Understanding the removal of precise geotagging in tweets

Yingjie Hu*
Department of Geography,
University at Buffalo, The State University of New York

Ruo-Qian Wang
Department of Civil and Environmental Engineering,
Rutgers, The State University of New Jersey

Twitter announced on June 18, 2019 that it would remove the precise geotagging feature in tweets. Protecting the location privacy of users, this change also affects human behavior studies based on geotagged tweets. We discuss the potential impact of Twitter's decision and how researchers can respond to this change.

On June 18, 2019, Twitter announced that it would remove the precise geotagging feature in tweets (https://twitter.com/TwitterSupport/status/1141039841993355264). According to Twitter, this decision was based on the observation that most people do not use precise geotagging. This announcement triggered heated discussions among the general public and the research community both for and against the decision. The discussions were so intense that Twitter made a follow-up three days later (https://twitter.com/twittersupport/status/1142130343715078144) clarifying that they only removed precise geotagging while general geotagging remained unchanged. So, what is geotagging and why did Twitter's decision draw so much attention? How does this decision affect researchers?

Precise and general geotagging

Geotagging broadly refers to assigning geographic locations to digital content, such as photos, text messages, and videos. For Twitter data, geotagging means assigning locations to tweets, which are usually the current locations of the Twitter users. Twitter further differentiates precise geotagging and general geotagging: precise geotagging means the tagged location is a pair of latitude and longitude coordinates usually obtained from the GPS receiver of a mobile device, whereas general geotagging means the location is a place, such as a restaurant, a park, a city, or a country. Compared to precise geotagging, general geotagging only shows a rough location, e.g., the bounding box of the tagged place. Therefore, removing precise geotagging protects location privacy and is welcomed by many people. Meanwhile, from a research perspective, geotagged tweets have been widely used in various scientific studies[1], and researchers may worry that Twitter's decision

*Corresponding author: Yingjie Hu (yhu42@buffalo.edu)

will result in a large drop in the number of precisely geotagged tweets and even the loss of a valuable data source. In full support of protecting user privacy, this Comment discusses the potential impact of Twitter's decision on related research.

Geotagged tweets in research

Geotagged tweets possess several characteristics valuable for human behavior related research. They provide real-time information about human activities on the ground, cover large geographic areas across countries, can be retrieved via Twitter API with low cost, and have high spatial and temporal resolutions. In addition, the tagged locations enable innovative research by allowing tweets to be linked with many other datasets (e.g., the official crime statistics of an area) using geography as a common key.

In recent years, geotagged tweets have been used in various studies. In natural hazard management, they have been used to locate the affected people and understand the impact of a disaster, such as the 2011 earthquake on the East Coast of the US[2]. In public health, geotagged tweets have been utilized to map and predict the locations of disease outbreaks, such as flu[3]. In human mobility and transportation studies, researchers have employed geotagged tweets to understand the movement of people with the goal of building better transport systems[4]. In politics and public policy research, geotagged tweets have been used to examine the reactions of people towards new policies and political events, such as the 2016 presidential campaigns[5]. There also exist many other studies based on geotagged tweets, such as understanding the attitudes of people in different geographic areas towards climate change[6]. Although the users who post geotagged tweets are not representative of the entire population[7], geotagged tweets can be a useful data source for asking research questions or providing complementary information to studies.

While many studies based on geotagged tweets contribute to the welfare of the society, they can also raise privacy and ethical concerns. For an individual Twitter user, the tagged geographic locations, especially recorded over a long time period, can reveal sensitive information about the location preferences and daily life trajectories of the user. Such information is highly personal and can become more sensitive when the tagged locations are in the form of precise longitudes and latitudes. Much research has been devoted into the important topic of geoprivacy, and one of the goals is to protect the privacy and anonymity of the individuals in location-identifiable datasets such as geotagged tweets, while preserving the ability of extracting meaningful knowledge from the data[8]. In addition to geoprivacy research, we as researchers have the ethical responsibility to protect the privacy of individuals when working with geotagged tweets.

Understanding the change

To understand the impact of Twitter's decision from a research perspective, we first check a few facts about its geotagging service. As of August 16, 2019, we can confirm that the option of directly tagging the precise GPS coordinates of a tweet is disabled, but a Twitter

user can still add location information to a tweet via one of the three following approaches. As illustrated in Fig. 1, a user can use either (1) the general geotagging option by assigning a place to a tweet (Fig. 1a), (2) the precise geotagging option but for only the photos captured by Twitter's mobile app (Fig. 1b), or (3) a third-party app connected to Twitter (e.g., Instagram) to assign precise coordinates to a tweet (Fig. 1c). Using Twitter's existing API, we find that the precise latitudes and longitudes of the tweets that are geotagged by Approach (2) or (3) can still be collected, but the location information tagged through Approach (1) is in the form of a bounding box containing the tagged place.

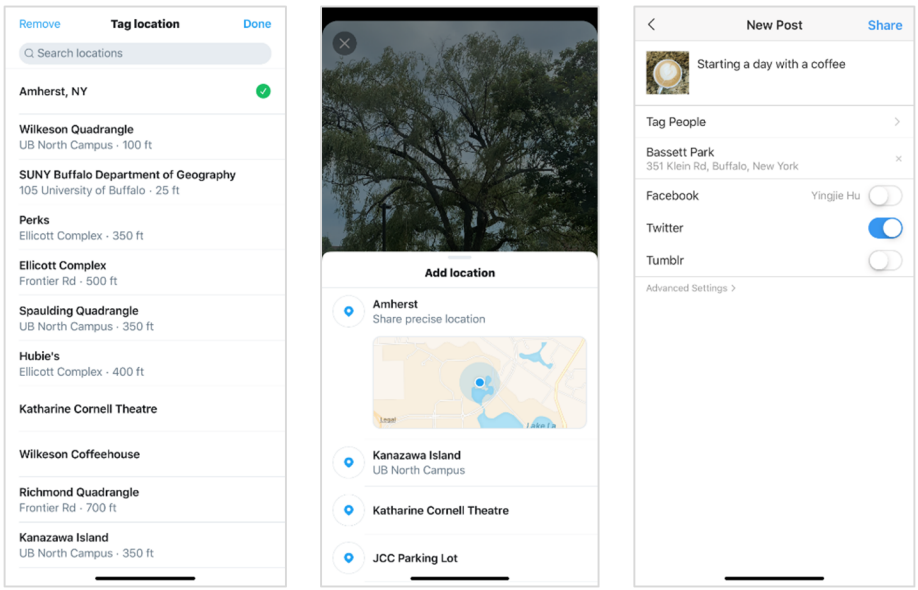

Fig. 1 The three remaining approaches of geotagging after Twitter's decision: (a) general geotagging with a place; (b) precise geotagging for photos only; (c) precise geotagging via a third-party app (Instagram as an example).

In Twitter's announcement on June 18, the company stated that "Most people don't tag their precise location ..." However, in aggregate, it was often not difficult to collect thousands or even millions of precisely geotagged tweets (see, for example [9]). If most people did not use precise geotagging, where did these geotagged tweets come from? One possibility is that most geotagged tweets, in fact, came from the third-party apps. On apps like Instagram, users can choose to enable precise geotagging and share the geotagged content onto other platforms (Fig. 1c). If most geotagged tweets were already from other apps, the removal of Twitter's own precise geotagging would have a limited impact on related research.

We look into three different datasets of geotagged tweets to check their sources. These three datasets were collected for other studies on people's reactions to extreme weather events from January to March 2019. Each tweet in these three datasets is tagged with a pair of latitude and longitude coordinates. We use Twitter API via the Twitter4J Java library to find out which source platforms these precisely geotagged tweets came from, and then classify these platforms into three categories: Twitter, third-party apps, and unknown platforms. Note that some tweets were deleted when

we were trying to find their source platforms, and thus, their platforms are labeled as unknown. We summarize the counts and percentages of the three categories of platforms in Table 1.

Target Event	Time Period	Total Tweets	From Twitter	From 3rd-party	From Unknown
Winter Storm Gia	01/05/2019-01/30/2019	25836	4976 (19.3%)	19875 (76.9%)	985 (3.8%)
Tornados in the southeastern US (1)	02/11/2019-03/03/2019	14117	1119 (7.9%)	12413 (87.9%)	585 (4.1%)
Tornados in the southeastern US (2)	02/23/2019-03/15/2019	14983	3704 (24.7%)	10767 (71.9%)	512 (3.4%)

Table 1: Three different datasets of precisely geotagged tweets and their sources.

Based on the three datasets in Table 1, about 72% to 88% of precisely geotagged tweets were from third-party apps, such as Instagram, whereas only about 8% to 25% were directly from Twitter. Although the three datasets cannot exhaust all possible datasets that can be retrieved through Twitter API, these results suggest that Twitter's decision may not have an earthshaking impact on the research relying on geotagged tweets. Besides, research that examines all tweets, rather than geotagged tweets alone, is less likely to be affected by the change of geotagging options, since geotagged tweets account for only about 0.85% of all tweets[7]. However, even if the data availability does not substantially change, researchers should continue to prioritize privacy and safety considerations when conducting tweet-based research.

Responding to the change

To respond to Twitter's decision, we need to both adjust research methods and improve privacy protection practices. From a research perspective, while researchers can continue (responsibly) using the remaining precisely geotagged tweets, there is nevertheless a decrease in this already sparse data resource. Two approaches can be employed to increase the number of precisely geotagged tweets. First, researchers can include the locations from general geotagging by e.g., calculating the geometric center of the bounding box of a tagged place. While this approach would add location uncertainty to the data, it can be acceptable for studies at large geographic scales, such as at country or world scales, when the tagged places are points of interest, neighborhoods, or cities. Second, researchers can leverage a method called geoparsing. This approach does not rely on any existing geotagged locations of tweets, but recognizes the place names mentioned in tweet content and geo-locates these names to their corresponding latitudes and longitudes using a gazetteer (a geographic dictionary with place names and their locations). This approach could largely expand the size of geotagged tweets since more than 10% tweets mention place names in their content[10]. While geoparsing does introduce more location uncertainty or even errors into the data, it can help us understand the locations that people are talking about instead of only where the tweets come from.

From a privacy protection perspective, Twitter's decision reflects the concerns of the

society in general on privacy issues. Researchers should increase our awareness of the potential privacy and safety issues that may exist in our data and research practice, and follow relevant guidelines, such as those from the institutional review board (IRB), to protect the privacy of individuals. For research based on geotagged tweets, using the remaining precisely geotagged tweets or adopting the two approaches discussed above to increase the number of geotagged tweets in data can still raise privacy concerns that contributed to Twitter's decision. Accordingly, researchers should carefully consider the dimension of privacy when continuing to analyze geotagged tweets and when using alternative methods for identifying location.

Conclusions

Twitter's decision may not result in a major loss of precisely geotagged tweets. By taking proper measures to protect user privacy, researchers can still responsibly use geotagged tweets for various studies that benefit our society. The removal of precise geotagging in tweets better protects user privacy in cases where the precise locations of users were revealed inadvertently. Meanwhile, the implications of Twitter's decision for the research community deserve our further thoughts. We are in an era when many researchers no longer collect data themselves but rely on the data from commercial platforms, because the required data collecting capability to answer some research questions is beyond what a single researcher or a research group has. A direct consequence is that a company's decision could largely affect the research agenda of many researchers. We need to diversify our data sources and increase the resilience of our research agenda to future changes. In addition, collaborations among researchers and between researchers and industry partners throughout the world may be further developed to create open and reliable data resources that both protect user privacy and support scientific research.

Competing Interests

The authors declare no competing interests.

References

1 Huang, B. & Carley, K.M. in *Proc. of Int. Conf. on Adv. Soc. Netw. Anal. Min.* 1-9 (2019).

2 Crooks, A., Croitoru, A., Stefanidis, A. & Radzikowski, J. *Trans. in GIS* **17**, 124-147 (2013).

3 Padmanabhan, A. *et al. Concurrency Computat.: Pract. Exper.* **26**, 2253-2265 (2014).

4 Rashidi, T.H., Abbasi, A., Maghrebi, M., Hasan, S. & Waller, T.S. *Transp. Res. Part C Emerg. Technol.* **75**, 197-211 (2017).

5 Hobbs, W. & Lajevardi, N. *Am. Polit. Sci. Rev.* **113**, 270-276 (2019).

6 Dahal, B., Kumar, S.A. & Li, Z. *Soc. Netw. Anal. Min.* **9**, 1-20 (2019).

7 Sloan, L. *et al. Socio. Res. Online* **18**, 1-11 (2013).

8 Keßler, C. & McKenzie, G. *Trans. in GIS* **22**, 3-19 (2018).

9 Phillips, N.E., Levy, B.L., Sampson, R.J., Small, M.L. & Wang, R.Q. *Socio. Meth. Res.*, 1-40 (2019).

10 Wallgrün, J.O., Karimzadeh, M., MacEachren, A.M. & Pezanowski, S. *Int. J. Geogr. Inform. Sci.* **32**, 1-29 (2018).