

# Geo-Privacy Beyond Coordinates

Grant McKenzie, Krzysztof Janowicz and Dara Seidl

**Abstract** The desire to share one's location with friends and family or to use location information for navigation and recommendations services is often overshadowed by the need to preserve privacy. As recent progress in big data analytics, ambient intelligence, and conflation techniques is met with the economy's growing hunger for data, even formerly negligible digital footprints become revealing of our activities. The majority of established geo-privacy research tries to protect an individual's location by different masking or perturbation techniques or by suppressing and generalizing an individual's characteristics to a degree where she cannot be singled out from a crowd. In this work we demonstrate that location privacy may already be compromised before these techniques take effect. More concretely, we discuss how everyday digital footprints such as timestamps, geosocial check-ins, and short social media messages, e.g., tweets, are indicative of the user's location. We focus particularly on places and highlight how protecting place-based information differs from a purely spatial perspective. The presented research is based on so-called semantic signatures that are mined from millions of geosocial check-ins and enable a probabilistic framework on the level of geographic feature types, here Points Of Interest (POI). While our work is compatible with leading privacy techniques, we take a user-centric perspective and illustrate how privacy-enabled services could guide the users by increasing information entropy.

**Keywords** Privacy · Place · Semantic signature · Location · Geosocial

---

G. McKenzie (✉) · K. Janowicz · D. Seidl  
STKO Lab, Department of Geography, University of California, Santa Barbara, USA  
e-mail: grant.mckenzie@geog.ucsb.edu

K. Janowicz  
e-mail: janowicz@ucsb.edu

D. Seidl  
Department of Geography, San Diego State University, San Diego, USA  
e-mail: dseidl@geog.sdsu.edu

© Springer International Publishing Switzerland 2016  
T. Sarjakoski et al. (eds.), *Geospatial Data in a Changing World*,  
Lecture Notes in Geoinformation and Cartography,  
DOI 10.1007/978-3-319-33783-8\_10

## 1 Introduction and Motivation

While data privacy continues to be an area of worry and confusion for many, recent concerns over the privacy of *location* information specifically have come to the societal forefront. With the increase in mobile devices, as well as technical advances in ambient intelligence powered by the Internet of Things (IoT), location information has become ubiquitous. It has been widely recognized that the resulting technological and social implications will change our understanding of privacy (Bohn et al. 2005; Weber 2010). In fact, personal location information is now arguably a commodity to be traded for services, e.g., for navigation applications, local search, and coupons. Social media have also had a role to play in the advancement of location information usage. An increasing number of social applications allow, and increasingly require, some aspect of location to be shared, be it through posts, messages, check-ins, or photos. While many of these services request location information to improve the user experience, e.g., to show nearby places recommended by friends, other services do not provide clear benefits to the user and collect a variety of personal data in the background (McKenzie and Janowicz 2014). A recent study, for instance, shows that smartphone users are still unaware of the extent and also the frequency at which their personal data are being collected and that they would benefit from more fine grained privacy settings and alerts (Almuhimedi et al. 2015). Even coarse location information can be revealing. In fact, 95 % of individuals can be uniquely identified by just 4 spatio-temporal fixes from cell antennas (de Montjoye et al. 2013).

Consequently, when discussing geo-privacy, people primarily think of geographic coordinates and positioning techniques such as Global Navigation Satellite Systems (GNSS), Wi-Fi-based positioning systems (WPS), Bluetooth Low Energy (BLE) beacons, or radio towers. There are, however, various other possibilities to infer somebody's location and, at least in terms of geo-privacy, some of them may be more revealing than geographic coordinates alone. Additionally, these approaches do not require access to the user's mobile device. This is particularly important as it dramatically increases the number of parties that may infer a user's location. In contrast to positioning techniques, these approaches rely on the notions of *place* and *place types* instead of merely focusing on geographic *space*. Intuitively, there are certain, often latent, place characteristics that emerge from human behavior towards these places and define them as being of a common type, e.g., *bar* or *office*. With respect to temporal characteristics, for instance, a place that is mostly visited during the evenings and weekends is more likely a bar than an office building. Similarly, a place where people predominantly talk about tacos, burritos, and tequila is more likely to be a Mexican restaurant than a Polish restaurant. In an analogy to remote sensing, a set of spatial, temporal, and thematic characteristics that jointly identify a type of place is referred to as the *semantic signature* of said type (Janowicz 2012).

In this work, we employ these signatures to demonstrate how apparently harmless digital footprints such as social media messages, check-in timestamps, and so forth can be used to compromise a user's geo-privacy before position masking techniques come into play. While our work is compatible with established methods for

location privacy, we focus on digital footprints here and how types of places impact geo-privacy. The concern in this case is that people should be aware that even if they don't explicitly share their geographic coordinates that their location can be probabilistically determined based on the words that they write, the timestamps that they make public, and a basic understanding of the spatial and platial<sup>1</sup> configuration of a city.

**The contributions of this work are as follows:**

1. We build on existing work in the area of geo-privacy to show how non-spatial content published by an individual can lead to the disclosure of information directly related to her location.
2. We demonstrate how semantic signatures, built from millions of geosocial footprints, can be used to infer the place type of the location someone is visiting. Moreover, we show that it is possible to quantify this inference and calculate the probability of determining one's location based on her content.
3. We offer a window into what is possible provided seemingly innocuous information. This work suggests ways that content publishers may adjust one or more pieces of published content in order to reduce the risk of revealing their location.

The remainder of the paper is organized as follows. Section 2 introduces related research relevant for the work at hand. Section 3 introduces the datasets used for our study and briefly reviews how the *semantic signatures* were constructed. Three different groups of semantic bands (spatial, temporal and thematic) are discussed in the section following this (Sect. 4). In Sect. 5, we implement our approach through a use case that demonstrates the importance of the semantic signatures in privacy preservation. Finally, we conclude with ideas for future work in Sect. 6.

## 2 Related Work

Geo-privacy research efforts in the GI science community have focused primarily on geomasking or obfuscation techniques, which introduce inaccuracy to geographic coordinates in an effort to balance the protection of location privacy and preservation of spatial information (Armstrong et al. 1999). Attention to the development and evaluation of geomasking procedures has given rise to a large body of work in recent years (Hampton et al. 2010; Zandbergen 2014; Keith C 2015; Kounadi and Leitner 2015; Seidl E. et al. 2015; Seidl 2015; Zhang et al. 2015). The foci of masking studies, which include the testing of distance thresholds and quantification of personal reidentification risk, remain unable to address the impact on location privacy of individuals generating location-bearing content outside a masked data set. A major missing component from these works is the consideration of other data disclosing personal locations even when geographic coordinates are omitted or masked to remain confidential.

---

<sup>1</sup>Following recent literature, we will use the term *platial* here for 'place-based' (Goodchild 2015).

Geo-privacy in masking studies is often defined as the right of the individual to determine how, when, and the extent to which his or her location data is shared with others (Duckham and Kulik 2006). This definition places an emphasis on human agency in privacy rights and is arguably unrealistic in a digital age characterized by frequent and rapid data exchange, where it is difficult to keep track of the parties to which personal data are transmitted. Setting a concrete definition of geo-privacy also opposes other frequently cited conceptual approaches that eschew specific definitions. The definition presented here, however, is in line with the purpose of this paper, which is to introduce unique means by which content publishers, e.g., social media users, may control the release of their location data, namely by considering what is possible with semantic signatures.

The measurement of privacy in a release of data is framed as the risk of identity disclosure. The principle of  $k$ -anonymity describes a release of data where each person in the data set is indistinguishable from  $k - 1$  other individuals in the same data set (Sweeney 2002). The  $k$ -anonymity property does not recognize the side information that an adversary might have about an individual in the database. Another development in information privacy studies is differential privacy, which addresses the problem auxiliary information outside a database poses to the notion of absolute disclosure prevention (Dwork 2011).

Compared to data collected and transferred to third parties in traditional data collection models, individuals do have some agency in the location information they share in user-generated content. The benefits of participation in location-sharing applications (LSAs) or other social networks tend to outweigh perceived privacy risks for users. Social influence is shown to have a strong impact on the adoption of a location sharing application (LSA) among university students (Beldad and Kusumadewi 2015), which extends from having friends or peers known to use the application. Users of the location check-in application Foursquare report that motivations for location sharing include coordination with friends, presentation of self, gaming aspects, and peace of mind or safety purposes (Lindqvist et al. 2011). Location reporting in other social media is not limited to GPS-assisted check-ins, and may be based on text content. Consider the message, “finally home,” which may be posted for peace of mind or coordination purposes. The site “Please Rob Me”<sup>2</sup> used a classifier predicting whether or not a Twitter user was home based on tweets to demonstrate how such information could be exploited by an adversary (Gambis et al. 2010).

Another consideration for this work is whether content publishers are likely to embrace new options for protecting their geo-privacy. A survey of location privacy preferences for personal GPS data finds that providing more complex privacy options, including setting temporal limits and specific locations that may not be shared, leads to more location sharing (Benisch et al. 2011). This provides support for developing an application that allows users to fine-tune privacy settings based on semantic signatures. It also debunks the idea that increased privacy support is at odds with information sharing.

---

<sup>2</sup><http://pleaserobme.com>.

### 3 Data and Semantic Signatures

For the analysis and examples used in this paper we accessed POI data from Foursquare's public facing application programming interface (API).<sup>3</sup> A total of 908,031 randomly selected Foursquare venues<sup>4</sup> were accessed, each categorized into one of 421 Foursquare-defined place types. These types are hierarchically organized into three levels, e.g., Arts and Entertainment > Movie Theater > Indie Movie Theater. Analyzing attributes of these POI and aggregating them to the type level allows us to derive *semantic signatures* (Janowicz 2012). Semantic signatures use digital footprints emitted from humans such as terms that are associated with certain place types, times at which places of a given type are typically frequented, and so forth.

To construct *temporal bands*, each POI in the dataset was accessed every hour for 4 months starting in October 2013. The number of *check-ins* was recorded and cleaned allowing for a *popularity* distribution to be calculated through aggregating data to the place type level. To further strengthen the temporal bands, the 4 months of check-ins were distilled down to hours of the day over the course of a single week. This produced an array of 168 temporal bands (24 h × 7 days). These bands can be further aggregated into coarser resolution bands which are discussed in Sect. 4.2.

*Thematic bands* are constructed from the unstructured textual content provided as *tips* by people that have visited POI. Tips are essentially reviews that a visitor uses to describe or comment on a place. All tips were accessed for each POI in the Foursquare venue dataset mentioned previously. The tips were combined based on place type, stemmed, and cleaned (punctuation and stop words were removed). To ensure robust data signatures, only those place types with 30 or more tips were included in this textual analysis. *Latent Dirichlet allocation (LDA)* (Blei et al. 2003) was used to mine topics from the text and assign probabilistic topic distributions to each of the place types. LDA analyzes documents (aggregate of tips by place types in this case) and extracts topics based on the co-occurrence of words. This allows place types to be described as a distribution of topics extracted from the textual content contributed by individuals to those place types. We call these topic distributions *thematic bands*. In this work, 200 topics (thematic bands) are used.

*Spatial bands* are developed by exploring the geospatial patterns within the POI data. A number of different approaches are used to create these bands. Spatial descriptive statistics such as *Ripley's K* function are used to estimate the deviation of POI place types from spatial homogeneity. In previous work these place type functions have been *binned* by distance and combined with other spatial dispersion techniques such as *Average Nearest Neighbors (ANN)* and *Voronoi place-type variance* to produce a range of spatial bands (McKenzie et al. 2014).

For the purposes of this research, further investigation into the role of semantic signatures in location privacy focuses specifically on examples in the greater Los Angeles region. The boundary of this region was determined through the 2014 U.S.

---

<sup>3</sup><https://developer.foursquare.com/docs/venues/search>.

<sup>4</sup>*Venue* in this case is the Foursquare-specific term for Point of Interest.

*census urban areas* dataset and the boundaries of 240 neighborhoods within this region were ascertained from the 2014 *census designated places* dataset.

## 4 Indicativeness of Digital Footprints

In this section, we present a number of ways that information shared by an individual could be used to expose her location. A multidimensional approach is outlined exploiting the spatial layout of POI, the unique temporal popularity distributions of place types, and the thematic structure that can be extracted from text. The impact of each group of semantic bands is discussed individually and implemented as a whole in Sect. 5.

### 4.1 Spatial Indicativeness

To start with an illustrative example, imagine a user publishing content via her favorite social networking application, stating that she is at a *Mexican restaurant* in neighborhood  $N$ . We assume for the purposes of this research that we have access to a complete POI gazetter for the greater Los Angeles region (e.g., Foursquare venue set).

If  $N$  is *East Los Angeles*, the probability of determining her location is quite low compared to other neighborhoods (Fig. 1a). East Los Angeles has one of the highest ratios of Mexican restaurants to all other POI types in the region, namely 50 out of 809 (0.062). In comparison, the probability of randomly selecting a Mexican restaurant in *Beverly Hills* (Fig. 1b) is merely 4 out of 900 (0.004).



**Fig. 1** Mexican restaurants compared to all POI in two greater Los Angeles neighborhoods. **a** East Los Angeles. **b** Beverly Hills

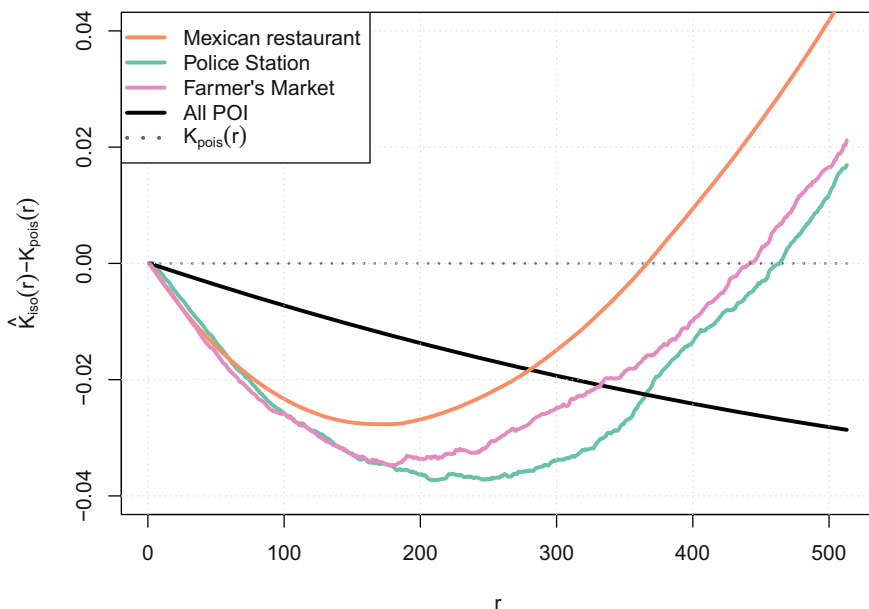
Consequently, knowing that a user is at a Mexican restaurant and in a specific neighborhood significantly impacts the ability to locate this individual. With access to a public POI dataset, the above example shows just how different two neighborhoods are with regards to platial privacy. In other words, the same place type can be revealing in one neighborhood, while it does not expose the user’s likely location in another neighborhood.

If an individual were to state the name of the establishment, e.g., indicate that she were at the chain restaurant *Chipotle Mexican Grill*, this would further increase the probability of determining her exact location within *Beverly Hills*. In this case, two of the four Mexican restaurants in Beverly Hills belong to the chain and therefore have the same name. In comparison, in *East Los Angeles*, no two Mexican restaurants have the same name. Thus, any indication of the place name on the part of the user immediately identifies her location to the place *instance* level.

Given the hierarchy of place types introduced in Sect. 3, we can increase location privacy by simply moving one level up in the place type hierarchy. For example, in the Foursquare place type vocabulary, *Food* is the category into which *Mexican Restaurant* is assigned (along with numerous other restaurant types, grocery stores, etc.). Comparing the number of POI categorized as *Food* to all POI in the dataset, the ability to locate someone in Beverly Hills based purely on place types drops considerably from 4 out of 900 POI (Mexican Restaurant) to 163 out of 900 (Food). Of the 240 neighborhoods in the greater Los Angeles region, Beverly Hills drops from 4th to 193rd with regards to its ability to locate someone based on place type. *East Los Angeles* on the other hand drops to a ratio of 0.234 (189 out of 809). This signifies a substantial decrease in identifiability, but not to the same extent as in Beverly Hills. Table 1 shows a sample of LA neighborhoods along with ratios for

**Table 1** A sample of neighborhoods in Los Angeles showing total POI within each neighborhood along with ratios for four different place types at two different levels in the place type hierarchy

Neighborhood	POI count	Mexican restaurant	Food	Museum	Arts and entertainment
Redondo Beach	948	0.014	0.217	0.000	0.023
Inglewood	998	0.025	0.200	0.000	0.024
Monterey Park	1,085	0.007	0.190	0.001	0.013
Torrance	2,731	0.011	0.168	0.001	0.017
Malibu	1,070	0.006	0.089	0.002	0.026
Santa Monica	1,443	0.016	0.243	0.001	0.038
Culver City	993	0.011	0.209	0.003	0.050
Stevenson Ranch	19	0.000	0.316	0.000	0.000
East Los Angeles	809	0.062	0.234	0.000	0.011
Beverly Hills	900	0.004	0.181	0.002	0.047
All POI	208,682	0.015	0.150	0.001	0.025



**Fig. 2** Plot of Ripley's K functions for three POI categories as well as all POIs in the greater Los Angeles region

*Mexican Restaurants* and *Museums* as well as their parent categories *Food* and *Arts and Entertainment* respectively.

The importance of spatial clustering within the POI dataset must also be considered. Simply knowing a place type and its prevalence within a region is valuable, but knowledge of the spatial distribution of the place type within the region may also lead to an increase in identifying a user's location. For example knowing that an individual is located at a place type that is highly clustered in a region minimizes the time necessary to find them (e.g., search and rescue operation).

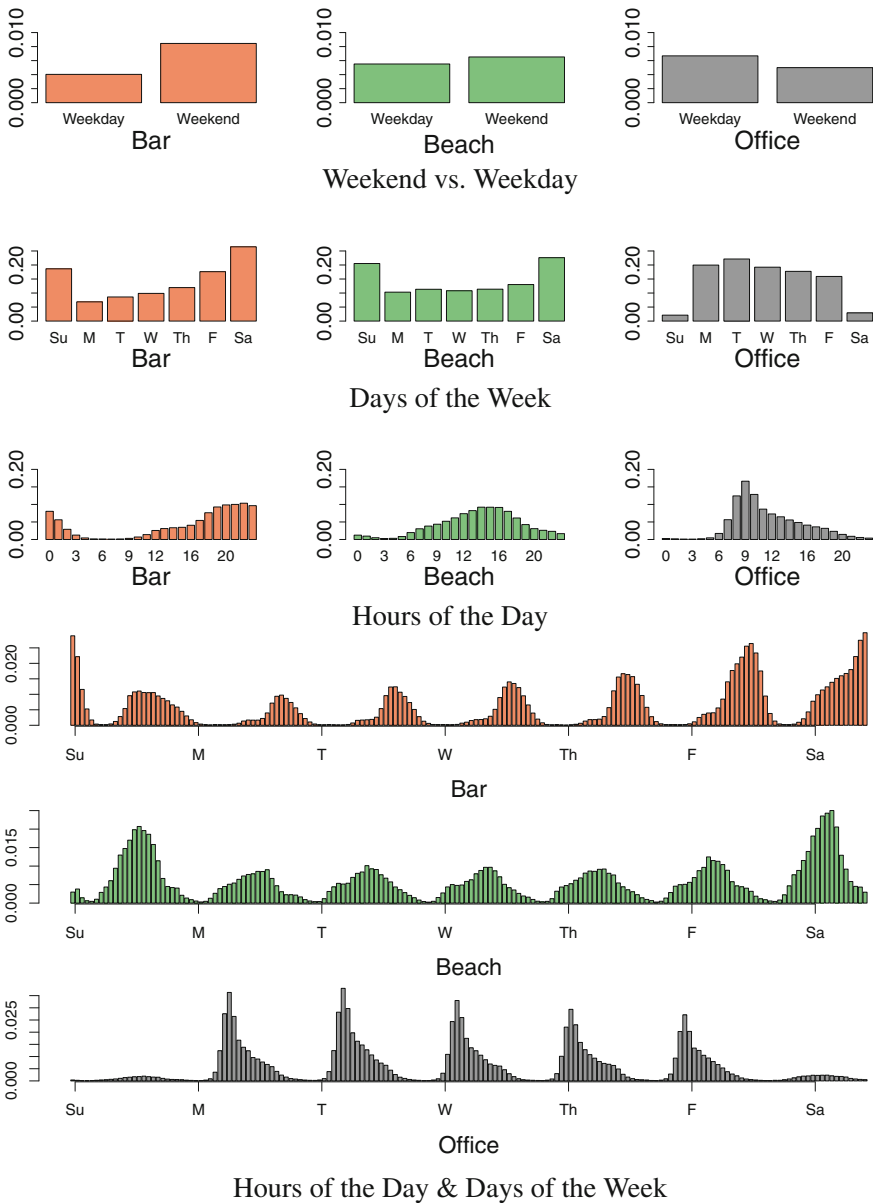
Figure 2 depicts Ripley's K statistics (Dixon 2002) for three place types as well as all places of interest in the Los Angeles. It shows the deviation from spatial homogeneity (shown as the dashed gray line in this figure). Naturally, place types such as Mexican restaurants show stronger clustering at a smaller distance than police stations or farmer's markets. Other methods for assessing the spatial indicativeness of a geospatial dataset have also proved valuable, including spatial entropy (Batty 1974).

## 4.2 Temporal Indicativeness

By way of another example, let us assume that an individual chooses not to publish the place type of the location but rather the *time* at which she is visiting a specific neighborhood  $N$ . Previous research has shown that time is highly indicative of



the types of places that people visit (McKenzie and Janowicz 2015). As one might expect, it is highly unlikely that someone posting from Los Angeles at 5 am on a Monday is at the *Department of Motor Vehicles*. Similarly, one is less likely to locate someone at a nightclub at 9 am on a Monday.



**Fig. 3** Temporal bands aggregated to different granularities and split by three example place types

Using the temporal bands we can probabilistically estimate an individual's location given a specific time. These probabilities can work at multiple levels of granularity. Figure 3 shows temporal signatures for three different place types with increasing levels of temporal granularity. Consulting the values in this Figure, an individual that is very precise in mentioning the time in an online post, e.g., 9 pm on a Friday night, would be more likely to be found at a *bar*, then at an *office building*. These bands can be aggregated based on the level of temporal granularity published. Say an individual solely mentioned the time of day, e.g., 9 am, and not the day of the week, then this method would return *office building* as the most probable place type.

Unsurprisingly, different temporal bands offer different amounts of information about the platial location of an individual. For instance, someone who only mentions 5 am on a Monday when publishing content is unlikely to be at *Department of Motor Vehicles*. Realistically, the probability of this person being anywhere except at home is rather small. On the other hand, if this person were to mention 6 pm on a Friday there is a much wider range of places this person could be given the activities that are possible at this time. To put it more formally, each temporal band can be defined by the unpredictability of the place types one might visit, which can be represented through *Information Entropy* (Claude E 1948). 5 am on a Monday has relatively low information entropy when compared to 6 pm on a Friday, given that one could more easily predict the place type of an individual in the first case, namely in some form of accommodation. Information entropy ( $E_T$ ) is defined in Eq. 1 where  $p_i$  is the probability of a given temporal band.

$$E_T = - \sum_i p_i \log_2(p_i) \quad (1)$$

Previous work (McKenzie et al. 2014) explored the amount by which the hourly temporal bands are unpredictable. Computing entropy across check-ins to all POI in the dataset showed that there is a statistical difference in the information that is presented between the hourly temporal bands (Table 2). This is important as the ability to determine the place where someone is can drastically increase depending on the time that she publishes content.

**Table 2** Information entropy for five lowest and five highest temporal bands

Low entropy			High entropy		
Day	Hour (AM)	Entropy	Day	Hour (PM)	Entropy
Monday	05:00:00	4.76	Thursday	07:00:00	5.97
Monday	04:00:00	4.87	Tuesday	07:00:00	5.96
Tuesday	04:00:00	4.93	Friday	06:00:00	5.95
Thursday	04:00:00	4.95	Friday	07:00:00	5.94
Tuesday	03:00:00	4.99	Saturday	12:00:00	5.93

### 4.3 Thematic Indicativeness

The words and language that people use when talking about the activities are indicative of the type of place they are doing the activity. Previous work in this area has shown that non-geographic terms and phrases can be geospatially indicative (Adams and Janowicz 2012; Mahmud et al. 2014). The results show that words in the English language can be tied to some region on the planet with varying levels of probability.

The thematic bands introduced in Sect. 3 define each place type in the Foursquare dataset as a distribution across topics. In short, the place types are defined by the language of the people that have visited them. Three examples of topics extracted from the unstructured natural language of the Foursquare tips are shown in Fig. 4 as word clouds of the topic’s most prevalent terms.

Using these thematic bands as the foundation, we use an LDA inference approach (McCallum 2002) to infer a distribution of these same topics for any new unstructured text-based document. For example, given content such as,

So glad I made it in to deposit my check at the ATM before they closed.

We, as humans, likely infer that the user is at a bank. From a computational perspective, an LDA model would need to construct a topic distribution for this text that would likely place a high probability on the topic related to banking (Fig. 4b), low probability on the topic related to Mexican food (Fig. 4a) and somewhere in the middle for the non-place type topic (Fig. 4c). It is also likely that the *bank* place type follows a very similar topic distribution to the topic distribution of the sentence above. *Jensen-Shannon distance (Jsd)* (Lin 1991) (Eq. 2) is used to measure the dissimilarity between our newly created topic distribution ( $P$ ) and each of the topic distributions for all 421 place types ( $Q$ ). *KLD* (Eq. 3) represents the *Kullback–Leibler divergence* and the lowercase  $d$  in *Jsd* signifies *Distance* instead of *Divergence*.  $M$  is equal to  $\frac{1}{2}(P + Q)$ . The smaller the dissimilarity value (bounded between 0 and 1), the more likely it is that our example content can be assigned to that place type. In this simplified example, the sentence above shows the least dissimilarity with the *bank* place type, and thus the user is said to be most likely at a bank. An implementation of this model is discussed in further detail in Sect. 5.

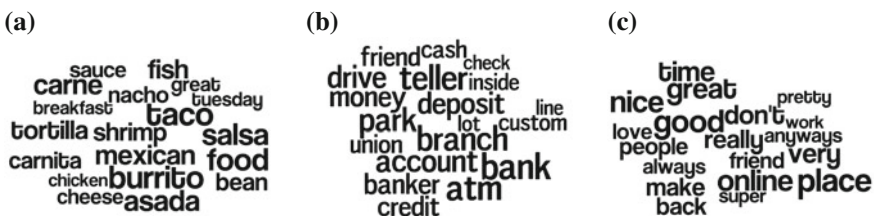


Fig. 4 Three example topics represented as word clouds of their most prevalent terms. **a** Terms related to Mexican food. **b** Banking related terms. **c** Non-place type specific terms

$$JSd(P \parallel Q) = \sqrt{\frac{1}{2}KLD(P \parallel M) + \frac{1}{2}KLD(Q \parallel M)} \quad (2)$$

$$KLD(P \parallel Q) = \sum_i P(i) \log_2 \frac{P(i)}{Q(i)} \quad (3)$$

## 5 Implementation: A Use Case

In the previous sections, we discussed the various bands of semantic signatures and the ways in which these bands contribute to determining the place where someone is. In this section, we bring the bands of the semantic signatures together to implement one approach that determines a user's place. An example use case is introduced, and the parameters are altered to show how sensitive the model is to changes. A first implementation of a formula is introduced to quantify the place-based privacy implications of the content.

### 5.1 Thematic Content

To start, let us imagine that an unknown individual publishes some small amount of unstructured content, e.g., a tweet. In this first iteration of the example, the content is both thematic and spatial but does not include any temporal property.

Excited for chicken tacos and delicious salsa in Beverly Hills. (1)

After stemming, a topic distribution for the text is inferred through an LDA topic inferencer based on the topic distributions (200 topics) learned from the 421 place types (thematic bands). A *JSd* dissimilarity value is then computed between the topic distribution for this text and each of the place type topic distributions. Note that this example uses a very small amount of text, so the inference model has a limited amount of data on which to infer the topic distribution. A greater amount of data would arguably lead to more accurate results. The top 10 least dissimilar place types are shown in Table 3.

The place types listed vary in their specificity. *Taco place* is a sub type of *Mexican restaurant* while *building* is a very generic place type. To put it another way, the descriptive content contributed as tips about taco places are narrower in their theme than the building place type which might include a wide range of themes related to places that exist within a building, e.g., restaurant types or car mechanics. Equation 4 shows how the thematic property of a place type ( $PT_{Theme}$ ) is quantified. Note that this function simply converts the dissimilarity value into a similarity value (higher value = better match).

$$PT_{Theme} = 1 - PT_{JSd} \quad (4)$$

**Table 3** Top 10 place types that are least dissimilar from the sample content (Quote 1)

Place type	JSd dissimilarity value
Mexican restaurant	0.267
Taco place	0.268
Food	0.301
Bar	0.302
Restaurant	0.309
American restaurant	0.317
Building	0.321
Miscellaneous shop	0.321
College cafeteria	0.329
Food and drink shop	0.330

## 5.2 Spatial Constraints

From a regional or *spatial* perspective, the content in Quote 1 indicates that the publisher is in *Beverly Hills*. We know from our gazetteer of places that there are four *Mexican restaurants* within the neighborhood boundary. Making the assumption that there is a certain region around an individual's point location that they can sense (e.g., visually, auditory), we construct a grid over a region. We expect that one would be able to locate something or someone reasonably quickly within this region. Provided this assumption, we overlay a  $500 \times 500$  meter cell grid over the Beverly Hills neighborhood in Los Angeles. Recording the presence or lack thereof of POI in each grid cell we find 115 out of 118 grid cells contain at least one POI. Of these, 2 grid cells contain at least one Mexican restaurant producing a ratio of  $2/115$  or 0.017.

Through these two data dimensions we are able to first determine the place type of the user and building off this constraint, spatially restrict the location possibilities. Using a rudimentary cell-based clustering technique we can further restrict the expected spatial locations of a content publisher.

## 5.3 Spatial Change

Building on the content of Quote 1, let us imagine that instead of sharing *Beverly Hills* as her location, this person mentions *East Los Angeles*. The textual content remains the same, so we have still determined that *Mexican restaurant* is the probable place type, but in this case, the number and spatial layout of place instances matching this criteria has changed. Overlaying the same  $500 \times 500$  meter cell grid over *East Los Angeles* we find that 112 out of 136 cells contain at least one POI and of these cells, 36 contain at least one Mexican restaurant resulting in a ratio of 0.321. So while the place type remains the same, the difference in spatial layout of these two

**Table 4** Effort values for two neighborhoods, Beverly Hills and East Los Angeles

Neighborhood	Mexican restaurant cells	Ratio	Total cells	Effort value ( $\times 10^4$ )
Beverly Hills	2	0.017	136	2.5
East Los Angeles	36	0.321	118	979.3
Greater Los Angeles region (Full area)	2,328	0.088	98,461	20.8

The Greater Los Angeles region is shown for comparison

neighborhoods means that there is a substantially lower chance of someone locating the user in East Los Angeles compared to Beverly Hills.

While the ratio is informative, the raw cell count is important here as well. Tasked with finding the publisher of the content a user would have to travel to 36 different regions (cells) in East Los Angeles but only 2 in Beverly Hills. Stepping back to the entire greater Los Angeles region, there are 98,461 cells that overlap neighborhood boundaries, and of these, 26,311 contain POI. Of the cells containing at least one POI, 2,328 contain at least one Mexican restaurant, producing a ratio of 0.088. Taking this ratio by itself implies that on average it is harder to locate someone at a Mexican restaurant in East Los Angeles than in the greater Los Angeles area overall. Though in this case, one would have to travel to 2,328 different regions (cells) in order to find the content publisher.

A relative *effort* value bounded between 0 and 1 is proposed by multiplying the number of likely cells by the ratio and dividing by the total possible set of cells over the regions. Table 4 lists the resulting effort values for the neighborhoods previously discussed.

## 5.4 Content Change

Again, let us slightly alter the published content and observe the implications on location privacy. Keep in mind that the actual location of the user (Beverly Hills) and activity (eating Mexican appetizers) remains the same. If instead of posting about the specific type of appetizer, the user generalizes her content as shown in Quote 2, what impact does this have on our ability to locate her?

Excited for great chicken appetizers in Beverly Hills. (2)

A topic distribution for this new content is again inferred from the existing LDA topic model and JSD is used to calculate the dissimilarity between this topic distribution and all place type topic distributions. The top ten least dissimilar place types are shown in Table 5.

Importantly, *Mexican restaurant*, presumably the place type the user is currently enjoying their food, appears nowhere in the list. The best match is instead, *food*, which is the parent category of *Mexican restaurant*, as well as many other place

**Table 5** Top 10 place types that are least dissimilar from the sample content (Quote 2)

Place type	JSd dissimilarity value
Food	0.263
Restaurant	0.268
American restaurant	0.275
Miscellaneous shop	0.276
Cafeteria	0.287
Cafe	0.305
Building	0.310
Assisted living	0.312
College cafeteria	0.313
General entertainment	0.322

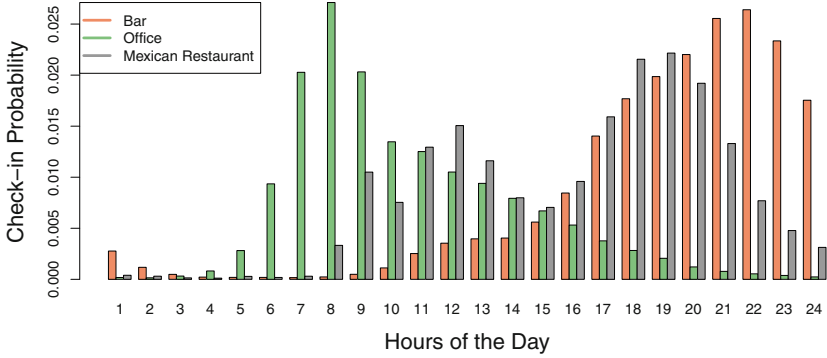
types. Instead of 4 possible locations in Beverly Hills, we are now faced with 163 possible locations. At least one *food* location exists in 44 of the 112 cells leading to a ratio of 0.393 and an *effort value* of 0.127. A similar adjustment is seen in East Los Angeles and for the greater Los Angeles region overall. Note that the broad activity of going out for food, even more specifically, appetizers, has not been lost through adjusting the text. By simply publishing a more generic term as part of her content, the publisher decreased her ability to be found in Beverly Hills dramatically.

## 5.5 Temporal Baseline

In addition to the textual and regional content specified in the examples above, one could imagine that someone might also tag their post with some type of temporal information. For example, a user might add the time *Friday at 7 pm* (e.g., as a meeting time) to the text.

In this example, the time is reported to a high granularity, permitting us to employ the 168 band temporal signatures in determining the place type probability. Taking the temporal signatures for each place type, we can directly compare the probabilities for Friday (Fig. 5) at 7 pm. For the purposes of this example, we have reduced our set of 421 place types to the three shown in this figure. Of these three, *Mexican restaurant* is the place type showing the highest probability at this time. Based on this information alone, we make the assumption that the user is at a *Mexican Restaurant* in Beverly Hills. This is in agreement with our text-based topic analysis discussed in Sect. 5.1.

This is not the entire story, however. While *Mexican restaurant* shows the highest temporal probability at 7 pm on a Friday, visually, it is followed quite closely by *bar* (Fig. 5). Computationally we can quantify this concern by referencing the *information entropy* for the hourly temporal signatures (a sample is shown in Table 2). Friday at 7 pm lists the fourth highest entropy value. The high entropy of this band tells us that in general, at 7 pm on a Friday night, people tend to be at quite a range of



**Fig. 5** Hour resolution temporal bands for *Bar*, *Office* and *Mexican Restaurant* on Friday

place types. Conceptually, this makes sense as this is the start of the weekend, and people could be engaging in a range of activities (e.g., watching a movie, at a bar, eating dinner, etc.). Knowledge of this high entropy reduces our certainty in determining the place type of the user and therefore has an impact on our overall ability to establish the spatial location of the user. The influence of temporal bands can be quantified using Eq. 5, where  $PT_{tp}$  represents the temporal probability of the given time band,  $max(tp)$  is the maximum temporal band value, and  $PT_E$  is the information entropy of the given time band.

$$PT_{Time} = PT_{tp}/max(tp) \times W + (1 - PT_E/max(E)) \times (1 - W) \quad (5)$$

If we set the weight component  $W$  equal to 0.5 and assume a time of 7 pm on Friday, *Mexican restaurant* produces a  $PT_{Time}$  value of 0.382, while *Bar* lists a value of 0.345. Importantly, the information entropy values remain the same in this case. This allows us to compare place types across different temporal bands.

What would happen if instead of Friday at 7 pm, the user tweets out her message 1 h later? The information entropy for 8 pm on a Friday is 5.852 (compared to 5.932 at 7 pm). The order of temporal probabilities has shifted as well with *bar* now slightly more probable than *Mexican restaurant*, 0.022 and 0.019 respectively. These changes lead to revised  $PT_{Time}$  values for the two place types. *Mexican restaurant* has dropped to 0.351 while *Bar* has risen to 0.389. Though minute, a 1 h adjustment has had a significant impact on determining the place type. At 8 pm on Friday, the temporal bands now indicate that the user is likely at a bar.

## 5.6 A Combined Approach: Thematic and Temporal Bands

We now need to combine the two values calculated through referencing the thematic and temporal bands into a single value which indicates the most likely place type for



**Table 6** Statistical approach to determining place type based on temporal and thematic bands

Time	Thematic		Temporal		Combined	
Place type	<i>Mex</i>	<i>bar</i>	<i>Mex</i>	<i>bar</i>	<i>Mex</i>	<i>bar</i>
Friday 7 pm	0.733	0.607	0.381	0.351	<b>0.558</b>	0.542
Friday 8 pm	0.733	0.607	0.345	0.389	0.521	<b>0.543</b>

the user. In the case of Friday at 7 pm, both the temporal band and thematic band indicate that the user is likely at a *Mexican restaurant*. One hour later offers a different perspective with the textual content indicating a *Mexican restaurant* and the temporal component suggesting a *bar*. A single value can be calculated through Eq. 6. Note that the equation gives the option of weighting one component over another.

$$PT_{Prob} = PT_{Theme} \times W + PT_{Time} \times (1 - W) \quad (6)$$

With equal weights of 0.5, Table 6 shows the resulting place types depending on time and theme. The thematic properties of both Mexican restaurant and bar remain the same across time, while the temporal properties change based on the values computed in Eq. 5. The combined value is calculated through Eq. 6. Not surprisingly, the results suggest that the user is likely at a Mexican restaurant on Friday at 7 pm, since both the thematic and temporal values agree. More interestingly, at 8 pm, this method determines that the user is slightly more likely to be at a *bar*, even though the content suggests that she is likely to be at a *Mexican restaurant*.

## 6 Conclusions and Future Work

In this work we discuss the use of semantic signatures for exposing location information about a user through the content that she publishes. These semantic signatures, described through various spatial, temporal, and thematic bands mined from user-generated geosocial content, have shown to be an important basis on which the place type of an individual's location can be determined. Despite omitting or masking geographic coordinates, the methods presented in this work show that a person's location can still be revealed through comparing the signatures to non-geotagged content published by an individual. We propose a method to compute the location indicativeness of the signatures, i.e., the ability to locate somebody based on their published content.

Our initial findings suggest that protecting a user's geographic coordinates and other potentially revealing characteristics, such as ethnicity, is not sufficient as everyday digital footprints can give away the user's location as well. These findings, for instance, could be used to develop mobile applications that helps users, e.g., political activists, to make small changes to their content in order to better protect their geo-privacy.

Future work in this area will focus on expanding the range of semantic signatures. For example, the data collection for check-ins is currently being expanded to look at yearly data with the goal of exploiting seasonal effects on place type check-ins. Furthermore, hyperlocal data such as events could be used to enhance the robustness of these signatures. In addition, we hope to expand this work into a prototype application or browser plug-in that reports on the level of location privacy that is attainable based on the content as well as spatial and temporal information that someone publishes.

## References

- Adams B, Janowicz K (2012) On the geo-indicativeness of non-georeferenced text. In: ICWSM, pp 375–378
- Almuhimedi H, Schaub F, Sadeh N, Adjerid I, Acquisti A, Gluck J, Cranor LF, Agarwal Y (2015) Your location has been shared 5,398 times!: a field study on mobile app privacy nudging. In: The 33rd annual ACM conference on human factors in computing systems (CHI). ACM, pp 787–796
- Armstrong MP, Rushton G, Zimmerman DL (1999) Geographically masking health data to preserve confidentiality. *Stati Med* 18(5):497–525
- Batty M (1974) Spat Entropy. *Geogr Anal* 6(1):1–31
- Beldad A, Kusumadewi MC (2015) Heres my location, for your information: the impact of trust, benefits, and social influence on location sharing application use among indonesian university students. *Comput Hum Behav* 49:102–110
- Benisch M, Kelley PG, Sadeh N, Cranor LF (2011) Capturing location-privacy preferences: quantifying accuracy and user-burden tradeoffs. *Pers Ubiquitous Comput* 15(7):679–694
- Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. *J Mach Learn Res* 3:993–1022
- Bohn J, Coroamă V, Langheinrich M, Mattern F, Rohs M (2005) Social, economic, and ethical implications of ambient intelligence and ubiquitous computing. In: *Ambient intelligence*. Springer, Heidelberg, pp 5–29
- Clarke CC (2015) A multiscale masking method for point geographic data. *Int J Geogr Inf Sci* 30(2):1–16
- de Montjoye YA, Hidalgo CA, Verleysen M, Blondel VD (2013) Unique in the crowd: the privacy bounds of human mobility. *Sci Rep* 3
- Dixon PM (2002) Ripley's K function. In: *Encyclopedia of environmetrics*
- Duckham M, Kulik L (2006) Location privacy and location-aware computing. *Dyn Mob GIS: Investigating Change Space Time* 3:35–51
- Dwork C (2011) Differential privacy. In: *Encyclopedia of cryptography and security*. Springer, pp 338–340
- Gambis S, Killijian MO, del Prado Cortez MN (2010) Show me how you move and i will tell you who you are. In: *Proceedings of the 3rd ACM SIGSPATIAL international workshop on security and privacy in GIS and LBS*. ACM, pp 34–41
- Goodchild MF (2015) Space, place and health. *Ann GIS* 21(2):97–100
- Hampton KH, Fitch MK, Allshouse WB, Doherty IA, Gesink DC, Leone PA, Serre ML, Miller WC (2010) Mapping health data: improved privacy protection with donut method geomasking. *Am J Epidemiol* 172(9):1062–1069
- Janowicz K (2012) Observation-driven geo-ontology engineering. *Trans GIS* 16(3):351–374
- Kounadi O, Leitner M (2015) Spatial information divergence: using global and local indices to compare geographical masks applied to crime data. *Trans GIS* 19(5):737–757
- Lin J (1991) Divergence measures based on the shannon entropy. *IEEE Trans Inf Theory* 37(1):145–151

- Lindqvist J, Cranshaw J, Wiese J, Hong J, Zimmerman J (2011) I'm the mayor of my house: examining why people use foursquare—a social-driven location sharing application. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, pp 2409–2418
- Mahmud J, Nichols J, Drews C (2014) Home location identification of twitter users. *ACM Trans Intell Syst Technol (TIST)* 5(3):47
- McCallum AK (2002) MALLETT: a machine learning for language toolkit. <http://mallet.cs.umass.edu>
- McKenzie G, Janowicz K (2014) Coerced geographic information: The not-so-voluntary side of user-generated geo-content. In: Eighth international conference on geographic information science
- McKenzie G, Janowicz K (2015) Where is also about time: a location-distortion model to improve reverse geocoding using behavior-driven temporal semantic signatures. *Comput Environ Urban Syst* 54:1–13
- McKenzie G, Janowicz K, Gao S, Yang JA, Hu Y (2015) POI pulse: a multi-granular, semantic signatures-based approach for the interactive visualization of big geosocial data. *Cartographica: Int J Geogr Inf Geovis* 50(2):71–85
- Seidl DE, Jankowski P, Tsou MH (2015) Privacy and spatial pattern preservation in masked GPS trajectory data. *Int J Geogr Inf Sci* 1–16
- Seidl DE, Paulus G, Jankowski P, Regenfelder M (2015) Spatial obfuscation methods for privacy protection of household-level data. *Appl Geogr* 63:253–263
- Shannon CE (1948) A note on the concept of entropy. *Bell Syst Tech J* 27:379–423
- Sweeney L (2002) k-anonymity: a model for protecting privacy. *Int J Uncertainty Fuzziness Knowl Based Syst* 10(05):557–570
- Weber RH (2010) Internet of things—new security and privacy challenges. *Comput Law Secur Rev* 26(1):23–30
- Zandbergen PA (2014) Ensuring confidentiality of geocoded health data: assessing geographic masking strategies for individual-level data. *Adv Med* pp 1–14
- Zhang S, Freundschuh SM, Lenzer K, Zandbergen PA (2015) The location swapping method for geomasking. *Cartography Geogr Inf Sci* 1–13