Population Bias in Geotagged Tweets

Momin M. Malik, Hemank Lamba, Constantine Nakos, and **Jürgen Pfeffer**
Institute for Software Research
School of Computer Science
Carnegie Mellon University

Abstract

Geotagged tweets are an exciting and increasingly popular data source, but like all social media data, they potentially have biases in who are represented. Motivated by this, we investigate the question, 'are users of geotagged tweets randomly distributed over the US population'? We link approximately 144 million geotagged tweets within the US, representing 2.6m unique users, to high-resolution Census population data and carry out a statistical test by which we answer this question strongly in the negative. We utilize spatial models and integrate further Census data to investigate the factors associated with this nonrandom distribution. We find that, controlling for other factors, population has no effect on the number of geotag users, and instead it is predicted by a number of factors including higher median income, being in an urban area, being further east or on a coast, having more young people, and having high Asian, Black or Hispanic/Latino populations.

'Geotagged' or 'geocoded' tweets, where users elect to automatically include their exact latitude/longitude geocoordinates in tweet metadata, provide data that are:

- High-quality: geotagging is automated, so there are fewer chances of data error such as from user specification (Graham, Hale, and Gaffney 2014; Hecht et al. 2011);
- Precise: geotags are down to a ten thousandth of a degree in latitude and longitude;
- Richly contextual: geotags are connected to tweets with all their temporal, semantic, and social content;
- Easily available, through the Streaming API;
- Large: using the Streaming API, a researcher can build a collection of tens of millions of tweets.

Unsurprisingly, this makes them an enormously attractive source for studying a wide range of human phenomena (Hong et al. 2012). Existing works have used geotagged tweets to study

- mobility patterns (Yuan et al. 2013; Cho, Myers, and Leskovec 2011),
- urban life (Doran, Gokhale, and Dagnino 2013; Frias-Martinez et al. 2012),
- transportation (Wang et al. 2014a),
- natural disasters, crises, and disaster response (Morstatter et al. 2014; Lin and Margolin 2014; Shelton et al. 2014; Sylvester et al. 2014; Kumar, Hu, and Liu 2014), and
- public health (Sylvester et al. 2014; Nagar et al. 2014; Ghosh and Guha 2013)

as well as the interplay between geography and

- language (Hong et al. 2012; Eisenstein et al. 2010; Kinsella, Murdock, and O'Hare 2011),
- discourse (Leetaru et al. 2013),
- information diffusion and flows (Kamath et al. 2013; van Liere 2010),
- emotion (Mitchell et al. 2013), and
- social ties (Stephens and Poorthuis 2014; Takhteyev, Gruzd, and Wellman 2012; Cho, Myers, and Leskovec 2011).

Furthermore, maps of geotagged tweets tend to look remarkably similar to maps of population density (figs. 1 and 2; see also Leetaru et al., 2013), even if there are differences at a finer scale (figs. 3 and 4). This naturally leads to the question: are Twitter users who send geotagged tweets (henceforth, 'geotag users') randomly distributed over the population? This is a critical question because, if users who elect to geotag are systematically different from people in general, the results of studying geotagged tweets will not have external validity.

We used the Twitter API to get a collection of 144,877,685 geotagged tweets from the contiguous US, from which we extract 2,612,876 unique twitter handles. We uniquely assign each handle to a *block group*, a geographic designation of the US Census Bureau that is the smallest geographic unit for which Census data is publicly available. We then link the counts of unique geotag users per block group to the 2010 Decennial Census population counts per block group, and create a statistical test for the null hypothesis that geotag users are randomly distributed over the US population. We find sufficient evidence to reject this null. Using other Census data, we then use a Simultaneous Autoregressive (SAR) model to test some candidate explanatory factors and investigate what is nonrandom about this distribution. This is, to our knowledge, the first paper to use

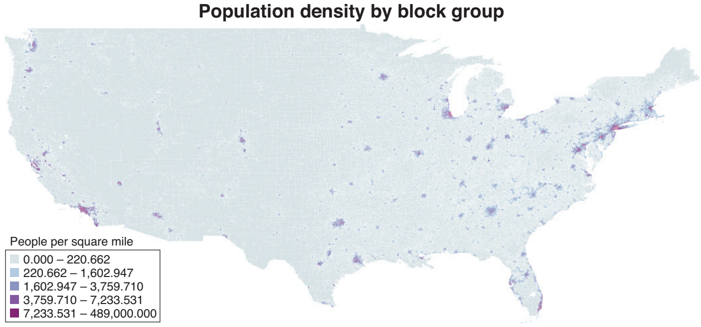

Figure 1: Quintiles of population per square mile by ‘block group’ (see below) in the 2010 Decennial Census.

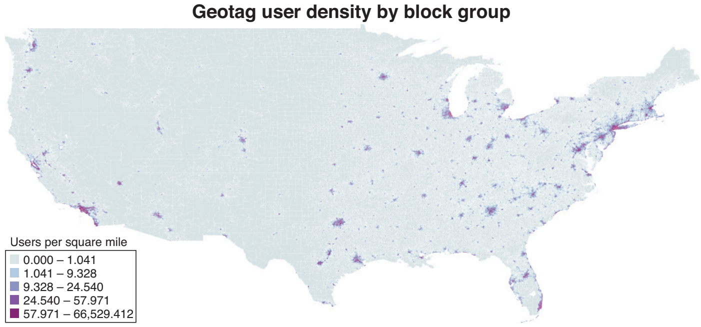

Figure 2: Quintiles of geotag users, uniquely assigned (see ‘mobile users’ below) per block group, divided by block group area.

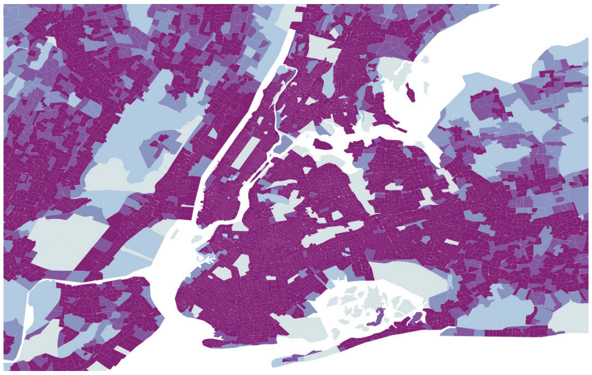

Figure 3: Detail of fig. (1) for New York.

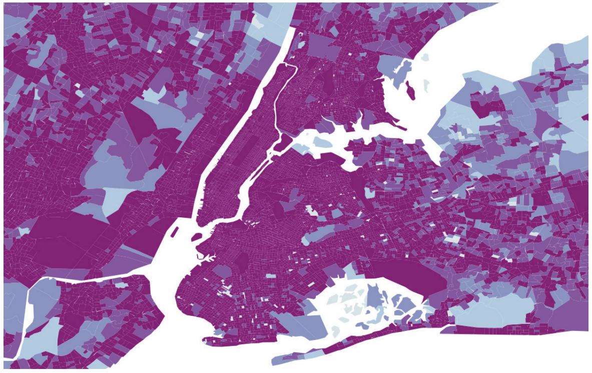

Figure 4: Detail of fig. (2) for New York.

statistical testing to establish population bias along multiple dimensions in geotagged tweets across the entire United States.

Background and Related Work

Our study relates to an increasing body of work about biases in who and what is represented in social media data. The first work with Twitter data was by Mislove et al. (2011), who found an overrepresentation of populous counties and an underrepresentation specifically of the Midwest, an undersampling in counties in southwest with large Hispanic populations, an undersampling in counties in the south and midwest with large Black populations, and an oversampling of counties associated with major cities with large White populations. However, these findings come from interpretations of distributions and county-level cartograms, rather than from statistical testing, and they rely on the user-defined ‘location’ field, which has been shown to have many inconsistencies (Graham, Hale, and Gaffney 2014; Hecht et al. 2011). Our study is on the one hand deeper because we use the far higher resolution of block groups and carry out statistical tests, but on the other hand not as general because our findings apply only to characteristics of *geotag users* within the US population rather than to geotag users within the Twitter population, or to Twitter users within the US population. Also worth noting is that Twitter has undergone large changes since the data used by Mislove et al., both in the governance and management of the platform itself (van Dijck 2013) and in patterns of user behavior (Liu, Kliman-Silver, and Mislove 2014).

More recently, Hecht and Stephens (2014) investigated urban biases across the US. Collecting 56.7m tweets from 1.6m users over a 25-day period in August and September 2013 and comparing it to Census data, they use a method of calculating a reduced effective sample size in order to correct for spatial dependencies. From this they calculate ratios of users per capita and find a bias towards urban areas, with 5.3 times more geotagged tweets per capita in urban regions as in rural ones, a magnitude even more pronounced in Foursquare data. Longley, Adnan, and Lansley (2015) investigate biases across a number of factors, focusing on the Greater London area. Using work on forename-surname pairs identifying gender, age and ethnicity, they parse usernames and other profile information to get a collection of estimated names, which they then compare to the 2011 UK Census and find an overrepresentation of young males, an underrepresentation of middle-aged and older females, an overrepresentation of White British users, and underrepresentation of South Asian, West Indian, and Chinese users, although tests of significance are not applied.

Coming from another methodological direction, a nationally representative survey study of smartphone owners (n=1,178) by Pew (Zickuhr 2013) looks at the demographics of location service users. Overall, 12% of those surveyed reported using what Pew terms ‘geosocial’ services (which includes geotagged tweets, and excludes informational services like Google Maps). Interestingly, the survey finds the the most frequent users of geosocial services are those of low*est* income and middle income; those of low*er* income use it less, and those of upper income use it least. More 18-26 year olds use geosocial services than older users, and almost double the proportion of hispanic (English- and Spanish-speaking) smartphone owners user geosocial services as compared to white and black (both non-hispanic) smartphone owners. However, out of the respondents who specified which geosocial services they use (n=141), most reported using Facebook (39%), Foursquare (18%) or Google Plus (14%); only 1%, or 1 respondent, used Twitter's geosocial services (i.e., geotagged tweets), such that it is not possible to make inferences about geotag users from the results of this study.

Our paper is answering the general call for stronger methodological investigations about the nature of population representation in social media data (Ruths and Pfeffer 2014; Tufekci 2014), as well as the specific call for combining geographic data from user-generated sources with non-user-generated sources, such as Twitter data with the Census (Crampton et al. 2013).

Method

Data collection

Geo-Coded Twitter Data. From Twitter's Streaming API, we collected 144,877,685 tweets from April 1 to July 1, 2013 using the geographic boundary box $[124.7625, 66.9326]W \times [24.5210, 49.3845]N$. This covers the contiguous US (i.e., the 48 adjoining US states and Washington DC but not Alaska, Hawaii, or offshore US territories and possessions). Consequently, all our tweets are geo-coded with lat/long GPS coordinates. As Morstatter et al. (2013) report from the Twitter Firehose, about 1.4% of tweets are geotagged; and elsewhere (Morstatter, Pfeffer, and Liu 2014) they report the Streaming API is more likely to be biased when the response to a query exceeds 1% of the total volume of tweets. Given also that North America accounted for only 22.32% of geotagged tweets in their collection, a fraction consistent with what Liu, Kliman-Silver, and Mislove report finding in a collection of decahose data covering the time period we consider, it is reasonable to assume that the use of the Twitter API to collect tweets geotagged in the US covers all or nearly all of geotagged tweets within the given time frame and geographic bounds.

Since the distribution of geotagged tweets over geotag users is characteristically long-tailed (fig. 5), with a minority of users sending out the majority of tweets, we decided that the relevant quantity was the number of geotag users rather than the number of tweets. We identified 2,612,876 unique user accounts in our data, which is the basis of our analysis.

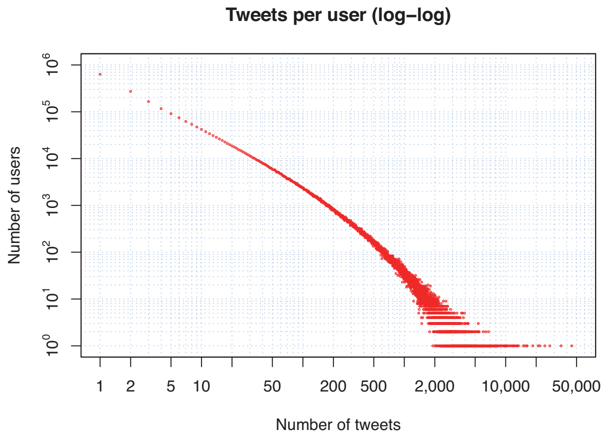

Figure 5: The usual long-tailed distribution of the number of users who have tweeted a certain number of tweets. Because of this skew, we focus on unique users alone, and ignore the volume of tweets.

Geospatial Data. The contiguous US plus Washington DC include 215,798[1] block groups (2010 specification) which range in size from .002 square miles to 7503.21 square miles. Block groups are designed by the Census Bureau to have roughly comparable population sizes. We verified this by noting that, in log scale, the distribution of populations per block group has a symmetric distribution and stable variance. Each block group has a unique identifier, the 12 digit *FIPS Code*, consisting of identifiers for state (first two digits), county (next three digits), tract (next six digits), and block group (last digit). For every state, the US Census Bureau provides geographic boundary files ('shapefiles') that includes the GPS coordinates of the borders of every block group within the state. We combined the shapefiles of the 48 contiguous states and the District of Columbia, deleting 364 block groups representing bodies of water (identifiable by being coded as having zero area, and having a FIPS code ending in zero[2]). With Python code (utilizing the `shapely` package) we identified the Census block group into which each tweet fell.

Socioeconomic Data. While the ideal would be to have rich and timely demographic data about the users who sent the tweets in our data, this is not realistic to collect for 2.6m users. But by aggregating data at the level of block groups, we can link Twitter data to the enormously rich demographic data the Census Bureau makes available at this level. We primarily use data from the 2010 Decennial Census, which we supplement with median income (not available in the Decennial Census) estimates from the 2009-2013 American Community Survey. For this ACS data, there were 1,224 block groups with missing values for median income, few enough that we filled these out as zeros rather than using imputation or smoothing. We also set 21 block groups with the value "2,500-" to 2,500, and 2,651 block groups

[1] Probably due to a rounding error in geographic calculations, we lost three small island block groups (2 in Florida, 1 in New York), such that our n = 215,795.

[2] https://www.census.gov/geo/reference/gtc/gtc_bg.html

with the value “250,000+” to 250,000. The 2009-2013 ACS had 54 block groups in the contiguous US whose boundaries (and FIPS) codes were from the 2000 Census, for which we found equivalent block groups in the 2010 Decennial Census to which to map. While the ACS 1-year estimates are more timely, they are more sparse and only at the county level (American Community Survey Office 2014), and we decided to prioritize the accuracy and completeness of values in the Decennial Census for this analysis. We similarly decided to not use the ACS 2009-2013 estimates for population quantities as there was more missing data, and there was high correlation between the 5-year estimates and 2010 Decennial Census figures across variables (generally around .95). Still, prioritizing timeliness over completeness, and looking at the county level with 2013 ACS 1-year estimates, may be the focus in future analysis.

Mobile users. Our construct of interest is the *number of potential geotag users*, for which population is the available proxy; there are cases where there are more geotag users than population, which points to tourists or, more generally, mobile users, as a complicating factor (Hecht and Stephens 2014).

Hecht and Stephens (2014) provide a useful review of techniques to uniquely assign users to a single geographic region. They identify two candidate techniques: temporal, where a user must send at least two tweets a set number of days apart in a region for the user to be located uniquely in that region, and ‘plurality rules,’ where the most frequently tweeted-from region is taken as the unique location of the user. Checking the ‘location’ field fails because of the low quality of the information there (Hecht et al. 2011). As one other option, Wang et al. 2014b use the location of the first geotagged tweet sent by a user as the location of the user. This is the simplest, but also has no motivation beyond convenience.

Despite the drawbacks of plurality not accounting for people local to two regions, our comparison is with the US Census which also does not account for this possibility. However, another problem is that foreign tourists are not counted in the US Census (unlike domestic tourists, who reside in some US block group), and of which there were 70m in the US in 2013[3]. This is substantial when compared to the total 2013 US population of 316m[4] (of which 307m are counted in the block groups we use). If many foreign tourists send geotagged tweets, it would introduce unaddressed bias; since our data collection only had geotagged tweets in the US, short of massive additional data collection we are unable to identify foreign tourists (such as by looking at the proportion of geotagged tweets outside of the US). This is a potential problem in our analysis that may be a topic for clarification in future work.

Additionally, we filter users by the number of tweets, considering only those with a certain number of tweets.[5] As the distribution of tweets per user (fig. 5) is smooth and has no natural break point, we arbitrarily pick 5 and 10 as cutoffs to use alongside all users.

Statistical Models

Random distribution over population. The basic relationship in which we are interested is between population and geotag users. In order to make a concrete test for random distribution, we suggest a model where there is a linear relationship between the population count and the number of users, i.e. users are drawn from the population at a constant rate subject to some noise. We can imagine the noise is heteroskedastic, which suggests the following data-generating process over population P, users U, and mean-zero noise term ε:

$$U = \alpha P + \varepsilon P \tag{1}$$

We transform both users and population to stabilize their variances, so this then becomes

$$\log U = \log \alpha + \log P + \log\left(1 + \frac{\varepsilon}{\alpha}\right) \tag{2}$$

Then, consider the linear model

$$\log U = \beta_0 + \beta_1 \log P + \varepsilon' \tag{3}$$

If eqn. (1) described the true data-generating process, from eqn. (3) we should get that $\hat{\beta}_1 = 1$, and then $\exp(\hat{\beta}_0)$ would estimate the value of the proportion α. That is, the $\log \alpha$ term is the intercept of the regression of $\log P$ onto $\log U$, and $\log\left(1 + \frac{\varepsilon}{\alpha}\right)$ is a mean zero error term now independent of P, and we have a null hypothesis $H_0 : \beta_1 = 0$. While this may seem unrealistic as a null model, other quantities that we would believe are randomly distributed proportional to population indeed match this. For example, we regressed log population onto log males and found it to be meaningful (presented below under results). With this validation, we argue that the model of eqn. (1) is a reasonable way of representing a quantity being randomly distributed over the population. Note that our interest is not in fitting this specific model and interpreting the parameters, but just having a way to test the null hypothesis of random distribution. Note also that we originally sought to compare log population density to log geotag user density as a way of treating measures on different block groups as equivalent (given that block groups are already designed to somewhat control for the variance in population density), but found that it produced excellent fits that did not disappear when the data was shuffled, suggesting that the dividing by area created artifactual relationships.

Model specification For comparison with analyses of race and Hispanic populations (Mislove et al. 2011; Zickuhr 2013), we use Census variables[6] P0030001 through P0030008 and P0040001 through P0040003. For comparison with analyses by age (Longley, Adnan, and Lansley 2015; Zickuhr 2013), we use P0120003 through P0120049 and aggregate across gender into the same age bins as in Zickuhr. Existing analyses by sex (Longley, Adnan, and Lansley 2015; Zickuhr 2013; Mislove et al. 2011) is based

[3] http://travel.trade.gov/view/m-2013-I-001/table1.html

[4] http://data.worldbank.org/indicator/SP.POP.TOTL

[5] We thank an anonymous reviewer for this fruitful suggestion.

[6] http://api.census.gov/data/2010/sf1/variables.html

on name-based inference or survey data; we decided that, while the Census does have sex data, the even distribution of sex across the US means that the sex ratio of a block group is not a meaningful proxy for geotag users who live there. For comparison with analyses of urban and rural populations (Hecht and Stephens 2014; Zickuhr 2013), we use P0020002 through P0020005.[7]

Thus, in total, we include terms for populations, the black population, the Asian population, the Hispanic/Latino population, the rural population, and respective populations of people ages 10-17, 18-29, 30-49, 50-64, and 65+. For all of these, we stabilize variance with a log transformation with add-one smoothing. We include median income (Zickuhr 2013), and test for a northern/eastern effect by including the (demeaned) latitudes and longitudes of block group centroids, and for a coastal effect by including terms for latitude and longitude squared.

Spatial autocorrelation. Discretization into uneven geographic units (as block groups certainly are) can cause statistical artifacts. Specifically, if the divisions do not correspond to the contours of the underlying spatial process (and there is little reason to believe they would), there will be dependencies between proximate geographic areas, and not accounting for this can inflate the R^2 statistic, shrink standard errors, and give misleadingly significant results. We use the standard statistic for measuring spatial autocorrelation, Moran's I,

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(X_i - \overline{X})(X_j - \overline{X})}{\sum_i (X_i - \overline{X})^2} \tag{4}$$

This is the empirical covariance, appropriately normalized, of the values of variable X between geographic units i and j. $W = [w_{ij}]$ is an $n \times n$ matrix of weights, discussed below. Rather than exploring autocorrelation in individual variables, we look for spatial autocorrelation in the residuals of a linear model (Anselin and Rey 1991). For management of spatial data and implementation of computation and estimation for spatial models, we used the R package `spdep` (Bivand and Piras 2015; Bivand, Hauke, and Kossowski 2013).

Weights matrix. Measuring spatial autocorrelation requires a 'weights matrix' of adjacencies between geographic units. There are multiple ways to generate this, and the choice of how to do so represents a substantive decision based on the problem at hand (Gaetan and Guyon 2012). However, given that we do not know in advance the form of the spatial autocorrelation, in practice we can test for autocorrelation over different choices of weights matrices to see which is most appropriate (Anselin, Sridharan, and Gholston 2007). Thus, we consider the following weights matrices:

- Queen contiguity (regions sharing a corner or edge are adjacent, equivalent to 8-connectivity in image processing);
- Rook contiguity (regions sharing an edge are adjacent, equivalent to 4-connectivity in image processing)
- k-nearest-neighbors for $k = \{2, 3, 4, 5, 6, 7, 8\}$, calculated from the midpoints of block groups.

For the contiguity cases, we consider both row-normalized (which normalizes the 'effect' of each neighboring unit such that they sum to one) and binary (which gives greater possibility for autocorrelation between a unit and its neighbors for units with more neighbors).

Spatial errors model. We model the relationship between population and geotag users using a Simultaneous Autoregressive (SAR) model, which is where one or more terms in the regression are correlated with itself. The main autoregressive model assumes that the residuals of unit i are correlated with the residuals of those units j adjacent to i, which is known in econometrics literature as a spatial errors model. The adjacencies are indexed exactly by the terms of the weights matrix. This gives the following two equations,

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{u} \tag{5}$$

$$\mathbf{u} = \lambda \mathbf{W}\mathbf{u} + \varepsilon \tag{6}$$

where u are the correlated residuals, $\varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ are the uncorrelated error terms, and the coefficient λ is the 'spatial multiplier' that captures the strength of the spatial autocorrelation (Anselin 2002). While there are other SAR models, we use spatial errors as the simplest to interpret and the most appropriate for our purpose.

Results and Discussion

Observational results. The block groups with the highest number of distinct users (before users are assigned uniquely) are major international airports and major tourist attractions (table 1).[8] The inclusion of several international airports on the list suggests that geotagging tweets during the process of travel is a common user behavior. There were some areas with zero population but nonzero users; out of these, the ones with the highest counts of distinct users are mostly the same: major airports and parks.[9]

Conversely, there were only 67 block groups from which nobody sent geotagged tweets; only 30 of these also had no population (these were national forests, minor airports, areas off highways, etc.). Of those that did have a population, the most populous was a block group with a population of 4,854 within San Quentin State Prison in California. The second-most populous block group is also a Corrections Department building in Texas, and third is a state prison in California (although not all prisons lack geotag tweet users; the block group of Rikers Island in New York has geotagged tweets from 22 users).

Out of the 2,612,876 unique users we identified, 2,216,219 (84.82%) had a single block group from which

[7] The Census API returned zero values for these, so we manually downloaded the variables of "P2. URBAN AND RURAL" for each state individually from factfinder.census.gov.

[8] Block groups may be looked up by their FIPS code at http://www.policymap.com/maps

[9] Interestingly, Central Park has a nonzero population (of 25), as do some airports. Some other tourist attractions (e.g., Universal Studios) also appear.

Table 1: Block groups from which the most users have sent geotagged tweets.

FIPS code	Users	Description
32 003 006700 1	28,280	Las Vegas Strip
06 037 980028 1	23,100	Los Angeles Int'l Airport
32 003 006800 4	16,748	McCarran Int'l Airport
13 063 980000 1	15,481	Atlanta Int'l Airport
12 095 017103 2	15,392	Walt Disney World
36 081 071600 1	15,067	JFK Int'l Airport
11 001 006202 1	14,906	National Mall
36 061 014300 1	14,605	Central Park
06 059 980000 1	14,576	Disneyland
17 031 980000 1	13,610	Chicago Int'l Airport

they tweeted most frequently. The others had ties for which block group was the highest; for these users, we uniquely assigned them to one of their block groups by randomization. We tried analyses on just the 84.82% as well, but found it made little substantive difference in the results.

In the terminology of Guo and Chen (2014), the most active accounts belong to 'non-personal users.'[10] In this case, the most active tweeter (44,624 tweets) seems to be a commercial service for travel, the second-most active (35,025) is an automatic news updater in Florida, etc. Starting from the 13th most active tweeter, with 12,922 tweets, there were accounts that appeared on inspection to be personal ones. As for number of block groups traversed, the top 'traveler' (23,547 block groups) is the same as the top tweeter, and others are similarly non-personal users. Across block groups, it is not until the 18th most mobile user, traversing 1,209 block groups, that there is a personal user.

How much mobility is there between units? Figures 6 and 7 show respectively that while there is minimal mobility between states, with only 22.39% of users sending geotagged tweets from more than one state and only 7.83% send from more than 2. However, there is a great deal of mobility between (possibly neighboring) block groups, with 65.24% of users sending geotagged tweets from more than one block group.

How well does unique assignment do? As one check, we consider the ratio of geotag users to population; there are 509 block groups where this ratio is greater than 1 (for users with 5 or more tweets only, there are 353, and for users with 10 or more tweets only, there are 290), indicating either the failure of population as proxy for potential geotag users or of the method of assigning mobile users. As we found the block groups with the largest ratios to be airports, it seems to be a case of the latter.

Bivariate regression model. We first test our null hypothesis of a linear regression yielding a coefficient of 1 to the logarithm of the population. Looking at the plot of the relationship of the logarithm of the two (fig. 8), there is a faint linear relationship, although the slope does not appear to be 1. An OLS regression fits slope $\hat{\beta}_1$ = .4916 (.002996) and intercept $\hat{\beta}_0$ = -1.219 (.02143),[11] although we should recall that the standard errors are not reliable under spatial autocorrelation.

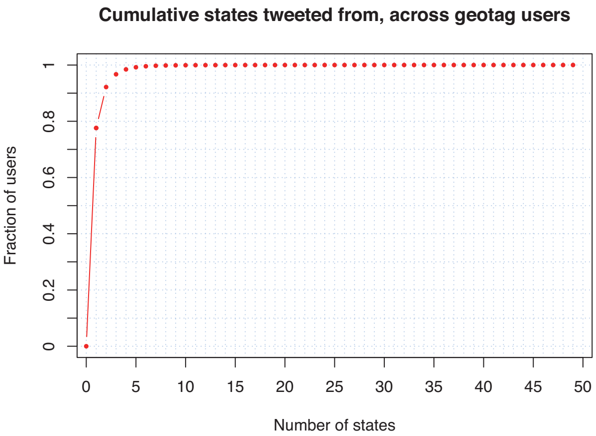

Figure 6: A full 77.61% of geotag users in our set tweeted only from one state, and having tweeted from 5 or fewer states accounts for 99.21% of users.

Compare this plot to the plot of our test case mentioned earlier, the distribution of males over the population, pictured in fig. (9). The true ratio of males to total population across the block groups we consider is .4915; according to our model, the exponential of the intercept should be this, and the coefficient of the log population term should be 1. Indeed, log(.4915) is within the 95% confidence interval (log(.4914), log(.4962)), and 1 is just outside the 95% confidence interval (.9980, .9994), but this is without accounting for how spatial autocorrelation shrinks estimated standard errors. The R^2 value of this model is also impressive at .975, although under spatial autocorrelation R^2 is inflated thereby not interpretable. Overall, our model fits the relationship of males to population exactly as we would expect it to fit to something randomly distributed over the population.

Using this as a validation of our statistical test, we can strongly reject the null hypothesis that $\hat{\beta}_1 = 1$ even without correcting for spatial autocorrelation. And the R^2 value for this regression is a paltry .109, too small to worry about being inflated. Thus, we can conclude that geotag users are not randomly distributed over the US population, and indeed that the population count is not very informative about the number of geotag users.

Weights matrix and spatial autocorrelation. Testing the residuals in our basic model for spatial autocorrelation using Moran's I against all weights matrices considered above, we

[10] They find that only 2.6% of geotag users are non-personal. This should be small enough to have no effect on results, so we did not employ filtering. However, this may be considered in a future work.

[11] Filtering for only those users who have 5 or more tweets and for those users with 10 or more tweets, the respective fitted slopes are .5192 (.002932) and .5136 (.2786).

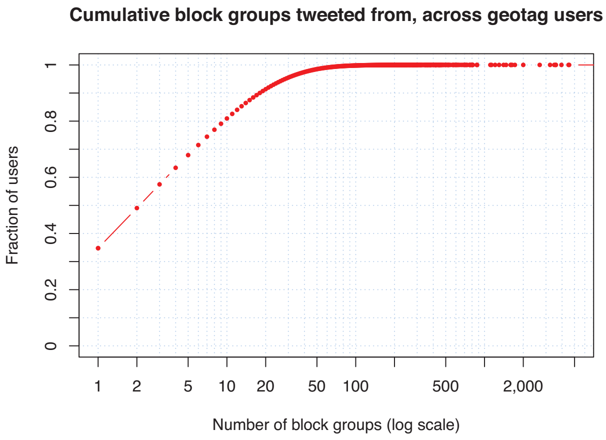

Figure 7: 34.76% of geotag users tweeted only from one block group. 27 or fewer block groups were 95%, 50 or fewer block groups were 99%. One outlier at 23,547 excluded.

find the results reported in table (2).

Table 2: Selected Values of Moran's I in residuals

	Population vs Users	Population vs Male
2nn	.3699	.2336
4nn	.3550	.2142
6nn	.3398	.1996
8nn	.3270	.1883
Rook	.4166 (b)	.2125 (b)
	.3992 (rn)	.2201 (rn)
Queen	.4151 (b)	.2097 (b)
	.3919 (rn)	.2154 (rn)

For the Rook contiguity case and the Queen contiguity case, binary (b) and row-normalized (rw) weights gave different values.

We found identical results of Moran's I for binary weights matrices and row-normalized weights matrices in the k-nearest neighbor case. For the two contiguity cases, row normalization made a difference, and we list both values. In all cases, an asymptotic test against the expected value of 0 was significant at $p < .0001$. The autocorrelation in the population-user model is stronger than in the 'null' population-male model. It appears, then, that the spatial autocorrelation is strong enough that the choice of weights matrix is not critical. For the population to user model fit on counts of users with 5 or more tweets, or 10 or more tweets, the spatial autocorrelation was similar (generally lower, but still higher than the autocorrelation of population vs. male).

Spatial errors model. The maximum likelihood method of fitting a SAR model involves computing the log determinant of the $n \times n$ matrix $|I - \lambda W|$, which is infeasible at our n of over 200,000. An alternative method finds the log determinant of a Cholesky decomposition of $(I - \lambda W)$, although this then requires W to be a symmetric matrix (Bivand, Pebesma, and Gómez-Rubio 2013). Since all of the candidate weights matrices picked up spatial autocorrelation at a significant level, we use a binary contiguity weights matrix. We tried both Rook and Queen, and they gave comparable fits, so we report only for Rook (3).

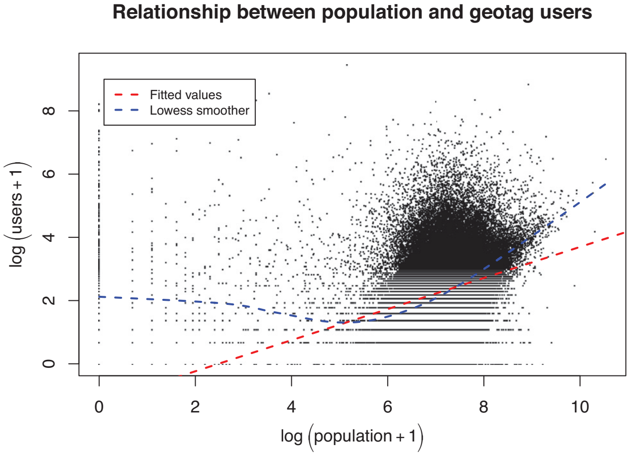

Figure 8: Eliminating zero-count observations reduces the artifacts visible at $x = 0$ and $y = 0$ but does not substantially change the fit.

Table 3: Spatial errors basic model, binary Rook contiguity

	Dependent variable:
	log(user + 1)
log(population + 1)	.4401*** (.002655)
Intercept	−1.138*** (.01890)
$\hat{\lambda}$:	.1107***
LR test value:	73,375
Numerical Hessian $\widehat{\text{se}}(\hat{\lambda})$:	8.4241e−06
Log likelihood:	−222,020.8
ML residual variance (σ^2):	.4206
Observations:	215,795
Parameters:	4
AIC:	444,050
Note:	***p<.0001

The spatial multiplier term is significant, although neither the coefficients nor the standard errors are substantively different than the previous model. However, calculating Moran's I on the residuals of this model gives a value of -.02367, with a p-value of 1, meaning we have successfully controlled for spatial autocorrelation.

We then investigate the full model specified above. We interpret this model in the standard way: for a log transformed explanatory variables X_i, a 1 percent change will predict a β_i percent change in $\mathbf{Y}$. We present the results of the regression on counts of only those users with 5 or more tweets.

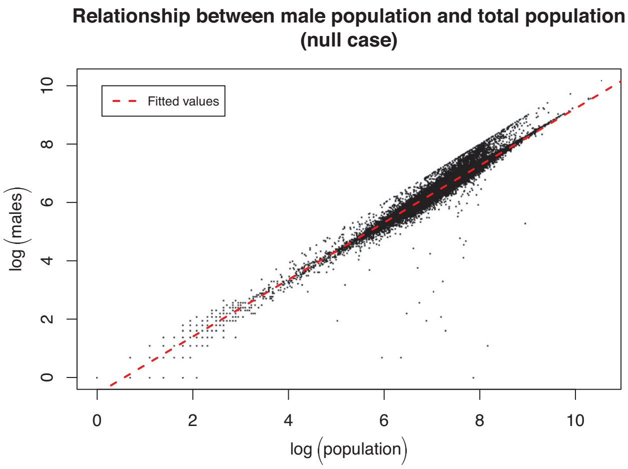

Figure 9: The relationship between males and total population behaves exactly as we expected of a quantity randomly distributed over the population, making it an effective null model against which to compare the observed distribution of geotag users.

This is shown in table (4).

As before, testing for spatial autocorrelation finds no significant amount, with a p-value of 1. Here we see that, after controlling for other factors, population loses its significance (this also points to the benefits of using a SAR model, as under OLS the population term is significant). The term for area included as a control is significant, with a one percent rise in block group area predicting a 15.56% rise in geotag users. It seems here that size overcomes the effects of population density (as mentioned above, block group population has stable variance only in log scale even though block groups are designed to enclose populations of roughly comparable size). Consistent with survey findings (Zickuhr 2013), a 1% larger Hispanic/Latino population predicts 1.533% more geotag users. However, the effect size is smaller than either that of the Asian population (a 1% rise predicting an 11.12% rise in geotag users) and, in contrast to survey findings, that of the Black population (a 1% rise predicting a 4.29% rise in geotag users). This might point to the Pew sample not including enough Twitter users, as there is an active Black community on Twitter that is gaining scholarly attention (Clark 2014; Florini 2014; Sharma 2013). The latitude, both in linear and quadratic effects, is not significant; however, the longitude is significant, pointing first to block groups further east having more geotag users, and second (from the positive sign of longitude squared) to a coastal effect where block groups on both the east and west coasts have more geotag users than in the center of the US. While we tried to test for nonlinearity in income, inclusion of a squared term for median income made the matrix computationally singular; however, inspecting the bivariate relationship did not yield any evidence for a nonlinear effect, and the linear effect is weak (a $10,000 rise in the median income predicts a 1.66% rise in the number of geotag users). Consistent with findings about urban biases (Hecht and Stephens 2014), we find that a 1% higher rural population predicts a 5.72% decrease in the number of geotag users. Lastly, also consistent with survey findings, 18-29 year olds are the most active geotag users, with a 1% higher population of this age group predicting 39.16% more geotag users. There is also a strong negative effect for the population of ages 50-64, with a one percent change predicting 17.93% fewer geotag users, but the teenage population surprisingly predicts fewer geotag users. Also surprisingly, there was a significant and positive effect from the population people 65 and older. These might be due to more complex interactions such as mixed populations. As is usual with logarithmic dependent variables, the intercept is not particularly interpretable as it would be a prediction for a block group at the center of the US with a population of 1.

Table 4: Spatial errors full model, binary Rook contiguity, users with >5 tweets only.

	Dependent variable:	
	log(user + 1)	s.e.
log(population + 1)	-.01218	(.008081)
log(area)	.1556***	(.001760)
log(asian + 1)	.1112***	(.001576)
log(black + 1)	.04292***	(.001576)
log(hispanic + 1)	.01533***	(.002066)
latitude (demeaned)	-.006992	(.0007052)
longitude (demeaned)	.02306***	(.0002739)
latitude^2	-.0001641	(.00009505)
longitude^2	.00008777***	(.00001411)
median income ($10K)	.01661***	(0006857)
log(rural + 1)	-.05722***	(.001096)
log(ages 10-17 + 1)	-.09831***	(.003712)
log(ages 18-29 + 1)	.3916***	(.004423)
log(ages 30-49 + 1)	.06362***	(.006731)
log(ages 50-64 + 1)	-.1793***	(.006953)
log(ages 65 and up + 1)	.09675***	(.003940)
Intercept	1.3382***	(.1916)
$\hat{\lambda}$:	.1009***	
LR test value:	36,577	
Num. Hessian $\widehat{\mathsf{se}}(\hat{\lambda})$:	0.0003456	
Log likelihood:	−207,923.5	
ML resid. var. (σ^2):	.3755	
Observations:	215,795	
Parameters:	19	
AIC:	415,890	
Note:	***p<.0001	

Running the SAR model using all users, instead of just those with 5 or more tweets, produces similar results, except that log population is significant with coefficient -.04196 (.007858); this suggests a nonlinear effect, and indeed, an added squared term for the log population came out as significant and positive at .06329 (.0008394). This points to some noise for those people who only 'try out' geotagged tweets but do not adopt their use that disappears if we maintain a minimum tweet threshold. When running the model on only those users with 10 or more tweets, results are again similar except the longitude squared term is no longer significant ($p = 0.1870$), and the latitude term becomes significant ($p = 0.02017$). This might be from the coasts having more users who try out geotagged tweets for a longer period of time before choosing not to continue. These subtle differences point to opportunities for modeling the demographics of different types of users (as determined by number of geotagged tweets or other factors), although we do not explore them more here.

Conclusion

Geotag users are not representative of the US population. Despite the volume of geotagged tweets and their impressive coverage (there were only 67 block groups out of 215,795 with no geotagged tweets), the users who send geotagged tweets are nonrandomly distributed over the population in subtle ways. These include predicable and already established biases towards younger users, users of higher income, and users in urbanized areas, as well as surprising biases towards Hispanic/Latino users and Black users that, in the latter case, have not seen in large-scale survey research. We also demonstrate an unsurprising but previously unreported coastal effect, where being located on the east or west coast of the US predicts more geotag users. Geotag users may not be a random sample of the population of any given block group, but given the fine level of detail and large-scale demographic variability, the demographics of a block group is a reasonable proxy for the demographics of geotag users located in that block group. Certainly, even with complications of uniquely assigning mobile users, it is enough to establish the nonrandom distribution of geotag users, and some candidate biases.

While from this study, we are unable to say whether or not geotag users are representative of the *Twitter* population, the more interesting question we address is whether geotagged tweets can be a useful proxy for the *general* population within the US. This is a critical question because geotagged Tweets are an enormously popular source of data for studying a wide variety of social and human phenomena. For future work, we emphasize that findings using geotagged tweets should not be assumed to generalize, and conclusions should be restricted only to geotag users with their population biases.

Future Work There are a number of directions for future work. One is to connect tweets to lower-resolution and lower-accuracy but more current 2013 ACS 1-year county-level estimates. Others are to see the effect of filtering out non-personal users, and to build ways to filter out foreign tourists and better uniquely place geotag users in the block group that is likely to be their residence. Modeling demographic differences between users of different levels of use is also possible with this data. We have applied one spatial model, but spatial modeling is a rich area with many other available techniques. For example, there are also relevant disease mapping models that break down incidence by various demographic strata (Bivand, Pebesma, and Gómez-Rubio 2013) that would be appropriate here, as well as non-parametric models that might better capture irregular effects. Furthermore, we elected to not consider the temporal aspect; there is work on spatio-temporal modeling (Longley, Adnan, and Lansley 2015; Sylvester et al. 2014; Nagar et al. 2014; Kamath et al. 2013) but it tends to be in the short-term window of a day or week. With reliable spatio-temporal models of how the prevalence of geotagged tweets per block group changes over longer periods of time and a better understanding of the demographic characteristics towards which geotag users are biased, we may be able to create models to provide a rapid and high-resolution proxy for demographic changes such as processes of gentrification, or urbanization, or urban decay; that is, utilize the very biases of social media data to make inferences about larger phenomena.

Funding

This material is based upon work supported by LexisNexis and HPCC Systems as well as the Army Research Office under Award No. W911NF1410481. Momin is supported by an award from the ARCS Foundation.

References

American Community Survey Office. 2014. ACS summary file technical documentation: 2013 ACS 1-year, 2011-2013 ACS 3-year, and 2009-2013 ACS 5-year data releases. Technical report, United States Census Bureau.

Anselin, L., and Rey, S. 1991. Properties of tests for spatial dependence in linear regression models. *Geographical Analysis* 23(2):112–131.

Anselin, L.; Sridharan, S.; and Gholston, S. 2007. Using exploratory spatial data analysis to leverage social indicator databases: The discovery of interesting patterns. *Social Indicators Research* 82(2):287–309.

Anselin, L. 2002. Under the hood: Issues in the specification and interpretation of spatial regression models. *Agricultural Economics* 27(3):247–267.

Bivand, R., and Piras, G. 2015. Comparing implementations of estimation methods for spatial econometrics. *Journal of Statistical Software* 63(18):1–36.

Bivand, R.; Hauke, J.; and Kossowski, T. 2013. Computing the Jacobian in Gaussian spatial autoregressive models: An illustrated comparison of available methods. *Geographical Analysis* 45(2):150–179.

Bivand, R. S.; Pebesma, E.; and Gómez-Rubio, V. 2013. *Applied spatial data analysis with R, Second edition*. Springer, NY.

Cho, E.; Myers, S. A.; and Leskovec, J. 2011. Friendship and mobility: User movement in location-based social networks. KDD '11, 1082–1090.

Clark, M. D. 2014. *To tweet our own cause: A mixed-methods study of the online phenomenon "Black Twitter"*. Ph.D. Dissertation, The University of North Carolina at Chapel Hill, School of Journalism and Mass Communication.

Crampton, J. W.; Graham, M.; Poorthuis, A.; Shelton, T.; Stephens, M.; Wilson, M. W.; and Zook, M. 2013. Beyond the geotag: Situating "big data" and leveraging the potential of the geoweb. *Cartography and Geographic Information Science* 40(2):130–139.

Doran, D.; Gokhale, S.; and Dagnino, A. 2013. Human sensing for smart cities. ASONAM '13, 1323–1330.

Eisenstein, J.; O'Connor, B.; Smith, N. A.; and Xing, E. P. 2010. A latent variable model for geographic lexical variation. EMNLP '10, 1277–1287.

Florini, S. 2014. Tweets, tweeps, and signifyin': Communication and cultural performance on "Black Twitter". *Television & New Media* 15(3):223–237.

Frias-Martinez, V.; Soto, V.; Hohwald, H.; and Frias-Martinez, E. 2012. Characterizing urban landscapes using geolocated tweets. PASSAT/SocialCom '12, 239–248.

Gaetan, C., and Guyon, X. 2012. *Spatial Statistics and Modeling*. Springer Series in Statistics.

Ghosh, D. D., and Guha, R. 2013. What are we tweeting about obesity?: Mapping tweets with topic modeling and geographic information system. *Cartography and Geographic Information Science* 40(2):90–102.

Graham, M.; Hale, S. A.; and Gaffney, D. 2014. Where in the world are you?: Geolocation and language identification in Twitter. *The Professional Geographer* 66(4):568–578.

Guo, D., and Chen, C. 2014. Detecting non-personal and spam users on geo-tagged Twitter network. *Transactions in GIS* 18(3):370–384.

Hecht, B., and Stephens, M. 2014. A tale of cities: Urban biases in volunteered geographic information. ICWSM '14, 197–205.

Hecht, B.; Hong, L.; Suh, B.; and Chi, E. H. 2011. Tweets from Justin Bieber's heart: The dynamics of the location field in user profiles. CHI '11, 237–246.

Hong, L.; Ahmed, A.; Gurumurthy, S.; Smola, A. J.; and Tsioutsiouliklis, K. 2012. Discovering geographical topics in the Twitter stream. WWW '12, 769–778.

Kamath, K. Y.; Caverlee, J.; Lee, K.; and Cheng, Z. 2013. Spatio-temporal dynamics of online memes: A study of geo-tagged tweets. WWW '13, 667–678.

Kinsella, S.; Murdock, V.; and O'Hare, N. 2011. "I'm eating a sandwich in Glasgow": Modeling locations with tweets. SMUC '11, 61–68.

Kumar, S.; Hu, X.; and Liu, H. 2014. A behavior analytics approach to identifying tweets from crisis regions. HT '14, 255–260.

Leetaru, K.; Wang, S.; Cao, G.; Padmanabhan, A.; and Shook, E. 2013. Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday* 18(5).

Lin, Y.-R., and Margolin, D. 2014. The ripple of fear, sympathy and solidarity during the Boston bombings. *EPJ Data Science* 3(1).

Liu, Y.; Kliman-Silver, C.; and Mislove, A. 2014. The tweets they are a-changin: Evolution of Twitter users and behavior. ICWSM '14.

Longley, P. A.; Adnan, M.; and Lansley, G. 2015. The geotemporal demographics of Twitter usage. *Environment and Planning A* 47(2):465–484.

Mislove, A.; Lehmann, S.; Ahn, Y.-Y.; Onnela, J.-P.; and Rosenquist, J. 2011. Understanding the demographics of Twitter users. ICWSM '11, 554–557.

Mitchell, L.; Frank, M. R.; Harris, K. D.; Dodds, P. S.; and Danforth, C. M. 2013. The geography of happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place. *PLoS ONE* 8(5):e64417.

Morstatter, F.; Pfeffer, J.; Liu, H.; and Carley, K. 2013. Is the sample good enough?: Comparing data from Twitter's streaming API with Twitter's firehose. ICWSM '13.

Morstatter, F.; Lubold, N.; Pon-Barry, H.; Pfeffer, J.; and Liu, H. 2014. Finding eyewitness tweets during crises. ACL LACSS '14, 23–27.

Morstatter, F.; Pfeffer, J.; and Liu, H. 2014. When is it biased?: Assessing the representativeness of Twitter's streaming API. WWW Companion '14, 555–556.

Nagar, R.; Yuan, Q.; Freifeld, C. C.; Santillana, M.; Nojima, A.; Chunara, R.; and Brownstein, S. J. 2014. A case study of the New York City 2012-2013 influenza season with daily geocoded Twitter data from temporal and spatiotemporal perspectives. *J Med Internet Res* 16(10):e236.

Ruths, D., and Pfeffer, J. 2014. Social media for large studies of behavior. *Science* 346(6213):1063–1064.

Sharma, S. 2013. Black Twitter?: Racial hashtags, networks and contagion. *New Formations: A Journal of Culture/Theory/Politics* 78(1).

Shelton, T.; Poorthuis, A.; Graham, M.; and Zook, M. 2014. Mapping the data shadows of Hurricane Sandy: Uncovering the sociospatial dimensions of 'big data'. *Geoforum* 52(0):167 – 179.

Stephens, M., and Poorthuis, A. 2014. Follow thy neighbor: Connecting the social and the spatial networks on Twitter. *Computers, Environment and Urban Systems*.

Sylvester, J.; Healey, J.; Wang, C.; and Rand, W. M. 2014. Space, time, and hurricanes: Investigating the spatiotemporal relationship among social media use, donations, and disasters. Technical Report Research Paper No. RHS 2441314, Robert H. Smith School.

Takhteyev, Y.; Gruzd, A.; and Wellman, B. 2012. Geography of Twitter networks. *Social Networks* 34(1):73–81.

Tufekci, Z. 2014. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. ICWSM '14, 505–514.

van Dijck, J. 2013. Chapter 4: Twitter and the paradox of following and trending. In *The Culture of Connectivity: A Critical History of Social Media*, 68–88. Oxford University Press.

van Liere, D. 2010. How far does a tweet travel?: Information brokers in the Twitterverse. MSM '10, 6:1–6:4.

Wang, D.; Al-Rubaie, A.; Davies, J.; and Clarke, S. 2014a. Real time road traffic monitoring alert based on incremental learning from tweets. EALS '14, 50–57.

Wang, X.; Gaugel, T.; ; and Keller, M. 2014b. On spatial measures for geotagged social media contents. MUSE '14, 35–50.

Yuan, Q.; Cong, G.; Ma, Z.; Sun, A.; and Thalmann, N. M. 2013. Who, where, when and what: Discover spatio-temporal topics for Twitter users. KDD '13, 605–613.

Zickuhr, K. 2013. Location-base services. Technical Report Pew Internet and American Life Project, Pew Research Center.