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, nfoutz@virginia.edu

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

(Working Paper February 4, 2020)

Consumer location tracking is becoming omnipresent on mobile devices, producing a vast volume 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 any location data collector, such as a mobile app owner or data aggregator, to balance consumer privacy risk and advertiser utility, when sharing consumer location data with any advertiser. This will also 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 a data collector to interpret the trade-off between the data's privacy risk to consumers and utility to an advertiser. To address this research gap, we propose a personalized and interpretable framework that enables a location data collector to optimize the risk-utility 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, an individual's home address can be accurately predicted within an average 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 each consumer's privacy risk (e.g., by 15% of inferring home address) with minimal (i.e., $< 1\%$) decrease in an advertiser's utility. As novel and powerful consumer location data become increasingly available, we demonstrate their utility to an advertiser and accompanying privacy risk to consumers, and most importantly, propose a personalized and interpretable framework to mitigate their risk while maximizing their utility.

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 consumers, 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, 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 GPS coordinates). Such individual-level 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 and portray rich contexts of a consumer's behavior and brand preference, broad lifestyle, socioeconomic status, and social relationship (Ghose et al. 2018). Such offline data become even more powerful if combined with a consumer's online footprints, such as click stream data or social media data, rendering a holistic online-offline consumer profile. Fourth, excellent location tracking and targeting across apps simplifies ad attribution of a location-based ad campaign. Each advertiser has access to a unique device ID associated with each smartphone, thus benefiting from reduced overhead to stitch together a consumer's location data

[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).

across sessions or apps and enjoying a holistic view of each consumer when measuring a campaign's effectiveness (Apple 2012). Fifth, geo-marketing by a butler advertiser 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 utility to an advertiser whose butler actions further benefit consumers, but also monetization opportunities to a location data collector who shares the data with the advertiser. 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 apps. 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 movement data offers important utility to an advertiser, and thus benefits to consumers and the data collector as well, they entail major privacy risks, such as home location inference. "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 a consumer's home address, daily trajectories, and broad lifestyle, as vividly depicted by two recent New York Times' articles (Valentino-Devries et al. 2018, Thompson and Warzel 2019). 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 discussion so far calls for any data collector, before sharing location data with an advertiser, to maintain a crucial trade-off between the utility to the advertiser and privacy risk to a consumer. 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

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

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 the risk-utility trade-off (or privacy-utility trade-off hereafter). We thus aim to develop a personalized, privacy-preserving framework that incorporates consumer heterogeneity and optimizes a data collector's risk-utility trade-off.

1.2. Research Agenda and Challenges

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

1. *Consumer:* is an individual who 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. *Advertiser:* is a firm that acquires data from a data collector to improve the targetability of its marketing campaigns. A subset of advertisers, or even a third party, with access to the location data, might have a stalker intent (stalker 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 and preserving a butler advertiser's utility such as POI recommendation (Muralidhar and Sarathy 2006). We aim to answer the following essential questions.

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

2. *Advertiser's utility:* What is the value of an obfuscated data set to a butler advertiser's utility? Specifically, what types of key behavioral information can a butler advertiser extract from the data to service or target consumers in a mutually beneficial way?

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

To accomplish the above, several methodological challenges need to be overcome. From a methodological standpoint, our research questions 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 (few overlaps of locations across consumers), 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 privacy 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 any specific business use case. These approaches, devised mostly from the Computer Science literature, measure an advertiser's utility with 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 an advertiser 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 each consumer's privacy risk, quantification of an advertiser's utility, and obfuscation scheme for a data collector.

Quantification of Consumer's Privacy Risk. 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 the other is "re-identification threat", where all locations visited by a consumer are being identified based on a subset of the locations (Samarati 2001, Pellungrini et al. 2018).

Quantification of Advertiser's Utility. While the utility of a mobile location data set to an advertiser is multi-faceted, we demonstrate one specific utility related to one arguably most popular and essential business goal examined by the literature – Point-of-Interest (POI hereafter) recommendation in mobile advertising (Ghose et al. 2018). Reliable predictions of a consumer's future locations would enable an advertiser to target the consumer with context relevant contents and lead to higher business revenues (Ghose et al. 2018). 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 similarity-based collaborative filtering recommendation model trained on the location data. The central idea of this recommender is to identify other consumers with similar historical behaviors in order to infer the focal consumer's future behavior (Bobadilla et al. 2011). Note that while focusing on POI recommendation, our framework can easily accommodate other utility measurements with various business goals as well.

Obfuscation Scheme for Data Collector. Acknowledging many potential solutions to the privacy-utility trade-off 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 identities of the locations suppressed for each consumer. The suppression, while reducing each consumer's privacy risk, also adversely impacts an advertiser's utility. Hence, we empirically identify the parameters that balance the privacy-utility trade-off through a structured grid search while leveraging the 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 baseline risk and the advertiser's baseline utility from the original, unobfuscated mobile location data (i.e., the full sample). These would also represent the counterfactual case when no privacy protection is performed. We expect the unobfuscated full sample to yield the maximum utility to the advertiser, yet incur the maximum privacy risk to the consumers. In Part B, we perform consumer-level obfuscation based on the suppression probabilities of each location computed from each consumer's baseline risk, measures of the informativeness of each location, and a grid parameter. We then calculate the mean risk and

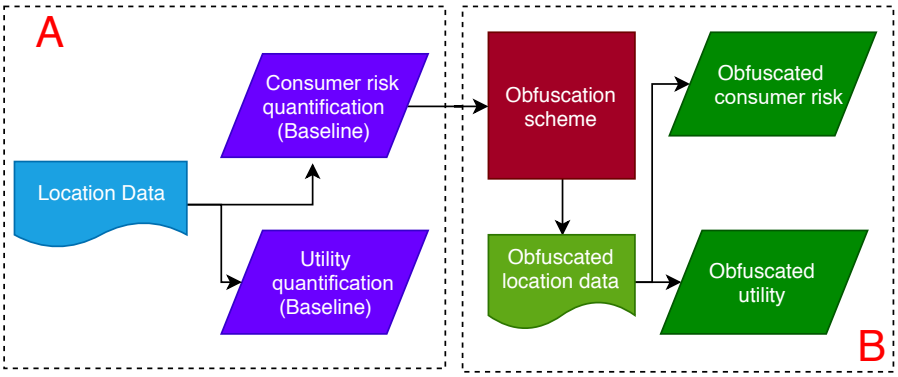


Figure 1 Overview of the proposed framework

utility across all consumers from the obfuscated data. Finally, we repeatedly obfuscate the original data by varying the grid parameter and recalculate the mean risk and utility on the corresponding obfuscated data to empirically determine the best risk-utility trade-off for the data collector. We will describe the details in the Methodology section.

As alluded to earlier, while we 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 location sequence or visit frequency inference, for which the risk may be quantified either analytically or via machine learning heuristics (Pellungrini et al. 2018). The framework may accommodate other types of advertiser use cases as well, for which the utility may be computed as the predictive accuracy of the specific business application of interest, such as when a consumer is most likely to convert into a paying customer given prior trajectories, 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 is 84% for inferring a consumer's home address and 82% for inferring mobile operating system[3]. On average, a consumer's home address can be predicted within a radius of 2.5

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

miles. Moreover, a consumer's entire location trajectories can be fully identified with a 49% success by knowing merely two randomly sampled locations visited by the consumer. It is noteworthy that these success probabilities of various privacy threats are all estimated based on machine learning heuristics, which require only the consumers' locations and corresponding timestamps as the inputs, as we will describe later. Hence, any entities, including advertisers, who have access to the location data could accomplish the same inferences.

Second, we find great value of the mobile location data to an advertiser. An advertiser aiming to target a consumers would be able to predict the next location most likely visited by the consumer with 25% success. This means that by analyzing the behavioral patterns revealed by the historical trajectories, for every one out of four customers, the advertiser is able to design a highly precise geo-targeting strategy.

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 options of risk-utility trade-off via a grid parameter to perform the obfuscation. Moreover, we find that the proposed framework presents a better choice set of risk-utility trade-off 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. These data exhibit distinctive properties, such as high dimensionality (resulting from massive numbers of locations), sparsity (with few overlaps across visited locations), and sequentiality (with temporal ordering of visited locations), hence imposing unique methodological challenges.

Conceptually, this research demonstrates the importance for any location data collector to preserve both consumer privacy and advertiser utility 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 (e.g., 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 a location data collector: quantification of each consumer's privacy risk, quantification of an advertiser's utility (i.e., value of mobile location data to an advertiser), and design of an intuitive and interpretable obfuscation scheme for a data collector. The framework requires only a single parsimonious input yet offers a data collector multiple, interpretable, and personalized options to protect consumer privacy while preserving an advertiser's utility, hence the data collector's overall monetization opportunity.

Methodologically, this framework (1) quantifies the privacy risk at a consumer level, instead of an aggregate or location level; and quantifies each consumer's privacy risk 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 an advertiser's utility associated with specific, real-world business use cases, such as POI recommendation shown to improve retailers' incremental revenues (Ghose et al. (2018)); (3) proposes an interpretable obfuscation scheme that requires merely one input from the data collector and suppresses locations at each consumer level to furnish the 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 with eight benchmarks.

Striking 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 allowed consumers 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 risk 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. Section 4 describes the details of the proposed framework (Figure 1). In Section 5, we discuss the empirical results and advantages of the proposed framework. We offer the 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 to 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 consumers' private information, the latest research has also investigated online, social media, and mobile advertising (Goldfarb and Tucker 2011a, Conitzer et al. 2012, Tucker 2013, Gardete and Bart 2018, Goldfarb and Tucker 2011c, Rafieian and Yoganarasimhan 2018, Goldfarb and Tucker 2011b). 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 volumes 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 main streams: consumer-, firm-, regulation-, and methodology- focused. We will concisely review each. The first stream takes on a consumers' perspective, and as a result, derives implications for firms to design privacy-friendly policies. For instance, a number of studies examine 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 website contents but with obtrusive format (Goldfarb and Tucker 2011c,b), or opt-in/out options of email marketing programs (Kumar et al. 2014). Other studies explore how normative and heuristic decision processes influence consumers' privacy decision making (Adjerid et al. 2016). Overall, these studies point to positive effects 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 stream of research also reveals 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 assumes a firms' perspectives, often using a game-theoretic approach to reach normative implications of firms' privacy policies. For instance, Chellappa and Shivendu (2010) derive an optimal design of personalization services for customers with heterogeneous privacy concerns. Gardete and Bart (2018) propose an optimal choice of ad content and communication when the firm withholds the customers' private information. Conitzer et al. (2012) reveal a monopoly's 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 firm's 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, these regulations are shown to impact firms' privacy-pertinent practices, technology innovations (Adjerid et al. 2016) and adoptions (Miller and Tucker 2009, 2017), and consumers' responses to e.g. the do-no-call registry (Goh et al. 2015). European Union (EU)'s privacy policy is shown to reduce the effectiveness of online display ads (Goldfarb and Tucker 2011a). Different components of a privacy law may also incur different effects, for instance, granting consumers controls over re-disclosure encourages genetic testing, whereas privacy notification deters it (Miller and Tucker 2017).

The fourth stream of research develops methodologies for regulatory bodies and firms 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 identifiable information (Tucker 2013), allowing customers to remain anonymous with a cost (Conitzer et al. 2012), or providing only 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 solicitations (Hann et al. 2008). Globally, governments are also providing regulatory protections, such as national do-no-call registries (Goh et al. 2015) and state genetic privacy laws (Miller and Tucker 2017). Other methodologies, on the other hand, leverage obfuscation of original data or query outputs. The premise is that an entity, data collector in our setting, would perform data obfuscation to preserve consumer privacy before releasing the data to a third party, an advertiser for instance, while ensuring that the data remain usable. Since such research is most related to our work, we will provide a more thorough survey of two sub-streams of this research based on the assumptions made when developing the relevant techniques (Clifton and Tassa 2013).

2.2. Privacy-preserving Methodology I: Syntactic Models

The assumption of syntactic models is that the entity performing the obfuscation knows the type of threat that a stalker or malicious entity intends 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 that 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 columnar values.

Studies have shown that k-anonymity is NP hard and suffers from the curse of dimensionality (Meyerson and Williams 2004). Variations of the concept of k-anonymity and heuristics to approximate k-anonymity have then been proposed (Aggarwal et al. 2005). Since k-anonymity primarily focuses on the re-identification threat, the method is susceptible to sensitive attribute inference when a stalker aims to only infer a particular column of a consumer rather than completely re-identify all the columnar values. ℓ-diversity (Machanavajjhala et al. 2006) and confidence bounding (Wang et al. 2007) are proposed to address these shortcomings. ℓ-diversity accomplishes this by obfuscating data such that sensitive attributes are well represented for each consumer, 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 are 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 developed 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 to obfuscate location data.

Extensions of the above traditional heuristics have been proposed 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. A variation of k-anonymity is further developed for moving object databases (MOD) based on the assumption that MODs do not have a fixed set of quasi-identifiers (QIDs) (Yarovoy et al. 2009). The authors define the timestamps of the locations as QIDs and propose two obfuscation techniques based on space generalization. These two studies aim at protecting consumers from re-identification threats.

More recently, suppression techniques have garnered attention in obfuscating location data (Chen et al. 2013, Terrovitis et al. 2008, 2017). For example, the seminal work by Terrovitis et al. (2008) presents a local suppression obfuscation technique assuming that a stalker has access to partial consumer trajectories, similar to the setting of the re-identification threat in our study. Built on this work, Terrovitis et al. (2017) further propose global suppression, separately from local suppression. Providing privacy guarantees against both identity and attribute linkage threats, Chen et al. (2013) develop $(K, C)_L$ privacy framework. The model requires three parameters from a data collector: a stalker's success probability thresholds in both types of threats and a parameter corresponding to a stalker's background knowledge. Instead of measuring the data utility with a rudimentary metric, the number of unique location points or frequent sequences preserved in the obfuscated data, as in Chen et al. (2013) and Terrovitis et al. (2008, 2017), our research captures the data utility by tying it to a popular business objective of an advertiser – POI recommendation.

2.3. Privacy-preserving Methodology II: Differentially Private Algorithms

This sub-stream of research is based on the concept of ϵ-differential privacy (Dwork and Lei 2009). Differentially private algorithms guarantee that a stalker would make the same inference from the shared data whether or not a focal individual is included in the data. Unlike syntactic models, they are not limited to a specific type of threats, thus presenting a much stronger privacy notion. The obfuscation performed on the data usually involves perturbation, 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 rows or columns, such as replacing a subset of a consumer's record with another consumer's record to minimize privacy risks. Various studies have leveraged perturbation, data shuffling, or a combination of them (Qian and Xie 2015). For instance, Garfinkel et al. (2002) perturb the answer of a database query to generate 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. Besides perturbation and data shuffling, public key encryption, digital certificate, and blinded signatures are also common privacy-friendly tools (Kalvenes and Basu 2006). All of the above methods focus on columnar data.

In the context of location data, the literature is sparse. A 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 involves summary statistics, such as the number of unique locations or frequent location patterns. Moreover,

owing to the stronger theoretical guarantees to be met, these techniques have been empirically shown to not preserve the truthfulness of the location data, hence hindering advertisers' abilities 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 collector and advertisers in exchange for personalized offers. So we take the route of syntactic models that are more likely to result in a higher data utility to an advertiser. We assume that a data collector has reasonable knowledge about the type of privacy threats that a consumer could be exposed to. To minimize the privacy threats, we propose an obfuscation scheme based on suppression that also ensures sufficient utility of the obfuscated location data to an advertiser.

As alluded to earlier, our study distinguishes itself from the prior research along several dimensions. Methodologically, 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). We also measure the utility of the location data in the context of real-world business applications, such as POI recommendation, instead of using the aggregate or rudimentary metrics from the literature, such as the number of unique locations or frequent sequences (He et al. 2015). From the standpoint of practical applicability, the proposed framework requires merely one parsimonious input, the number of locations already known to a stalker (Section 4.1.3). Therefore, it is intuitive and interpretable to the data collector or any manager. We also provide a data collector with multiple options of the risk-utility trade-off. Finally, most prior studies have validated their recommendations only 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) with various data limitations described earlier, such as accuracy or representativeness. In contrast, we validate our proposed framework on granular mobile location data from a large population over time.

2.4. Location-based Mobile Marketing

Finally, our work is related to the research on location-based mobile marketing. Using randomized field experiments, researchers have demonstrated that mobile advertisements based on the 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 the utility of the location data to an advertiser by considering a popular business application, POI recommendation. Identifying the next location most likely visited by a consumer based on his or her prior trajectories is crucial to perform behavioral targeting. Ghose et al. (2018) design a POI-based mobile recommendation based on similarities of consumers' mobile trajectories and demonstrate that such a strategy can lead to a significant improvement in a

retailer's incremental revenues. Other recent studies have revealed 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 choices (Fong et al. 2015a, Dubé et al. 2017), is also critical to marketers' evaluations of the effectiveness of mobile marketing. Another group of studies have further examined consumers' perceptions and attitudes toward location-based mobile marketing (Bruner and Kumar 2007, Xu 2006). In the next section, we will describe the mobile location data under analysis.

3. Data

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

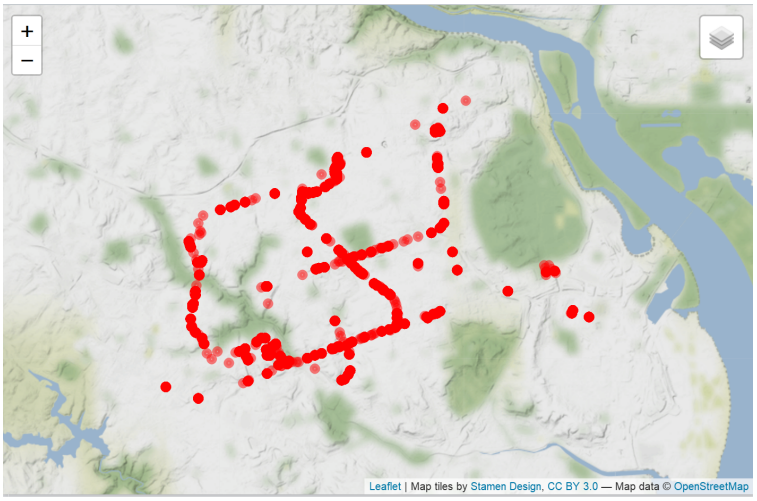


Figure 2 **An example of a consumer's footprints with 732 unique locations over the five-week sample period**

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 contains information about

- Consumer ID: a unique identifier of each consumer;
- Platform ID: an identifier of the consumer's mobile operating system (Android or iOS);
- Latitude and longitude (i.e., geo-coordinates) of the location visited;
- Timestamp: the beginning time at the location.

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

- 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 (Section 5 and Appendix A). Based on the models and parameters trained, we then conduct the focal analysis using the remaining 50% of the 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 change 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 has tested the methodologies on 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 only 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 that often has 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 privacy-preserving frameworks that can address these challenges.

4. Methodology

The proposed framework enables a location data collector to share data in a privacy preserving manner while ensuring sufficient utility to an advertiser from the shared data. Consistent with the premise of syntactic models, a data collector has some knowledge about the types of potential privacy threats (Clifton and Tassa 2013). While the proposed framework accommodates various types of privacy threats, we illustrate two commonly encountered types - sensitive attribute inference and re-identification threat. We will introduce the notations first and then formulate the privacy preservation in the context of the location data.

Definition 1. *A trajectory T_i of a consumer 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 $l_i^k = (x_i^k, y_i^k)$ is a location k visited by consumer i with geo-coordinates (i.e., a pair of longitude and latitude) x_i^k and y_i^k, t_i^k is the corresponding timestamp, and n_i is the total number of locations tracked of consumer i.*

Problem Formulation. We frame the problem of preserving privacy in location data at a consumer level. Let r_i denote a consumer i's privacy risk associated with trajectory T_i for a specific type of privacy threat, and u_i the advertiser's utility from leveraging consumer i's trajectory. A data collector aims to find a transformation $T_i \rightarrow \mathcal{P}(T_i)$, where $\mathcal{P}(T_i)$ is consumer i's obfuscated trajectory that the data collector shares with an advertiser by minimizing r_i while maintaining u_i. The transformation is based on suppressing the locations in T_i given two suppression parameters. One is $\vec{s_i}$, the suppression weight corresponding to each unique location in T_i. It is assigned based on various measures of the informativeness of each location, such as the consumer's frequency, recency, and time spent at each location. The more informative a location is, the more likely it is suppressed. The other is z_i, the suppression score for consumer i, which controls the number of locations in T_i to be suppressed. It is assigned based on the consumer's privacy risk. The higher the risk for consumer i, the more locations are suppressed in T_i. Both parameters contribute to the final suppression probabilities assigned to each location in T_i. 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. The corresponding risk and utility of the obfuscated trajectory $\mathcal{P}(T_i; \{\vec{s_i}, z_i\})$ 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 privacy threat and business objective of the advertiser, respectively.

Overall, for a set of N consumers' trajectories $T = \{T_1, ..., T_N\}$, the data collector aims to find a transformation of T, $T \to \mathcal{P}(T; \{\vec{s_i}, z_i\}_{i=1}^N)$, to produce the obfuscated trajectories to be shared with the advertiser that minimize the expected privacy risk $E(r_i)$ across all consumers while maintaining the expected data utility $E(u_i)$ to the advertiser. Consistent with our focal research questions and overview of the three components of the proposed framework (Fig. 1), we further break down the data collector's problem into three sub-problems below. The first two pertain to the estimation of u_i and r_i based on $\mathcal{PR}$ and $\mathcal{U}$, respectively; and the third is to identify the suppression parameters $\{\vec{s_i}, z_i\}$.

Research Question 1. ***Quantification of Consumer's Privacy Risk:*** *Given the consumers' trajectories T and a privacy threat $\mathcal{PR}$, we quantify each consumer's risk $\{r_1, ..., r_N\}$, where each $r_i \in [0, 1]$ indicates the stalker's success rate in inferring the private information from consumer i's trajectory T_i.*

Research Question 2. ***Quantification of Advertiser's Utility:*** *Given the consumers' trajectories T and a business objective $\mathcal{U}$, we quantify each trajectory's utility to an advertiser $\{u_1, ..., u_N\}$.*

Research Question 3. ***Obfuscation Scheme for Data Collector:*** *Given consumer trajectories T and their corresponding risks, for an advertiser's business objective $\mathcal{U}$, we identify an obfuscation scheme $T \to \mathcal{P}(T; \{\vec{s_i}, z_i\}_{i=1}^N)$ to balance the average risk and utility across consumers.*

Next, we will illustrate the quantification of two classes of privacy risks in Section 4.1 and quantification of the data's utility to an advertiser in one business application of POI recommendation in Section 4.2. Finally, in Section 4.3, we will propose an obfuscation scheme that provides a balance between the privacy risks and data utility.

4.1. Quantification of Consumer's Privacy Risk

The first step of the proposed framework is quantifying each consumer's privacy risk. To accomplish this, we simulate a stalker's actions and assign its success rate in obtaining a consumer's sensitive information as the consumer's privacy risk. Privacy threats could range from using simple heuristics, such as querying the consumers' trajectories, to leveraging more robust machine learning heuristics to predict consumers' sensitive 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 the most commonly encountered stalker threats. The first type is "sensitive attribute inference", where a stalker could employ robust machine learning heuristics to infer sensitive information, such as home address (Yang et al. 2018). The second type is "re-identification threat", where a stalker aims to infer a consumer's complete set of locations T_i, that is, identify consumer i, from the published trajectories $\mathcal{P}(T)$ (Pellungrini et al. 2018). With some background knowledge, such

as a subset of a consumer's locations $\overline{T}_i \in T_i$, a stalker could query the published trajectories $\mathcal{P}(T)$ to identify a subset of J consumers who have visited all locations in $\overline{T}_i$. A lower J indicates a higher re-identification risk. Next, we will describe the features that an stalker could extract from the published trajectories and quantify each of the two types of privacy risks discussed above.

4.1.1. Trajectory Feature Extraction. To replicate a stalker's adversarial actions and assess each consumer's privacy risks, we extract a comprehensive set of features from the trajectories examined by the literature, $\mathcal{F}(T)$[4] (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 5.1, will also help a data collector interpret which features contribute the most to the privacy risks, gain insights on possible obfuscation schemes, and quantify and interpret the data utility to an advertiser. We will categorize these features as below.

1. **Consumer Mobility**: This set of features captures a consumer's aggregate mobility patterns based on the locations visited in T_i, such as the consumer's frequency to, time spent at (Pappalardo et al. 2016), and distance traveled to 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 features is listed in Table 2.

2. **Consumer-Location Affinity**: Leveraging the literature on learning significant locations from predicting movements across trajectories (Ashbrook and Starner 2003, Zheng et al. 2010), we build three arguably most straightforward consumer-location tensors: the frequency to, time spent at, and total distance traveled from the immediate prior location to each location by a consumer at a weekly level. Each of these three tensors is of order three—consumer by unique location by week. We then extract consumer specific, lower dimensional representations by performing a higher order singular value decomposition (HOSVD) on the three tensors separately (De Lathauwer et al. 2000). HOSVD is typically 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 visited by these consumers, the extracted features would effectively capture the affinity of the consumers to significant locations.

3. **Consumer-Consumer Affinity**: Prior studies have also predicted consumer network or social links based on trajectories (Wang et al. 2011). We thus quantify the consumers' co-location behaviors by building consumer-consumer affinity tensors based on the locations that the consumers share at a weekly level. Each tensor would of order three —consumer by consumer by week. We

[4] To simplify the notation, we use $\mathcal{F}(T)$ to refer to $\mathcal{F}(\mathcal{P}(T)$. As we will see later in Section 4.3, both $\mathcal{P}(T)$ and T are a set of trajectories as defined Def. 1. Hence, any operation ($\mathcal{F}$ here) performed on T is applicable to $\mathcal{P}(T)$ as well.

populate three such tensors with the average frequency to, total time spent at, and distance traveled to each co-visited location within a week, respectively. Next, we perform a HOSVD on each of these three tensors to extract the consumer specific low dimensional representations indicative of the affinity to other consumers. The incremental benefit of the affinity features is discussed in Appendix E.

Stylized Example. We illustrate the above consumer-location and consumer-consumer affinity features using a stylized example. Consider three consumer 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 weekly level. That is, $T = \{T_1, T_2, T_3\}$ reveals that these three consumers visited four unique locations over a period of two weeks. Each of the three consumer-location tensors discussed above would be of size $[3 \times 4 \times 2]$ for the 3 consumers, 4 unique locations, and 2 weeks For instance, the frequency matrix of the first consumer with T_1 is $\begin{pmatrix} 1 & 1 & 0 & 0 \\ 2 & 0 & 0 & 0 \end{pmatrix}$, where the rows and columns correspond to the 2 weeks and 4 unique locations, respectively, and each entry in the matrix captures the number of times that this consumer visited each of the four locations during that week. Each of the three consumer-consumer location tensors described above would be of size $[3 \times 3 \times 2]$ for the 3 consumers by 3 consumers by 2 weeks. For instance, the frequency matrix for the first consumer with T_1 would be $\begin{pmatrix} 1 & \frac{(1+2)}{2} & \frac{(1+1)}{2} \\ 1 & 0 & 0 \end{pmatrix}$, where the rows and columns correspond to weeks and the consumer pairs 1-1, 1-2, and 1-3. Each entry in this matrix is the average frequency of the co-visited locations within each consumer pair. For instance, during week 1, $(A,1)$ was co-visited by consumers 1 and 2, and $(B,1)$ by consumers 1 and 3. The time and distance tensors are similarly constructed. We then perform a HOSVD on these tensors separately and use the first five principal components that capture a majority of the variance. Hence, for each consumer and tensor, we have five lower dimensional representations that capture the corresponding consumer-location and consumer-consumer affinities. Next, we imitate how a stalker would use the extracted features from the published trajectories to orchestrate privacy threats.

4.1.2. Sensitive Attribute Inference. Leveraging the published trajectories $\mathcal{P}(T)$ and extracted features, a stalker could infer various sensitive attributes, thus posing a privacy threat (Li et al. 2007). We train a supervised model $\mathcal{M}_{proxy}$ with the extracted features as a proxy for the stalker's model $\mathcal{M}$ to infer the sensitive attributes (Yang et al. 2018). Each consumer's risk is quantified as the certainty of identifying a sensitive attribute from the consumer's published trajectory using $\mathcal{M}_{proxy}$. We illustrate the method by inferring two sensitive attributes, home address and mobile operating system.

Specifically, we enlist Random Forest as $\mathcal{M}_{proxy}$ in light of its flexibility in handling regression and classification tasks, and its competitive performance across a wide range of supervised learning

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 consumer 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 consumer 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 consumer 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 consumer 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 consumer mobility features

algorithms (Breiman 2001, Liaw et al. 2002). For each sensitive attribute, we learn a Random Forest using the extracted features[5]. The risk is then calculated as the certainty of $\mathcal{M}_{proxy}$ in identifying the corresponding sensitive attribute, that is, the probability of correctly identifying the attribute in classification, or negative root mean square error in regression. We also perform a 0-1 normalization in regression such that $r_i \in [0, 1]$.

4.1.3. Re-identification Threat. Adapting the risk notion that a stalker is able to identify a consumer and associate the consumer with a record in the published data (Samarati 2001, Samarati and Sweeney 1998), we define re-identification threat in the context of location data. Here, a stalker tries to re-identify all locations visited by a consumer based on some prior knowledge of an (often small) subset of locations visited by the consumer, such as employer address from a membership registration form. Formally, this problem can be defined as follows:

[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 provids the best out-of-sample performance.

Definition 2. *Given the published trajectories $\mathcal{P}(T)$ and a subset of consumer i's trajectory $\bar{T}_i \subseteq T_i$, the stalker aims to identify T_i from $\mathcal{P}(T)$.*

Since a data collector does not know consumer i under threat or the subset locations $\bar{T}_i$ a-priori, to quantify the consumer's risk r_i, the data collector would need to account for all $\binom{|T_i|}{|\bar{T}_i|}$ possible subsets of T_i, where $|T_i|$ is the total number of unique locations visited by a consumer i. For each such subset, the probability of a consumer being identified is $\frac{1}{J}$, where J denotes the number of all consumers among N who have visited all locations in $\bar{T}_i$. If no such consumer exists other than i, then the probability of identifying consumer i would be 1 for the subset considered. We quantify a consumer's re-identification risk as the maximum of these probabilities over all such subsets.

A 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 $|\bar{T}_1| = 2$. To compute the risk for consumer 1 across both weeks, we consider the subset $\{(A,B),(B,C),(A,C)\}$. Then the corresponding probabilities of identifying consumer 1 for each of these three subsets would be $\{\frac{1}{3},\frac{1}{2},\frac{1}{2}\}$, since across both weeks, 3 consumers have visited (A,B), 2 visited (B,C), and 2 visited (A,C). Thus, consumer 1's re-identification risk 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 $|\bar{T}_i|$, it is computationally inefficient with a complexity of $O(\binom{|T_i|}{|\bar{T}_i|} \times N)$. To speed up the computation, we leverage a recent study (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, by building a set of mobility features similar to $\mathcal{F}(T)$ discussed in Section 4.1.1. We adopt this idea by further augmenting the mobility features with our consumer-consumer and consumer-location affinity features. We then analytically compute the risks for a subset of the consumers and use the trained model to approximate the risks for the rest of the consumers (see Appendix A for the technical details).

4.2. Quantification of Advertiser's Utility

Having quantified each consumer's privacy risks associated with the two commonly encountered privacy threats (our research question 1), we next examine the utility that an advertiser would derive from the published trajectories (research question 2). The behavior-rich nature of the location data enables advertisers to derive great insights and perform various targeted marketing activities to reap monetary benefits. In this work, we consider a popular business application, POI recommendation (Ashbrook and Starner 2003). The underlying idea is to leverage the historical consumer preferences revealed in the trajectories to predict the locations that a consumer is most likely to visit in the future. This would enable an advertiser to target the consumer with relevant, contextualized marketing messages (Ghose et al. 2018). To this end, we quantify an advertiser's

utility by learning a recommendation model. Intuitively, more accurate POI predictions will render better targeting and thus a higher utility for the advertiser. Hence, we quantify u_i, the utility of consumer $i's$ trajectory, as the predictive accuracy of the recommendation model.

Most recommendation models leverage collaborative filtering to identify other consumers with similar historical preferences to infer the focal consumer's preference (Bobadilla et al. 2011). This idea is consistent with human social behavior: people tend to account for their acquaitances' tastes, opinions, and experiences when making own decisions. We thus imitate an advertiser's use of the location data for POI recommendation and compare a number of recommendation models examined in the literature (Appendix A). We focus the following discussion on the best performing nearest neighborhood (NN) based learning technique. Simply put, the main idea of NN is to identify the m consumers most similar to the focal consumer, namely m neighbors, and utilize their locations to predict the focal consumer's future locations. The similarity is computed based on the visited locations that reveal each consumer's preference by leveraging the set of features extracted from the published trajectories described in Section 4.1.1. To find the m most similar consumers, we compute the cosine similarity between two consumers' features $\mathcal{F}(T_i)$ and $\mathcal{F}(T_j)$:

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

After identifying the m most similar consumers to a consumer i, denoted as M_i, we aggregate and rank the unique locations visited by M_i based on a combination of visit frequency and these m consumers' similarities to consumer i. Specifically, for each consumer $j \in M_i$, location $l \in T_j$, let f_j^l denote the number of times that consumer j visited location l, then the rank of a location l for consumer j is determined by:

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

In the above equation, $\frac{f_j^l}{\sum_l f_j^l}$ is the normalized visit frequency at a consumer 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 consumer. We further aggregate o_{ij}^l across all the consumers who visited the location l in M_i by computing the mean of o_{ij}^l:

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

where $1(j \in T_j) = 1$ if consumer j has visited location l and zero otherwise. The higher the value of o_i^l, the more likely that a consumer i visits location l in the future. The next k locations (ordered by time) most likely visited by consumer i hence correspond to the top k such ranked locations. The

utility of consumer i's trajectory T_i to the advertiser is then measured as the predictive accuracy of the recommendation model for the different values of k, measured by the widely used information retrieval metrics that assess the quality of the recommendations: Average Precision at k ($AP@k$ or AP_i^k) and Average Recall at k ($AR@k$ or AR_i^k) (Yang et al. 2018). Specifically, let $L_i = \{l_i^1, l_i^2, , l_i^{k'}\}$ be the actual next k' locations visited by consumer i and $\overline{L_i} = \{\overline{l_i^1}, \overline{l_i^2}, ..., \overline{l_i^k}\}$ be the top k locations predicted by the NN recommendation model as described above. Then AP_i^k and AR_i^k are:

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

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

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 the expected utility of all consumers' trajectories to the advertiser $E(u_i)$ is calculated as $MAP@k$ and $MAR@k$, i.e., the mean AP_i^k and mean AR_i^k, respectively, across all consumers. Also, the parameter m (number of the most similar neighbors) is selected by performing a five-fold cross-validation aimed at maximizing the accuracy of the recommendations (details in Section 5.2), a technique commonly used in the statistical learning literature to ensure a good out-of-sample performance (Friedman et al. 2001).

4.3. Obfuscation Scheme

The last step in our framework is to address the third research question – devising an obfuscation scheme for the data collector that would balance the privacy risks to the consumers and the utility of the published trajectories to the advertiser. As discussed earlier, given the unique properties of trajectory data, such as high dimensionality, sparsity, and sequentiality, 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 utility 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) requires multiple parameter inputs from a data collector, including the threshold of the stalker's success probability and the stalker's background knowledge in each type of threat. LSup (Terrovitis et al. 2017) requires similar inputs. Given the complex nature of such heuristics, setting these parameters and interpreting the resulting obfuscations for practical purposes is non-trivial. Moreover, the current techniques do not provide the flexibility

for a data collector to choose among multiple obfuscation schemes. Addressing these critical challenges, we develop $T \rightarrow P(T, \{\vec{s_i}, z_i\}_{i=1}^{N}))$, a personalized consumer-level suppression technique that is interpretable to the data collector. It requires no input parameter for the sensitive attribute inference and merely one input parameter for the re-identification threat – the number of a consumer's locations already known to the stalker $|\overline{T_i}|$. Furthermore, the data collector will enjoy the flexibility of choosing among multiple interpretable obfuscations for each type of privacy threat.

In our obfuscation scheme, a consumer's trajectory T_i is suppressed based on two consumer-specific suppression parameters $\{\vec{s_i}, z_i\}$. As described earlier, the suppression score z_i controls the number of locations to be suppressed in a consumer i's trajectory T_i, and the suppression weights $\vec{s_i}$ denote the likelihood for each unique location to be suppressed. A naive approach to identify $\{\vec{s_i}, z_i\}$ that balance the risk and utility is to search over a random grid of positive values of $\vec{s_i}$ and z_i. However, this would be computationally inefficient, contingent on the grid of values chosen, and potentially resulting in no parameters that could satisfactorily balance the risk and utility and hence requires a more sophisticated grid search.

A more structured approach to identify the parameters would be to consider a grid that ensures reduction in consumer's risk and assesses the corresponding reduction in utility to pick a specification that satisfactorily balances the risk-utility trade-off. Intuitively, more locations suppressed would mean lower risks to the consumers and lower utility to the advertiser; and in the extreme scenario of no trajectories published, both risk and utility would be zero. Also, to ensure similar risk reduction for a high-risk and a low-risk consumer, the number of locations suppressed would need to be proportional to the consumer's privacy risk r_i, that is, $z_i = r_i \times p$, where $p \in [0, 1]$ is a grid parameter.

While z_i ensures that the number of locations suppressed is proportional to the consumer's risk r_i, to further limit the information available to perform a stalker threat, the more informative locations within T_i would need to be suppressed with higher probabilities. Since the informativeness is related to the possible features that can be extracted by a stalker from T_i (Section 4.1.1), we assign the suppresion weights $\vec{s_i}$ based on the key features capturing the informativeness - frequency, recency, or time spent at each location. 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 based on the corresponding frequencies $\{f_i^1, f_i^2, ..., f_i^{k_i}\}$ *are* $\vec{s_i} = \{\frac{f_i^1}{\sum_{j=1}^{k_i} f_i^j}, \frac{f_i^2}{\sum_{j=1}^{k_i} f_i^j}, ..., \frac{f_i^{k_i}}{\sum_{j=1}^{k_i} f_i^j}\}$.

Combining the two parameters $\{\vec{s_i}, z_i\}$ described above, we can calculate the suppression probability of each unique location in T_i. Then the unique locations are independently suppressed with Bernoulli trials given the following probabilities:

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

For a value of p, the base suppression probability (z_i) ensures that consumers at higher risks would have more locations suppressed. The additional term ($z_i \times s_i^j$) ensures that a more informative location j is suppressed with a higher probability ($z_i + z_i \times s_i^j$). Since each consumer i's risk r_i and the suppression weights $\vec{s_i}$ can be computed apriori from the original unobfuscated data, the suppression probabilities above depend only on the grid parameter p. Suppressing the location data to limit a stalker's ability to invade private information would also adversely affect a butler advertiser's utility derived from $\mathcal{P}(T)$. For instance, in the extreme scenario when each consumer's risk $r_i = 1$ and p is reasonably high, all locations would be suppressed (i.e., complete suppression[6]: $\{\mathcal{P}(T_i)\} = \mathcal{P}(T) = \emptyset$), resulting in no utility to the advertiser, nor threat to consumer privacy. A similar inference can be made when $p = 0$ (i.e., no suppression: $\mathcal{P}(T) = T$), resulting in high data utility and also high privacy risk. Noting these two extreme scenarios, we empirically determine the suppression parameters $\{\vec{s_i}, z_i\}$ by varying the grid parameter p to derive the published trajectories $\mathcal{P}(T)$ that balance the risk and utility.

The proposed obfuscation scheme has two main advantages. First, the structured grid search by varying the grid parameter p provides the data collector with multiple trade-off choices. Second, the identified $\{\vec{s_i}, z_i\}$ provide the data collector with consumer level interpretability of the obfuscation. By fine-tuning $\{\vec{s_i}, z_i\}$, our ultimate goal is to understand, quantify, and optimize the trade-off between the data utility ($\mathcal{U}$) and privacy risk ($\mathcal{PR}$) in a meaningful way.

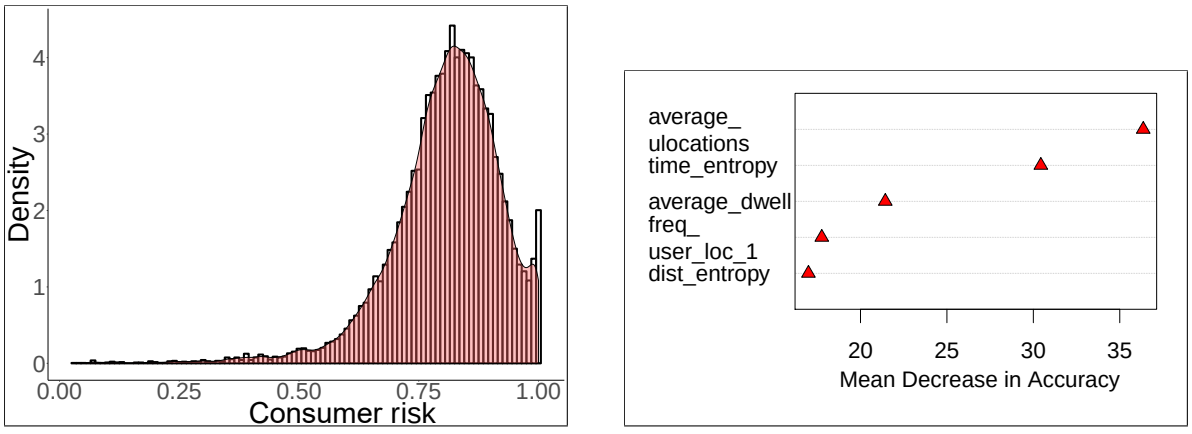
5. Empirical Study

Consistent with the proposed framework (Part A of Figure 1), prior to obfuscation, we first compute each consumer i's baseline risk r_i (Section 4.1) and suppression weights $\vec{s_i}$ (Section 4.3) on the unobfuscated data. We also compute the baseline data utility $MAP@k$ and $MAR@k$ across all consumers on the unobfuscated data (Section 4.2). Then for each $p \in \mathcal{G}_p = \{0, 0.1, ..., 1\}$, we obfuscate each consumer's trajectory based on the suppression probabilities computed from the above r_i, $\vec{s_i}$ and p (Equation 6); and re-compute the mean risk and utility across all consumers on the corresponding obfuscated data to assess the percentage decrease in the mean risk and utility from the baseline mean risk and baseline utility, respectively (Part B of Figure 1). This repeated process with varied p will offer the data collector multiple options to balance the risk and utility. We will report the details and key findings below.

[6] Note that $s_i \in [0, 1]$; and $z_i \in [0, 1]$ because $r_i \in [0, 1]$ and $p \in [0, 1]$. Nonetheless, the corresponding location is suppressed with probability 1 whenever $(z_i + z_i \times s_i) > 1$.

5.1. Quantification of Consumer's Privacy Risk

As described above, for each type of threat, we quantify each consumer's baseline risk r_i and suppression weights $\vec{s}_i$ without obfuscation. Then based on the suppression probabilities calculated from these r_i, $\vec{s}_i$, and each $p \in \{0, 0.1, ..., 1\}$, we perform consumer-level obfuscation. Each p leads to a different set of obfuscated trajectories and hence re-computation of the mean risk and utility across consumers. To obtain consistent estimates of the mean risk and utility, we use bootstrapping with 20 trials for each p. In the sensitive attribute threat, we consider two sensitive attributes of home address and mobile operating system. To train the predictive model, we mimic a stalker with access to a training sample of known trajectories and sensitive attributes. We split the data into two random samples: 50% training set (T_{train}) with 20,000 consumers to train the predictive model, and 50% test set (T_{test}) with 20,012 consumers. As described in Section 4.1, we use Random Forest regressor to predict the risk of inferring home location and use Random Forest classifier to predict mobile operating system and the re-identification risk. 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 on T_{test} in each privacy threat. In Figure 4, we report the average risk across all consumers in T_{test} for each p. To compute the re-identification risk, we assume the number of locations in each consumer's trajectory already known to a stalker is 2, that is, $|\bar{T}_i| = 2$ in Definition 2, to illustrate our approach.



(a) Consumer risk density - OS inference threat (b) Feature importance, OS inference model

Figure 3 Personalized Risk Management Insights

A data collector can gain a host of insights from the initial step of quantifying consumers' privacy risks prior to obfuscation, such as which consumers are at the greatest risk, what is the severity of

each privacy risk, which feature is most informative to a stalker and hence should be suppressed. For example, Figure 3a offers the data collector a visual of the distribution of the consumers' risks if a stalker were to infer their operating systems from the unobfuscated trajectory data. It shows that the majority of the consumers carry a relatively high risk (≥ 0.75 chance of success for a stalker) of their sensitive attribute of operating system being inferred if no obfuscation were performed. Also, the average risk of home address inference is 0.84. By assessing the error of the Random Forest regressor learned to predict the home address, we find that on average a stalker could successfully identify a consumer's home address within a radius of 2.5 miles (Appendix A). Further, the average risk of re-identifying an individual's entire trajectory by knowing merely two randomly sampled locations is 0.49, that is, a 49% chance of success for a stalker. In addition, the data collector can assess the worst cases associated with the top-risk consumers in each of the above threats.

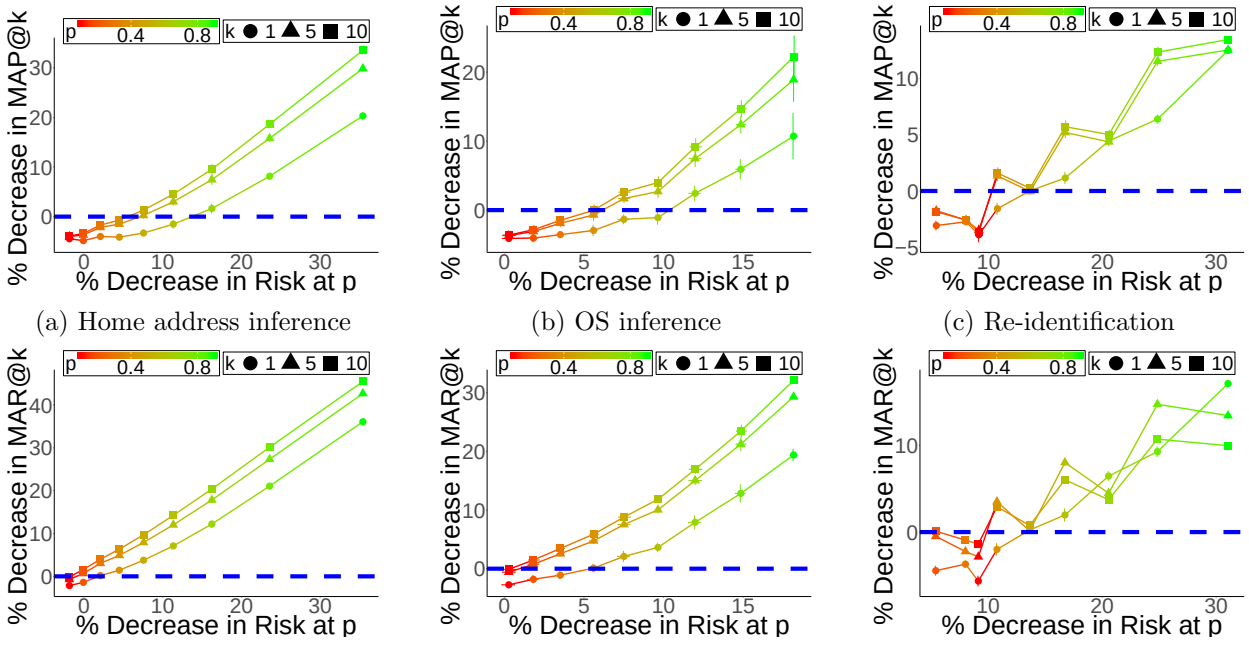
Despite these paramount privacy risks arising from unobfuscated location data, they can be curtailed by a data collector using the proposed framework. For instance, the risk associated with the operating system inference could be reduced by 10% while fully preserving the data utility on the POI@1 performance (Figures 4b, 4e, $p = 0.6$). 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 consumer-specific privacy risk to better understand the customer lifetime value (Berger and Nasr 1998) and personalize customer relationship management.

In addition, a data collector may look at the feature importance prior to obfuscation. For instance, Figure 3b displays the top five most important features of the Random Forest trained to compute the consumers' risks in Figure 4b. A data collector can infer that the temporal information of the trajectories (`time_entropy` and `average_dwell`) contributes most to the model's predictive performance. Hence, a possible obfuscation scheme that removes (even partially) the timestamps in the trajectories would prevent the stalker from constructing the temporal features and hence considerably reduce the consumers' risks. Similar insights can be gained by analyzing the risk scores related to other stalker threats - home address inference and re-identification threat considered in the work.

5.2. Quantification of Advertiser's Utility

Next, we compute the data utility to a butler advertiser by leveraging a collaborative filtering recommendation heuristic to predict each consumer's future locations. To assess the predictive accuracy, we use the locations actually visited by each consumer in the fifth week as the ground truth and train the recommendation model to predict the locations. A neighborhood-based recommender is learned on a grid of $\{5, 10, 25, 50, 100, 200\}$ to tune the number of neighbors via a

five-fold cross-validation on the obfuscated data (Bobadilla et al. 2011). The model ranks the locations that a consumer 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 to maximize the prediction accuracy. Then, we compute the average utility for the advertiser across all consumers, $MAP@k$ and $MAR@k$, for $k = \{1, 5, 10\}$ to illustrate the method's efficacy, where k is the next k locations. The model can also be used to compute $MAP@k$ and $MAR@k$ for other values of k. We perform 20 trials for each p and report the mean and 95% confidence intervals of the utility (Figure 4). A more detailed explanation of the utility computation is available in the Appendix E.



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

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

Figure 4 **Proposed framework - $MAP@k$ and $MAR@k$ for varying p**

5.3. Obfuscation Scheme for Data Collector

In Figures 4a, 4b, and 4c, we visualize the risk-utility trade-off based on $MAP@k$. As described earlier, the locations in each T_i are suppressed based on the suppression probabilities computed from p and $\{\vec{s_i}, z_i\}$. We will focus on discussing the results where the suppression weights $\vec{s_i}$ are computed based on the frequency to each location, although we have also computed $\vec{s_i}$ based on recency and time spent at each location (Appendix C). In Figures 4a, 4b, and 4c, the X and Y axes display the percentage decrease in the mean risk and $MAP@k$ from the baseline risk and baseline $MAP@k$ for each $p \in \mathcal{G}_p$. 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, with different k and p. 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 4d, 4e, and 4f where we compare the percentage decreases in $MAR@k$ to the percentage decreases in the mean risk.

In all graphs in Figure 4, we observe that as we increase p, the values along both axes increase. This is expected since an increase in p, for the same consumer risk scores, more locations get suppressed, thus 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 consumer risk scores that capture the success of a privacy 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.

5.4. Model Comparison

We compare the proposed obfuscation scheme with eight different baselines corresponding to three types of obfuscation approaches – obfuscation rules derived from timestamps of consumer locations, alternate suppression schemes based on consumer risk and the latest work in syntactic models LSup and GSup (Terrovitis et al. 2017).

5.4.1. Comparison to Rule-based Obfuscations. We derive a few practical rules for obfuscation based on the timestamps of the locations in the 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 the locations in sleeping hours and working hours (9 AM - 6 PM) on weekdays, or 3) remove the timestamps of the 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 adversely affect the advertiser's utility. For instance, if the timestamps of the location data were to be removed, both

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

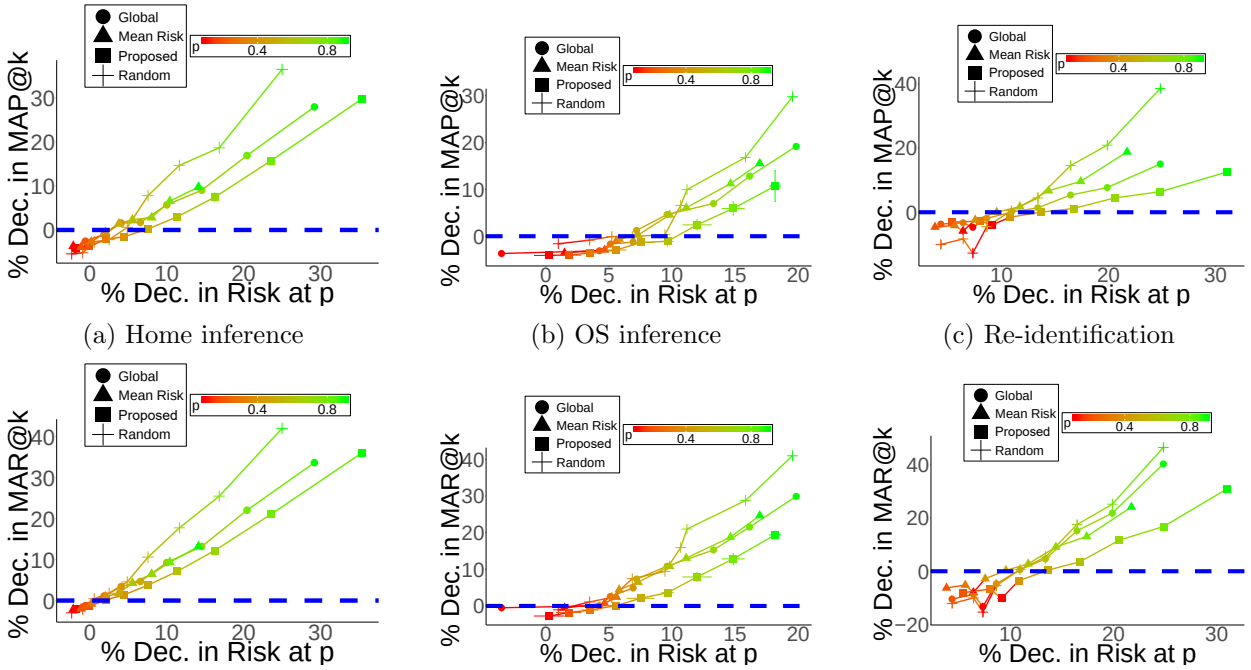
the mobility features, `time_entropy`, `time_rog`, `average_dwell` (from Table 2) and the consumer-consumer, consumer-Location affinity features (Section 4.1.1) based on time spent by a consumer at a location cannot be computed.

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 the risk and utility. In the home address inference threat (Figure 4a, $p = 0.7$, $k = 1$), we find that a risk to consumer privacy can be reduced by 15% (maximum decrease when compared to the rule-based heuristics) with less than 1% decrease in $MAP@1$ (minimum decrease). A similar trend is observed in the re-identification threat (Figures 4c, 4f). In the operating system inference (Figure 4a, $p = 0.9$, $k = 1$), we observe that the risk is reduced by $\approx 18\%$ compared to 25.49% when the timestamps are removed. However, this is achieved with a lesser decrease in the utility $\approx 10\%$ using the proposed framework when compared to 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.

5.4.2. Comparison to Risk-based Obfuscations. We further compare the proposed obfuscation scheme to three other risk-based suppression baselines. These baselines are devised to show the efficacy of the quantification of each consumer's risk and personalized suppression achieved by introducing and identifying consumer-specific parameters $\{\vec{s_i}, z_i\}$ in our framework.

1. **Random** - In this baseline, we do not perform suppression of locations at a consumer 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 the proposed obfuscation scheme to make it comparable.

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



(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** $MAR@1$

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 varied risk levels across consumers. This is different from the proposed obfuscation scheme where a tuple may not be suppressed for a lower risk consumer but has been suppressed for a higher risk consumer. For each tuple, we assign the mean of all consumers' risk scores as the tuple's risk score, vary p, and perform suppression.

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

5.4.3. Comparison to Latest Suppression Models. Finally, we compare the proposed framework to the latest 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 comes with an added benefit that the proposed obfuscation scheme only requires one input parameter corresponding to the number of locations known to a stalker in the re-identification threat, as compared to the larger sets of parameters required by LSup or GSup. The detailed comparisons are described in Appendix B.

6. Conclusion

Smartphone location tracking has created a wide range of opportunities for data collectors to monetize location data (Valentino-Devries et al. 2018, Thompson and Warzel 2019). Leveraging the behavior-rich location data for targeting is proven to be an effective mobile marketing strategy and 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 yet under-studied topic from a data collector's perspective. Specifically, we identify three key challenges facing a data collector and propose an end-to-end framework to enable a data collector to monetize and share location data with an advertiser while preserving consumer privacy.

The existing literature on privacy preservation is unsuited for this new type of data with distinct characteristics, not interpretable to the data collector, or not personalized to an individual level. Our research fills this gap. Specifically, we propose a framework of three components, each addressing a key challenge facing a data collector. First, we quantify each consumer's risks, exemplified by two common types of stalker behaviors – sensitive attribute inference 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 data to an advertiser by considering a popular business use case - POI recommendation. 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 risk levels associated with each privacy threat and the informativeness of each location in their trajectories. The proposed obfuscation scheme also 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 individual's 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 most to the risk associated with each privacy 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 consumer 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 the proposed framework with 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 consumer. 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 considers 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 consumers' online click streams or social media comments, become linked to their mobile location data, more sophisticated privacy preservation methodologies may be developed. Lastly, in the proposed obfuscation scheme, the location data is obfuscated by assuming one consumer privacy threat and one advertiser objective at a time. That is, a composition of privacy threats or business objectives is not addressed. This calls for further methodological research to address this in the location data sharing paradigm.

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.

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.

Stuart A. Thompson and Charlie Warzel. Twelve Million Phones, One Dataset, Zero Privacy. https://www.nytimes.com/interactive/2019/12/19/opinion/location-tracking-cell-phone.html, 2019.

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.

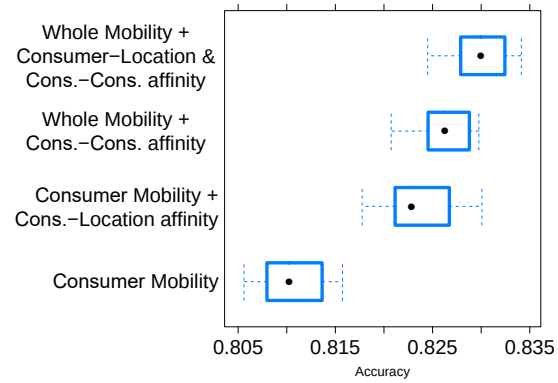
Dingqi Yang, Bingqing Qu, and Philippe Cudre-Mauroux. Privacy-preserving social media data publishing for personalized ranking-based recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 2018.

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.

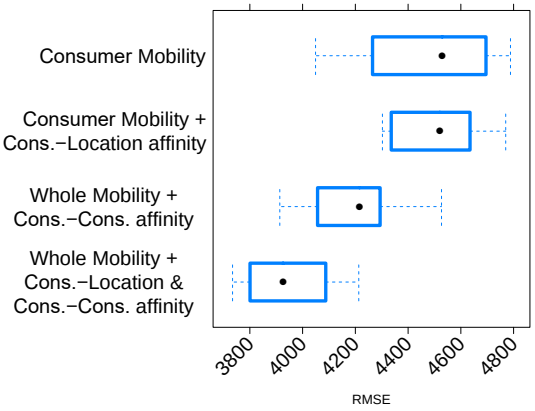
Yu Zheng, Xing Xie, and Wei-Ying Ma. Geolife: A collaborative social networking service among user, location and trajectory. *IEEE Data Eng. Bull.*, 33(2):32–39, 2010.

Appendix A: Model Choices in the Proposed Framework

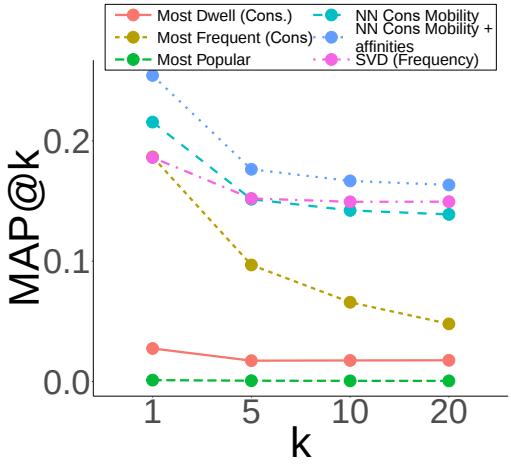
We empirically justify our model choices in the proposed framework. All choices are made based by assessing the performance of different machine learning heuristics used in our framework on the unobfuscated data. First, in Figures 6a and 6b, we show the incremental benefit of the affinity features discussed in extracting the features $\mathcal{F}(T)$. Figure 6a shows the accuracy of the Random Forest classifier in predicting each consumer's operating system. The model is regularized by performing a grid search on the maximum number of features $\{.25, .5, .75, 1\}$ and trees $\{50, 100, 200\}$ via five-fold cross-validation. The best performing model has an accuracy of 82% which indicates the success that a stalker would have in inferring the unpublished operating system of a consumer from the trajectory data. In Figure 6b, we plot the RMSE of the Random Forest regressor trained to predict home addresses.[7]



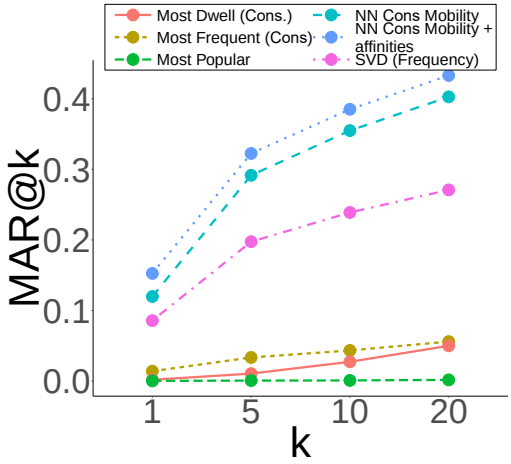
(a) OS inference threat - Model Accuracy



(b) Home address inference - Model RMSE



(c) $MAP@k$ POI Recommendation model



(d) $MAR@k$ POI Recommendation model

Figure 6 Proposed framework model choices

[7] We treat each consumer's most frequently visited location 10pm-6am as the ground truth of home location. The results remain robust across alternative time periods, e.g. 11pm-5am. We do not save these home locations to preserve consumer privacy.

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 and 6b), we notice that the consumer-consumer and consumer-location affinity features incrementally improve the performance of both proxy models learned. In Figures 6c and 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), locations that the consumer spent the most time in (Most Dwell (consumer)), visited most frequently (Most Frequent (consumer)), and a singular value decomposition (SVD) on the consumer-location matrix populated with visit frequency. We observe that the NN based model performs better in both metrics compared to the baselines, justifying the choice. The RMSE, 3,900 meters $\approx$ 2.46 miles indicates the success that a stalker would have in identifying a consumer's home location from the unobfuscated data. Further, we also notice the incremental benefit of the affinity features in the recommendation performance (See NN consumer Mobility vs NN consumer Mobility + affinities in Figures 6c and 6d).

Appendix B: 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 models obfuscate the location data to reduce the re-identification threat while 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 the sensitive attribute inference. Second, a location is suppressed either globally across all consumers (GSup) or locally for a subset of consumers (LSup). In contrast, our suppression scheme, due to the introduction of the two consumer specific parameters $\{\vec{s_i}, z_i\}$, suppression may be performed at a consumer level with varying suppression probabilities assigned to each location visited by the consumer. In addition, compared to the parsimonious input that our proposed framework requires, both models in consideration require multiple input parameters, such as the number of adversaries $\mathcal{A}$, background knowledge of each adversary in $\mathcal{A}$, and P_{br} that controls the number of locations suppressed either locally (LSup) or globally (GSup). The higher the value of P_{br}, the lower is the number of locations suppressed. In our comparison, we follow the empirical evaluation framework of the authors to set the number of adversaries $\mathcal{A}$, set the background knowledge of each adversary in $\mathcal{A}$, and merely vary P_{br}.

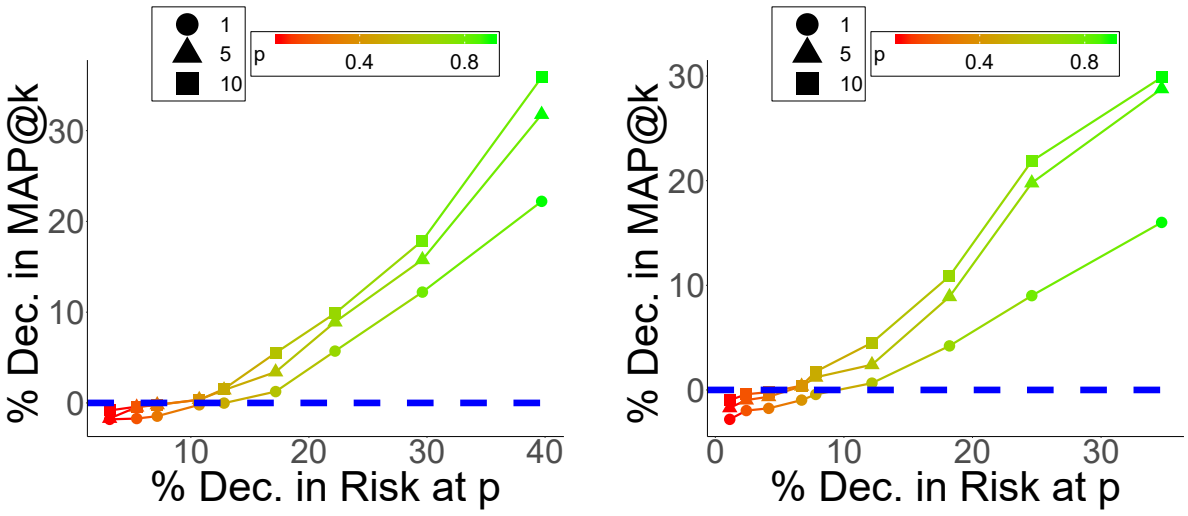
In Table 4, we present the decrease in the consumer risk from the unobfuscated trajectories for the two types of privacy threats - re-identification and sensitive attribute inference[8] (operating system and home address inference) and the corresponding measures of advertiser's utility as $MAP@1$, $MAR@1$. To identify the obfuscation scheme that provides the better/worse trade-off, we compute the slope ($\frac{Y}{X}$ in Figure 4 —%

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

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)**

decrease in the utility divided by % decrease in the risk) for different decreases in the 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 comes with an added benefit that the proposed framework only requires one input parameter – the number of locations already known to a stalker in the re-identification threat, as compared to the great sets of parameters required by 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.**

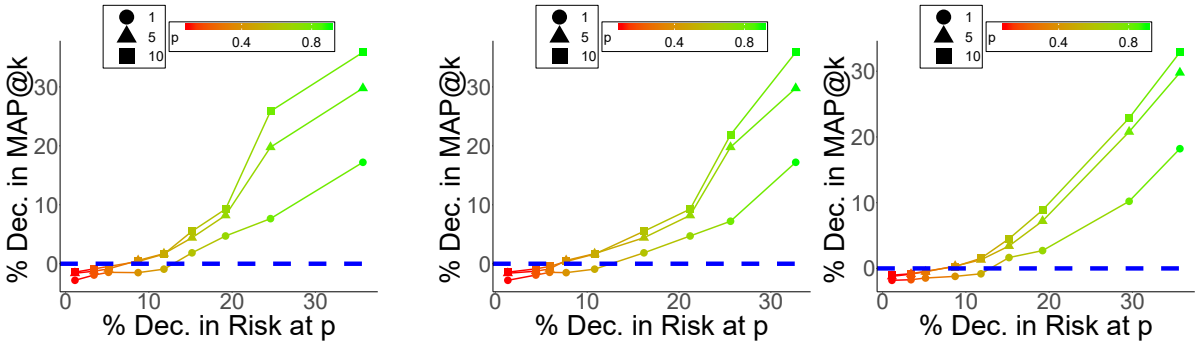
Appendix C: Suppression based on Recency and Time Spent

In the proposed suppression scheme (Section 4.3), we introduce and provide a structured grid search by varying the grid parameter p to identify the two consumer specific parameters $\{\vec{s_i}, z_i\}$. Recall that z_i controls 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 the

Figure 4 of our empirical study (Section 5), we assign $\vec{s_i}$ based on the frequency of the location visited in T_i. Here, we further augment the empirical study and showcase the flexibility of the proposed suppression scheme by assigning the $\vec{s_i}$ based on the 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 and 7a. Similar to Figure 4, we observe that for a given percentage decrease in the risk, there is a lesser corresponding percentage decrease in the utility.

Appendix D: Varying Sample Sizes

To test for the robustness of the results discussed in Figure 4, 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 the location visited by a consumer (similar to Figure 4) for the home address inference threat. The resulting plots comparing the percentage decreases in the consumer risk and advertiser utility from the baselines calculated on the unobfuscated data are visualized in Figures 8a, 8b, and 8c. We note that even at smaller samples, the slope (i.e., the % decrease in the utility divided by the % decrease in the risk) at different values of p is similar to that of the full sample (Figure 4a).



(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

Appendix E: Utility Measurement

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

Based on the consumer risks, we obfuscate T_{train} by varying $p \in \mathcal{G}_p$. We learn a neighborhood-based recommendation model (Bobadilla et al. 2011) by tuning the number of neighbors via five-fold cross-validation on the obfuscated training sample $\mathcal{P}(T_{train})$. The model is learned to rank the locations that a consumer is likely to visit during the fifth week of the observation period. That is, we build the features $\mathcal{F}(P(T_{train}))$ on first four weeks' data and tune the number of neighbors by using a grid of $\{5, 10, 25, 50, 100, 200\}$ to

maximize the predictive accuracy. Then, we compute the data utility, $MAP@k$ and $MAR@k$, on T_{test} for $k = \{1, 5, 10\}$ to illustrate the efficacy of the proposed method. The learned recommendation model can be used to compute $MAP@k$ and $MAR@k$ for other values of k as well. Intuitively, $MAP@1$ and $MAR@1$, for example, represent an advertiser's utility to predict the next location most likely visited by a consumer in the fifth week based on the recommendation model 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 a consumer's true preference revealed in the unobfuscated test sample. Similar to the risk computation, we perform 20 trials for each p and report the mean and 95% confidence intervals of the utility metrics in Figure 4.