Forthcoming at *Information Systems Research*

Personalized Privacy Preservation in Consumer Mobile Trajectories

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

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

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

Anindya Ghose
New York University (NYU) - Leonard N. Stern School of Business, ag122@stern.nyu.edu

Ubiquitous mobile technologies have been producing massive swaths of consumer location data, giving rise to an elaborate multi-billion-dollar ecosystem. In this ecosystem, some consumers share personal data in exchange for receiving economic benefits including personalized recommendations, data aggregators curate and monetize data by sharing data with advertisers, and advertisers often utilize such data for location-based marketing. While these various entities can benefit from such data sharing, privacy risks can prevail. This creates an opportunity for data aggregators to implement an effective privacy preserving framework to balance potential privacy risks to consumers and data utilities to advertisers before sharing data with advertisers. We hence propose a personalized and flexible framework that quantifies personalized privacy risks, performs personalized data obfuscation, and flexibly accommodates a variety of risks, utilities, and acceptable levels of risk-utility trade-off. Leveraging machine learning methods, we illustrate the power of the framework with two privacy risks and two utilities. Validating the framework on one million consumer trajectories, we demonstrate potential privacy risks in the absence of data obfuscation. Outperforming ten baselines from the latest literature, the proposed framework significantly reduces each consumer's privacy risk while preserving an advertiser's utility. As industries increasingly unleash the power of location big data, this research offers an imperatively needed framework to balance privacy risks and data utilities, and to sustain a secure and self-governing multi-billion dollar location ecosystem.

Key words: consumer privacy, privacy preservation data publishing, mobile location data, machine learning, location-based marketing

1. Introduction

Location Ecosystem and Advertiser Utility Massive volumes of mobile location data are being generated daily. This is catalyzed by wide adoptions of smartphones (76% in advanced and 45% in emerging economies) and location-based services (90% in the U.S.), such as navigation, ride

share, and food delivery (Taylor and Silver 2019). These data represent the latest form of marketing intelligence in the historical evolution of consumer data, from surveys to click stream, search, social media, and location (Wedel and Kannan 2016). Location data (or trajectory data hereafter) embed rich, granular, and spatio-temporal consumer behavior, such as visits to restaurants, gyms, and hospitals, hence enabling applications of commercial value, including geo-targeting, or point-of-interest (POI hereafter) recommendations of restaurants or location-based advertising (Luo et al. 2014, Andrews et al. 2016, Kelsey 2018, Ghose et al. 2019).[1] Location-based marketing is rapidly becoming a primary venue for campaign planning and consumer targeting, enriching both traditional and digital marketing strategies. Also, the global market for location analytics alone is projected to reach $25.5 billion by 2027.[2] As a result, an elaborate multi-billion-dollar ecosystem of location data collection, sharing (or publishing hereafter), analytics, and applications has emerged. Three central entities occupy this space.

1. *Data collector:* is either an app owner who has access to location data via its own mobile app, once its customers opt in to location tracking and data sharing; or more commonly, a data aggregator who purchases and aggregates location data from multiple app owners within its app network, and then sells the aggregated data in bulk to interested advertisers (or other end users) in compliance with consumer agreements and other privacy regulations, such as the General Data Protection Regulation (GDPR) and California Consumer Privacy Act (CCPA). Compared to a single app owner, a data aggregator can offer location data with much more comprehensive consumer and POI coverage, hence of much higher value to advertisers. Examples of leading data aggregators include SafeGraph, Xmode Social, and PlaceIQ. Most of them specialize in data curation and publishing, instead of data analytic services, in order to grant advertisers greater flexibility and broader use cases beyond geo-targeting, such as customer segmentation or new store site selection. While some data aggregators share POI-level data (such as total numbers of visits to a store), others share individual-level data of greater value, accruing from advertisers 0.5 to 2 cents per consumer-month (Valentino-Devries et al. 2018). Each location record commonly includes a device ID, timestamp, longitude, latitude, speed, and dwell time at each visited location.[3]

[1] Numerous other business applications (i.e., utilities to advertisers), such as geo-fencing, re-targeting, behavior-based insurance, advertising attribution, retail site selection, stock prediction, and non-business applications, such as smart city, event planning, COVID epidemiology, have been exemplified on various data aggregators' websites, such as `https://xmode.io/data-licensing/` and `https://www.safegraph.com/industries/retail`.

[2] `https://www.fortunebusinessinsights.com/location-analytics-market-102041`.

[3] Mobile phone numbers are not shared to preserve consumer privacy. Also, device IDs or advertising IDs were originally introduced by the iOS and Android mobile operating systems for the core purpose of advertising. These IDs can be easily reset by a consumer on his/her own phone's setting. Also, interested readers may refer to `http://bit.ly/3IJgOBN` for more information about mobile location tracking.

2. *Consumer:* is an individual who owns a smartphone with apps installed that can transmit his/her locations to the data collector if he/she opts in to sharing location data. Each consumer may choose to opt in to share their data when installing an app, or at any time while using the app, primarily to enjoy the app's location-based services, such as map navigation, location-based search, geo-targeted advertising, restaurant recommendations, and so on. As is well known in the privacy literature, consumers are increasingly willing to share their personal data with companies in return for economic benefits or convenience(Ghose 2017) and there is substantial heterogeneity in consumers' privacy preferences and valuations(Ghose 2017). They may also choose to opt out of location tracking at any time within each single app, or opt out of all apps in the data aggregator's app network on the data aggregator's app or website, or simply opt out of all location tracking on his/her smartphone using the privacy setting provided by the mobile operation system.

3. *Advertiser:* is a firm, such as a retail store, that gains access to location data, often from a data aggregator, to accomplish various marketing objectives, such as geo-targeting, POI recommendation, customer acquisition, or market research. For instance, the retail store may derive highly valuable marketing insights from location data – understanding customers' broad lifestyle, trajectory sequence before and after visiting its store, patronage to its competitors, and timing and frequency of those present and potential customers entering its vicinity. Then it may target the consumers of interest with coupons or other marketing messages by sending these consumers' device IDs to common mobile targeting platforms. Location data acquired from a data aggregator thus grant an advertiser a holistic view of its past, present, and potential customers, as well as their broad lifestyle and patronage to all POIs of relevance, hence offering much richer information than what an advertiser would normally access from its own app or stores.

Privacy Risks. Overall, access to location data can benefit all three entities above by allowing consumers to enjoy location-based services, advertisers to accomplish targeting and other marketing goals, and data aggregators to monetize the data. Nonetheless, location data can also entail potential privacy risks or privacy costs upon some consumers. "Privacy", defined by the Merriam-Webster dictionary as "the quality or state of being apart from company or observation", broadly pertains to the protection of personally identifiable information (PII hereafter), such as name and home address, and enactment of privacy policies and regulations. Privacy violation can occur when the use or release of PII violates an individual's reasonable expectation of confidentiality, or violates any law protecting such information.

A subset of advertisers, or more broadly a third party ("stalker" or "adversary" hereafter), may perform malicious acts using the published data or in combination with other sources of data, largely for short-term revenue gains. For instance, a stalker may infer a consumer's home location to launch excessive direct mail campaigns, predict political ideology to deliver political ads, perform

privacy-invasive targeting (e.g., a baby brand targeting a consumer who visited a pregnancy clinic), or identify celebrities, their estates, or those visiting the Playboy Mansion overnight, abortion clinics, mosques, and queer locations (Valentino-Devries et al. 2018, Thompson and Warzel 2019). Privacy risks may also arise if location data are linked to other sources of consumer data, such as an individual's media consumption, online search behavior, social networking activities, credit card transactions, online check-ins, ride shares, or wearable technologies. Overall, the media coverage of location tracking has elevated consumers' privacy concerns and triggered stricter mobile opt-in policies. For example, Apple's latest App Tracking Transparency policy requires all iOS apps to ask users for permission to share data. This change has critically impacted the $189 billion digital advertising industry worldwide.[4]

Need for Personalized and Flexible Framework in Privacy Preserving Data Publishing. Situated at the center of the two-sided market, with consumers on one side as data generators and advertisers on the other as data users, a data aggregator has both a responsibility and vested interest in preserving consumer privacy while maintaining data utilities to advertisers (Katsomallos1 et al. 2019). As apparent in the examples above, preserving privacy is key for consumers to willingly opt in and continuously offer their location data – the lifeline for the data aggregator's monetization and the foundation of the entire location ecosystem.[5] Meanwhile, maintaining data utilities is key for advertisers to accomplish targeting and other marketing objectives, and to continue purchasing location data from the data aggregator. On the contrary, failure to preserve consumer privacy could diminish consumers' willingness to provide location data, which would entail monetary and non-monetary losses to all entities in the location ecosystem (Pew 2018). Hence, it is in the data aggregator's best interest to balance the risks and utilities, and to share location data with advertisers while protecting consumer privacy.

More importantly, the data aggregator needs a ***personalized and flexible framework*** to balance *diverse* types of risks and utilities for heterogeneous consumers and advertisers, each with *individualized* needs for privacy protection and business applications (Primault et al. 2019, Li et al. 2020, Cunha et al. 2021). One, from the consumers' perspective, *diverse types of privacy risks* (or threats hereafter), hence diverse consumer needs for privacy protection, frequently arise, such as home inference risk, or re-identification of a consumer and his/her sensitive activities (Li et al. 2020, Jiang et al. 2022). Even for the same type of privacy risk, each consumer's *risk level is heterogeneous* due to heterogeneous mobility patterns. Therefore, a consumer with a higher privacy risk, such as being the only individual in the data who has visited an abortion clinic, needs stronger protection.

[4] `http://bit.ly/3Zem7VL`; `https://bit.ly/3ZhvIev`.

[5] `https://gtnr.it/3IHPPvr`.

Two, from the advertisers' perspective, *diverse types of data utilities*, hence diverse advertiser needs, prevail in business applications, such as geo-targeting versus personalized recommendations (Yang et al. 2018, Ghose et al. 2019). Thus, the framework needs to accommodate a wide range of data utilities and advertisers' use cases. Finally, *diverse acceptable levels of risks* to consumers or consumer advocacy groups[6], including zero risk, and similarly *diverse acceptable levels of utilities* to advertisers, are commonplace. For instance, some advertisers demand a high utility from the data for finessed geo-targeting, while others an intermediate-to-high utility for coarser consumer segmentation. Hence, the academic literature and location industry imperatively covet not just a framework that balances a single risk and a single utility, but a ***personalized and flexible framework*** for the data aggregator to agilely fulfill and balance *individualized* and *diverse* needs from both the consumer side and advertiser side, thus sustaining a healthy, secure, and mutually beneficial location data ecosystem. We aim to fulfill this urgent need.

Research Objectives and Findings. In particular, while the emerging literature on privacy preserving data publishing (PPDP hereafter) of location data has focused on a single privacy risk or non-business utility (e.g., similarities between the original and published trajectories) at a time, and also global data obfuscation at the sample- instead of individual-level, we develop a novel framework that performs *personalized* obfuscation while *flexibly* accommodating different types, and different acceptable levels, of risks and business utilities. Specifically, we take a data aggregator's perspective and address the following core research questions:

1. *How can we quantify privacy risk to a consumer (personalized risk quantification)?*
2. *How can we quantify data utility to an advertiser (personalized utility quantification)?*
3. *Most importantly, how can we personalize data obfuscation to accommodate diverse types of risks, utilities, and acceptable levels of risk-utility trade-off (personalized and flexible obfuscation)?*

To accomplish the above, we develop a machine learning-based framework with three components: (1) quantification of each consumer's privacy risk; (2) quantification of an advertiser's utility; and (3) design of a ***personalized and flexible obfuscation scheme*** – our key contribution. The core idea of this scheme is to suppress a subset of locations visited by a consumer based on his/her personalized suppression parameter proportional to his/her risk level, and to further leverage flexible structured grid search by varying a grid parameter to accommodate different types and different acceptable levels of risks and utilities (such as a 1% reduction in utility upon data suppression with a 10% reduction in risk). We specifically illustrate the agility of the framework with two most prevalent types of privacy risks: home inference from the published location data and re-identification of a consumer and his/her entire trajectory from a subset of locations known

[6] `https://bit.ly/3EQWTVo`.

a priori to a stalker, and two essential types of advertiser utilities: next location prediction and activity-timing prediction. We further validate the proposed framework on one million trajectories generated by 40,000 consumers over five weeks from a major U.S. metropolitan area.

A few key findings arise. First, an absence of obfuscation, i.e., no steps taken to preserve consumer privacy, indeed entails high privacy risks to consumers. For instance, a stalker may infer a consumer's home location with 0.84 normalized haversine distance (varying between 0 and 1). This means that the home location can be inferred within a radius of 2.5 miles, indicating a high home inference risk. A stalker may also identify a consumer and his/her entire trajectory with 49% success by knowing a priori merely two randomly sampled locations visited by the consumer. Second, location data offer great value to an advertiser, who may predict a consumer's next location with 25% success, and next activity (such as dining or fitness) and its timing (such as Saturday morning) with 26% success. Finally, a data aggregator can effectively curtail a potential invasion of consumer privacy by performing personalized data obfuscation without sacrificing the utility of the obfuscated data to an advertiser. The data aggregator may also fulfill personalized and diverse demands from both the consumer side and advertiser side by flexibly accommodating multiple types of risks and utilities, as well as a wide array of acceptable levels of a specific risk, utility, and risk-utility trade-off. Our extensive benchmark comparisons and robustness checks also confirm the superior and robust performance of the proposed *personalized and flexible* framework.

Key Contributions. In summary, this study contributes to a core research domain across multiple disciplines – privacy preserving data publishing (PPDP) – by designing a ***personalized and flexible PPDP framework for location data***. It harnesses the power of the latest form of unstructured big data with rich business applications, while simultaneously preserving consumer privacy. Consumer privacy and business utility stand as two essential pillars sustaining the long-term health and self-governance (without solely relying on regulatory interventions) of the location data ecosystem. The proposed ***personalized*** data obfuscation, facilitated by personalized risk quantification, mitigates each consumer's distinct privacy risk. The ***flexible*** accommodation of multiple types of risks and utilities, and multiple acceptable levels of a risk, utility, and risk-utility trade-off (i.e., objective function), allows a data aggregator to fulfill the diverse needs of risk protection and utility preservation from both the data provider and data user sides, with even multiple trade-off solutions to each specific type of risk and utility. With skyrocketing business applications of novel location big data, the proposed personalized and flexible PPDP framework hence fills a critical void in a multi-disciplinary literature on PPDP of location data. The framework also offers a direly needed solution to the most quintessential challenge confronting the multi-billion-dollar location data and location-based marketing industries – a balance between privacy risks and data utilities.

2. Literature Review

With the rising needs for inter-organizational sharing and offline mining of consumer data (e.g., between healthcare providers and CDC, Netflix Prize and academics, location-based services like Google Map and location-based marketers), a multi-disciplinary literature on PPDP is rapidly emerging (Fiore et al. 2020). For instance, the Information Systems and Marketing literature has designed PPDP methods for numeric and categorical structured data (Li and Sarker 2011, Li and Sarkar 2013, 2014, Menon and Sarkar 2016, Chen et al. 2022) and more recently unstructured data, such as text (Li and Qin 2017) and image data (Zhou et al. 2020). As different data types require different PPDP methods (Cunha et al. 2021), our research extends this small yet growing business literature on PPDP into one of the latest form of unstructured data, consumer location data.

Meanwhile, the literature from Computer Science and related disciplines has developed a variety of PPDP methods specifically for location data, such as vehicle movements (Abul et al. 2008, Yarovoy et al. 2009), social media check-ins (Terrovitis et al. 2017, Yang et al. 2018), and simulated location data (Abul et al. 2008, Yarovoy et al. 2009, Chen et al. 2013). Nonetheless, large-scale business applications on real-world location data remain sparse (Primault et al. 2019). More importantly, the methods thus far primarily (a) preserve global (sample-level) instead of *personalized* (individual-level) privacy in location data; and (b) focus on a single risk (most commonly re-identification risk) and non-business utilities (e.g., the distance between the original and published trajectories). They are hence not *personalized* or *flexible* to accommodate heterogeneous types of privacy risks, business utilities, or acceptable levels of risks and utilities quintessential in business applications, such as location-based marketing. Below we will zoom into this literature of PPDP of location data and highlight the key contribution of our research as offering a *personalized* and *flexible* PPDP framework for location data in *business* applications. This literature falls under two mainstreams: *differential privacy models* and *syntactic models*, to which our framework belongs.

Differential Privacy Models. Based on ϵ-differential privacy, these models aim at reaching the same inference regardless whether a focal individual is included in the data or not (Dwork and Lei 2009). They demand zero risk, hence offering stronger privacy protection. Nonetheless, this limits their applications and data utilities to primarily data queries, instead of broad business applications where non-zero risks are prevalent. For instance, consumers often have to share their location data, and as a result tolerate non-zero privacy risks, in exchange for location-based services, such as map navigation, food delivery, and POI recommendations. Moreover, the specific obfuscation techniques used to accomplish stronger privacy protection in differential privacy models, such as perturbation (adding noise to data), do not preserve data truthfulness required by many downstream business applications (Terrovitis et al. 2017, Fiore et al. 2020, Jin et al. 2022). For instance,

after perturbation of a location data set, it becomes challenging to answer even simple questions, such as how many customers visited my restaurant today.[7]

Syntactic Models. These models permit non-zero risks, hence well-versed for business applications (Jin et al. 2022). Moreover, suppression (of a subset of data) commonly used in syntactic models preserves the truthfulness of the remaining unsuppressed and subsequently published data (Chen et al. 2013, Terrovitis et al. 2017), making syntactic models a top choice for a wide range of downstream business tasks that demand data truthfulness (Jin et al. 2022). We hence take the route of syntactic models when designing our framework, and focus on reviewing the syntactic models for location data below. Syntactic models ensure k-anonymity, that is, at least k records share the same value of an attribute (e.g., visiting the same location). Then (k, δ) anonymity is proposed to offer stronger privacy preservation (Abul et al. 2008, Yarovoy et al. 2009). For instance Abul et al. (2008) perform space generalization on location data and transform the trajectories so that k of them lie in a cylinder of radius δ. Variants of k-anonymity and (k, δ) anonymity further relax the assumptions of the earlier methods (Chow and Mokbel 2011, Huo et al. 2012, Hwang et al. 2013, Gao et al. 2014, Brauer et al. 2022).

While importantly advancing PPDP of location data, these methods originating from Computer Science and related disciplines are not particularly apt at business applications (Primault et al. 2019, Jiang et al. 2022, Jin et al. 2022). Specifically, business use cases of location data demand a number of important qualities in the PPDP methods. First, with the well established business philosophy of *personalization* (Tong et al. 2020, Chandra et al. 2022), PPDP of location data requires *personalized* risk quantification and acceptable level of risk for heterogeneous consumers, *personalized* utility and acceptable level of utility for heterogeneous advertisers and use cases, and *personalized* data obfuscation (Li et al. 2020, Cunha et al. 2021, Jiang et al. 2022). Personalization also enhances transparency and interpretability of PPDP in business applications (e.g., which locations are obfuscated for which consumers and why). Second, the rapid emergence of diverse types of privacy risks and business utilities, such as with the rise of 5G, blockchain, and metaverse, calls for *flexible* frameworks that can accommodate many types of risks, utilities, and acceptable levels of a specific risk, utility, and risk-utility trade-off (Primault et al. 2019, Fiore et al. 2020, Cunha et al. 2021, Jiang et al. 2022). Finally, business applications demand PPDP that maintains data truthfulness for downstream tasks, such as geo-targeting. In contrast, the above syntactic models hide a user inside a crowd, hence not *personalized* (Primault et al. 2019). They are also

[7] Besides these two mainstreams, ad hoc methods such as *mix zone* and *dummy* (adding dummy users) (Primault et al. 2019, Jin et al. 2022, Jiang et al. 2022) are also developed. Nonetheless, similar to differential privacy models, these methods often do not preserve data truthfulness well, nor do they offer sufficient privacy protection. Therefore, they are not pertinent to business applications, or the focus of our review.

designed to tackle a single risk, hence not *flexible* to accommodate diverse risks (Jin et al. 2022). They are also tested only on simulated or small data (Primault et al. 2019). These important gaps are also accentuated by the latest surveys of this literature (Primault et al. 2019, Fiore et al. 2020, Li et al. 2020, Cunha et al. 2021, Jin et al. 2022, Jiang et al. 2022).

As a result, newer methods have been developed, aiming to incorporate either personalized risks or alternative risks beyond re-identification. Specifically, as the literature starts to recognize the significance of fulfilling personalized risk requirements, developing *personalized* PPDP methods for location data has become an active area of research in recent years (Gao et al. 2014, Komishani et al. 2016, Qiu et al. 2021, Mahdavifar et al. 2022). Nonetheless, these methods (a) focus on personalized risks, without considerations of personalized utilities across diverse data users; (b) examine only non-business utilities (Jin et al. 2022), such as information loss, query error (Komishani et al. 2016), correlation (Gao et al. 2014) or similarity between original and published trajectories (Qiu et al. 2021); (c) address merely one or a small number of privacy risks, and cannot flexibly accommodate heterogeneous types of risks (or utilities); and (d) publish trajectories that do not well preserve the truthfulness of the original trajectories for downstream business applications (Mahdavifar et al. 2022).

The literature is also advancing along the *flexibility* dimension. For example, methods such as t-closeness (Li et al. 2007b) and ℓ-diversity have been proposed to go beyond the *re-identification risk* to address the *sensitive attribute inference* (Pelekis et al. 2011), such that sensitive locations are well represented for each consumer (Machanavajjhala et al. 2006), or a stalker's confidence in inferring a sensitive attribute is limited to a threshold (Wang et al. 2007). Frameworks accommodating more than one risk also emerge (Komishani et al. 2016, Yao et al. 2021). However, these studies remain incapable of flexibly accommodating a wide range of risks and business utilities, and are not personalized. Our research thus makes distinct contributions to this literature as fulfilling *personalization* and *flexibility*, both widely recognized by the literature as important future research directions.

Synthetic Trajectory Models. Besides the above two mainstream methods, synthetic trajectory approaches that generate *dummy* trajectories using a trained generative model start to show promise in preserving consumer privacy in location data. These approaches typically involve two phases: training and generation. For instance, PPMTF (Murakami et al. 2019) aims to preserve the population-level distribution of the visits and transition matrix across POIs using factor matrices trained via posterior sampling. Trajectories are then generated from the reconstructed tensors. LSTM-TrajGAN (Rao et al. 2020) sets up a generative adversarial task with a loss metric that quantifies the similarities between the trajectories. The trained GAN is then used to generate the synthetic trajectories of a consumer. Conceptually, these models are trained with a set of consumer

trajectories to generate synthetic trajectories to be shared with an advertiser. Since the generators are trained solely to preserve the statistical properties of the original trajectories, these methods do not emphasize any specific utility, and hence may not fulfill every advertiser's needs. Moreover, the published *dummy* trajectories do not preserve the truthfulness of the data.

Table 1 summarizes the above non-exhaustive, yet most representative, PPDP methods for location data from Computer Science and related disciplines. Interested readers may further refer to the latest surveys for a more comprehensive coverage of this literature (Katsomallos1 et al. 2019, Primault et al. 2019, Fiore et al. 2020, Li et al. 2020, Cunha et al. 2021, Jin et al. 2022, Jiang et al. 2022). Table 1 further highlights our two key contributions – ***personalization*** and ***flexibility***. First, while the existing literature largely performs obfuscation at the group- or sample-level, the proposed framework administers ***personalized*** obfuscation, facilitated by the individual-level risk quantification. Such personalization nimbly addresses each consumer's heterogeneous level of risk given each distinct type of privacy risk, as well as heterogeneous privacy preference and risk tolerance. Personalization also importantly grants transparency and interpretability to data aggregators (e.g., on which features escalate the risk for which individual) and offers managerial insights to advertisers (Primault et al. 2019). Moreover, both the utility and risk-utility trade-off in our framework are also personalized, fulfilling diverse advertiser and data aggregator needs.

A second critical benefit of the proposed framework arises from its *flexibility*. In contrast to the existing methods that are largely tailored to tackle a single risk or a single (and also non-business) utility, the proposed framework ***flexibly*** accommodates multiple types of risks and utilities (particularly business utilities), and also multiple acceptable levels of each risk, utility, and risk-utility trade-off. As we will demonstrate in detail next, it accomplishes so by leveraging a generalizable objective function and also a structured grid search, both of which are not tied to any specific risk, utility, or their functional forms. In contrast, prior studies resort to an optimization function linked to a specific functional form of a specific risk or utility. Therefore, these methods need to modify the optimization function for alternative risks or utilities case by case, also without a guarantee of an optimal solution that often depends on the specific functional forms of the risk and utility. Moreover, structured grid search permits and produces multiple trade-offs of each type of risk and type of utility (corresponding to different values of the grid parameter), such as 1% decrease in the utility for 10% decrease in the risk, or 2% decrease in the utility for 15% decrease in the risk. Therefore, it is capable of fulfilling a multitude of acceptable levels of a specific risk, utility, and risk-utility trade-off. Overall, such flexibility empowers a data aggregator to satisfy diverse demands of risk mitigation from the consumer side and simultaneously utility conservation from the advertiser side. As discussed in the Introduction, such flexibility is critical for business applications.

	Risk	Utility	Data
Syntactic Models:			
Abul et al. (2008)	(k, δ) anonymity	Deviation from true trajectories	Simulated data
Yarovoy et al. (2009)	k-anonymity	Information loss	Car trajectories
Chow and Mokbel (2011)	k-anonymity	Deviation from true trajectories	NA
Pelekis et al. (2011)	Sensitive attribute	Streaming KNN queries	Simulated data
Huo et al. (2012)	(k, δ) anonymity	Information loss	155 consumers' trajectories
Hwang et al. (2013)	r-anonymity	Number of consumers in a region	Taxi trajectories
Chen et al. (2013)	Re-identification risk Sensitive attribute inference	Frequent sequences	Simulated data
Gao et al. (2014)	k-anonymity	Information loss	Simulated data
Komishani et al. (2016)	Sensitive attribute	Information loss	Simulated data
Terrovitis et al. (2017)	Re-identification risk	Frequent sequences	Social network trajectories
Qiu et al. (2021)	Re-identification risk, sensitive attribute	original-obfuscated trajectory similarities	182 consumers' trajectories
Yao et al. (2021)	l-diversity enhanced	Information loss	Simulated data
Mahdavifar et al. (2022)	Re-identification risk	original-obfuscated trajectory similarities	Simulated data
Brauer et al. (2022)	k site-unidentifiability	Jensen-Shannon Distance	182 consumers' trajectories
Synthetic Trajectories:			
Murakami et al. (2019)	NA	POI distribution	Foursquare
Rao et al. (2020)	NA	Trajectory similarity	Foursquare
Proposed Framework	**Re-identification risk** **Home inference risk** **Flexible to accommodate other risks**	**Location prediction** **Activity-timing prediction** **Flexible to accommodate other utilities**	**1 million consumer trajectories**

Table 1 Comparison of proposed framework with syntactic and synthetic trajectory models

3. Data

We partner with a leading data aggregator that integrates the location data across 400+ commonly used mobile apps, such as news, weather, map, and fitness, from one-quarter of the U.S. population in compliance with privacy regulations such as GDPR and CCPA[8]. The data are representative of the U.S. population given the company's detailed research. The sample under analysis covers a major U.S. metropolitan region. Figure 1 displays the region's map (purposefully blurred to preserve privacy) and an example of an individual's footprints with 732 unique locations visited during the five-week sampling period between September and October of 2018. The entire sample includes 940,000 locations from 40,012 consumers. Each data record corresponds to a location visited by an individual and contains an anonymized device ID, mobile operating system (Android or iOS), and the timestamp, latitude, longitude, and dwell time at the location. Table 2 displays the summary statistics of the sample. Each individual's location is recorded every five to fifteen minutes, or when the geo-coordinates (longitude/latitude) change substantially, to reduce device battery drainage and data redundancy.

Prior studies have tested their methodologies on simulated data, vehicle movements, or social media check-ins over a small sample and a short period such as 24 hours. We make an initial effort to validate the proposed framework on the newly available, population-scale, individual-level mobile location data. These data are automatically generated in real time via GPS, Wi-Fi, Bluetooth, and Beacon multi-technology multi-lateration with an accuracy radius of 20 meters. They are hence

[8] These apps install a proprietary SDK designed by the company, which tracks location data (via GPS, Wi-Fi, Bluetooth, and Beacon multi-technology multi-lateration) more accurately than other sources, including individual app owners. This SDK also greatly reduces battery drainage in mobile devices when location data are being tracked in real time. Interested readers may refer to https://www.tamoco.com/blog/location-data-info-faq-guide/ for further details. Consumers are always asked to double opt-in to the specific app in use and to the data aggregator's SDK for location tracking and data sharing.

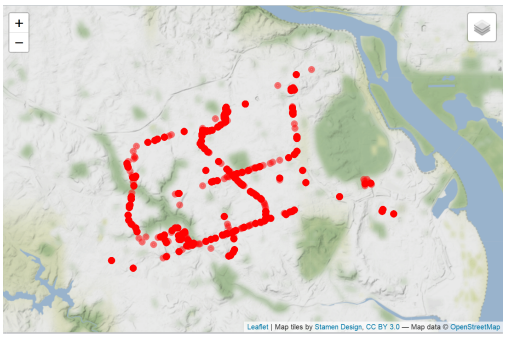


Figure 1 An example consumer's footprints (732 unique locations over five weeks)

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 (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 (km)	1.89 (3.89)	0.02 (75.49)

Table 2 Summary statistics

Concept	Definition	Model	Measurement	Findings
Risk(1): Home inference	Stalker infers a consumer's home location from shared data	Random Forest, LSTM	Normalized haversine distance between predicted and actual home locations (Section 4.1.1)	Average home inference risk is 0.84 (Section 5.1)
Risk(2): Re-identification	Stalker knowing a consumer's subset of locations a priori identifies the consumer (and his/her entire trajectory) from shared data	Random Forest	Max. prob. of identifying a consumer over all subsets of locations known a priori, then averaged across consumers (Section 4.1.2)	Average re-identification risk is 0.49 by knowing a consumer's two random locations a priori (Section 5.1)
Utility (1): Location prediction	Advertiser predicts a consumer's next k locations from shared data	Nearest Neighbor	MAP@k, MAR@k (Section 4.2)	Next location is predicted with 25% accuracy (Section 5.2)
Utility (2): Activity-timing prediction	Advertiser predicts a consumer's next k activities and activity timing from shared data	LSTM	MAP@k, MAR@k (Section 4.2.2)	Next activity and timing is predicted with 26% accuracy (Section 4.2.2)

Table 3 Overview of key concepts, definitions, measurements, and findings

much more precise than cell tower tracking with an accuracy radius in kilometers, social media geo-tags known for sparsity and inaccuracy, or social media check-ins that rely on consumers' self-reports. Also, unlike taxi or public transportation data that only capture the consumers using these transportation modes, our data are representative of the U.S. population and everyday consumer behavior, hence more valuable to advertisers.

4. Methodology

The framework consists of three core components: quantification of each consumer's privacy risk, quantification of the advertiser's utility, and personalized and flexible obfuscation scheme for the data aggregator. When describing each component below, we will formulate the problem and then propose the solution. We will empirically evaluate the solution in the subsequent Section 5. Table 3 summarizes all key definitions, measurements, models, and findings discussed below.

4.1. Quantification of Consumer's Privacy Risk

The first step of the proposed framework is the quantification of each consumer i's privacy risk r_i by simulating a stalker's adversarial action on the published trajectories $\mathcal{P}(T)$. While the framework can accommodate different privacy risks, we will illustrate two specific risks of vital concern to consumers: home inference risk and re-identification risk. That is, as detailed below, the stalker would infer the consumer's home, or re-identify the consumer and hence his/her entire trajectory T_i with the background knowledge of a subset of his/her locations $\overline{T_i} \in T_i$. In this process, the stalker could leverage either simple heuristics, such as trajectory queries as illustrated in the re-identification risk, or robust machine learning models as illustrated in both the home inference risk and re-identification risk (Li et al. 2007a, Yang et al. 2018).

4.1.1. Home Inference Risk

Definition: Home inference risk refers to a stalker's inference of a consumer's home location from the published data (Li et al. 2007a, Tucker 2013, Gardete and Bart 2018, Rafieian and Yoganarasimhan 2021). Once a consumer's home location is identified, then his/her identity, including name and potentially other personal information, is readily revealed, as many public databases, such as voter registration data or real estate data, associate an individual's name with his/her home location. Even a simple Google search of a home address returns the name and other personal information of the resident at the location. Such a connection between a consumer's identity and his/her entire trajectory history via home inference hence entails privacy risks to the consumer, such as the identification of sensitive locations visited, or excessive mail advertising.

Measurement: A consumer i's home inference risk r_i is quantified by a stalker's predictive accuracy of a consumer's home location from the published trajectories $\mathcal{P}(T)$ using a machine learning model. This predictive accuracy, as detailed next, is specifically measured by (a) the haversine distance between the predicted and actual home locations, and (b) a normalized haversine distance, which lies between 0 (low risk) and 1 (high risk).

Model: As shown next, we first extract a set of mobility features from the location data. Then using these features, we train an ensemble (Sagi and Rokach 2018) of two Random Forest (RF hereafter) regressors (Breiman 2001) to predict a consumer's home latitude and longitude. Finally, we use the resulting predictive accuracy described in the above **Measurement** to quantify a consumer's home inference risk.

Trajectory Feature Extraction. We first mimick a stalker's adversarial action of extracting trajectory features $\mathcal{F}(T)$[9] from the published trajectories prior to inferring a consumer's home location.

[9] To simplify the notation, we use $\mathcal{F}(T)$ to refer to $\mathcal{F}(\mathcal{P}(T))$. Since both $\mathcal{P}(T)$ and T are a set of trajectories, any operation (such as $\mathcal{F}$ here) performed on T is also applicable to $\mathcal{P}(T)$.

These features include the consumer mobility features commonly used in the literature (Pellungrini et al. 2018), as well as the richer features that we further incorporate, consumer-location and consumer-consumer affinities. Such feature extraction, also known as feature engineering, is routinely performed prior to a prediction task in the machine learning literature, since the extracted features capture important high-level patterns in the data (e.g., visit frequency and visit duration) and thus offer more predictive power than the noisier raw data (e.g., latitude/longitude) (Zheng et al. 2010, Wang et al. 2011, Williams et al. 2015, Pappalardo et al. 2016). This is to offer the stalker an advantage and allow us to estimate the worst-scenario home inference risk. We further illustrate the advantage of using the extracted features in Online Appendix D[10]. Note that these extracted features are also subsequently used in other tasks within the proposed framework, such as quantification of the re-identification risk, interpretation of each feature's contribution to a specific risk, quantification of the data's utility to the advertiser, computation of each location's suppression probability, and estimation of the benchmark models to be compared with the proposed framework.

The set of *consumer mobility features*, summarized in Table 4, captures a consumer's basic 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), as well as richer mobility patterns, such as the entropy (Eagle and Pentland 2009) and radius of gyration (Gonzalez et al. 2008). In addition, adapting from the literature on identifying significant locations to predict consumer mobility (Ashbrook and Starner 2003, Zheng et al. 2010), we build three *consumer-location affinity* tensors: a consumer's weekly frequency to, time spent at, and total distance traveled from the immediate prior location to each location. Each tensor is of order three: consumer by unique location by week. We then extract the consumer-specific lower dimensional representations by performing a higher order singular value decomposition (HOSVD) on the three tensors separately (De Lathauwer et al. 2000). Finally, prior studies have also predicted consumer network or social links based on trajectories (Wang et al. 2011). We thus quantify the consumers' co-locations by building the *consumer-consumer affinity* tensors based on the locations shared among consumers at a weekly level. Each tensor is of order three: consumer by consumer by week. We populate three such tensors with respectively the weekly average frequency to, total time spent at, and distance traveled to each co-visited location, before performing a HOSVD on each tensor to extract the consumer-specific lower dimensional representations indicative of the consumer-consumer affinity.

We illustrate the above consumer-location and consumer-consumer affinity features using a stylized example here. Consider three consumer trajectories: $T_1 = \{(A,1),(B,1),(A,2),(A,2)\}$, $T_2 =$

[10] All Online Appendices are available at `https://bit.ly/privacy_2022`.

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} = geographical coordinates l^i_{cm} = center of mass of the consumer w_{ij} = 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 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, and distance

Table 4 Consumer mobility features

$\{(C,1),(A,1),(A,1)\}$, and $T_3 = \{(D,1),(B,1),(C,2)\}$, where A, B, C, D are location identifiers, and 1 and 2 are week identifiers. 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 the weeks and 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, $(A,1)$ is 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 each tensor and use the first five principal components that capture a majority (92%) 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 to orchestrate privacy threats.

Home Inference Risk. For the home location prediction, we first transform the raw latitude/longitude of each location into the Universal Transverse Mercator (UTM) coordinates. Since the latitude/longitude coordinate system uses angular measurements to describe a position on the surface of the earth, the distance covered by one degree of longitude differs when moving from the equator

to the poles[11]. The UTM coordinate system, in contrast, divides the world into 60 zones, each of 6 degrees of longitude wide, hence offering a constant distance anywhere on a map.[12] An example of the UTM-transformed coordinates for the latitude/longitude pair (38.969210, -77.105650) is UTM Easting/Northing of (317580.40, 4315468.41) in meters.

We then use an ensemble of two Random Forest (RF) models, one to predict the UTM-transformed latitude and the other to predict the UTM-transformed longitude of a consumer's home location. The two models are concatenated and jointly trained using the machine learning training pipeline in Python's sklearn package[13]. The loss used to jointly train these two models is the haversine distance (or equivalently Euclidean distance as a result of the UTM transformation) between the predicted home and actual home of a consumer[14]. That is, the dependent variables are the UTM-transformed latitude and longitude, and the independent variables are the consumer mobility features, consumer-location affinity, and consumer-consumer affinity.

Then consumer i's home inference risk is measured by the resulting predictive accuracy, i.e., the haversine distance between i's predicted home and actual home (e.g., 500 meters), denoted as h_i. h_i is continuous, and a *smaller* h_i indicates a *higher* home inference risk. Therefore, we further calculate a relative measure of the home inference risk r_i by using the normalized haversine distance: $r_i = \frac{max(h_i)-h_i}{max(h_i)-min(h_i)}$, where $max(h_i)$ and $min(h_i)$ respectively represent the maximum and minimum haversine distance among all consumers under analysis. This normalization is similar to the widely used min-max normalization. Given $max(h_i)$ and $min(h_i)$, a smaller h_i will produce a higher r_i. Hence, a consumer i's home inference risk r_i now lies between 0 and 1, where a *lower* r_i indicates a *lower* risk (e.g., $r_i = 0$ when $h_i = max(h_i)$) and a *higher* r_i indicates a *higher* risk ($r_i = 1$ when $h_i = min(h_i)$). We subsequently focus on this normalized, instead of raw, haversine distance between the predicted and actual homes as the focal measure of the home inference risk.[15]

We choose RF as the stalker model for several reasons. One, it can be used for either regression (predicting a home's UTM-transformed latitude/longitude) or classification (predicting if a location visited is a consumer's home = yes/no), hence offering greater flexibility. Two, it merges a forest of decision trees, each of which might be a weaker predictor, to form an overall much more

[11] Interested readers may view the definitions of latitude and longitude at `https://www.maptools.com/tutorials/lat_lon/definitions`.

[12] Interested readers may view the definition of UTM and its advantages over raw latitude/longitude at `https://bit.ly/3YcK8LD` and `https://www.maptools.com/tutorials/utm/why_use_utm`.

[13] `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.StackingRegressor.html`.

[14] Interested readers may view the history and formula of haversine distance at `https://en.wikipedia.org/wiki/Haversine_formula` and Euclidean distance at `https://en.wikipedia.org/wiki/Euclidean_distance`.

[15] Note well that another way to define risk would be $r_i = 1 - \frac{h_i - min(h_i)}{max(h_i)-min(h_i)}$ which captures the prediction accuracy of the stalker, lower the accuracy, lower the risk.

accurate model. It is hence widely used across disciplines in many business- or non-business-related applications. Albeit context-specific, RF and its variants have demonstrated superior predictive performance relative to other commonly used bagged or boosted machine learning models, such as XGBoost (Chen and Guestrin 2016) or AdaBoost (Hastie et al. 2009) across various contexts (Said and Mouazen 2017, Jhaveri et al. 2019, Huang et al. 2020, Bhakta et al. 2021). Our choice of RF as the stalker model hence gives the stalker an advantage and allows us to estimate the worst-scenario privacy risk. Three, RF holds other advantages, such as minimal requirement for data pre-processing (e.g. no need for data re-scaling or transformation), computationally inexpensive for big data or data with high dimensionality since it can parallelize subsets of trees, robustness against noise, outliers, or non-linearity, automatic handling of missing value or unbalanced data, lower risk of over-fitting, easier to tune due to very few hyper-parameters, and simpler to implement and visualize than XGBoost or AdaBoost. Four, compared to deep learning models that boast high predictive accuracy, such as Long Short Term Memory (LSTM) (Hochreiter and Schmidhuber 1997), RF is easier to implement, computationally faster, less prone to over-fitting, similarly or even more accurate in many settings (in our study too, cf. Online Appendix D), and more interpretable with feature importance (also illustrated Section 5.1) or most representative trees, hence more actionable to decision makers (e.g., data aggregators) (Fernández-Delgado et al. 2014, Ahmad et al. 2017, Weinberg and Last 2019, Chen et al. 2019, Bhakta et al. 2021). Nonetheless, our framework is flexible: one may freely implement alternative methods other than RF to quantify the home inference risk, or to quantify the home inference risk with alternative measures. For instance, we demonstrate in the Online Appendix D the use of an alternative model, LSTM, to quantify the home inference risk, as well as an alternative measure of the home inference risk, the predictive accuracy (MAP@1, MAR@1) of a binary classification (yes/no) of whether a location visited is a consumer's home.

4.1.2. Re-identification Risk

Definition: Consistent with the literature, we define the re-identification risk as the re-identification of a consumer and hence his/her entire trajectory, including sensitive locations visited, from the published location data, by a stalker with background knowledge of a subset of his/her locations (i.e., partial trajectory), such as home, work, or a store visited, that is already linked to the consumer's identity (Fiore et al. 2020, Jin et al. 2022).

A stalker could access a consumer's partial trajectories from a wide range of internal and external sources (Fiore et al. 2020, Jiang et al. 2022). Internally, a stalker (an advertiser or a retail brand) commonly hosts its own customer databases, such as customer purchase and shipping records, catalog mailing lists, loyalty program registration forms, customer surveys, or customer service

records, all containing each customer's home or work address, or both, as well as potentially other location information. Externally, the stalker could easily access numerous free consumer self-reports or public databases, such as social media (Meta, LinkedIn, Twitter) accounts, social media check-ins, voter registrations[16], and large numbers of other public records, such as marriage registrations and divorce records[17]. The stalker may also purchase, with low costs, from a large number of vendors the commercial databases containing millions of consumers' names, home and work addresses, and other location histories. Examples of such databases include consumer or resident mailing lists[18], background checks, and commercial people search tools[19]. Lastly, the widespread data breaches, 7500+ incidences annually worldwide[20], have also granted stalkers access to massive volumes of individual-level partial trajectories, such as breached employee databases[21], credit card records of visited stores[22], hotel stays[23], and taxi rides[24]

Measurement: The re-identification risk r_i of a consumer i is measured by the maximum probability of identifying i and thus his/her entire trajectory from the published trajectories $\mathcal{P}(T)$, among all possible subsets of i's locations known a priori to the stalker $\bar{T}_i \subseteq T_i$.[25]

Model: Without knowing the subset of locations $\bar{T}_i$ a priori, the data aggregator needs 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 consumer i. For each such subset, the probability of i 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 consumers exist other than i, then the probability of identifying i would be 1 for the subset considered. Ultimately, a consumer i's re-identification risk is the maximum of the probabilities of identifying i over all subsets of locations known a priori to the stalker. We further employ a Random Forest classifier with a speed-up heuristic to reduce the computational complexity (Online Appendix C).

Here is a stylized example. Three consumers' trajectories over a two-week period, $T_1 = \{(A,1),(B,1),(C,2),(C,2)\}$, $T_2 = \{(A,1),(B,1),(A,2)\}$, and $T_3 = \{(A,1),(B,1),(C,2)\}$, suggest

[16] `http://bit.ly/3J7rWyD`.

[17] `https://bit.ly/3YgLogI` and `http://bit.ly/3Zv7CfW`.

[18] Here are a few examples: `http://bit.ly/3IOgV3Y`; `http://bit.ly/3YvGKff`; and `http://bit.ly/3IJhtYP`.

[19] `https://www.supereasy.com/how-to-find-out-where-someone-works/`.

[20] `http://bit.ly/41GjZYt`; and `https://bit.ly/3Zc8HcW`.

[21] `https://www.theverge.com/2022/5/27/23144418/hacker-verizon-employee-database`.

[22] `https://usa.visa.com/content/dam/VCOM/download/merchants/visa-merchant-data-standards-manual.pdf`.

[23] `https://techcrunch.com/2022/07/06/marriott-breach-again/`.

[24] `http://bit.ly/3kFDDmS`.

[25] The assumption here is that a stalker obtains the location data from a data aggregator only once, and also does not collude with other stalkers to combine data. We invite future research on these interesting yet more complex scenarios.

that all three consumers visited the location subset (A, B). Two of them (consumers 1 and 3) visited (B, C), and two (consumers 1 and 3) visited (A, C). Then given each of these location subsets, the corresponding probabilities of identifying consumer 1 are $\{\frac{1}{3}, \frac{1}{2}, \frac{1}{2}\}$, resulting in consumer 1's re-identification risk as $max(\frac{1}{3}, \frac{1}{2}, \frac{1}{2}) = \frac{1}{2}$. Given a similar number of unique locations visited across consumers, an individual who visits more unique locations not visited by others would then have a higher re-identification risk.

4.2. Quantification of Advertiser's Utility

The second step of the proposed framework is the quantification of an advertiser's utility when leveraging the published data for location-based marketing. While an advertiser's utility is multi-faceted, and the proposed framework is capable of flexibly accommodating different utilities, we demonstrate two most common and essential types of utilities: location prediction, which focuses on the utility related to spatial factors, and activity-timing prediction, which further integrates temporal and contextual factors. We will define each below.

4.2.1. Location Prediction Utility

Definition: Location prediction utility pertains to an advertiser's utility of predicting a consumer's next k locations from the published data. This utility enables an advertiser to target a consumer with spatially relevant contents, such as POI recommendations, leading to higher revenues (Ghose et al. 2019). For instance, a chain restaurant may target a consumer with a discount coupon if being able to predict the consumer's visit of a location near one of its outlets. A gym may target a potential customer if others similar to this consumer have visited the gym often.

Measurement: This utility is measured by the advertiser's predictive accuracy, specifically the Mean Average Precision ($MAP@k$) and Mean Average Recall ($MAR@k$), of the next k locations from a collaborative filtering recommendation model, that is, by identifying other consumers with similar trajectories to infer the focal consumer's future locations (Bobadilla et al. 2011).

Model: We select the collaborative filtering recommendation model since consumers with similar historical trajectories are more likely to visit similar locations in the future (Bobadilla et al. 2011). Specifically, we focus on the best performing model, the nearest neighbor (NN) model, based on the comparison against the common recommendation models in the literature (Online Appendix D). This model identifies the m consumers most similar to the focal consumer i (i.e., m neighbors, denoted as M_i) and uses their locations to predict i's future locations. The similarity between any pair of consumers is computed as the cosine similarity between the features of the two consumers' trajectories $\mathcal{F}(T_i)$ and $\mathcal{F}(T_j)$:

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

Then we rank the unique locations visited by M_i based on visit frequencies and these m neighbors' similarities to consumer i. Specifically, for each consumer $j \in M_i$ and location $l \in T_j$, let f_j^l denote the visit frequency of consumer j to location l, then the rank of a location l for consumer j is determined by:

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

where $\frac{f_j^l}{\sum_l f_j^l}$ is the normalized visit frequency. Equation 2 ensures that consumer i will most likely visit the most frequently visited location by the most similar consumer. We further aggregate o_{ij}^l across all consumers in M_i who have visited location l 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 0 otherwise. A higher valued o_i^l suggests that consumer i will more likely visit location l in the future. The next k locations most likely visited by consumer i hence correspond to the top k 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 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 $L_i = \{l_i^1, l_i^2, \ldots, l_i^{k'}\}$ 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:

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

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

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

4.2.2. Activity-Timing Prediction Utility

Definition: The activity-timing prediction utility refers to an advertiser's utility of predicting a consumers' next k activities (such as commute, work, shopping, or dining) and the timing of these activities (time of the day, day of the week) from the shared data. These daily activities capture rich spatial, temporal, and contextual information about a consumer. Therefore, accurate predictions of these activities and the timing of these activities, such as when a consumer will most likely engage in leisure (instead of necessity) shopping, will allow an advertiser to deliver time-sensitive and context-aware marketing messages. Compared to the location prediction utility that focuses on spatial factors, this utility further incorporates temporal (sequential) and contextual dimensions of the data. It also makes predictions based on a consumer's own prior trajectories, instead of leveraging the nearest neighbors as in the location prediction utility.

Measurement: This utility is measured by the predictive accuracy, specifically $MAP@k$ and $MAR@k$, of a robust machine learning model that could further incorporate the temporal and contextual information, in addition to the spatial information.

Model: We leverage the sequential deep learning model, LSTM, to jointly predict each consumer's next activity and timing of the activity. We choose LSTM for a number of reasons. One, a type of recurrent neural network, LSTM is designed to decipher patterns entrenched in long sequences of data. It is hence particularly suited for data, such as ours, with time and sequence information, and is widely used in many applications, such as stock price prediction or traffic forecasting (Hochreiter and Schmidhuber 1997). Two, it is end-to-end, relying only on raw data without the need for feature engineering (Sarkar and De Bruyn 2021). Three, it can easily predict multiple outputs, such as time of the day, day of the week, and type of activity in our activity-timing prediction task, hence superior to alternate models that require joint training of multiple models, one for each output (Pascual and Bonafonte 2016, Zhou et al. 2019).

Before the prediction task, we query the Google Places API to identify the place type of each location (second column of Table 5), semantically group these place types into 14 activities that capture a consumer's daily activities, such as fitness and leisure shopping, and then transform each consumer's trajectory T_i into an activity trajectory A_i comprising these 14 activities (first column of Table 5). We then use LSTM to jointly predict the consumer's next k activities, time of the day (morning, afternoon, evening), and day of the week (weekend/weekday) of each activity:

1. Input Layer: The input triplet, (a, t, w), captures the state of a consumer. The activity a is a one-hot vector of length 14 corresponding to the 14 activities. The time of the day $t \in$ {Morning, Afternoon, Evening} is a vector of length 3. And the day of the week $w \in 0/1$ stands for {Weekend/Weekday}. The input sequences are concatenated into a vector of length 18 and fed into the next LSTM layer.

Activity	Place type
hospital	hospital, doctor
health	physiotherapist, pharmacy, dentist, drugstore
ncessityshopping	store, supermarket, convenience_store, home_goods_store, grocery_or_supermarket, hardware_store
fitness	gym
publictransport	transit_station, train_station, bus_station, light_rail_station, subway_station
owntransport	car_wash, car_repair, parking, gas_station, taxi_stand
religious	church, mosque, hindu_temple, synagogue
recreation	amusement_park, tourist_attraction, zoo, park, theatre, sports_stadium, concert, bowling_alley, art_gallery, aquarium, museum, movie_rental, book_store, library, movie_theater, campground
travel	hotel, lodging, rv
personalcare	beauty_salon, spa, hair_care
leisureshopping	clothing_store, department_store, shopping_mall, shoe_store, electronics_store, furniture_store
unhealthyactivities	casino, liquor_store, bar, night_club, cigarette
restaurant	restaurant, food, meal, bakery, cafe, meal_delivery, meal_takeaway
other	locations for which a place type was not identified by Places API

Table 5 Consumer activities

2. LSTM Layer: This layer has a hidden state to store the historical information and is carried forward to the subsequent time-steps.

3. Dropout Layer: The output of the LSTM layer is fed through a dropout layer to prevent the model from over-fitting on the training data by setting the activations of a certain percent of neurons (i.e., dropout rate) to zero.

4. Activated Dense Layer: The output of the dropout layer is fed into a dense layer which outputs a vector of length 18, representing the state triplet (a, t, w). We apply a SoftMax activation over the first 14 elements and the next three elements separately, representing the probability assigned to each activity and the time of the day of the activity. The last element has a rectified linear unit activation applied on it to represent the day of the week of the activity.

4.3. Personalized and Flexible Obfuscation Scheme for Data Aggregator

The third step of the proposed framework is the design of a personalized and flexible obfuscation scheme for the data aggregator.

Definition: The data aggregator's problem is to identify a transformation of each consumer i's original unobfuscated trajectory T_i into an obfuscated trajectory $\mathcal{P}(T_i)$, in order to balance the risk r_i and utility u_i before publishing $\mathcal{P}(T_i)$.

Acknowledging potential alternative specifications, we here introduce a simple, intuitive, yet flexible, and generalizable objective function. It offers a broad framing of the risk-utility trade-off as finding a grid parameter $p \in [0, 1]$ that governs which data are obfuscated to minimize the % decrease in the utility $\frac{(E^*(u_i)-E(u_i))}{E^*(u_i)}$ while maximizing the % decrease in the risk $\frac{(E^*(r_i)-E(r_i))}{E^*(r_i)}$:

$$O(p) = Min(\frac{E^*(u_i) - E(u_i)}{E^*(u_i)} - \frac{E^*(r_i) - E(r_i)}{E^*(r_i)}), \tag{6}$$

where $E^*(u_i)$ and $E^*(r_i)$ respectively denote the expected utility and risk computed a priori from i's original unobfuscated trajectory. $E(u_i)$ and $E(r_i)$, or more precisely $E(u_i(p))$ and $E(r_i(p))$, respectively refer to the expected utility and risk from i's obfuscated trajectory. Equation 6 is equivalent to

$$O(p) = Min(\frac{E(r_i)}{E^*(r_i)} - \frac{E(u_i)}{E^*(u_i)}. \tag{7}$$

This objective function is *flexible* to accommodate any type of personalized risk and utility. That is, the objective here is not minimizing a specific risk, but to reach an individualized acceptable trade-off between any type of risk and utility. In contrast, the objective of most prior studies is to minimize a single specific risk, and then evaluate the impact on a non-business utility post hoc (Yang et al. 2018, Primault et al. 2019, Fiore et al. 2020, Cunha et al. 2021, Jiang et al. 2022). As a result, each study needs to be modified for a different risk case by case, also without a guarantee of an optimal solution since such a solution depends on the functional form of the specific risk or utility[26].

Measurement: Given $E^*(u_i)$ and $E^*(r_i)$ computed a priori from the original unobfuscated data, we empirically solve the objective function by varying $p \in [0, 1]$. Specifically, for each different value of p, we repeatedly obfuscate the original data[27] and re-calculate $E(u_i)$ and $E(r_i)$ on the obfuscated trajectory, to identify the obfuscated trajectory that fulfills the objective function. In fact, any p that results in a greater % decrease in the risk than utility could be a viable solution depending on the consumers' and advertiser's individualized needs, hence producing multiple acceptable trade-off options and further exemplifying the *flexibility* of the proposed framework.

Model (Obfuscation Scheme): Acknowledging potential alternatives, we propose a personalized and flexible obfuscation scheme below grounded on repeated personalized *suppressions* of a subset of locations by varying the grid parameter p, and *structured grid search* (Coope and Price 2001) over the obfuscated trajectories to identify the solution to the objective function.

Suppression. We choose suppression to obfuscate the published location data because it holds a number of key advantages that make it best suited for business applications. One, compared to other obfuscation techniques, suppression is easier to implement and interpret for the data aggregator, for instance, without the need for a reverse algorithm to recover the original data after the obfuscation is performed. Two, it reduces the advertiser's utility the least, since unlike

[26] Note that depending on the functional forms and properties of $E(u_i)$ and $E(r_i)$, there may not always exist a single optimal solution that minimizes the objective function. For instance, how to optimize the difference of a convex and non-convex function is currently an active area of research in the optimization community. Also, this objective function may be easily adapted to alternative specifications, such as using the weighted % decrease in the utility versus risk.

[27] This process can be performed for a single individual, or for all individuals in the sample one by one before calculating the average utility across all individuals to determine the best risk-utility trade-off.

in other methods, such as perturbation, the remaining data not suppressed remain unaltered and thus best preserve the truthfulness of the data. Three, the suppressed data are widely accepted, and sometimes even preferred, by advertisers. For instance, it is common that advertisers acquire and leverage location data that are suppressed and aggregated to a POI-level, as in the case of the SafeGraph data, to accomplish geo-targeting and other marketing tasks. Lastly, suppression reduces the data storage costs, but not revenues, of the data aggregator, since advertisers compensate the data aggregator by the number of tracked individuals, instead of number of location records, in the shared data (Thompson and Warzel 2019). Suppression also saves advertisers the data storage and analysis costs.

To suppress each consumer i's original trajectory T_i, we first calculate the suppression probability of each location visited by i. Specifically, T_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 $l_i^k = (x_i^k, y_i^k)$ contains the geo-coordinates x_i^k and y_i^k (longitude and latitude) of a location k visited by consumer i, t_i^k is the corresponding timestamp, and n_i is the total number of locations visited by consumer i. Intuitively, more locations should be suppressed for a consumer with a higher privacy risk. Also, more informative locations, such as those visited often by the consumer, should be suppressed with higher probabilities. We thus formulate the personalized suppression probability for each location j visited by consumer i as driven by (a) consumer i's baseline privacy risk r_i calculated a priori on the original unobfuscated data, which has accounted for intra- and inter-individual mobility patterns as described earlier; (b) a grid parameter $p \in [0, 1]$ that scales up or down the likelihood of a location being suppressed or the total number of locations suppressed for i; and (c) location j's informativeness (i.e., suppression weight) s_i^j given consumer i's mobility patterns. Acknowledging potential alternative formula, we specify these personalized suppression probabilities for the unique locations in T_i, $L_i = \{l_i^1, l_i^2, ..., l_i^{k_i}\}$, $k_i \leq n_i$, as:

$$r_i \cdot p \cdot (1 + s_i^1), r_i \cdot p \cdot (1 + s_i^2), ..., r_i \cdot p \cdot (1 + s_i^{k_i}). \tag{8}$$

Here the informativeness of these locations $\vec{s_i} = \{s_i^1, s_i^2, ...s_i^{k_i}\}$ can be captured by the trajectory features extracted from T_i (Section 4.1.1), such as the frequency, recency, and time spent at each location. For instance, $\vec{s_i}$ based on the frequency f (and similarly recency and time spent) at these locations would be $\vec{s_i} = \{\frac{f_i^1}{\sum_{j=1}^{k_i} f_i^j}, \frac{f_i^2}{\sum_{j=1}^{k_i} f_i^j}, ..., \frac{f_i^{k_i}}{\sum_{j=1}^{k_i} f_i^j}\}$. Other features could also be incorporated, such as a location's distance to the individual's home, which produces a similar performance. Overall, this specification suggests that for a specific grid parameter p, the base suppression probability $r_i \cdot p$ ensures that a consumer at a higher baseline risk would have more locations suppressed; and the additional term $r_i \cdot p \cdot s_i^j$ ensures that a more informative location j is suppressed with an

even higher probability[28]. Now, each unique location visited by consumer i can be independently suppressed (or conversely, kept) as a Bernoulli trial based on its respective suppression probability specified in Equation 8. A Bernoulli trial, also known as binomial trial, is commonly used to capture events with exactly two possible outcomes (e.g., suppressed or kept a location in our context) (Papoulis 1984).

Here is a stylized example of suppression. Consider Alice whose trajectory is $\{H, B, H, H, M, W, E\}$ where H is her home and W is her work; and Bob whose trajectory is $\{\overline{W}, \overline{W}, Y, M, D, \overline{H}, \overline{W}\}$, where $\overline{H}$ is his home and $\overline{W}$ is his work. And M is a mall. The home inference risk is 0.75 for Alice and 0.35 for Bob. Given similar informativeness of the locations based on the frequency distributions (H=3 times, B=1, M=1, W=1, E=1 for Alice and $\overline{W} = 3$, Y =1, M =1, D=1, $\overline{H} = 1$ for Bob), and a higher home inference risk for Alice, more locations are then suppressed for Alice than Bob for a specific value of p. This will result in the obfuscated trajectories of, for example, $\{B, M, E\}$ for Alice and $\{Y, M, D, \overline{H}\}$ for Bob. Next, we will describe how to vary the grid parameter p between 0 and 1 in a structured grid search to identify all possible obfuscated trajectories $\mathcal{P}(T_i)$ that balance the risk and utility.

Structured Grid Search. As r_i and s_i^j are computed a priori from the original unobfuscated data, the suppression probability specified above depends solely on the grid parameter $p \in [0, 1]$. In an extreme scenario where consumer i's risk $r_i = 1$ and the grid parameter p is reasonably high, all locations visited by i would be suppressed. Then this complete suppression where $\{\mathcal{P}(T_i)\} = \mathcal{P}(T) = \emptyset$) would result in zero risk to consumer i, yet also zero utility to the advertiser.[29] Conversely, $p = 0$ results in no suppression ($\mathcal{P}(T) = T$), hence a high utility to the advertiser, but also high risk to the consumer. Noting these two extreme scenarios, we vary the grid parameter p from 0 and 1. For each value of p, we re-calculate the suppression probabilities based on r_i, p, and $\vec{s_i}$, suppress a subset of locations for each consumer based on these re-calculated suppression probabilities, and then re-calculate the risk and utility on the obfuscated trajectories, to be compared with the baseline risk and utility calculated a priori on the original unobfuscated data. This comparison allows us to evaluate the % reductions in the risk versus utility, given each scenario of suppression (i.e., each different value of p), and thus identify all possible obfuscated trajectories $\mathcal{P}(T)$ that could balance the risk and utility. Structured grid search reduces computational intensity compared to alternative

[28] While the suppression probability leverages each consumer's baseline risk, not utility, both are repeatedly calculated on each obfuscated trajectory to identify the risk-utility balance. This set-up circumvents the unnecessary complication of considering both the baseline risk and utility when calculating the suppression probability. Such complication could arise, for instance, from the risk and utility being both correlated with a shared set of trajectory features, such as visit frequency.

[29] Note that $s_i^j \in [0, 1]$, and $r_i \cdot p \in [0, 1]$ because $r_i \in [0, 1]$ and $p \in [0, 1]$. Nonetheless, the corresponding location is suppressed with probability 1 whenever $r_i \cdot p \cdot (1 + s_i^j) > 1$.

methods, such as a gradient descent-based approach. We also develop an early stopping heuristic with simulated annealing to further speed up the search (Online Appendix A).

In summary, the proposed obfuscation scheme is ***personalized*** with the individual-specific suppression probabilities despite a global grid parameter p. It is also ***flexible*** with a generalizable objective function that accommodates different types of risks or utilities, together with a structured grid search that is independent of the risk or utility quantification and also permits multiple acceptable levels of the risk and utility. The framework hence empowers a data aggregator to satisfy diverse demands of risk preservation from the consumer side and utility preservation from the advertiser side. It also offers transparency to the data aggregator regarding which locations are suppressed and why, for which consumer of which risk level. Moreover, compared to the literature that often requires multiple input parameters, the scheme requires no input parameter for the home inference risk, and merely one input parameter for the re-identification risk (the number of a consumer's locations known a priori to a stalker $|\overline{T}_i|$). The scheme further accounts for key characteristics of location data, such as high spatial dimensionality (a large number of unique locations) via for instance dimension reduction on the extracted features, and high temporal dimension via for instance integrating the temporal information in the activity-timing prediction (Section 4.2.2).

In summary, Figure 2 visualizes the entire proposed framework. In Part A, each consumer's baseline risk r_i and the advertiser's baseline utility u_i from the original unobfuscated data are calculated, capturing the counterfactual case with no privacy preservation, hence maximum utility yet maximum risk. In Part B, the personalized obfuscation is performed, and risk and utility from the obfuscated data re-computed repeatedly to determine the best or acceptable risk-utility trade-off. In the next section, we will describe the empirical evaluation of the proposed framework.

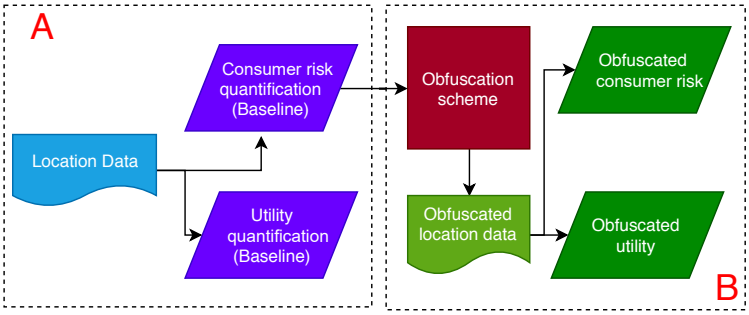


Figure 2 **Framework overview**

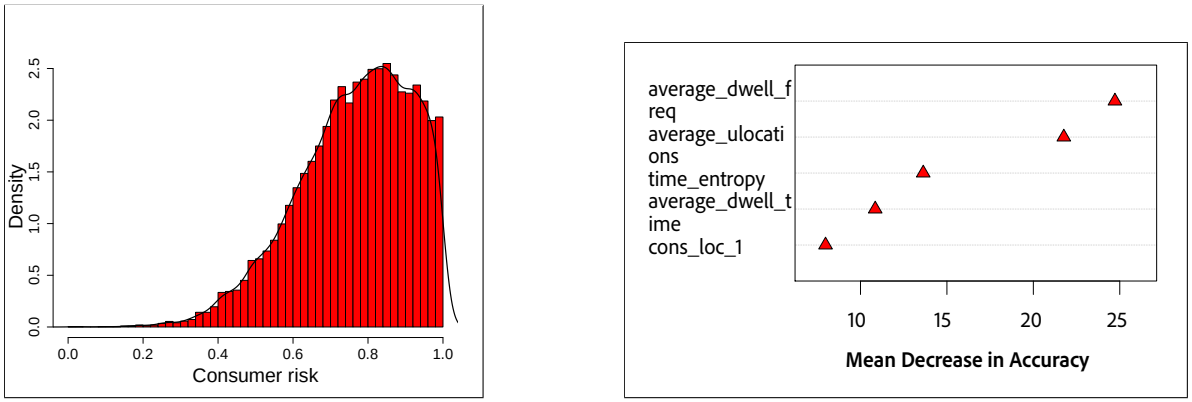
5. Empirical Study

To consistently estimate the risk and utility, we use bootstrapping with 20 trials for each p. We also randomly sample 50% of the data (20,000 consumers) as our training set (T_{train}) for training and

cross-validation (Online Appendix D), and the remaining 50% (20,012 consumers) as the test set (T_{test}) for the focal analysis. We subsequently examine alternative ways to split the training/test sets (Online Appendix E). Additional robustness studies are reported in Online Appendix G.

5.1. Quantification of Consumer's Privacy Risk

As described earlier, to compute each consumer's home inference risk, an ensemble Random Forest is trained on the training set, then used to predict the home location on the test set. Also, to compute the re-identification risk, two locations in each consumer's trajectory are assumed to be known to the stalker to illustrate our approach. Figure 4 displays the average home inference risk and average re-identification risk across all consumers in T_{test} for each p.



(a) Consumer risk density (b) Feature importance

Figure 3 Personalized obfuscation: home inference risk

The data aggregator may already garner great insights from just this initial step of quantifying the consumers' privacy risks prior to the data obfuscation, such as *which consumers are at the greatest risk; what is the severity of each privacy risk; and which feature is most informative to a stalker and should be suppressed?* Using the home inference risk as an example, Figure 3a offers a visual of the distribution of the consumers' home inference risks if the stalker were to infer their home locations from the unobfuscated data. It shows that the majority of the consumers carry relatively high risks of home inference if no obfuscation were performed. On average, the normalized haversine distance between the predicted and actual home locations (varying from 0 low risk to 1 high risk) is on the high end of 0.84. A stalker may identify a consumer's home within a small 4km (2.5 mile) radius of the actual home (Online Appendix D). While not displayed here, the average risk of re-identifying a consumer and hence his/her entire trajectory by knowing merely two randomly sampled locations visited by the consumer is 0.49, that is, 49% chance of success for the stalker. Finally, the data aggregator can assess the worst cases associated with the top-risk

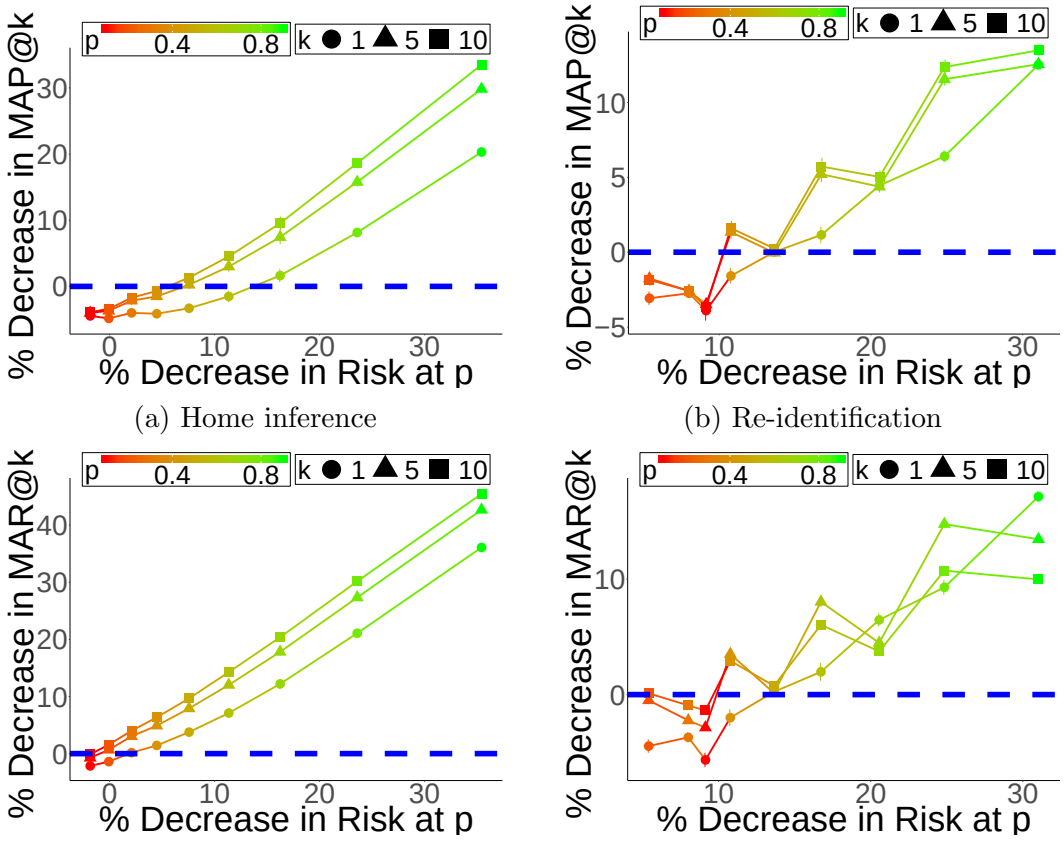
consumers in each risk. Such transparent, personalized risk assessment would enable marketers to trade off against varying utilities or marketing campaigns. Despite the serious privacy risks with the unobfuscated data, the data aggregator can curtail these risks by implementing the proposed framework. For instance, the home inference risk could be reduced by 15% while fully preserving the advertiser utility of location prediction on the POI@1 performance (Figures 4a and 4c at $p = 0.7$).

Besides inspecting the risks, the data aggregator may also assess the importance of the extracted trajectory features (Table 4) prior to the obfuscation. Again, using the home inference risk as an example, the top five most important features used in the Random Forest regressor include: a consumer's mobility patterns and consumer-location affinity (Figure 3b). Specifically, the average number of unique locations visited by a consumer (`average_ulocations`), mobility entropy that measures the predictability of a consumer's trajectory (`time_entropy`), average time spent at each location (`average_dwell_freq`), and consumer-location affinity (`cons_loc_1`, i.e., the first principal component in the consumer-location affinity tensor) are the most important features in estimating a consumer's home inference risk. These indicate to the data aggregator that the temporal information of the trajectories (`time_entropy` and `average_dwell_freq`) contributes significantly to the stalker model's predictive performance, and hence the consumer's home inference risk. Then a possible obfuscation scheme removing (even partially) the timestamps in the trajectories would potentially prevent the stalker from constructing the temporal features, and as a result, reduce the consumer's home inference risk.

5.2. Quantification of Advertiser's Utility

As described earlier, to compute the advertiser's utility in the location prediction, we leverage a collaborative filtering recommendation model. The locations visited by each consumer in the fifth week are used as the ground truth to train the recommendation model. Without obfuscation, a consumer's next location may be predicted with 25% success ($MAP@1 = 0.25$, Figure 7 in Online Appendix D). With obfuscation, the average utilities across all consumers in the test set, $MAP@k$ and $MAR@k$, for the next k locations where $k = \{1, 5, 10\}$ are computed to illustrate the method's efficacy. $MAP@k$ and $MAR@k$ for other values of k may also be computed. We perform 20 trials for each p and report the mean and 95% confidence intervals of the utility in Figure 4.

To compute the activity-timing prediction utility using the LSTM for each consumer in the training set, we randomly select one week of his/her activities as the ground truth. Before training, a hold-out validation set of one week of activities per consumer is randomly separated out from the training data set. To prevent over-fitting, the models are trained until their performances on the validation set reach a maximum. While training, we compute three losses: categorical cross-entropy



(a) Home inference (b) Re-identification

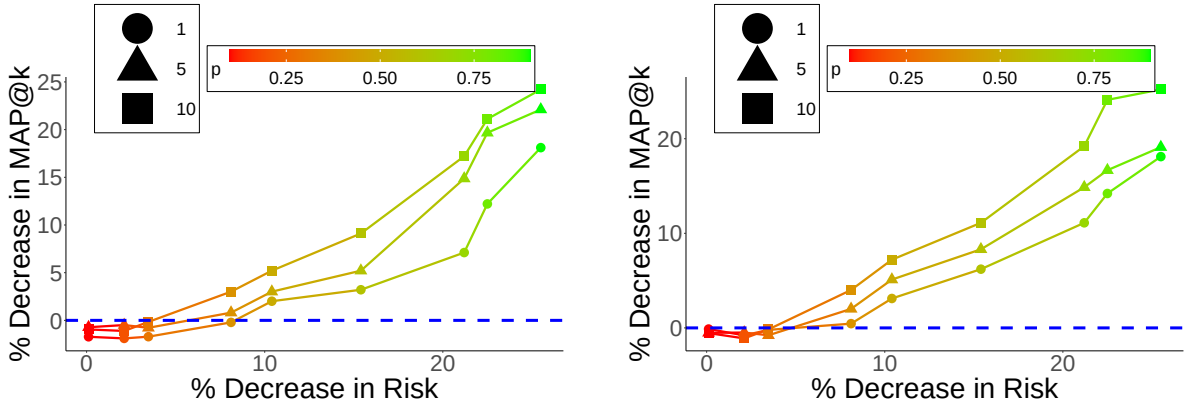
(c) Home inference (d) Re-identification

Figure 4 **Proposed framework: $MAP@k$ and $MAR@k$ for varying p in location prediction utility**

for the activity, time of the day, and binary cross-entropy for day of the week. Since the accuracy of the next activity prediction is more salient compared to that of time of the day prediction and day of the week prediction, we assign disproportionate weights $(\lambda, 1-\lambda, 1-\lambda)$, where $\lambda \in [0,1]$, while summing across these losses and empirically selecting λ with the best average validation accuracy (grid $\lambda \in [0.5, 0.6, 0.7]$). In addition to λ, we also tune the number of hidden states in the LSTM layer (grid: 64, 128, 256) and the drop-out rate (linear grid: 0.05 to 0.4, in increments of 0.05). The models are trained for a maximum of 500 epochs. The model with the best validation accuracy is then chosen to be evaluated on the test set.

We compute the AP_i^k and AR_i^k (Equations 4 and 5) for the three prediction tasks – next activity, time of the day, and day of the week. The expected utility of all consumers' trajectories $E(u_i)$ is calculated as the average across the three prediction tasks across all consumers in the test set, and is denoted as $MAP@k$ and $MAR@k$. Without obfuscation, a consumer's next activity and its timing may be predicted with 26% success ($MAP@1 = 0.26$). With obfuscation, the results for varying p, k, and their corresponding decreases in the risks for both the home inference risk and the re-identification risk are displayed in Figure 5. Similar to Figure 4, for a given % decrease in the

risk, we observe a lesser corresponding % decrease in the utility. This emphasizes the robustness of the proposed obfuscation scheme under the prediction heuristics (activity-timing prediction) that further incorporate the temporal and contextual information.



(a) Home inference (b) Re-identification

Figure 5 Proposed framework: activity-timing prediction

5.3. Personalized and Flexible Obfuscation Scheme for Data Aggregator

We focus on discussing the results using the suppression weights $\vec{s_i}$ based on the frequency at each location. Online Appendix G.1 reports the results based on the recency and time spent at each location. Figures 4a and 4b visualize the risk-utility trade-off based on $MAP@k$, and figures 4c and 4d based on $MAR@k$. Corresponding to each $p \in \mathcal{G}_p$, the X-axis displays the % decrease in the mean risk from the baseline risk (second term in Equation 6), and Y-axis the % decrease in the utility from the baseline utility (first term in Equation 6). Therefore, a higher X-value means more decrease in the risk and hence better privacy preservation. A higher Y-value means more decrease in the utility, hence worse off for the advertiser. We can see that for each $k = \{1, 5, 10\}$ (corresponding to each curve in each graph), as p increases (a curve's color changing from red to green), more locations are suppressed, and hence more reductions in the risk, but also more decreases in the utility. As a result, the data aggregator is presented with multiple options of k and p to balance a specific type of risk and utility: e.g., as long as the decrease in the risk is larger than the decrease in the utility (i.e., the area in the first quadrant below the 45-degree line). A special case is the blue line (Y = 0), with decreases in the risk while zero sacrifice for the utility.[30]

[30] The figure shows that in some cases (e.g., below the blue line), the decrease in the utility ($MAP@k$ and $MAR@k$) is actually negative. That is, the utility has actually increased upon the obfuscation. This could happen because the obfuscation has potentially removed some noise from the location data, leading to a better model performance.

In summary, these results showcase the flexibility of the proposed framework in furnishing the data aggregator with multiple options in obfuscations, to satisfy diverse demands from various entities in the location ecosystem. An auxiliary benefit of the framework is that data suppression actually reduces the data aggregator's and advertiser's data storage and analytic costs. Finally, while the obfuscation is performed offline by the data aggregator before sharing the location data with the advertisers, we still report the time complexity and clock-time of the proposed obfuscation scheme in Online Appendix B.

5.4. Model Comparison

We compare the proposed framework with ten baselines representing the core PPDP methods for location data reviewed in Section 2 and also most comparable to our proposed framework. The first group includes 3 rule-based baselines that obfuscate locations with specific timestamps, a simple method commonly used in the early PPDP literature (Online Appendix H). The second group showcases 3 global risk-based obfuscations to demonstrate the advantage of *personalized* risk quantification and *personalized* obfuscation only examined by a few recent studies in Table 1 (Online Appendix I). The third group contains 2 baselines representing some of the latest syntactic models in Table 1, LSup and GSup (Terrovitis et al. 2017). And the fourth group covers 2 newest synthetic trajectory methods, PPMTF (Murakami et al. 2019) and LSTM-TrajGAN (Rao et al. 2020). We will focus on the last two groups below. A method is considered superior if the % decrease in the utility is less than the % decrease in the risk.

Obfuscation Method	% decrease in home inference risk	% decrease in re-identification risk	% decrease in location prediction utility ($MAP@1$)	% decrease in location prediction utility ($MAR@1$)
GSup ($P_{br} = 0.2$)	18.12	14.52	7.74	8.31
GSup ($P_{br} = 0.5$)	7.25	7.29	4.49	3.42
LSup ($P_{br} = 0.2$)	22.16	31.56	5.31	7.12
LSup ($P_{br} = 0.5$)	9.15	10.91	-1.65	0.86

Table 6 Proposed framework vs LSup and GSup:
green/red indicate proposed framework offers a better/worse trade-off

Latest Syntactic Models. LSup and GSup (Terrovitis et al. 2017) obfuscate data to mitigate the re-identification risk. Methodologically, these models differ from the proposed framework in three important ways. First, they quantify only the re-identification risk, whereas the proposed framework can *flexibly* accommodate other risks. Second, these methods suppress a location either globally across all consumers (GSup, similar to the mean-risk baseline in Figure 13) or locally for a subset of consumers quantified as risky (LSup, e.g., all consumers visiting the same mall). In

contrast, the proposed scheme suppresses a location at a consumer-level with the consumer-specific, i.e., *personalized*, parameters $\{\vec{s_i}, z_i\}$. Third, these methods 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), where a higher P_{br} results in fewer locations suppressed. In comparison, the proposed framework requires either no input parameter (for home inference risk) or a single parsimonious input (for re-identification risk, the number of locations known a priori to a stalker), making the framework easy to implement in practice.

In model comparison, we follow these authors' empirical evaluation framework to set the number of adversaries $\mathcal{A}$ and background knowledge of each adversary in $\mathcal{A}$; and merely vary P_{br}. When comparing the slope $\frac{Y}{X}$ in Figure 4 that capture the % decrease in the utility divided by % decrease in the risk for the different decreases in the utility ($MAP@1$) of LSup and GSup (Table 6[31]), we observe that in most (6 out of 8) cases, the proposed framework provides a better trade-off (denoted by green color in Table 6) than LSup and GSup. While it does not always outperform LSup and GSup, it can flexibly accommodate multiple types of risks and utilities – a key feature of a PPDP framework sought by the location ecosystem, whereas LSup and GSup only focus on a single re-identification risk. Another benefit of the proposed framework is that it requires zero or only one input parameter, as discussed earlier.

Synthetic Trajectory Generation Models. Finally, we compare the proposed method with the synthetic trajectory generators, PPMTF (Murakami et al. 2019) and LSTM-TrajGAN (Rao et al. 2020). We execute the publicly available implementations of these methods[32], and adhere to the pre-processing and parameter suggestions made by the respective authors. Using the output synthetic trajectories, we compute the % decrease in the home inference and re-identification risks, respectively, from the original trajectories (Section 4.1), and the % decrease in the advertiser utility for the location prediction and activity-timing prediction, respectively (Section 4.2). Table 7 shows that in 7 out of 8 cases, the proposed framework provides a better trade-off (denoted by the green color) compared to both PPMTF and LSTM-TrajGAN.

[31] Terrovitis et al. (2017) consider four values of P_{br}: {0.2, 0.25, 0.33, 0.50} and conclude that for a fixed number of adversaries, a higher data utility occurs at higher P_{br} values (less locations suppressed) while ensuring reduction in the re-identification risk. The best value suggested in their work is $P_{br} = 0.5$. Our choice of P_{br} is based on these experiments and observations. Also, since these models do not address home inference, we obfuscate the data to reduce the re-identification risk and use the same obfuscated data to quantify the reductions in the home inference risk and the re-identification risk.

[32] PPMTF: `https://github.com/PPMTF/PPMTF`. LSTM-TrajGAN: `https://github.com/GeoDS/LSTM-TrajGAN`.

Method/Utility	% decrease in home inference risk	% decrease in re-identification risk	% decrease in location prediction utility ($MAP@1$)	% decrease in location prediction utility ($MAR@1$)
PPMTF Location prediction	21.51	26.73	12.61	11.27
LSTMTrajGAN Location prediction	15.81	13.69	17.65	15.28
PPMTF Activity-timing prediction	21.51	26.73	12.51	11.83
LSTMTrajGAN Activity-timing prediction	15.81	13.69	7.43	2.45

Table 7 Proposed framework vs synthetic trajectory models: green/red indicate proposed framework offers a better/worse trade-off

6. Conclusion and Discussion

Mobile location-based technologies have generated massive volumes of user trajectory data over recent years, and access to such location-based data has produced a wide range of benefits to various entities in the ecosystem: consumers, data aggregators, and advertisers. Meanwhile, for some entities, potential privacy concerns may arise from access to these granular data, calling for data aggregators to adopt an effective, personalized, and flexible PPDP framework, to address consumers' heterogeneous privacy demands while also meeting advertisers' heterogeneous needs. This research tackles this essential yet under-studied topic, and contributes to an important area of research that cuts across multiple disciplines – PPDP of consumer mobile location data – by proposing a ***personalized*** and ***flexible*** framework to unleash the potential of location-based marketing while protecting consumer privacy. The ***personalized*** data obfuscation approach accounts for the heterogeneity in consumers' privacy preferences and data valuations. The ***flexible*** accommodation of different types of risks (costs) and utilities (benefits), as well as different acceptable levels of risk and utility, permits a data aggregator to satisfy diverse needs from both the consumers and advertisers, as well as its own need for varied computational efficiency. Overall, this research fills a critical void in the literature and offers an important tool for the privacy-aware practices of big data location-based apps and services.

Specifically, we illustrate the framework with two potential privacy costs to consumers (home inference risk and re-identification risk) and two potential benefits to advertisers (location prediction and activity-timing prediction). We further validate the framework on a million real world individual-level trajectories. Our analyses demonstrate that the proposed framework accounts for distinct characteristics of the population-scale individual-level location data (high spatial and temporal dimensions), and outperforms multiple benchmark methods from the latest literature. Our findings indicate that potential privacy risks can prevail in the absence of data obfuscation. And the personalized risk quantification enables a data aggregator to identify high-risk individuals, and the data features that contribute the most to each potential risk. Furthermore, with the proposed obfuscation scheme, a data aggregator can nimbly accomplish a more beneficial privacy-utility

trade-off compared to the existing methods from the literature. For instance, the home inference risk can be reduced by 15% with less than 1% decrease in the utility.

The proposed framework also offers important managerial and policy implications. For instance, it can incorporate various types of potential privacy risks, such as location sequence inference, or visit frequency inference, for which the risks may be quantified analytically or via machine learning heuristics (Pellungrini et al. 2018). It also permits various types of utilities or benefits from business applications, such as predictions of the likelihood and timing of customer conversion given past trajectories, or evaluations of the incremental revenue from a location-based marketing campaign. Such versatility can allow advertisers and other downstream data users to harness the power of the individual-level location big data, while also protecting consumer privacy.

From the perspective of the policy makers, our study examines the potential benefits and costs of access to location data, reinforces the importance of maintaining the risk-utility balance, and provides a powerful and interpretable solution that can benefit all entities in the digital ecosystem, especially consumers. More generally, policy makers may want to consider broadening the focus of privacy preservation from merely issuing a blanket ban for limiting data access, to adopting a privacy friendly data sharing and utilization approach, i.e., PPDP, particularly in light of consumers' willingness to share data in exchange for various economic or societal benefits (Ghose et al. 2020), as well as recognizing the existence of the privacy paradox wherein there is a disconnect between what consumers claim vs. how they behave (Ghose 2017). Finally, this research sheds additional light on the broader discussions pertaining to privacy enhancing techniques on the internet and also illustrates a viable solution.

Despite the contributions, this research has some limitations, which can be fruitful sources for future research. For example, our data contain no information about individual consumers' demographics. When such data become available, greater insights may be garnered into which demographics are associated with higher potential privacy costs. This is important because heterogeneity in privacy preferences and valuations exists across demographics, both within and across time and context. Also, our analysis considers the longitudes and latitudes, but not the names of the locations of the entities. Hence future research may further differentiate sensitivity levels across locations. Furthermore, as other sources of consumer-level data, such as click streams or search queries become linked to the location data, more sophisticated privacy preservation methodologies will be needed because there is substantial heterogeneity in a given consumer's privacy preferences and valuations across different kinds of data originating from different sources. In addition, we demonstrate the framework with two potential risks and two utilities. Future research may explore alternative situations, such as social relationship detection or public surveillance, and alternative utilities, such as lifestyle profiling or potential customer identification. It would also be interesting

to assess the cost of accommodating different types of risks and utilities. Lastly, the proposed framework considers only one-shot data sharing with the advertiser. Future research may explore more complex scenarios with multiple risks, multiple utilities, or when a advertiser combines multiple batches or sources of shared data.

References

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

Ahmad, M.W., M. Mourshed, Y Rezgui. 2017. Trees vs neurons: Comparison between random forest and ann for high-resolution prediction of building energy consumption. *Energy and Buildings* **147** 77–89.

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

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

Bhakta, Arnav, Yeunjoo Kim, Pamela Cole. 2021. Comparing machine learning-centered approaches for forecasting language patterns during frustration in early childhood URL `https://arxiv.org/pdf/2110.15778.pdf`.

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

Brauer, Anna, Ville Mäkinen, Axel Forsch, Juha Oksanen, Jan-Henrik Haunert. 2022. My home is my secret: concealing sensitive locations by context-aware trajectory truncation. *International Journal of Geographical Information Science* 1–29.

Breiman, L. 2001. Random forests. *Machine Learning* **45** 5–32.

Chandra, Shobhana, Sanjeev Verma, Weng Marc Lim, Satish Kumar, Naveen Donthu. 2022. Personalization in personalized marketing: Trends and ways forward. *Psychology & Marketing* **39**(8) 1529–1562. doi:https://doi.org/10.1002/mar.21670.

Chen, Peng, Niu Aichen, Jiang Wei, Liu Duanyang. 2019. Air pollutant prediction: Comparisons between lstm, light gbm and random forest. *Geophysical Research Abstracts* **21** EGU2019–3121–1.

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

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

Chen, Xi, David Simchi-Levi, Yining Wang. 2022. Privacy-preserving dynamic personalized pricing with demand learning. *Management Science* **68**(7) 4878–4898.

Chow, Chi-Yin, Mohamed F Mokbel. 2011. Trajectory privacy in location-based services and data publication. *ACM Sigkdd Explorations Newsletter* **13**(1) 19–29.

Coope, Ian D, Christopher John Price. 2001. On the convergence of grid-based methods for unconstrained optimization. *SIAM Journal on Optimization* **11**(4) 859–869.

Cunha, Mariana, Ricardo Mendes, João P Vilela. 2021. A survey of privacy-preserving mechanisms for heterogeneous data types. *Computer science review* **41** 100403.

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

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

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

Fernández-Delgado, M., E. Cernadas, S. Barro. 2014. Do we need hundreds of classifers to solve real world classifcation problems? *Journal of Machine Learning Research* **15** 3133–3181.

Fiore, Marco, Panagiota Katsikouli, Elli Zavou, Mathieu Cunche, Françoise Fessant, Dominique Le Hello, Ulrich Matchi Aivodji, Baptiste Olivier, Tony Quertier, Razvan Stanica. 2020. Privacy in trajectory micro-data publishing: a survey. *Transactions on Data Privacy, IIIA-CSIC* **13** 91–149.

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

Gao, Sheng, Jianfeng Ma, Cong Sun, Xinghua Li. 2014. Balancing trajectory privacy and data utility using a personalized anonymization model. *Journal of Network and Computer Applications* **38** 125–134.

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

Ghose, A., B. Li, M. Macha, C. Sun, N. Z. Foutz. 2020. Trading privacy for public good: How did america react during covid-19? *SSRN Working Paper* .

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

Ghose, Anindya, Beibei Li, Siyuan Liu. 2019. Mobile targeting using customer trajectory patterns. *Management Science* **65**(11) 4951–5448.

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

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

Hochreiter, S., J. Schmidhuber. 1997. Long short-term memory. *Neural Computation* **9**(8).

Huang, Jiun-Chi, Yi-Chun Tsai, Pei-Yu Wu, Yu-Hui Lien, Chih-Yi Chien, Chih-Feng Kuo, Jeng-Fung Hung, Szu-Chia Chen, Chao-Hung Kuo. 2020. Predictive modeling of blood pressure during hemodialysis: a comparison of linear model, random forest, support vector regression, xgboost, lasso regression and ensemble method. *Computer Methods and Programs in Biomedicine* **195** 105536. doi:https://doi.org/10.1016/j.cmpb.2020.105536.

Huo, Zheng, Xiaofeng Meng, Haibo Hu, Yi Huang. 2012. You can walk alone: trajectory privacy-preserving through significant stays protection. *International conference on database systems for advanced applications*. Springer, 351–366.

Hwang, Ren-Hung, Yu-Ling Hsueh, Hao-Wei Chung. 2013. A novel time-obfuscated algorithm for trajectory privacy protection. *IEEE Transactions on Services Computing* **7**(2) 126–139.

Jhaveri, S., I. Khedkar, Y. Kantharia, S. Jaswal. 2019. Success prediction using random forest, catboost, xgboost and adaboost for kickstarter campaigns. *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)* 1170–1173doi:https://doi.org/10.1109/ICCMC.2019.8819828.

Jiang, Hongbo, Jie Li, Ping Zhao, Fanzi Zeng, Xiao Zhu, Arun Iyengar. 2022. Location privacy-preserving mechanisms in location-based services: A comprehensive survey. *ACM Computing Surveys* **54**(1) 1–36. doi:https://doi.org/10.1145/3423165.

Jin, Fengmei, Wen Hua, Matteo Francia, Pingfu Chao, Maria Orlowska, Xiaofang Zhou. 2022. A survey and experimental study on privacy-preserving trajectory data publishing. *IEEE Transactions on Knowledge and Data Engineering* 1–1doi:10.1109/TKDE.2022.3174204.

Kahneman, Daniel, Amos Tversky. 1979. Prospect theory: An analysis of decision under risk. *Econometrica* **47**(2) 263–291.

Katsomallos1, Manos, Katerina Tzompanaki1, Dimitris Kotzinos1. 2019. Privacy, space, and time: a survey on privacy-preserving continuous data publishing. *JOURNAL OF SPATIAL INFORMATION SCIENCE* **19** 57–103.

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

Komishani, Elahe Ghasemi, Mahdi Abadi, Fatemeh Deldar. 2016. Pptd: Preserving personalized privacy in trajectory data publishing by sensitive attribute generalization and trajectory local suppression. *Knowledge-Based Systems* **94** 43–59.

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

Li, Ninghui, Tiancheng Li, Suresh Venkatasubramanian. 2007b. t-closeness: Privacy beyond k-anonymity and ldiversity. *IEEE 23rd International Conference on. IEEE, 2007* .

Li, Songyuan, Hong Shen, Yingpeng Sang. 2020. A survey of privacy-preserving techniques on trajectory data. Hong Shen, Yingpeng Sang, eds., *Parallel Architectures, Algorithms and Programming*. Springer Singapore, Singapore, 461–476.

Li, Xiao-Bai, Jialun Qin. 2017. Anonymizing and sharing medical text records. *Information Systems Research* **28**(2) 332–352.

Li, Xiao-Bai, Sumit Sarkar. 2013. Class-restricted clustering and microperturbation for data privacy. *Management Science* **59**(4) 796–812.

Li, Xiao-Bai, Sumit Sarkar. 2014. Digression and value concatenation to enable privacy-preserving regression. *MIS Quarterly* **38**(3) 679–698.

Li, Xiao-Bai, Sumit Sarker. 2011. Protecting privacy against record linkage disclosure: A bounded swapping approach for numeric data. *Information Systems Research* **22**(4) 774–7891.

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

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

Mahdavifar, Samaneh, Fatemeh Deldar, Hassan Mahdikhani. 2022. Personalized privacy-preserving publication of trajectory data by generalization and distortion of moving points. *Journal of Network and Systems Management* **30**(10).

Menon, Syam, Sumit Sarkar. 2016. Privacy and big data: Scalable approaches to sanitize large transactional databases for sharing. *MIS Quarterly* **40**(4) 963–982.

Murakami, Takao, Koki Hamada, Yusuke Kawamoto, Takuma Hatano. 2019. Privacy-preserving multiple tensor factorization for synthesizing large-scale location traces with cluster-specific features. *arXiv preprint arXiv:1911.04226* .

Papoulis, A. 1984. *Probability, Random Variables, and Stochastic Processes*.

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

Pascual, S., A. Bonafonte. 2016. Multi-output rnn-lstm for multiple speaker speech synthesis and adaptation. *2016 24th European Signal Processing Conference (EUSIPCO)* 2325–2329, doi: 10.1109/EUSIPCO.2016.7760664.

Pelekis, Nikos, Aris Gkoulalas-Divanis, Marios Vodas, Despina Kopanaki, Yannis Theodoridis. 2011. Privacy-aware querying over sensitive trajectory data. *Proceedings of the 20th ACM international conference on Information and knowledge management*. 895–904.

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

Pew. 2018. 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/`.

Primault, Vincent, Antoine Boutet, Sonia Ben Mokhtar, Lionel Brunie. 2019. The long road to computational location privacy: A survey. *IEEE Communications Surveys Tutorials* **21**(3) 2772–2793. doi:10.1109/COMST.2018.2873950.

Qiu, Guoying, Deke Guo, Yulong Shen, Guoming Tang, Sheng Chen. 2021. Mobile semantic-aware trajectory for personalized location privacy preservation. *IEEE Internet of Things Journal* **8**(21) 16165–16180. doi:10.1109/JIOT.2020.3016466.

Rafieian, Omid, Hema Yoganarasimhan. 2021. Targeting and privacy in mobile advertising. *Marketing Science* **40**(2) 193–218.

Rao, Jinmeng, Song Gao, Yuhao Kang, Qunying Huang. 2020. Lstm-trajgan: A deep learning approach to trajectory privacy protection. *arXiv preprint arXiv:2006.10521* .

Sagi, O., L. Rokach. 2018. Ensemble learning: A survey. *Wiley Interdisciplinary Review: Data Miniing and Knowledge Discovery* **8** e1249.

Said, N., A.M. Mouazen. 2017. Comparison between random forests, artificial neural networks and gradient boosted machines methods of on-line vis-nir spectroscopy measurements of soil total nitrogen and total carbon. *Sensor* **17** 2428–2450.

Sarkar, Mainak, Arnaud De Bruyn. 2021. Lstm response models for direct marketing analytics: Replacing feature engineering with deep learning. *Journal of Interactive Marketing* **53** 80–95. doi:https://doi.org/10.1016/j.intmar.2020.07.002.

Taylor, Kyle, Laura Silver. 2019. 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/`.

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

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

Tong, Siliang, Xueming Luo, Bo Xu. 2020. Personalized mobile marketing strategies. *Journal of the Academy of Marketing Science* **48** 64–87.

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

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

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

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

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

Weinberg, AI, M. Last. 2019. Selecting a representative decision tree from an ensemble of decision-tree models for fast big data classification. *J Big Data* **6**(1:23).

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

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

Yao, Lin, Zhenyu Chen, Haibo Hu, Guowei Wu, Bin Wu. 2021. Sensitive attribute privacy preservation of trajectory data publishing based on l-diversity. *Distributed and parallel databases* **39**(3) 785–811.

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

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

Zhou, Yanlai, Fi-John Chang, Li-Chiu Chang, I-Feng Kao, Yi-Shin Wang. 2019. Explore a deep learning multi-output neural network for regional multi-step-ahead air quality forecasts. *Journal of Cleaner Production* **209** 134–145. doi:https://doi.org/10.1016/j.jclepro.2018.10.243.

Zhou, Yinghui, Shasha Lu, Min Ding. 2020. Contour-as-face framework: A method to preserve privacy and perception. *Journal of Marketing Research* **57**(4) 617–639.

Appendix A: Early Stopping with Simulated Annealing

As described earlier, the exhaustive grid-based approach offers two key advantages: (1) flexibility to incorporate different types of risks and different types of utilities; and (2) reduced computational intensity compared to a descent-based approach. Nonetheless, it comes with its own shortcomings. Specifically, the discretization of the grid p would impact the best trade-off achieved. While considering a finer discretization of p can remedy this, computational issues might occur in estimating $E(r_i)$ and $E(u_i)$. We partially address this by estimating $E(r_i)$ and $E(u_i)$ for different values of $p \in \{0, 0.1, ..., 1\}$ in parallel. However, an exhaustive search over a finer grid would require constraining the search space to remain computationally efficient.

To alleviate this, we propose an early stopping heuristic that improves the current grid-based search by starting from coarser grid intervals of p, instead of a fixed grid of points, iteratively estimating $E(r_i)$ and $E(u_i)$ for a finer grid of values efficiently using simulated annealing, guided by an acceptable decrease in the advertiser's utility. We will outline the early stopping heuristic below.

Input: N consumer trajectories $\{T_i\}$, estimators for $E(r_i)$ and $E(u_i)$, where $r_i = \mathcal{PR}(T_i; \{\vec{s_i}, z_i\}), u_i = \mathcal{U}(T_i; \{\vec{s_i}, z_i\})$, an acceptable relative decrease in the advertiser utility U^{acc}.

Output: The obfuscated consumer trajectories $\{P(T_i)\}$.

1. Start with a coarse set of grid intervals for $G_p \in \{[0, 0.1], [0.1, 0.2], \ldots, [0.9, 1]\}$ and a possible set of pre-computed $\vec{s_i}$ based on frequency, recency, and time spent at each location in T_i.

2. Set $G_{prune} = \phi$.

   (a) **Estimation**: In parallel, repeatedly sample N_s consumers from N.

   i. In each iteration, for each $\vec{s_i}$, compute the best choice of p using simulated annealing[33] p_g^{sim}, $\forall$ $g \in G_p$.

   ii. Compute $E(r_i)$ and $E(u_i)$ over M iterations by suppressing the locations using Eq 8). The average of the M iterations, each computed at their respective p_g^{sim}, correspond to the estimates for a grid in G_p.

   (b) **Pruning**:

   i. If $\frac{(E^*(u_i) - E(u_i))}{E^*(u_i)} < U^{acc}$, add the corresponding g to G_{prune}.

   ii. In G_{prune}, keep top $\lceil \frac{|G_{prune}|}{2} \rceil$ grids based on the increasingly sorted $\frac{(E^*(r_i) - E(r_i))}{E^*(r_i)}$.

   iii. **Stopping criterion**: If the M paired estimates of $\frac{(E^*(r_i) - E(r_i))}{E^*(r_i)}$ of the top two grid intervals do not have a statistically significant difference under paired t-test statistic, or if G_{prune} is empty, pick the obfuscation parameters with the highest $\frac{(E^*(r_i) - E(r_i))}{E^*(r_i)}$ in Step 2.1 ii), obfuscate $\{T_i\}$ and return $\{P(T_i)\}$.

   (c) **Candidate Set**:

   i. Construct finer grids for each $g \in G_{prune}$ by splitting each g at the average value of p_g^{sim} across M iterations, resulting in a maximum of $|G_p| + 1$ candidate sets.

   ii. Set $G_p = G_{prune}$ and go to 2.2.

Overall, the algorithm starts off with a coarse-grained set of grid intervals G_p, instead of a fixed set of points. We instantiate $|G_p|$ independent parallel threads at Step 2), similar to the parallel computation in the proposed fixed grid approach. Each thread is responsible for computing the estimates of $E(u_i)$ and $E(r_i)$

[33] For simulated annealing, we use an initial temperature of 10 and a total of 100 iterations to converge to p_g^{sim}.

for a grid interval g in G_p, across M repeated sample trajectories of size N_s, which are again executed in parallel. In the fixed grid approach, we compute this estimate by averaging across twenty trials for a fixed value of p (Section 4.1 and 4.2) on all N consumer trajectories.

Next, each child thread involves performing simulated annealing[34] in a grid interval. This involves the following steps,

- Start with an initial candidate of p in the grid interval (E.g.: 0.21 in grid [0.2, 0.3]), computing $E(u_i)$ and $E(r_i)$ across the three specifications of $\vec{s_i}$.
- Generate another candidate within the bounds of the grid, compute the temperature (we start with an initial temperature of 10) and the metropolis acceptance criterion for the current iteration (we perform a total of 100 iterations).
- Update the candidate if the metropolis criterion is met, repeat until maximum number of iterations, return the optimal value returned by simulated annealing p_g^{sim}.

The average of the M resulting estimates at the corresponding p_g^{sim} for each g is used to prune G_p to remain computationally efficient and generate finer grid intervals in the successive iterations, thus performing an exhaustive search of the parameter space.

In Step 2.2) i), an optional parameter – acceptable relative decrease in the advertiser's utility U^{acc} is used to prune the grid intervals in G_p into G_{prune}. We further prune G_p by dropping the grid intervals corresponding to the bottom quantile of the relative decreases in the risk $\frac{(E^*(r_i)-E(r_i))}{E^*(r_i)}$. If the resulting G_{prune} is empty, or if the means of the top two estimates in G_{prune} are not statistically significant under the paired t-test statistic, we stop the search and obfuscate $\{T_i\}$ based on the parameters that result in the highest relative decrease in the risk in Step 2.1) ii). Next, we generate a finer candidate set based on the resulting non-empty G_{prune} by splitting each g at the average over M iterations of p_g^{sim}. This results in a maximum of $|G_p|+1$ candidate sets for the next iteration which happens when no pruning is done due to U^{acc} and if $|G_p|$ is odd.

[34] Interested readers may learn more about the implementation of simulated annealing at `https://machinelearningmastery.com/simulated-annealing-from-scratch-in-python/`.

Utility	Risk	k	Acceptable decrease in u	Best p	% decrease in r
Location prediction	Home inference	1	5	0.76	20.3
		5	5	0.62	9.2
		10	5	0.59	11.5
		1	10	0.83	23.1
		5	10	0.86	18.2
		10	10	0.81	19.4
	Re-identification	1	5	0.73	21.2
		5	5	0.53	16.1
		10	5	0.57	18.2
		1	10	0.93	29.6
		5	10	0.80	21.2
		10	10	0.84	20.4

Table 8 Early stopping heuristic with Simulated Annealing: POI@k

Utility	Risk	k	Acceptable decrease in u	Best p	% decrease in r
Activity-timing prediction	Home inference	1	5	0.65	15.9
		5	5	0.52	12.2
		10	5	0.45	10.2
		1	10	0.79	21.2
		5	10	0.65	16.2
		10	10	0.58	13.7
	Re-identification	1	5	0.59	17.2
		5	5	0.52	13.1
		10	5	0.47	9.4
		1	10	0.70	22.1
		5	10	0.63	18.3
		10	10	0.53	13.9

Table 9 Early stopping heuristic with Simulated Annealing: activity-timing prediction@k

Appendix B: Complexity Analysis

We envision that data obfuscation will be performed offline by the data aggregator before sharing the data with advertisers. The proposed obfuscation scheme requires computing features $\mathcal{F}(T_i)$ and inference of u_i and r_i for a trajectory T_i (or an obfuscated trajectory $\mathcal{P}(T_i)$) from a trained machine learning heuristic. Denote these inference times for a single consumer trajectory T_i as $O(F_i)$, $O(u_i)$ and $O(r_i)$. Note that these vary depending on the data aggregator's choice of the risk and utility functions. To compute the estimates presented in Figure 4, we vary the grid parameter $p \in \{0, 0.1, ..., 1\}$ and estimate $E(u_i)$ and $E(r_i)$ for twenty trials. This involves $20 \times 10 \times N \times O(F_i) \approx N \times O(F_i)$ time for feature computation. Once the features are built, these are fed into the corresponding risk and utility estimations - $20 \times 10 \times N \times O(u_i) \times \mathrm{O(r_i)} \approx N \times O(u_i) \times O(r_i)$. Hence the total time complexity is bounded by $O(N(F_i + u_i r_i))$.

1. In the case of Random Forests employed for the home inference risk and in the sped-up heuristic for the re-identification threat (Section 4.1.2), since $N \gg d$, the inference complexity is bounded by $O(N)$, hence $O(u_i) \approx 1$. For the POI prediction, the inference is again linear to compute the nearest k locations for each consumer using a selection algorithm. Hence, the overall complexity is bounded by $O(NF_i)$ for both privacy threats in the location prediction.

2. In the activity-timing prediction, we employ a LSTM to quantify the utility. The inference time is linear in the number of points. While we do not need feature computation for the activity-timing prediction, these still need to be computed to quantify the risk. Hence, the complexity is the same as earlier, bounded by $O(NF_i)$.

In the proposed early stopping heuristic, additional computation overhead arises from spanning across a finer grid of parameters, averaging over $M = 50$ repeated trials, until a stopping criterion is met. However, this overhead is offset since the estimates are computed on a subset of the data N_s. This is observed in Figure 6 and Table 10, where the early stopping heuristic is on average four times faster than the fixed-point grid-based search. We repeat all experiments with the proposed early stopping heuristic and report the resulting relative decreases in the risk and utility in Tables 8 and 9.

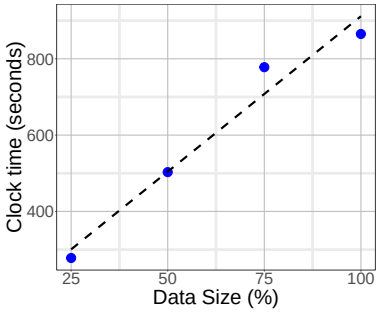


Figure 6: Clock time for home location inference and location prediction

Utility	Risk		Clock time (seconds)	
		Data	Full grid	Early stop
Location prediction	Re-identification	100%	865	226
Location prediction	Home inference	100%	978	258
Activity-timing prediction	Re-identification	100%	1390	312
Activity-timing prediction	Home inference	100%	1543	396
Location prediction	Re-identification	50%	503	136
Location prediction	Home inference	50%	645	187
Activity-timing prediction	Re-identification	50%	790	210
Activity-timing prediction	Home inference	50%	832	225

Table 10: Clock time of proposed heuristic

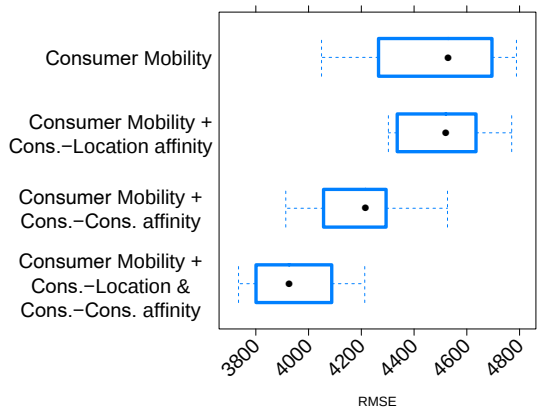
Appendix C: Speed-up Heuristic

While the re-identification risk can be exactly computed for a given $|\bar{T}_i|$, it is computationally inefficient with a complexity of $\binom{O(|T_i|}{|\bar{T}_i| \times N)}$. 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 consumers and use the trained model to approximate the risks for the remaining consumers (see Online Appendix D for the technical details).

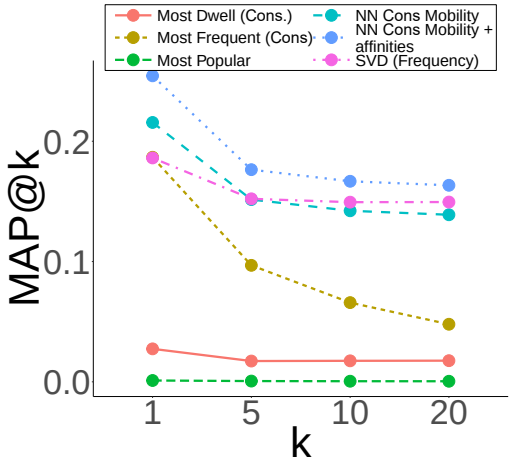
We compute the data utility under different obfuscations and by computing the performance of an NN-based collaborative filtering recommendation model to accurately predict future consumer locations. To assess the accuracy of in location and activity-timing predictions, we treat the locations visited by each consumer and correspondingly the activities by each consumer in the fifth week as the ground truth and train the models to predict these locations/activities. Based on the risks, we obfuscate T_{train} by varying $p \in \mathcal{G}_p$. We learn an NN-based recommendation model (Bobadilla et al. 2011) by tuning the number of neighbors via five-fold cross-validation on the obfuscated training set $\mathcal{P}(T_{train})$. 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 the 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 twenty trials for each p and report the mean and 95% confidence intervals of the utility metrics in Figure 4.

Appendix D: Model Choices and Robustness

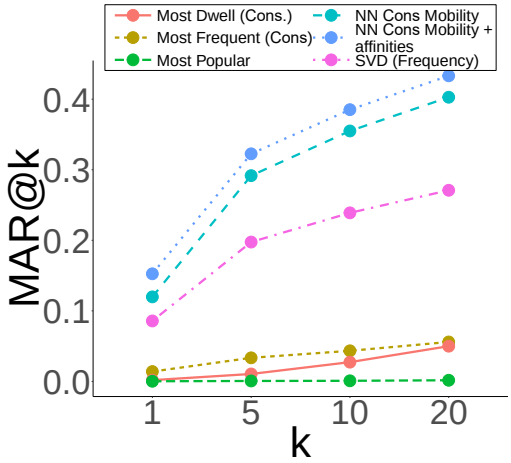
Here we will empirically justify our model choices in the proposed framework. All choices are made based on assessing the performance of different machine learning heuristics used in our framework on the unobfuscated data. In Figure 7a, we show the incremental benefit of the affinity features discussed in extracting the features $\mathcal{F}(T)$. In Figure 7a, we plot the mean haversine distance of the Random Forest regressor trained to predict home locations.[35] 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.



(a) Home inference - mean haversine distance model



(b) $MAP@k$ location prediction model



(c) $MAR@k$ location prediction model

Figure 7 Proposed framework: model choices

Next, we learn two regression models to predict each consumer's home latitude and longitude in UTM with similar hyper-parameter tuning as earlier. The error estimate is the haversine distance between the

[35] We treat the most frequently visited location 10pm-6am over weekdays by each consumer as the ground truth of home location. The results remain robust across alternative operationalizations, such as 11pm-5am. We do not save these home locations to preserve consumer privacy.

Utility	Risk	k	Acceptable decrease in u	Best p	% decrease in r
Location prediction	Home inference with LSTM	1	5	0.69	16.7
		5	5	0.61	11.2
		10	5	0.57	8.1
		1	10	0.78	20.1
		5	10	0.75	15.8
		10	10	0.68	13.2

Table 11 LSTM for home inference: early stopping: $POI@k$

predicted home latitude/longitude and actual home latitude/longitude. From the box plots of the re-sampled performance measures (Figure 7a), we notice that the consumer-consumer and consumer-location affinity features incrementally improve the performance of both proxy models learned. In Figures 7b and 7c, we visualize the $MAP@k$ and $MAR@k$ of the NN-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 at (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 mean haversine distance, 3.9 km $\approx$ 2.5 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 7b and 7c).

For robustness, we further implement an **LSTM** model to predict the home location, both as a classification task and regression task. In classification, we pass a location visited by a consumer i through a LSTM encoder and train it to decode a 1 (yes, i's home) or 0 (not i's home). The performance is poor ($MAP@1 = 0.12$, $MAR@1 = 0.56$). In regression, the architecture is similar to that in the activity-timing prediction (Section 4.2.2). The predictive performance (an average of 4.2 km between the predicted and actual location locations) is inferior to the ensembled Random Forest model (4.0 kilometers) described in Section 4.1.1 and Section 5.1. While context specific, Random Forest in many contexts actually outperforms LSTM in predictive performance (Fernández-Delgado et al. 2014, Ahmad et al. 2017, Weinberg and Last 2019, Chen et al. 2019). LSTM is also more prone to over-fitting and more sensitive to input or different random weight initializations.

Furthermore, we perform an obfuscation via early stopping (Online Appendix A) using LSTM instead of Random Forest for the utility computation. Table 11 shows that the obfuscation scheme is capable of achieving a trade-off between the consumer risk for the varying levels of acceptable decrease in the utility, demonstrating the generalizability of the proposed approach. However, this trade-off is a bit worse when compared to Table 8.

Appendix E: Robustness of Training/Testing Sets Split

As detailed in Section 5.1, our T_{train} and T_{test} have overlapping time periods across the five weeks. This split was suggested by the data partner on how a data aggregator is envisioned to use the proposed framework. For instance, a data aggregator could perform the obfuscation on a small sample of consumers, decide the best p among the different choices available, and then perform the obfuscation with the chosen p for the entire sample before sharing the location data with an interested party.

In this Section, we also present the results when there are no overlapping periods across the five weeks of location data between the training and testing sets. Specifically, we use the locations from the 1st and 2nd week to train the model and then predict the activities over the 3rd week. Finally, the testing performance is measured by utilizing the 3rd and 4th weeks' location data to predict the 5th week's locations.

With the modified training/testing sets split, the NN-based algorithm achieves similar performance in terms of $MAP@k$ and $MAR@k$. This is expected since the features that go into the NN do not explicitly factor in the sequence of weeks. They only capture the aggregate weekly behavior. Hence, as long as there is no huge deviation between the week-to-week behavior, the features are likely to be predictive of the next week's behavior. The obfuscation results are thus similar to those in Figure 4.

However, we do notice a drop in the performance (about 10% decrease in both $MAP@k$ and $MAR@k$) in the activity-timing prediction (Section 4.2.2). This is expected since LSTM explicitly factors in the sequence of the location data. In Figure 8, we present the obfuscation trade-off with the varying levels of p with the newly trained LSTM model, with no overlapping time periods and the home inference risk. The proposed framework is still capable of providing a reasonable trade-off between the consumer risk and the advertiser utility.

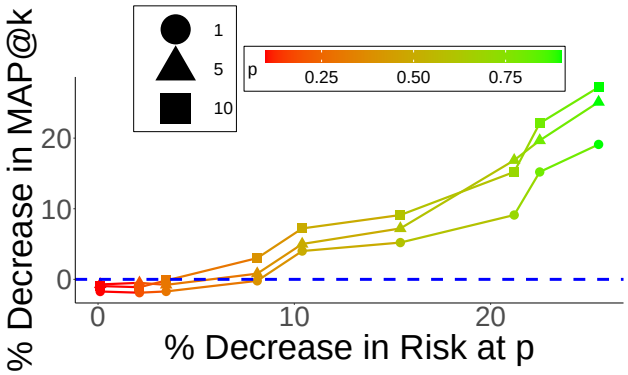


Figure 8 **Trade-off between home inference risk and activity-timing utility**

Appendix F: Cost-Benefit Analysis

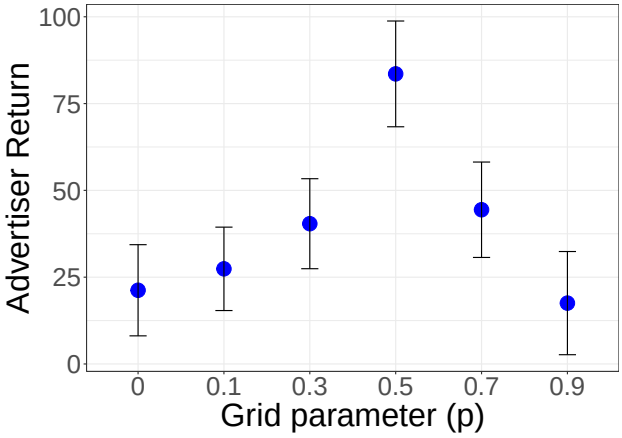


Figure 9 Proposed framework: cost-benefit analysis

We supplement the empirical analysis with a cost-benefit analysis, designed to closely simulate the application of the proposed framework in practice. Consider a consumer whose location data have been acquired by an advertiser from a data aggregator. Suppose that the advertiser engages POI recommendations and allocates a marketing budget in hope of a return. Based on the obfuscation parameter p employed by the data aggregator, the model accuracy (i.e. utility) and the risk would both change (Figure 4). In the simulation, we factor both into the desired return as follows: (1) the advertiser makes a return if the location prediction model is accurate in recommending a consumer's likely next location (i.e., generating a benefit from being a proper advertiser); and (2) the advertiser incurs a loss if the consumer perceives the targeting to be a privacy invasion (i.e., loss from being considered as a stalker). We model the return from an accurate prediction in (1) as a normal distribution, and perceived loss by the consumer in (2) based on the value function of the Prospect Theory (Kahneman and Tversky 1979). Formally, if the model makes an accurate prediction, the expected return from each consumer is modeled as $\mathcal{N}(\mu, \sigma^2)$ - $\lambda(-r_i)^v$, where r_i is consumer i's risk and μ, σ, λ, and v are the simulation parameters. If the targeting model makes an inaccurate prediction, the return is assumed to be zero.

Figure 9 displays the advertiser's returns and the corresponding standard errors based on 1,000 simulations. For each simulation, we sample 10,000 consumers and their corresponding trajectories. For each consumer trajectory, we infer the next likely location based on the location prediction model (Section 4.2) and the corresponding risk of home inference (Section 4.1). To imitate the obfuscation performed by the data aggregator before sharing the data, we also run the analyses for different specifications of p. We set parameters $\mu = 1$, $\sigma^2 = 0.05$, $\lambda = 1$, and $v = 0.6$. to compute the return from a consumer in an accurate prediction. The advertiser's return displays an increasing trend as p increases, peaking at $p = 0.5$ and then decreasing as p increases (Figure 9). Notably, the return at $p = 0.5$ is significantly higher when compared to no obfuscation ($p = 0$). The initial increase in the returns is attributed to the decrease in the risk r_i while maintaining a similar prediction accuracy accomplished by the proposed obfuscation scheme (Figure 4). After

$p = 0.5$, the model performance declines more, compared to the decrease in r_i, leading to lower returns. Overall, deploying the obfuscation scheme leads to higher returns to the advertiser, further incentivizing the data aggregator to obfuscate the data before sharing the data with the advertiser.

F.1. Grid Search versus Gradient Descent-based Search

Next, we supplement the cost-benefit analysis by formulating an objective function similar to the one presented in the obfuscation scheme discussed in Section 4.3, with several simplifying assumptions. We then use gradient descent in lieu of the proposed grid-based approach to maximize the expected return. Finally, we employ the early stopping based grid-based approach presented in Online Appendix A to compare with the expected return from the descent based approach.

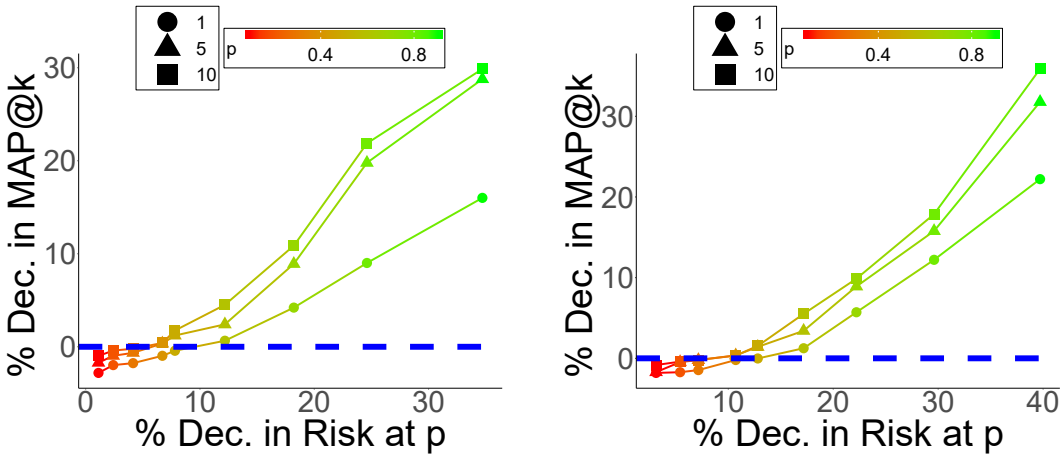
Formally, we have $I(p)\mathcal{N}(\mu, \sigma^2)$ - $\lambda(-r_i(p))^v$, where $r_i(p)$ is consumer i's risk based on the obfuscation p, $I(p)$ is an indicator of a correct prediction for a given utility, p, the obfuscation parameter. As we know, as p increases, $I(p)$ and $r_i(p)$ both decrease. To stay consistent with this monotonic property, we assume $I(p)$ as a $Beta(1,3)$ distribution. For $r_i(p)$, we consider linear $r_i(p) = r_i(1-p)$ and a concave $r_i(p) = r_i(1-p)^2$ functional formulations. Given these, we maximize the objective function using the standard gradient descent approach over the parameter p.

When compared to the early stopping grid-based approach, the minimum value attained is 5.42% lower (when r_i is linear), and 8.65% lower when $r_i(p)$ is convex. At the outset, this is expected given the grid-based approach is a heuristic in comparison to a structure descent based approach. However, the descent based approach comes with simplifying assumptions for $I(p)$ and $r_i(p)$ which heavily restrict the generalizability of such a obfuscation scheme – one of our key contributions in this work. Note that we present this analysis for illustrative purposes only. A more accurate way to capture such cost and benefit is by conducting counterfactual analysis in a structural model framework, wherein an agent (here, the data aggregator) maximizes its profit. We leave this for future research.

Appendix G: Robustness Tests and Additional Analyses

G.1. Suppression based on Recency and Time Spent

The proposed obfuscation scheme uses a structured grid search by varying the grid parameter p to identify the two consumer-specific parameters $\{\vec{s_i}, z_i\}$ that balance the risk and utility. Recall that $\vec{s_i}$ captures the informativeness of each location in T_i based on the visit frequency of the location (Figure 4). Here, we further augment the empirical study and showcase the scheme's flexibility by assigning $\vec{s_i}$ based on the recency and time spent at each location in T_i. For brevity, we only consider the home inference risk and visualize the risk-utility trade-off in Figures 10a and 10b. Similar to Figure 4, for a given % decrease in the risk, there is a lesser corresponding % decrease in the utility.



(a) Suppression based on recency (b) Suppression based on time spent

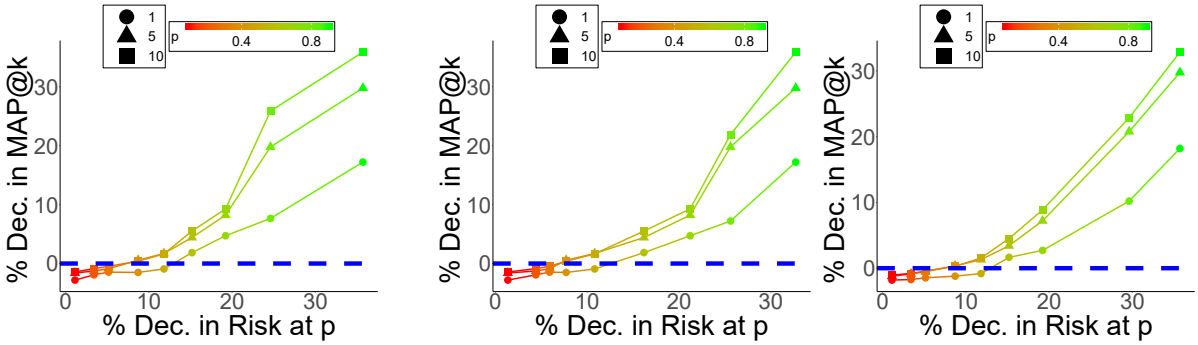
Figure 10 **Proposed framework for home inference risk, location prediction utility, and suppression based on recency and time spent**

G.2. Varying Sample Sizes

To verify the robustness of the results in Figure 4, we repeat the same empirical exercise on three random samples: 25%, 50% and 75% of the consumers in the data. For brevity, the suppression is performed only for the home inference risk based on the visit frequency of each location (similar to Figure 4). Figures 11a, 11b, and 11c show that even with smaller samples, the $\frac{Y}{X}$ slope (i.e., the % decrease in the utility divided by the % decrease in risk) at different values of p remains similar to that in the full sample (Figure 4a).

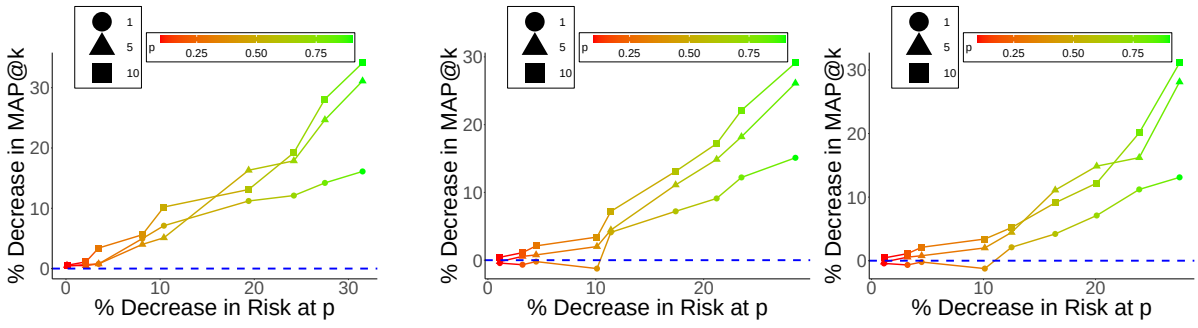
G.3. Varying Dimensionalities

We further demonstrate the framework's robustness by varying the dimensionality of the trajectories. For each trajectory, we perform 25%, 50%, or 75% truncation for each consumer each day, and repeat the empirical exercise. For brevity, the suppression is performed for the home inference risk based on the visit frequency of each location (similar to Figure 4). Figures 12a, 12b, and 12c suggest that the proposed framework performs reasonably well on the sparser dimensions (25% and 50%) and remains comparable to the full sample (Figure 4a) on the 75% of the trajectories considered.



(a) 25% of consumers (b) 50% of consumers (c) 75% of consumers

Figure 11 **Proposed framework for home inference risk, location prediction utility, and varying sample sizes**



(a) 25% of daily trajectories (b) 50% of daily trajectories (c) 75% of daily trajectories

Figure 12 **Proposed framework for home inference risk, location prediction utility, and varying dimensionality**

Appendix H: Rule-based Obfuscations

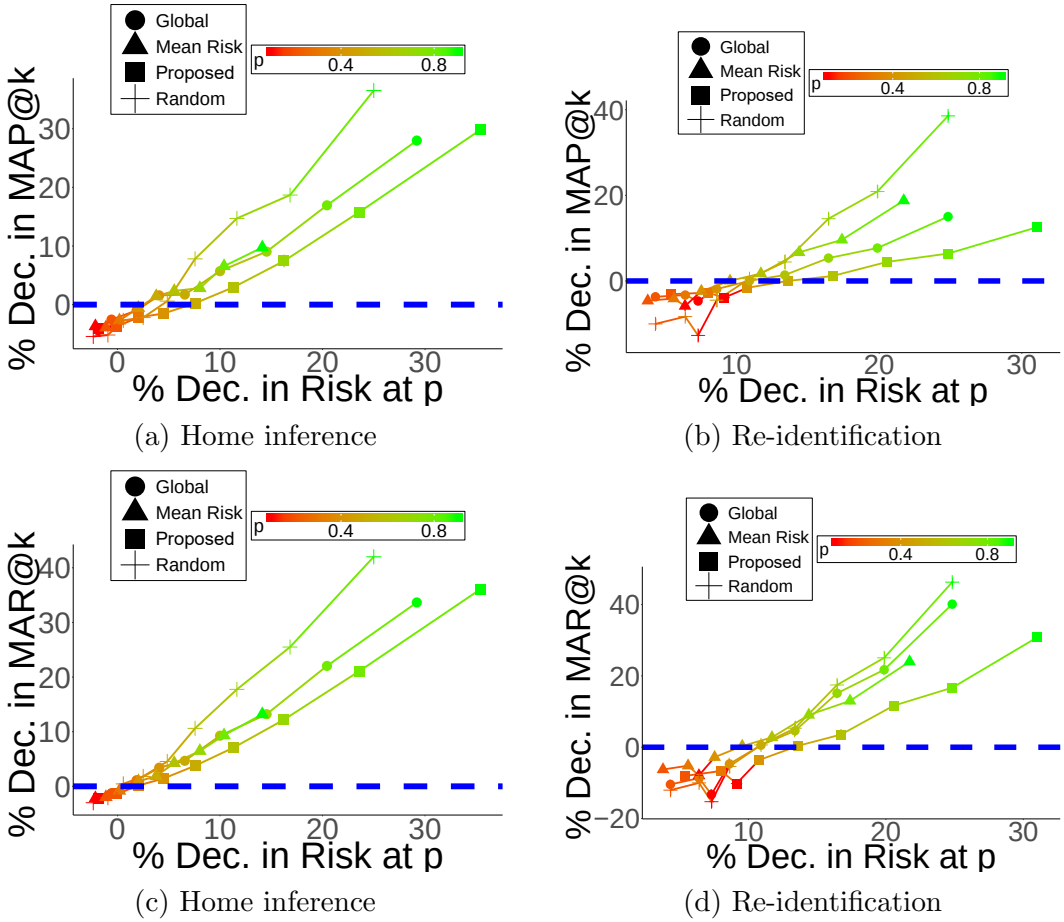
The first three baselines use practical rules of obfuscating locations with certain timestamps. Specifically, the data aggregator could remove all locations during the sleep hours (10 PM - 7 AM daily), sleep and work hours (9 AM - 6 PM on weekdays), or all timestamps of the entire data. These obfuscations would reduce the amount of information that can be extracted from the obfuscated data, and hence the data's utility to the advertiser. For instance, if all timestamps were removed, both the mobility features, `time_entropy`, `time_rog`, `average_dwell` (Table 4), and consumer-consumer, consumer-location affinity features based on time spent at a location cannot be computed. When compared to these baselines (Table 12), the proposed framework (Figure 4) offers a better choice set of the risk-utility trade-off. For instance, the proposed method can reduce the home inference risk by 15% with merely 1% decrease in the utility (Figure 4a, $p = 0.7$, $k = 1$), whereas for a comparable 13.45% decrease in the risk, removing all timestamps reduces the utility by as high as 33.16% (Table 12); conversely, for the same 33.16% decrease in the utility, the proposed method can reduce the risk by 34% (Figure 4a). Similar patterns are observed across all methods in Table 6 when compared to the proposed method (Figures 4b, 4d).

Obfuscation rule	**% decrease in home inference risk**	**% decrease in re-identification risk**	**% decrease in location prediction utility (MAP@1)**	**% decrease in location prediction utility (MAR@1)**
Remove sleep hours	2.43	1.41	11.83	12.69
Remove sleep and work hours	10.72	21.49	34.45	23.72
Remove all timestamps	13.45	0	33.16	32.97

Table 12 Alternative obfuscation schemes: rule-based obfuscation

Appendix I: Risk-based Obfuscations

We further compare the proposed method with three alternative risk-based obfuscations that also leverage the same mobility features, to demonstrate the advantages of *personalized* risk quantification and *personalized* suppression via the consumer-specific parameters $\{\vec{s_i}, r_i\}$. The first is a random baseline, where the suppression is performed randomly instead of based on consumer-level parameters, although the same number of locations are suppressed for comparability. The second is a mean-risk baseline, where the suppression is based on the mean risk without variations across consumers, and hence not personalized. That is, we replace r_i with $\bar{r} = \frac{1}{N}\sum_i r_i$ and suppress locations using $\bar{r}$, p and $\vec{s_i}$ for each T_i, as in most prior studies. The third is a global baseline, where all locations in T have the same chance (based on the mean risk) to be suppressed irrespective of the risk variations across consumers. For a given decrease in the risk (same X-value), the proposed obfuscation displays the least decrease in the utility across all risks (lowest Y-values in Figure 13). The random baseline performs the worst, justifying the need for the risk quantification either at a location-level (global baseline) or consumer-level (mean-risk baseline and proposed method).



(a) Home inference (b) Re-identification

(c) Home inference (d) Re-identification

Figure 13 **Proposed framework vs risk-based obfuscations:** $MAP@1$ **and** $MAR@1$