**Data Collection**

We collected Twitter user timelines using the public Streaming and REST APIs (https://dev.twitter.com/docs). Twitter does not provide a simple way of randomly sampling users, so we used the following methodology to collect as diverse a set as possible. First, we queried the Streaming API with a set of common English words (i, u, you, he, she, we, they, it, a, an, the, one, all, this, that, these, his, her, their, m, is, s, are, was, were, be, do, have, has, had, will, would, can, could, should, may, might, there, what, where, who, which, if, as, up, in, on, at, by, of, for, out, from, with, to, about, and, or, but, so, not). Second, we identified the authors who had published at least one tweet that was embedded with GPS coordinates. Lastly, we downloaded the most recent tweets (up to a limit of 3,000) of these authors. We successfully downloaded 248,953 user timelines during June to December 2016. We restricted to English tweets to simplify filtering based on keywords, and we required GPS both to allow comparison of user patterns across geographic regions, and to map posting times to user’s local time zones.

**Bot Removal**

While randomly checking the downloaded user timelines, we noticed bot-like accounts devoted to 4 specific types of content. These user timelines served the purposes of providing information regarding local businesses, jobs, traffic, and weather, but rarely contained individual opinion or reflected personal life experiences. We assumed that these accounts are automated by computer scripts or jointly operated by a group of people instead of an individual, and they should not be included in our analysis.

We utilized the BotOrNot service (http://truthy.indiana.edu/botornot) for classifying whether our downloaded user timelines belong to human or bot accounts. The service returns a *bot score* that estimated the probability that the user in question was a bot. We were able to obtain the bot scores for 50,334 users that we have downloaded.

We also applied a bag-of-words approach for identifying these accounts. For each of the 4 content types, we first reviewed all tweets from one user timeline that we found suspicious and identified a set of keywords that were common to many of the tweets. Then we reviewed other suspicious tweets that did not contain any of the identified keywords to explore more keywords. We iteratively identified more keywords until there were less than 25 tweets that did not contain any of the keywords, and repeated this process on at least three user timelines until no further keywords could be added. The set of identified keywords were assumed to mark a specific type of content. We found “earthquake, humidity, cloudy, sunny, weather, sunrise, visibility, rain, temperature, sunset, clouds” to be commonly shared by Twitter accounts that were devoted to weather updates; “delay, caution, accident, road work, miles, fire rescue, serious, hazard, heavy traffic, breakdown, construction zone, ramp, blocked, vehicle” for traffic status; and “job, hiring” for the job postings. Also, we found that word patterns “X% off, $X off, half off, save $X, save X%, $X for” (where X denotes any numeric values) were commonly shared among the tweets for promoting local business. We excluded user timelines for which a sizable proportion of the tweets seemed to match any of the 4 types. For each type, we derived the ratio of keyword-containing tweets to all tweets among all user timelines. We calculated the first Jenk’s natural breaks from the ratios that we derived above as the threshold for each type. Any user timeline that had any ratio higher than its corresponding threshold was considered a non-human and was excluded. We removed 17,673 user timelines through this process.