Conducting Research on Twitter: A Call for Guidelines and Metrics

Patrick Gage Kelley
University of New Mexico
Albuquerque, NM 87131 USA
pgk@cs.unm.edu
@patrickgage

Manya Sleeper
Carnegie Mellon U niversity
Pittsburgh, PA 15213 USA
msleeper@cs.cmu.edu

Justin Cranshaw
Carnegie Mellon University
Pittsburgh, PA 15213 USA
jcransh@cs.cmu.edu
@compurbanist

Abstract
Twitter has been broadly adopted by the privacy research community. However,Twitter research has limitations, and missteps often occur. Issues involve data access restrictions, user sampling and filtering, as well as legal and ethical concerns. Developing guidelines around these recurring problems as a community would help us better standardize and improve the quality of our work.

Author Keywords
Twitter; data; privacy; regret; API; IRB

ACM Classification Keywords
H.5.m [Information interfaces and presentation (e.g., HCI)]: Miscellaneous.

Introduction
The rapid growth of social networking sites (SNSs) like Twitter has drastically changed how we interact with technology and each other, altering how we document our lives, chat, gossip, and network.

Users often share personal data on SNSs in transparent, archived ways. As they do, we, the community of researchers studying online privacy, are interested in tracking metrics to better understand their evolving behaviors, both to examine emergent privacy issues and to

develop mechanisms to support their changing needs.

Such data tracking has been performed on a range of SNSs, but Twitter's relative ease of access has allowed it to emerge as the most commonly used [3] both broadly and for privacy researchers. However, despite its value, limitations to the data provided by the site as well as potential pitfalls for data collection lead to limitations that often go unacknowledged in Twitter-related research.

In this brief, we consider the level of access Twitter truly provides relative to desired metrics. We also discuss proper data sampling and legal/ethical Twitter data collection, two issues that frequently arise and are often not properly acknowledged. We believe that the community needs to recognize and develop guidelines around Twitter data reporting and collection to improve research comparability and quality.

Data We Can and Can't Have

Twitter is often viewed as a public platform, largely because of Twitter user data provided through APIs [7]. However, the data Twitter provides is limited because of a combination of data volume, user privacy expectations, and Twitter's business interests. In this section, we detail some of the major types of data that are and are not technically accessible through Twitter's current APIs, as well as some limitations on provided data.

Full Tweets— Researchers often want access to the full stream of all current tweets. Because of the volume of tweets, access to the full stream is not widely distributed. Partnerships can be made with the Twitter business team but are difficult. Smaller streams of data, available through the Global Public Streaming API are typically used instead. This feed is estimated at about 1% of published tweets, although the sampling process is unclear. This limitation should be understood and noted.

Researchers may want tweets that match specific criteria, for which there are also a number of limitations. Twitter provides a Search API to retrieve very-recent tweets with certain strings of text. On high volume, and possibly other, searches, tweets are sampled and not returned in full. If tweets are pulled for a specific users, there is also a cap on how many historic tweets can be retrieved for a given user.

Individual accounts are also rate-limited. In a certain time period, only a certain number of API calls can be made. Researchers often set up calls at regular intervals; however, as these calls are capped, data may be lost between calls. Polling details should be reported, and checks for data loss should be performed.

Log Data— Researchers also often want to know users' viewing patterns, such as when and on what devices users read Twitter, tweets clicked on and read, as well as messages partially typed but not sent. However, this data is not shared publicly. Similarly, while Twitter formerly displayed the client from which a user posted, information still available in the API, it has recently been removed from the website, and may become obfuscated from the public.

Private Data— Finally, researchers might want access to information users provided with an expectation of privacy. Direct messages (DMs) are private and not available to researchers unless a given user authenticates with an application registered to the research team. The details for users with "protected" accounts are also inaccessible to researchers who do not have such access.

Sampling, Filtering, and Finding The Twitter Users You Intended To

When gathering data, researchers must define and select a set of Twitter users to study. However, because of API limitations, as well as the varied uses of Twitter, this can be difficult. This section details some common pitfalls in sample selection and suggestions for avoiding such pitfalls, as well as the need for community-level guidelines.

Random Sampling— For privacy research, a random sample of users is often desirable either for a onetime sample, or for tracking users over time. Randomly sampling Twitter users is challenging. The most frequent method seems to be randomly sampling users from the public stream (drawn from already-sampled data). This approach may be biased by time of sampling, by favoring more active users, and by Twitter's preselection criteria.

Researchers also often expand their samples to include users who interact with the original set of participants, initiating a "snowball" effect. This technique, as well as snowball sampling, in which researchers use Twitter users with high numbers of followers to tweet links to studies, is biased around interconnected groups and dependent on the points of origin [2, 8]. Such methods, and their limitations, must be thoroughly described in Twitter research.

Active Users— Researchers also often designate subjects as "active users;" however, this metric is challenging to define and not commonly agreed upon. One major issue is that Twitter users' posting behaviors often ebb and flow over time. Short studies are thus unlikely to capture consistent patterns, as many users tweet rarely and with irregular frequency [1]. Also, sampling on the public stream may be biased toward users who post more frequently. Metrics to measure and define active users vary from paper to paper. Community-level guidelines (e.g., number of followers, follows, total tweets, tweets in a given timespan) around the definition of an "active user" or tiers of user activity would increase consistently, comparability, and repeatability across research.

Types of Users— Researchers also often seek to define the set of users included in their work. For example, on Twitter, spam-bots, which are accounts that only post ads, links, porn, or malware, are prevalent [5]. Spam-bots are often filtered out, however filtering methods are often not clearly defined and again vary by study. In other cases, researchers want to filter out, or select for, celebrity or organizational Twitter accounts (e.g., businesses or government entities), although there are not consistent profiles for such accounts. Developing community-level guidelines for selecting different types of accounts would again increase the quality of work community-wide.

Language Filtering— It is also often desirable to select Twitter users who speak a specific language. A user-specified language can be pulled from the API. However, it is not always an accurate representation of users' actual tweets. Post-hoc language filtering can also be performed. While most researchers generally seem to report on only English-speaking users, it is often unclear how this filtering is done. Again, community-level guidelines on language filtering and reporting would increase comparability and quality.

We Can Do This, But Should We? Ethical Considerations

After Twitter provides data, in a limited form, and researchers have decided on a means to gather the desired data, ethical and legal issues relevant to accessing and

using the data must be considered.

Terms of Service— The Twitter Terms of Service (TOS) are legal documents that govern how users, developers, and researchers access and use Twitter content on the site and via the API. The TOS are meant "to strike a balance between encouraging interesting development and protecting both Twitter's and users' rights" [6]. It is important to note that these documents are constantly evolving, sometimes disrupting established businesses whose practices become misaligned with TOS changes.

The TOS stipulate not only how developers may access Twitter content but also what they are allowed to do with the content after download. Broadly, the TOS cover how frequently and from how many machines developers are allowed to request content, the kinds of applications developers are allowed to build, as well as the data developers are allowed to access and how long they are allowed to store the data.

Often questions researchers want to answer with Twitter are impossible to study without violating the TOS. For example, despite Twitter's request that developers not "aggregate, cache, or store" [6] Twitter's geographic content, dozens of academic papers study the geography of tweets.

There are several reasons why researchers sidestep the TOS. For example, because researchers primarily collect data for analysis and discovery, they rarely build long-standing applications using the platform. Thus, adverse consequences from TOS breaches, beyond potentially being asked to delete downloaded content, may seem less likely. There is also a general belief that the research contribution benefits society and Twitter itself. This may perhaps be why Twitter, at least thus far, seems to condone academic breaches to the TOS. But this practice raises question. Is such behavior ethical? Legal? The lack of a dialog within the community on this topic, or established guidelines for researchers on these important issues, poses a serious threat to the long term sustainability of privacy research using Twitter data.

Institutional Review Boards— Although most Institutional Review Boards (IRBs) in the United States do not consider data harvested publicly from the Internet as human subject data, this distinction becomes more complex when the data is behind a log-in. Even though tweets are publicly visible, to access them through the API, one must authenticate with developer credentials. It is unclear whether IRBs are aware of this fact, and in general how this would affect perception of Twitter as human subject data.

Many studies also create custom Twitter clients or applications to monitor participants' Twitter behaviors by downloading and analyzing their tweets and direct messages. Most such studies, because they involve direct user participation, require IRB approval. However it is unclear if IRBs are aware of the distinction between generic tweets, which are regarded as public, and direct messages which are private and akin to email or SMS messages. If IRBs knew of the private nature of direct messages they might not approve studies that collected them. As a community we need to consider the ethical issues around working with data collected from custom Twitter clients to develop best practice guidelines.

Furthermore, we need to consider how IRBs, who are often charged with not only protecting human subjects but also minimizing liabilities to their institutions, would consider studies if they were fully aware of TOS violations [4].

Conclusion

As a research community, we would like our work to be more efficient, repeatable and have meaningful impact. Twitter is a wonderful platform for research and as such has been broadly adopted. We can improve the caliber of Twitter research if we can create community agreed upon standards and metrics for explaining Twitter methodologies, including how we sample, filter, and acquire data. As ethical researchers we must also consider the broader impact of collecting and using this data. Specifically, as privacy researchers, we should consider four high-level questions as a community:

1. When using Twitter data, how should we report measures taken to account for limitations (technical and otherwise) in the data available through the Twitter API?
2. What are best practices for sampling and filtering Twitter data and for reporting these processes?
3. How can we account for violations of the Twitter TOS when conducting our research?
4. What are best practices for Twitter privacy research when considering the ethics of human subjects research?

We hope that this workshop will allow for initial discussions leading toward community-driven best practice guidelines and metrics around each of the above areas. We also hope that as our group considers and proposes other metrics for studying privacy on social networks these questions will be taken into account.

References

[1] Abel, F., Gao, Q., Houben, G.J., and Tao, K. 2011. Analyzing User Modeling on Twitter for Personalized News Recommendations. UMAP 2011, LNCS 6787, pp. 112, 2011. http://ktao.nl/pdf/2011-wis-twitter-um-umap.pdf

[2] Biernacki, P., and Waldorf, D. 1981. Snowball Sampling: Problems and Techniques of Chain Referral Sampling. Sociological Methods & Research. vol 10, issue 2, pp. 141-163.

[3] boyd, d.m. 2012. Bibliography of Research on Twitter & Microblogging. http://www.danah.org/researchBibs/twitter.php.

[4] Code of Federal Regulations. Title 45. Part 46. Protection of Human Subjects. http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.html

[5] Grier, C., Thomas, K., Paxson, V., and Zhang, M. 2010. @spam: the underground on 140 characters or less. In Proceedings of the 17th ACM conference on Computer and communications security (CCS '10). ACM, New York, NY, USA, 27-37.

[6] Twitter. 2012. Terms of Service. http://twitter.com/tos.

[7] Twitter. 2012. The Streaming APIs. https://dev.twitter.com/docs/streaming-apis.

[8] Wu, S., Hofman, J.M., Mason, W.A., and Watts, D.J. 2011. Who Says What to Whom on Twitter. WWW 2011. pp. 705-714.