	+56%	2	3
2	(2)	twittascope.com	21,160	(6.8%)	-5.2%	300,510	-
3	(4)	k.kakaocdn.net	3,565	(1.1%)	-32%	-	-
4	(6)	linkedin.com	1,661	(0.53%)	-26%	75	6
5	(5)	img1.daumcdn.net	1,535	(0.49%)	-69%	-	-
6	(-)	dolk.jp	1,322	(0.43%)	-	-	-

TABLE IV: Top-6 most frequently linked domains: Bitly.

		Domain	Occurrences			Alexa	Maj.
1	(1)	twittascope.com	9,374	(18%)	-60%	300,510	-
2	(3)	k.kakaocdn.net	2,449	(4.7%)	-55%	-	-
3	(4)	img1.daumcdn.net	1,085	(2.0%)	-79%	-	-
4	(6)	t1.daumcdn.net	777	(1.5%)	-58%	-	-
5	(-)	rbeiv.com	697	(1.3%)	-	-	-
6	(19)	drive.google.com	684	(1.3%)	-13%	-	36



(a) CDF (b) CCDF

Fig. 2: Distribution of the cumulative fraction of links that different ranked domains are responsible for. (2020 dataset).

tweets in the 2020 dataset. The twittascope.com posts typically contain a generic horoscope-of-the-day statement with a Bitly link to their sign's daily horoscope. This represents a 60% drop in the relative Bitly usage (when normalizing for the number of Bitly links in each dataset). The Twitter user @Twittascope itself typically uses a mix of bit.ly and ow.ly links. Similar to twittascope.com posts, most users posting links to k.kakaocdn.net and img1.daumcdn.net (owned by Kakao Corp.) are only observed once (spreading the usage over many users). However, in these cases, the links typically point to specific images. For k.kakaocdn.net, 85.5% of the links point to .gif images and 13.3% to .jpg images.

Power-law-shaped popularity distributions: The high skew observed for the top-domains shown in the above tables is even clearer when considering the full distributions of the fraction of links that different ranked domains are responsible for. Figures 2(a) and 2(b) show the corresponding Cumulative Distribution Function (CDF) and Complementary CDF

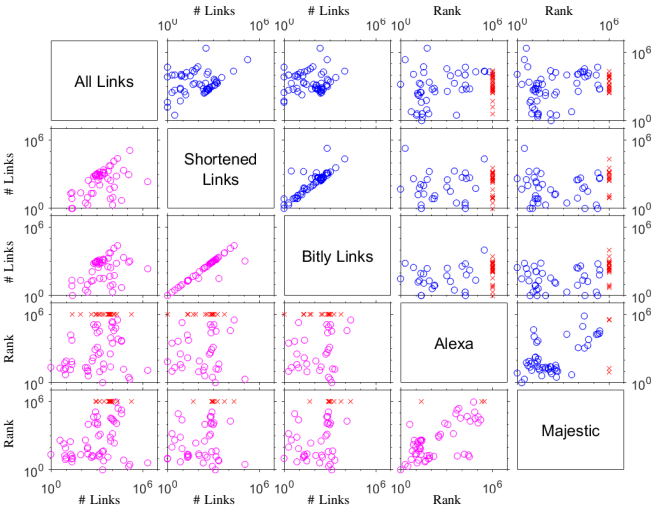


Fig. 3: Pairwise scatter plots showing the frequencies and ranks of the top-25 domain sets based on all links (# such links/domain), shortened links (# such links/domain), Bitly links (# such links/domain), Alexa rank (for the domain), and Majestic rank (for the domain). Frequency/rank comparisons from 2019 dataset (pink) are shown below vertical divider and 2020 comparisons (blue) are shown above. Red crosses '×' at rank 10^6 are used to illustrate domains with unknown rank.

(CCDF), respectively. Note that the distributions of the three classes of links have almost identical shapes and that the top-10 domains are responsible for more than 90% of the links (in each category). Furthermore, the straight-line shape of the CCDF suggests that the distributions are power-law like. While we only show results for the 2020 dataset, these findings have been found invariant between the two datasets.

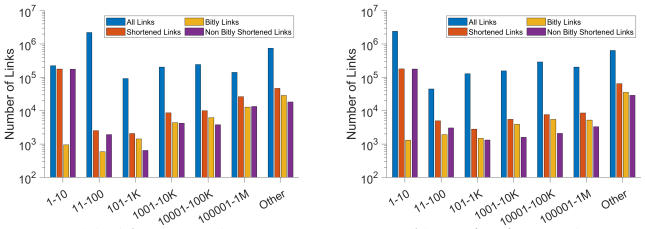
Link usage correlations: A yet closer look at Tables II-IV suggests (i) that there are considerable overlaps between the domains that are popular in each category, and (ii) that some of the popular domains in the three link categories also have high global popularity rank with Alexa and Majestic. It may therefore be tempting to assume that there are significant correlations between the domain usage across the link categories, as well as between the link usage and the global popularity. To glean insight into the validity of these two hypotheses, Figure 3 shows pairwise scatter plots of five usage and popularity measures: (i) domain occurrences among all links, (ii) domain occurrences among all shortened links, (iii) domain occurrences among all Bitly links, (iv) global Alexa ranking, and (v) global Majestic ranking. Here, we only plot

these metrics for the domains that are ranked among the top-25 domains in at least one of these five categories. If rank is not known, we set it to 1 million.

Before analyzing the above hypotheses, note that also here, only limited changes can be observed between 2019 and 2020. This suggests that the relative usage and popularity of the set of linked domains have remained relatively stable, even if the usage/popularity of some individual domains may have changed substantially [10]. This can be seen by the symmetry when comparing the occurrences/ranks for 2019 (in pink, below the diagonal) and 2020 (in blue, above the diagonal).

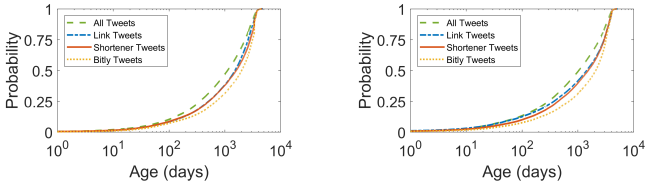
Let us now consider correlations among the domain occurrences using the three link categories. The top-left 3×3 (sub)matrix of the 5×5 matrix shown in Figure 3 shows the pairwise scatter plots of the domain occurrences using these three link types. With Bitly being responsible for 34.4% of the shortener links in the 2020 dataset (49.3% in the 2019 dataset) and 82.2% (2020) of all shortener links after removing the youtu.be links, it is perhaps not surprising that we see high correlation between the links observed for the "Shortened" and "Bitly" categories. (Most markers fall along the diagonal for these pairings.) Furthermore, since the set of all Bitly links is a subset of the shortener links and the set of all shortener links itself is a subset of all links, there are no points below (2019) or above (2020) the diagonal seen in the 3×3 plots. Instead, any point away from the diagonal illustrates a domain with additional links in the larger of the two sets. Based on this observation, it is interesting to note that the two biggest outliers are youtube.com (with 1,288 bit.ly links in 2019 and 704 in 2020) and flickr.com (144 and 82), in addition to all their youtu.be and flic.kr links. These figures also clearly show that there are many domains with significant link-shortener usage that also frequently are linked directly. These domains are represented by points offset from the diagonal in the "Shortened" vs "All" plots. The biggest outlier here is twitter.com which is linked a few times using third-party shorteners (mostly Bitly).

Domain popularity correlations: Next, let us compare the most frequently linked domains of each link category (first three rows/columns) with the Alexa ranks (fourth row/column) and Majestic ranks (fifth row/column). Comparing the correlations between the ranks themselves (bottom-right 2×2 matrix), which each are obtained in vastly different ways [10], we see very limited correlations. This suggests that the frequency that links to the domains are shared on Twitter (using the three link categories) may not correlate strongly with the domains' global (popularity) rank. Having said that, we next highlight two interesting observations when comparing ranks and frequencies. First, the top-three ranked domains in both Alexa (Google, YouTube, Facebook) and Majestic (Google, Facebook, YouTube) are present in all categories. Second, the highest-ranking domain (google.com or its domains) is never the most frequently occurring domain in any of the link sets at the same time as the domain that has the lowest rank among all observed domains that were ranked in the Bitly-Alexa plot (twittascope.com) is the domain that has the most



(a) Alexa top-1M. (b) Majestic top-1M.

Fig. 4: Link popularity distribution to domains of different popularity classes. (2020 dataset)



(a) 2019 (b) 2020

Fig. 5: Age distribution of user accounts at the time of tweets.

occurrences in this set. twittascope.com had 9,374 (17.8%) of the bit.ly links and an Alexa rank of 300,501. In contrast, various google.com domains were only linked 942 times using bit.ly. (615 of these were to drive.google.com.)

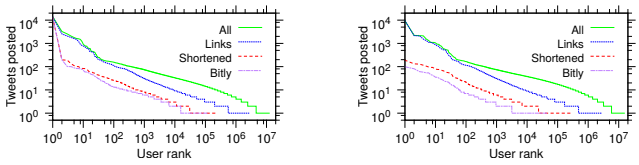
Global popularity of linked domains: Thus far we have considered each domain individually. As the links usage of individual domains can be noisy, Figure 4 shows aggregate link usage in 2020 to domains with rank ranges: 1-10, 11-100, 101-1K, 1001-10K, 100001-100K, 100001-1M, and non-ranked domains ("other"). The 2019 results are very similar. In addition to the three previously considered link categories, we also include results for the set of all non-Bitly shortener links. The high top-10 usage of this category (first purple bar) is due to a large number of YouTube links. The high Majestic top-10 usage for general links are dominated by Twitter links. Otherwise, we see a steadily increasing link usage for the "popularity buckets" containing lower-ranked domains. This shows that the long tail of less popular domains (that by definition includes many more domains than the high-ranked buckets) is responsible for a substantial fraction of the observed links (for each category).

C. User, usage, and click analysis

The observed popularity distributions clearly are a product of the choices and successes of the aggregate set of Twitter users. We next analyze the users of link shorteners and whether their user profiles differ from general posters on Twitter.

Bitly (and link) users have older accounts: The age of the user accounts associated with links typically are somewhat older. This bias is largest among posts that include Bitly links. Figure 5 shows CDFs of the account age of the tweeters of the tweets in each category. Note the clear separation of the lines, despite the x-axis being on log scale. These differences are invariant between the two datasets.

Skewed link usage with long tails: Figure 6 shows a rank-frequency plot of the number of observed tweets per user,



(a) 2019 (b) 2020

Fig. 6: Number of observed tweets per user, when users are ranked based on the same metric.

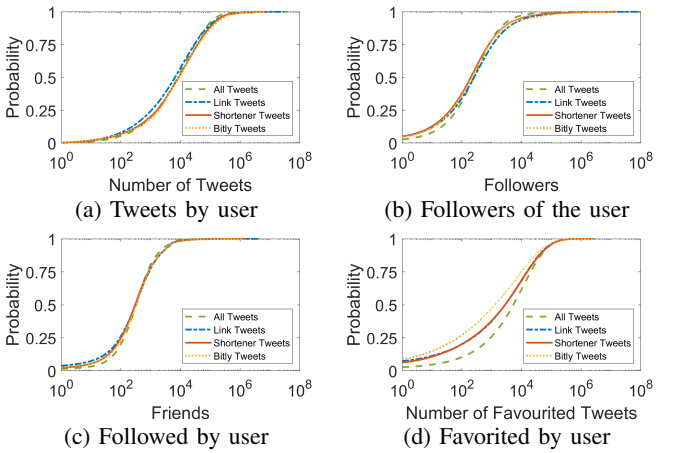
broken down for each link type. First, note that the "All" curve (all tweets) and "Links" curve (tweets with a link) almost overlap for the top-ranked positions. This shows that the most frequent tweeters use links in most of their tweets. Second, for all four categories, there is a long tail of users that have few tweets. All three of the link categories appear to have a Zipf-like tail (straight-line behavior on log-log scale). In contrast, the "all" curve have a "bump". As of today, we do not have a good explanation what process may cause the bump. Third, inline with general Bitly usage trends, the most frequent tweeter in our 2019 dataset, which exclusively used Bitly links in 2019, have since reduced its Bitly usage. This user (@akiko_lawson) used Bitly for all 13,944 tweets observed by this user in the 2019 dataset (rank 1 for all curves in 2019), but only used Bitly in 99 out of 194 observed tweets in the 2020 dataset. This is a substantial drop in usage, but still allowed it to be the top-Bitly user in 2020 too. The top tweeter in 2020 (@famima_reply) used links in all 9,393 observed tweets.

The most frequent link tweeters are Japanese: Interestingly, the most frequent posters (as noted above) share links in all their tweets, and where almost exclusively associated with Japanese websites and posted in Japanese. For example, four out of the top-5 (2019) and five out of the top-5 (2020) posters in the categories "all" and "links" were associated with Japanese websites. Also for the other tweet categories ("shorteners" and "Bitly") two out of the top-5 where Japanese both in 2019 and 2020.

Languages used: Only 1.12% (2019) and 0.96% (2020) of the tweets were geo-tagged. This number was even smaller among the tweets that contained link shorteners (0.60% and 0.56%) and Bitly links (0.62% and 0.22%). To understand differences in who the tweeters in the three categories were, we instead analyzed the language used in the tweets and the language listed by the users posting the tweets. Most noticeable, the top-3 languages in 14 out of 16 cases (2 metrics $\times$ 2 datasets $\times$ 4 types of tweets) were: (1) English, (2) Japanese, and (3) Spanish (in that order). Table V summarizes the usage of these three languages used in tweets, with the only exception marked with an asterisk (*). In the exception case, Spanish had rank 5 and Korean had rank 3. The language usage differences (rank and magnitude) when comparing across the tweet types were primarily associated with rank 4 and below (Korean stood out as always having rank 3 or 4 for the "shortener" and "Bitly" categories, but never obtaining that high rank for "All" and "Links") and the relative usage of the top-3 languages (English dominating the Bitly usage).

TABLE V: Top-3 lists of languages used in different tweets.

	All		Links		Shortener		Bitly	
Language	2019	2020	2019	2020	2019	2020	2019	2020
1. English	31.0%	29.4%	38.3%	38.0%	38.4%	38.1%	42.3%	50.6%
2. Japanese	19.7%	17.2%	15.1%	18.0%	21.3%	22.9%	21.9%	14.3%
3. Spanish	8.5%	8.0%	8.9%	9.1%	9.6%	6.9%*	10.3%	8.8%



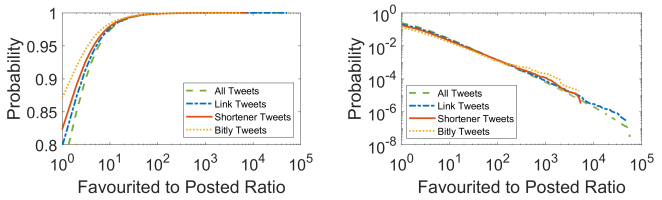
(a) Tweets by user (b) Followers of the user

(c) Followed by user (d) Favorited by user

Fig. 7: Tweets posted, followers, users followed, and number of favorited tweets by the user. (2020 dataset)

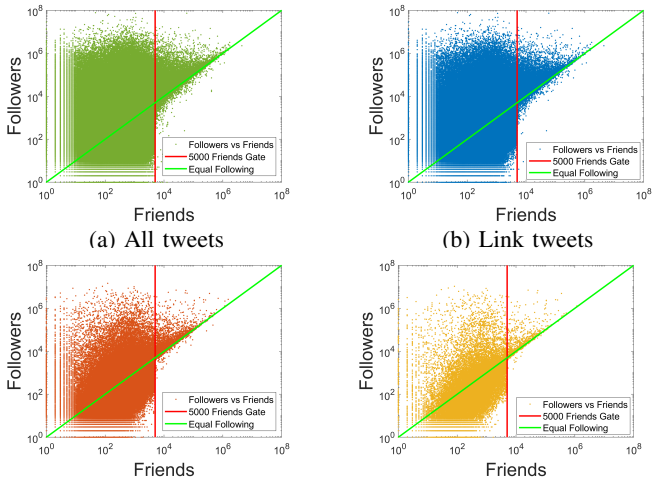
Similar tweet activity and number of followers: Except for the extreme users mentioned above, the general distributions for the number of tweets and followers associated with the posting accounts of tweets in the different categories look very similar across the different categories. Figures 7(a), 7(b), and 7(c) show CDFs for the number of tweets, followers, and the number of users followed by the users that posted tweets, respectively. All three distributions are similar across the different tweet subsets, although tweeters of link shorteners typically post slightly more tweets than tweeters of non-shortened links (e.g., comparing medians) and have slightly fewer followers (e.g., median comparison). Observations are consistent between 2019 and 2020.

Less interactive publishers: Twitter users can interact in several ways. One way is to favorite others' posts. This can be seen as "liking" a post, and encourages the poster. We have found that the top tweeters, regardless of category, tweet substantially more than they favorite. We next compare such interaction for posters associated with each link category. Figure 7(d) shows how much the tweeters of each tweet category had favorited at the time of their tweet and Figure 8 shows the favorite-to-tweet ratio at that time. First, note that Bitly tweets more frequently are posted by users that have favorited less and that these users have lower favorite-to-tweet ratio. This suggests that users of Bitly are more likely to focus more on publishing than interacting and encouraging others. Second, the tail of the favorite-to-tweet ratio has a power-law-like shape (as suggested by the relatively straight CCDF curves in Figure 8(b)). Third, all tails have roughly the same slope (slightly steeper than -1.2), with the curves only shifted somewhat up or down (capturing the above-mentioned biases between the sample sets). The last two observations are interesting since parallel lines here suggests that users with large favorite-to-tweet ratios are equally likely to use Bitly



(a) CDF (b) CCDF

Fig. 8: Distribution of the favorite-to-tweet ratio at the time posting their tweets. (2020 dataset)



(a) All tweets (b) Link tweets

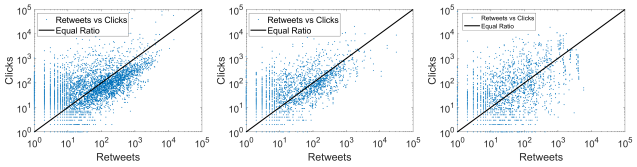
(c) Shortener tweets (d) Bitly tweets

Fig. 9: Followers-to-following ratios for users at the time of posting their tweet, separated by tweet category. (2020 dataset)

regardless of where in the tail they are (after a ratio 1-2 or higher) and that the bias instead is due to publish-oriented users (with small ratios) being relatively more likely to use Bitly. The above observations are consistent for both datasets.

Followers vs. users followed: Despite Bitly users typically having lower favorite-to-tweet ratios, we have not found any significant differences when looking at the follower-to-following ratios (Figure 9). This may partially be due to Twitter's threshold policy [11] for the friend-to-follower relationship. As illustrated in the figure, this policy requires that users must have more followers than what they are following themselves after they follow more than 5,000 users (red line). Here, the equality-ratio is shown in green and we use different colors to show the increasingly specific subsets of links. The occurrences of users with more than 5,000 friends but more friends than followers can be explained by users losing followers while keeping their friends. Again, except for the number of points, all four scatter plots are similar.

Verified users more likely to use links, shorteners, and Bitly links: Verified user accounts belong to users of public interest [12]. Based on our analysis, compared to non-verified users, verified users appear more likely to use links, link shorteners, and Bitly links in their posts. For example, in 2020, 18.6% of all the 15,247,424 observed users used links, 1.76% used link shorteners, and 0.03% used Bitly links. In contrast, the corresponding numbers for the 61,614 observed verified users were 52.9%, 3.85%, and 1.26%, respectively.



(a) 2019 (b) 2020 (c) Covid-19

Fig. 10: Scatter plots of Bitly click-to-retweets-ratios.

For 2019, the corresponding numbers were 23.7% vs 53.9% (all links), 2.00% vs 5.36% (shorteners), and 0.92% vs 3.48% (Bitly links). In all cases, the verified users were significantly more likely to also have tweets associated with the three different link types. While our tweet-based data collection (rather than user-based sampling) only allows us to form a conjecture, similar observations are seen when considering the relative fraction of tweets in each category that were posted by a verified user. For example, while the verified users only make up 0.40% (0.44%) of the unique users posting tweets in the 2020 (2019) dataset, they make up a relatively larger fraction of the users posting tweets belonging to the three link categories: 1.15% (0.99%) of the unique users posting a link, 0.89% (1.16%) of the unique users using a shortener, and 1.67% (1.65%) of the unique users using a Bitly link.

Retweeting without reading (Bitly) links: Figures 10(a) and 10(b) show scatter plots of the retweets and clicks for all Bitly links that were no older than 10 minutes when the tweets were posted. These sets represent the newly created Bitly links in the 2019 and 2020 datasets. Although we consider a more general set of tweets than Holmström et al. [2] (they only considered links to certain news pages) and despite our methodology providing a conservative retweet-to-click ratio, similar to them, we observe a large fraction of tweets for which there are more retweets than clicks (i.e., points below the diagonal). To protect users, some browsers provide users additional information about shortener links (before clicking the link) and Twitter tries to remove links that may present danger to end users [13]. However, a combination of automatic analysis (comparison with Phishtank dataset [14], containing reported phishing links, did not result in any matches) and manual analysis of individual links (revealed some suspicious domains) suggests that Twitter's filtering is not perfect. This further highlights the risk of people retweeting shortener links that they themselves do not click/check.

Unfortunately, the fraction of tweets that fall below the diagonal line (indicating an equal number of clicks and retweets) does not improve when looking closer at tweets associated with covid-19; a topic that has seen a significant amount of misinformation spread on Twitter and other social media. For example, Figure 10(c) shows the corresponding scatter plot for all Bitly tweets that contain covid-19 or corona [virus] in the long URL or in a hashtag of the tweet. (We also considered two other one-week-long datasets from 2020: Mar. 18-25 and Apr. 1-8. These resulted in very similar results.) While the fraction at first glance looks less dense below the diagonal, 51.6% (1,358/2,632) have a retweet-to-click ratio greater than 1 (i.e., points below the diagonal). In contrast, for the complete

2020 set of links, the 35.9% (18,841/52,517) had a ratio greater than 1, and for the 2019 set 42.8% (68,094/159,143) had a ratio greater than 1. The higher retweet-to-click ratios for the covid-19 topic are concerning as spread of misinformation about this topic can have severe consequences. While we do not study misinformation explicitly here, the higher retweet-to-click ratio of tweets related to covid-19 may suggest that the topic itself has more "viral" properties (pun intended) than the other tweets. For example, we note that "fake news" has been found to be spread faster (and wider) than other tweets about regular news stories [3]. Given the potential consequences of misinformation about covid-19, we believe that these observations provide further support for the importance of policies that try to prevent news from spreading faster than users click/read shared links. For example, Twitter could simply require users to *read before sharing*.

V. Related Work

Most prior work on link shorteners have focused on security-related aspects. For example, Maggi et al. [15] studied the link shortener links clicked by 7,000 users (over a two-year period) and the security threats that the links exposed the users to. Others have focused in specifically on the use of link shorteners for spam [16]–[21], phishing [22], [23] or other malicious usages [24], [25]. These aspects have resulted in Twitter maintaining and using their own URL blacklists, which effectiveness Bell et al. [13] recently evaluated. Others have shown that even benign origin URLs can expose users of ad-based shortening services (that gives link creators money for any clicks they generate and present an ad to users before directing them to the origin URL) can expose users to further risks and malicious content [26]. In contrast to these works, we present a popularity-based analysis of the general usage.

In this regard, the seminal 2011 paper by Antoniades et al. [27] is closely related, as it studied some popularity aspects (e.g., access frequencies, click distributions, and the most popular websites accessed using shorteners) among other things (e.g., byte overheads). While their work provides some insights into the shortener usage ten years ago, a lot has happened since then. Here, we also provide a deeper popularity-based analysis of what shorteners are used, the domains they point to, and the Twitter users using these services.

Click through rates: There is a limited number of studies that consider both the retweet and click through rates on Twitter. Most closely related are the works by Gabielkov et al. [9] and Holmström et al. [2]. Similar to us, both these works combine the use of the Twitter and Bitly API. However, their focus is different than ours. Gabielkov et al. [9] highlight differences in how many times Bitly links are retweeted compared to how many times they are clicked, and Holmström et al. [2] perform a temporal analysis of the retweets and clicks to news articles associated with a very limited number of news websites. In contrast, we consider all visible Bitly links. Others have used click-through-rates and similar metrics to measure the quality of ads [28], [29].

Asymmetric user behaviors and influences: While it is difficult to precisely quantify influence [30], [31], it has been shown that large number of followers does not necessarily translate into retweets [30]. Instead, retweets appear to be driven by the content of the tweets, and mentions appear to be driven by the popularity of the users. It has also been shown that users that have many followers (e.g., more than 250 followers) post status updates more often than those that follow many people (e.g., more than 250 people) [32].

Biases and Twitter's streaming API: In orthogonal research, others have studied the potential biases in the streams provided by Twitter's streaming API [6], [33] and the impact that channel selection can have on the results [34]. For example, Morstatter et al. [6] compared the data obtained using the free streaming API (used here) with random sampling from Twitter's payed firehose API service, with results suggesting that there may be some hidden biases in the samples obtained using the streaming API (especially when predicting the popularity of top-N lists, where N is small) and that geo-tagged tweets are over-represented. While we are aware of these biases and acknowledge that they likely impact the exact sets and frequencies reported in our top-N lists, we argue that these biases should not impact the general conclusions presented. Campan et al. [33] study the impact of filtering, with results suggesting that care should be taken when using filtering in combination with the streaming API. We did not use any filtering, and simply collected the full 1% stream.

VI. Conclusion

The link usage on Twitter gives an important window into users' information sharing habits. This paper presents a measurement framework and a novel characterization of the third-party link sharing usage on Twitter. The framework combines two Twitter APIs and the Bitly API, and allows us to collect detailed statistics about tweets, their posters, their link usage, and the retweets and clicks 24 hours after the tweets first are published. Using two one-week-long datasets (labelled "2019" and "2020") collected one year apart, we then identify and characterize important difference in link usage among such users, the domains that different users and link shorteners direct their users too, and compare the click rates of such links with the retweet rates of the corresponding tweets, conditioned on different user categories.

Similar to several other popularity-based contexts, we observe a significant skew also in link usage, including with regards to who posts the most links, which shorteners are most often selected, and which domains are most frequently linked. Interestingly, the most tweeted, retweeted, and clicked domains often are not the most popular domains on the internet (e.g., as ranked by Alexa and Majestic). Instead, they are often services (from all ranks) well-suited to be shared via Twitter (e.g., YouTube videos, Spotify playlists, daily horoscopes, or honesty surveys). While most of our observations are invariant over the datasets, we also observed some changes/trends (e.g., reduced Bitly usage, increased use of website-specific shorteners) that may warrant interesting future work.

In summary, the identified properties have implications on information sharing, highlights differences in how different types of content are shared and clicked, and help build support for policies that would limit the propagation of news that spread faster than users click/read the news stories.

ACKNOWLEDGEMENTS

This work was funded in part by the Swedish Research Council (VR). We also thank our shepherd Gareth Tyson and the anonymous reviewers for constructive suggestions.

REFERENCES

[1] M. Grothaus, "Twitter q2 2020 earnings: Revenue falls 19% yoy, but daily active users up 39% yoy to 186 million." [Online]. Available: https://www.fastcompany.com/90531571

[2] J. Holmstrom, D. Jonsson, F. Polbratt, O. Nilsson, L. Lundstrom, S. Ragnarsson, A. Forsberg, K. Andersson, and N. Carlsson, "Do we read what we share? analyzing the click dynamic of news articles shared on Twitter," in *Proc. IEEE/ACM ASONAM*, 2019.

[3] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[4] Twitter, "Getting started," https://help.twitter.com/en/twitter-guide, 2019, [Online; accessed 17-05-2019].

[5] ——, "About twitter's apis," https://help.twitter.com/en/rules-and-policies/twitter-api, 2019, [Online; accessed 17-05-2019].

[6] F. Morstatter, J. Pfeffer, H. Liu, and K. M. Carley, "Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose," *ICWSM*, 2013.

[7] Twitter, "Get statuses/lookup," https://developer.twitter.com/en/docs/tweets/post-and-engage/api-reference/get-statuses-lookup.html, 2019, [Online; accessed 17-05-2019].

[8] Bitly, "Rate limiting," https://dev.bitly.com/rate_limiting.html, 2018, [Online; accessed 17-05-2019].

[9] M. Gabielkov, A. Ramachandran, A. Chaintreau, and A. Legout, "Social clicks: What and who gets read on twitter?" *ACM SIGMETRICS*, 2016.

[10] Q. Scheitle, O. Hohlfeld, J. Gamba, J. Jelten, T. Zimmermann, S. D. Strowes, and N. Vallina-Rodriguez, "A long way to the top: Significance, structure, and stability of internet top lists," *Proc. ACM IMC*, 2018.

[11] Twitter, "About following on twitter," https://help.twitter.com/en/using-twitter/twitter-follow-limit, 2019, [Online; accessed 17-05-2019].

[12] ——, "About verified accounts," https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts, 2019, [Online; accessed 17-05-2019].

[13] S. Bell *et al.*, "Catch me (on time) if you can: Understanding the effectiveness of twitter url blacklists," *arXiv:1912.02520*, 2019.

[14] PhishTank, "Phishtank dataset," 2019. [Online]. Available: https://www.phishtank.com/

[15] F. Maggi, A. Frossi, S. Zanero, G. Stringhini, B. Stone-Gross, C. Kruegel, and G. Vigna, "Two years of short urls internet measurement: security threats and countermeasures," in *Proc. WWW*, 2013.

[16] V. Kandylas and A. Dasdan, "The utility of tweeted URLs for web search," *Proc. WWW*, 2010.

[17] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammers on twitter," *Proc. ACSAC*, 2010.

[18] N. Gupta, A. Aggarwal, and P. Kumaraguru, "bit.ly/malicious: Deep dive into short url based e-crime detection," in *Proc. eCrime*, 2014.

[19] C. Cao and J. Caverlee, "Detecting spam urls in social media via behavioral analysis," in *Proc. ECIR*, 2015.

[20] D. Wang, S. B. Navathe, L. Liu, D. Irani, A. Tamersoy, and C. Pu, "Click traffic analysis of short url spam on Twitter," in *Proc. IEEE CollaborateCom*, 2013.

[21] F. Klien and M. Strohmaier, "Short links under attack: geographical analysis of spam in a url shortener network," in *Proc. ACM HT*, 2012.

[22] S. Le Page, G.-V. Jourdan, G. V. Bochmann, J. Flood, and I.-V. Onut, "Using url shorteners to compare phishing and malware attacks," in *Proc. eCrime*, 2018.

[23] S. Chhabra, A. Aggarwal, F. Benevenuto, and P. Kumaraguru, "Phi.sh/$oCiaL: The phishing landscape through short URLs," in *Proc. CEAS*, 2011.

[24] N. Gupta, A. Aggarwal, and P. Kumaraguru, "bit. ly/malicious: Deep dive into short url based e-crime detection," in *Proc. eCrime*, 2014.

[25] Y. Kokubun and A. Nakamura, "Analysis of malicious urls on twitter," in *Proc. IEEE CSCI*, 2018.

[26] N. Nikiforakis, F. Maggi, G. Stringhini, M. Z. Rafique, W. Joosen, C. Kruegel, F. Piessens, G. Vigna, and S. Zanero, "Stranger danger: exploring the ecosystem of ad-based url shortening services," in *WWW*, 2014.

[27] D. Antoniades, I. Polakis, G. Kontaxis, E. Athanasopoulos, S. Ioannidis, E. P. Markatos, and T. Karagiannis, "we.b: The web of short URLs," in *Proc. WWW*, 2011.

[28] A. Farahat and M. C. Bailey, "How effective is targeted advertising?" in *WWW*, 2012.

[29] H. B. McMahan, G. Holt, D. Sculley, M. Young, and D. Ebner, "Ad click prediction: a view from the trenches," in *Proc. ACM KDD*, 2013.

[30] M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi, "Measuring user influence in Twitter: The million follower fallacy," in *ICWSM*, 2010.

[31] E. Bakshy, J. Hofman, W. Mason, and D. Watts, "Everyone's an influencer: quantifying influence on Twitter," in *ACM WSDM*, 2011.

[32] B. Krishnamurthy, P. Gill, and M. Arlitt, "A few chirps about twitter," *Proc. WOSN*, 2008.

[33] A. Campan, T. Atnafu, T. M. Tuta, and J. Nolan, "Is data collection through twitter streaming api useful for academic research?" *Proc. IEEE Big Data*, 2018.

[34] C. Llewellyn and L. Cram, "Distinguishing the wood from the trees: Contrasting collection methods to understand bias in a longitudinal Brexit Twitter dataset," *Proc. ICWSM*, 2017.

APPENDIX

High-level implementation: For data collection, we run multiple parallel processes. The master (M) is responsible for scheduling and starting up the other processes. For the first phases, it relies on three processes. The Tweet Gatherer (TG) has the sole task of opening up a stream to the Twitter API and constantly receive tweets from it. We then pipe these tweets to the Tweet Queuer (TQ), who adds these tweets to a queue. (By separating these two processes we could ensure that TG does not fall behind and therefore become disconnected from the API.) Finally, the Tweet Writer (TW) reads from the queue and writes them to a file (associated with the current block) on the File System (FS).

For the second phase, M delegates the responsibility to a Second Phase Data Collection (SPDC) process that for each new block that should be processed, starts two new parallel sub processes: (i) the Retweets Retriever (RR) and (ii) the Bitly Retriever (BR). RR reads tweets from FS and uses batch requests to the Twitter API to look up retweet information about (up-to) 100 tweets per request. BR extracts Bitly tweets from the same file and looks up information about these links directly using the Bitly API. For every Bitly link, we collect the full URL that the Bitly link redirects to, when the shortened link was created, and how many clicks it had received from different sources during different time periods. In particular, we used one call to obtain (i) all clicks since the creation of the link and (ii) all clicks generated via Twitter since this same time instances, and another call to obtain (iii) all clicks since the tweet was posted and (iv) all clicks generated via Twitter since the same instance. Using these combined calls, we reduced the number of calls to the Bitly API from six to four calls per Bitly link. To speed up BR further, we created four threads for each Bitly link and executed each of the calls to the Bitly API on parallel threads. When the information was returned for all four requests, it was written to a CSV file.

Dataset: For the analysis presented in this paper, we used several one-week long datasets. Each such dataset consisted of resulted in 42 four-blocks, which after aggregation resulted in a datasets of a few GB each, each containing information about a few million tweets. From the first phase we stored away (fields not always available in italics): the tweet ID, when tweet was posted, *ID of the place where the user posted from*, *name of the place where the user posted from*, *country of the place where the user posted from*, *coordinates of the user when the tweet was posted*, language of the tweet, *list of hashtags in the tweet*, *list of URLs in the tweet*, whether the tweet is a retweet or not, whether the tweet is in reply to another tweet, user ID of poster, when user account was created, how many followers user has, how many users the user is following, how many tweets the user has tweeted, how many tweets the user has favourited, whether or not the user is verified, and what language the user uses. From the second phase, the following retweet information is included: the number of retweets the tweet has received since posting, and the time when the retweet count was retrieved. Finally, the following Bitly related information is included (in the case it is a Bitly link): a list of the total number of clicks the Bitly links has received, a list of the total number of clicks the Bitly links has received that originate from Twitter, a list of the number of clicks the Bitly links has received since the tweet was posted, a list of the number of clicks the Bitly links has received that originate from Twitter since the tweet was posted, a list of the URLs the Bitly links redirect to, a list of time stamps when the Bitly links were posted, and the time when the Bitly data was retrieved. For easy analysis, all of the above fields were merged into one big dataset where every row of the CSV file is a tweet and if there is no data for a specific column for a tweet it is left empty. Code and example datasets can be found here: https://www.ida.liu.se/~nikca89/papers/tma21.html.