


Platial k -Anonymity: Improving Location Anonymity Through Temporal Popularity Signatures

Grant McKenzie ✉ 🏠 

Platial Analysis Lab, McGill University, Montréal, Canada

Hongyu Zhang ✉ 🏠 

Platial Analysis Lab, McGill University, Montréal, Canada

Abstract

While it is increasingly necessary in today's digital society, sharing personal location information comes at a cost. Sharing one's precise place of interest, e.g., Compass Coffee, enables a range of location-based services, but substantially reduces the individual's privacy. Methods have been developed to obfuscate and anonymize location data while still maintaining a degree of utility. One such approach, spatial k -anonymity, aims to ensure an individual's level of anonymity by reporting their location as a set of k potential locations rather than their actual location alone. Larger values of k increase spatial anonymity while decreasing the utility of the location information. Typical examples of spatial k -anonymized datasets present elements as simple geographic points with no attributes or contextual information. In this work, we demonstrate that the addition of publicly available contextual data can significantly reduce the anonymity of a k -anonymized dataset. Through the analysis of place type temporal visitation patterns, hours of operation, and popularity values, one's anonymity can be decreased by more than 50 percent. We propose a platial k -anonymity approach that leverages a combination of temporal popularity signatures and reports the amount that k must increase in order to maintain a certain level of anonymity. Finally, a method for reporting platial k -anonymous regions is presented and the implications of our methods are discussed.

2012 ACM Subject Classification Security and privacy → Privacy protections; Information systems → Location based services; Information systems → Geographic information systems

Keywords and phrases location anonymity, location privacy, geoprivacy, place, temporal, geosocial

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.9

Supplementary Material *Other (Code)*: <https://github.com/ptal-io/platial-k-anonymity>
archived at `swb:1:dir:d359af2244e4dc123d656fe613a9c2a3d3d6f985`

Funding Fonds de Recherche du Québec – Société et culture (Award Number NP-281897).

1 Introduction

In 2014, a student used time-stamped paparazzi photographs of celebrities exiting taxicabs in New York City (NYC) to identify their home locations using a supposedly anonymized dataset of taxicab trips [34]. This raised privacy concerns about the dataset [9] and forced the NYC Taxi & Limousine Commission to revisit their anonymization process and obfuscate trip origins and destinations in latter data releases. The lesson to be learned from this privacy debacle is that even though a dataset may have been anonymized, it does not exist in a vacuum. Rather, these data exist in a world where other information pertaining to the same subject may be available. Through these additional sources of information, one may be able to reduce the anonymity of the anonymized dataset. This is referred to as a *linkage-attack* [33] and the dilemma is that one likely does not know what additional sources of information exist, or will be created.



© Grant McKenzie and Hongyu Zhang;
licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 9; pp. 9:1–9:15

Leibniz International Proceedings in Informatics



LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Driven by the ubiquity of context-aware technologies, and the data they collect, we have seen a shift towards the development of computational approaches that leverage these data to model *places* [22, 26]. Through these approaches, photographs, audio recordings, temperature sensors, etc. are being used in combination with geographic data to provide more holistic representations of our environment and the places we inhabit. The irony is that the same data used to generate increasingly intricate models of the world can be used to violate one’s privacy and de-anonymize personal location information.

Within the privacy and anonymity domains, there have been considerable efforts on developing techniques that provide a trade-off between privacy preservation and data utility. Driven by the needs of individuals, most privacy models parameterize that trade-off, permitting users to exhibit control over their privacy based on their personal comfort levels. One of the most popular privacy preservation method for individual data sharing is *k-anonymity* [32]. The objective of this approach is to anonymize a data point such that it cannot be differentiated from $k-1$ other data points. Within geographic domains, these data points tend to be locations. For a wide variety of reasons (see [2]), an individual may want to obfuscate their location by reporting a set of locations (including their own), rather than their actual position alone. *Spatial k-anonymization* was introduced to address a number of challenges unique to geographic content [1, 7].

Much of the existing methodological work on *k-anonymity* and location privacy research is designed to be domain agnostic. Researchers overwhelmingly approach locations as simple geometric objects. In real-world scenarios, however, these objects represent entities that have a variety of properties and relationships. Furthermore, these entities do not exist solely in this dataset, i.e., other sources of related information exist. In this work, we explore such a real-world scenario and demonstrate how the privacy guarantee of a spatial *k-anonymized* dataset can be violated through the inclusion of external data. The real-world scenario of interest to us, is the process of sharing one’s location. This is a process that happens millions of times a day as people *check-in* to a location through social media, share their favorite restaurant with friends, or tag their location in a photograph. In these scenarios, *location* refers not to one’s geographic coordinates but rather the *place* that one is visiting, e.g., Mel’s Diner. The dilemma is in the trade-off between preserving privacy and sharing location data to gain utility. While I may be content to publicly share my visit to a trendy restaurant I may not wish to disclose the location of my teenager with anyone other than immediate family. It may still be useful, however, for my teenager to share anonymized location information, such as a set of k possible places, in order to receive recommendations for events nearby, for example.

The complexity of using places in a *k-anonymity* model is that an extraordinary amount of information is publicly available about most places, information that can be used to reduce the anonymity of someone sharing their platial location. In this work, we leverage the fact that different types of places have different visiting behavior and different hours of operation. For instance, people typically visit restaurants for lunch and dinner and more so on weekends than weekdays. The place types themselves also vary in popularity, regardless of time of day. For instance sports bars consistently receive more visitors than dentist offices.

While companies like Foursquare and Google collect the opening hours, popular visitation times, and overall popularity of most places in the world, access to this volume of data is unrealistic for most. For this work, we aggregate such data to the level of place type (e.g., Coffee shop) instead of place instance (e.g., Compass Coffee on 14th St.) and demonstrate that even a sample of place instances aggregated to this level can significantly reduce the anonymity of a place in a *k-anonymized* spatial dataset. More specifically, we will address the following three research questions (RQ).

- RQ1** Does the ability to identify an individual's location within a set of locations increase if we know the time the individual visited the location? Specifically, we investigate the degree to which temporal visitation patterns (signatures) can be used to reduce the efficacy of the spatial k -anonymity technique.
- RQ2** Do all temporal popularity signatures have an equal impact on the de-anonymization of a k -anonymized spatial dataset? We compare three types of temporal patterns and popularity values to identify which of them has the largest impact on the anonymity of an individual. We then determine if a weighted combination of these temporal popularity signatures can outperform the individual signatures.
- RQ3** Given a set of weighted temporal popularity signatures, by how much must we increase the number of places (k) in order to maintain the same level of anonymity promised by a non-enhanced k -anonymized spatial dataset? Furthermore, if a set of places are reported as a geographic region, what impact does the increase in k have on the average *size* of the reported region?

2 Related Work

A large body of literature pertaining to computational approaches to location privacy and anonymity has been published over the past few decades. Computational science research has mostly approached this from a geometric perspective [14, 16] whereas human geographers have typically taken a more qualitative approach [38, 12].

The concept of k -anonymity was first proposed by Sweeney and Samarati in 1998 [28] and later formalized as a property of certain anonymized datasets. *Relational k -anonymity* was then proposed as an approach for database privacy and disclosure control. A table is said to be k -anonymized if each record is indistinguishable from at least $k-1$ other records [32]. Within relational k -anonymity, *generalization* is often applied to reduce the uniqueness of each record, thus preserving a level of anonymity. While k -anonymity was originally designed with anonymity of the individual (or record) in-mind, an extension, ℓ -diversity [18], was proposed with the objective of preserving the *sensitivity of the values* associated with the records or individuals. This is addressed by introducing ℓ "well-represented" sensitive attribute values in each anonymized group. Li et al. [15] discovered that in some cases (e.g., skewed distributions or similar attributes) ℓ -diversity is insufficient in privacy protection. As a result, they propose t -closeness [15] to overcome the limitation of ℓ -diversity.

Spatial k -anonymity incorporates location information into the discussion of anonymity and privacy preservation. While relational k -anonymity is static and often involves a single k , the spatial version was designed to be dynamic with variable k [6].

Existing research on this topic has leveraged spatial k -anonymity for the development of *k -anonymized spatial regions* that include an anonymized set of locations consisting of an anonymized user and at least $k-1$ other users [6, 23, 8]. Early work by Kalnis et al. [11, 10] developed a series of cloaking techniques (e.g., Hilbert cloak, center cloak) with the goal of reducing vulnerabilities in basic spatial k -anonymity algorithms. Additional efforts have introduced techniques that consider the temporal connectivity of location-based services [5]. To date, the majority of research from computational scientists has approached spatial k -anonymity through the introduction of spatial-temporal cloaking and tree-based spatial indices, predominantly focusing on the geometric properties of the data.

Aside from *spatial k -anonymity*, additional methods of *geomasking* have been developed to obfuscate location information. While not strictly anonymity approaches, these are typically categorized into aggregation-based or perturbation-based with aggregation methods

being similar to anonymized spatial region’s *spatial-temporal cloaking* but often leverage existing geographic units such as administrative boundaries [3], Voroni polygons [30, 25], or census tracts [17]. Others have built aggregation techniques based on geometric shapes or centroids [36]. *Perturbation* geomasking methods displace individual data points to nearby locations using random distance and direction and various kernels [36, 37]. Finally, efforts have been made to combine the two types of geomasking methods. *Adaptive areal elimination* first aggregates population polygons into anonymized spatial regions, then randomly displaces data points within the newly formed anonymizing regions [13]. Charleux and Schofield [4] proposed *adaptive areal masking* that replaces longest border shares in adaptive aerial elimination with Euclidean distance ranks.

In recent years we have seen a substantial increase in computational approaches to define and understand *places*. As geographic information science has evolved, researchers are not only exploring the Aristotelian view of place (i.e., objects in the Euclidean space), but also the Platonic view (i.e., relationships and experiences in the environment) [27]. A growing body of work has been examining and modeling the concept of place from multiple dimensions [35, 29, 22]. As increased availability of large heterogeneous datasets from a variety of sensors has allowed geospatial scientists to move from spatial studies to the multidimensional concept of place, so too have the geoprivacy and spatial anonymity domains.

3 Data

3.1 Temporal visitation, hours of operation, and place popularity

Two different data sources were used in these analyses. First, all of the place types (e.g., Bars, Parks, Police Stations) published by the local place recommendation service, *Foursquare*, were identified.¹ We randomly selected 20 places of interest (POI) from across the United States in each of the place types. The Foursquare application programming interface (API)² was used to request the number of check-ins to each of these POI every hour over the course of 3 months. These check-in counts were grouped by place type and aggregated to hour of the week producing a set of 168 (24×7) temporal signatures (T_F) for each Foursquare place type. Hours of operation were accessed from the API for each of the Foursquare POI in our sample. These data consist of a binary value for each hour of a typical week. As before, these were grouped by place type and aggregated (median) to the hour of the week producing an hours of operation signature (T_H). Foursquare also offers a popularity value for each POI which is computed based on foot-traffic and user ratings.³ Using the Foursquare API, we accessed the popularity values for each POI in our sample dataset, and averaged them by place type. This produced a mean popularity value, Pop , for each place type in the Foursquare dataset.

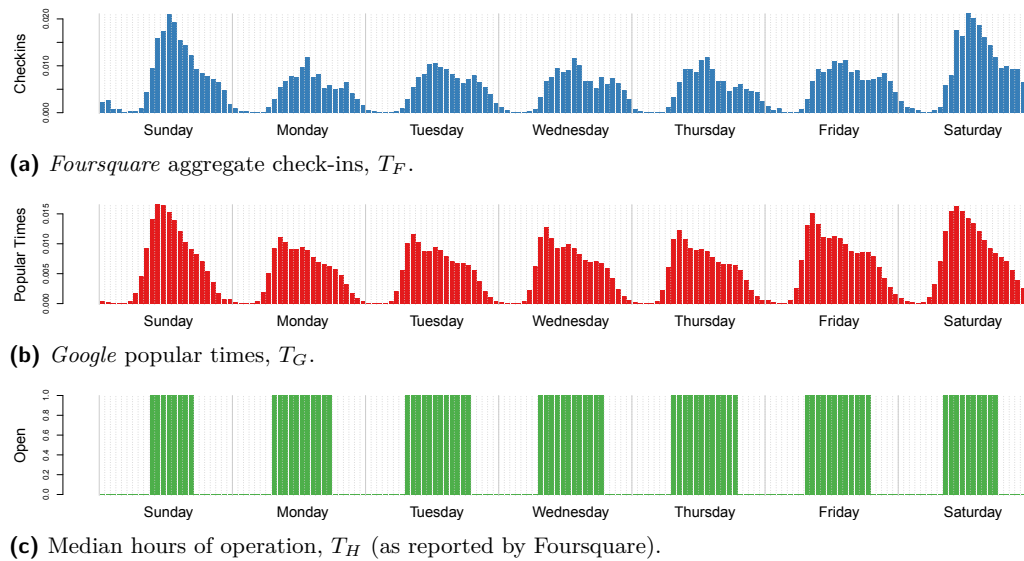
We then accessed popular times data for 185,600 *Google Places* POI across the United States.⁴ The popular times data are constructed through passive collection of location information accessed from the mobile devices of Google’s location service users. Similar to the process used for the Foursquare data, these popular times were groups by Google’s place

¹ A full list is available at <https://location.foursquare.com/places/docs/categories>.

² <https://developer.foursquare.com/>

³ <https://medium.com/foursquare-direct/tagged/engineering>

⁴ Data collection script available at https://github.com/apollojain/popular_times



■ **Figure 1** Example temporal signatures for the place type *Café*.

type⁵ (different from *Foursquare*'s) and aggregated by hour of the week. This approach produced a set of temporal signatures (T_G) for each *Google* place type. Finally, all three temporal signatures (T_F , T_H , T_G) were normalized individually producing a distribution of temporal values that sum to 1 (Figure 1). This normalization process was necessary so that each signature was evenly weighted at the start of analysis.

3.2 Place type alignment

Given the two sources of POI data, the first task was to align the place type schemas. We leveraged our previous work on this topic [20] to identify alignments between place types. The process involved collecting the same POI representations (e.g., the same restaurant) from *Foursquare* and *Google* via their APIs. POI matching was done by comparing the names and geographic distances between place representations. We took an overly conservative approach by only accepting matches for POI where there was an exact name match and the geographic distance was less than 100 meters. We then generated a matrix counting the occurrence of place type matches. The place types that had the largest number of POI matches were accepted as an alignment. For instance, *Foursquare* has a place type *Coffee Shop* while *Google* does not. Through our alignment process, we identified *Google*'s *Café* place type as a match. As a final step, we manually reviewed the alignment results and made minor adjustments to the place type alignments where appropriate.

3.3 Validation data

To validate our approach we required access to a large sample of data where an individual recorded their real-world visit to a location, including the time and place type they visited. While geosocial media *check-ins* are suitable for this task, access to a large and randomized

⁵ https://developers.google.com/maps/documentation/places/web-service/supported_types

sample is not possible directly through Foursquare or its gaming application, *Swarm*.⁶ Users of both Foursquare and Twitter, are able to connect their two accounts allowing them to publish their Foursquare check-ins on their public Twitter feed. Leveraging this knowledge, we used the public Twitter API⁷ to randomly sample 17,909,516 geotagged tweets within the continental United States between May 2017 and May 2022. The tweets were filtered to select only those whose source was Foursquare’s Swarm application. All of these tweets contained the information necessary to access a user’s geosocial check-in. Each check-in consists of the Foursquare POI name, place type, geographic coordinates, and timestamp of the visit. A total of 54,568 check-ins to 22,206 unique POI were identified after cleaning. To reduce POI bias we elected to only include one check-in (randomly selected) for each POI in our analysis.

Through the Foursquare API, we requested the closest set of POI to each of the 22,206 check-in POI. The API sets an upper limit of 50 POI per *Nearby* request. The maximum of 50 POI was not always returned, so in order to maintain a robust set of data, we removed all check-ins with fewer than 29 nearby POI from further analysis. This resulted in a final validation dataset of 19,478 check-ins to the same number of unique POI and a total of 584,340 nearby POI.

4 Analysis

Our first task in addressing RQ1 was to determine the degree to which one of our temporal signatures impacted our ability to identify an individual’s POI location from within a set of k POI. We will refer to an individual’s actual location as p_l , nearby locations as p_n , and the larger set of all 30 POI in a region as P_{30} , where $p \in P$. Each p has a *place type* and each visit to a p_l occurred at some time, reported as the hour of the week. Three temporal signatures and the popularity value were assigned to each p based on its place type.

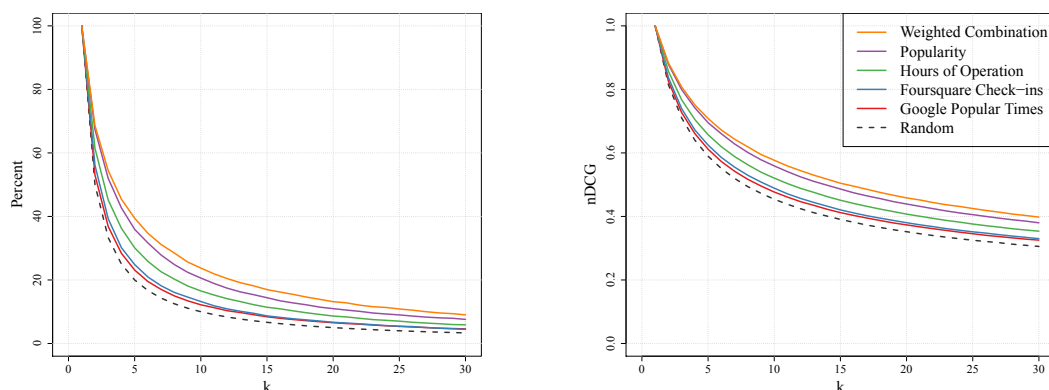
4.1 Spatial k -anonymization

We started with a baseline k -anonymized spatial dataset, one that includes p_l and a set of p_n nearby POI, but ignores the place type property or temporal signatures of each p . We set a range for k from 1 to 30 for our analysis. For each of the 19,478 check-in in our dataset, we selected a subset of the k closest POI (P_k) to p_l , including p_l itself. For instance, $k = 3$ means that P consisted of 3 p , including 2 p_n and our p_l . To determine the level of k -anonymity in our set, we randomly selected a p from the set of P_k . This was done for all 19,478 check-ins and all values of k . The average number of times p_l was correctly identified in P_k was recorded. The results are shown as the dashed black line in Figure 2a (the other lines will be discussed in Section 4.2).

Provided no other information on which to select a p from P_k , the results are random with the percentage equating to $1/k \times 100$. While informative, this method of only counting instances where p_l is correctly identified ignores position ranking. For example, a model that identifies p_l as the second most likely place is better than a model that identifies p_l as the 20th most likely place. This is irrelevant for the random model, but will play a role in assessing the temporal signature approaches. To account for differences in rank, we calculated the *normalized Discounted Cumulative Gain* (nDCG) (Equations 1 and 2). nDCG

⁶ <https://www.swarmapp.com/>

⁷ <https://developer.twitter.com/>



(a) Percentage of the time p_l is correctly identified in P_k .

(b) Normalized Discounted Cumulative Gain based on position of p_l in ranked P_k .

■ **Figure 2** Percentage of POI that were identified correctly, or where they ranked, using different approaches, shown as k increases.

considers the rank of a prediction by penalizing incorrect p_l selections at a \log_2 rate based on their position i in the ranking, where rel_i is the graded relevance of p in P_k . $IDCG_p$ is the idealized ranking where p_l is correctly identified in the first ranked position. The results of the nDCG assessment of the random spatial k -anonymity approach are shown in Figure 2b.

$$nDCG_p = \frac{DCG_p}{IDCG_p} \quad (1) \quad DCG_p = \sum_{i=1}^{|REL_p|} \frac{rel_i}{\log_2(i+1)} \quad (2)$$

4.2 Temporal signature enhancement

We then designed a method to reduce the anonymity of a k -anonymized spatial dataset through the inclusion of place type temporal signatures. Our answer to RQ1 depends on whether the ability to identify someone increases with the inclusion of this temporal dimension.

To start, we limited our analysis to include temporal check-in behavior as reported by Foursquare at the place type level. Remember that each p in our dataset has a place type, and each place type has a Foursquare temporal signature, T_F . The check-in time for each of our p_l was recorded and used to identify the temporal probability of an individual visiting a p based on the temporal signature. For example, the temporal probability at 20:00 on a Friday is higher for the *Restaurant* place type than *Bank*. Figure 3 represents these temporal probabilities as graduate symbols.

As before, a subset of POI, P_k closest to p_l were selected. The p in this subset were then ranked based on the visitation probability at the indicated time (temporal signature). Given that P_k may contain multiple p of the same place type, these p have the same temporal probability value. Order was randomized between places of the same type. The p with the highest temporal probability was flagged as the predicted location of p_l . This was done for all known POI visits in our dataset and the average accuracy was reported for each value of k both as the *correctly* identified p and the nDCG. These are shown as the blue lines in Figure 2.

The results of this analysis indicate that the inclusion of place type temporal signatures increases one's ability to identify an individual's location in a set of k -anonymized POI. Averaged across all selected values of k (1-30), the inclusion of Foursquare's temporal signatures, T_F , decreased the anonymity of p_l in P_k by 31%. As shown in Table 1 (column T_F), the percentage of de-anonymization increases with larger values k .



■ **Figure 3** Places of interest in a region shown with graduating symbology representing the temporal probability at 20:00 on Friday. The black star marker indicates the actual location of the individual. Base map by Carto.

■ **Table 1** Average percentage improvement in correctly identify an individual (p_i), above random selection from a k -anonymized spatial dataset. The three different temporal signature-based approaches are reported along with the popularity value method and the weighted combination of temporal popularity signatures, $TPop$.

k	T_F (%)	T_G (%)	T_H (%)	Pop (%)	$TPop$ (%)
2	12.2	6.1	12.0	34.7	37.2
5	24.2	15.7	27.6	79.5	97.6
10	32.1	21.8	35.0	106.0	137.5
15	30.4	26.2	38.5	116.2	154.4
20	33.6	30.8	40.4	119.4	163.2
25	36.0	34.5	41.8	125.0	171.5
30	38.2	38.0	43.4	127.7	172.3

4.3 Comparing temporal signature and popularity approaches

Knowing that a place type temporal signature can be used to decrease the anonymity of an individual in a k -anonymized spatial dataset, we compared temporal signatures from different sources as well as the atemporal place type popularity values.

4.3.1 Temporal signatures

Having developed a model based on Foursquare's temporal signature in the previous section, we conducted the same analysis for the Google popular times signatures T_G and the place type averaged hours of operation, T_H . As shown in Figure 2, ranking POI based on the probability of an individual visiting them at a given time improved the place prediction in all cases and for all values of k . In other words, location privacy was reduced through the inclusion of any temporal signature data. In comparing the results of analysis using different temporal signatures, T_G has the lowest impact, reporting an average decrease in anonymity of 26.3% across all values of k . Similar to T_F , the percentage increased with larger values of k . The T_H signatures produced the largest impact on anonymity with an average decrease of 35.3%. The percentage decrease in anonymity is shown for select values of k in Table 1.

4.3.2 Popularity

In addition to place type temporal signatures, the popularity of place types can also be used to reduce anonymity of a user's location in a set of POI. While the previous data signatures reported a relative change in visitation popularity over time, our popularity values, Pop , are atemporal and represents a comparison between place types, ranging from 0 (least popular) to 1 (most popular). These place type popularity values were assigned to their respective p in P_k and were ranked based on this popularity. We again randomly order p of the same place type within this ranking. As shown in Table 1, this approach results in a greater percentage of anonymity decrease than each of the temporal signatures alone. If we examine $k = 8$, for instance, there is a 1 in 8 (12.5%) chance of randomly selecting an individual's actual location in a spatial k -anonymized dataset. Through the inclusion of place type popularity, this doubles to 1 in 4 (25.0%). These results, along with those from the previous section address the first portion of RQ2, namely that all of these data signatures decrease anonymity by different amounts.

4.3.3 A weighted combination of signatures

In addition to assessing each of the temporal signatures and the popularity values independently, we also computed a weighted combination of the signatures. In addressing the second portion of RQ2, we question whether combining the signatures and popularity value will outperform, with respect to de-anonymization, each signature alone. The combined approach is shown in Equations 3 and 4. In our analysis, applied all combinations of weights, incrementing by 0.1 so that 285 combinations were applied to all temporal signatures and the popularity value. This was done for all 19,478 check-ins and all values of k between 1 and 30.

$$w_1(T_F) + w_2(T_G) + w_3(T_H) + w_4(Pop) \quad (3) \quad w_1 + w_2 + w_3 + w_4 = 1 \quad (4)$$

The results of this weighted approach, with all combinations of weights are provided in the project repository. The weight combination that produced the highest number of correct POI identifications, and highest nDCG, consisted of a weight of 0.3 for each of the temporal signatures and a weight of 0.1 for the average place type popularity. We refer to this weighted combination as the temporal popularity signature, $TPop$. On average, this approach decreased anonymity by 143.3% with exact values shown in Table 1. This is a substantial amount as compared to each of the temporal signatures and popularity independently.

We further investigated the results of this analysis by ordering all weighted combinations by their average accuracy across all values of k . Our top model of 0.3 for all temporal signatures and 0.1 for popularity values was ranked 1 out of 285 possible combinations. The first combination of weights to not include average place type popularity ($w_4 = 0$) was at rank 220. This suggest that the inclusion of popularity in our model is essential for a large decrease in anonymity, but that the actual weight is less important. Also of note, the best performing combination placed equal weight on each of the temporal signatures, indicating that each temporal signature represents a unique aspect of place visitation behavior and that all are needed in order to produce the best approach for de-anonymization of a k -anonymized dataset.

4.4 Platial k -anonymization

The results of the previous sections demonstrate that an attacker with access to temporal and/or popularity data reported at a place type level can considerably decrease the anonymity of an individual's reported location within a k -anonymized set of POI. The accessibility to,

and inclusion of, such contextual data requires that the number (k) of POI in a k -anonymized set be increased in order to guarantee the same level of anonymity promised by the original spatial k -anonymity model. In addressing RQ3, we establish these new values for k proposing that the values be labelled *Platial k* or k_p .

Through referencing the results of our analysis in Section 4.3, we can match accuracy percentages between a spatial k -anonymized (random selection) approach and our most accurate platial approach, *TPop*, taking the k value from our most accurate model as k_p . In other words, how many k_p are needed in order to guarantee the same level of anonymity that was promised by a k -anonymized dataset that assumed no temporal popularity data were available? Table 2 shows the value of k from a standard k -anonymized dataset along with the k_p values necessary to achieve the same level of anonymity using our temporal popularity signatures.

■ **Table 2** k number of POI along with the k_p number of POI needed to preserve k -anonymity given the temporal signatures, popularity values, or combination temporal popularity signature.

k	$k_p T_F$	$k_p T_G$	$k_p T_H$	$k_p Pop$	$k_p TPop$
2	3	3	3	4	4
5	7	6	7	11	13
10	14	13	14	23	29
15	21	20	22	>30	>30
20	28	28	29	>30	>30

For instance, in order to limit one’s exposure to a 20% chance of being randomly identified in a set of POI (the equivalent of a k -anonymity of 5), one would need to include 13 POI, or a k_p-1 of 12. As shown in Table 2, in some cases, the number of POI needed to preserve k_p -anonymity was greater than the 30 POI we had in each of our check-in sample sets. Using these results, we can report k_p as a function of k , namely $k_p = 2.54k + 0.04k^2 - 0.88$.

4.5 Reporting platial k -anonymity through geographic regions

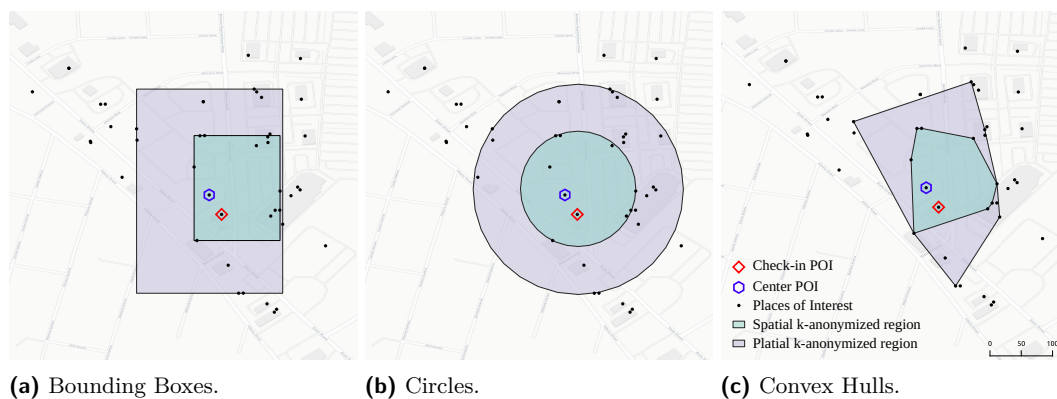
What do these results mean in practice though? The application of k -anonymity specifically deals with sets and spatial k -anonymity situates the elements of a set in geographic space. In real-world scenarios, anonymized spatial data are often reported through a location-based service as geographic regions, typically polygons that include the set of k -anonymized locations. Depending on the user’s privacy preferences, they set a large or small value for k which in turn determines the size of the reported polygon.

There are several ways to generate polygons that encompass a set of points. Here we identify geometric shapes based purely on the POI set, rather than political, social, or environmental boundaries. Such boundaries could also be used, but are not the focus of this work. The most common geometric shapes are a circle, bounding box, or convex hull. The centroid of these regions also varies. The simplest option is to set the centroid of the region on the known location and expand the radius or perimeter until k points are contained within the region. From an anonymity perspective, this approach falls victim to a *center-of-anonymized spatial region* attack, where an attacker would assume, given the geometry and centroid, that the actual location of an individual is the center most POI [11]. To avoid this, many current approaches [24] offset the centroid of the region by taking the n^{th} -nearest neighbor.

In generating a k -anonymized platial region, we have two options. One is *generalized* and involves simply referencing Table 2 or the k_p function to generate a polygon that contains k_p POI. This is a general approach as it uses the average k_p as reported through our analysis of

19,478 check-ins. While this can be used for any set of POI that contain place type attributes, platial k -anonymity can also be computed for an individual scenario. This is the *local* option. In this case, we assume the attacker has knowledge of the local region, knows the time someone visited a location, and has access to the place types of all POI. In this case, our set of P_k must include those that report a combined temporal popularity probability, $TPop$, greater than or equal to that of the actual check-in POI. The local platial k -anonymized region is the region that contains all of these POI. This may be better explained through an example. Let us set $k = 10$ and specify that our actual check-in POI has a $TPop$ probability value of 0.5. At a minimum, our platial k -anonymized region needs to include the 9 nearest POI with a $TPop$ probability greater than or equal to 0.5. Depending on the shape of the region, it may also include other POI with $TPop$ probability less than 0.5. All of these POI together sum to our local k_p .

Figure 4 shows examples of the various polygonal representations for a $k = 10$ anonymized set of POI as well as the k_p equivalent region. In these examples, the center point (blue hexagon), from which the shapes are determined, is the nearest neighbor to the actual *check-in* POI (red diamond).⁸ The geometries are generated by expanding the search radius from the center point until k POI are enclosed within the region. The smaller green regions show the minimum areas that encompass the specified k number of points (10 in this example), limited by the shape specifications. The larger purple regions represent the minimum areas that include P_k that are equal to or greater than the temporal popularity signature probability of the check-in POI at a given time. For a k of 10, k_p will always be at least 10.



■ **Figure 4** Polygonal representations of k -anonymity as well as the temporal popularity enhanced k -anonymity. These use a *local* approach based on the place types signatures of the actual POI.

For our sample set of 19,478 check-ins, we calculated the area of all three shapes that contain k and k_p POI. For all shapes and values of k , the areas of the platial k -anonymized regions are greater than the spatial k -anonymized regions. The difference in percentage decreases as k increases. The average percentage increase in area for k 1-20 ranges from 170.1% for a convex hull to 193.7% for a circle. Table 3 shows the median percentage increases in area. k is limited to 20 in this Table as we have seen that corresponding values of k_p can be considerably larger.

⁸ We use the first nearest neighbor here, but second or third could be used to increase privacy.

■ **Table 3** Median percentage increase in area between spatial k -anonymized regions and platial k -anonymized regions.

k	<i>Convex Hull</i> (%)	<i>Bounding Box</i> (%)	<i>Circle</i> (%)
2	0	520.9	845.7
5	394.0	343.6	324.0
10	105.9	96.9	120.9
15	73.9	67.4	105.5
20	32.7	37.8	56.8

5 Discussion

In the real-world, a geographic dataset does not exist in isolation. Additional information is available about all aspects of our lives, including the places that we visit. The times of day and days of the week that people interact with places in their environment follow patterns that can be discriminated at the categorical, or place type level. This knowledge can be leveraged and patterns can be used to estimate the locations of individuals. For the privacy-conscious among us, this is problematic. The spatial k -anonymity of a dataset states that an individual sharing a set of k places is guaranteed a level of anonymity.

In this work, we demonstrate that through the inclusion of place type temporal visitation patterns and popularity values, the presumed level of anonymity is violated. The results of RQ1 indicate that a place identification model built using publicly available temporal visitation signatures can significantly reduce the anonymity of a user sharing their location as a set of POI. Temporal signatures extracted from social media check-ins perform slightly better than those collected through passive data collection such as Google’s location services. Access to the average hours of operation for different place types outperform both of the activity-based temporal signatures. It is unclear exactly why hours of operation outperformed the behavior-based temporal signatures. One possible reason is that the hours of operation data were the least nuanced of the temporal data and by taking the median, the data were quite restrictive in reporting opening and closing times. It appears that for our sample of check-in data, these restrictive time periods were beneficial in predicting an individual’s location. By far the most useful information is the relative popularity of a place type. On average, access to these values substantially decreases the anonymity of an individual in a shared set of locations. This is worth noting as it suggests that the nuance of when a person visits a location, while important, is less important (on its own) than the overall, non-temporal popularity of a place. In identifying a weighted combination of these temporal signatures and popularity values (RQ2), we demonstrated that each of the different dimensions contributes to an improved model for de-anonymization. For instance, the probability of identifying an individual’s location out of a set of five POI ($k = 5$) is nearly 80% greater given access to popularity data and 100% greater using our weighted combination approach, compared to a model that did not include any additional data.

This equates to a meaningful decrease in individual privacy brought about by analysis of publicly available data. These signatures and popularity values are aggregated to the place type level, not the individual place instance, suggesting that they can be applied to k -anonymized POI datasets anywhere in the world. While research on temporal signatures has shown that roughly 50% of these temporal patterns vary regionally, some of the more common place types such as drug stores and restaurants, do not [19]. The results of our analyses demonstrate that k does not accurately represent the anonymity of a dataset given access to other sources of related data. To address this, we propose a *platial* value, k_p , that

represents the number of POI necessary to guarantee k -anonymity given an attacker may have access to these contextual data sources (RQ3). In this paper, we provide a reference for those developing place-based obfuscation applications, recommending a baseline k_p number necessary to ensure actual k -anonymity in a set of POI. Importantly, our proposed measure of k only assumes access to the three temporal signatures and one relative popularity set for a given set of place types. There are undoubtedly additional sources of information that can be used to further reduce the anonymity of a user sharing an anonymized dataset. In this work, we simply highlight some of the ways this can be done, and report the magnitudes of de-anonymization.

Our analysis reports that regions built from k_p -anonymized datasets are considerably larger in area than k -anonymized datasets. What was surprising was the dramatic increase in area reported on average. For instance, at $k = 5$, the average platial k -anonymized region was roughly 350% larger than the spatial k -anonymized region. Larger regions equate to a reduction in utility. While we argue that the anonymity of a user remains in-tact through our improved approach, the trade-off in utility must be acknowledged. All of this demonstrates a need for further critical investigation of how we choose to obfuscate location information.

The biases of the datasets used in this work must be mentioned. All of the data used in these analyses were contributed by individuals of geosocial media applications or a location service provider. While these data have been used in a wide variety of research, they do represent a biased subset of the population. Though check-ins were randomly sampled, the types of people that choose to check in and share their geographic locations are a unique subset of the population. They tend to be tech-savvy and predominantly live in urban areas. The data most often do not adequately reflect the activity patterns of the elderly, lower-income individuals, and those in rural communities. Any application or policy that uses the results of this work, should consider the biases and act accordingly.

A limitation of this work is the alignment of the two different place type vocabularies. Since Google and Foursquare use different terms and concepts to label their categories, alignment was necessary. As mentioned previously, the alignment was achieved through identifying co-occurrence of place instances. In some cases, a place type from one service would align with multiple place types from the other service. We took the place type that had the largest number of place instance matches, but sometimes the difference was a single POI. A manual check was done to ensure that the matches made sense, but any manual alignment introduces bias on the part of the person doing the aligning.

Future work in this area will involve the inclusion of additional contextual data such as the change in temporal behavior due to weather and local events. Our approach will be integrated with other efforts in the location privacy domain that leverage socio-economic, demographic, and mobility data. Additional efforts will be made in the application of this approach to real-world scenarios and privacy-preservation platforms, similar to projects such as *MaskMy.XYZ* [31] and *PrivyTo* [21].

6 Conclusion

In this paper, we identify some of the ways that the k -anonymity of an individual's reported location can be reduced by using existing publicly available place-based data. Specifically, our work shows that knowledge of place type temporal visitation patterns, average hours of operation, and relative popularity can substantially decrease the anonymity of one's location in a set of places of interest. Through analysis of 19,478 place check-ins we developed a platial k -anonymity approach that aims to improve anonymity, acknowledging that an attacker may have access to contextual information. Using this platial k -anonymized approach, we show that sets reported as geospatial regions must increase in area in order to preserve their

presumed degree of anonymity. Overall, this work demonstrates the need to be aware of the additional data that is increasingly available, publicly accessible, and can be used to reduce the anonymity of individuals sharing their seemingly obfuscated personal location information.

References

- 1 Charu C Aggarwal. On k -anonymity and the curse of dimensionality. In *Proceedings of the 31st International Conference on Very Large Data Bases*, pages 901–909, 2005.
- 2 Marc P Armstrong and Amy J Ruggles. Geographic information technologies and personal privacy. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 40(4):63–73, 2005.
- 3 Marc P Armstrong, Gerard Rushton, and Dale L Zimmerman. Geographically masking health data to preserve confidentiality. *Statistics in medicine*, 18(5):497–525, 1999.
- 4 Laure Charleux and Katherine Schofield. True spatial k -anonymity: areal elimination vs. adaptive areal masking. *Cartography and Geographic Information Science*, 47(6):537–549, 2020.
- 5 Bugra Gedik and Ling Liu. Location privacy in mobile systems: A personalized anonymization model. In *25th IEEE International Conference on Distributed Computing Systems (ICDCS'05)*, pages 620–629. IEEE, 2005.
- 6 Gabriel Ghinita, Keliang Zhao, Dimitris Papadias, and Panos Kalnis. A reciprocal framework for spatial k -anonymity. *Information Systems*, 35(3):299–314, 2010.
- 7 Aris Gkoulalas-Divanis, Panos Kalnis, and Vassilios S Verykios. Providing k -anonymity in location based services. *ACM SIGKDD explorations newsletter*, 12(1):3–10, 2010.
- 8 Marco Gruteser and Dirk Grunwald. Anonymous usage of location-based services through spatial and temporal cloaking. In *Proceedings of the 1st international conference on Mobile systems, applications and services*, pages 31–42, 2003.
- 9 Alex Hern. New york taxi details can be extracted from anonymised data, researchers say. *The Guardian*, June 2014. (Accessed on 01/16/2023).
- 10 Panos Kalnis and Gabriel Ghinita. Spatial anonymity. In LING LIU and M. TAMER ÖZSU, editors, *Encyclopedia of Database Systems*, pages 2685–2690. Springer, 2009.
- 11 Panos Kalnis, Gabriel Ghinita, Kyriakos Mouratidis, and Dimitris Papadias. Preventing location-based identity inference in anonymous spatial queries. *IEEE transactions on knowledge and data engineering*, 19(12):1719–1733, 2007.
- 12 Carsten Kießler and Grant McKenzie. A geoprivacy manifesto. *Transactions in GIS*, 22(1):3–19, 2018.
- 13 Ourania Kounadi and Michael Leitner. Adaptive areal elimination (AAE): A transparent way of disclosing protected spatial datasets. *Computers, Environment and Urban Systems*, 57:59–67, 2016.
- 14 John Krumm. A survey of computational location privacy. *Personal and Ubiquitous Computing*, 13(6):391–399, 2009.
- 15 Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t -closeness: Privacy beyond k -anonymity and l -diversity. In *2007 IEEE 23rd international conference on data engineering*, pages 106–115. IEEE, 2007.
- 16 Bo Liu, Wanlei Zhou, Tianqing Zhu, Longxiang Gao, and Yong Xiang. Location privacy and its applications: A systematic study. *IEEE access*, 6:17606–17624, 2018.
- 17 Yongmei Lu, Charles Yorke, and F Benjamin Zhan. Considering risk locations when defining perturbation zones for geomasking. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 47(3):168–178, 2012.
- 18 Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkatasubramanian. l -diversity: Privacy beyond k -anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):Article 3, 2007.

- 19 Grant McKenzie, Krzysztof Janowicz, Song Gao, and Li Gong. How where is when? On the regional variability and resolution of geosocial temporal signatures for points of interest. *Computers, Environment and Urban Systems*, 54:336–346, 2015.
- 20 Grant McKenzie, Krzysztof Janowicz, and Carsten Keßler. Uncovering spatiotemporal biases in place-based social sensing. *AGILE GIScience Series*, 1:14, 2020.
- 21 Grant McKenzie, Daniel Romm, Hongyu Zhang, and Mikael Brunila. Privyto: A privacy-preserving location-sharing platform. *Transactions in GIS*, 26(4):1703–1717, 2022.
- 22 Franz-Benjamin Mocnik. Putting geographical information science in place—towards theories of platial information and platial information systems. *Progress in Human Geography*, 46(3):798–828, 2022.
- 23 Mohamed F Mokbel, Chi-Yin Chow, and Walid G Aref. The new casper: Query processing for location services without compromising privacy. In *VLDB*, volume 6, pages 763–774, 2006.
- 24 Dilay Parmar and Udai Pratap Rao. Privacy-preserving enhanced dummy-generation technique for location-based services. *Concurrency and Computation: Practice and Experience*, 35(2):e7501, 2023.
- 25 Fiona Polzin and Ourania Kounadi. Adaptive Voronoi Masking: A Method to Protect Confidential Discrete Spatial Data. In Krzysztof Janowicz and Judith A. Versteegen, editors, *11th International Conference on Geographic Information Science (GIScience 2021) - Part II*, volume 208, pages 1–17, 2021.
- 26 Ross S Purves, Stephan Winter, and Werner Kuhn. Places in information science. *Journal of the Association for Information Science and Technology*, 70(11):1173–1182, 2019.
- 27 Stéphane Roche. Geographic information science ii: Less space, more places in smart cities. *Progress in Human Geography*, 40(4):565–573, 2016.
- 28 Pierangela Samarati and Latanya Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, Data Privacy Lab Report, 1998.
- 29 Simon Scheider and Krzysztof Janowicz. Place reference systems. *Applied Ontology*, 9(2):97–127, 2014.
- 30 Dara E Seidl, Gernot Paulus, Piotr Jankowski, and Melanie Regenfelder. Spatial obfuscation methods for privacy protection of household-level data. *Applied Geography*, 63:253–263, 2015.
- 31 David Swanlund, Nadine Schuurman, and Mariana Brussoni. MaskMy. XYZ: An easy-to-use tool for protecting geoprivacy using geographic masks. *Transactions in GIS*, 24(2):390–401, 2020.
- 32 Latanya Sweeney. k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems*, 10(5):557–570, 2002.
- 33 Zhouxuan Teng and Wenliang Du. Comparisons of k-anonymization and randomization schemes under linking attacks. In *Sixth International Conference on Data Mining (ICDM'06)*, pages 1091–1096. IEEE, 2006.
- 34 J.K. Trotter. Public NYC taxicab database lets you see how celebrities tip, October 2014. (Accessed on 10/14/2022).
- 35 Daniel Wagner, Alexander Zipf, and Rene Westerholt. Place in the giscience community—an indicative and preliminary systematic literature review. In *Proceedings of the 2nd International Symposium on Platial Information Science (PLATIAL'19)*, pages 13–22. Zenodo, 2020.
- 36 Jue Wang, Junghwan Kim, and Mei-Po Kwan. An exploratory assessment of the effectiveness of geomasking methods on privacy protection and analytical accuracy for individual-level geospatial data. *Cartography and Geographic Information Science*, pages 1–22, 2022.
- 37 Paul A Zandbergen. Ensuring confidentiality of geocoded health data: assessing geographic masking strategies for individual-level data. *Advances in medicine*, 2014:1–14, 2014.
- 38 Hongyu Zhang and Grant McKenzie. Rehumanize geoprivacy: from disclosure control to human perception. *GeoJournal*, 88(1):189–208, 2022.