Perils of Location Tracking? Personalized and Interpretable Privacy Preservation in Consumer Mobile Trajectories

Meghanath Macha
Information Systems and Management, Carnegie Mellon University, mmacha@cmu.edu

Beibei Li
Information Systems and Management, Carnegie Mellon University, beibeili@andrew.cmu.edu

Natasha Zhang Foutz
McIntire School of Commerce, University of Virginia, ynf8a@comm.virginia.edu

Anindya Ghose
Leonard N. Stern School of Business, New York University (NYU), aghose@stern.nyu.edu

(Working Paper December 20, 2019)

Consumer location tracking is becoming omnipresent on mobile devices, producing vast volumes of behavior-rich location trajectory data. These data have enabled a wide range of opportunities for monetization, such as location-based targeting. An advertiser, however, could either use the acquired location data to play the role of a "butler" who understands consumer needs and provides valuable personalized services, or goes overboard with marketing campaigns and misuses the location data by invading consumer privacy and becoming a real-life "stalker". This calls attention for regulatory bodies and location data collectors, such as mobile app owners or data aggregators, to identify ways to balance privacy risks to consumers and values to advertisers when sharing consumer location data with advertisers. This will curtail the stalker intent while facilitating the butler service. Existing approaches to privacy preservation are unsuited for location data as they are largely not personalized and difficult for data collectors to interpret the trade-off between privacy risks to consumers and values to advertisers. To address this research gap, we propose a personalized and interpretable framework that enables location data collectors to optimize the risk-value trade-off. Validating the framework on nearly one million location trajectories from more than 40,000 individuals, we find that high privacy risks indeed prevail in the absence of data obfuscation. For instance, on average an individual's home address can be accurately predicted within a radius of 2.5 miles and mobile operating system with an 82% success. Moreover, 49% individuals' entire location trajectories can be fully identified by knowing merely two randomly sampled locations visited by the individual. Outperforming multiple baselines, the proposed framework significantly reduces the consumers' privacy risks (e.g., by 15% of inferring home address) with minimal (i.e., $< 1\%$) decrease in the value to an advertiser. As novel and powerful consumer location trajectory data become increasingly available, we demonstrate their value to advertisers and accompanying privacy risks to consumers, and most importantly, propose a personalized and interpretable framework to mitigate their risks while maximizing their values.

Keywords: consumer privacy, GPS tracking, location data, mobile trajectory, machine learning, data obfuscation, mobile advertising

1. Introduction

1.1. Smart Tracking, Targeting, and Privacy

According to the latest Pew Research (Taylor and Silver 2019), 76% and 45% of the current population in the advanced and emerging economies, respectively, own a smartphone. These percentages continue to rise rapidly. Among the U.S. smartphone users, over 90% use location services such as Google Maps (Pew 2016). The fast penetration of smartphones, combined with the wide adoption of location services, has produced a vast volume of behavior-rich mobile location data (or location data, trajectory data hereafter). These data represent one of the latest, and most important, information sources available to marketers in the evolution of marketing data, from surveys to online clickstream and social media data (Wedel and Kannan 2016). It has also opened up $21 billion sales opportunities for advertisers, ranging from e-commerce retailers sending discount coupons to individuals in the vicinity of a store, commonly known as geo-fencing, to personal injury lawyers targeting those in emergency rooms (Luo et al. 2014, Andrews et al. 2016, Ghose et al. 2018, Kelsey 2018).

Geo-marketing based on mobile location data is attractive to advertisers for multiple reasons. First, mobile location data are straightforward to collect – an app permission away and tracked in the background in most mobile ecosystems – and readily accessible to advertisers.[1] Second, mobile location data are superior to alternative location data. The built-in sensors of mobile devices can provide continuous tracking of the movement trajectory of an individual (i.e., a sequence of fine-grained location GPS coordinates). Such individual-level location trajectory data are more precise and granular than social media geo-tags and consumer self check-ins. They are also more representative of the population than the less granular taxi and public transportation location data. Third, mobile location data offer an extensive profile of a consumer (or user hereafter, whose location are being tracked by a smartphone) for targeting purposes, such as his/her spatial-temporal movements, thus rich contexts of his/her behaviors and brand preferences, broad lifestyle patterns, socioeconomic conditions, and social relationships (Ghose et al. 2018). Such offline data become even more powerful if combined with the consumer's online footprints, such as click stream data or social media data, rendering a holistic online-offline consumer portfolio. Fourth, owing to excellent

[1] While both Apple and Android have taken measures to limit the collection of location data, guidelines remain ambiguous about the sales of such data to advertisers (Apple 2014, Verge 2019).

tracking and targeting of mobile advertising, attributing the success of a location-based ad campaign is simplified. Advertisers have access to a unique device ID associated with each smartphone, thus reducing their overhead to stitch together consumers' location data across sessions or apps when measuring the success of a campaign and gain a holistic view of each consumer (Apple 2012). Fifth, geo-marketing by butler advertisers also benefits consumers (Ghose 2017), such as allowing consumers to receive enhanced services, personalization (Chellappa and Shivendu 2010), and financial benefits such as coupons (Luo et al. 2014, Ghose et al. 2018) or lower insurance premiums (Soleymanian et al. 2019).

Mobile location data not only provide values to advertisers whose butler actions further benefit consumers, but also monetization opportunities to location data collectors who share data with advertisers. Despite of the existence of diverse sources and varieties of mobile location data, the backbone of this rapidly growing mobile location data economy is the huge number of mobile applications (apps hereafter). App owners and location data aggregators serve a two-sided market with consumers on one side and advertisers on the other, collecting location data to offer better services to consumers and to monetize with advertisers. For example, a recent article by the New York Times reported that mobile location data collectors accrue half to two cents per consumer per month from advertisers (Valentino-Devries et al. 2018).

Meanwhile this powerful new form of human location and movement data offer important values to advertisers (and as a result to consumers and data collectors as well), they entail major consumer privacy concerns, such as the inference of home locations. "Privacy" is defined as "the quality or state of being apart from company or observation" in Merriam-Webster. In business contexts, privacy broadly pertains to the protection of individually identifiable information online or offline, and the adoption and implementation of privacy policies and regulations. It is a key driver of online trust (Hoffman et al. 1999). More than three-quarters of consumers believe that online advertisers hold more information about them than they are comfortable with; and approximately half of them believe that websites ignore privacy laws (Dupre 2015). For offline location data, privacy risks are exemplified by identifications of an individual's home address, daily movement trajectories, and broad lifestyle, as vividly depicted by a recent New York Times' article (Valentino-Devries et al. 2018). These risks are arguably more concerning than those associated with other forms of consumer data, such as an individual's media habit or social media content.

The discussions so far call for data collectors, before sharing location data with advertisers, to maintain a crucial trade-off between the values to advertisers and privacy risks to consumers. This responsibility falls primarily upon data collectors as they are situated right between advertisers and consumers, and hold vested interests in continuously maintaining consumers' trust in order to collect and monetize location data.[2] This notion is also consistent with the extant literature across multiple disciplines on data sharing (Li et al. 2012, Terrovitis et al. 2008, Li and Sarkar 2009, Chen et al. 2013, Yarovoy et al. 2009, Machanavajjhala et al. 2009). The unique properties of, and hence challenges entailed by, the increasingly accessible and important mobile location data to be detailed next, nonetheless, call for novel frameworks to accomplish this trade-off. We thus aim to develop a personalized, privacy-preserving framework that incorporates consumer heterogeneity and optimizes a data collector's trade-off between values to advertisers and privacy risks to consumers.

1.2. Research Agenda and Challenges

As discussed earlier, there are three key entities in our business setting.

1. *Consumers:* are individuals who each owns a smartphone with one or more of the apps installed that transmit the individual's location data to the data collector. Each consumer has the option to opt out of any app's location tracking, with some potential downsides of restricted use of certain app functions, such as maps or local restaurant finders.

2. *Advertisers:* are firms that acquire data from a data collector to improve the targetability of their marketing campaigns. A subset of these advertisers, or even a third party with access to the location data, might have a stalker intent (stalkers hereafter) to perform malicious activities on the location data that invade consumer privacy, such as overly aggressive marketing or ignoring privacy concerns.

3. *Data collector:* is an app owner that collects consumers' location data from its mobile app, or a data aggregator that integrates location data from multiple apps. The data are collected in real time and may then be shared with or sold to advertisers interested in targeting the consumers.

In this work, we take a data collector's perspective and propose a framework for the data collector to balance between protecting consumer privacy against any stalker intent of an advertiser and preserving butler advertisers' targeting capabilities via data accessibility, accuracy, and usability (Muralidhar and Sarathy 2006). We aim to answer the following essential questions.

1. *Consuemers' privacy risks:* What are some of the key privacy risks of mobile location data to consumers due to an advertiser's potential stalker intent? Can these risks be quantified at each consumer level? Since a data collector has limited purview of how an advertiser could link the location data to a consumer's private information, understanding and quantifying the risks associated with various types of stalker behaviors (or threats hereafter) is a crucial first step to perform necessary data obfuscations.

[2] Cambridge Analyticas misuse of consumer data exemplifies severe backlash on the data collector, Facebook, whose privacy practices resulted in loss of both consumers and advertisers (Pew 2018).

2. *Advertisers' utilities:* What are some of the values of raw and obfuscated mobile location data to butler advertisers' targeting abilities (for simplicity, advertisers' utilities hereafter)? That is, what types of key behavioral information can butler advertisers extract from mobile location data to target consumers in a mutually beneficial way?

3. *A data collector's trade-off between consumers' privacy risks and advertisers' utilities:* Is there a reasonable privacy-utility trade-off? If yes, what are the necessary steps that a data collector needs to take?

To accomplish the above, several methodological challenges need to be overcome. Our research questions, from a methodological standpoint, broadly fall under the paradigm of Privacy-Preserving Data Publishing (PPDP) widely studied in the context of relational databases (Fung et al. 2010). Nonetheless, the unique properties of mobile location data, such as high dimensionality (due to a large number of locations visited), sparsity (fewer overlap of locations across users), and sequentiality (order of locations visited), pose additional challenges (Chen et al. 2013). For example, traditional k-anonymity, which ensures an individual's record is indistinguishable from at least $k-1$ records, and its extensions face the curse of high dimensionality while dealing with granular, sometimes second-by-second, location data (Aggarwal 2005). $\epsilon-$differential privacy anonymization, which ensures adding or deleting a single consumer record has no significant impact on analysis outcomes, and other randomization-based obfuscation techniques (Machanavajjhala et al. 2006), fail to preserve the truthfulness of location data, rendering obfuscated data less useful for an advertiser's visual data mining tasks. More recent local obfuscation techniques (Chen et al. 2013, Terrovitis et al. 2017) that suppress locations with lower risk-utility trade-off provide a good privacy-utility balance. However, the obfuscation mechanisms are often complex for a data collector to interpret and apply in practice. For instance, the $(K,C)_L$ privacy framework (Chen et al. 2013) requires multiple parameters from a data collector, such as the probability thresholds of a stalker threat to succeed in different types of behaviors. LSup (Terrovitis et al. 2017) requires similar input parameters. Given the complex nature of these approaches, understanding and setting such parameters are non-trivial for a data collector. Hence, a more interpretable framework is needed to assist a data collector. Furthermore, the extant approaches do not tie a butler advertiser's utility to specific business use cases. These approaches, devised mostly from the Computer Science literature, measure an advertiser's utility by simply the number of unique locations or location sequences preserved in the obfuscated data (Chen et al. 2013, Terrovitis et al. 2017). These measures are rather rudimentary and impractical for advertisers to interpret or link to monetary decision-making. This challenge thus needs to be tackled by tying the advertiser's utility to real-world business contexts. We will next overview the proposed framework that intends to address the above challenges.

1.3. Overview of Proposed Framework

We provide a brief overview of the proposed framework that consists of three main components: quantification of consumers' privacy risks, quantification of advertisers' utilities, and obfuscation scheme for a data collector.

Quantification of Consumers' Privacy Risks. While the proposed framework may accommodate a variety of privacy risks, we illustrate the framework by computing two specific risks of vital concerns to consumers. One is "sensitive attribute inference", where a consumer's sensitive attributes, such as home address, is being inferred (Li et al. 2007, Tucker 2013, Gardete and Bart 2018, Rafieian and Yoganarasimhan 2018). And two is "re-identification threat", where all locations visited by a consumer are being identified based on a subset of the locations visited by the consumer (Samarati 2001, Pellungrini et al. 2018).

Quantification of Advertisers' Utilities. While the utilities of mobile location data to advertisers are multi-faceted, we demonstrate a specific utility related to one arguably most popular and essential business goal examined in the literature – Point-of-Interest (POI hereafter) recommendations in mobile advertising (Ghose et al. 2018). Reliable predictions of a consumer's future locations would enable an advertiser to target the consumer with relevant contextual contents. For instance, if a chain restaurant can accurately predict that a consumer is going to be in the vicinity of one of its outlets, it may target the consumer with a discount coupon of value to the consumer. We hence quantify this utility as the accuracy of a neighborhood-based collaborative filtering recommendation model trained on the location data. The central idea of this recommender is to identify other consumers with similar historical preferences in order to infer the focal consumer's future preferences (Bobadilla et al. 2011)

Obfuscation Scheme for a Data Collector. Acknowledging many potential solutions to the privacy-utility trade-off facing a location data collector may emerge, we propose an obfuscation scheme grounded on the idea of suppressing a subset of a consumer's locations, given the consumer's specific privacy risk and the frequency, recency, and time that the consumer spent at each location. We achieve this by introducing consumer-specific parameters that control the number and identity of the locations suppressed for each consumer. The suppression, while reducing the consumers' privacy risks, also adversely impacts advertisers' utilities. Hence, we empirically identify the parameters that balance the privacy-utility trade-off through a structured grid search while leveraging the earlier risk quantification for each consumer.

In summary, Figure 1 illustrates the proposed framework encompassing the three components discussed above. In Part A, we compute each consumer's privacy risk and the advertisers' utility associated with the non-obfuscated mobile location data (i.e., the full sample). These would also represent the counterfactual case when no privacy protection is performed. We expect the full

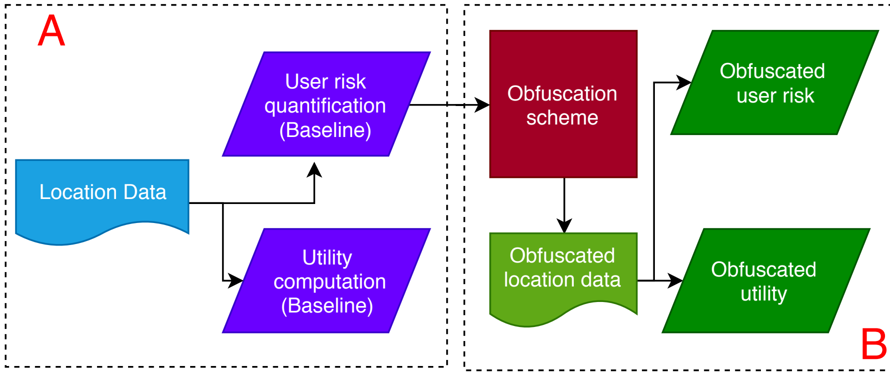


Figure 1 **Overview of the proposed framework**

non-obfuscated sample to yield the maximum utility to advertisers and incur maximum privacy risk to consumers if the data were shared as is. In Part B, we perform consumer-level obfuscation by computing the number and identity of locations to be suppressed for each consumer using each consumer's risk scores computed on the full non-obfuscated sample, as well as the frequency, recency, and time spent by the consumer at each location. We then repeat the above risk and utility computations by varying the parameters to empirically determine the best trade-off.

As alluded to earlier, while we hereby illustrate the power and value of the proposed framework by examining two key types of privacy risks and one key advertiser application, the framework is flexible to accommodate other types of privacy risks, such as the inference of a consumer's location sequence or visit frequency for which the risk may be quantified either analytically or via machine learning heuristics (Pellungrini et al. 2018). It may accommodate other types of advertiser use cases as well, for which the utility may be computed as the predictive accuracy for the business application of interest, such as when a consumer is most likely to convert given the consumer's past location trajectory, or how much is an advertiser's incremental revenue from geo-marketing. The framework is also applicable to other contexts, for instance, when the data collector conducts geo-marketing for itself or for advertisers without sharing location data. We will summarize the key findings next.

1.4. Summary of Key Findings

We validate the proposed framework on a unique data set of nearly one million mobile locations from over 40,000 individuals in a major mid-Atlantic metropolitan area in the U.S. over a period of five weeks in 2018. The main findings are summarized as follows.

First, we find that the absence of an obfuscation scheme, that is, no steps taken by a data collector to ensure consumer privacy, indeed entails high privacy risks to consumers. On average, the success probability of a privacy threat is 83% for infering a consumer's sensitive information and 49% for re-identifying a consumer's entire location trajectory, respectively. Specifically, on average a customer's home address can be predicted within a radius of 2.5 miles and mobile operating system[3] with an 82% success. [Meghanath - is this 82% related to the 83% above? What does the 83% above mean?] Moreover, 49% individuals' entire location trajectories can be fully identified by knowing only two randomly sampled locations visited by a consumer. [Meghanath, let us be clear about what these numbers mean, e.g. 49% individuals or 49% times successful when inferring any trajectory among all trajectories pooled across all individuals?] It is noteworthy that these success probabilities of various privacy threats are all estimated based on machine learning heuristics, which require merely the consumers' locations and corresponding timestamps as the inputs, as we will describe later. Hence, any entities, including advertisers, that have access to the location data could accomplish the same.

Second, we find significant values of the mobile location data to advertisers. An advertiser aiming to target consumers by identifying the next location most likely visited by a consumer would be able to predict it with 25% success. This means that by analyzing the behavioral patterns revealed by the historical location trajectories, for every one out of four customers, advertisers are able to design a highly precise geo-targeting strategy. [Meghanath - again, what does this 25% mean, 25% consumers or 25% times an advertiser may accurately predict a consumer's next location?]

Finally, a data collector could curtail the potential invasion of consumer privacy by performing data obfuscation. Using the proposed obfuscation scheme, where we suppress each consumer's locations based on the consumer's privacy risks and frequency, recency, and time spent at each location, a data collector may choose from multiple trade-off options to perform the obfuscation. Moreover, we find that the proposed framework presents a better choice set of trade-offs when compared to eight baselines obfuscation schemes of various types, including the rule based, consumer risk based, and latest suppression techniques such as Terrovitis et al. (2017). For instance, when the privacy threat is to predict a consumer's home address, the proposed obfuscation scheme reduces the risk by 15%, which represents the maximum decrease when compared to the baselines, with a minimum decrease of less than 1% in an advertiser's utility. We will present a more detailed discussion of the empirical findings and comparisons with the baseline obfuscation schemes in Section 5.

1.5. Summary of Key Contributions

We propose an interpretable framework built upon the principle of personalized data obfuscation for the emerging and increasingly critical mobile location data, which exhibit distinctive properties, such as high dimensionality, sparsity, and sequentiality, hence impose unique methodological challenges.

[3] Previous surveys have shown a strong relationship between mobile operating system and consumer demographics.(eMarketer 2013)

Conceptually, this research demonstrates the importance for location data collectors to preserve both consumer privacy and data utility to advertisers on a two-sided market. It hence presents a systematic framework to accomplish this privacy-utility balance. It also stands among the first research to demonstrate the immense business values of the novel mobile location data that capture granular human movements and are increasingly leveraged by marketers and other entities, such as municipalities (for instance, for smart city planning). This research simultaneously illustrates the significant privacy risks associated with these data if no framework were in place to preserve consumer privacy.

Managerially, this framework tackles three inter-related, critical challenges facing location data collectors: quantification of consumers' privacy risks, quantification of advertisers' utilities (i.e., values of mobile location data to advertisers), and design of an intutitve and interpretable obfuscation scheme for a data collector. It offers any data collector multiple, interpretable, and personalized options that require only parsimonious input to protect consumer privacy while preserving advertisers' utilities, hence overall monetization opportunities for the data collector.

Methodologically, this framework (1) quantifies consumers' privacy risks by extracting a comprehensive set of features from the mobile location data, thus accommodating various types of privacy risks and allowing identifications of which features contribute the most to the privacy risks; (2) measures advertisers' utilities associated with specific, real-world business use cases, such as POI recommendations shown to improve retailers' incremental revenues (Ghose et al. (2018)) [Meghanath - can we get ride of these extra parentheses in all citations throughout the manuscript?]; (3) proposes an interpretable obfuscation scheme that suppresses locations at each consumer level to facilitate a data collector with multiple intuitive options to maintain the privacy-utility trade-off; (4) demonstrates efficacy by validating the proposed framework on a massive, real-world mobile location data set and comparing the suggested obfuscation scheme with eight benchmarks.

A balance between consumer privacy and geo-marketing constitutes part of a broader debate over tracking and targeting on digital platforms. This debate has resulted in actions from both industries and regulatory bodies. For instance, Apple, with 44.6% US smartphone market share (Statista 2018), introduced limited ad tracking (LAT) in 2016, which allows individuals to opt out of tracking indefinitely (Apple 2016). Following suit, Android, the second most adopted mobile ecosystem rendered more controls to each consumer to limit tracking in its latest software update (Verge 2019). European Union's General Data Protection Regulation (GDPR, Regulation (2016)), effective from May 2018, requires individuals to opt-in (rather than opt out of) behavioral targeting and to give explicit permission for their data to be shared across firms. Balancing the benefit and privacy risks of consumer location data is increasingly becoming a key concern and top priority for firms and regulatory bodies. Besides strengthening privacy regulations, more research is called for to develop privacy-friendly data storage, processing, and analysis technologies (Wedel and Kannan 2016). Against this background, our research provides empirical evidence and practical solutions to inform the ongoing debate over mobile location tracking and location-based targeting. The rest of the manuscript is organized as follows. In Section 2, we review the literatures from various disciplines that are relevant to our research questions. In Section 3, we provide details of our business setting and discuss sampling and summary statistics of the mobile location data under analysis. In Section 4, we introduce the proposed framework (Figure 1). In Section 5, we discuss the empirical results and detail the advantages of the proposed framework. We offer concluding remarks in Section 6.

2. Literature Review

We will concisely review the most relevant Marketing, Management, Information Systems (IS), and Computer Science (CS) literature on consumer privacy, privacy-preserving methodologies, and location-based mobile advertising.

2.1. Literature on Consumer Privacy

The literature, particularly from Marketing, has a historical, and newly revived, interest in consumer privacy. As different forms of consumer data emerge over time, the literature has examined consumer privacy concerns that arise from many business contexts and data forms, such as marketing research like surveys (Mayer and White Jr 1969, De Jong et al. 2010, Acquisti et al. 2012), direct marketing via phones or emails (Hann et al. 2008, Kumar et al. 2014, Goh et al. 2015), offline retail sales (Schneider et al. 2018), subscription services and various customer relationship management (CRM) programs (Conitzer et al. 2012), online personalization services in computers and mobile devices (Chellappa and Shivendu 2010), online search and e-commerce transactions (Bart et al. 2005), online social networks (Adjerid et al. 2018). Prior studies have also examined privacy topics related finance and healthcare, such as crowd-funding (Burtch et al. 2015), credit transactions, insurance (Garfinkel et al. 2002, Soleymanian et al. 2019), and healthcare (Garfinkel et al. 2002, Miller and Tucker 2009, 2017). As advertisers commonly target consumers by leveraging their private information, the latest research has also investigated online, social media, and mobile advertising (Goldfarb and Tucker 2011a,b,c, Conitzer et al. 2012, Tucker 2013, Gardete and Bart 2018, Rafieian and Yoganarasimhan 2018). Broadly speaking, any circumstances that involve customer databases would entail privacy concerns and needs for privacy protection (Garfinkel et al. 2002, Martin et al. 2017, Muralidhar and Sarathy 2006, Qian and Xie 2015). As a result, even business-to-business (B2B) platforms incur privacy concerns and require effective strategies to address these concerns (Kalvenes and Basu 2006). Nonetheless, as massive amounts of novel mobile location data emerge, which offer unparalleled opportunities to examine large populations' granular lifestyles and generate debatably more severe privacy concerns, more research is needed to quantify consumer privacy risks and devise privacy-preserving strategies.

Marketing research on consumer privacy falls into four broad streams: consumer-, firm-, regulation-, and methodology- focused. We will concisely review each.

The first stream takes consumers' perspectives, and as a result, derives implications for firms to design privacy-friendly policies. For instance, this literature has studied how consumers respond to privacy concerns or make privacy choices about privacy-intruding survey questions (Acquisti et al. 2012), platform provided privacy settings (Burtch et al. 2015, Adjerid et al. 2018), online display ads that match the website contents but with obtrusive format (Goldfarb and Tucker 2011c,b), or opt-in/out options of email marketing programs (Kumar et al. 2014). It also studies how normative and heuristic decision processes influence consumers' privacy decision making (Adjerid et al. 2016). Overall, this research points to the positive effect of granting consumers enhanced controls over their own privacy, such as increasing their likelihood of responding to sensitive survey questions or click on personalized ads (Tucker 2013). Interestingly, this literature also shows that consumers behave in a way that reflects a "privacy paradox": claiming to care about their personal data yet more than willing to exchange the data for concrete benefits, such as convenience, personalization, or discounts (Acquisti and Grossklags 2005, Chellappa and Sin 2005, Awad and Krishnan 2006, Xu et al. 2011, Ghose 2017, Luo et al. 2014, Ghose et al. 2018), lower insurance premiums (Soleymanian et al. 2019), or a wider reach to audiences on social media for information acquisition or propagation (Adjerid et al. 2018). This paradox conversely indicates the potential for butler advertisers to leverage the newest mobile location data for geo-marketing to consumers in a mutually beneficial manner.

The second stream of literature take firms' perspectives, often using a game-theoretic approach to reach normative implications of firms privacy policies. For instance, Chellappa and Shivendu (2010) derive the optimal design of personalization services for customers with heterogeneous privacy concerns. Gardete and Bart (2018) propose the optimal choice of ad content and communications when firm withholds customers private information. Conitzer et al. (2012) reveal a monopolys optimal cost of privacy for customers to remain anonymous. Hann et al. (2008) show that consumers' different actions toward preserving their privacy, such as address concealment or deflecting marketing, impact a firms actions to either shifting marketing toward other consumers or reduce marketing overall. Adding competition to the picture, this stream of research also suggests optimal competitive strategies when profiting from disclosing customer information (Casadesus-Masanell and Hervas-Drane 2015), or designing a B2B market which preserves privacy to incentivize competitor participation (Kalvenes and Basu 2006). Other studies have also conceptualized the differential importance of privacy to different platforms (Bart et al. 2005) and assessed the impact of data breaches on firms financial performances (Martin et al. 2017). Interestingly, this stream of research also demonstrates that firms, such as an ad network, do have innate incentives to preserve customer privacy even without privacy regulations (Rafieian and Yoganarasimhan 2018).

The third stream of research focuses on privacy regulations. For example, privacy regulations are shown to impact firms privacy-pertinent practices, technology innovations (Adjerid et al. 2016) and adoptions (Miller and Tucker 2009, 2017), and consumer responses such as to do-no-call registry (Goh et al. 2015). (Goldfarb and Tucker 2011a) show that the European Union (EU)s privacy policy reduces the online display ads effectiveness. On the other hand, different components of a privacy law may incur different effects, for instance, granting users control over re-disclosure encourages the spread of adoption of genetic testing, whereas privacy notification deters individuals from obtaining genetic tests (Miller and Tucker 2017).

The fourth stream of research examines a wide range of methodologies that regulatory bodies and firms may engage to address privacy concerns. These methods fall under two broad categories: without data obfuscation and with as in our research. Without data obfuscation, these methods largely involve firms altering consumers privacy perceptions, hence alleviating privacy concerns. Examples include altering the order of survey questions (Acquisti et al. 2012), revealing other consumers attitudes towards privacy (Acquisti et al. 2012), altering the labels of privacy-protecting options (Adjerid et al. 2018), offering opt-in/out options (Kumar et al. 2014), granting enhanced privacy controls over, for instance, personally identification information (Tucker 2013), allowing customers to remain anonymous with a cost (Conitzer et al. 2012) or granting aggregate instead of granular information (Sandıkçı et al. 2013). Consumers themselves may also take actions to preserve privacy, such as declining to answer certain survey questions, concealing addresses, or deflecting marketing solicitation (Hann et al. 2008). Governments all around the world are also providing regulatory protections, such as do-no-call registry (Goh et al. 2015) and instating state genetic privacy laws (Miller and Tucker 2017).

Other methodologies, on the other hand, leverage obfuscation to the original data or query outputs. The premise in this type of research is that an entity, data collector in our setting, would perform data transformation to preserve consumer privacy before releasing the data to a third party, an advertiser for instance, while ensuring that the data remain useable. Our work is closely related to this stream of work. Hence, we provide a more thorough survey of this stream next, which can be further broken down into two sub-streams based on the assumptions made when developing the relevant techniques (Clifton and Tassa 2013).

2.2. Privacy-preserving Methodology - Syntactic Models

In syntactic models, the assumption is that the entity performing the transformation knows the type of threat that a stalker or a malicious entity is going to perform on the shared data, and accordingly transforms the data to curtail that specific threat. The seminal work in this area was the concept of k-anonymity (Samarati and Sweeney 1998) aimed at columnar data to ensure that given a column, there would be at least k records which take the same columnar value. This would ensure that a consumer is protected from a re-identification threat, that is, his/her record cannot be completely identified even if a stalker has some background knowledge, usually a subset of the consumer's column values.

Studies have shown that k-anonymity is NP hard and suffer from the curse of dimensionality (Meyerson and Williams 2004). Variations of the concept of k-anonymity and heuristics to approximate k-anonymity were hence proposed (Aggarwal et al. 2005). Since k-anonymity primarily focuses on the re-identification threat, it is susceptible to sensitive attribute inference when a stalker aims to only infer a particular column of a user rather than completely re-identify all the columnar values. ℓ-diversity (Machanavajjhala et al. 2006) and confidence bounding (Wang et al. 2007) were proposed to address these shortcomings. ℓ-diversity accomplishes this by obfuscating data such that sensitive attributes are well represented for each user, while confidence bounding limits a stalker's confidence of inferring a sensitive value to a certain threshold. In the context of mobile location data, the above methodologies were shown to suffer from the curse of high dimensionality (*Aggarwal* 2005), reducing an advertiser's utility. To address this, variations of k-anonymity such as k^m-anonymity (Terrovitis et al. 2008) and complete k-anonymity (Bayardo and Agrawal 2005) have been proposed for high dimensional transaction data. However, these techniques only address re-identification threats and are still vulnerable to sensitive attribute inference. Further, while these techniques work well for high dimensional data, they do not explore obfuscation of temporal information crucial in extracting behavioral information from location data. Next, we will review some of the recent syntactic models proposed for obfuscation of location data.

Researchers, primarily from Computer Science, have proposed extensions of these traditional heuristics to preserve privacy in simulated/synthetic location data (Chen et al. 2013, Terrovitis et al. 2008, Abul et al. 2008, Yarovoy et al. 2009), truck/car movements (Abul et al. 2008, Yarovoy et al. 2009), or social media check-in data (Terrovitis et al. 2017, Yang et al. 2018). The seminal work by (Abul et al. 2008) proposes (k, δ) anonymity to perform space generalization on location data. In other words, the trajectories are transformed so that k of them lie in a cylinder of the radius δ. (Yarovoy et al. 2009) further present a variation of k-anonymity for moving object databases (MOD) based on the assumption that MODs do not have a fixed set of quasi-identifiers (QIDs). They define the timestamps of locations as QIDs and propose two obfuscation techniques based on space generalization. These two studies aim at protecting users from re-identification threats. In our problem, we consider both sensitive attribute inference and re-identification threat.

More recently, suppression techniques have garnered attention in obfuscating location trajectory data (Chen et al. 2013, Terrovitis et al. 2008, 2017). In the seminal work, Terrovitis et al. (2008) develop a local suppression obfuscation technique assuming that a stalker holds partial user trajectories similar to the setting of re-identification threat in our study. Building on this work, the authors further propose global suppression, separate from local suppression Terrovitis et al. (2017). Chen et al. (2013) propose $(K, C)_L$ privacy framework that provides privacy guarantees against both identity and attribute linkage threats. The model requires three parameters from a data collector - probability thresholds of a stalker succeeding in the two types of threats along with a parameter corresponding to a stalker's background knowledge. The utility of the data in Chen et al. (2013) and Terrovitis et al. (2008, 2017) is measured by a rudimentary metric - the number of unique location points or frequent sequences preserved in the obfuscated data. In our work, we capture the utility of an advertiser by tying it to a popular business objective of advertisers, POI prediction/recommendation.

2.3. Privacy-preserving Methodology - Differentialy Private Algorithms

This group of research is based of the concept of ϵ-differential privacy proposed by (Dwork and Lei 2009). Differentially private algorithms guarantee that a stalker would make the same inference from the shared data whether or not the individual is included in the data. Since it, unlike syntactic models, does not limit itself to a specific type of attack, it has a much stronger privacy notion than the syntactic models. The transformations performed on the data usually involve pertubation, that is, adding a noise to the data before sharing them (Muralidhar and Sarathy 2006). Another related method is data shuffling, which is usually performed across user rows or columns, such as replacing a subset of a user's record with another user's record to minimize privacy risk. Perturbation, data shuffling, or a combination of them, have also been proposed (Qian and Xie 2015). For instance, Garfinkel et al. (2002) perturb the answer of database query to have the correct answer probabilistically or deterministically embedded in the range of the perturbed answers. Muralidhar and Sarathy (2006) employ data shuffling for confidential numerical data where the values of the condential variables are shuffled among observations, while preserving a high level of data utility and minimizing the risk of disclosure. Schneider et al. (2018) develop a Bayesian probability model to produce synthetic data. Public key encryption, digital certificate, and blinded signatures are also common privacy-friendly tools (Kalvenes and Basu 2006). All of these methods are aimed at columnar data.

In the context of location data, the literature is sparse. Few techniques have been developed to generate synthetic trajectories from a series of differentially private queries (He et al. 2015, Chen et al. 2012). The utility of the data preserved while generating these trajectories usually involve summary statistics, such as the number of unique locations or frequent location patterns. Moreover, owing to the stronger theoretical guarantees that need to be met in developing these techniques, they have been empirically shown to not preserve the truthfulness of the location data, hindering marketer's ability to perform sophisticated data mining tasks (Terrovitis et al. 2017).

In our research, the consumers have explicitly opted in to share their location data with the data collectors and advertisers to enjoy the benefits of behavioral targeting. Hence, we take the route of syntactic model, which is more likely to result in a higher data utility for advertisers and thus consumer than alternative models. We assume that a data collector has sufficient knowledge about the type of stalker threats that a consumer could potentially face by sharing the location data. To minimize the privacy threats, we propose an obfuscation scheme based on suppression that also ensures that an advertiser has enough utility from the obfuscated location data.

Our study distinguishes itself from the prior research in several aspects. From a methodological point of view, we quantify the privacy risk at a consumer level, instead of an aggregate or location level as in the prior literature (Terrovitis et al. 2008, 2017). Next,we measure the utility of the location data by real-world business applications, such as POI recommendation, instead of aggregate, rudimentary measures, such as the number of unique location points or frequent sequences, used in the literature (He et al. 2015). From the standpoint of practical applicability, the proposed framework requires merely one parsimonious input regarding the cardinality of partial trajectories (see Section 4.1.3 for a more detailed explanation) in order to compute the privacy risk and obfuscate the trajectory data. Thus it is intuitive and interpretable to the data collector or any manager. We introduce a heuristic decision factor to provide a data collector with multiple trade-off options between consumers' privacy risk and advertisers' utility. Finally, the majority of the previous work has validated the recommendations on synthetic data (Chen et al. 2013, Terrovitis et al. 2008, Abul et al. 2008, Yarovoy et al. 2009), vehicle movements (Abul et al. 2008, Yarovoy et al. 2009), or social media check-ins (Terrovitis et al. 2017, Yang et al. 2018). In contrast, we validate the proposed framework on the detailed mobile location data of a massive number of consumers spanning across five weeks.

2.4. Location-based Mobile Advertising

Finally, our work is related to mobile marketing and location-based advertising research. Using randomized field experiments, researchers have demonstrated that mobile advertisements based on location and time information can significantly increase consumers likelihood of redeeming geo-targeted mobile coupons (Fang et al. 2015, Molitor et al. 2019, Fong et al. 2015b, Luo et al. 2014). In our framework, we measure utility of location data to an advertiser by considering a popular business application - POI Recommendation. Identifying the next location a consumer is likely to visit based on consumer location trajectories is crucial to perform such behavioral targeting. In a recent paper by Ghose et al. (2018), the authors designed a POI-based mobile recommendation based on similarity in consumer mobile trajectories and demonstrated that such a strategy can lead to a significant improvement in retailer's incremental revenue. Recent studies have also shown that understanding consumers' hyper-context, for example the crowdedness of their immediate environment (Andrews et al. 2016), weather (Li et al. 2017) or the competitive environment (Fong et al. 2015a, Dubé et al. 2017), is critical to marketers' evaluation of mobile marketing effectiveness. Previous studies have also examined consumer perceptions and attitudes toward mobile location-based advertising (Bruner and Kumar 2007, Xu 2006).

3. Data

We partner with a leading U.S. data collector that aggregates location data across hundreds of commonly used mobile apps varying from news, weather, map, to fitness and covering one-quarter of the U.S. population across Android and iOS operating systems.

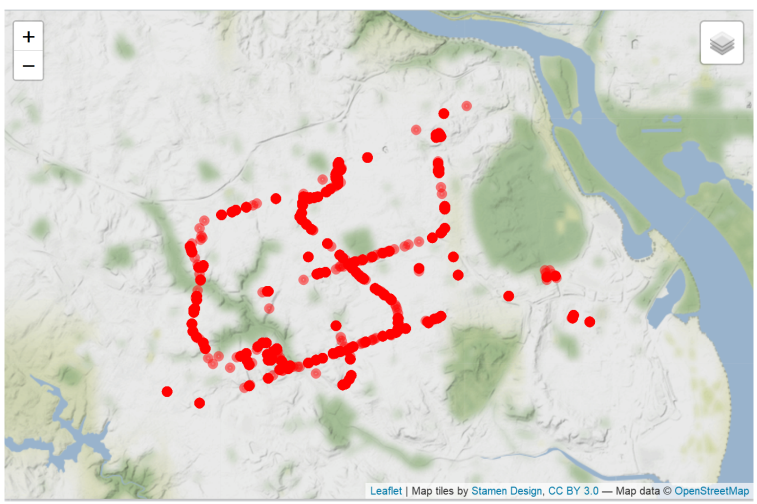


Figure 2 An example of a consumer's footprints with 732 unique locations over the five-week sample period

The data sample under analysis covers a major mid-Atlantic metropolitan region in the U.S. Figure 2 displays the region's map (blurred on purpose due to a confidentiality agreement) and an example of a consumer's footprints with 732 unique locations visited during our five-week sampling period between September and October, 2018. The entire sample includes 940,000 locations from 40,012 consumers. Each row of the data corresponds to a location recorded for a consumer and also contains information about

- Consumer ID: a unique identifier of each consumer.
- Platform ID: Identifier of the consumer's mobile operating system (Android or iOS).

Description	Mean (S.D.)	Min (Max)
Number of locations per person	23.47 (50.26)	2 (1104)
Number of unique locations per person	14.25 (38.12)	2 (963)
Overall duration (in hours)	272.97 (278.25)	0.05 (759.27)
Duration at each location (minutes)	27.96 (45.99)	1.6 (359.23)
Distance between locations (in km)	1.89 (3.89)	0.02 (75.49)

Table 1 Summary statistics of the location data sample under analysis

- Latitude and longitude (i.e., geo-coordinates) of the location.
- Timestamp: the beginning time at the location.
- Time spent: The amount of time spent at the location.

We randomly sample 50% of all consumers in the data (20,000 consumers) and all their location data for training and cross-validating our machine learning models, with details offered in Section 5.1 and Appendix A. Based on the models and parameters trained, we then conduct the focal analysis using the remaining 50% of the data with 20,012 consumers and their location data. Table 1 displays the summary statistics of the data. On average, a consumer visited from 2 to 963 unique locations tracked by the data. To reduce smartphone battery drainage, data redundancy, and storage cost, each consumer's smartphone is pinged frequently, but only recorded a location when there is a substantial shift in the geo-coordinates. The average duration at each location is 27.96 minutes. And the Euclidean distance between any two consecutively tracked locations is 1.89 km on average after converting the locations' latitudes and longitudes to the Universal Transverse Mercator (UTM) coordinates.

The literature on privacy-preserving sharing of location data, primarily from the Computer Science discipline, has tested the methodologies on either simulated data (Chen et al. 2013, Terrovitis et al. 2008, Abul et al. 2008, Yarovoy et al. 2009), vehicle movements (Abul et al. 2008, Yarovoy et al. 2009), or social media check-ins (Terrovitis et al. 2017, Yang et al. 2018), also merely over a short period such as 24 hours. We make an initial effort to develop a privacy-preserving framework for, and validate it on, a real-world human physical movement data across a large population. Such data are automatically tracked in real time by mobile devices, often via wifi, beacons, and GPS etc. multi-technology multilateration with an accuracy radius of merely 20 meters. They are thus much more precise than cell tower tracking often with an accuracy radius of a few kilometers, social media geo-tags known for its sparsity and inaccuracy, or consumers' self check-ins that rely on consumers' manual labor and willingness to check in at any location. The mobile location data under our study are also more representative of the general population than taxi or public transportation data, hence much more valuable to advertisers and other data users. On the other hand, these data's massive scale and high dimensionality, in our case nearly one million mobile location over just five weeks from one metropolitan region, also entail unique challenges, as discussed earlier, hence imminent needs to develop new frameworks to address these challenges.

4. Methodology

Following the syntactic model's premise in the literature, we consider the definitive notions of privacy in this research. That is, we assume that a data collector has knowledge of the specific type of privacy threats that a stalker advertiser could potentially perform on the shared location data (Clifton and Tassa 2013). While the framework proposed can be extended to other types of stalker threats, we illustrate the proposed framework by considering two commonly encountered types of stalker behaviour - sensitive attribute inference and re-identification threat. We formally define these later in the context of location data in Section 4.1.

Our framework aims at enabling a location data collector to publish data in a privacy preserving manner while ensuring that an advertiser gets sufficient utility. We will introduce the notations first and then formulate the privacy preservation in the context of the location data.

Definition 1 (Trajectory). *A trajectory T_i of a user i is defined as a temporally ordered set of tuples $T_i = \{(l_i^1, t_i^1), ..., (l_i^{n_i}, t_i^{n_i})\}$, where $l_i^j = (x_i^j, y_i^j)$ is a location where x_i^j and y_i^j are the geo-coordinates (i.e., a pair of longitude and latitude) of the geographic location, and t_i^j is the corresponding timestamp, n_i is the number of locations tracked of user i.*

Problem Formulation: We frame the problem of preserving privacy in location data at a consumer level. Assume r_i be the risk of a consumer associated in sharing the trajectory T_i for a specific type of stalker threat, u_i be the advertiser's utility, which relates to the benefit a consumer gets from T_i. A data collector aims to find a transformation $T_i \rightarrow \mathcal{P}(T_i)$, where $\mathcal{P}(T_i)$ is obfuscated consumer's trajectory shared with an advertiser, that minimizes r_i while maintaining u_i. Our transformation is based on suppression of locations in T_i and has two parameters. One is z_i, the suppression score for a user i assigned based on the user's privacy risk, which controls the number of locations in T_i to be suppressed. And the other is $\vec{s_i}$, the suppression weights corresponding to each location in T_i and assigned based on a user's frequency to, time spent at, and distance traveled to each location. In Section 4.3, we will detail a structured grid search to fine tune these two parameters, which do not need to be input by a data collector. Both parameters contribute to the final suppression probabilities assigned to each location in T_i. Hence, the corresponding obfuscated risk and utility of $\mathcal{P}(T_i; \{\vec{s_i}, z_i\})$ are functions of the two suppression parameters

$$r_i = \mathcal{PR}(T_i; \{\vec{s_i}, z_i\})$$

$$u_i = \mathcal{U}(T_i; \{\vec{s_i}, z_i\})$$

where $\mathcal{PR}(.)$ and $\mathcal{U}(.)$ depend on the type of stalker threat and business objective of the advertiser respectively. A data collector's job would be to identify $\mathcal{P}(T_i; \{\vec{s_i}, z_i\})$ that minimize r_i while maintaining u_i. More generally, aggregating over all the users, privacy preservation in location data for our problem setting can be stated as - Given a set of N user trajectories $T = \{T_1, ..., T_N\}$, a data collector aims to find a function or a transformation of T, the trajectories to be shared, $T \to \mathcal{P}(T; \{\vec{s_i}, z_i\}_{i=1}^N)$, such that the privacy risk $E(r_i)$ is minimized while maintaining the data utility $E(u_i)$. [Natasha - is this E an expected value or a function of r or u?] We break this down into the following three sub-problems. The first two problems pertain to the estimation of u_i and r_i based on $\mathcal{PR}$ and $\mathcal{U}$ respectively. The third problem is the identification of the suppression parameters $\{\vec{s_i}, z_i\}$.

Research Question 1 (Risk quantification)**.** ***Given*** *user trajectories T and a stalker threat $\mathcal{PR}$, we aim to* ***find*** *the user-specific risk, $\{r_1, ..., r_N\}$ associated with T. $r_i \in [0, 1]$ would indicate the success rate of a stalker in inferring private information from user i's trajectory T_i.*

Research Question 2 (Utility measurement)**.** ***Given*** *user trajectories T and a business objective $\mathcal{U}$, we aim to* ***measure*** *the utility of the trajectories, u_i to an advertiser.*

Research Question 3 (Obfuscation scheme)**.** ***Given*** *user trajectories T and their corresponding risks, for an advertiser's data utility, $\mathcal{U}(T)$, we aim to* ***find*** *$T \to \mathcal{P}(T; \{\vec{s_i}, z_i\}_{i=1}^N)$, where $\mathcal{P}(T)$ selects a subset of location tuples from each T_i based on the parameters $\{\vec{s_i}, z_i\}$, balancing r_i and utility u_i.*

We introduce our framework with the three components discussed in Fig. 1 addressing the three sub-problems. In Section 4.1, we detail risk quantification for two classes of privacy threats (RQ 1), in Section 4.2, we introduce our business application and computation of location data utility (RQ 2). Finally, in Section 4.3, we propose an obfuscation scheme that provides a balance between privacy risk and data utility. (RQ 3)

4.1. Risk Quantification

The first step of our framework is quantifying each user's privacy risks. To accomplish this, we simulate a stalker's actions and assign its success in obtaining individual's sensitive information as a user privacy risk. Privacy threats could range from using simple heuristics, such as querying the user trajectories, to leveraging more robust machine learning heuristics to predict sensitive user attributes (Li et al. 2007, Yang et al. 2018). In our framework, we consider both simple and sophisticated heuristics. Specifically, we will examine two types of most commonly encountered stalker behaviors. In the first type, a stalker aims to infer the complete set of locations of an individual T_i from the published trajectories $\mathcal{P}(T)$[4] $= \{T_1, , T_N\}$, orchestrating an "re-identification

[4] We denote $\mathcal{P}(T; \{\vec{s_i}, z_i\}_{i=1}^N)$ in short as $\mathcal{P}(T)$

threat" (Pellungrini et al. 2018). [Meghanath - please incorporate this footnote into the text.] With some background knowledge, such as a subset of user's locations $\overline{T_i} \in T_i$, a stalker could query the published trajectories $\mathcal{P}(T)$ to identify a subset of users' $\overline{T} = \{T_1, , T_J\}, \overline{T_i} \in \overline{T_j}; \forall j \in [1 : J], J \leq N$, where $\overline{T}$ consists of all the individuals who have visited the locations in $\overline{T_i}$. The success of the stalker in obtaining T_i would depend on the value of J - lower the value of J, higher the success. Next, exploiting obtained complete individual trajectory, T_i, a stalker could employ a robust machine learning heuristic and extract spatial and temporal information from them to further infer other sensitive information ("sensitive attribute inference") such as home/work address of the user (Yang et al. 2018). In the following sub-sections, we first discuss the information an stalker could extract from published trajectories and then quantify each of these privacy threats.

4.1.1. Trajectory Feature Extraction ($\mathcal{F}(T)$) Before we assess each user's privacy risks, we extract a set of features from the users' trajectories. To accomplish this, we assume that a subset of those who acquire the data from a data collector, thus having access to the complete set of trajectories, exhibit invasive behavior, orchestrate stalker threats. Additionally, from the published trajectories, the stalker could obtain other spatial and temporal information. To replicate the adversarial actions of a stalker, we extract a comprehensive set of features from trajectories studied by the literature (Gonzalez et al. 2008, Eagle and Pentland 2009, Williams et al. 2015, Pappalardo et al. 2016, Ashbrook and Starner 2003, Zheng et al. 2010, Wang et al. 2011). These extracted features, as we will see later in Section 5.3 also help a data collector interpret which features contribute to user risk, gain insights on possible obfuscation schemes, and measure and interpret data utility to advertisers. We categorize the features as,

1. **User Mobility** : This set of features captures the aggregate user mobility patterns based on the locations visited in T_i. We consider the frequency, time spent (Pappalardo et al. 2016), and distance traveled by a user to visit a location (Williams et al. 2015). We also compute other richer mobility features, such as entropy (Eagle and Pentland 2009) and radius of gyration (Gonzalez et al. 2008). A detailed description of these user mobility features is listed in Table 2.

2. **User-Location affinity** : Leveraging the literature on learning significant locations from predicting movement across user trajectories (Ashbrook and Starner 2003, Zheng et al. 2010), we build three user-location tensors: the time spent at each location, frequency of visiting each location, and total distance traveled to each location from the immediate prior locations by a user at a weekly level. Each of these three tensors is of order three—user by unique location by week. We then extract user specific, lower dimensional representations by performing a higher order singular value decomposition (HOSVD) (De Lathauwer et al. 2000) on the three tensors separately. HOSVD is usually applied to extract features from multivariate data with temporal and spatial dimensions

similar to ours (Fanaee-T and Gama 2015). Since the tensors are populated over the locations that these users have each visited, the extracted features, would effectively capture affinity of the users to significant locations.

3. **User-User Affinity** : Prior studies have also predicted user network or social links based on trajectories (Wang et al. 2011). We thus quantify the users' co-location behavior and build user-user affinity tensors based on locations that the users share at a weekly level. We again populate the tensors with the total weekly time spent, average frequency, and distance traveled to the co-visited locations within a week. These would be third order tensors as well—user by user by week. Next, we perform a HOSVD on the three tensors to extract the user specific low dimensional representations indicative of their affinity to other users. The incremental benefit of the affinity features is discussed in Section 5.

Stylized example: We illustrate the above user-location and user-user affinity features using a stylized example. Consider three user trajectories as defined in Definition 1 —$T_1 = \{(A,1),(B,1),(A,2),(A,2)\}$, $T_2 = \{(C,1),(A,1),(A,1)\}$, $T_3 = \{(D,1),(B,1),(C,2)\}$, where A,B,C,D are location identifiers and the granularity of the timestamps is at a week level. That is, $T = \{T_1,T_2,T_3\}$ reveals that these three users visited four unique locations over a period of two weeks. The three user-location tensors would be of size $[3 \times 4 \times 2]$. The frequency matrix for user corresponding to T_1 is $\begin{pmatrix} 1 & 1 & 0 & 0 \\ 2 & 0 & 0 & 0 \end{pmatrix}$, where the rows and columns correspond to the weeks and unique locations respectively. The three user-user location tensors would be of size $[3 \times 3 \times 2]$ and the frequency matrix for user corresponding to T_i would be $\begin{pmatrix} 1 & \frac{(1+2)}{2} & \frac{(1+1)}{2} \\ 1 & 0 & 0 \end{pmatrix}$ where rows and columns correspond to weeks and users in T. The matrix for user 1's trajectory, T_1, is populated based on the co-visited locations, $(A,1)$ with user 2 and $(B,1)$ with user 3. The time and distance tensors can be similarly built in both cases. We then perform a HOSVD on these tensors separately to extract user specific features. We use the first five principal components of the decomposition. Hence for each user and tensor, we have five lower dimensional representation capturing the user-location and user-user affinity. Next, we imitate how a stalker would use the information from the published trajectory data (captured by $\mathcal{F}(T)$) to orchestrate stalker threats.

[Natasha - Meghanath; I think maybe it's a good idea to add a concise summary here about where we have used these features in our subsequent analysis, e.g. use these features as regressors or inputs for sensitive attribute identifications, re-identification task, next POIs predictions when calculating the advertiser utility, and calculating suppression weights - what do you think?]

4.1.2. Sensitive Attribute Inference Using the published trajectories $\mathcal{P}(T)$ and the information extracted from them $\mathcal{F}(T)$, a stalker could aim to infer various sensitive attributes, such as age and gender, from the published trajectories $\mathcal{P}(T)$, thus posing a threat to user's privacy. In

Feature	Description
`average_locations`	Number of locations in T_i averaged weekly.
`average_ulocations`	Number of unique locations in T_i averaged weekly.
`average_distance`	Distance travelled by a user to visit locations in T_i, averaged weekly.
`average_dwell`	Time spent at locations in T_i averaged weekly.
`avg_max_distance` (Williams et al. 2015)	Average of the maximum distance travelled by a user each week.
`freq_rog`, `time_rog`, `dist_rog` (Gonzalez et al. 2008)	Radius of gyrations is the characteristic distance traveled by an individual. $rog_i = \sqrt{\frac{1}{\lvert T_i\rvert}\sum_{j=1}^{\lvert T_i\rvert} w_{ij}(l_{ij} - l^i_{cm})^2}$ $l^i_{cm} = \frac{1}{\lvert T_i\rvert}\sum_{j=1}^{j=\lvert T_i\rvert} l_{ij}$, l_{ij} are the geographical coordinates l^i_{cm} is the center of mass of the user w_{ij} are weights obtained based on frequency, time & distance w.r.t to l_{ij}
`freq_entropy`, `time_entropy`, `dist_entropy` (Eagle and Pentland 2009)	Mobility entropy measures the predictability of user trajectory. $E_i = -\sum_{j=1}^{\lvert T_i\rvert} p_{ij} log_2 p_{ij}$, p_{ij} computed from w_{ij} for time, frequency & distance.

Table 2 Description of user mobility features

this type of threats (Li et al. 2007), the assumption is that a stalker has a model $\mathcal{M}$ to infer the attributes from $\mathcal{P}(T)$ (Yang et al. 2018). Intuitively, the more certain a stalker is about a user's sensitive attribute based on $\mathcal{M}$, the higher is the risk to the user's privacy. To replicate these actions by a stalker, we train a supervised model, $\mathcal{M}_{proxy}$ with the features $\mathcal{F}(T)$ extracted from the trajectories. This acts as a proxy for stalker's $\mathcal{M}$ to perform inferences. To quantify each user's risk, we assign the certainty of identifying a sensitive attribute from the user's published trajectory data using $\mathcal{M}_{proxy}$. In this research, we illustrate the method by inferring two sensitive attributes, home address and mobile operating system.

Proxy stalker model ($\mathcal{M}_{proxy}$) In light of its flexibility of handling regression and classification tasks, and its competitive performance across a wide range of supervised learning algorithms, we use Random Forest (Breiman 2001, Liaw et al. 2002) as $\mathcal{M}_{proxy}$. For each sensitive attribute, we learn a Random Forest, using trajectory features $\mathcal{F}(T)$[5]. The risk associated is then assigned as

[5] We have also compared Random Forest with a number of tree-based and boosting classification methods – xGBoost (Chen and Guestrin 2016), Conditional inference trees (Hothorn et al. 2015), Adaboost (Hastie et al. 2009); and found that Random Forest provided the best out-of-sample performance for our data.

the certainty of $\mathcal{M}_{proxy}$ in identifying a sensitive attribute. For a classification task, this is the probability of correctly identifying the sensitive attribute. For regression, we use the negative root-mean-square error as the risk. The higher error indicates more uncertainty for the stalker about the sensitive attribute. We also perform a 0-1 normalization in case of regression, such that $r_i \in [0,1]$.

4.1.3. Re-identification Threat Adapting the notion of the risk that a stalker is able to identify a user and associate him/her to a record in published data (Samarati 2001, Samarati and Sweeney 1998), we define re-identification threat in the context of location data. In this type of threat, a stalker tries to re-identify all the locations visited by a user based on some prior knowledge of an (often small) subset of locations visited by the user. For example, in reality, an advertiser may likely have knowledge about a customer's name, home and work addresses from a membership program registration form. Based on this "background knowledge" of two frequently visited location points, a potential malicious activity would be to link such knowledge to the anonymized location trajectory data and try to re-identify the customer's entire location trajectory associated with the name. Formally, this problem can be defined as follows:

Definition 2. *Given the published trajectory data, $\mathcal{P}(T)$ and a subset of trajectory from user i under threat $\bar{T}_i \subseteq T_i$, $|\bar{T}_i| = k$, the stalker aims to identify T_i from $\mathcal{P}(T)$.*

Notice that a data collector does not know user i under threat or the subset $\bar{T}_i$ a-priori. Hence, to quantify the user risk r_i, one would need to account for all possible $\binom{|T_i|}{k}$ subsets of T_i for all N users. For each such subset, $\bar{T}_i$, the probability of an individual being identified completely (i.e., to infer T_i) is $\frac{1}{J}$, where J denotes number of users in N who have visited the locations in $\bar{T}_i$. If a user has not visited locations in $\bar{T}_i$, then the probability of identification would be zero for the subset considered. We quantify a user' re-identification risk as the maximum of these probabilities over all such subsets.

Stylized example: Let $T_1 = \{(A,1),(B,1),(C,2),(C,2)\}$, $T_2 = \{(A,1),(B,1),(A,2)\}$, $T_3 = \{(A,1),(B,1),(C,2)\}$. Assume $|\bar{T}_i| = 2$; to compute risk for T_1 across both weeks, we consider the subset $\{(A,B),(B,C),(A,C)\}$. Corresponding probabilities of user identification are $\{\frac{1}{3},\frac{1}{2},\frac{1}{2}\}$, risk assigned is $max(\frac{1}{3},\frac{1}{2},\frac{1}{2}) = \frac{1}{2}$

Speed-up heuristic : While the re-identification risk can be exactly computed for a given k, it is computationally inefficient with a complexity of $O(\binom{n_i^{max}}{k} \times N)$ where n_i^{max} is maximum number of unique locations visited by a user. To speed up the computation, we leverage a recent work (Pellungrini et al. 2018) that empirically shows the predictability of the re-identification risk for a given k using mobility features. The main idea is to learn a supervised algorithm (Random Forest regressor in (Pellungrini et al. 2018)) by building a set of mobility features, similar to $\mathcal{F}(T)$ discussed in Section 4.1.1. We adopt this idea by augmenting the mobility features with the user-user and user-location affinity features to learn a similar model. We analytically compute the risks for a subset of the users and then use a trained model to approximate the risk for the rest of the users in our empirical study. We will detail this and justify our model choices in Section A.

This completes assignment of risk $\{r_1, ..., r_N\}$, $r_i \in [0,1]$ for a given set of trajectories T, under the above two types of privacy threats $\mathcal{PR}$ as defined in RQ 1.

4.2. Location Data Utility

Having quantified the user risks associated with two commonly encountered, definitive privacy threats on location data, we next examine the utility that an advertiser would derive from published trajectories. The behaviorally rich nature of location data enables advertisers to derive insights and perform various targeted marketing activities leading to monetary benefits. In this work, we consider a popular business use case for an advertiser - POI recommendations (Ashbrook and Starner 2003). The underlying idea is to leverage historical user preferences revealed in user trajectories to predict locations that a user is most likely to visit in the future. Predicting a user's future locations accurately would enable an advertiser to target him/her with relevant, contextualized marketing campaigns (Ghose et al. 2018). To this end, we quantify an advertiser's utility in the context of POI-based targeting by learning a recommender model. Intuitively, more accurate POI predictions will render better targeting with the learned recommendation model, leading to a higher utility for the advertiser. Hence, we quantify the data utility u_i for a consumer as the accuracy of the future predictions made by the recommendation model. Next, we discuss the POI model in more detail.

4.2.1. POI Recommendation Model Most recommendation models focus on the collaborative filtering process—identifying other users with similar historical preferences to infer a user's future preferences (Bobadilla et al. 2011). This is consistent with human social behavior—people tend to consider their acquaitances' tastes, opinions, and experiences when making their own decisions. We thus imitate an advertiser's leverage of the location data for POI-based targeting and compare a number of recommendation models in the literature (comparison in Appendix A). We focus our discussion on the best-performing nearest neighborhood-based (NN) learning technique. Simply put, the main idea in NN is to identify the m users most similar to an individual (namely m neighbors) and utilize their locations to predict the focal individuals future preferences. The similarity is computed based on the visited locations that reveal each user's preferences. The set of features extracted from the user trajectories in Section 4.1.1 $\mathcal{F}(T)$ serves this purpose. And to find the m most similar users, we compute the cosine similarity between two users i, j with trajectories T_i, T_j represented by features $\mathcal{F}(T_i), \mathcal{F}(T_j)$

$$sim(\mathcal{F}(T_i), \mathcal{F}(T_j)) = \frac{\mathcal{F}(T_i) \cdot \mathcal{F}(T_j)}{||\mathcal{F}(T_i)||||\mathcal{F}(T_j)||} \tag{1}$$

After identifying the m most similar users for an individual i, denoted by M_i, we aggregate and rank the locations visited by M_i based on a combination of visit frequency and their similarity to individual i. For each user $j \in M_i$, location $l \in T_j$, let f_j^l denote the number of times the user j visited location l, the rank of a location l for a user j is determined by the following quantity

$$o_{ij}^l = \sum_{l=1}^{|T_j|} \frac{f_j^l}{\sum_l f_j^l} sim(\mathcal{F}(T_i), \mathcal{F}(T_j)) \tag{2}$$

In the above equation, $\frac{f_j^l}{\sum_l f_j^l}$ is the normalized frequency of visits at a user level for a location. Intuitively, Equation 2 ensures that an individual i is most likely to visit the most frequently visited location of the most similar user. We further aggregate o_{ij}^l across all the users who visited the location l in M_i by computing the mean of o_{ij}^l , that is

$$o_i^l = \frac{1}{\sum_{j=1}^{|M_i|} 1(l \in T_j)} \sum_{j=1}^{|M_i|} 1(l \in T_j) \cdot o_j^l \tag{3}$$

where $1(j \in T_j)$= 1 if user j has visited location l and zero otherwise. Higher the value of o_i^l, more likely that an individual i visits location l in the future. The next k locations an individual i is most likely to visit correspond to the top k ranked locations in M_i. The utility of an advertiser for each consumer location trajectory T_i is measured as the accuracy of the predictions made by the recommender for the different values of k. This is computed by the widely used information retrieval metrics that assess the quality of the recommendations - Average Precision at k ($AP@k$) (Equation 4), Average Recall at k ($AR@k$) (Equation 5) (Yang et al. 2018). These are further aggregated to compute $E(u_i)$ by averaging across all the users $MAP@k$ and $MAR@k$.

More specifically, let $L_i = \{l_i^1, l_i^2, , l_i^{k'}\}$ be the true locations that a user has visited in the future and $\overline{L_i} = \{\overline{l_i^1}, \overline{l_i^2}, ..., \overline{l_i^k}\}$ be the predictions made by the NN recommender ordered by the ranks based on Equation 3. First, the average precision AP_i^k and average recall AR_i^k for user i with top k recommended locations are given by

$$AP_i^k = \frac{1}{|L_i \cap \overline{L}_i|)} \sum_{j=1}^{k} \frac{|L_{1:j} \cap \overline{L}_{1:j}|}{|L_{1:j}|)} \tag{4}$$

$$AR_i^k = \frac{1}{|L_i \cap \overline{L}_i|)} \sum_{j=1}^{k} \frac{|L_{1:j} \cap \overline{L}_{1:j}|}{|L_i|)} \tag{5}$$

The intuition is that AP_i^k measures the proportion of the recommended locations that are relevant, while AR_i^k measures the proportion of relevant locations that are recommended. Then, $MAP@k$ and $MAR@k$ are computed by averaging AP_i^k and AR_i^k across all the users. The parameter m, number of the most similar neighbors used for location predictions, is selected by performing a five-fold cross-validation aimed at maximizing the accuracy of the recommendations. This is a technique commonly used in statistical learning literature to ensure a good out-of-sample performance (Friedman et al. 2001). We will detail how we have selected the number of the most similar neighbors in our empirical setting in Section 5.1.2.

4.3. Suppression Scheme

The last step in our framework is to address RQ 3 – devising an obfuscation scheme that would balance the privacy risks and utilities of the published trajectories to advertisers. Given the unique properties of the location data of high dimensionality (due to the large number of locations visited by the users), sparsity (fewer overlap across the locations visited by the users) and sequentiality (the temporal order of the visited locations by users), employing the traditional obfuscation techniques proposed for relational data, such as k-anonymity (Samarati and Sweeney 1998), ℓ-diversity (Machanavajjhala et al. 2006), and confidence-bounding (Wang et al. 2007) would be computationally prohibitive and significantly reduce the utilities of the resulting obfuscated data (Aggarwal 2005).

On the other hand, those techniques devised specifically for trajectory data are often complex for a data collector to interpret and apply in practice. For instance, the $(K, C)_L$ privacy framework (Chen et al. 2013), an anonymization scheme built on the principles of k-anonymity requires multiple parameter inputs from a data collector. The parameters may include the threshold of the stalker's success probability, or the stalker's background knowledge in each type of threat. For instance, LSup (Terrovitis et al. 2017) that suppresses the trajectories by quantifying the trade-off between the utility to advertisers and the risks associated with each location visited by a user requires similar inputs. Given the complex nature of such heuristics, setting these parameters and interpreting the resulting obfuscations for practical purposes is non-trivial. Further, the current techniques do not provide the flexibility for a data collector to make a choice between multiple obfuscation schemes.

Addressing these critical challenges, we develop $T \to P(T, \{\vec{s_i}, z_i\}_{i=1}^N)$, a personalized user-specific suppression technique that is interpretable to the data collector since it requires merely a single parameter input that corresponds to a stalker's background knowledge for re-identification threat – the cardinality of the set $\overline{T_i}$ as defined in Definition 2. Intuitively, this means that our method requires only one parameter input regarding the number of a user's locations already known to the stalker (and thus used as the background knowledge to re-identify the user's full trajectory). In addition, the suppression technique requires no input parameters for the sensitive attribute threat. Furthermore, we will offer the data collector the flexibility to choose across multiple, interpretable obfuscations for each type of stalker threat by performing a structured grid search on the two user-specific suppression parameters $\{\vec{s_i}, z_i\}$.

4.3.1. Personalized Data Suppression ($\mathcal{P}$) Our obfuscation scheme is based on the idea of suppressing a user's location trajectory T_i. We introduce and identify two consumer specific parameters $\{\vec{s_i}, z_i\}$ to achieve this. The parameter z_i, the suppression score assigned to each user controls the number of locations to be suppressed for a given T_i and within T_i, we further assign weights to each tracked location through $\vec{s_i}$ to denote the likelihood of a specific location being suppressed. By definition, $\vec{s_i} \in \mathbb{R}^+$ and $z_i \in \mathbb{R}^+$. Given the setting, a data collector, to identify $\{\vec{s_i}, z_i\}$ that reduce the consumer's risk r_i and maintain u_i could perform a search by considering a random grid of positive values for z_i and $\vec{s_i}$ and assess r_i and u_i for each such specification of parameters. However, this naive search would be computationally inefficient and depending on the grid of values chosen, there may not be a set of parameters $\{\vec{s_i}, z_i\}$ that satisfactorily balance the risk-utility trade-off. Hence, a sophisticated grid search is needed.

A more structured approach to identify $\{\vec{s_i}, z_i\}$ would be to consider a grid that would ensure a reduction in consumer's risk and then assess u_i to pick the specification that balances the risk-utility trade-off. Intuitively, more the number of location suppressed in T_i, lesser the trajectories published $\mathcal{P}(T_i)$, meaning a stalker has lesser information to infer private information. For instance, in the extreme scenario when no trajectories are published, both re-identification and sensitive attribute risk would be zero. Further, to ensure similar reduction in risk for a high risk and a low risk individual, the amount of information suppressed would need to be proportional to the corresponding privacy risk. Factoring this observation, in our search strategy, we assign the grid for z_i, the suppression score for a user that controls the number of locations suppressed for each user as $z_i = r_i \times p$, $r_i \in [0, 1]$ where p is a positive grid parameter.

While z_i ensures that individuals are suppressed proportionally based on the risk score, to further limit the information available to perform a stalker threat, the more informative locations within T_i would need to be suppressed with a higher probability. Since the informativeness would be related to the possible features that can be extracted by a stalker from T_i (Section 4.1.1), we assign the weights $\vec{s_i} = \{w_i^1, w_i^2, ..., w_i^{|T_i|}\}$ based on either of the three qualitative measures used in extracting the user-user and user-location affinity – frequency, time spent and distance traveled. To exemplify, let $L_i = \{l_i^1, l_i^2, ..., l_i^{k_i}\}$, be the unique locations in $T_i = \{(l_i^1, t_i^1), ..., (l_i^{n_i}, t_i^{n_i})\}$, $k_i \leq n_i$. Then the weights $\vec{s_i}$ assigned based on the corresponding frequencies $L_i = \{f_i^1, f_i^2, ..., f_i^{k_i}\}$ of L_i are given by $\vec{s_i} = \{\frac{f_i^1}{\sum_{j=1}^{k_i} f_i^j}, \frac{f_i^2}{\sum_{j=1}^{k_i} f_i^j}, ..., \frac{f_i^{k_i}}{\sum_{j=1}^{k_i} f_i^j}\}$. Combining the search strategies of the two parameters $\{\vec{s_i}, z_i\}$, we assign the suppression probabilities for each location in T_i.

In our obfuscation scheme $P(T, \{\vec{s_i}, z_i\}_{i=1}^N))$, for a trajectory T_i, the unique locations are independently suppressed with probabilities

$$z_i + z_i \times w_i^1, z_i + z_i \times w_i^2, ..., z_i + z_i \times w_i^{k_i}, w_i \in \vec{s_i} \tag{6}$$

Note that with the proposed search strategy since r_i and w_i can be computed apriori[6], the suppression probabilities for each location in T_i would depend only on the grid parameter p. For a value of p, the probabilities would ensure that users at higher risk have more informative locations suppressed in their published trajectories compared to lower risk users. This search strategy constructed to limit a stalker's ability to invade private information would also adversely affect the ability of a butler advertiser's utility from $\mathcal{P}(T)$. For instance, in the extreme scenario when all user risks $r_i = 1$ and p is reasonably high(complete suppression[7]), all locations would be suppressed, $\{\mathcal{P}(T_i)\} = \mathcal{P}(T) = \emptyset$ resulting in no utility to the advertiser or threat to user privacy. A similar inference can be made when $p = 0$ (no suppression) when $\{\mathcal{P}(T_i)\} = \mathcal{P}(T) = T$ resulting in the high data utility and high privacy risks. Noting these two extreme scenarios, we empirically determine the specification of suppression parameters $\{\vec{s_i}, z_i\}$ by varying the grid parameter p and the corresponding shared trajectories $\mathcal{P}(T)$ that provide a utility-risk balance accounting for both risks detailed in Section 4.1.

Our obfuscation scheme has two main advantages. First, the structured grid search strategy based on varying the grid parameter p provide a data collector with multiple trade-off choices to pick from. Second, the two identified parameters $\{\vec{s_i}, z_i\}$ provide consumer level interpretability of the obfuscation of T_i to a data collector setting us apart from the earlier obfuscation schemes tackling a similar problem. By fine-tuning $\{\vec{s_i}, z_i\}$, our final goal is to understand, measure, and optimize the trade-off between data utility ($\mathcal{U}$) and privacy risk ($\mathcal{PR}$) in a meaningful way.

5. Results

5.1. Empirical Setup

An overview of our empirical setup is presented in Figure. 1. As discussed above, in our obfuscation scheme, a user's locations are suppressed based on the probabilities assigned to each T_i (Equation 6). These probabilities are derived from $\{\vec{s_i}, z_i\}_{i=1}^{N}$ that depend on a single[8] parameter p. In our empirical study, we vary $p \in \mathcal{G}_p = \{0, 0.1, ..., 1\}$ to compute the quantities that reflect the overall data utility ($MAP@k$, $MAR@k$) and privacy risk for each p. We discuss these in detail below.

5.1.1. Risk Computation We estimate the overall user risk for various levels of obfuscations ($p \in \mathcal{G}_p$). For each type of threat ,we first quantify the user-level risk by extracting features $\mathcal{F}(T)$ for the entire sample ($p = 0$) without any obfuscation. These risks are baseline individual risks when

[6] r_i estimated as detailed in Section 4.1 and w_i from the data depending on frequency, recency or time spent at each location

[7] Note that $r_i \in [0,1]$ and $w_i^j \in [0,1]$ by construction. In our empirical study, whenever $(z_i + z_i \times w_i^j) \times p > 1$, we suppress them with probability 1

[8] We compute r_i and $\vec{s_i}$ prior to obfuscation. r_i is estimated as detailed in Section 4.1 and $w_i, w_i \in \vec{s_i}$ from the data depending on frequency, recency or time spent at each location

the location data are stored or shared as is (Part A in Figure 1). Based on these baseline risks, we perform user-level obfuscation by varying $p \in \mathcal{G}_p$ (Part B in Figure 1). Note that each p would result in a new specification of the suppression parameters $\{\vec{s_i}, z_i\}_{i=1}^{N}$, assignment of suppression probabilities to trajectories in T, resulting in varied obfuscations $\mathcal{P}(T)$. We then repeat the process – extracting features and re-calculating the user risks on each obfuscated trajectory sample $\mathcal{P}(T)$. Finally, we compute the overall privacy risk $E(r_i)$ by averaging the risks across all users. To obtain a consistent estimate, we use bootstrapping with 20 trials for each $p \in \mathcal{G}_p$.

In the sensitive attribute threat , we consider two sensitive attributes - home address and mobile operating system predictions from the original trajectory data. To train the predictive model, we assume, consistent with the literature, that a stalker has access to the sensitive attributes of a subset of users, such as via surveys (Yang et al. 2018, du Pin Calmon and Fawaz 2012). This assumption is as if the stalker has access to a training sample with known trajectories and sensitive attributes, thus capable of inferring sensitive attributes from new trajectories.

To simulate this process, we split the data into two random samples: 50% training set (T_{train}) with 20,000 users, and 50% test set (T_{test}) with 20,012 users. We use T_{train} to train the predictive model for the privacy risk. Recall that in our risk quantification (Section 4.1), we use Random Forest regressor to predict the risks of home locations and use Random Forest classifier to predict mobile operating system and re-identification risks. To avoid over-fitting, we perform a five-fold cross-validation on T_{train} and pick two optimal hyper-parameters specific to the Random Forest – the maximum number of features in the tree and the number of trees (see the Appendix A for more details). Cross-validation ensures that the model produces better out-of-sample predictions (Friedman et al. 2001). Once the model is trained, we apply it to estimate the risk in each privacy threat on T_{test}. In Figure 3, we report the mean of these risks computed for each p across all 20,012 users in T_{test}.

To compute the privacy risk of re-identification threat, we need to know the number of locations in each user's trajectory that a stalker has already known. In our empirical study, we assume this number to be 2 and hence set $|\bar{T}_i| = 2$ in Definition 2.[9]

5.1.2. Utility Measurement Next, we compute the data utility under different obfuscations. To do so, we compute the performance of a neighborhood-based collaborative filtering recommendation heuristic to accurately predict future user locations. To assess the predictive accuracy, we treat the locations actually visited by each user in the fifth week as the ground truth and train the recommender model to predict these locations.

[9] We make this assumption for our empirical study. The model learning is generalizable for other values of $|\bar{T}_i|$.

Based on the user risks, we perform varying levels of obfuscation, $p \in \mathcal{G}_p$. We learn a neighborhood-based recommender (Bobadilla et al. 2011) that tunes the number of neighbors via a five-fold cross-validation on the obfuscated data. The model is learned to rank the locations that a user is likely to visit in the fifth week of the observation period. That is, we build the features (discussed in Section. 4.1.1) on the first four weeks of the obfuscated data and tune the number of neighbors[10] to maximize the prediction accuracy. Then, we compute the overall utility $E(u_i)$ —$MAP@k$, $MAR@k$ for $k = \{1, 5, 10\}$[11]. Intuitively, $MAP@1$ and $MAR@1$, for example, represent an advertiser's utility to predict the next location that a user is most likely to visit in the fifth week based of the recommender model learned on the obfuscated data. Similar to what we have done for the risk metrics, we perform 20 trials for each p and report the mean and 95% confidence intervals of the utility metrics (Figure 3). A more detailed explanation of the utility computation is available in the Appendix B.

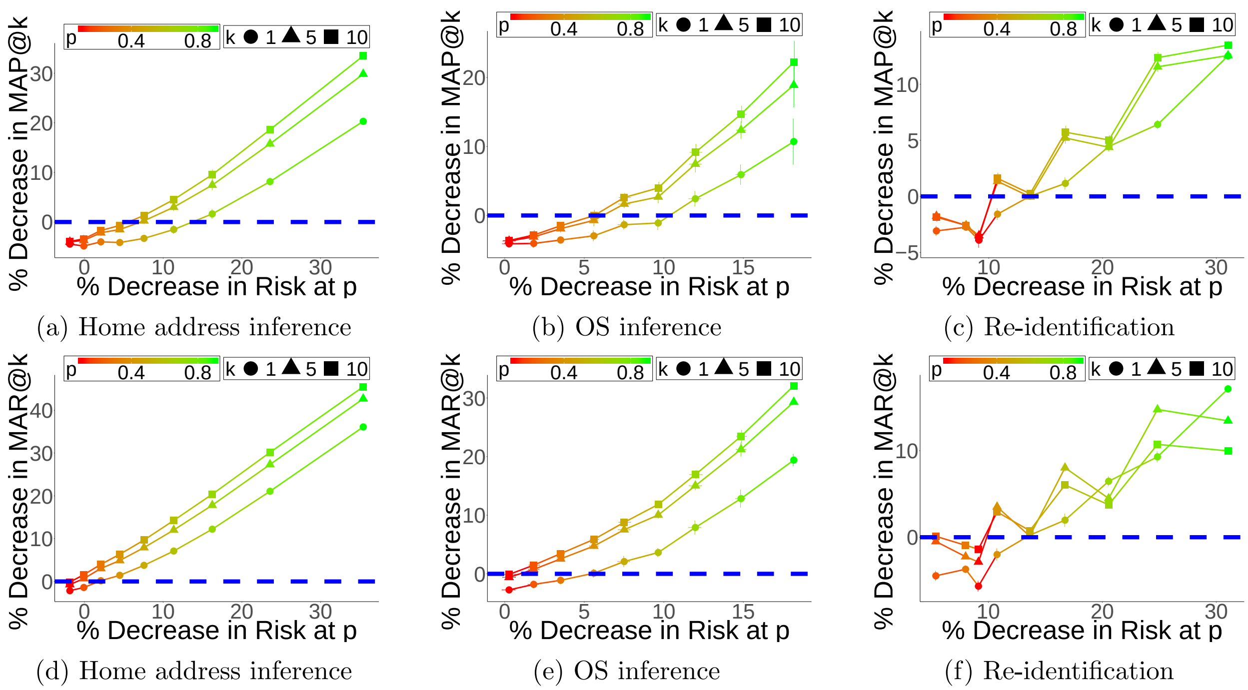


Figure 3 **Proposed framework** - $MAP@k$ **and** $MAR@k$ **for varying** p

5.2. Privacy-preserving Obfuscation

In Figures 3a, 3b, 3c we visualize the trade-off between data utility and privacy risk. The locations are suppressed based on the suppression probabilities assigned to each T_i that depend on the

[10] We use the following grid for number of neighbors - $\{5, 10, 25, 50, 100, 200\}$

[11] The learned recommender model can be used to compute $MAP@k$, $MAR@k$ for other values of k as well. We consider $k = \{1, 5, 10\}$ to illustrate the method's efficacy.

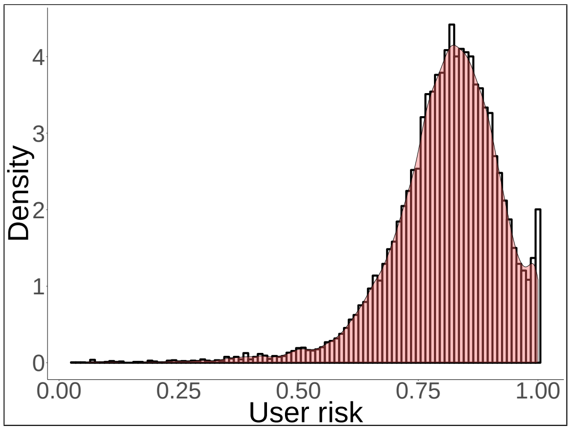
single parameter p, derived from the two parameters $\{\vec{s_i}, z_i\}_{i=1}^{N}$ in our obfuscation scheme (Eq 6, Section 4.3). We discuss the results where $\vec{s_i}$ (the suppression weights assigned to each location of a consumer's trajectory) is assigned based on the frequency of their visits to each location[12]. In the Figures 3a, 3b, 3c, the X and Y axes display the percentage decrease in the aggregate risk and $MAP@k$ from the original sample ($p = 0$) with no obfuscation for each $p \in \mathcal{G}_p$ respectively. We plot these for $k = \{1, 5, 10\}$. Intuitively, the higher the value of X-axis, the more the decrease in the overall risk and hence better preservation of privacy. On the other hand, the lower values of Y-axis correspond to a lesser decrease in the utility of the obfuscated data compared to the original data, suggesting a similar utility for the advertiser even after obfuscation. A data collector who aims to trade off between utility and privacy, is thus presented with multiple choices in our framework. Ideally, a good choice for obfuscation would be the values of p that correspond to a higher value along the X-axis and a lower value along the Y-axis. In the figures, the horizontal blue line, with no decrease in data utility from obfuscation indicates these choices. Similar insights can be drawn from figures 3d, 3e, 3f where we compare the percentage decreases in $MAR@k$ to the percentage decreases in the aggregate risk.

In all the figures in Figure 3, we observe that as we increase p, both the quantities decrease in aggregate risk (X-axis) and decrease in performance measures (Y-axis) increase. This is expected since an increase in p, for the same user risk scores, more locations get suppressed, meaning more information loss to an advertiser's utility as well as a privacy threat. For a given percentage decrease in risk, we observe a lesser corresponding percentage decrease in performance. This can be explained by the framework's obfuscation parameters $\{\vec{s_i}, z_i\}_{i=1}^{N}$ which are varied based on the user risk scores that capture the success of a stalker threat. This risk-based obfuscation would penalize and cause more information loss to the stalker's adversarial intent compared to the utility. The figures also emphasize the proposed framework's flexibility to provide a data collector with several interpretable choices for obfuscation. Further, since our obfuscation scheme works by suppressing a set of location tuples instead of randomization (Yang et al. 2018) or splitting (Terrovitis et al. 2017), this would also have potential benefits to the server costs incurred by an advertiser in storing and analyzing the location data.

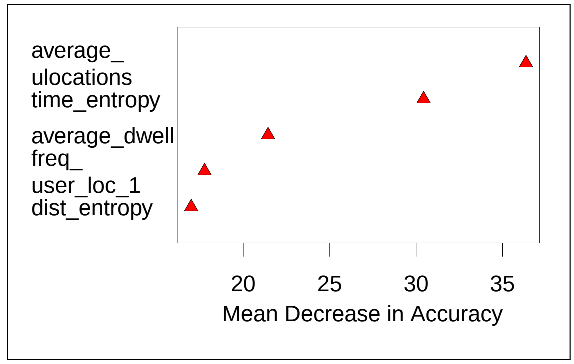
5.3. Personalized Risk Management

Here, we discuss insights a data collector can gain from the risk quantification we perform for different types of threats prior to obfuscation. In Section 4.1 we extract a comprehensive set of features ($\mathcal{F}(T)$) and then build predictive models to assess user privacy risk associated with two

[12] We have also conducted an empirical study where we assign $\vec{s_i}$ based on recency and time spent at each location which are presented in Appendix D



(a) User risk density plot - OS inference threat



(b) Feature importance, OS inference model

Figure 4 Personalized Risk Management Insights

specific privacy threats—sensitive attribute (individual home address and operating system) and re-identification threats. In Figure 4a, we visualize the density plot of user risk on non-obfuscated trajectory data. The user risks were computed by mimicking an stalker's aim to infer the operating system of a user. A data collector, by looking at Figure 4a can have a precise estimation of the distribution of user privacy risk. For example, a majority of users have a relatively high risk ($\geq$ 0.75) of their sensitive attribute being inferred if there was no obfuscation performed.

Similarly, we find that in the case of home address inference, the average estimated user risk is 0.84. By assessing the error of the Random Forest regressor learned to predict the home address, we observe that a stalker on average could successfully identify a consumer's home address within a radius of 3,900 metres $\approx$ 2.5 miles (Discussed in more detail in Appendix A). Further, the average risk of re-identifying an individual's entire trajectory by knowing only two randomly sampled locations is 0.49, meaning a 49% chance of success for a stalker.

These privacy risks associated with non-obfuscated data can be curtailed by a data collector by using our framework's obfuscation scheme. For instance, the risk associated with operating system inference could be bought down by 10% (Refer to Figures 3b, 3e $p = 0.6$) while still being able to fully preserve the data utility with regard to the POI@1 performance. As a follow-up step, by implementing the POI recommendation strategy in the real world, a data collector can also measure the monetary value of an individual trajectory, and compare it with the user-specific privacy risk to better understand customer lifetime value (Berger and Nasr 1998) and personalize customer relationship management.

In addition, prior to obfuscation, a data collector could look at feature importance in our model. In Figure 4b we visualize the top five important features of the Random Forest we train to compute user risks we plot in Figure 3b. From the figure, a data collector can infer that temporal information of the trajectories (`time_entropy`, `average_dwell`) contributes significantly to model predictive performance. Hence, a possible obfuscation scheme that removes the timestamps (even partially) in the trajectories preventing the stalker to construct the temporal features could reduce the user risk considerably. Similar insights can be gained by analyzing the risk scores related to other stalker threats.

5.4. Model Comparisons

We compare our framework's obfuscation scheme with eight different baselines corresponding to three types of obfuscation schemes – obfuscation rules derived from timestamps of user locations, alternate suppression schemes based on user risk and the most recent work in syntactic models LSup and GSup (Terrovitis et al. 2017).

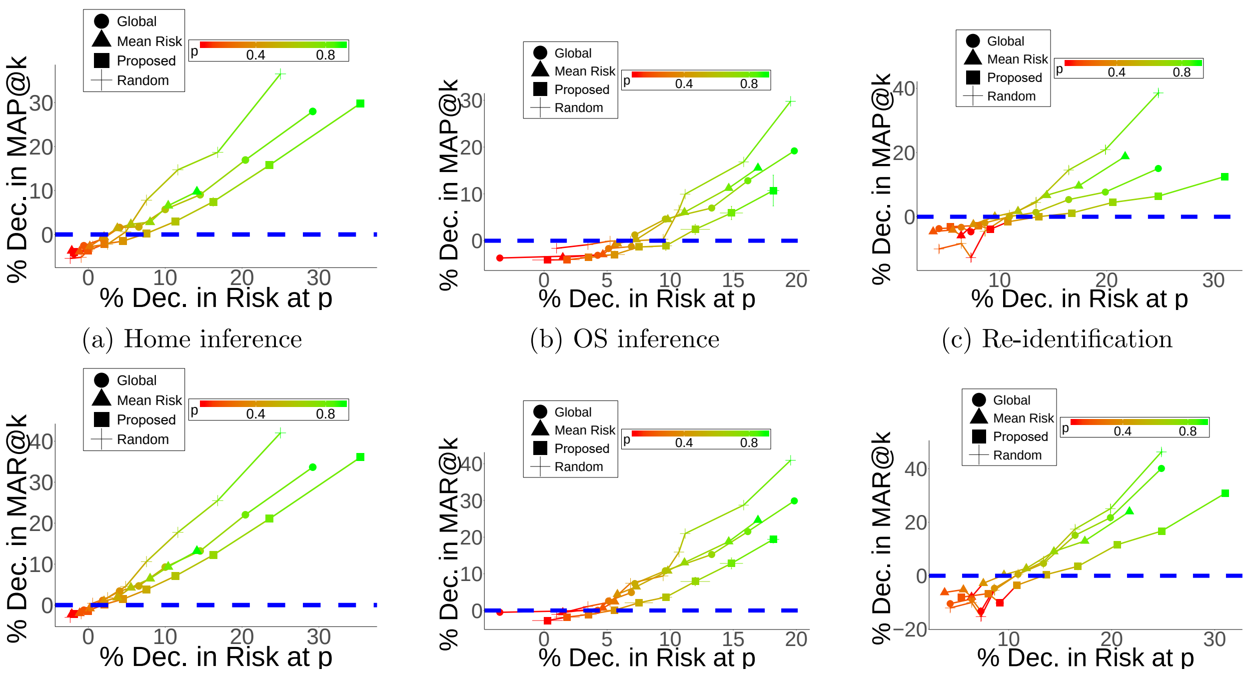
Obfuscation rule	% Decrease Home address risk	% Decrease Operating system risk	% Decrease Re-identification risk	% Decrease Utility (MAP@1)	% Decrease Utility (MAR@1)
Remove Sleep hours	2.43	12.51	1.41	11.83	12.69
Remove Sleep and working hours	10.72	21.84	21.49	34.45	23.72
Remove time stamps	13.45	25.49	0	33.16	32.97

Table 3 Alternative Schemes: Rule-based Obfuscation

5.4.1. Comparison to Rule-based Obfuscations: We derive a few practical rules for obfuscation based on the timestamps of user locations in the location data. In the absence of a privacy-friendly framework, a data collector could perform obfuscation by choosing to 1) Remove all the locations during the usual sleeping hours (10 PM - 7 AM) on all days, 2) Remove locations in sleeping hours and working hours (9 AM - 6 PM) on weekdays, or 3) Remove timestamps of locations entirely before sharing the data. The three time-based rule obfuscations would reduce the amount of information that can be extracted from the shared location data and hence would adversely affect the advertiser's utility and adversarial intent to invade consumer's privacy. For instance, if the timestamps of the location data were to be removed both mobility features (refer Table 2) – `time_entropy`, `time_rog`, `average_dwell` and the User-User, User-Location affinity features (refer Section 4.1.1 based on time spent by a user at a location cannot be computed.

The decrease in risks for the two threats and decrease in utility for each of these obfuscations are presented in Table 3. As expected, there is a decrease in both risk and utility. In the home address inference threat (Figure 3a, $p = 0.7$, $k = 1$), we find that a risk to user privacy can be reduced by

15% (maximum decrease when compared to rule-based heuristics) with less than 1% decrease in $MAP@1$ (minimum decrease). A similar trend is observed in re-identification threat (Figures 3c, 3f). In the operating system inference (Figure 3a, $p = 0.9$, $k = 1$), we observe that risk is reduced by $\approx$ 18% compared to 25.49% when timestamps are removed. However, this is achieved with a lesser decrease in utility $\approx$ 10% using the proposed framework when compared to the 33%. Overall, we find a better choice set for the trade-off justifying a need for a privacy-friendly framework to assist a data collector to share location data in a privacy-friendly way.



(a) Home inference (b) OS inference (c) Re-identification

(d) Home inference (e) OS inference (f) Re-identification

Figure 5 Proposed framework vs risk-based obfuscations - $MAP@1$ **and** $MAR@1$

5.4.2. Comparison to Risk-based Obfuscations We compare the proposed obfuscation framework to three alternate suppression baselines. These are devised to show the efficacy of user risk quantification and personalized local suppression (achieved by introducing and identifying user-specific parameters $\{\vec{s_i}, z_i\}$) of trajectories performed in our framework.

1. **Random** - In this baseline, we do not perform suppression of locations at a user level. Instead of hiding location tuples in T_i based on $z_i = r_i \times p$ and suppression weights $\vec{s_i}$, we randomly suppress locations in T. We suppress the same number of location tuples as in our framework's obfuscation scheme to make it comparable.

2. **Mean Risk** - Here, we perform a user-specific suppression without any variation across users here. We replace the user risk score r_i with the mean $\bar{r} = \frac{1}{N}\sum_i r_i$ as r_i and hide locations using $z = \bar{r} \times p$ and suppression weights $\vec{s_i}$ as described in Section 4.3 for each T_i.

3. **Global** - In this baseline, we suppress a location tuple globally. That is, a tuple in any T has the same chance of being suppressed irrespective of a different user risk threat. This is different from our obfuscation scheme where a tuple may not be suppressed for a less risky but has been suppressed for a high risk user. For each tuple, we assign the mean of user risk scores as tuple risk score, vary p and perform suppression.

We empirically compare the proposed obfuscation scheme to the baselines listed and visualize $MAP@1$ and $MAR@1$ in Figure 5. We observe that, for a given decrease in risk, our framework's obfuscation has the least decrease in utility gain across all three threats. Random baseline, which is an ablation of our obfuscation scheme without the risk quantification step performs the worst among competing models. This justifies a need for threat quantification either at a user-level (Mean Risk and proposed obfuscation) or at a location tuple level (Global). Better performance than Mean Risk baseline shows that a personalized level of obfuscation for each user is necessary. Finally, a higher utility gain over Global baseline emphasizes the need for quantifying and suppressing locations at a user level compared to a tuple level.

5.4.3. Comparison to prior suppression models Finally, we compare the proposed framework to the most recent suppression based syntactic models LSup and GSup proposed by Terrovitis et al. (2017). We observe that in a majority (10 out of 12) of the cases, the proposed framework provides a better trade-off (denoted by green color in Table 4) compared to both LSup and GSup. This improved trade-off come with an added benefit that the obfuscation scheme of the proposed framework only requires one input parameter corresponding to the number of locations of a stalker in the re-identification threat compared to the various parameters required for LSup and GSup. Due to space limitations, we discuss the details of the comparison in Appendix C.

6. Conclusion

Smartphone location tracking has created a wide range of opportunities for data collectors to monetize location data (Valentino-Devries et al. 2018). Leveraging the behavior-rich location data for targeting is proven to be an effective mobile marketing strategy to increase advertisers' revenues (Ghose et al. 2018). However, these monetary gains come at the cost of potential invasion of consumer privacy. In this research, we tackle this important and under-studied topic from a data collector's perspective. We identify the key challenges faced by a data collector and propose an end-to-end framework to enable a data collector to leverage location data while preserving consumer privacy.

The existing literature on privacy preservation, primarily from the Computer Science discipline, are either unsuited for this new type of data with distinct challenges, or not interpretable or personalized to an individual level. Our research fills this gap. Specifically, we propose a framework of three components, each addressing a key topic facing a data collector. First, we quantify each consumer's risks, exemplified by two common types of stalker behaviours – sensitive attribute threat and re-identification threat. These risks are intuitively modeled as the stalker's success probabilities in inferring the consumer's private information. Second, we measure the utility of the location trajectory data to an advertiser by considering a popular business use case - POI recommendations. The utility is estimated by the accuracy of using the location data to infer a consumer's future locations. Finally, to enable a data collector to trade off between consumer risk and advertiser utility, we propose an obfuscation scheme suppressing consumers' trajectories based on their individual risks associated with each privacy threat and informativeness of each location in their trajectories. The proposed obfuscation scheme provides multiple options for the data collector to choose from based on specific business contexts.

We validate the proposed framework on a unique data set containing nearly a million mobile locations tracked from over 40,000 individuals over a period of five weeks in 2018. To our best knowledge, this research reflects an initial effort to analyze such a rich, granular, newly available human trajectory data; and for the purpose of privacy preservation. We find that there exists a high risk of invasion of privacy in the location data if a data collector does not obfuscate the data. On average, a stalker could accurately predict an individuals home address within a radius of 2.5 miles and mobile operating system with an 82% success. The proposed risk quantification enables a data collector to identify high risk individuals and those features contributing to the risk associated with each stalker threat. Furthermore, using the proposed obfuscation scheme, a data collector can achieve better trade-off between consumer privacy and advertiser utility when compared to several alternative rule-based and risk-based obfuscations. For instance, in the home address inference threat, we find that a risk to user privacy can be reduced by 15%, a maximum decrease when compared to rule-based heuristics, with less than 1% decrease in utility, a minimum decrease. Further, we compare our proposed framework to eight baselines and exemplify the performance gains in balancing the privacy-utility trade-off. In summary, this study presents conceptual, managerial, and methodological contributions to the literature and business practice, as summarized in the Introduction. Besides offering a powerful tool to data collectors to preserve consumer privacy while maintaining the usability of the increasingly accessible form of rich and highly valuable location data, this research also informs the ongoing debate of consumer privacy and data sharing regulations.

Despite the contributions, there are limitations of this research, thus calling for continued explorations of this rich and promising domain. For example, our data contain device IDs, but no detailed demographics, associated with each individual. When such data become available, one may, for instance, develop deeper insights into which demographic sub-populations are most vulnerable to privacy risks. Also, our analysis considered the locations' longitudes and latitudes, but not names (such as Starbucks) or types (such as hospital). Hence future research may further distinguish varied sensitivity levels across locations in privacy preservation. Furthermore, as other data, such as the same individuals' online clickstreams or social media comments, become linked to their mobile location data, more sophisticated privacy preservation methodologies may be developed.

References

Osman Abul, Francesco Bonchi, and Mirco Nanni. Never walk alone: Uncertainty for anonymity in moving objects databases. In *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*, pages 376–385. Ieee, 2008.

Alessandro Acquisti and Jens Grossklags. Privacy and rationality in individual decision making. *IEEE security & privacy*, 3(1):26–33, 2005.

Alessandro Acquisti, Leslie K John, and George Loewenstein. The impact of relative standards on the propensity to disclose. *Journal of Marketing Research*, 49(2):160–174, 2012.

Idris Adjerid, Eyal Peer, and Alessandro Acquisti. Beyond the privacy paradox: Objective versus relative risk in privacy decision making. *Available at SSRN 2765097*, 2016.

Idris Adjerid, Alessandro Acquisti, and George Loewenstein. Choice architecture, framing, and cascaded privacy choices. *Management Science*, 2018.

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. VLDB Endowment, 2005.

Gagan Aggarwal, Tomas Feder, Krishnaram Kenthapadi, Rajeev Motwani, Rina Panigrahy, Dilys Thomas, and An Zhu. Approximation algorithms for k-anonymity. *Journal of Privacy Technology (JOPT)*, 2005.

Michelle Andrews, Xueming Luo, Zheng Fang, and Anindya Ghose. Mobile ad effectiveness: Hyper-contextual targeting with crowdedness. *Marketing Science*, 35(2):218–233, 2016.

Apple. Apple Has Quietly Started Tracking iPhone Users Again, And Its Tricky To Opt Out, 2012. `http://www.businessinsider.com/ifa-apples-iphone-tracking-in-ios-6-2012-10`, 2012.

Apple. Requesting Permission. `https://developer.apple.com/design/human-interface-guidelines/ios/app-architecture/requesting-permission/`, 2014.

Apple. iOS 10 to Feature Stronger Limit Ad Tracking Control, 2016. `https://fpf.org/2016/08/02/ios-10-feature-stronger-limit-ad-tracking/`, 2016.

Daniel Ashbrook and Thad Starner. Using gps to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous computing*, 7(5):275–286, 2003.

Naveen Farag Awad and Mayuram S Krishnan. The personalization privacy paradox: an empirical evaluation of information transparency and the willingness to be profiled online for personalization. *MIS quarterly*, pages 13–28, 2006.

Yakov Bart, Venkatesh Shankar, Fareena Sultan, and Glen L Urban. Are the drivers and role of online trust the same for all web sites and consumers? a large-scale exploratory empirical study. *Journal of marketing*, 69(4):133–152, 2005.

Roberto J Bayardo and Rakesh Agrawal. Data privacy through optimal k-anonymization. In *null*, pages 217–228. IEEE, 2005.

Paul D Berger and Nada I Nasr. Customer lifetime value: Marketing models and applications. *Journal of interactive marketing*, 12(1):17–30, 1998.

Jesus Bobadilla, Antonio Hernando, Fernando Ortega, and Jesus Bernal. A framework for collaborative filtering recommender systems. *Expert Systems with Applications*, 38(12):14609–14623, 2011.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Gordon C Bruner and Anand Kumar. Attitude toward location-based advertising. *Journal of interactive advertising*, 7(2):3–15, 2007.

Gordon Burtch, Anindya Ghose, and Sunil Wattal. The hidden cost of accommodating crowdfunder privacy preferences: a randomized field experiment. *Management Science*, 61(5):949–962, 2015.

Ramon Casadesus-Masanell and Andres Hervas-Drane. Competing with privacy. *Management Science*, 61 (1):229–246, 2015.

Ramnath K Chellappa and Shivendu Shivendu. Mechanism design for free but no free disposal services: The economics of personalization under privacy concerns. *Management Science*, 56(10):1766–1780, 2010.

Ramnath K Chellappa and Raymond G Sin. Personalization versus privacy: An empirical examination of the online consumers dilemma. *Information technology and management*, 6(2-3):181–202, 2005.

Rui Chen, Gergely Acs, and Claude Castelluccia. Differentially private sequential data publication via variable-length n-grams. In *Proceedings of the 2012 ACM conference on Computer and communications security*, pages 638–649. ACM, 2012.

Rui Chen, Benjamin CM Fung, Noman Mohammed, Bipin C Desai, and Ke Wang. Privacy-preserving trajectory data publishing by local suppression. *Information Sciences*, 231:83–97, 2013.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM, 2016.

Chris Clifton and Tamir Tassa. On syntactic anonymity and differential privacy. In *2013 IEEE 29th International Conference on Data Engineering Workshops (ICDEW)*, pages 88–93. IEEE, 2013.

Vincent Conitzer, Curtis R Taylor, and Liad Wagman. Hide and seek: Costly consumer privacy in a market with repeat purchases. *Marketing Science*, 31(2):277–292, 2012.

Martijn G De Jong, Rik Pieters, and Jean-Paul Fox. Reducing social desirability bias through item randomized response: An application to measure underreported desires. *Journal of Marketing Research*, 47(1): 14–27, 2010.

Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications*, 21(4):1253–1278, 2000.

Flávio du Pin Calmon and Nadia Fawaz. Privacy against statistical inference. In *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*, pages 1401–1408. IEEE, 2012.

Jean-Pierre Dubé, Zheng Fang, Nathan Fong, and Xueming Luo. Competitive price targeting with smartphone coupons. *Marketing Science*, 36(6):944–975, 2017.

Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 371–380. ACM, 2009.

Nathan Eagle and Alex Sandy Pentland. Eigenbehaviors: Identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63(7):1057–1066, 2009.

eMarketer. US Smartphone OS Race Still Close, as Men, Younger Users Favor Android. `https://www.emarketer.com/article.aspx?R=1009961&RewroteTitle=1`, 2013.

Hadi Fanaee-T and João Gama. Eigenevent: An algorithm for event detection from complex data streams in syndromic surveillance. *Intelligent Data Analysis*, 19(3):597–616, 2015.

Zheng Fang, Bin Gu, Xueming Luo, and Yunjie Xu. Contemporaneous and delayed sales impact of location-based mobile promotions. *Information Systems Research*, 26(3):552–564, 2015.

Nathan M Fong, Zheng Fang, and Xueming Luo. Geo-conquesting: Competitive locational targeting of mobile promotions. *Journal of Marketing Research*, 52(5):726–735, 2015a.

Nathan M Fong, Zheng Fang, and Xueming Luo. Real-time mobile geo-conquesting promotions. *Journal of Marketing Research*, 2015b.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

Benjamin C. M. Fung, Ke Wang, Rui Chen, and Philip S. Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Comput. Surv.*, 42(4):14:1–14:53, June 2010. ISSN 0360-0300. doi: 10.1145/1749603.1749605. URL `http://doi.acm.org/10.1145/1749603.1749605`.

Pedro M Gardete and Yakov Bart. Tailored cheap talk: The effects of privacy policy on ad content and market outcomes. *Marketing Science*, 37(5):733–752, 2018.

Robert Garfinkel, Ram Gopal, and Paulo Goes. Privacy protection of binary confidential data against deterministic, stochastic, and insider threat. *Management Science*, 48(6):749–764, 2002.

Anindya Ghose. *TAP: Unlocking the mobile economy.* MIT Press, 2017.

Anindya Ghose, Beibei Li, and Siyuan Liu. Mobile targeting using customer trajectory patterns. *Management Science*, Forthcoming, 2018.

Khim-Yong Goh, Kai-Lung Hui, and Ivan PL Png. Privacy and marketing externalities: Evidence from do not call. *Management Science*, 61(12):2982–3000, 2015.

Avi Goldfarb and Catherine Tucker. Online display advertising: Targeting and obtrusiveness. *Marketing Science*, 30(3):389–404, 2011a.

Avi Goldfarb and Catherine Tucker. Rejoinderimplications of online display advertising: Targeting and obtrusiveness. *Marketing Science*, 30(3):413–415, 2011b.

Avi Goldfarb and Catherine E Tucker. Privacy regulation and online advertising. *Management science*, 57 (1):57–71, 2011c.

Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *nature*, 453(7196):779, 2008.

Il-Horn Hann, Kai-Lung Hui, Sang-Yong T Lee, and Ivan PL Png. Consumer privacy and marketing avoidance: A static model. *Management Science*, 54(6):1094–1103, 2008.

Trevor Hastie, Saharon Rosset, Ji Zhu, and Hui Zou. Multi-class adaboost. *Statistics and its Interface*, 2(3): 349–360, 2009.

Xi He, Graham Cormode, Ashwin Machanavajjhala, Cecilia M Procopiuc, and Divesh Srivastava. Dpt: differentially private trajectory synthesis using hierarchical reference systems. *Proceedings of the VLDB Endowment*, 8(11):1154–1165, 2015.

Donna L Hoffman, Thomas P Novak, and Marcos Peralta. Building consumer trust online. *Communications of the ACM*, 42(4):80–85, 1999.

Torsten Hothorn, Kurt Hornik, and Achim Zeileis. ctree: Conditional inference trees. *The Comprehensive R Archive Network*, pages 1–34, 2015.

Joakim Kalvenes and Amit Basu. Design of robust business-to-business electronic marketplaces with guaranteed privacy. *Management Science*, 52(11):1721–1736, 2006.

Kelsey. US Local Mobile Local Social Ad Forecast. `https://shop.biakelsey.com/product/2018-u-s-local-mobile-local-social-ad-forecast/`, 2018.

V Kumar, Xi Zhang, and Anita Luo. Modeling customer opt-in and opt-out in a permission-based marketing context. *Journal of Marketing Research*, 51(4):403–419, 2014.

Chen Li, Houtan Shirani-Mehr, and Xiaochun Yang. Protecting individual information against inference attacks in data publishing. In *International Conference on Database Systems for Advanced Applications*, pages 422–433. Springer, 2007.

Chenxi Li, Xueming Luo, Cheng Zhang, and Xiaoyi Wang. Sunny, rainy, and cloudy with a chance of mobile promotion effectiveness. *Marketing Science*, 36(5):762–779, 2017.

Tiancheng Li, Ninghui Li, Jian Zhang, and Ian Molloy. Slicing: A new approach for privacy preserving data publishing. *IEEE transactions on knowledge and data engineering*, 24(3):561–574, 2012.

Xiao-Bai Li and Sumit Sarkar. Against classification attacks: A decision tree pruning approach to privacy protection in data mining. *Operations Research*, 57(6):1496–1509, 2009.

Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

Xueming Luo, Michelle Andrews, Zheng Fang, and Chee Wei Phang. Mobile targeting. *Management Science*, 60(7):1738–1756, 2014.

Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, and Muthuramakrishnan Venkitasubramaniam. ℓ-diversity: Privacy beyond κ-anonymity. In *null*, page 24. IEEE, 2006.

Ashwin Machanavajjhala, Johannes Gehrke, and Michaela Götz. Data publishing against realistic adversaries. *Proceedings of the VLDB Endowment*, 2(1):790–801, 2009.

Kelly D Martin, Abhishek Borah, and Robert W Palmatier. Data privacy: Effects on customer and firm performance. *Journal of Marketing*, 81(1):36–58, 2017.

Charles S Mayer and Charles H White Jr. The law of privacy and marketing research. *Journal of Marketing*, 33(2):1–4, 1969.

Adam Meyerson and Ryan Williams. On the complexity of optimal k-anonymity. In *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 223–228. ACM, 2004.

Amalia R Miller and Catherine Tucker. Privacy protection and technology diffusion: The case of electronic medical records. *Management Science*, 55(7):1077–1093, 2009.

Amalia R Miller and Catherine Tucker. Privacy protection, personalized medicine, and genetic testing. *Management Science*, 64(10):4648–4668, 2017.

Dominik Molitor, Philipp Reichhart, Martin Spann, and Anindya Ghose. Measuring the effectiveness of location-based advertising: A randomized field experiment. *Available at SSRN 2645281*, 2019.

Krishnamurty Muralidhar and Rathindra Sarathy. Data shufflinga new masking approach for numerical data. *Management Science*, 52(5):658–670, 2006.

Luca Pappalardo, Salvatore Rinzivillo, and Filippo Simini. Human mobility modelling: exploration and preferential return meet the gravity model. *Procedia Computer Science*, 83:934–939, 2016.

Roberto Pellungrini, Luca Pappalardo, Francesca Pratesi, and Anna Monreale. A data mining approach to assess privacy risk in human mobility data. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 9(3):31, 2018.

Pew. More Americans using smartphones for getting directions, streaming TV. `http://www.pewresearch.org/fact-tank/2016/01/29/us-smartphone-use/`, 2016.

Pew. Americans complicated feelings about social media in an era of privacy concerns. `https://www.pewresearch.org/fact-tank/2018/03/27/americans-complicated-feelings-about-social-media-in-an-era-of-privacy-concerns/`, 2018.

Yi Qian and Hui Xie. Drive more effective data-based innovations: enhancing the utility of secure databases. *Management Science*, 61(3):520–541, 2015.

Omid Rafieian and Hema Yoganarasimhan. Targeting and privacy in mobile advertising. 2018.

General Data Protection Regulation Regulation. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46. *Official Journal of the European Union (OJ)*, 59(1-88):294, 2016.

Pierangela Samarati. Protecting respondents identities in microdata release. *IEEE transactions on Knowledge and Data Engineering*, 13(6):1010–1027, 2001.

Pierangela Samarati and Latanya Sweeney. Generalizing data to provide anonymity when disclosing information. In *PODS*, volume 98, page 188. Citeseer, 1998.

Burhaneddin Sandıkçı, Lisa M Maillart, Andrew J Schaefer, and Mark S Roberts. Alleviating the patient's price of privacy through a partially observable waiting list. *Management Science*, 59(8):1836–1854, 2013.

Matthew J Schneider, Sharan Jagpal, Sachin Gupta, Shaobo Li, and Yan Yu. A flexible method for protecting marketing data: An application to point-of-sale data. *Marketing Science*, 37(1):153–171, 2018.

Miremad Soleymanian, Charles B Weinberg, and Ting Zhu. Sensor data and behavioral tracking: Does usage-based auto insurance benefit drivers? *Marketing Science*, 2019.

Statista. Share of smartphone users that use an Apple iPhone in the United States from 2014 to 2019. `https://www.statista.com/statistics/236550/percentage-of-us-population-that-own-a-iphone-smartphone/`, 2018.

Kyle Taylor and Laura Silver. Smartphone Ownership Is Growing Rapidly Around the World, but Not Always Equally. `http://www.pewglobal.org/2019/02/05/smartphone-ownership-is-growing-rapidly-around-the-world-but-not-always-equally/`, 2019.

Manolis Terrovitis, Nikos Mamoulis, and Panos Kalnis. Privacy-preserving anonymization of set-valued data. *Proceedings of the VLDB Endowment*, 1(1):115–125, 2008.

Manolis Terrovitis, Giorgos Poulis, Nikos Mamoulis, and Spiros Skiadopoulos. Local suppression and splitting techniques for privacy preserving publication of trajectories. *IEEE Trans. Knowl. Data Eng*, 29(7): 1466–1479, 2017.

Catherine E Tucker. Social networks, personalized advertising, and privacy controls. *Journal of Marketing Research*, 50(5):546–562, 2013.

Jennifer Valentino-Devries, Natasha Singer, Michael H. Keller, and Aaron Krolik. Your Apps Know Where You Were Last Night, and Theyre Not Keeping It Secret. `https://www.nytimes.com/interactive/2018/12/10/business/location-data-privacy-apps.html?smid=re-nytimes`, 2018.

Verge. Android Q leak reveals system-wide dark mode and bigger emphasis on privacy . `https://www.theverge.com/2019/1/16/18185763/android-q-leak-dark-mode-new-privacy-settings`, 2019.

Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti, and Albert-Laszlo Barabasi. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1100–1108. Acm, 2011.

Ke Wang, Benjamin CM Fung, and S Yu Philip. Handicapping attacker's confidence: an alternative to k-anonymization. *Knowledge and Information Systems*, 11(3):345–368, 2007.

Michel Wedel and PK Kannan. Marketing analytics for data-rich environments. *Journal of Marketing*, 80 (6):97–121, 2016.

Nathalie E Williams, Timothy A Thomas, Matthew Dunbar, Nathan Eagle, and Adrian Dobra. Measures of human mobility using mobile phone records enhanced with gis data. *PloS one*, 10(7):e0133630, 2015.

David Jingjun Xu. The influence of personalization in affecting consumer attitudes toward mobile advertising in china. *Journal of Computer Information Systems*, 47(2):9–19, 2006.

Heng Xu, Xin Robert Luo, John M Carroll, and Mary Beth Rosson. The personalization privacy paradox: An exploratory study of decision making process for location-aware marketing. *Decision support systems*, 51(1):42–52, 2011.

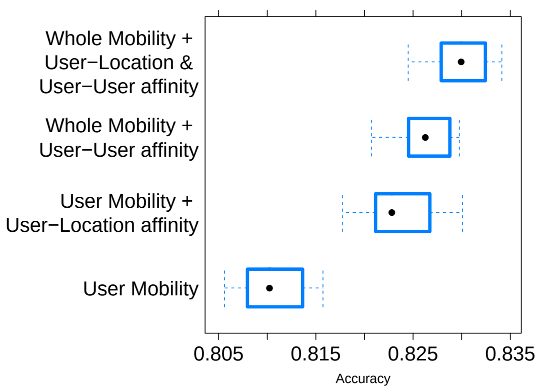
Dingqi Yang, Bingqing Qu, and Philippe Cudre-Mauroux. Privacy-preserving social media data publishing for personalized ranking-based recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 2018.

Roman Yarovoy, Francesco Bonchi, Laks VS Lakshmanan, and Wendy Hui Wang. Anonymizing moving objects: How to hide a mob in a crowd? In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, pages 72–83. ACM, 2009.

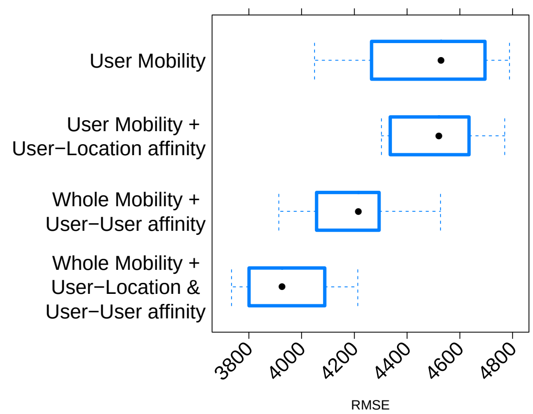
Yu Zheng, Xing Xie, and Wei-Ying Ma. Geolife: A collaborative social networking service among user, location and trajectory. *IEEE Data Eng. Bull.*, 33(2):32–39, 2010.

Appendix A: Model choices in proposed framework

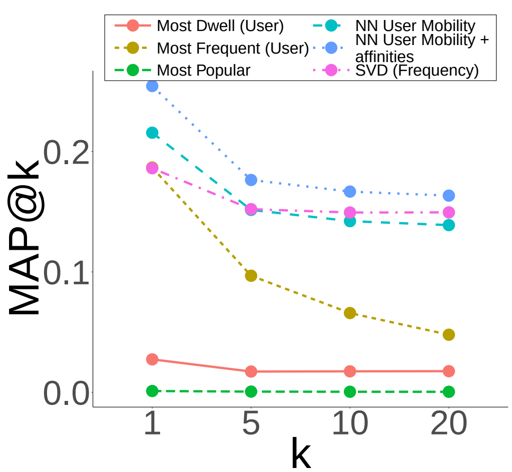
We empirically justify the model choices made in our methodology. All the choices were made based by assessing the performance of different machine learning heuristics used in our framework on non-obfuscated data. First, in Figures 6a, 6b we show the incremental benefit of the affinity features discussed in the feature extraction $\mathcal{F}(T)$. Figure 6a shows the accuracy of the Random Forest classifier to predict the operating system of a user. The model was regularized by performing a grid search on the maximum number of features and trees [13] via five-fold cross-validation. The best performing model has an accuracy of 82% which indicates the success a stalker would have in inferring the unpublished operating system of a user from trajectory data. In Figure 6b, we plot the RMSE of the Random Forest regressor trained to predict the home address of a user.[14]



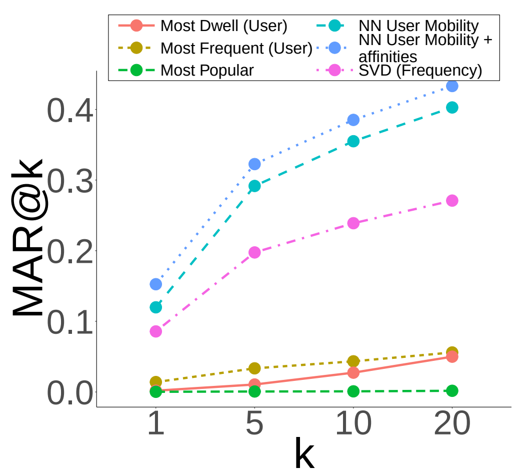
(a) OS inference threat - Model Accuracy



(b) Home address inference - Model RMSE



(c) $MAP@k$ POI Recommendation model



(d) $MAR@k$ POI Recommendation model

Figure 6 Proposed framework model choices

[13] Grid for fraction of features - $\{.25, .5, .75, 1\}$, trees - $\{50, 100, 200\}$

[14] We infer the ground truth of home location in our data by assigning this to be the most frequently visited location during 10 PM - 6 AM for each user. We have already tested alternative time periods such as 11 PM - 5 AM, and the results remain robust. We also delete the inferred home location from all user trajectories for our experiments.

Next, we learn two regression models to predict the Universal Transverse Mercator (UTM) transformed latitude and longitude of the home location with similar hyperparameter tuning as earlier. The error estimate is the Euclidean distance between the estimated and assigned home UTM coordinates. From the box plots of the re-sampled performance measures (Figures 6a, 6b), we notice that the User-User and User-Location affinity features incrementally improve the performance of both the proxy models learned. In Figures 6c, 6d, we visualize the $MAP@k$ and $MAR@k$ of the neighborhood-based recommendation model learned by tuning the number of neighbors.

We compare the performance with several baselines - recommendations based on the most popular locations (Most Popular), based on the locations that the user spent the most time in (Most Dwell (User)), visited most frequently (Most Frequent (User)) and an SVD on the user-location matrix populated with frequency. We observe that the NN based model performs better in both the metrics compared to the baselines justifying the choice. The RMSE, 3,900 meters $\approx$ 2.46 miles indicates the success a stalker would have in identifying the home location of a user from non-obufuscated data. Further, we also notice the incremental benefit (See NN User Mobility vs NN User Mobility + affinities in Figures 6c, 6d) of the affinity features in the recommendation performance.

Appendix B: Utility measurement

We compute the data utility under different obfuscations. We estimate this by computing the performance of a neighborhood-based collaborative filtering recommendation heuristic to accurately predict future user locations. To assess the accuracy of predictions made, we treat the locations visited by each user in the fifth week as ground truth and train the recommender model to predict these locations.

Based on the user risks, we obfuscate T_{train} by varying $p \in \mathcal{G}_p$. We learn a neighborhood-based recommender (Bobadilla et al. 2011) tuning number of neighbors by five-fold cross-validation on the obfuscated training sample $\mathcal{P}(T_{train})$. The model is learned to rank locations a user is likely to visit in the fifth week of the observation period. That is, we build features, $\mathcal{F}(P(T_{train}))$ on first four weeks and tune number of neighbors[15] to maximize prediction accuracy. Then, we compute utility —$MAP@k$, $MAR@k$ on T_{test} for $k = \{1, 5, 10\}$[16]. Intuitively, $MAP@1$ and $MAR@1$, for example, represent advertiser's utility to predict next location a user is most likely to visit in the fifth week based of the recommender model that was learned on the obfuscated data. A key detail in the utility estimation is that we do not perform any obfuscation on T_{test} for any value of p since our aim is to quantify the ability of obfuscated data, $\mathcal{P}(T_{train})$ to learn true preferences of a user, which are revealed in the non-obfuscated test sample. Similar to risk, we perform twenty trials for each p and report mean and 95% confidence intervals of utility metrics in Figure 3.

[15] Grid for number of neighbors - $\{5, 10, 25, 50, 100, 200\}$

[16] The learned recommender model can be used to compute $MAP@k$, $MAR@k$ for other values of k as well. We consider $k = \{1, 5, 10\}$ for illustration of the method's efficacy.

Appendix C: Comparison to LSup and GSup

Continuing our comparison to different types of baselines from Section 5.4, here, we compare the proposed framework to the most recent syntactic models LSup and GSup proposed by (Terrovitis et al. 2017). Both the models obfuscate the location data to reduce the re-identification threat by maintaining utility. Methodologically, these models differ from the proposed framework (Section 4.3) in two ways. First, in both LSup and GSup, the consumer risk is only quantified for one threat (re-identification) whereas our framework additionally considers sensitive attribute inference. Second, the suppression is either performed globally, that is a location is suppressed across all the users(GSup) or locally (LSup), location suppressed for a subset of the users. In our suppression scheme, thanks to the introduction of the two consumer specific parameters $\{\vec{s_i}, z_i\}$, suppression occurs at a consumer level with varying suppression probabilities assigned to each location a user has visited. In addition, compared to the parsimonious inputs that our proposed framework requires, both the models in consideration require multiple input parameters P_{br}, number of adversaries $\mathcal{A}$ and background knowledge of each adversary in $\mathcal{A}$. P_{br} controls the number of locations suppressed either locally (LSup) or globally (GSup). Higher the value of P_{br}, lower the number of location suppressed. In our comparison, we follow the empirical evaluation framework of the authors to set the number of adversaries $\mathcal{A}$ and background knowledge of each adversary in $\mathcal{A}$ and vary P_{br}.

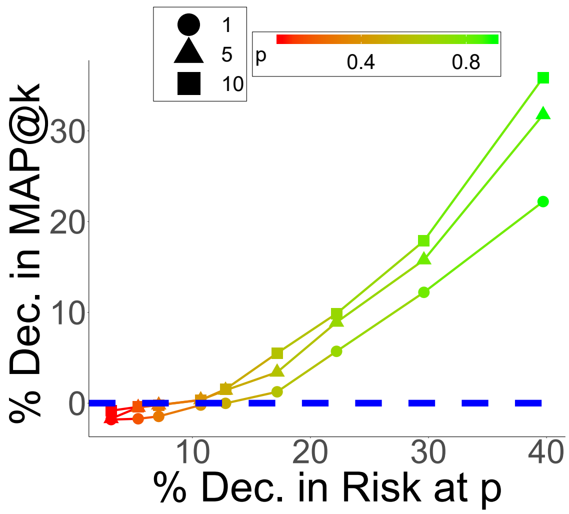
Obfuscation Method	% Decrease Home address risk	% Decrease Operating system risk	% Decrease Re-identification risk	% Decrease Utility (MAP@1)	% Decrease Utility (MAR@1)
GSup ($P_{br} = 0.2$)	18.12	9.26	14.52	7.74	8.31
GSup ($P_{br} = 0.5$)	7.25	3.11	7.29	4.49	3.42
LSup ($P_{br} = 0.2$)	22.16	14.56	31.56	5.31	7.12
LSup ($P_{br} = 0.5$)	9.15	4.01	10.91	-1.65	0.86

Table 4 LSup **and** GSup **comparison. (Green/Red indicate proposed framework provides a better/worse trade-off)**

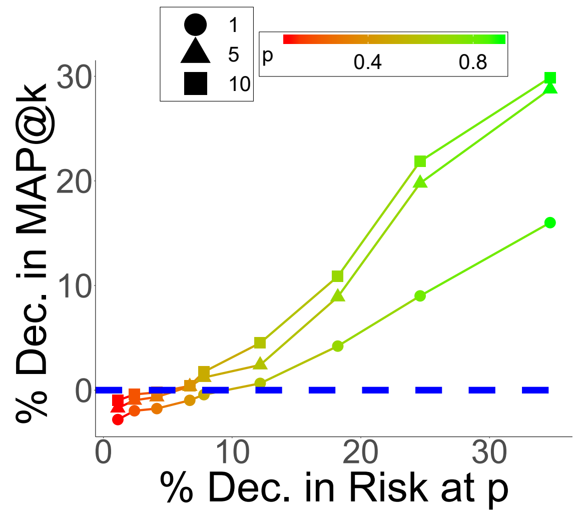
In Table 4, we present the the decrease in consumer risk from the non-obfuscated trajectories for the two types of privacy threats - re-identification and sensitive attribute inference[17] (operating system and home address inference) and the corresponding measures of advertiser's utility as $MAP@1$, $MAR@1$. To identify the obfuscation scheme that provides the better/worse trade-off, we compute the slope ($\frac{Y}{X}$ in Figure 3 —% Decrease in utility divided by % Decrease in risk) for different decreases in utility ($MAP@1$) of LSup and GSup. We observe that in a majority (10 out of 12) of the cases, the proposed framework provides a better trade-off (denoted by green color in Table 4) compared to both LSup and GSup. This improved trade-off

[17] Since the considered models do not handle sensitive attribute inference, we obfuscated the data to reduce re-identification threat and use the same obfuscated data to quantify the reduce in consumer risk for the two types of attacks.

come with an added benefit that the proposed framework only requires one input parameter corresponding to the number of locations of a stalker in the re-identification threat compared to the various parameters required for LSup and GSup.



(a) Suppression based on time spent



(b) Suppression based on recency

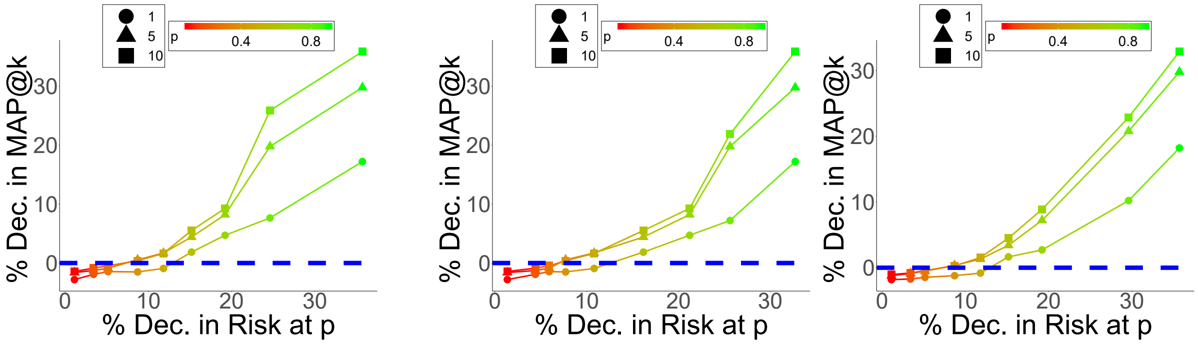
Figure 7 **Proposed framework : Home address inference, suppression by recency and time spent.**

Appendix D: Suppression based on recency and time spent

In our suppression scheme detailed in Section 4.3, we introduce and provide a structured grid search by varying the grid parameter p to identify two consumer specific parameters $\{\vec{s_i}, z_i\}$, where z_i captures the number of locations to be suppressed for a given consumer trajectory T_i and within T_i, we assign weights to each tracked location through $\vec{s_i}$ to denote the likelihood of a specific location being suppressed. In our empirical study detailed in Section 5, in Figure 3, we assign $\vec{s_i}$ based on the frequency of the location visited in T_i. Here, we augment the empirical study and showcase the flexibility of the proposed suppression scheme by assigning the $\vec{s_i}$ based on time spent by a consumer at each location in T_i and the recency of the locations in T_i. For brevity, we only consider the sensitive attribute threat where a stalker aims to infer the home address of a consumer and visualize the privacy-utility trade-off in figures 7b,7a. Similar to Figure 3, we observe that for a given percentage decrease in risk, there is a lesser corresponding percentage decrease in performance in both the figures.

Appendix E: Varying sample sizes

To test for the robustness of the results discussed in Figure 3, we repeat our empirical exercise on three random samples - 25%, 50% and 75% of the full 40,000 consumer trajectory data. For brevity and to avoid repetition of similar plots, the suppression is performed based on the frequency of location visited by a consumer (similar to Figure 3) for the home address inference threat. The resulting plots comparing the percentage decreases in consumer risk and advertiser's utility from the baselines (non-obfuscated data) are visualized in Figures 8a,8b,8c. We note that even at smaller samples, the slope (% Decrease in utility divided by % Decrease in risk) at different values of p is similar to the full sample (Figure 3a).



(a) 25% of data sample (b) 50% of data sample (c) 75% of data sample

Figure 8 Proposed framework : Home address inference, varying sample sizes