JOURNAL OF SPATIAL INFORMATION SCIENCE
Number 19 (2019), pp. 105–129
doi:10.5311/JOSIS.2019.19.510

RESEARCH ARTICLE

Exploring the effectiveness of geomasking techniques for protecting the geoprivacy of Twitter users

Song Gao, Jinmeng Rao, Xinyi Liu, Yuhao Kang, Qunying Huang, Joseph App

Department of Geography, University of Wisconsin–Madison, USA

Received: March 10, 2019; returned: June 23, 2019; revised: November 4, 2019; accepted: November 13, 2019.

Abstract: With the ubiquitous use of location-based services, large-scale individual-level location data has been widely collected through location-awareness devices. Geoprivacy concerns arise on the issues of user identity de-anonymization and location exposure. In this work, we investigate the effectiveness of geomasking techniques for protecting the geoprivacy of active Twitter users who frequently share geotagged tweets in their home and work locations. By analyzing over 38,000 geotagged tweets of 93 active Twitter users in three U.S. cities, the two-dimensional Gaussian masking technique with proper standard deviation settings is found to be more effective to protect user's location privacy while sacrificing geospatial analytical resolution than the random perturbation masking method and the aggregation on traffic analysis zones. Furthermore, a three-dimensional theoretical framework considering privacy, analytics, and uncertainty factors simultaneously is proposed to assess geomasking techniques. Our research offers insights into geoprivacy concerns of social media users' georeferenced data sharing for future development of location-based applications and services.

Keywords: privacy, geoprivacy, geomasking, social media, digital footprints, uncertainty

1 Introduction

The availability of location-based services has made the collection of large-scale individual-level location data through the use of mobile phones, GPS devices, and geotagged social

media commonplace [34, 58]. While such location-based big data provides new opportunities to study human mobility patterns and transportation models [6, 13, 21, 32], complex human-environment interactions [18,26,34,39,49], socioeconomic characteristics [22,31,35], urban spatial structure and changes [20, 59], disaster responses [23, 54, 55], and location business intelligence [48], it also introduces challenges regarding the protection of location privacy [51]. Furthermore, there are increasing concerns about the social, ethical, legal, and behavioral implications of geoprivacy caused by user identity de-anonymization and location exposures [5, 27, 50].

Figure 1: The spatial distribution of geotagged tweets around a Twitter user's home.

Generally speaking, geoprivacy refers to an individual's rights to prevent the disclosure of personal sensitive locations including but not limit to their home, workplace, daily activity places, and travel trips [30]. However, the majority of people are unaware of how the underlying location-related technologies work and what can be inferred from an individual's location records that are collected when people use various location-based services [27] . Figure 1 shows the spatial distribution of geotagged tweets around a Twitter user's home. Obviously, the home location of this individual can be easily identified through his/her digital footprint on social media with high confidence [21]. As a result, researchers have developed a number of statistical approaches and technical solutions aimed to protect individuals from being identified through their location records. A common practice for preserving data confidentiality is aggregation such that detailed individual records are merged into anonymized large-group characteristics. For example, aggregating individual home location into geographic or administrative units. Aggregating raw address points into such identical polygons makes the inference of original locations hard and user privacy becomes a k-anonymity problem [8, 14, 42]. There exist several location obfuscation approaches for

achieving k-anonymity [8,15,28]. However, aggregation may reduce the spatial resolution of analysis that can be conducted and reduce the effectiveness of the analysis [30]. Another family of approaches is called geomasking in which the original location may be hidden or modified for geoprivacy protection but the spatial point patterns are not significantly affected.

There is a rich history of literature on leveraging geographical masking to preserve the confidentiality of health records and trajectory data. With child leukemia lymphoma data from North Humberside, England, Armstrong et al. [4] described and evaluated several types of geographical masks to protect personal privacy as well as to allow the conduct of valid spatial analyses. Kwan [30] examined the effects of random perturbation masks on the results of a spatial analysis using data on lung-cancer deaths. Three different random perturbation masks were implemented with each at three different levels of introduced error. Hampton et al. [17] extended existing methods of random perturbation by developing an adaptive geomasking technique known as the donut method. This method guarantees that each geocoded address is not randomly assigned on or too near its original location. Compared with random perturbation method, the performance of k-anonymity using the proposed donut method was at least 42.7% higher in geoprivacy measures and was less than 4.8% in cluster detection measures. Seidl et al. [46] examined the grid masking and random perturbation techniques for anonymizing the GPS trajectory data and tested the preservation of both privacy and spatial patterns. They found that as the distance thresholds for grid masking and random perturbation increase, the correlation between density patterns decreases.

However, the use of geographical masking methods to prevent the disclosure of sensitive locations of social media users is still not well addressed. Location-based social media data is different from other existing data sources (e.g., health survey and GPS trajectories) due to its innate characteristics such as data sparsity and sampling bias, spatiotemporal distribution heterogeneity, and location representativeness and uncertainty [21,33]. To this end, we aim to investigate the effectiveness of geomasking techniques for protecting the geoprivacy of active Twitter users who frequently share geotagged tweets in their home or work location. To the best of our knowledge, this work is a first attempt in this direction using individual-level location-based social media data. Additionally, a theoretical framework considering privacy, analytics, and uncertainty factors simultaneously is proposed to assess different geomasking techniques.

2 Related work

Geomasking has been used in public health and spatial analysis for decades in order to protect sensitive information. Much of the literature on geomasking has been done on data with a fairly coarse spatial and temporal resolution. Twitter data, on the other hand, are frequent and may occur in a relatively small geographic area. In order to inform ourselves on the nature of obfuscating a varying density of geospatial data, we need to investigate novel and traditional geomasking techniques.

Voronoi masking relies on the creation of Voronoi polygons around individual point features, and then, for those points to be relocated to the nearest edge of its bounding polygon. This method is shown to be robust with lower resolution spatial data, about 23 persons/km^2 of population density [47]. It is also effective in reducing the likelihood

of false identification of true household location because the points are often relocated to boundaries of parcels. Since Voronoi masking is not randomly generated and dependent on the spatial structure of points, it may preserve the original locations, however, as polygons. Given the nature of geotagged tweets, it would not benefit user privacy to create hundreds of polygons which still lie on or near the location of concern, whether it be home or work. This issue is inherent in high resolution spatial data. It may be beneficial to repeat the Voronoi masking process a second time. The nearest edge to a polygon centroid may be the nearest edge for more than one centroid and therefore it is possible that the number of unique locations will be reduced after the initial masking. This process would lower the resolution of the dataset and possibly reduce the true location detection accuracy after two or more iterations.

On the other hand, Seidl et al. [47] show that grid masking is not an effective method for preserving spatial analysis at the aforementioned low resolution. In this case, the assignment of points along a uniform grid amounts to aggregation over the area of the grid. This may be a beneficial method at high resolutions as we are able to set the size of the grid to a much smaller area and in essence create our own minor aggregation units without displacing the points nearly as far [47].

A multiscale geomasking technique by which locations are converted to Military Grid Reference System (MGRS) coordinates provides a unique amount of control over the adjusted locations [7]. MGRS eastings and northings provide 5 levels at which to mask data in increments of powers of ten from 1, 10, 100, 1000, and 10000 meters. Points are displaced along axes from the original point along the grid system. Tests show the method is invertible and, after Level 3, loses almost all overlap between masked and unmasked points indicative of personal location information. These tests were conducted on 2,000 randomly generated points in GIS software. This method also resembles grid masking such that the displacement of points is done along the eastings and northings from the origin [7]. The ability to control random perturbations along MGRS easting and northings is a form of high resolution grid masking that is worthwhile to compare with traditional grid masking.

A further consideration for the preservation of spatial characteristics as well as privacy, is topology. Given a set of parcels or an easily obtained base map such as OpenStreetMap (OSM), we can ground truth residential, work, or school locations based on spatiotemporal tweet patterns. Relocating points just outside of parcels or to a road center line, was shown via survey to introduce more uncertainty among participants as to actual location points [45]. Those points displaced within a parcel or along a parcel boundary induced less uncertainty. In addition to cluster detection, the reduction in map-user confidence is a unique measure for determining the effectiveness of geomasking. This method may be useful in distorting user perceptions of point clusters and reduce the likelihood of inferring a Twitter user's home or place of employment [45]. Geoprivacy is not limited to the users' geometric coordinate information [40]. The user-generated social media content includes rich semantic signatures (i.e., spatial, temporal, and thematic patterns) [1, 24, 38, 39, 62], which may also reveal distinct place-based patterns and cause potential privacy risks. McKenzie et al. [40] illustrate how protecting place-based information differs from a purely spatial perspective using location-based social networking check-in data.

In the statistics and computer science communities, the trade-off between utility and the level of differential privacy guaranteed by a processing mechanism has been considered in several privacy-preserving learning approaches such as private support vector machine (SVM) learning [44] and private Bayesian inference [61]. The key concern in the study

of differential privacy is whether the published aggregation information from a statistical database would disclose private individual information. Regarding location-based systems or services (LBS), a mechanism to draw random noise to the user's location from a planar Laplace distribution has been proposed to guarantee geo-indistinguishability [3]. In [19], a differential private pattern mining algorithm was proposed for geographic location discovery using a combination of region quadtree spatial decomposition and a density-based clustering algorithm. The experiments were conducted using synthetic datasets and showed the feasibility of their proposed algorithm to achieve the differential privacy goal. In addition, privacy-preservation can be achieved through the process of obfuscation with degrading the quality of information about a person's location using spatial and temporal cloaking [9, 15]. A geographic graph model of obfuscation for protecting an individual's location privacy in LBS was demonstrated in [8]. Moreover, a comprehensive survey of computational location privacy for broader implications was conducted in [28].

3 Methods

In this research, we are concerned with individual user coordinate information and survey the effectiveness of three popular geomasking techniques: *Aggregation*, *Random Perturbation*, and *Gaussian Perturbation*, for the preservation of Twitter users' location privacy [4]. One open-source geomasking implementation in R can be accessed via the GitHub repository[1].

Aggregation: merges individual geotagged tweet points into polygons into which those points fall. Different types of administrative boundaries such as census blocks, tracts, and traffic analysis zones or vague cognitive regions (e.g., downtown) could be candidate polygons [12]. And the centroids of those spatially overlaying polygons are used as the coordinates of those tweets.

Random Perturbation: is a geomasking approach in which each point is displaced in space by a randomly determined distance and direction [4, 30]. A distance threshold is typically added to set the allowed maximum displacement distance in the case of uniform geomasking. As shown in Figure 2, the original posted locations of the geotagged tweets of a Twitter user are randomly displaced within a 1km distance radius.

Gaussian Perturbation: uses a two-dimensional isotropic (i.e., circularly symmetric) Gaussian kernel to control the random displacement process of a point set such that the distribution of those displaced points follows a two-dimensional Gaussian ("bell-shaped") form [11, 60]:

$$G(x, y) = (1/2\pi\delta^2)e^{-(x^2+y^2)/2\delta^2} \tag{1}$$

$$\delta = \sqrt{\frac{\sum_i^n (x_i - \bar{x})^2}{n} + \frac{\sum_i^n (y_i - \bar{y})^2}{n}} \tag{2}$$

Where (x, y) is the 2D coordinates of each location after displacement, and δ specifies the standard deviation (SD) of the positional error, ($\bar{x}$,$\bar{y}$) is the mean center of a point set, and n is the total number of points. As shown in Figure 2, with the increment of the standard deviation, the displacement of those points spreads more widely. The derived spatial

[1] https://github.com/claudiofronterre/geomask

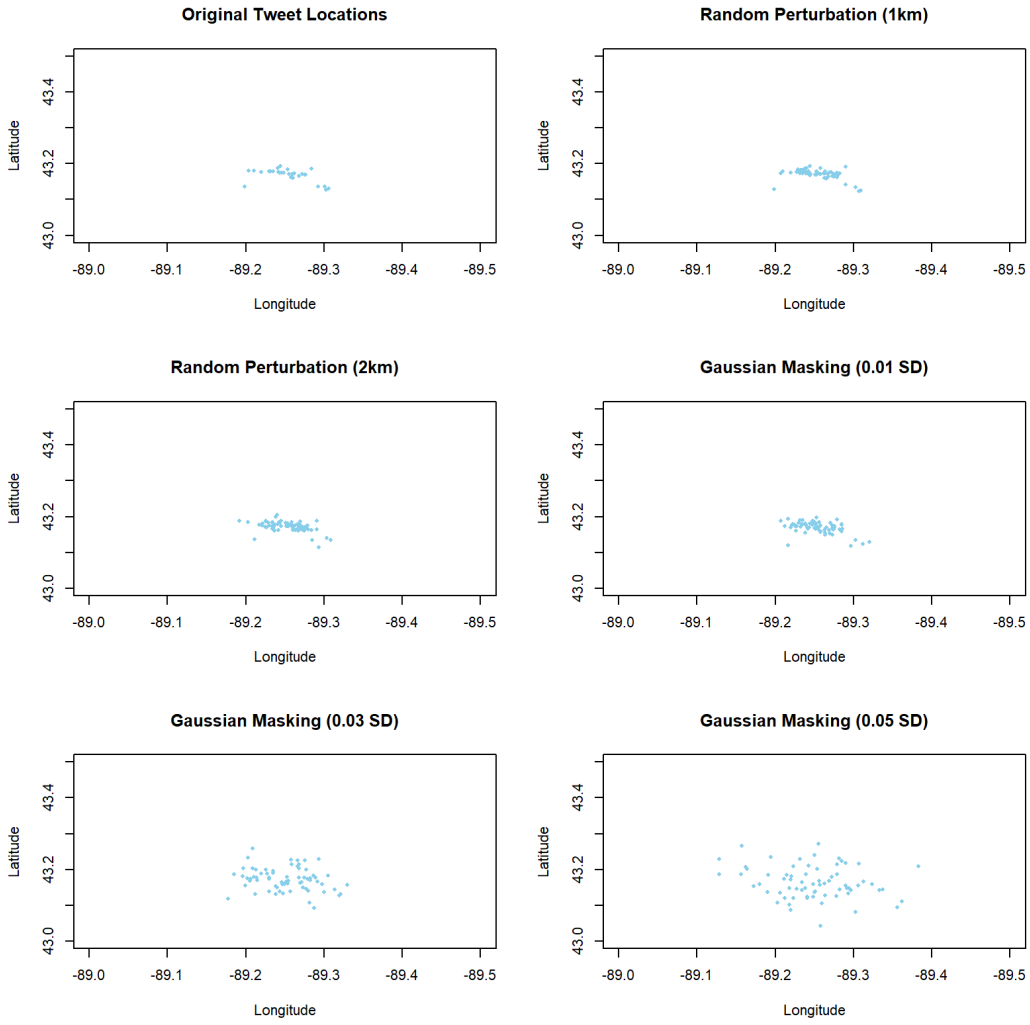


Figure 2: The Gaussian geomasking with different standard deviations (SD) and the random perturbation with 1km and 2km threshold of a user's geotagged tweets.

point patterns with a large standard deviation may not capture the original spatial density distribution of an individual's digital footprints.

After the perturbation processing of the original locations, we need to further determine whether users' home or work location (two of the most sensitive places for an individual's geoprivacy) can still be identified through state-of-the-art location detection algorithms. Specifically, we explored different parameter calibrations for the density-based spatial clustering with noise (DBSCAN) [10,21,35] that has been widely used in spatial clustering and the identification of significant human activity places. The DBSCAN algorithm requires

two parameters: the searching radius of a cluster (Eps) and the minimum number of points (MinPts) within a cluster. The different combinations of Eps and MinPts values may get different spatial clustering results [20, 37]. In the case of detecting Twitter users' home or work location, the parameter calibration may generate different candidate clusters or distance shifts from the actual location. Therefore, we have explored different scenarios with varying parameter values for the perturbation and the spatial clustering steps. In each operation of the perturbation and the clustering, two representative centers (i.e., centroid and medoid) are calculated for further calculation of shift distance from the true home or work location. The centroid is the weighted sum of geotagged tweets' coordinates in a cluster and it might not be one of the original locations, while the medoid can be defined as the point of a cluster whose average distance to all the objects in the cluster is minimal [52, 53].

Evaluation Measures: In addition to the shift distance to ground-truth locations, the following quality measures are also used in this study to evaluate the effectiveness of different geomasking approaches. Those measures are defined in terms of the following four cases: true positive (TP), true negative (TN), false positive (FP), and false negative (FN) [2, 29, 43]. TP is the number of points correctly identified as home or work locations after geomasking and cluster analysis in each period (daytime or nighttime). TN is the number of points correctly identified as not-home or not-work locations. FP is the number of points incorrectly identified as home or work locations. FN is the number of points incorrectly identified as non-home or non-work locations.

- Overall accuracy = (TN+TP)/(TN+TP+FN+FP) is the ratio between the number of correctly identified home (work) and not-home (not-work) cluster points (including both TP and TN cases) and the total number of points in each period. It serves as a general accuracy measure. Usually, the higher the value is, the worse the associated geomasking is in protecting geoprivacy.
- Precision = TP/(TP+FP) is the ratio between the number of correctly identified home or work cluster points and the total number of locations that are identified as home or work locations (all positive predictions). A high precision shows that, among all the positive predictions, the method gets more home or work cluster points that are correctly identified than the home or work cluster points that are incorrectly identified.
- Sensitivity = TP/(TP+FN) (also known as *Recall*) is the ratio between the number of correctly identified home or work cluster points and the total number of true home or work locations. A high recall shows that the results discover a larger fraction of the home or work cluster points.
- Specificity = TN/(TN+FP) is the ratio between the number of correctly identified non-home or non-work cluster points and the total number of true non-home or non-work locations. It shows how good a method is for detecting a user's non-home or non-work location after geomasking.
- Balanced accuracy = (Sensitivity+Specificity)/2 is a measure that combines *sensitivity* and *specificity*. It considers the imbalance of a dataset and shows a balanced performance on how accurate a method is. If the data is imbalanced, then the balanced accuracy is suggested to be used as an accuracy measure.
- F1-score = 2*Precision*Recall/(Precision+Recall) is a measure that combines *precision* and *recall*. It shows a balanced performance on how effective a method is for detecting a user's home or work cluster location after geomasking.

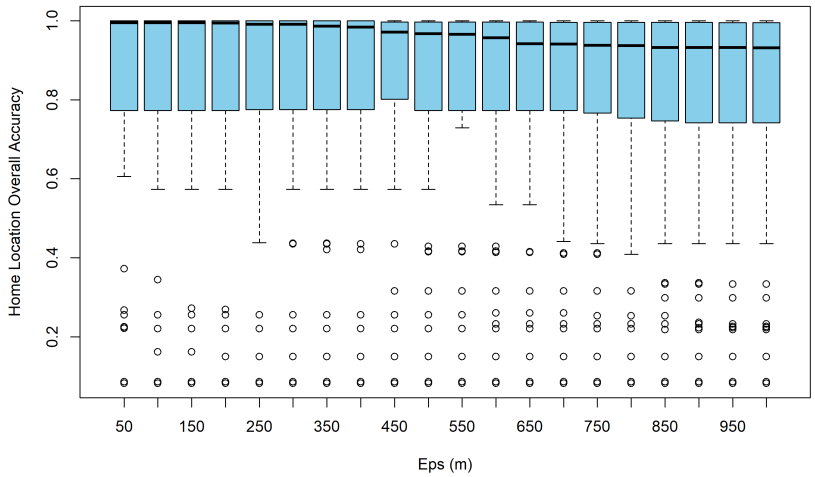


Figure 3: The boxplot of overall accuracy changes of home cluster detection with different DBSCAN parameters (without geomasking).

4 Experiments and results

We selected 93 active Twitter users who have frequently posted geotagged tweets in three U.S. urban areas: two metropolitan areas (Washington DC and Los Angeles) and one smaller urban area (the City of Madison, Wisconsin). Over 38,000 geotagged tweets were collected only from users' mobile phone devices such that their location information is most accurate for human mobility studies [13, 22, 33]. We selected these three areas (two large cities located on the east coast and the west coast respectively, and one city located in the Midwest) as representations of the U.S. urban areas. Another reason was our familiarity of the geographic backgrounds of the three cities, which helped the location ground-truth labeling and validation process. Up to 3,200 tweets can be fetched for each individual Twitter user due to the API access limit. The anchor points (i.e., the location of home and work) [16, 41, 56, 57] are two most important locations for an individual and are chosen as the target place type for geoprivacy protection. We manually identified their home and work locations as the ground-truth by overlaying their nighttime (8pm-7am) and daytime (9am-5pm) geotagged tweets onto the high-resolution (about 2m-4m) Digital Globe aerial images and the OpenStreetMap (OSM) points of interest layer. Another important rule for the ground-truth labeling is to check whether the same location cluster persists across multiple days. Among those users, 70 users' home location and 60 users' work location can be manually identified.

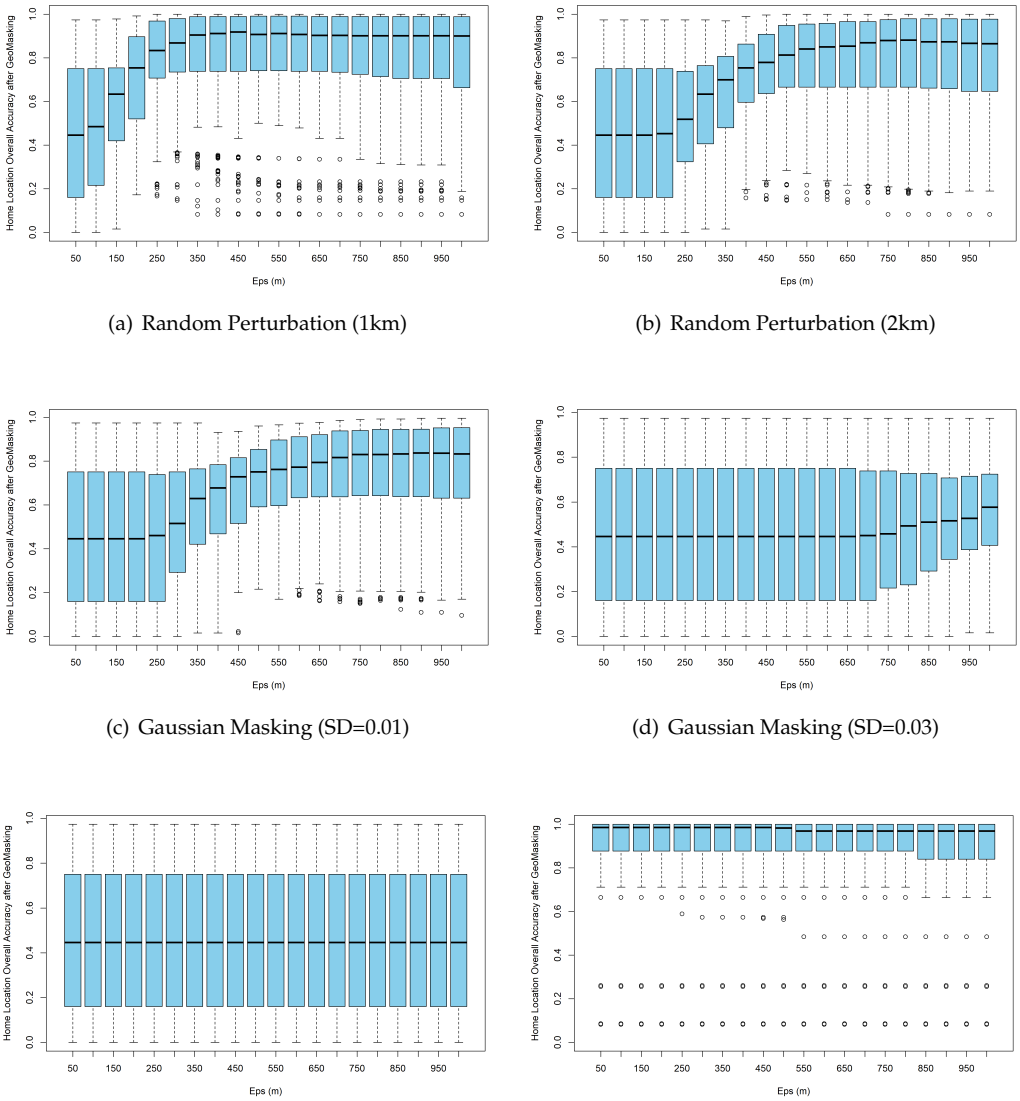
The impact of Eps and MinPts: Before applying the geomasks, we first tested how the choice of *MinPts* and *Eps* in DBSCAN would impact the effectiveness of identifying the home clusters of those Twitter users. We chose the *MinPts* ranging from 4 points to the square root of the total number of tweets in each period (nighttime or daytime), and the search radius *Eps* in a range of 50m to 1000m with a step of 50m. As shown in Figure 3,

Measures / Methods	RM (H)	GM (H)	TAZ (H)	RM (W)	GM (W)	TAZ (W)
Mean overall accuracy	0.742	0.459	0.867	0.661	0.729	0.880
Median overall accuracy	0.840	0.446	0.980	0.722	0.803	0.989
Mean balanced accuracy	0.686	0.500	0.852	0.551	0.500	0.917
Median balanced accuracy	0.667	0.500	0.980	0.500	0.500	0.991
Mean sensitivity (recall)	0.978	0.000	0.966	0.988	0.000	0.970
Median sensitivity (recall)	1.000	0.000	1.000	1.000	0.000	1.000
Mean specificity	0.704	1.000	0.898	0.705	1.000	0.865
Median specificity	0.819	1.000	0.973	0.800	1.000	0.982
Mean precision	0.789	N/A	0.951	0.627	N/A	0.849
Median precision	0.936	N/A	0.990	0.837	N/A	0.983
Mean F1-score	0.836	N/A	0.949	0.704	N/A	0.858
Median F1-score	0.954	N/A	0.991	0.898	N/A	0.987
Median shift to the medoid	42m	N/A	403m	90m	N/A	360m
Median shift to the centroid	737m	N/A	485m	798m	N/A	460m

Table 1: The geoprivacy effectiveness measures using different geomasking methods (Random perturbation with 1km threshold and Gaussian perturbation with 0.05 SD; H: Home, W: Work, GM: Gaussian Masking, RM: Random Masking, TAZ: Aggregation by traffic analysis zones, and N/A means results are not available).

we grouped the home cluster detection results based on the *Eps*, and each sub-boxplot represents the overall accuracy with varying *MinPts* in DBSCAN. Not surprisingly, the mean and median of overall accuracy is over 0.836 and 0.970 and keeps high values (basically over 0.8) regardless of the parameter choices. It also indicates the potential risk of location exposures of those active users as their home location cluster can be easily identified even without parameter calibration.

Comparing the effectiveness of different geomasking techniques: First, we explored the impact of the random perturbation geomasking with different thresholds. Existing studies have found that the choice of Eps=200m to 300m could generate good spatial clustering results for urban areas of interest and human activity zones [20, 22, 33]. Therefore, we are interested in whether the geomasking process could protect users' home or work location privacy within such a distance range. However, our experiments show that small-distance (such as within 300m or even 1km) random perturbations don't help the protection of users' geoprivacy because their home location clusters can still be correctly identified with over 0.80 overall accuracy. Moreover, the mean of sensitivity (recall) for detecting home clusters is 0.978 and the median is even higher to 1.0; the mean of precision for detecting home clusters is 0.789 and the median is 0.936; and the mean of F1-score is 0.836 and the median is 0.954. All these quality measures show that the users' home locations are exposed to the general public after random perturbation masking with a 1km distance threshold. Even when the displacement threshold reaches 2km, the mean of overall accuracy using the ran-



(a) Random Perturbation (1km)

(b) Random Perturbation (2km)

(c) Gaussian Masking (SD=0.01)

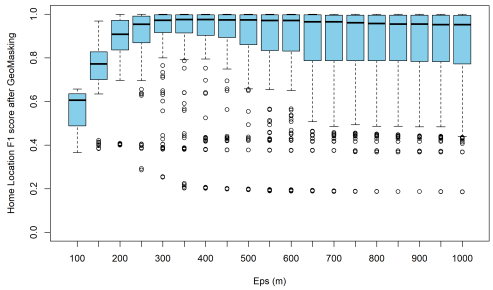
(d) Gaussian Masking (SD=0.03)

(e) Gaussian Masking (SD=0.05)

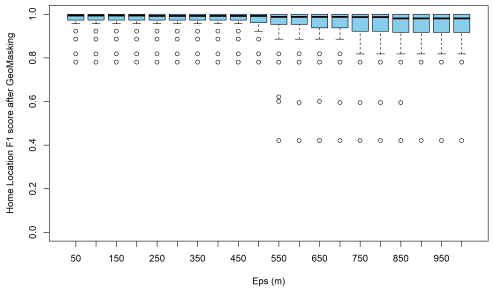
(f) TAZ-based aggregation (Madison)

Figure 4: The boxplot of overall accuracy of home cluster detection with different DBSCAN parameters with random perturbation, Gaussian masking, and the TAZ-based aggregation.

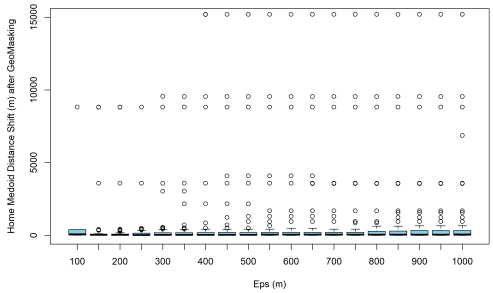
dom perturbation mask is still over 0.70. The 2km random perturbation mask is effective for protecting users' home locations within the search radius of 200m, and only less than 0.5 of overall accuracy can be achieved for identifying the users' home locations (in Figure 4). As for the displacement distance, as shown in Table 1, the median of shift distances from the true home location to the centroid of home clustering results is about 737m and to



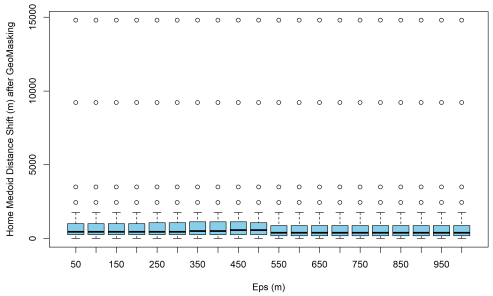
(a) F1-score by Random Perturbation (1km)



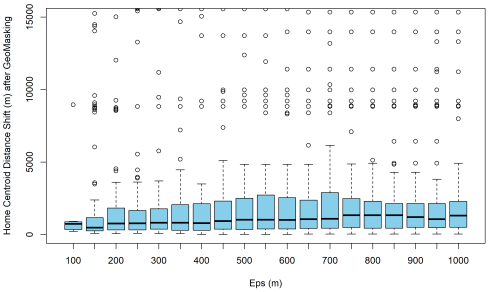
(b) F1-score by TAZ-based Aggregation



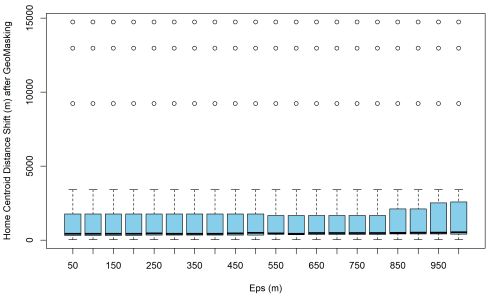
(c) Medoid Shift by Random Perturbation (1km)



(d) Medoid Shift by TAZ-based Aggregation



(e) Centroid Shift by Random Perturbation (1km)

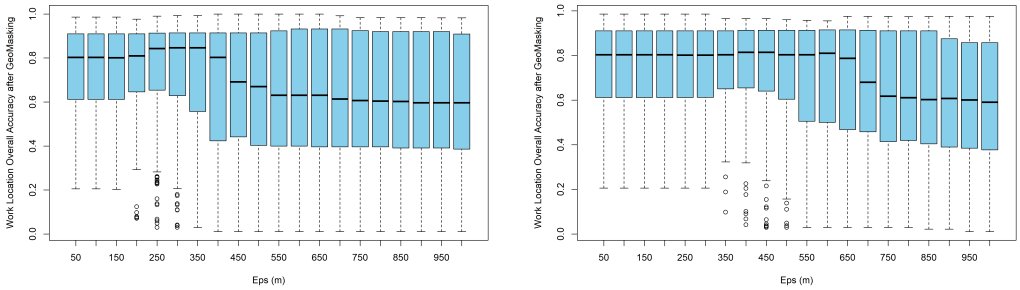


(f) Centroid Shift by TAZ-based Aggregation

Figure 5: The boxplot of F1-score, medoid and centroid distance shifts of home cluster detection with different DBSCAN parameters with random perturbation and the TAZ-based aggregation.

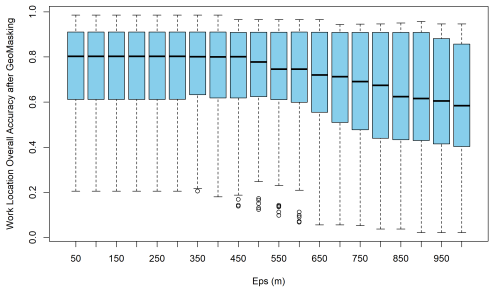
the medoid is only about 42m. The median shift distance is a more stable measure rather than the mean shift distance considering the outliers in spatially dispersed point patterns. The boxplot of F1-score, the medoid distance shift and the centroid distance shift of home

cluster detection after random perturbation (1km) can be seen in Figure 5(a), Figure 5(c), and Figure 5(e) respectively.

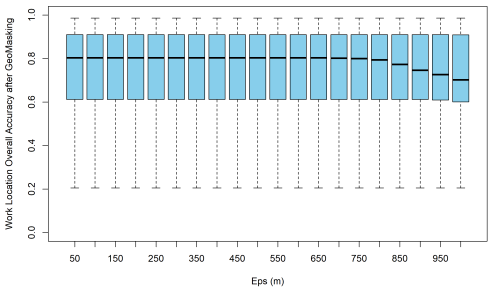


(a) Random Perturbation (1km)

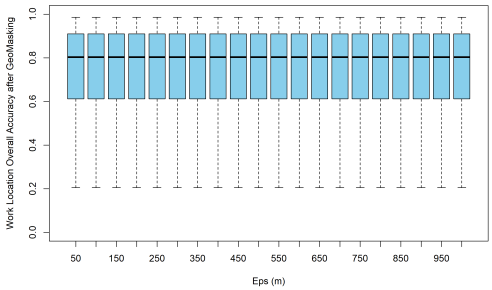
(b) Random Perturbation (2km)



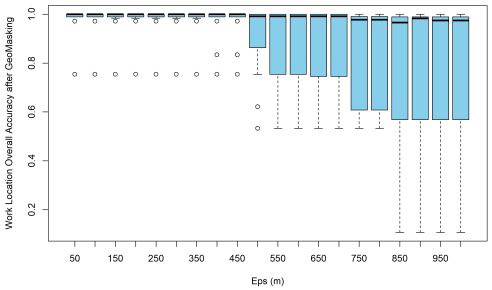
(c) Gaussian Masking (SD=0.01)



(d) Gaussian Masking (SD=0.03)



(e) Gaussian Masking (SD=0.05)



(f) TAZ-based aggregation (Madison)

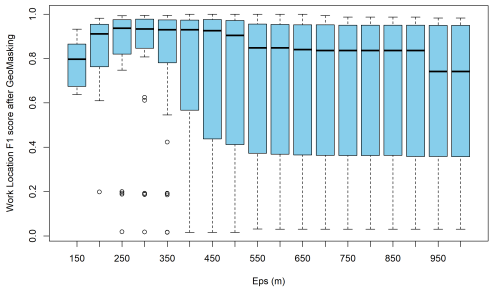
Figure 6: The boxplot of overall accuracy of work cluster detection with different DBSCAN parameters with random perturbation, Gaussian masking, and the TAZ-based aggregation.

As for the Gaussian perturbation, we found that it is effective for protecting users' home locations with proper parameters. The mean of overall accuracy for identifying users' home clusters using two-dimensional Gaussian kernels with 0.05 standard deviation (SD) is less

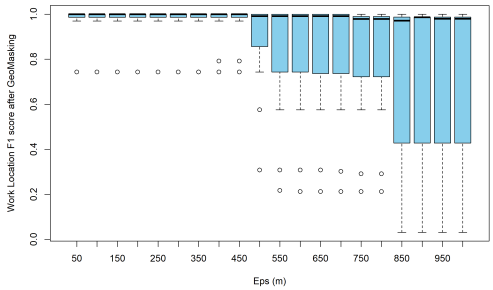
than 0.50 in average and the median is also less than 0.50 regardless of the DBSCAN parameter choices. And the Gaussian maskings with larger SD get more dispersed spatial point patterns and have very low prediction accuracy for home location identification. However, given the nature of sparse spatial distributions of digital footprints, the SD of geotagged tweet distributions of a user is often large (about 5~10km), and so is the distance shift from original points to displaced points after the Gaussian masking process. This is part of the reason why we could not correctly identify the home location clusters after Gaussian masking.

Results regarding geoprivacy protection of work locations during daytime (in Figure 6), differ from the home location case. The mean and median overall accuracy decreased to 0.661 and 0.722 using the random perturbation (1km) approach. The reason might be the diverse spatial patterns of human activity locations in daytime [33]. However, high median sensitivity 1.000 and F1-score 0.898 demonstrate that high percentage of true work location clusters can be successfully identified. The median of shift distances from the true work location to the medoid of work clustering results increased to about 90m (and shift to the centroid: 798m). The boxplot of F1-score, the medoid distance shift and the centroid distance shift of work cluster detection after random perturbation (1km) can be seen in Figure 7(a), Figure 7(c), and Figure 7(e) respectively. In addition, the overall prediction accuracy of work location decreases as the Eps increases using the random perturbation method. This is mainly because of the imbalanced work location data problem (i.e., the number of exposed work locations is much fewer than the non-work locations during the daytime) and the decrease of true negative predictions, which will be discussed later in Section 5. To deal with the imbalanced data, the balanced accuracy is also reported to measure the accuracy of the clustering results. The mean and median of the balanced accuracy are all 0.500, and in most cases the sensitivity is 0.000 and the specificity is 1.000, which shows that, despite the high overall accuracy, no work clusters after Gaussian masking are correctly identified. It is worth noting that the displacement of each tweet location might vary in different operations of perturbation process. But the overall quality measures for geoprivacy preservation in multiple operations did not change much (about 2% difference in our experiment) and the overall accuracy measures reported are stable.

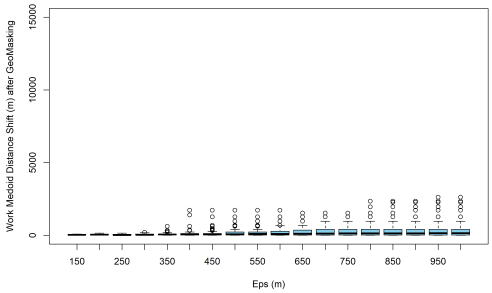
In addition, we also conducted the traditional aggregation-based masking analysis at the traffic analysis zones (TAZs) in the three urban areas (as shown in Figure 8). As a lot of human mobility and transportation studies using geotagged social media data are based on the home-work trips at the TAZ level, such a scale meets the spatial resolution requirement for urban transportation analysis. Also, we demonstrate here the aggregation results on the Madison area (as shown in the TAZ(H) and TAZ(W) columns of Table 1), since we are more familiar with Madison's traffic conditions and urban spatial layout. Also, unlike other users in Washington DC and Los Angeles, many of whom are tourists and thus have a very wide range of activities (even nationwide), the Madison users are more locally active and their tweet locations match Madison TAZ better, enabling more accurate and representative aggregation analysis. The TAZ-based aggregation method, however, still cannot protect the geoprivacy well of those active Twitter users given a high median overall accuracy of identifying home cluster (0.980) and work cluster (0.989). The boxplot of overall accuracy, F1-score, the medoid distance shift and the centroid distance shift of home cluster detection after TAZ-based aggregation can be seen in Figure 4(f), Figure 5(b), Figure 5(d), and Figure 5(f) respectively. Also, the same measures of work cluster detection can be seen in Figure 6(f), Figure 7(b), Figure 7(d), and Figure 7(f). All the accuracy measures align well



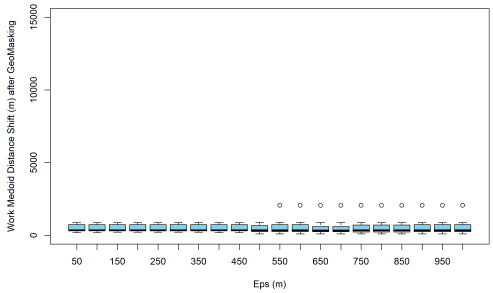
(a) F1-score by Random Perturbation (1km)



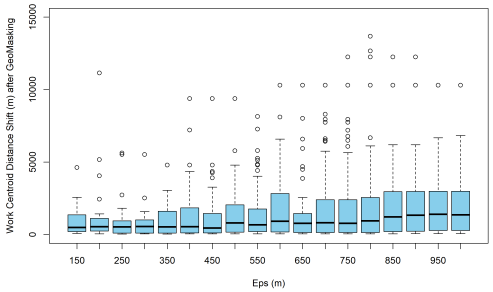
(b) F1-score by TAZ-based Aggregation



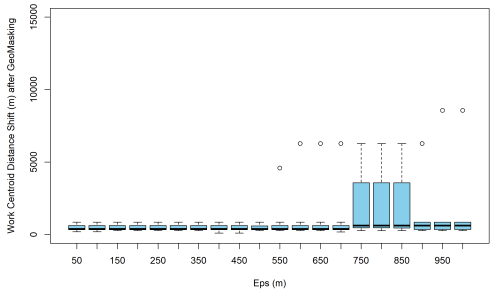
(c) Medoid Shift by Random Perturbation (1km)



(d) Medoid Shift by TAZ-based Aggregation



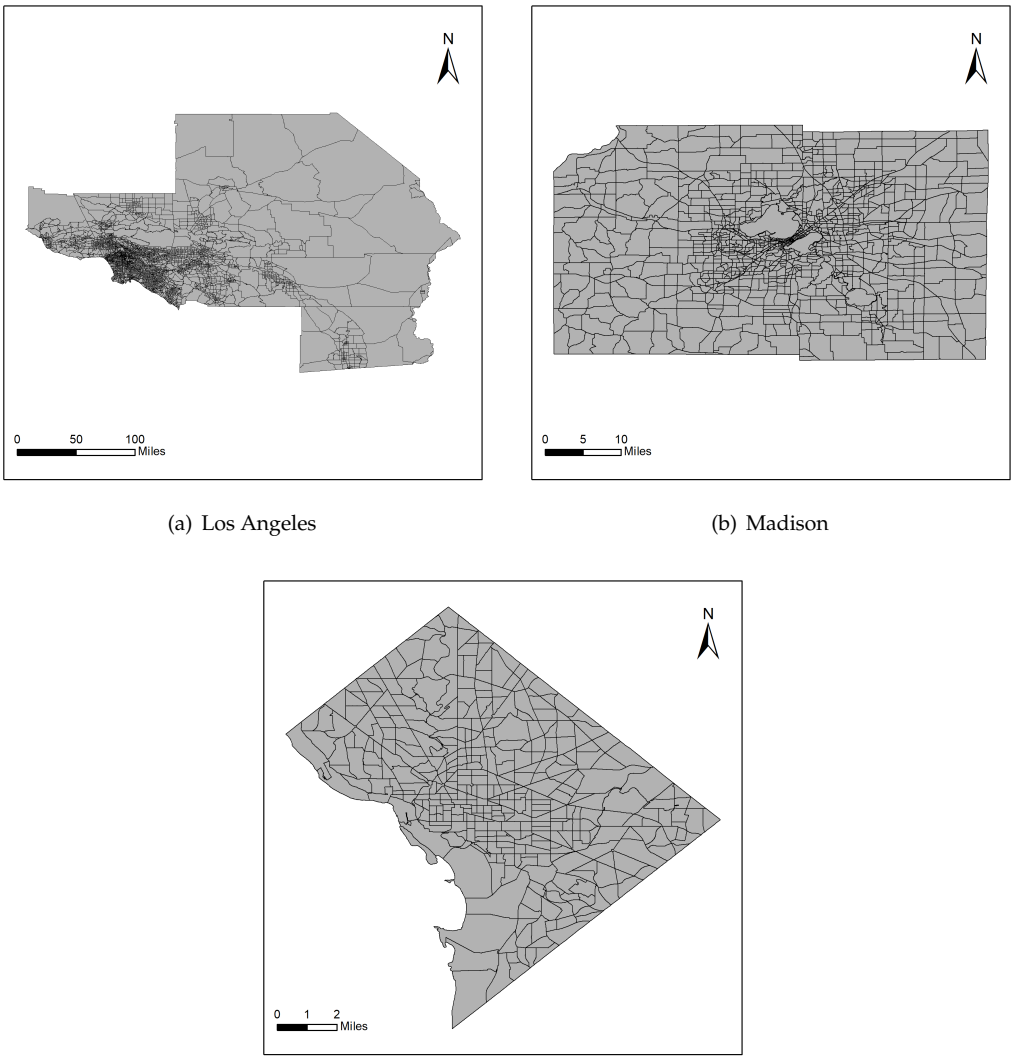
(e) Centroid Shift by Random Perturbation (1km)



(f) Centroid Shift by TAZ-based Aggregation

Figure 7: The boxplot of F1-score, medoid and centroid distance shift of work cluster detection with different DBSCAN parameters with random perturbation and the TAZ-based aggregation.

regardless of the setting for DBSCAN search radius Eps and confirm the ineffectiveness of geoprivacy preservation at the TAZ level using aggregation. It is worth noting that the "ineffectiveness" is in a sense for the protection of home cluster identity rather than the actual home location as there is still a possible distance shift between the true home



(a) Los Angeles

(b) Madison

(c) Washington DC

Figure 8: Traffic Analysis Zones (TAZs) of the three cities in this research.

location and the centroid of a TAZ polygon. From this perspective, it might be effective for protecting the true home (work) location, but the distance shift really depends on the spatial distribution of a home (work) location within the TAZs (i.e., the proximity to the TAZ center). The results show that the median of shift distances from the true home (work) location to the centroid of home (work) TAZ is about 485m (work: 460m) and to the medoid is about 403m (work: 360m).

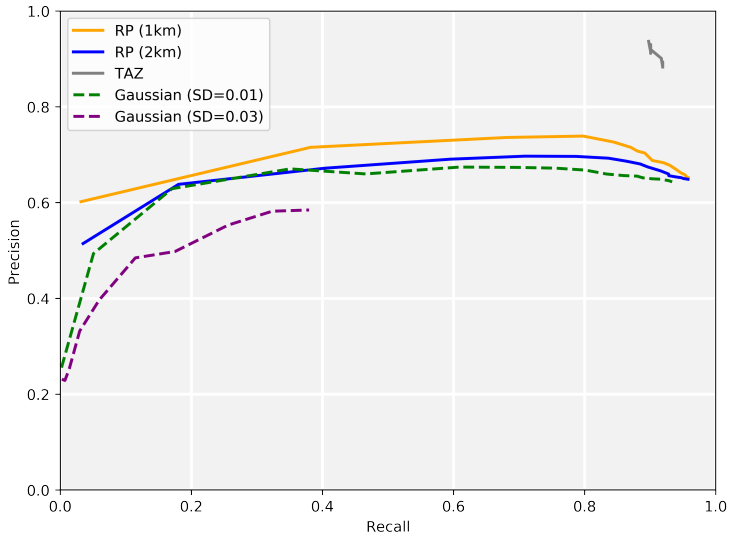
5 Discussion

5.1 Imbalanced data and geomasking performance

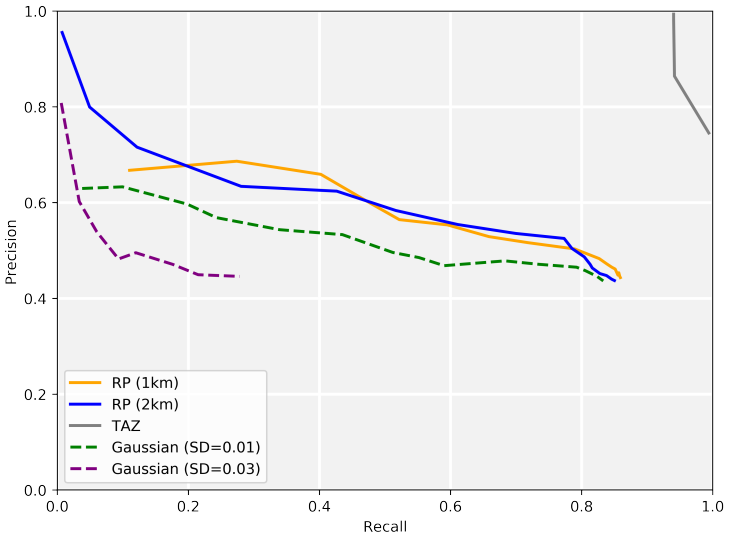
As shown in the results, the geomasking effectiveness differs largely on home locations and work locations due to their different spatiotemporal patterns. During the nighttime, the users often stay at home, and thus most of their nighttime tweet locations can represent their home locations. However, during the daytime, the users have more diverse activity space and do not necessarily stay at their work locations all the time. Thus there are many non-work tweets that are posted at other places, which results in an imbalanced data problem between work locations and non-work locations. The number of home locations (6,700) and non-home locations (6,053) during the nighttime is almost equal (about 1:1), which is more balanced and explains why the overall accuracy is still able to reach around 0.50 but with 0.0 recall rate after the Gaussian masking. However, the number of work locations (3,168) and non-work locations (8,904) during the daytime differs substantially (about 1:3), and the number of true negative predictions is therefore large enough so that the overall accuracy could be around 80% even when no true work location cluster is correctly detected. As the Eps (the search radius threshold for the DBSCAN cluster algorithm) increases, however, more true negative predictions become false positive predictions, resulting in a decreasing overall accuracy.

In this regard, focusing only on the overall accuracy alone is meaningless, and we need to take both the precision and recall into account to evaluate the geomasking performance. Therefore, we further computed and drew the Precision-Recall curves (PR-curves) for different geomasking methods based on different Eps (from 50m to 1000m) in Figure 9. The PR-curve is a comprehensive tool to measure the model performance even on imbalanced data since it takes both the precision and recall into consideration at the same time. Each curve represents the performance of a geomasking method. The higher the curve stays when moving from left to right, the higher precision the home (work) cluster detection algorithm gets, and therefore the worse the geomasking method performs. As shown in Figure 9, for both home locations and work locations, the random perturbation methods (both 1km and 2km) and TAZ-based aggregation methods have higher overall precisions and recalls, whereas the Gaussian masking methods (SD=0.01 and 0.03) are able to suppress both the precision and recall at the same time. Note that the Gaussian masking method with 0.05 SD doesn't appear in the Figure since it protects the geoprivacy well and thus there is no precision or recall rates for drawing its PR curve. As such, with proper parameter settings, the Gaussian geomasking method could have a better effectiveness for protecting the location privacy of Twitter users than the random perturbation method and the TAZ-based aggregation method.

In addition, we also explored the differences between the settings of distance threshold (for random perturbation) and the standard deviations (for Gaussian geomasking) as well as their influences on the geomasking performance. As shown in the violin plot (Figure 10), the values of distance shifts of tweet locations (daytime and nighttime) after random perturbation are evenly distributed within the distance threshold, while after the Gaussian geomasking, the distribution of distance shifts is much closer to a normal distribution with a wider range. For the random perturbation (2km) and the Gaussian geomasking (SD=0.01), although they have a similar average distance shift (about 1081m and 1211m respectively), the latter still has a significantly better performance due to its high uncertainty



(a) Home cluster



(b) Work cluster

Figure 9: Precision-Recall curves of different geomasking methods.

level by comparing their PR-curves in Figure 9. It shows that the perturbation that follows a normal distribution would have a larger but more natural influence on the geotagged tweet locations than the random perturbation, and we believe this partially explains why the Gaussian masking method with proper standard deviation settings could have a better performance than the random perturbation method.

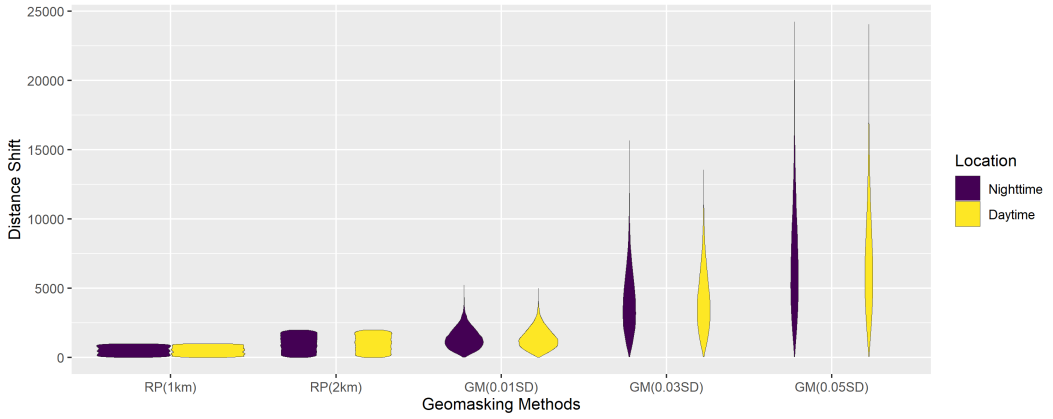


Figure 10: The violin plot of distance shifts of tweet locations after geomasking.

5.2 Implications among privacy, analytics, and uncertainty

The results in our experiment demonstrate that one geomasking method could effectively protect users' geoprivacy but may reduce the spatial analysis capability and introduce uncertainty to further analytics. Inspired by several theoretical frameworks in geovisualization and geospatial semantics studies [25, 36], we herein present a three-dimensional visualization framework (as shown in Figure 11) including privacy, analytics, and uncertainty as a tool to evaluate and inform the selection of appropriate geomasking methods under different contexts. The first dimension is about the capability to protect users' geoprivacy from low to high. The second dimension is the spatial resolution of geospatial analytics from coarse to fine. And the third dimension is uncertainty level from low to high [21]. Two presented geomasking methods (Gaussian and random perturbations) with different parameter settings and the TAZ-based aggregation method in our experiments using geotagged tweets are tentatively placed in this 3D cube. It is worth noting that the placement of each method is estimated from the results of our case study shown in Table 1 (e.g., based on the accuracy measures and the distance shifts). We think that such a 3D cube visualization can serve as an assessment tool for evaluating other geomasking methods from the three aspects simultaneously as well. For instance, one may add the donut masking, Voronoi masking, and other geomasking techniques into this framework with a different application domain (e.g., public health). While we mainly focus on the privacy preservation aspect in this research, the exploration of other two dimensions (i.e., spatial analytics and uncertainty) requires more investigation in future work.

5.3 Limitations

Several limitations exist in our current study. First, our manual labeling approach has uncertainty about users' actual home or work locations without their interview confirmation. Thus, it is possible that the labeling results may not be able to reveal the ground truth of all the actual home or work locations of those Twitter users. However, we try our best to ensure the quality of the labels using a set of comprehensive rules mentioned in Section 4.

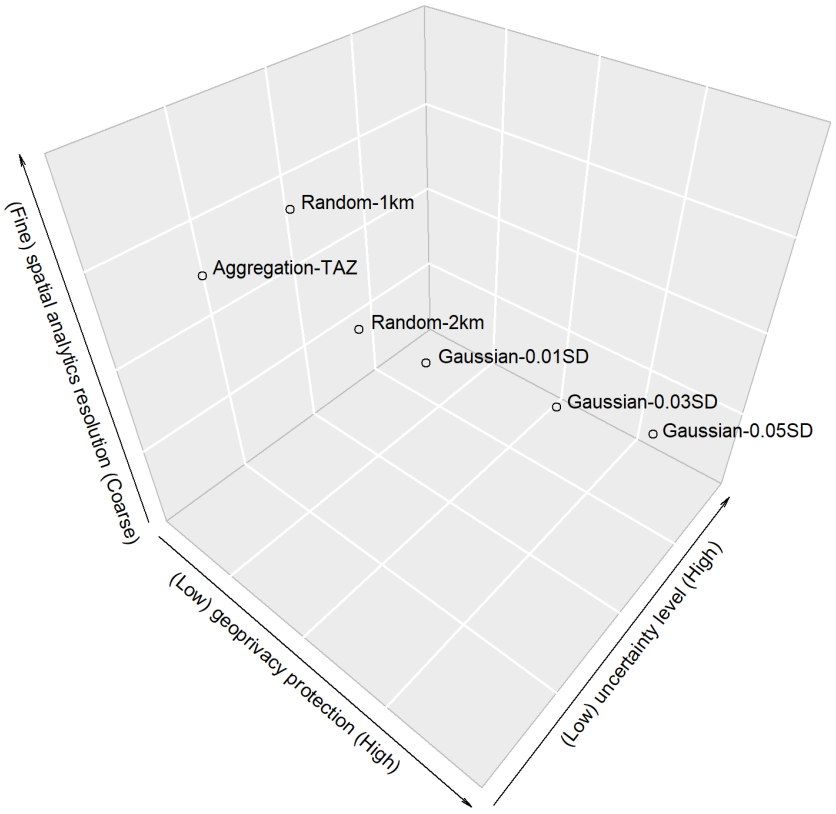


Figure 11: A 3D-cube framework for assessing different geomasking techniques; the position of each method is estimated from the results of our case study.

Second, such an approach is also limited on the sample size since it needs labor-intensive labeling process. Even within the same study group, we agree on most of the manual labels but individual differences do exist towards the concordance of labeled training data. Third, the location cluster detection results depend heavily on the number of tweets of a user, and their particular tweeting behavior. If a user posts a large number of tweets from home (or work location) then it is easier to identify his/her home (or work location) compared to the users who tweet rarely. Last but not least, the presented home (work)-detection method only relies on the DBSCAN spatial clustering for geotagged tweets. Other approaches such

as the detection of home-work locations using recurring trips also exist. The coordinate information may not be as important as the spatial interaction frequency among those points using the trip-based detection approaches.

5.4 Broader impacts

In fact, Twitter removed support for precise geotagging since June, 2019. However, the metadata of historical tweets prior to the policy change may still reveal precise GPS coordinates. In addition, when a user deletes a geotagged tweet[2], Twitter does not guarantee the information will be completely removed from all copies of the data on third-party applications or in external search results. Even if the precise GPS location is not available anymore, Twitter users are still able to add place tags (e.g., a city, office building, apartment, landmark, and many other types of places) to their geotagged tweets, which can be converted to the GPS coordinates (often using the centroid as a representation location). This is similar to the aforementioned aggregation-based masking approach, thus we may still be able to get users' sensitive locations based on fine-scale place tags. People should be aware that sharing or publishing such kind of location data involve geoprivacy issues and the geomasking technique provides a way to help mitigate the problem not only for Twitter users but also for other social media platforms such as Facebook, Flickr, Weibo, and Instagram where geotagging or place-tagging is accessible, as well as for mobile applications that track individual locations.

6 Conclusions and future work

In this work, we have explored the effectiveness of three popular geomasking techniques for protecting the geoprivacy of active Twitter users who frequently share geotagged tweets in their home or work locations. Based on our experiments, the two-dimensional Gaussian masking with proper standard deviation settings is found to be more effective on hiding or shifting social media user's home location than the random perturbation and the aggregation masks. However, the Gaussian masking may also lower the spatial resolution of geospatial analytics given the sparsity nature in geotagged social media data. Our experiments show that small-distance (such as within 1km or 2km) random perturbations do not sufficiently help the protection of users' geoprivacy because the majority of their home or work locations can still be correctly identified with high accuracy and very small median shift distance from the ground-truth locations. Our research offers insights into the geoprivacy concern of social media users' georeferenced data sharing for future development of location-based applications and services.

For future work, one direction would be what is the impact of these geoprivacy enhancements on the user experience comparing with simply removing the benefit to the user of posting geotagged tweets. Another direction is about the protection of geoprivacy using the spatiotemporal information and among other activity place types (e.g., shopping, entertainment) of social media users. In addition, we would like to extend our workflow to other cities to test whether our conclusion drawn from our case study is generalizable. Although Twitter decided to remove the precise location coordinate of each tweet while keeping the place tagging function, a precise location is very critical in some application

[2]https://help.twitter.com/en/using-twitter/tweet-location

scenarios such as disaster response and crime investigation. The trade-off between the requirement of spatial analysis resolution and the privacy preservation capability requires more research on different scenarios.

Acknowledgments

Support for this research was provided by the University of Wisconsin–Madison Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation.

References

[1] ADAMS, B., AND JANOWICZ, K. Thematic signatures for cleansing and enriching place-related linked data. *Intl. Journal of Geographical Information Science 29*, 4 (2015), 556–579. doi:10.1080/13658816.2014.989855.

[2] ALTMAN, D. G., AND BLAND, J. M. Diagnostic tests. 1: Sensitivity and specificity. *BMJ: British Medical Journal 308*, 6943 (1994), 1552. doi:10.1136/bmj.308.6943.1552.

[3] ANDRÉS, M., BORDENABE, N., CHATZIKOKOLAKIS, K., AND PALAMIDESSI, C. Geo-indistinguishability: Differential privacy for location-based systems. In *20th ACM Conference on Computer and Communications Security* (2013), ACM, pp. 901–914. doi:10.1145/2508859.2516735.

[4] ARMSTRONG, M. P., RUSHTON, G., ZIMMERMAN, D. L., ET AL. Geographically masking health data to preserve confidentiality. *Statistics in Medicine 18*, 5 (1999), 497–525.

[5] BERESFORD, A. R., AND STAJANO, F. Location privacy in pervasive computing. *IEEE Pervasive Computing 2*, 1 (2003), 46–55. doi:10.1109/MPRV.2003.1186725.

[6] CAO, G., WANG, S., HWANG, M., PADMANABHAN, A., ZHANG, Z., AND SOLTANI, K. A scalable framework for spatiotemporal analysis of location-based social media data. *Computers, Environment and Urban Systems 51* (2015), 70–82. doi:10.1016/j.compenvurbsys.2015.01.002.

[7] CLARKE, K. C. A multiscale masking method for point geographic data. *Intl. Journal of Geographical Information Science 30*, 2 (2016), 300–315. doi:10.1080/13658816.2015.1085540.

[8] DUCKHAM, M., AND KULIK, L. A formal model of obfuscation and negotiation for location privacy. In *Intl. conference on pervasive computing* (2005), Springer, pp. 152–170. doi:10.1007/11428572_10.

[9] DUCKHAM, M., AND KULIK, L. Location privacy and location-aware computing. In *Dynamic and Mobile GIS*. CRC press, 2006, pp. 63–80.

[10] ESTER, M., KRIEGEL, H.-P., SANDER, J., XU, X., ET AL. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD* (1996), vol. 96, pp. 226–231.

[11] FRONTERRÈ, C. Spatial analysis of geomasked and aggregated data. *Ph.D. thesis, Corso di Dottorato di Ricerca in Scienze Statistiche* (2018), 1–56.

[12] GAO, S., JANOWICZ, K., MONTELLO, D. R., HU, Y., YANG, J.-A., MCKENZIE, G., JU, Y., GONG, L., ADAMS, B., AND YAN, B. A data-synthesis-driven method for detecting and extracting vague cognitive regions. *Intl. Journal of Geographical Information Science 31*, 6 (2017), 1245–1271. doi:10.1080/13658816.2016.1273357.

[13] GAO, S., YANG, J.-A., YAN, B., HU, Y., JANOWICZ, K., AND MCKENZIE, G. Detecting origin-destination mobility flows from geotagged tweets in greater Los Angeles area. In *Eighth Intl. Conference on Geographic Information Science (GIScience'14)* (2014), Citeseer.

[14] GEDIK, B., AND LIU, L. Protecting location privacy with personalized k-anonymity: Architecture and algorithms. *IEEE Transactions on Mobile Computing 7*, 1 (2008), 1–18. doi:10.1109/TMC.2007.1062.

[15] GRUTESER, M., AND GRUNWALD, D. Anonymous usage of location-based services through spatial and temporal cloaking. In *Proc. of the 1st Intl. Conference on Mobile Systems, Applications and Services* (2003), ACM, pp. 31–42. doi:10.1145/1066116.1189037.

[16] HÄGERSTRAAND, T. What about people in regional science? *Papers in Regional Science 24*, 1 (1970), 7–24. doi:10.1111/j.1435-5597.1970.tb01464.x.

[17] HAMPTON, K. H., FITCH, M. K., ALLSHOUSE, W. B., DOHERTY, I. A., GESINK, D. C., LEONE, P. A., SERRE, M. L., AND MILLER, W. C. Mapping health data: Improved privacy protection with donut method geomasking. *American Journal of Epidemiology 172*, 9 (2010), 1062–1069. doi:10.1093/aje/kwq248.

[18] HAN, S. Y., TSOU, M.-H., AND CLARKE, K. C. Do global cities enable global views? Using twitter to quantify the level of geographical awareness of US cities. *PLOS ONE 10*, 7 (2015), e0132464. doi:10.1371/journal.pone.0132464.

[19] HO, S.-S., AND RUAN, S. Differential privacy for location pattern mining. In *Proc. of the 4th ACM SIGSPATIAL Intl. Workshop on Security and Privacy in GIS and LBS* (2011), ACM, pp. 17–24. doi:10.1145/2071880.2071884.

[20] HU, Y., GAO, S., JANOWICZ, K., YU, B., LI, W., AND PRASAD, S. Extracting and understanding urban areas of interest using geotagged photos. *Computers, Environment and Urban Systems 54* (2015), 240–254. doi:10.1016/j.compenvurbsys.2015.09.001.

[21] HUANG, Q., AND WONG, D. W. Modeling and visualizing regular human mobility patterns with uncertainty: An example using twitter data. *Annals of the Association of American Geographers 105*, 6 (2015), 1179–1197. doi:10.1080/00045608.2015.1081120.

[22] HUANG, Q., AND WONG, D. W. Activity patterns, socioeconomic status and urban spatial structure: What can social media data tell us? *Intl. Journal of Geographical Information Science 30*, 9 (2016), 1873–1898. doi:10.1080/13658816.2016.1145225.

[23] HUANG, Q., AND XIAO, Y. Geographic situational awareness: Mining tweets for disaster preparedness, emergency response, impact, and recovery. *ISPRS Intl. Journal of Geo-Information 4*, 3 (2015), 1549–1568. doi:10.3390/ijgi4031549.

[24] JANOWICZ, K., MCKENZIE, G., HU, Y., ZHU, R., AND GAO, S. Using semantic signatures for social sensing in urban environments. In *Mobility Patterns, Big Data and Transport Analytics*. Elsevier, 2019, pp. 31–54. doi:10.1016/B978-0-12-812970-8.00003-8.

[25] JANOWICZ, K., VAN HARMELEN, F., HENDLER, J. A., AND HITZLER, P. Why the data train needs semantic rails. *AI Magazine* (2014). doi:10.1609/aimag.v36i1.2560.

[26] KANG, Y., WANG, J., WANG, Y., ANGSUESSER, S., AND FEI, T. Mapping the sensitivity of the public emotion to the movement of stock market value: A case study of manhattan. *Intl. Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences 42* (2017), 1213–1221. doi:10.5194/isprs-archives-XLII-2-W7-1213-2017.

[27] KESSLER, C., AND MCKENZIE, G. A geoprivacy manifesto. *Transactions in GIS 22*, 1 (2018), 3–19. doi:10.1111/tgis.12305.

[28] KRUMM, J. A survey of computational location privacy. *Personal and Ubiquitous Computing 13*, 6 (2009), 391–399. doi:10.1007/s00779-008-0212-5.

[29] KUHN, M. Building predictive models in R using the caret package. *Journal of Statistical Software 28*, 5 (2008), 1–26. doi:10.18637/jss.v028.i05.

[30] KWAN, M.-P., CASAS, I., AND SCHMITZ, B. Protection of geoprivacy and accuracy of spatial information: How effective are geographical masks? *Cartographica: The Intl. Journal for Geographic Information and Geovisualization 39*, 2 (2004), 15–28. doi:10.3138/X204-4223-57MK-8273.

[31] LIU, L., ANDRIS, C., AND RATTI, C. Uncovering cabdrivers' behavior patterns from their digital traces. *Computers, Environment and Urban Systems 34*, 6 (2010), 541–548. doi:10.1016/j.compenvurbsys.2010.07.004.

[32] LIU, Q., WANG, Z., AND YE, X. Comparing mobility patterns between residents and visitors using geo-tagged social media data. *Transactions in GIS 22*, 6 (2018), 1372–1389. doi:10.1111/tgis.12478.

[33] LIU, X., HUANG, Q., AND GAO, S. Exploring the uncertainty of activity zone detection using digital footprints with multi-scaled DBSCAN. *Intl. Journal of Geographical Information Science* (2019), 1–28. doi:10.1080/13658816.2018.1563301.

[34] LIU, Y., LIU, X., GAO, S., GONG, L., KANG, C., ZHI, Y., CHI, G., AND SHI, L. Social sensing: A new approach to understanding our socioeconomic environments. *Annals of the Association of American Geographers 105*, 3 (2015), 512–530. doi:10.1080/00045608.2015.1018773.

[35] LUO, F., CAO, G., MULLIGAN, K., AND LI, X. Explore spatiotemporal and demographic characteristics of human mobility via twitter: A case study of chicago. *Applied Geography 70* (2016), 11–25. doi:10.1016/j.apgeog.2016.03.001.

[36] MACEACHREN, A. M., AND KRAAK, M.-J. Research challenges in geovisualization. *Cartography and Geographic Information Science 28*, 1 (2001), 3–12. doi:10.1559/152304001782173970.

[37] MAI, G., JANOWICZ, K., HU, Y., AND GAO, S. ADCN: An anisotropic density-based clustering algorithm for discovering spatial point patterns with noise. *Transactions in GIS 22*, 1 (2018), 348–369. doi:10.1111/tgis.12313.

[38] MCKENZIE, G., AND JANOWICZ, K. Where is also about time: A location-distortion model to improve reverse geocoding using behavior-driven temporal semantic signatures. *Computers, Environment and Urban Systems 54* (2015), 1–13. doi:10.1016/j.compenvurbsys.2015.05.003.

[39] MCKENZIE, G., JANOWICZ, K., GAO, S., YANG, J.-A., AND HU, Y. POI pulse: A multi-granular, semantic signature-based information observatory for the interactive visualization of big geosocial data. *Cartographica: The Intl. Journal for Geographic Information and Geovisualization 50*, 2 (2015), 71–85. doi:10.3138/cart.50.2.2662.

[40] MCKENZIE, G., JANOWICZ, K., AND SEIDL, D. Geo-privacy beyond coordinates. In *Geospatial Data in a Changing World*. Springer, 2016, pp. 157–175. doi:10.1007/978-3-319-33783-8_10.

[41] MILLER, H. J. A measurement theory for time geography. *Geographical Analysis 37*, 1 (2005), 17–45. doi:10.1111/j.1538-4632.2005.00575.x.

[42] NIU, B., LI, Q., ZHU, X., CAO, G., AND LI, H. Achieving k-anonymity in privacy-aware location-based services. In *IEEE INFOCOM 2014-IEEE Conference on Computer Communications* (2014), IEEE, pp. 754–762. doi:10.1109/INFOCOM.2014.6848002.

[43] POWERS, D. M. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies 2*, 1 (2011), 37–63. doi:10.9735/2229-3981.

[44] RUBINSTEIN, B. I., BARTLETT, P. L., HUANG, L., AND TAFT, N. Learning in a large function space: Privacy-preserving mechanisms for SVM learning. *Journal of Privacy and Confidentiality 4* (2012), 65–100. doi:10.29012/jpc.v4i1.612.

[45] SEIDL, D. E., JANKOWSKI, P., AND NARA, A. An empirical test of household identification risk in geomasked maps. *Cartography and Geographic Information Science 46*, 6 (2019), 475–488. doi:10.1080/15230406.2018.1544932.

[46] SEIDL, D. E., JANKOWSKI, P., AND TSOU, M.-H. Privacy and spatial pattern preservation in masked GPS trajectory data. *Intl. Journal of Geographical Information Science 30*, 4 (2016), 785–800. doi:10.1080/13658816.2015.1101767.

[47] SEIDL, D. E., PAULUS, G., JANKOWSKI, P., AND REGENFELDER, M. Spatial obfuscation methods for privacy protection of household-level data. *Applied Geography 63* (2015), 253–263. doi:10.1016/j.apgeog.2015.07.001.

[48] SIJTSMA, B., QVARFORDT, P., AND CHEN, F. Tweetviz: Visualizing tweets for business intelligence. In *Proc. of the 39th Intl. ACM SIGIR conference on Research and Development in Information Retrieval* (2016), ACM, pp. 1153–1156. doi:10.1145/2911451.2911470.

[49] SOLIMAN, A., SOLTANI, K., YIN, J., PADMANABHAN, A., AND WANG, S. Social sensing of urban land use based on analysis of twitter users' mobility patterns. *PLOS ONE 12*, 7 (2017), e0181657. doi:10.1371/journal.pone.0181657.

[50] SONG, C., QU, Z., BLUMM, N., AND BARABÁSI, A.-L. Limits of predictability in human mobility. *Science 327*, 5968 (2010), 1018–1021. doi:10.1126/science.1177170.

[51] TSOU, M.-H. Research challenges and opportunities in mapping social media and big data. *Cartography and Geographic Information Science 42*, sup1 (2015), 70–74. doi:10.1080/15230406.2015.1059251.

[52] VAN DER LAAN, M., POLLARD, K., AND BRYAN, J. A new partitioning around medoids algorithm. *Journal of Statistical Computation and Simulation 73*, 8 (2003), 575–584. doi:10.1080/0094965031000136012.

[53] VELMURUGAN, T., AND SANTHANAM, T. Computational complexity between K-means and K-medoids clustering algorithms for normal and uniform distributions of data points. *Journal of Computer Science 6*, 3 (2010), 363. doi:10.3844/jcssp.2010.363.368.

[54] WANG, Z., AND YE, X. Social media analytics for natural disaster management. *Intl. Journal of Geographical Information Science 32*, 1 (2018), 49–72. doi:10.1080/13658816.2017.1367003.

[55] WANG, Z., YE, X., AND TSOU, M.-H. Spatial, temporal, and content analysis of twitter for wildfire hazards. *Natural Hazards 83*, 1 (2016), 523–540. doi:10.1007/s11069-016-2329-6.

[56] XU, Y., SHAW, S.-L., ZHAO, Z., YIN, L., FANG, Z., AND LI, Q. Understanding aggregate human mobility patterns using passive mobile phone location data: A home-based approach. *Transportation 42*, 4 (2015), 625–646. doi:10.1007/s11116-015-9597-y.

[57] XU, Y., SHAW, S.-L., ZHAO, Z., YIN, L., LU, F., CHEN, J., FANG, Z., AND LI, Q. Another tale of two cities: Understanding human activity space using actively tracked cellphone location data. *Annals of the American Association of Geographers 106*, 2 (2016), 489–502. doi:10.4324/9781315266336-27.

[58] YAO, X. A., HUANG, H., JIANG, B., AND KRISP, J. M. Representation and analytical models for location-based big data. *Intl. Journal of Geographical Information Science 33*, 4 (2019), 707–713. doi:10.1080/13658816.2018.1562068.

[59] YIN, J., SOLIMAN, A., YIN, D., AND WANG, S. Depicting urban boundaries from a mobility network of spatial interactions: A case study of great britain with geo-located twitter data. *Intl. Journal of Geographical Information Science 31*, 7 (2017), 1293–1313. doi:10.1080/13658816.2017.1282615.

[60] ZANDBERGEN, P. A. Ensuring confidentiality of geocoded health data: Assessing geographic masking strategies for individual-level data. *Advances in Medicine 2014* (2014), 1–14. doi:10.1155/2014/567049.

[61] ZHANG, Z., RUBINSTEIN, B. I., AND DIMITRAKAKIS, C. On the differential privacy of bayesian inference. In *Proc. of the Thirtieth AAAI Conference on Artificial Intelligence* (2016), Phoenix, Arizona, February 12-17, 2016, pp. 2365–2371.

[62] ZHU, R., HU, Y., JANOWICZ, K., AND MCKENZIE, G. Spatial signatures for geographic feature types: Examining gazetteer ontologies using spatial statistics. *Transactions in GIS 20*, 3 (2016), 333–355. doi:10.1111/tgis.12232.