Uniqueness in the City: Urban Morphology and Location Privacy*

HANCHENG CAO, JIE FENG, and YONG LI, Beijing National Research Center for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, China
VASSILIS KOSTAKOS, University of Melbourne, Australia

We investigate the potential for privacy leaks when users reveal their nearby Points-of-Interest (POIs). Specifically, we investigate whether and how a person's location can be reverse-engineered when that person simply reveals their nearby POI types (e.g. 2 schools and 3 restaurants). We approach our analysis by introducing a "Location Re-identification" algorithm that is computationally efficient. Using data from Open Street Map, we conduct our analysis on datasets of multiple representative cities: New York City, Melbourne, Vancouver, Zurich and Shanghai. Our analysis indicates that urban morphology has a clear link to location privacy, and highlights a number of urban factors that contribute to location privacy. Our findings can be used in any systems or platforms where users reveal their proximal POIs, such as recommendation systems, advertising platforms, and appstores.

CCS Concepts: • **Security and privacy → Human and societal aspects of security and privacy**; • **Information systems** → *Data mining*;

Additional Key Words and Phrases: Location Privacy, re-identification, POI, uniqueness, urban Morphology

ACM Reference Format:
Hancheng Cao, Jie Feng, Yong Li, and Vassilis Kostakos. 2018. Uniqueness in the City: Urban Morphology and Location Privacy. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 2, Article 62 (June 2018), 20 pages. https://doi.org/10.1145/3214265

1 INTRODUCTION

This paper investigates a deceivingly simple question: by revealing the types of nearby Points-of-Interests (POIs), do we reveal our actual location? For instance, by simply revealing that there are 2 schools and 3 restaurants nearby a user, is it possible to reverse-engineer the actual location of the user? There are many systems that rely on knowledge of nearby POIs, such as recommendation systems and advertising platforms[33, 39, 40]. Even appstores, for example, can promote travel apps when the user is near an area with tourist attractions[34]. Furthermore, POIs are freely and easily accessible through a variety of platforms, hence they are becoming increasingly used for conveying users' context.

As disclosing user locations can give away sensitive information about individuals' homes/working places, lifestyles, economic status, and even politic beliefs and health information, location privacy is drawing increasing attention in academia where researchers have been trying to develop privacy-aware location based services

*This work was supported in part by the National Nature Science Foundation of China under 6171101425, 61621091 and 61673237, and research fund of Tsinghua University - Tencent Joint Laboratory for Internet Innovation Technology.

Authors' addresses: Hancheng Cao, chc14@mails.tsinghua.edu.cn; Jie Feng, feng-j16@mails.tsinghua.edu.cn; Yong Li, liyong07@tsinghua.edu.cn, Beijing National Research Center for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing, 100084, China; Vassilis Kostakos, University of Melbourne, Australia, vassilis.kostakos@unimelb.edu.au.

2474-9567/2018/6-ART62 $15.00
https://doi.org/10.1145/3214265

that avoid direct sharing of user GPS information [10, 14]. Conventionally, an assumption has been made that recommendation systems can preserve privacy by only requiring knowledge of the types of POIs near a user, and not the actual location coordinates of the user. For instance, Yu et al. [34] have proposed a system that makes app usage prediction for app promotion on the list of nearby POI types without disclosing GPS information and states that the method ensures privacy. However, as we show in this paper, some more can be inferred, depending on the city within which the user is, and the actual uniqueness of the user's location. Our findings therefore provide key insight for building future privacy-preserving systems.

Our analysis uses the notion of location uniqueness: a unique location is one that has a unique combination of POIs. We expect that such locations are easier to reverse-engineer, when users are near POIs of which there are not many in a city. In fact, we intuitively refer to the concept of location uniqueness in our everyday life. For instance, when speaking of a person just arriving in New York City by air, New Yorkers will immediately think of the person being at LGA, JFK or EWR — any of the three airports in New York. Similarly, when a user is near a theatre, Broadway will be a likely location. Thus, places with distinct functions are very likely to have a high level of location uniqueness (and therefore relatively low privacy when revealing nearby POIs). These anecdotal observations inspire us to consider that location uniqueness in cities may be much higher than we may assume.

To make our analysis computationally tractable, we propose a "Location Re-identification" algorithm and develop a model to study the lower bounds of location uniqueness. Instead of trying to reverse-engineer the exact GPS coordinates of a user, we simply try to figure out the candidate areas surrounding any existing POIs. This insight enables us to avoid intractable calculations, allows pre-computation, and thus is much faster and computational efficient for our purposes.

Using POI data from five representative cities of New York, Melbourne, Vancouver, Zurich and Shanghai, we conduct an extensive assessment on location uniqueness through our proposed model, and systematically analyze the role of spatial granularity, POI density, POI composition as well as geographical factors contributing to location uniqueness. Our results show that locations can be highly unique in different cities, where a spatial granularity of radius 2km can uniquely identify 75%, 53%, 87%, 64% and 72% of all locations in New York, Melbourne, Vancouver, Zurich and Shanghai, respectively. Meanwhile, location uniqueness is closely related to multiple factors: denser POI regions tend to contain more infrequent POI types, and regions closer to the city centre tend to be more unique. The extensive results highlight that sharing or revealing nearby POIs can lead to privacy breaches, and we highlight a number of factors and strategies that can be used to mitigate this.

2 RELATED WORK

2.1 POI Identification and Recommendation

Recent studies analyzed the POI identification with crowd-sourced data to generate entries for landmarks [12], discover places with events happening [13], and identifying users' visited POIs from their trajectories [24]. Moreover, some research focused on individual level POI recommendation. Bin et al. [15, 16] studied personalized recommendation by leveraging users' multi-aspect preferences, and Xin et al. [18] consider users' preference transition to achieve relevant recommendation. Other previous work has considered POIs in the context of geographical and temporal influence [33, 39, 40]. Instead of identifying and recommending POIs, we study how POIs can uniquely identify the locations of city. Our findings provide useful guidance for building privacy-aware location-based service systems in the future.

2.2 POI-based Urban Property Study

POIs have been utilized for urban land-use in transportation research since the early 20th century. POIs have been used to describe the characteristics of travel behaviour between different types of land use, such as the traffic between residential zones and industrial zones. Voorhees [27] described how travel between different types of

origins and destinations roughly follows gravitational laws, with different types of destinations generating certain types of "pull" towards the origins. In fact, it is suggested that individuals organize spatial knowledge according to anchor points, POIs, or generally salient locations that form the cognitive map that the individual uses to navigate [22]. Besides geographical points, such as landmarks, anchor points can be path segments, nodes or even distinctive areas, similar to urban properties categorized by Lynch [21]. McGowen et al. [23] tested the feasibility of a model that predicts activity types based solely on GPS data from personal devices, GIS data and individual or household demographic data. Ye et al. [32] proposed a framework which uses a mixed hidden Markov model to predict the category of user activity at the next step and then predict the most likely location given the estimated category distribution. Yang et al. [30] first modelled the spatial and temporal activity preference separately, and then used a principle way to combine them for preference inference. Other models treat the urban as a system of markets to simulate the linkages between different industrial sectors[28]. While some other dynamic models capture the dynamics of urban growth and land use changes by both rule-based models [8] and empirical estimation models [11, 20]. Different from these static or dynamic models to characterize the urban environment, we utilize POIs to distinguish the uniqueness of the locations of the city. Our findings highlight the surprising high location uniqueness in cities and analyze the relationship between location uniqueness and multiple factors, and therefore contribute to our understanding of urban property from a novel angle.

2.3 POI-based Human Behavior Study

More broadly, land use affects various aspects of travel behaviour, such as trip generation, distance travelled and choice of mode of transport [4]. Thus, POIs have also been utilized for human behavior study.

POI have often been used to estimate travel behaviour since the early 1990s [2]. Such kind of method is usually with high complexity and intense data requirements [1], and it has even been noted that it is difficult to find a representative set of participants willing to commit to a long-term data gathering effort [3]. POI is also utilized to predict human future demand [19] and understand urban mobility with neighborhood characterizations [25]. Based on the understanding of user behavior inferred from the POI, lots of applications can be supported such as route recommendation[37], tourism recommender systems[17, 36], human mobility investigation [35], novel urban computing applications [38][9, 41, 42] and even smartphone app usage prediction[34]. Rather than attempt to develop applications leveraging POI, we instead explore the problem of how the POI information indicate the uniqueness of a city, which has never been studied before.

2.4 Location Privacy

Our work is also closely related to another line of research on location privacy. Recent years have seen a growing concern on privacy issue related to location based services, e.g., POI recommender system and location based social network, where the services are largely dependent on sharing of individuals' location data [10, 14]. Montjoye et al. [6, 7] demonstrated the high uniqueness of individual movement using mobile phone and transaction data, where four spatio-temporal points are shown sufficient to uniquely identify over 90% of the individuals. Xu et al. [29] showed that aggregated mobility data does not ensure user privacy. Meanwhile, a number of frameworks have been proposed to ensure user location privacy from multiple angles, including privacy-aware system architecture building [5], data obfuscation [31] and data merging [26]. While most research deals with location privacy issues in mobility data, we demonstrate a new kind of threat to location privacy when users simply reveal nearby POI types, and we highlight how they correlate with urban morphology.

Table 1. Basic statistics for each city in our analysis.

	Country	Population(*million*)	Urban area(km^2)	Num of POIs	Num of POI types	Studied area (km^2)
New York	U.S.	8.538	1213.37	26202	125	118*89
Melbourne	Australia	3.848	9992.5	17735	145	151*144
Vancouver	Canada	0.647	114.97	5267	74	22*16
Zurich	Switzerland	0.391	87.88	22000	147	41*39
Shanghai	China	24.2	6341	9618	94	395*365

3 DATASET

In our analysis we use publicly available POI data on Mapzen Metro Extract (https://mapzen.com/data/metro-extracts/), which is extracted and updated weekly from OpenStreetMap (http://www.openstreetmap.org/). Compared with other platforms such as Google Map, Baidu Map and Foursquare, OpenStreetMap is the most easily accessible data source, covers a wide range of user around the world and reveals actual urban morphology instead of user check-in. To investigate a variety of POI compositions and morphologies, we carefully selected five cities to represent cities of different sizes and functional roles in 5 countries around the globe: New York, Melbourne, Vancouver, Zurich and Shanghai. We provide a description of these cities as follows:

New York As a global city, New York is the largest city in the United States, with the world's largest natural harbour and the most active financial market. Because of its special status, and size, we choose New York City as the representative of western metropolis.

Melbourne Melbourne is the second-most populous city in Oceania. Compared to New York City, Melbourne has half the population but ten times the land area. We choose Melbourne as a representative Oceanian metropolis.

Vancouver Vancouver is the third-largest metropolitan area in Canada. Although the population and land area of Vancouver are smaller than New York, Vancouver's population density is similar to New York. We choose Vancouver as the representative of middle-size western city.

Zurich Zurich is the biggest city of Switzerland. With a similar population size to Vancouver, Zurich is also a typical median city. Meanwhile, because of the popularity of OpenStreetMap in Europe, the POI data of Zurich is expected to be richer than the other cities.

Shanghai Shanghai is the largest city in China, and one of the most populous cities in the world. As a global financial center and transport hub, we select Shanghai as a representative of Asian metropolis in developing country. Due to the aggregation of cities in China, the Mapzen dataset also records cities near Shanghai (e.g., Suzhou, Hangzhou) since these cities lie in greater Shanghai metropolitan area.

We summarize the basic information (population, urban area) of these five cities and the statistics of their recorded POI data (Number of POIs, Number of POI types and studied area) from Mapzen in Table 1. As presented in the table, POI data from Mapzen is fine-grained in composition, where the selected cities record around 100 different POI types, and each city records a few thousand POIs. We also present the top 10 most frequent POI types in the five cities in Table 2. It can be observed that parking, restaurant, cafe, bench, school, etc. frequently appear in all cities, which are common POI types concerned in location based services.

4 METHOD

Here we describe our approach to measure location uniqueness. To quantify location uniqueness, we first define the problem of location re-identification. Then, we present our computationally efficient re-identification model.

Table 2. Most popular POI types in studied cities.

City	Top 10 most popular POI types
New York	bicycle parking, restaurant, school, place of worship, cafe, fast food, bench, bank, fire station, drinking water, post box
Melbourne	restaurant, bench, cafe, fast food, toilets, post box, parking, drinking water, bicycle parking, waste basket, telephone
Vancouver	bench, restaurant, bicycle parking, cafe, fast food, post box, waste basket, bank, toilets, bicycle rental, drinking water
Zurich	bench, restaurant, drinking water, waste basket, vending machine, post box, parking, parking entrance, recycling, bicycle parking, fast food
Shanghai	bicycle rental, restaurant, bank, cafe, toilets, fast food, parking, community centre, fuel, school, bench

4.1 Problem Definition

Location Re-identification: We count the frequency of all POI types within a given radius r around a particular location l, which gives us a POI type distribution vector P ($P = [n_{p_1}, n_{p_2}, ...n_{p_m}]$), where n_{p_i} represents the frequency of POI type p_i within radius r around l. Then, we try to re-identify this location through P from a location pool L and result in a number of possible locations L_C, known as *candidate*.

Through the proposed location re-identification task, we are able to study location uniqueness: locations with unique surrounding POIs will end up with few *candidate* locations, while other locations with common POIs nearby will result in lots of *candidate* locations across the city.

As we analyze location uniqueness through POIs, it is unnecessary to consider regions without POIs in the location pool L since these places can never be identified. We therefore represent location pool L as a number of spatial circles centering at each POI, which ensure that all instances in L have at least one POI and may be used for re-identification. Therefore, the number of possible locations location pool L in a city corresponds to the number of POIs recorded for that city.

4.2 Model

It is indeed true that one can develop an algorithm to re-identify a location and search for all its *candidate*s with high accuracy, say comparing the POI type distribution vector P through a sliding circle of the same radius in the city. However, such brute-force algorithms turn out to be impractical for our purposes, since its computational requirement is exponential: at every move, the distance between the sliding circle center and every POI in the city has to be recalculated, and even an improved approach like checking locations around POIs also involve lots of calculations and comparisons.

So as to tackle the computational challenge and provide a first glance into location uniqueness, we relax our demand on accuracy. Instead of identifying the exact location, we attempt to find larger regions within which those possible locations lie in. In this way, we provide a lower bound in location uniqueness.

Specifically, our model takes in a given POI type distribution vector P around location l within radius r, and first sorts P based on the overall POI frequency observed in the entire city. Our model then selects the most infrequent/unique POI type p_l in P and finds all p_l in the city from location pool L. As p_l is observed within radius r around location l, conversely every *candidate* location must be within a distance of r around each p_l, which therefore forms a bounding circle of radius r around p_l. Thus, the POI type distribution vector P must result from an observation of a radius r circle lying within a big circle of radius $2r$ centered at each p_l. We then add further constraints: since P comes from an observation of radius r circle lying within a big circle of radius $2r$

centered at each p_l, to be eligible for a *candidate*, the radius $2r$ circle around p_l should overlay no fewer than $n_{(p_i)}$ for POI type p_i, thus further narrowing down the *candidate* locations. The model is illustrated in Fig.1.

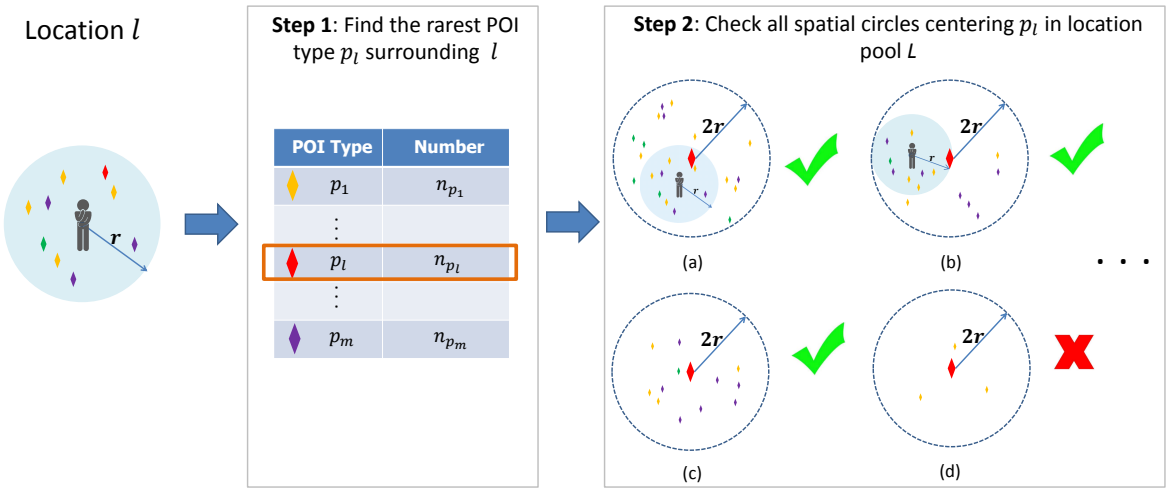


Fig. 1. Illustration of our proposed location re-identification model. Different POI types are represented by diamonds of different colors. We first pick out the rarest POI type p_l found near location l within radius r (denoted by "red" diamond in the figure), then we select all spatial circles centering at POI p_l from the location pool L. The spatial circles have radius $2r$ since the user should be no farther from the center than r so that he/she can "see" p_l, therefore the radius of $2r$ ensures that all POIs the user "sees" lie within the big circle. We further check if the $2r$ circle involves all POIs revealed by the user. The model will return possible *candidate* locations, including l (subfigure (a)), locations sharing exact POI compositions within radius r (subfigure (b)), as well as locations sharing similar POI components (subfigure (c)), but filter out spatial circles where it is impossible to reveal POI type distribution vector P (subfigure (d)). Thus the proposed model provides a lower bound of location uniqueness in a time efficient manner.

Since our proposed model does not involve many distance calculations and comparisons, but instead focuses on the POI statistics within radius $2r$ circle around p_l, the computational complexity is greatly reduced. Our model also makes possible pre-calculation: POI statistics within $2r$ around each POI points can be pre-calculated once and saved as a dictionary since this process is independent of the location-to-be-re-identified, and thus further speeds up the entire process.

5 ANALYSIS

We conduct multiple analyses on our datasets from the 5 cities. In addition to measuring location uniqueness in these cities, we also look at the role of different factors in shaping location uniqueness, including spatial granularity, POI density, POI composition as well as distance to city center.

5.1 Experiment Setup

To best capture the city structure, for each city, we uniformly sample 1,200,000 unique locations lying inside the bounding box of the city given by Mapzen Metro Extract. We calculate POI type statistics for each location using a varying radius of 0.1 km, 0.25 km, 0.5 km, 1 km, 2 km, 4 km to represent different levels of spatial

granularity. Intuitively, this radius reflects the range that a user considers when reporting nearby POIs in real world applications such as recommender systems. In some cases, a user may use a short range (e.g. 100 meters), while in other cases a longer range may be considered (e.g. 4 kilometers).

Using a varying radius in our analysis ensures that we preserve the multi-level POI compositions in the city: when the radius is small, fine-grained POI composition is captured while a larger radius captures coarse-grained properties. We filter out those locations with few surrounding POIs (which represent unpopulated areas such as sea, mountain and fields) and retain those locations in populated regions. We use the following filtering criterion: for each experimental location in the five cities, we calculate the POI density within radius r, and filter out location with POI density lower than threshold $50/\pi$ (i.e., there are at least 50 POIs around the location when r is 1km). The visualizations in Fig. 6 and Fig. 7 suggest that the proposed criterion has successfully preserved locations with relatively high POI density across the city while filtering out unpopulated locations, such as rivers and sea. Finally, we adopt our proposed attack model to measure and analyze a lower bound of location uniqueness in cities.

5.2 Location Uniqueness

We first examine location uniqueness in the five cities. The results of our experiments show that location uniqueness in these cities is surprisingly high. As presented in Fig. 2, under spatial granularity r = 2km, over 75% of randomly selected locations in New York can be uniquely identified through their surrounding POI types, while the same figure for Melbourne, Vancouver, Zurich and Shanghai is 53%, 87%, 64% and 72%, respectively. Meanwhile, the percentage of randomly chosen locations which can be narrowed down to fewer than three candidate areas of similar POI compositions are 84%, 68%, 91%, 72% and 77% in New York, Melbourne, Vancouver, Zurich and Shanghai, and the percentage of being identified within three candidate areas are 92%,75%, 93%, 76% and 85%. These results suggest that POI composition, or urban functional structure, is highly unique in cities all over the world, and consequently can pose a considerable threat to location privacy.

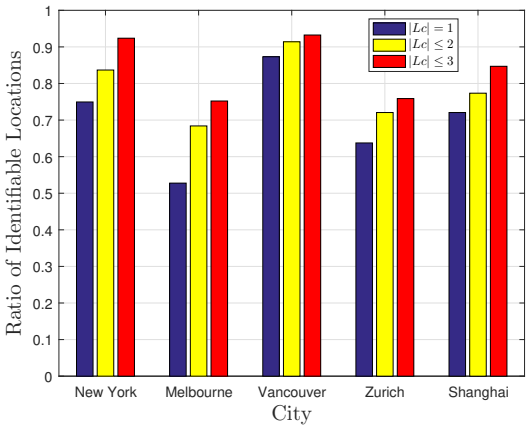


Fig. 2. Location Uniqueness in New York, Melbourne, Vancouver, Zurich and Shanghai, spatial granularity $r = 2km$. The chart shows the ratio of locations that can be re-identified down to a single ($|L_c| = 1$), two ($|L_c| = 2$) and three ($|L_c| = 3$) locations.

5.3 Spatial Granularity

We look at how location uniqueness varies under different spatial granularity, where spatial granularity is represented by radius r. Fig. 3 shows the percentage of locations which can be identified within one or two candidate locations, with respect to radius. It can be observed that as the radius increases, the location uniqueness increases. When the radius is 0.1km, 0.25km and 0.5km, approximately 10% to 30% of locations can be uniquely identified. When the radius is 1km, around 40% of locations can be uniquely identified, while almost all locations can be uniquely identified when the radius gets to 4km. In other words, if one stands in a city and reveals his nearby POIs within 0.1km, the chance of successfully finding the user is around 10% while if he/she reveals POIs within a radius 4km, his/her location is very easy to accurately identify. Thus users should be especially careful to reveal POIs around them with a large radius. Interestingly, we notice that Shanghai and New York show higher degree of location uniqueness than Zurich at larger radius, even though the studied areas in the two cities are much larger in size. This could be explained by different city's distinct urban morphology. In fact, location uniqueness is a direct consequence of urban structure, which is not only shaped by the size of the city, but also urban planning, cultural customs, etc. Therefore, location uniqueness cannot be explained by city size alone. Moreover, the relatively fewer POI records in Shanghai might also be a factor in affecting the results. Since OpenStreetMap is not quite popular in China, Shanghai's city structure might not be as well captured as western cities.

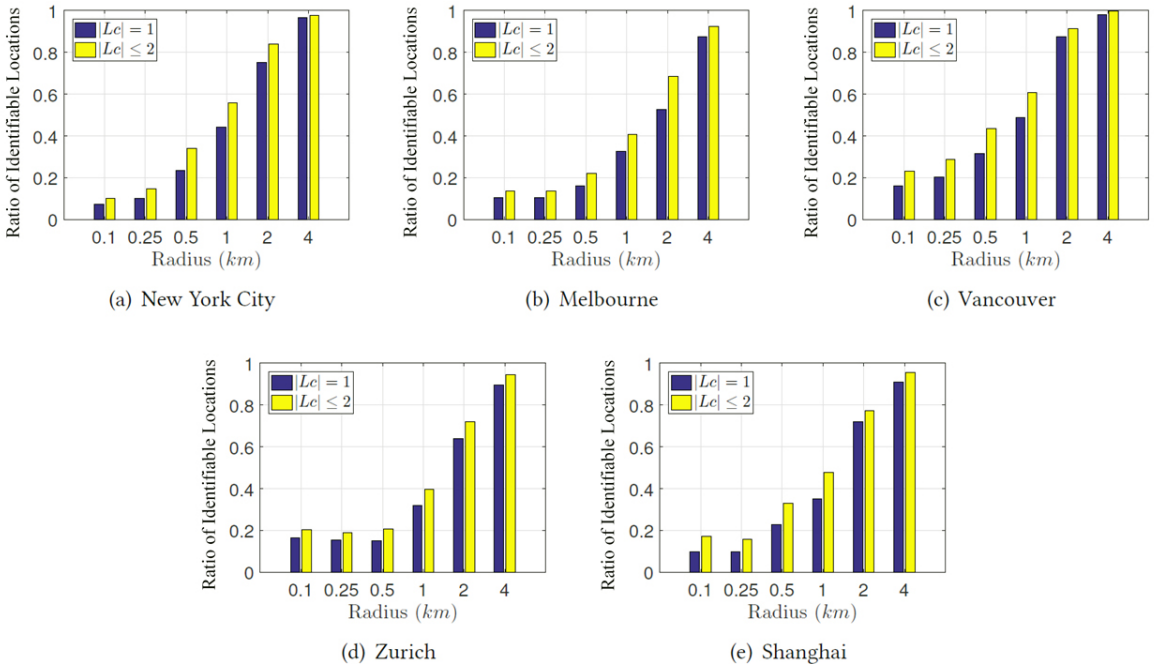
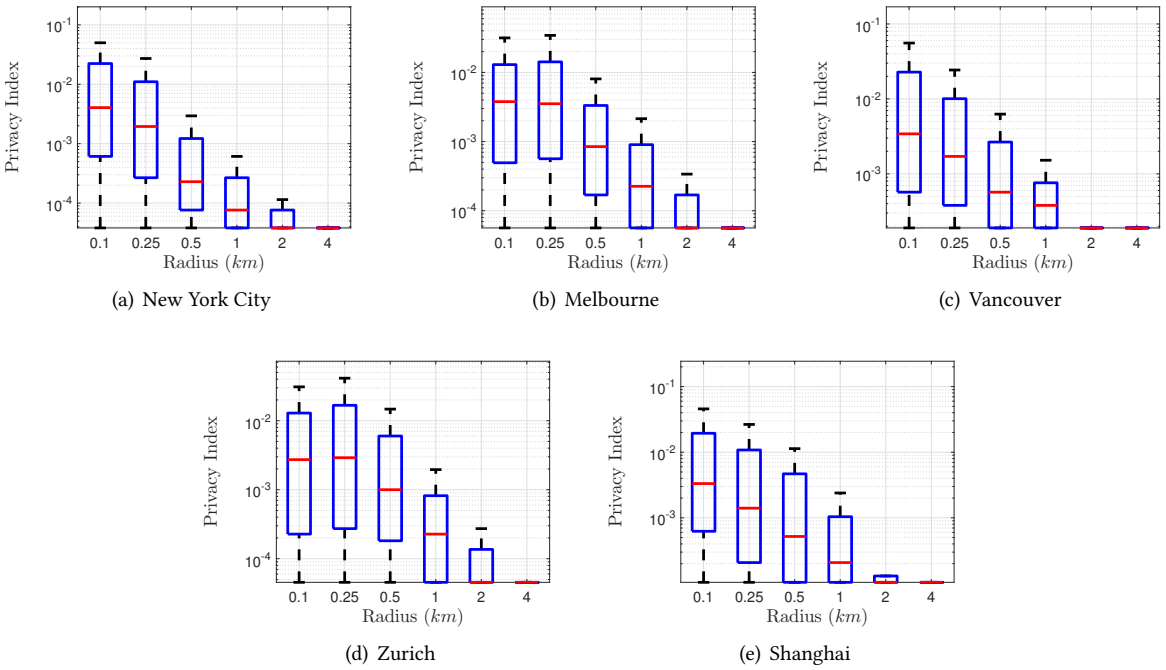


Fig. 3. Percentage of locations which can be identified within one ($|L_c| = 1$) or two ($|L_c| = 2$) possible regions in the city with respect to radius.

To obtain a global picture of location uniqueness in these four cities, we also generate boxplots of the mean privacy index with respect to radius, as shown in Fig.4. We define the privacy index of a location as the number of candidate re-identified locations divided by the total number of locations in the pool L (equals to the number of POIs recorded for the city). This metric aims to describe "what is the percentage of other areas in the city which share highly similar POI composition as this location?". Thus, the smaller the privacy index, the greater level of uniqueness of that location, and therefore less location privacy.



(a) New York City (b) Melbourne (c) Vancouver

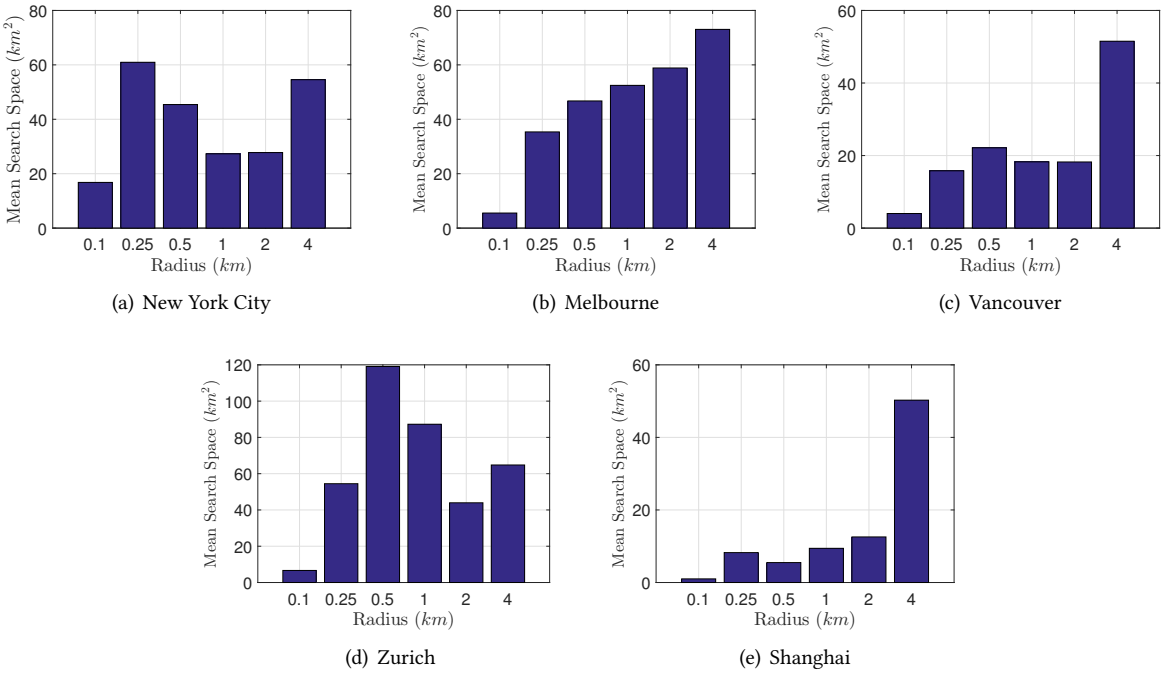
(d) Zurich (e) Shanghai

Fig. 4. Privacy index (location uniqueness) with respect to Radius.

Fig. 4 shows that under most circumstances, the privacy index decreases as the radius increases. In other words, a location becomes easier to identify as the radius increases. If one only looks at their immediate surroundings it is hard to uniquely identify that location. This result is in accordance with our intuition that "there are many locations in a city where a restaurant can be found within 100 meters; but locations where there is a restaurant and a factory within 250 meters are scarcer". It is worth mentioning that despite the general trend, certain locations can be easily identified with a small radius, as indicated by the fact that the lower quartile values in Fig. 4 can be quite low. This in turn motivates us to further analyze the relationship between location uniqueness and POI density as well as POI composition afterwards.

Besides the privacy index of each location, which is effectively a probability, we can quantify the privacy of each location by considering the actual geographic search space (in square kilometers) associated with each location. This measure is especially important from a privacy point of view since it represents the mean area an attacker will have to search in order to accurately locate the user, and is therefore a useful metric to measure the level of privacy for each user. As shown in Fig. 5, we analyze the relationship between spatial granularity (i.e. r) and the mean geographic search space of the locations that have identical nearby POIs (defined as $mean(\pi r^2 | l_c |)$). We observe that there is a considerable difference between the cities. For Melbourne, Vancouver and Shanghai, a

larger radius results in larger search spaces. In contrast, for New York City and Zurich, location privacy is best preserved with a radius of 0.25km and 0.5km, respectively. We attribute these differences to the morphological differences between the cities.



(a) New York City (b) Melbourne (c) Vancouver

(d) Zurich (e) Shanghai

Fig. 5. Mean search space with respect to radius.

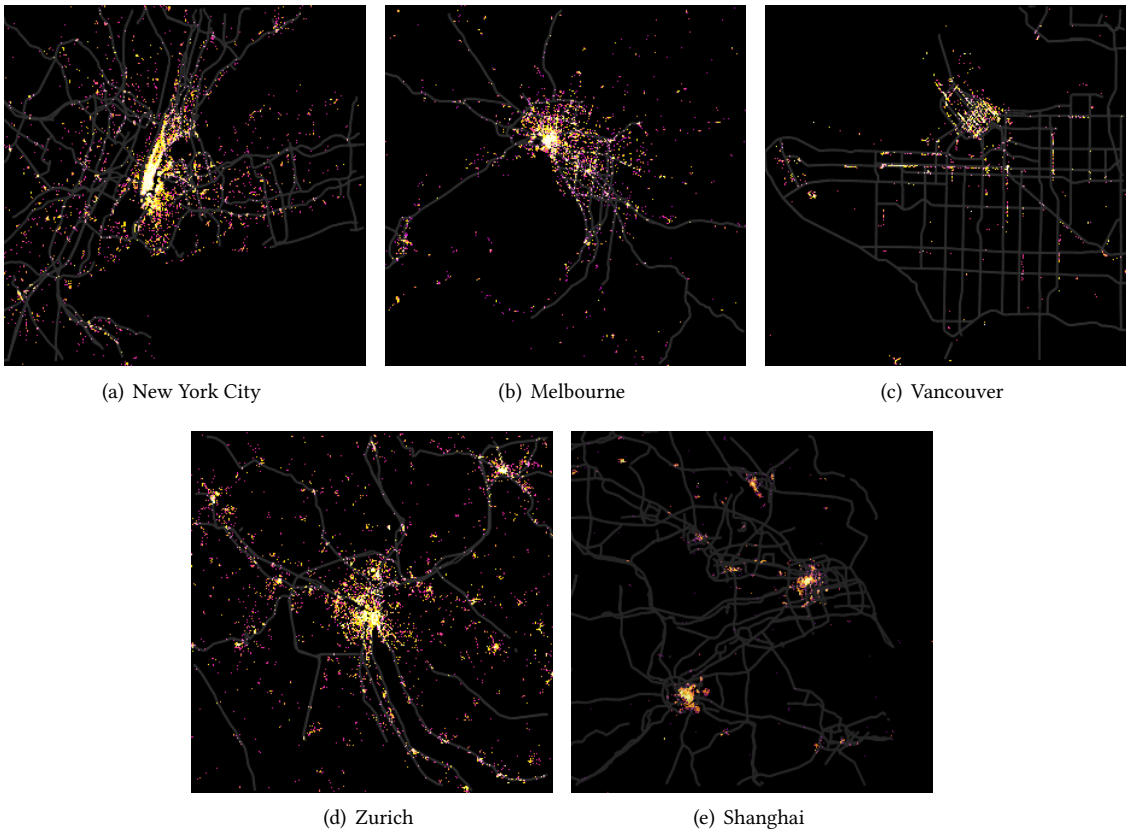
5.4 POI Density

Next, we analyze the relationship between POI density and location uniqueness. Fig. 6 presents the spatial distribution of POI density in the five cities, where the black background represents land and the greyish lines shows the shape of the primary road network. In Fig. 6, locations with more POIs are highlighted by a brighter color. As Fig. 6 shows, POIs usually spread around the road network and are clustered in the downtown area of city. Similarly, Fig. 7 shows the spatial distribution of location uniqueness. It is possible to notice that these two spatial distributions are visually similar, which suggests that the distribution of unique locations are very likely relevant to the distribution of POI.

We further analyze the relationship between location uniqueness and POI density in Fig. 8. Taking New York for example, as Fig. 8(a) shows, locations become increasingly unique with the growth of POI density. This conclusion is supported by the results from different cities with different search radius (r) in Fig. 8. In other words, it is easier to re-identify a location in denser regions.

5.5 POI Popularity

We investigate the effect of infrequent POIs on location uniqueness. As shown in Fig. 9, if there are infrequent (i.e. rare) POIs near a location, the location tends to have higher level of uniqueness. In contrast, if only frequent



(a) New York City (b) Melbourne (c) Vancouver

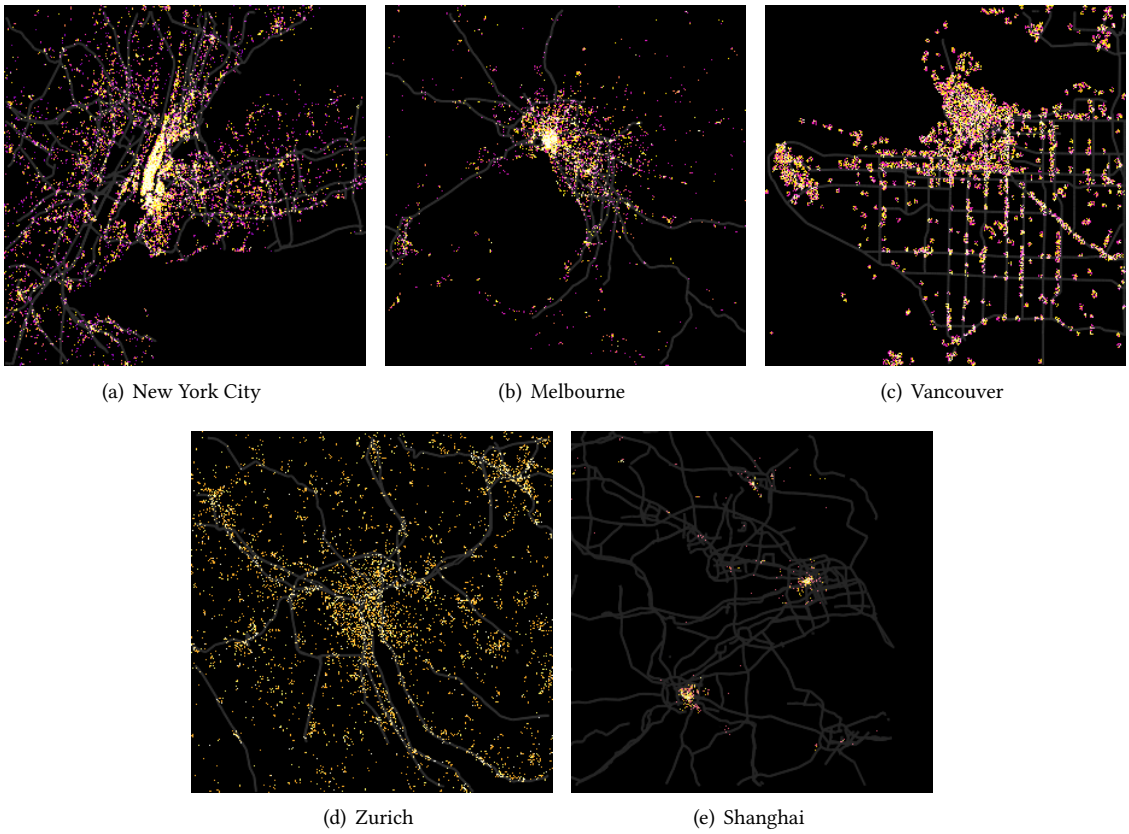
(d) Zurich (e) Shanghai

Fig. 6. Spatial distribution of POI density. Brighter color means higher density.

Table 3. Selected Infrequent POIs for different cities.

City	Infrequent POIs
New York	green market, hunting stand, music venue, catering
Melbourne	video store, surgery, statue, water point
Vancouver	nursery, waste disposal, co-op housing
Zurich	nightclub, research institute, barn, spa, conference center

POIs appear near a location, the uniqueness level is low. The rare POIs for each city are shown in Table 3. It is worth noting, however, that under certain circumstances, even if the nearby POIs are frequent, the location can still be quite unique. This is shown, for example, in the Melbourne plot: when the POI rank is around 140, certain locations can become quite unique.



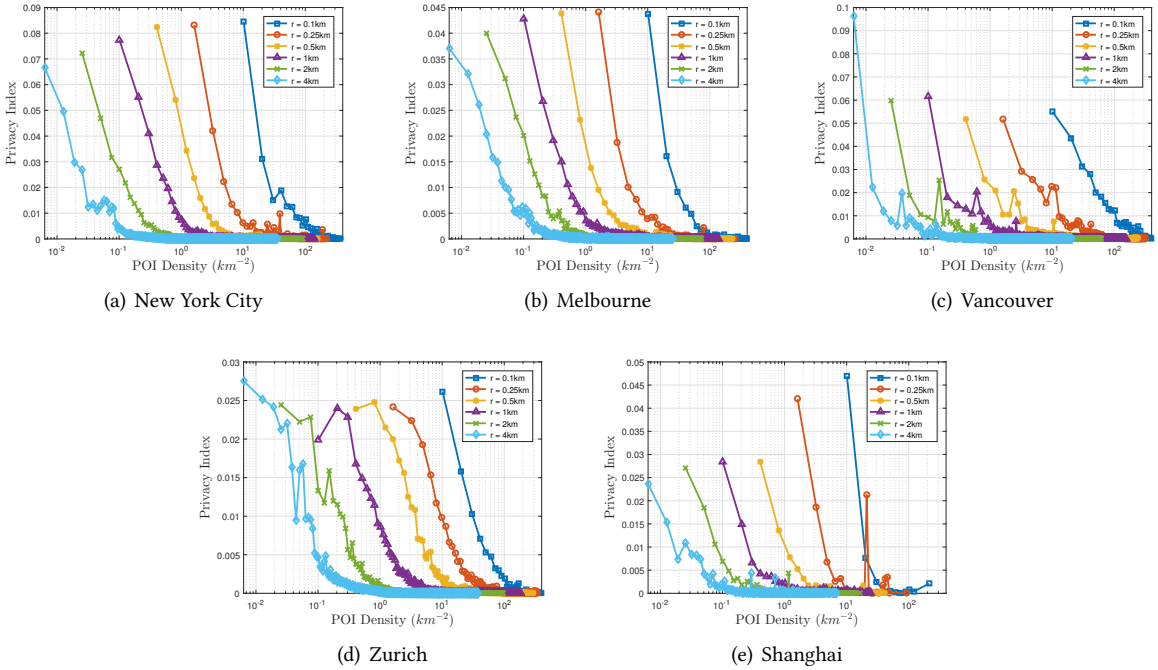
(a) New York City (b) Melbourne (c) Vancouver

(d) Zurich (e) Shanghai

Fig. 7. Spatial distribution of location uniqueness. Brighter color means higher uniqueness.

5.6 Distance to City Center

As illustrated in Fig.10, we find a trend that location uniqueness decreases when the location gets farther from the city center. This can be explained by the fact that city center normally exhibits higher density of POIs. Vancouver, however, exhibits some interesting pattern: when radius is small (0.1km and 0.25km), location uniqueness is lowest at the city centre and get highest roughly 2km from the city center, while under larger radius, location uniqueness follows regular pattern that decreases when getting farther from city center. This phenomenon may be explained by the fact that the downtown area (where POIs show high density) of Vancouver is much smaller than other cities and the functional structure at city center is simpler than other cities. Therefore, with small radius, users are more likely than elsewhere to observe generic POIs such as restaurant. When the radius gets larger, however, the POI structure gets more complex as one can observe more POIs, thus the uniqueness level increases.



(a) New York City (b) Melbourne (c) Vancouver

(d) Zurich (e) Shanghai

Fig. 8. Relationship between privacy index (location uniqueness) and POI density.

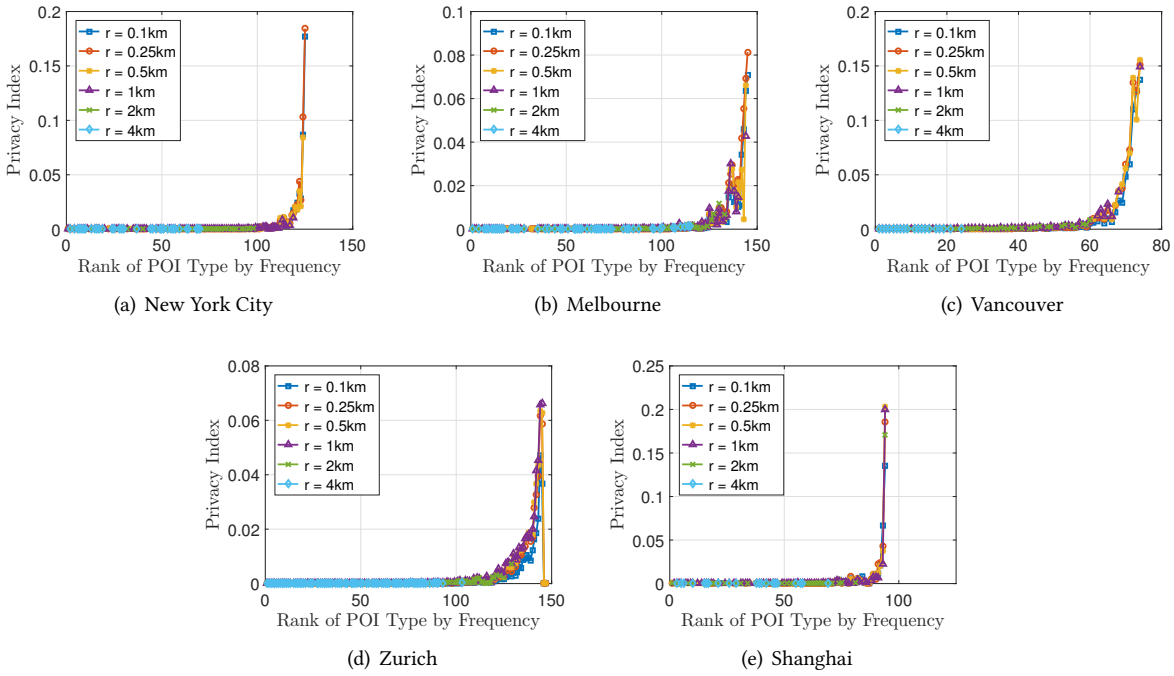
6 DISCUSSION

6.1 Strategies for Reducing Privacy Leaks

Our analysis reveals the extent to which privacy leaks can occur when sharing one's nearby POIs. For instance, this happens when an appstore makes app recommendations based on the list of nearby POI types [34]. We focus on cases where the phone only reveals nearby POI types to the service provider, not the actual GPS coordinates. Therefore, our analysis provides important guidelines for future privacy-aware system building. For example, we found that when one reveals the POIs within a 2km range, there is an 87% chance that they can be precisely pin-pointed if they are in Vancouver, while this can be as low as 53% in Melbourne. One way to protect privacy is to use a shorter radius. We found that using a radius of 100 meters, there is a 10% chance that the location can be uniquely identified. When the radius is much larger, say 4km, then almost all locations can be uniquely identified.

In other words, our findings show that a location is not that unique if one only reveals their immediate surroundings. However, the location becomes increasingly unique as one considers a larger radius. It is important to note that in our analysis we have only considered one-off single-sharing scenarios. Yet, it is quite conceivable that a mobile user, roaming across the city, may be sharing their nearby POIs on multiple occasions. Our analysis suggests that if those instances of sharing happen close in time, then an attacker may combine the two shared datasets (thus effectively considering one larger radius) and possibly have a better chance of identifying the location of the user.

Finally, the results show that it is easier to re-identify a location in denser regions, or when there are rare POIs nearby. There are a couple of sharing strategies that can take these findings into account to preserve privacy. For future ubiquitous location based device/service, a higher level of location privacy will be preserved if sharing



(a) New York City (b) Melbourne (c) Vancouver

(d) Zurich (e) Shanghai

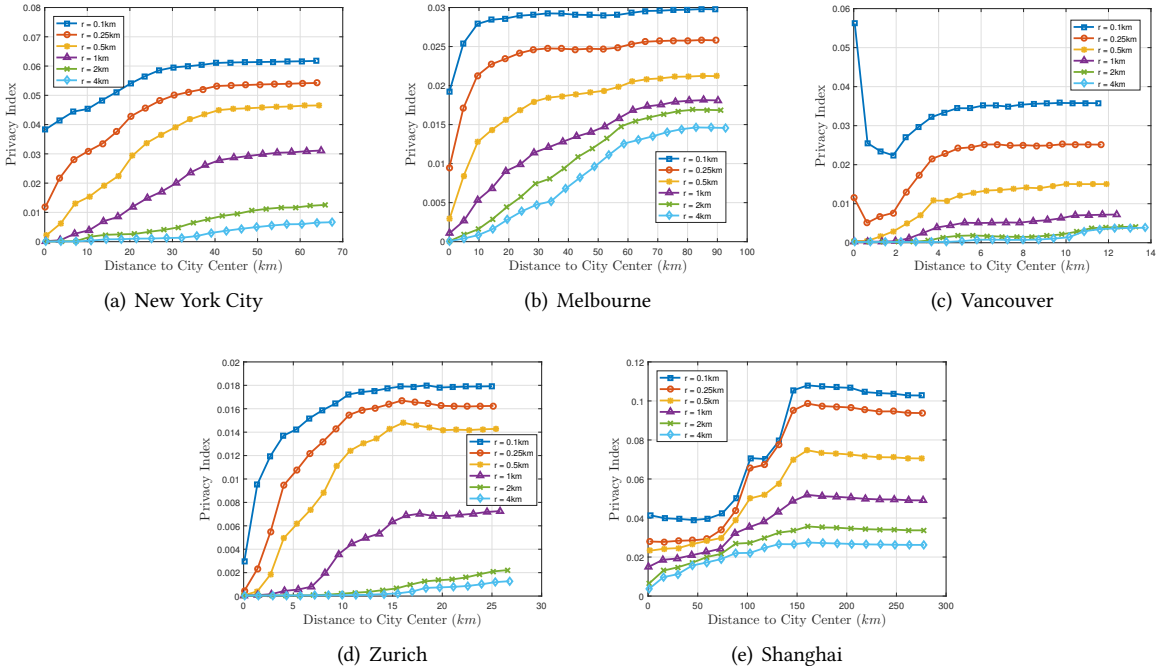
Fig. 9. Relationship between privacy index (location uniqueness) and POI Uniqueness/Ranking. POIs of rank 1 are the most popular in their city.

of POIs is reduced when in POI-dense locations, or the device/service switches to a shorter radius when in dense locations. In addition, sharing could be masked by avoiding to reveal rare or infrequent nearby POIs. Furthermore, a strategy could combine all these parameters non-deterministically, and randomly choose certain POIs to hide (preferably infrequent POIs). It may also be of help to precompute a safe radius (which meets certain criteria such as k-anonymity) in different locations and construct location privacy maps in different cities, so that device/service may self-adapt to ensure user location privacy.

6.2 Generalizability of Our Findings

In our analysis we have considered multiple cities in different continents, for the purpose of ensuring that our findings are reliable and transferable. We have found a number of patterns that exist in all cities, thus establishing stronger confidence in our results. However, it is also interesting to consider the differences across cities, so that we can better understand which of our findings are likely to be localized and contextualized for each city independently.

There are certain general patterns evident in all cities in our analysis: all cities show high level of location uniqueness and therefore face the same general threats to location privacy. As the radius for spatial granularity increases, locations become more unique. Moreover, POI density is a reliable indicator of location uniqueness: higher POI density tends to suggest higher level of uniqueness and thus more susceptible to privacy attacks. Meanwhile, the uniqueness of POIs also greatly determines the level of location uniqueness: unique POIs that do not frequently appear in the city make it easier to re-identify a location. Since POIs generally aggregate near the



(a) New York City (b) Melbourne (c) Vancouver

(d) Zurich (e) Shanghai

Fig. 10. Relationship between privacy index (location uniqueness) and distance to city center.

city center, the distance to the city center can also help distinguish location uniqueness: the closer to the city center, the location is more likely to be unique.

Yet there are some interesting differences between the cities in our analysis. On one hand, smaller cities as Vancouver and Zurich show greater location uniqueness when the spatial granularity is high. Perhaps this is due to the fact that the chance of observing infrequent POIs with high spatial granularity is greater in smaller cities.

Furthermore, the mean search area shows different patterns for each city: Melbourne, Vancouver and Shanghai show greater location privacy when the radius is larger, while New York and Zurich have higher level of location privacy when the radius is neither too small nor too big (r=0.25km or 0.5km). Furthermore, even though all cities demonstrate that the existence of rare POIs lead to reduced location privacy, the actual types of these POIs vary between different cities since fine-grained city structure differs. For example, as a middle size city, Zurich has only a few conference centers and research institutes, in contrary to New York, as shown in Table 3. As a result, the privacy pattern in one city cannot be directly generalized to other cities, and therefore researchers should pay special attention to the city difference when designing privacy-aware mechanisms.

In essence, our work highlights the separation between morphological and functional characteristics of cities, and how they contribute to the privacy index we have proposed. We show that the importance of core morphological aspects such as POI density and radius size remains invariant across cities, and therefore are probably the drivers of our findings. On the other hand, the importance functional characteristics, such as the precise purpose of POIs, vary substantially between cities, and we therefore conclude that they are likely to be driven by morphological characteristics, such as the size of the city, the size of city center, and the spatial distribution of different functions.

6.3 POIs as Context

POIs are a commonplace and publicly available data resource. Many online map services, some of which are free, provide access to this information, including Google Maps, FourSquare, Open Street Map and Baidu Map. POIs are simply categories or labels that have been geo-coded and attached to specific coordinates across cities.

While a lot of UbiComp research has focused on localization techniques, context-aware computing has largely not shown great interest in the potential benefits of considering POIs. Perhaps this has been the case because freely open and high-quality POI databases have only recently become relatively popular, and due to crowdsourcing efforts they are becoming increasingly complete.

Yet, POIs have the potential to enrich a system's contextual understanding, especially for location-aware systems. In some ways, GPS and localization coordinates are valuable for navigation, but our community still lacks techniques that can leverage nearby POIs to build a contextual model of a user's location: for a user to know "where am I" is not only a matter of localization, but also contextualization. For instance, visiting a new city requires localization for navigation purposes, but contextualization is required for exploring and experiencing the city.

Our results show that revealing our immediate surroundings does tell much about where we are. While our work has considered this finding from the perspective of an attacker, it is also interesting to consider it from the perspective of context-aware systems. Conceptually, our findings suggest that if our personal device can only "see" or know our immediate surroundings, it may not have a good idea about the type of location where we are. We show that by gradually expanding the range of consideration, from 100 meters to 4 kilometers, it becomes increasingly easier to re-identify the user's location.

It could be argued that this finding puts a bound on location-based context awareness, and the geographic range that should be considered when developing location-based services. Consider the scenario where a user's device wishes to personalise its behaviour based on the nearby POIs. For example, this could be the case for a travel app that detects nearby tourist attractions. The device may simply query OpenStreetMaps, or FourSquare, to find out what nearby POIs exist. A key question here is to determine what range should the query consider. Should it consider only immediate POIs, or farther ones too? Our findings show that if only immediate surroundings are considered, then the personalisation that can be achieved is quite "generic", since not enough can be inferred about a person's location when using a limited range. As that range grows to 4 kilometers, we show that the information is rich enough to uniquely identify the location, and therefore offer a richer personalisation experience. Similarly, in dense location a shorter range can be considered.

6.4 Computational Complexity

In this paper, we show that it is possible to consider a city's POIs as a means of investigating the nature of location uniqueness. In our analysis we consider whether given a location and its surrounding POIs, can we identify similar places in this city? Our results provide key insights for location privacy: is it safe for users if location-based service providers such as POI recommendation systems make public their surrounding POIs? Should users themselves reveal their nearby POIs? If not, under what conditions is such kind of sharing acceptable?

Despite its great significance, quantifying location uniqueness is far from computationally trivial. The key challenge comes from the computational complexity. Brute-force methods generally require huge amounts of distance computation and comparisons between different location points and their nearby POIs so as to judge if they share similar POI compositions, which becomes too computationally expensive to get a fine-grained understanding of location uniqueness in different cities. Instead, we propose a simple enough model to give a global understanding of location uniqueness and its implications for city structure and location privacy with fine-grained spatial granularity, which is memory and computational efficient.

6.5 Limitations and Future Work

We make use of POI data of five cities in North America, Oceania, Europe and Asia for this study. There is a limitation in the dataset selection as Open Street Map does not provide sufficient POI information, especially for developing countries in Asia and Africa. In the future, we plan to adopt multiple data sources, such as Google Map, Baidu Map and Foursquare to study location uniqueness in the world on a more global scale. Furthermore, we have not taken into account the effect of multi-level POI granularity on location privacy since Open Street Map only provides one level of POI granularity, which is a possible direction for future study on other data sources. We also plan to study the roles of more factors, such as road network property, district economic level, population, etc. in shaping location uniqueness. What's more, it would be interesting to jointly model urban morphology, radius and location privacy. For instance, currently we have not taken into consideration the difference between 4km*4km regions downtown and 4km*4km in the suburbs on level of privacy.

In this work, we mainly aim at taking the first glance into the problem of location uniqueness and privacy. Therefore, we propose a computationally efficient model to study a lower bound of location uniqueness so as to provide preliminary results on the problem. A more accurate model can be used to study the problem and conversely design a framework to ensure user location privacy when sharing POI information. For instance, POI distribution in the city can be modelled with a kernel density distribution, and the task of location re-identification can be treated as a document search problem. Also, we focus on scenario when users expose POIs nearby within a radius. It would be interesting to investigate privacy issues under different location sharing framework, e.g., the user expose the nearest k POI types around without exposing a radius. Of course, it is of the greatest importance to investigate how to design a location service framework that maintains high quality of personalization experience while maintain user location privacy.

Finally, our work only considers static and honest users. Mobility has not been taken into account, and it is conceivable that mobile users who share their nearby POIs on multiple occasions may be revealing even more information that makes it possible to retrace their steps. Honesty has been assumed throughout our analysis, meaning that users share all nearby POIs within a radius, and do not selectively hide POIs.

7 CONCLUSION

In this paper we investigate the possibility of re-identifying the actual location of a person based on their surrounding POIs. This type of analysis quantifies the potential for privacy breaches when revealing nearby POIs. We propose a computationally efficient "Location Re-identification" method, and conduct extensive experiments on POI datasets of four representative cities of New York, Melbourne, Vancouver and Zurich from Open Street Map. Our results highlight the surprisingly high location uniqueness in these cities. We further analyze the relationship between location privacy and spatial granularity, POI density, POI combination and distance to city center, so that we help understand under what circumstances location privacy may be compromised. Our findings contribute to the understanding of urban morphology and provide guidance for future location privacy-preserving system and platforms where users reveal their nearby POIs, such as recommendation systems, advertising platforms, and appstores.

APPENDIX

We provide a pseudocode of our proposed Location Re-Identification Algorithm as shown in Algorithm 1:

ALGORITHM 1: Location Re-Identification Algorithm

Input: Location $l(x, y)$, POI locations in the city $poi(label, x, y)$, radius r
Output: Candidate location $l_c(x, y)$
// get the POI distribution vector P at l within radius r;
P = zeros(1,num(poi.label));
for *j in poi* **do**
  if *distance*$(l.x, l.y, j.x, j.y)$<r **then**
    $P(j.label) = P(j.label)$+1;
  end
end
p_l = rarest($P.label$) // get the label of rarest POI type found in P ;
l_c = [] ;
$C = poi(poi.label == p_l)$ // find all p_l type POIs in the city ;
C_p = cell(num(C)) // save nearby POI type distribution within radius $2r$ centering at POI type p_l ;
for *c in C* **do**
  for *j in poi* **do**
    if *distance*$(c.x, c.y, j.x, j.y)$<$2 * r$ **then**
      $C_p(c)(j.label) = C_p(c)(j.label)$+1 ;
    end
  end
end
// check if the original location may appear in the radius $2r$ circle centering at POI type p_l ;
for *c in C* **do**
  if *($C_p(c)(j)$>= $P(j)$ for all j in poi.label)* **then**
    l_c.append($[C.x, C.y]$)
  end
end

REFERENCES

[1] Kay W Axhausen. 1998. Can we ever obtain the data we would like to have. *Theoretical foundations of travel choice modeling* (1998), 305–323.

[2] Kay W Axhausen and Tommy Gärling. 1992. Activity-based approaches to travel analysis: conceptual frameworks, models, and research problems. *Transport reviews* 12, 4 (1992), 323–341.

[3] Kay W. Axhausen, Andrea Zimmermann, Stefan Schönfelder, Guido Rindsfüser, and Thomas Haupt. 2002. Observing the rhythms of daily life: A six-week travel diary. *Transportation* 29, 2 (2002), 95–124. https://doi.org/10.1023/A:1014247822322

[4] Marlon G. Boarnet, Kenneth Joh, Walter Siembab, William Fulton, and Mai Thi Nguyen. 2011. Retrofitting the Suburbs to Increase Walking: Evidence from a Land-use-Travel Study. *Urban Studies* 48, 1 (2011), 129–159. https://doi.org/10.1177/0042098010364859

[5] Chi-Yin Chow and Mohamed F Mokbel. 2009. Privacy in location-based services: a system architecture perspective. *Sigspatial Special* 1, 2 (2009), 23–27.

[6] Yves-Alexandre De Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. 2013. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports* 3 (2013), 1376.

[7] Yves-Alexandre De Montjoye, Laura Radaelli, Vivek Kumar Singh, et al. 2015. Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science* 347, 6221 (2015), 536–539.

[8] Yongjiu Feng, Yan Liu, and Michael Batty. 2016. Modeling urban growth with GIS based cellular automata and least squares SVM rules: a case study in Qingpu–Songjiang area of Shanghai, China. *Stochastic environmental research and risk assessment* 30, 5 (2016), 1387–1400.

[9] Yanjie Fu, Guannan Liu, Spiros Papadimitriou, Hui Xiong, Yong Ge, Hengshu Zhu, and Chen Zhu. 2015. Real estate ranking via mixed land-use latent models. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 299–308.

[10] Gabriel Ghinita. 2013. Privacy for location-based services. *Synthesis Lectures on Information Security, Privacy, & Trust* 4, 1 (2013), 1–85.

[11] Zhiyong Hu and CP Lo. 2007. Modeling urban growth in Atlanta using logistic regression. *Computers, Environment and Urban Systems* 31, 6 (2007), 667–688.

[12] Lyndon S Kennedy and Mor Naaman. 2008. Generating diverse and representative image search results for landmarks. In *Proceedings of the 17th international conference on World Wide Web*. ACM, 297–306.

[13] Slava Kisilevich, Florian Mansmann, and Daniel Keim. 2010. P-DBSCAN: a density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos. In *Proceedings of the 1st international conference and exhibition on computing for geospatial research & application*. ACM, 38.

[14] John Krumm. 2009. A survey of computational location privacy. *Personal and Ubiquitous Computing* 13, 6 (2009), 391–399.

[15] Bin Liu, Yanjie Fu, Zijun Yao, and Hui Xiong. 2013. Learning geographical preferences for point-of-interest recommendation. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1043–1051.

[16] Bin Liu and Hui Xiong. 2013. Point-of-interest recommendation in location based social networks with topic and location awareness. In *Proceedings of the 2013 SIAM International Conference on Data Mining*. SIAM, 396–404.

[17] Qi Liu, Enhong Chen, Hui Xiong, Yong Ge, Zhongmou Li, and Xiang Wu. 2014. A cocktail approach for travel package recommendation. *IEEE Transactions on Knowledge and Data Engineering* 26, 2 (2014), 278–293.

[18] Xin Liu, Yong Liu, Karl Aberer, and Chunyan Miao. 2013. Personalized point-of-interest recommendation by mining users' preference transition. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. ACM, 733–738.

[19] Yanchi Liu, Chuanren Liu, Xinjiang Lu, Mingfei Teng, Hengshu Zhu, and Hui Xiong. 2017. Point-of-Interest Demand Modeling with Human Mobility Patterns. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 947–955.

[20] Xinjiang Lu, Zhiwen Yu, Leilei Sun, Chuanren Liu, Hui Xiong, and Chu Guan. 2016. Characterizing the life cycle of point of interests using human mobility patterns. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 1052–1063.

[21] Kevin Lynch. 1960. *The image of the city*. MIT press.

[22] E.J. Manley, J.D. Addison, and T. Cheng. 2015. Shortest path or anchor-based route choice: a large-scale empirical analysis of minicab routing in London. *Journal of Transport Geography* 43 (2015), 123 – 139. https://doi.org/10.1016/j.jtrangeo.2015.01.006

[23] Patrick Tracy McGowen and Michael G McNally. 2007. *Evaluating the potential to predict activity types from GPS and GIS data.* Technical Report.

[24] Kyosuke Nishida, Hiroyuki Toda, Takeshi Kurashima, and Yoshihiko Suhara. 2014. Probabilistic identification of visited point-of-interest for personalized automatic check-in. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 631–642.

[25] Souneil Park, Marc Bourqui, and Enrique Frias-Martinez. 2017. MobInsight: Understanding Urban Mobility with Crowd-Powered Neighborhood Characterizations. In *IEEE International Conference on Data Mining Workshops*. 1312–1315.

[26] Zhen Tu, Kai Zhao, Fengli Xu, Yong Li, Li Su, and Depeng Jin. 2017. Beyond k-anonymity: protect your trajectory from semantic attack. In *Sensing, Communication, and Networking (SECON), 2017 14th Annual IEEE International Conference on*. IEEE, 1–9.
[27] Alan M. Voorhees. 2013. A general theory of traffic movement. *Transportation* 40, 6 (2013), 1105–1116. https://doi.org/10.1007/s11116-013-9487-0
[28] Michael Wegener. 1994. Operational urban models state of the art. *Journal of the American planning Association* 60, 1 (1994), 17–29.
[29] Fengli Xu, Zhen Tu, Yong Li, Pengyu Zhang, Xiaoming Fu, and Depeng Jin. 2017. Trajectory recovery from ash: User privacy is not preserved in aggregated mobility data. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1241–1250.
[30] Dingqi Yang, Daqing Zhang, Vincent W Zheng, and Zhiyong Yu. 2015. Modeling user activity preference by leveraging user spatial temporal characteristics in LBSNs. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 45, 1 (2015), 129–142.
[31] Jing Yang, Zack Zhu, Julia Seiter, and Gerhard Tröster. 2015. Informative yet unrevealing: Semantic obfuscation for location based services. In *Proceedings of the 2nd Workshop on Privacy in Geographic Information Collection and Analysis*. ACM, 4.
[32] Jihang Ye, Zhe Zhu, and Hong Cheng. 2013. What's your next move: User activity prediction in location-based social networks. In *Proceedings of the 2013 SIAM International Conference on Data Mining*. SIAM, 171–179.
[33] Mao Ye, Peifeng Yin, Wang-Chien Lee, and Dik-Lun Lee. 2011. Exploiting geographical influence for collaborative point-of-interest recommendation. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM, 325–334.
[34] Donghan Yu, Yong Li, Fengli Xu, Pengyu Zhang, and Vassilis Kostakos. 2018. Smartphone App Usage Prediction Using Points of Interest. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 174.
[35] Gang Yu, Junsong Yuan, and Zicheng Liu. 2012. Predicting human activities using spatio-temporal structure of interest points. In *Proceedings of the 20th ACM international conference on Multimedia*. ACM, 1049–1052.
[36] Zhiwen Yu, Huang Xu, Zhe Yang, and Bin Guo. 2016. Personalized travel package with multi-point-of-interest recommendation based on crowdsourced user footprints. *IEEE Transactions on Human-Machine Systems* 46, 1 (2016), 151–158.
[37] Jing Yuan, Yu Zheng, Liuhang Zhang, XIng Xie, and Guangzhong Sun. 2011. Where to find my next passenger. In *Proceedings of the 13th international conference on Ubiquitous computing*. ACM, 109–118.
[38] Nicholas Jing Yuan, Yu Zheng, Xing Xie, Yingzi Wang, Kai Zheng, and Hui Xiong. 2015. Discovering urban functional zones using latent activity trajectories. *IEEE Transactions on Knowledge and Data Engineering* 27, 3 (2015), 712–725.
[39] Quan Yuan, Gao Cong, Zongyang Ma, Aixin Sun, and Nadia Magnenat Thalmann. 2013. Time-aware point-of-interest recommendation. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 363–372.
[40] Quan Yuan, Gao Cong, and Aixin Sun. 2014. Graph-based point-of-interest recommendation with geographical and temporal influences. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. ACM, 659–668.
[41] Y Zheng, F Liu, and HP Hsieh. 2013. When urban air quality inference meets big data. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining pp.(1436-1444). ACM.*
[42] Yu Zheng, Tong Liu, Yilun Wang, Yanmin Zhu, Yanchi Liu, and Eric Chang. 2014. Diagnosing New York city's noises with ubiquitous data. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 715–725.

Received November 2017; revised February 2018; accepted April 2018