Gibbs et al. *EPJ Data Science* ( 2026) 15:17
https://doi.org/10.1140/epjds/s13688-025-00611-4

EPJ Data Science

RESEARCH Open Access

# Impact of federated data with local differential privacy for human mobility modeling

Hamish Gibbs[1,4*], Mirco Musolesi[2,3], James Cheshire[4*] and Rosalind M. Eggo[5]

Handling Editor: Rossano Schifanella

*Correspondence:
h.gibbs@northeastern.edu;
james.cheshire@ucl.ac.uk
[1]Network Science Institute, Northeastern University, Boston, USA
[4]Department of Geography, University College London, London, UK
Full list of author information is available at the end of the article

**Abstract**

With increasing awareness of the privacy risks posed by mobile phone location data, researchers need ways to use mobility data while offering stronger privacy guarantees to the individuals included in this data. A promising approach to this challenge is the creation of privacy-preserving mobility insights from decentralized location data using Local Differential Privacy (LDP). However, mobility data generated with LDP, based on the introduction of noise by individual mobile devices, is limited by the volume of noise required to achieve individual privacy. In this paper, we provide a fully reproducible model of the accuracy of mobility networks generated with LDP compared to mobility network data generated with more traditional privacy mechanisms: Central Differential Privacy (CDP) and K-anonymity. Using a simulated mobile phone mobility dataset informed by real-world travel patterns in the USA, we explore the trade-off between privacy and data utility provided by different parameters in a federated system with LDP. We also explore the impact of spatial and temporal aggregation on data accuracy, showing that long-standing considerations regarding the appropriate units of analysis for geographic data play a key role in determining the utility of federated mobility data with LDP. Our paper facilitates an in-depth understanding of the trade-offs between privacy and data utility entailed by the future adoption of a federated approach which uses LDP to generate insights from decentralized mobility data.

**Keywords:** Human mobility; Differential privacy; Local differential privacy; Ethics; Federated analytics; GPS; Call detail records

## 1 Introduction

Location data from mobile phones is a commonly used source of information on human mobility patterns. This data has been adopted in a wide range of disciplines including epidemiology [1, 2], natural disaster response [3], international development [4], and the study of urban systems [5, 6]. In these contexts, mobile phone mobility data provides valuable, near real-time information on dynamic patterns of human behavior. Common applications of mobility data include predicting the spread of infectious disease [7], identifying locations requiring humanitarian intervention following natural disasters [8], and measuring the effect of social inequalities in urban environments [9, 10]. The use of mobility data

from mobile phones accelerated during the COVID-19 pandemic, which saw the public release of aggregated indices of mobility activity produced by major technology platforms [11, 12]. In the rapidly evolving context of the pandemic, mobile phone data was one of the few available sources of information on near real-time population dynamics and was used as a proxy for rates of social contact [13], inputs to spatial models of infectious disease transmission [7], and was used to estimate the degree of adherence to non-pharmaceutical interventions such as travel restrictions and lockdowns [14, 15].

While mobility data provides clear value for researchers studying human behavior, this data can also pose significant risks to individual privacy. Individual mobility patterns are highly unique [16], and even incomplete information on individual movement can reveal sensitive characteristics such as a person's health conditions, sexual orientation, or religious affiliation [17–19]. Moreover, individual mobility trajectories remain highly identifiable even with the application of noise or coarse spatio-temporal aggregation intended to preserve individual privacy [16, 20]. The large-scale collection and use of mobile phone mobility data by commercial organizations, governments, and researchers, has increased concern about the privacy risks posed by large collections of detailed mobility trajectories, and the risk that even aggregate mobility measures could expose sensitive personal information [21, 22]. These data privacy concerns have fostered legal restrictions on the collection and processing of mobile phone location data, included in the EU General Data Protection Regulation (GDPR), the California Consumer Privacy Act (CCPA), and regulations by the US Federal Trade Commission, among others [23–25]. These regulations attempt to enforce a balance between the societal value of mobility and other forms of personal data, and the privacy risks that they pose to individuals.

While legal frameworks provide guidelines for processing sensitive data, there is significant interest in technical solutions which can achieve robust privacy protections for large-scale data and provide an up-front guarantee of privacy compliance within existing legal requirements. A number of technical solutions have been proposed, including Secure Multi-party Computation (SMPC) in which mobile devices collaborate to produce aggregate statistics without disclosing their individual data, typically using Homomorphic Encryption [26]; as well as Differential Privacy (DP), a formal technique for guaranteeing the privacy of aggregate statistics through the addition of calibrated random noise [27, 28]. In practice, because of its formal privacy guarantees and lower computational overhead, DP has been more widely adopted in the release of mobility datasets including urban mobility networks [29, 30], COVID-19 related activity indices from Google [31, 32], and indices of distances traveled by Facebook users [33]. DP can provide strong privacy guarantees for aggregate statistics, however, DP implementations still require the collection of individual mobility traces in a central database prior to aggregation and privatization, which can pose risks to the security of individual data.

Outside the analysis of human mobility, new technologies have sought to eliminate the need to collect potentially sensitive data from individuals before producing aggregate statistics [34, 35]. The emerging field of Federated Analytics (FA) with Local Differential Privacy (LDP) achieves stronger privacy protections over current DP approaches by processing individual data on the device where it was collected, and privatizing data with Differential Privacy before data sharing [36]. Privacy-preserving FA with LDP has already been implemented in real-world systems at the scale of millions of devices to identify high frequency words typed in mobile keyboards and record web domains which cause signif-

icant memory usage in web browsers [37]. While scalable LDP mechanisms exist, they can result in much higher noise required to achieve individual-level privacy compared to traditional DP. The balance between utility and data privacy is an active area of LDP research, with novel algorithms reducing noise by calibrating privacy mechanisms using publicly available information on domain specific data distributions [38], or by adaptively applying noise to specific data points depending on their sensitivity [39].

In principle, LDP mechanisms provide a promising potential solution to the challenge of collecting privatized mobility data from mobile phones, without disclosing individually identifiable location information. There is an increasingly pressing need to understand the significance of LDP mechanisms for human mobility researchers, in light of a trend towards the 'decentralization' of mobility data. For example, in December 2023 Google Maps announced that user location data would be stored on users' mobile devices by default [40]. This raises the question of how researchers will access mobility data for future research on human dynamics, and, for data generated using LDP mechanisms, how the noise introduced by LDP will impact the accuracy of future mobility statistics.

In this paper, we explore the creation of origin-destination (OD) matrices, a general purpose form of mobility data, which describes the total number of travelers between spatial regions, within a privacy-preserving federated system with LDP. We benchmark the accuracy of OD matrices generated with LDP against existing approaches to privacy-preserving human mobility modeling and analysis: Central Differential Privacy (CDP), which has increasingly been applied to aggregate mobility networks used in research, and K-anonymity, a well-established approach to preserving privacy through the suppression of low journey counts in mobility networks [41]. Our analysis aims to provide an open-source benchmark for researchers using data from a FA system with LDP who are curious about the impact of LDP techniques on the reliability of mobility data for downstream analysis. Our study is based on a simulated individual-level mobile phone dataset informed by empirical travel networks from the USA, producing county-to-county mobility networks similar to those used to model national-level disease transmission during the COVID-19 pandemic [42]. Our use of synthetic mobility data provides a unique opportunity to understand the impact of the full range of choices involved in the design of a privacy-preserving FA system with LDP, with full transparency into the impact that these have on the accuracy of resulting mobility data. This understanding will be crucial to future uses of mobility data, allowing researchers to understand the implications of the range of choices defining a given system for generating privacy-preserving insights from decentralized mobility data.

Our paper provides an overview of the simulation procedure used to create a fully reproducible dataset of individual-level mobility displacements, followed by a methodological summary of the different privacy techniques compared in this study. We indicate the parameters associated with each privacy technique, the architectural level at which privacy protection is applied, and the role of different parameters in each privacy technique. We then present analytical results comparing mobility networks made private using the three different privacy methods: K-anonymity, CDP, and LDP, followed by an in-depth analysis of the impact of different privacy parameters on error introduced by LDP. Finally, we explore the crucial role played by the choice of spatial and temporal units of aggregation for the accuracy of mobility insights made private using LDP.

## 2 Methods and data

### 2.1 Mobility data

This work is based on a simulated mobility dataset calibrated using empirical OD travel networks from the United States. To compare different privacy mechanisms in subsequent analysis, we aggregate the simulated individual mobility histories using different privacy mechanisms, and assess the correspondence between aggregated travel networks after privacy protection has been applied, and a mobility network without privacy protection.

We simulated a series of individual mobility trajectories based on a 2019 dataset produced by the US-based mobile phone mobility data aggregator SafeGraph (Fig. 1a) [43]. This dataset is based on the location histories of individuals using a panel of mobile applications, and is aggregated to describe the daily flow of individuals between US counties (Fig. 1b, c). To simulate individual-level travel, we used the Density-Exploration and Preferential Return (d-EPR) model (Fig. 1d) [44, 45], which combines a mechanistic description of individual mobility, with information on the population-level distribution of travel. This allowed individual-level simulations to be calibrated by empirical travel network data, increasing the realism of the individual mobility simulation. We fit the model of collective mobility to empirical data in a Bayesian framework, and performed sensitivity analyses, ensuring that the model of population-level travel accurately captured mobility patterns from the empirical travel network (Supplemental Sect. 1).

The main text of this paper shows analysis for April 8th, 2019 for the US Census-defined Middle Atlantic Division, which includes the states of New York, Pennsylvania, and New Jersey [46]. We perform a sensitivity analysis to identify possible variability in collective model predictions by assessing the accuracy of the collective mobility model informed by empirical mobility data for the first week of each month in 2019, as well as all US federal holidays, in four other regions of the US (Supplemental Sect. 2). Because the complexity

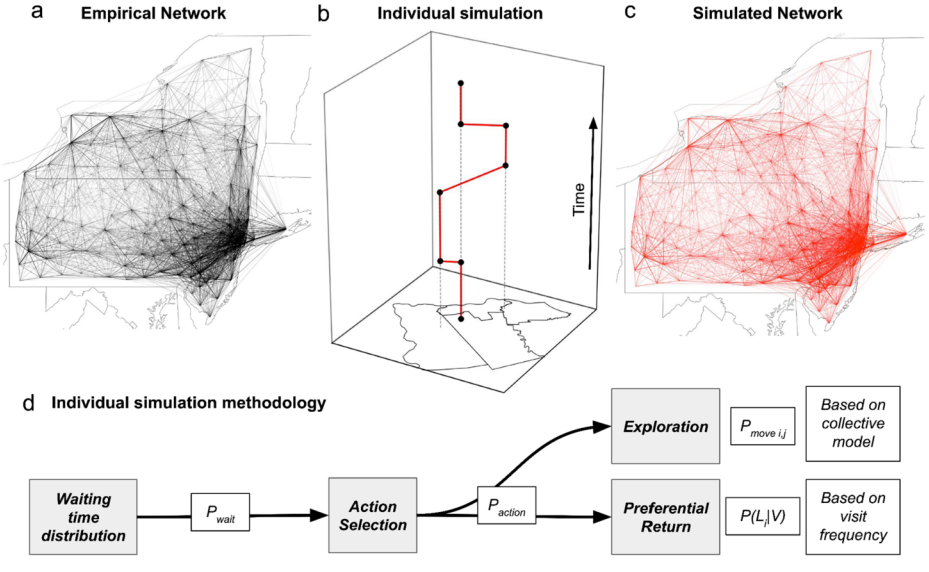


**Figure 1** *Diagram of individual mobility trajectory simulation procedure.* a) Empirical mobility data used to inform the trajectory simulation. b) Visualization of an individual simulation over time showing movements between counties around Philadelphia. c) Simulated individual trajectories, aggregated without privacy protection for comparison with empirical mobility data. d) A process diagram describing the construction of the d-EPR model of individual mobility

of our simulation model is sensitive to the number of counties in the simulation, we select these regions to maintain an approximately constant number of counties, while providing a diversity of spatial contexts and maintaining the geographic contiguity of states within each region.

## 2.2 Design of a federated analytics system

After simulating individual-level location histories, we construct a federated system for producing OD matrices. In this system, simulated individual location histories are collected and stored by individual mobile devices, producing a decentralized mobility dataset $D$, where each individual holds a subset of the dataset $D = \{D_i\}_{i=1}^{N_{clients}}$ where $N_{clients}$ is the number of simulated individuals.

In a federated system, a central server makes a query of the decentralized dataset (Fig. 2a). The query is a function which the central server sends to each mobile device. Mobile devices then compute the results of this query on their own dataset, and share these results with the central server, which aggregates the query results for all devices [35].

In the context of this study, the query requests the total count of OD pairs $f_{i,j}$ held by all mobile devices, providing the total volume of travel between regions for all devices participating in the federated system.

In a basic FA system, a query computes the true number of OD pairs across all devices, without privacy protection. This means that devices provide potentially sensitive location information to the central server, which then aggregates this information into a population-level OD matrix. Sharing detailed location information with the central server poses the obvious risk that the server could use this information to reveal sensitive information about individuals. However, even after aggregation, an OD matrix produced

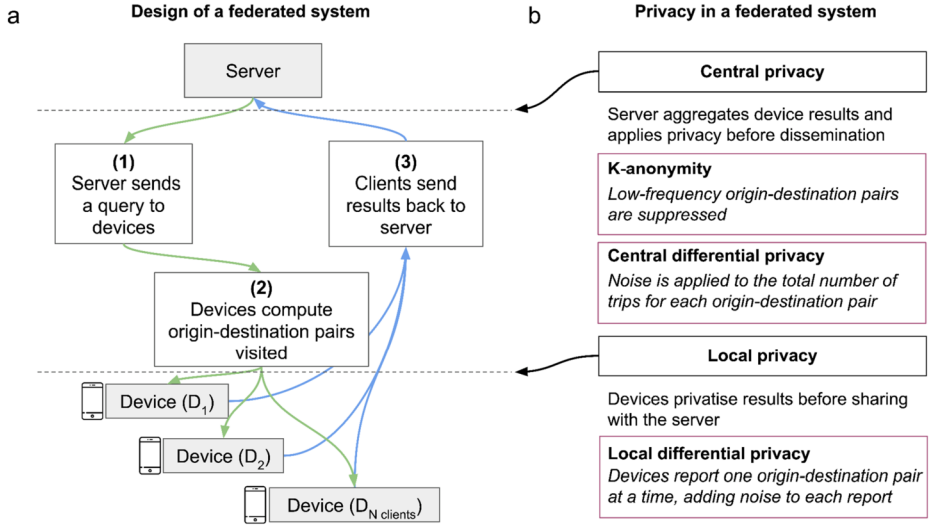


**Figure 2** *The generation of federated mobility data with different privacy mechanisms*. a) Overview of the communication steps in a federated system. Communication begins with a query by the central server, followed by the requested computation on individual mobile devices, and ending with the aggregation of individual results across all devices by the server. b) Different privacy mechanisms in a federated system. We compare two central privacy mechanisms introduced at the server level (K-anonymity, and CDP), and one local privacy mechanism introduced by individual mobile devices (LDP). Local privacy can provide stronger privacy guarantees compared to central privacy mechanisms because individual-level data is never shared beyond the device that collected it. For LDP, this comes at the cost of greater noise added to privatized data

**Table 1** *Parameter choices for different privacy-preserving FA systems*. A description of the different parameters required to define a federated system for mobility analytics using the three privacy-preservation techniques compared in this study

| System construction | Architectural level | Parameters |
|---|---|---|
| K-anonymity | Central Privacy | $K$: Anonymity threshold |
| CDP | Central Privacy | $s$: Sensitivity<br>$\varepsilon_c$: Privacy budget |
| CMS | Local Privacy | $d$: Number of hash functions (depth)<br>$w$: Hash size (width)<br>$\varepsilon_l$: Privacy budget<br>$u$: Number of user records shared with server |

without privacy protection has the potential to reveal sensitive information for devices with highly unique travel patterns [16, 47], and could permit Membership Inference Attacks [22], which seek to identify a specific individual's presence in the aggregated dataset. A privacy-preserving system for federated mobility analytics must prevent the disclosure of personally identifiable information, both for individual results shared by devices with the central server, or for aggregate statistics produced by the server.

Privacy protection can be introduced at two architectural levels of a federated system (Fig. 2b). Central privacy requires that devices report their individual OD pairs to the server, where these pairs are aggregated and privatized. By contrast, local privacy involves devices privatizing the precise OD pairs in their dataset prior to data sharing. We assess two central privacy mechanisms, K-anonymity and CDP, and one local privacy mechanism, LDP. For LDP, we use an existing implementation, the Private Count Mean Sketch algorithm (CMS), which is a general-purpose algorithm for federated frequency estimation with LDP [37]. We choose CDP and LDP mechanisms that provide representative examples of differential privacy algorithms used to achieve central and local privacy protection. In particular, CMS offers a representative example of an efficient LDP mechanism implemented at scale, which is appropriate for large input domains such as the edges in a mobility network [37]. Despite advances in LDP mechanisms in the literature [38, 39, 48], CMS remains actively studied [49, 50] and offers a baseline mechanism of comparison which is representative of LDP mechanisms which could be applied by organizations which produce mobility network data. In the Supplemental Information, we provide additional analysis comparing error distributions produced by CMS with Hadamard Response (HR), a more recent LDP mechanism which aims to reduce communication costs while maintaining LDP [51] (Supplemental Sect. 3). Table 1 provides an overview of the different parameters involved in each privacy mechanism, and the architectural level at which they are introduced in the federated system.

### 2.3 Low count suppression with K-anonymity

As a baseline for privacy protection in a federated system, we use a traditional approach, called K-anonymity, where the counts of OD pairs are suppressed below a chosen threshold $K$. K-anonymity achieves privacy for individuals by ensuring that a specific record in an aggregated dataset can be attributed to a minimum of $K$ individuals. In mobility research, $K$ is often chosen heuristically. Higher values of $K$ suppress more low-frequency edges, which is assumed to yield higher privacy. Based on convention in publicly available mobility datasets, we choose $K = 10$ [52]. It is important to note, however, that K-anonymity

is sufficient to protect privacy only in a dataset where individual devices contribute only one OD transfer. In the context of OD matrices, where individuals can make multiple inter-region transfers, the privacy guaranteed by K-anonymity is dependent on the scale of spatial and temporal aggregation, the number of records reported by each device, and the distribution of individual travel patterns [41]. For example, K-anonymous data could still reveal the identity of an individual making $N_{trips} > K$ between regions with no other travellers.

### 2.4 Central differential privacy

As an example of current state-of-the-art privacy protection applied to aggregated OD matrices, we use CDP, where calibrated random noise is added to aggregated OD travel counts to prevent disclosure of information on individuals' patterns of travel. Compared to K-anonymity, CDP allows for stronger privacy guarantees and a formalization of the trade-off between privacy and data utility. DP uses a randomized function $A$ to add noise to the records of a dataset $D$, with the intention of making it impossible to identify whether any individual record is included in the dataset. Formally, differential privacy aims to achieve *statistical indistinguishability* between two datasets $D_1$ and $D_2$ which differ by a single element, by adding noise to the dataset records according to a privacy budget $\varepsilon_c$ [28].

$$\frac{Pr\left[A(D_1) \in S\right]}{Pr\left[A(D_2) \in S\right]} \leq e_c^{\varepsilon} \tag{1}$$

The strength of the privacy protection provided by differential privacy depends on $\varepsilon_c$, where values of $\varepsilon_c$ closer to 0 provide higher levels of privacy protection by increasing the magnitude of noise added to the dataset. Differential privacy encapsulates the balance between data privacy and data accuracy. The noise introduced to ensure the privacy guarantee, as specified by $\varepsilon_c$, determines the precision of the differentially private data relative to the original dataset [27].

Standard implementations of differential privacy assume that individuals can contribute only one record to the privatized dataset. In the context of this study, mobile devices can produce multiple OD pairs, meaning that a single individual's dataset could produce multiple records in the aggregate dataset. We account for this possibility by restricting the total number of OD pairs a device can contribute to a fixed value $s$, and include this value in the privacy budget which defines noise added to aggregated counts of travel in the OD matrix. We add noise to the aggregated mobility dataset with CDP by perturbing the count of OD travel $f_{i,j}$ using Laplace-distributed random noise defined by a distribution mean $\mu$ ($\mu = 0$) and scale parameter $b$. The scale parameter $b$ is specified by a chosen privacy budget $\varepsilon$ and a sensitivity $s$, indicating the maximum variation of the output of the aggregation function that could be caused by a given input: $b = \frac{s}{\varepsilon_c}$.

$$f(x \mid \mu, b) = \frac{1}{2b} exp(-\frac{|x-\mu|}{b}) \tag{2}$$

Noise $\eta_{i,j}$ is applied independently to each aggregated count of OD travel $f_{i,j}$, producing a differentially private count $f'_{i,j}$.

$$f'_{i,j} = f_{i,j} + \eta_{i,j} \tag{3}$$

To adhere to the maximum number of trips that devices can contribute to the aggregate OD counts, individuals sample $s$ OD pairs in their dataset with uniform probability, then aggregate these counts before sharing with the server.

We note that in the DP literature, Laplace-distributed random noise is the standard approach used to achieve $\varepsilon$-DP in aggregate statistics. However, practical applications have used "approximate DP" (also known as $(\varepsilon, \delta)$-DP), a relaxed privacy definition which offers $\varepsilon$-DP with a small probability $\delta$ that the privacy mechanism will fail in extreme cases [53, 54]. In practice, $(\varepsilon, \delta)$-DP often relies on Gaussian, rather than Laplace-distributed random noise, and achieves greater utility at the cost of a small probability of privacy violation. As our study compares privacy mechanisms under fairly standard conditions and focused on a strict comparison of CDP and LDP, we choose a standard CDP formulation relying on a Laplace noise distribution.

### 2.5 Local differential privacy with CMS

The key to improving privacy guarantees to mobile devices participating in a federated system for generating mobility data is to achieve local privacy protection with LDP, where devices privatize the location information they share with the server. LDP means that devices can contribute to aggregate measures of mobility without sharing any personally identifiable location information with the central server. We apply LDP in the federated system for OD matrix generation using the CMS algorithm developed for frequency estimation with LDP [36, 37]. The CMS algorithm requires that devices share OD pairs from their local dataset with the central server one-at-a-time. To achieve local privacy protection, each device encodes the OD pair into a fixed-size representation using a hash function chosen at random from a set of $d$ three-wise independent hash functions. These hash functions encode the OD pair into a one-hot vector of size $w$. To achieve LDP, each element of this vector is independently flipped with probability $p = \frac{1}{e^{\varepsilon_l/2}+1}$ according to a given privacy budget $\varepsilon_l$. A fixed number $u$ of the encoded vectors are sampled with uniform probability from a device's dataset and shared with the central server, along with the selected hash function used to encode the OD pair. For a given OD pair, the server then computes an unbiased estimate of the frequency of travel by computing the average frequency of the hashed value across all hash functions selected by devices.

### 2.6 Quantifying the privacy-utility trade-off

We compute the empirical noise introduced by different privacy mechanisms by computing the difference between privatized and true counts of travel for each OD pair $i, j$.

$$\eta_{i,j} = f_{i,j} - f'_{i,j} \tag{4}$$

To assess the relative error of privatized counts, we compute the Absolute Percentage Error for each OD pair $i, j$.

$$Error(\%) = \frac{|\eta_{i,j}|}{f_{i,j}} \tag{5}$$

### 2.7 Sensitivity to privacy parameters: LDP with CMS

To understand the implications of parameter values defining the CMS algorithm, we conduct a sensitivity analysis of the empirical error distribution introduced by combinations

of the four CMS parameters: $\varepsilon_l$, $u$, $w$, $d$. $u$ defines the number of OD pairs shared by devices with the server, $w$ sets the size of the hash domain used to encode OD pairs in a fixed-size representation, $\varepsilon_l$ specifies the privacy budget of privatized hashes, and $d$ indicates the number of hashes used to privatize individual OD pairs, with hash functions chosen randomly by mobile devices. We choose four values of $\varepsilon_l = 0.1, 1, 5, 10$ and base the selection of parameter values for $u$, $w$, and $d$ on the characteristics of the simulated mobility dataset: $w$ is varied in proportion to the total number of OD pairs in the simulated dataset, while $d$ is varied in proportion to the number of clients in the simulated dataset. In our analysis: $u = 1, 2, 5, 10$, $w/N_{OD} = 0.005, 0.01, 0.1, 0.5$, and $d/N_{clients} = 0.0001, 0.001, 0.01, 0.1$. Based on the $N_{clients}$ and $N_{OD}$ presented in the main text of this paper, this produces $d = 20, 205, 2056, 20{,}561$ and $w = 117, 234, 2340, 11{,}704$.

To understand the relative contribution of each parameter to the error distribution of privatized OD counts, we perform a multiple linear regression relating the standard deviation of the error distribution $sd(\eta)$ across all OD pairs $i$, $j$ for a given combination of CMS parameters.

$$sd(\eta) \sim a + \beta_0 \cdot \varepsilon_l + \beta_1 \cdot u + \beta_2 \cdot w + \beta_3 \cdot d \tag{6}$$

We then compute the partial $R^2$ for each parameter, measuring the relative importance of different parameter choices on the empirical error distribution. We also compute 95% confidence intervals around partial $R^2$ estimates using bootstrap sampling with 1000 iterations.

To ensure the generalizability of our results outside of the simulated dataset used in this study, we repeat our analysis using real-world trajectory data produced by Yahoo Japan Corporation [55] to study the effect of CMS parameters on noise distributions required to achieve LDP (Supplemental Sect. 4).

### 2.8 The effect of spatial and temporal scale

In a federated system for generating OD matrices with LDP, noise is defined by the privacy parameters governing the privacy mechanism, in this case CMS. However, the relative error of privatized OD counts is defined by the magnitude of the count. One of the key decisions in the design of a federated system for generating mobility data is the selection of units of spatial and temporal aggregation, which can play a significant role in defining the relative error of OD counts in the system. We test the importance of spatial and temporal aggregation for data accuracy by producing OD matrices from a selection of spatial tessellations with decreasing spatial granularity, and for increasing temporal periods (ranging from 1 – 7 days).

We define a series of spatial tessellations based on hierarchical aggregations of counties into regions of decreasing spatial granularity (Supplemental Fig. 9). To achieve these tessellations, we perform hierarchical agglomerative clustering using Ward's method based on a distance matrix describing the distance between county centroids [56, 57]. There are 150 counties in the Middle Atlantic Division presented in the main text of this paper. We therefore select tessellations with between 145 clusters (high spatial granularity) and 15 clusters (low spatial granularity) with a step size of 10 (Supplemental Fig. 10). This produces a series of spatial tessellations with an average spatial area between 1980 $km^2$ and 18,850 $km^2$ (Supplemental Fig. 11).

We aggregate OD pairs for all combinations of spatial tessellation and time period, and compute the differentially private count of travelers between OD pairs using the CMS algorithm with $\varepsilon_l = 5$, $u = 1000$, $d = 0.1 \cdot N_{clients}$, and $w = 0.5 \cdot N_{OD}$.

## 3 Results

### 3.1 Quantifying error for different privacy mechanisms

We compare the quality of mobility data generated in a theoretical privacy-preserving federated system by calculating the error of privatized counts relative to true counts of travel between OD pairs for different privacy mechanisms providing central and local privacy. The ultimate goal of a federated system for generating mobility data is to achieve strong privacy protection with local privacy, while maintaining acceptable levels of accuracy in the resulting mobility data. Figure 3a shows the OD network privatized with K-anonymity with a suppression threshold $K = 10$. K-anonymity is a traditional approach to privacy protection in OD matrices. In the K-anonymous OD network, 39.52% of OD pairs have travel volumes above $K$, representing 98.58% of the total travel volume in the network.

Figure 3b shows an example of CDP, with privacy protection parameters $\varepsilon_c = 1$ and $s = 10$. CDP privatizes OD counts by applying random noise to travel volumes aggregated in the central server. Based on a theoretical definition of "acceptable" error, with error $\leq$10% for privatized values relative to true values, 19.19% of OD pairs representing 92.47% of the total travel in the network have error $\leq$10% in the CDP-privatized network. We also assess the CMS algorithm, a privacy mechanism which achieves LDP, where devices privatize individual OD pairs before they are shared with the aggregation server. While this privacy mechanism achieves strong individual privacy guarantees, it also requires the addition of greater noise compared to the CDP mechanism. The CMS-privatized mech-

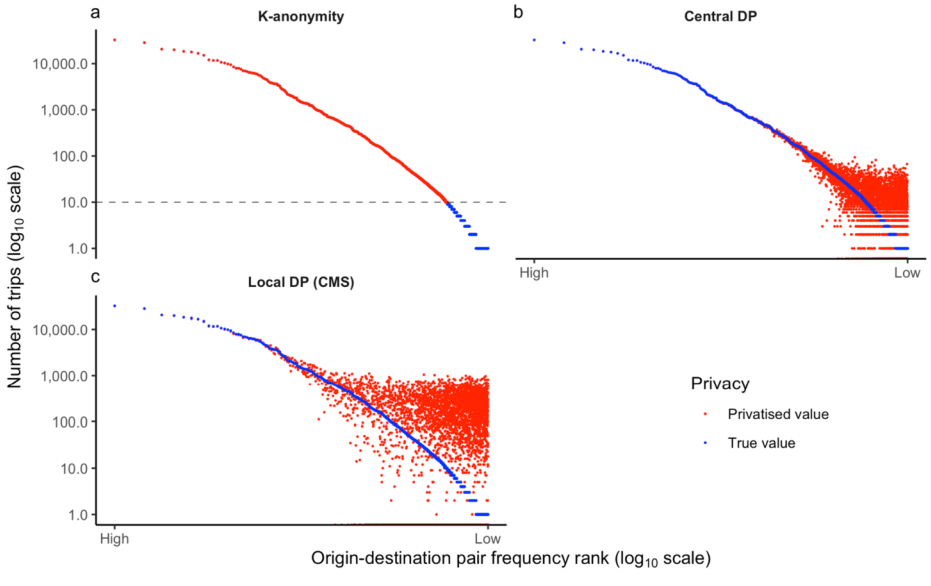


**Figure 3** *Comparison of mobility analytics computed with different privacy mechanisms.* Zipf's plots comparing mobility analytics computed without privacy-preservation (blue) to FA computed with different privacy mechanisms (red): a) K-anonymity with threshold $K = 10$, b) Central differential privacy ($\varepsilon_c = 1$, $s = 10$), c) LDP with the CMS algorithm ($w = 2340$, $d = 205$, $\varepsilon_l = 5$, $u = 10$). Plots show error introduced by different privacy mechanisms, with higher error introduced by the CMS algorithm. All negative values have been truncated to 0. OD pairs are ranked by travel volume in the non-privacy preserving dataset

anism with parameters $\varepsilon_l = 5$, $u = 10$, $d = 205$, and $w = 2340$ produces only 1.76% of OD pairs (56.14% of total network travel) with $\leq$10% error compared to true OD counts. Comparison of the CDP and CMS privacy mechanisms demonstrates the key challenge in improving privacy guarantees through the use of LDP, namely, LDP introduces significantly more noise compared to a central privacy model, even for higher privacy budgets $\varepsilon_l$. These results are supported by an additional sensitivity analysis comparing CMS with another LDP mechanism: HR. Our analysis shows a similar degree of noise required to achieve LDP using HR, when compared to CMS for a given privacy budget (Supplemental Sect. 3).

### 3.2 The challenge of local differential privacy for mobility analytics

Figure 3b, c highlight a key aspect of differential privacy. With differentially private mechanisms, noise is introduced to privatize OD counts independent of the total volume of travel for a given OD pair. While the noise distribution is constant across all OD pairs in the travel network, the CMS algorithm (with $\varepsilon_l = 5$, $u = 10$, $d = 205$, and $w = 2340$) introduces noise with an average absolute error roughly 23$\times$ greater than the error of the CDP mechanism (with $\varepsilon_c = 1$ and $s = 10$) (Fig. 4a). In the context of a mobility network, where true travel volumes are highly skewed, the noise added to achieve differential privacy produces a very large error for low-frequency journeys, but can achieve relatively high accuracy ($\pm$10% of true values) for high frequency journeys. Understanding the magnitude-independent nature of noise in the differential privacy mechanisms, and the empirical noise distribution produced by a given mechanism and associated parameters, allows for the estimation of the uncertainty distribution associated with a given true volume of OD travel (Fig. 4b). Here, we compute the absolute percentage error for each true OD travel count in the underlying travel network, and show the difference in uncertainty introduced by CDP, compared to LDP.

Given that the reported volume of travel only has error $\leq$10% for the highest frequency network connections, a federated system with LDP may severely limit the forms of mobil-

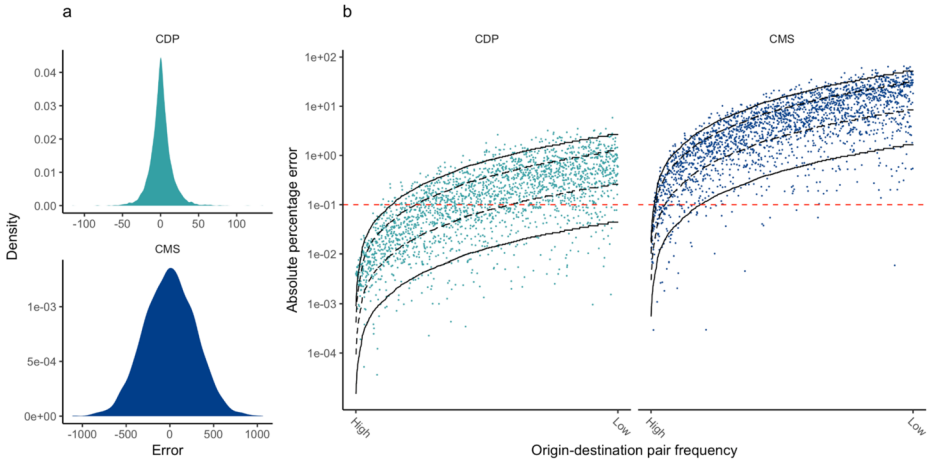


**Figure 4** *Error distributions for differential privacy mechanisms.* a) The distribution of error introduced by noise added to counts of OD travel privatized with the CDP and CMS mechanisms. b) Absolute percentage error across the OD frequency distribution. Solid lines show 90% density based on empirical noise distribution, dashed lines show 50% density. Note that the frequency distribution in panel b shows OD pairs with travel volume >10 for clarity. Red dashed line shows a theoretical threshold of $\leq$10% error for privatized values relative to true values

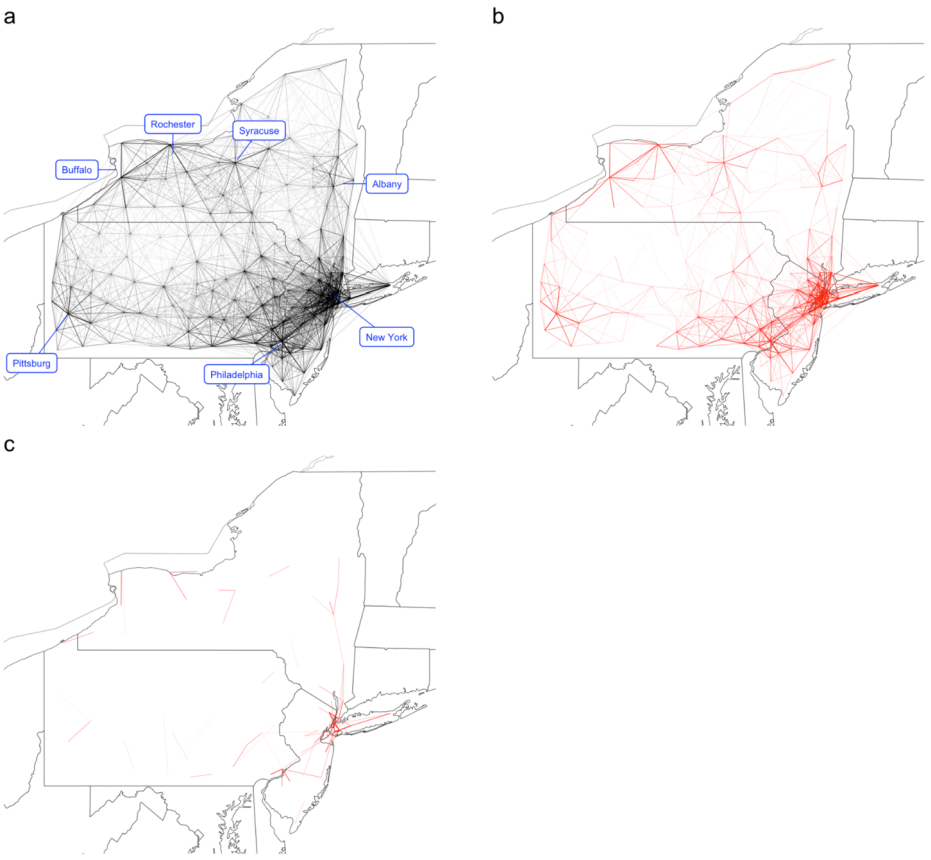


**Figure 5** *Error in journeys produced by different privacy mechanisms*. a) The original travel network. Blue labels indicate major cities. b) Travel network produced with CDP ($\varepsilon_c = 1$, $s = 10$), showing the privatized OD pairs with error $\leq$10% (red). c) Travel network produced with CMS ($w = 2340$, $d = 205$, $\varepsilon_l = 5$, $u = 10$), showing the privatized OD pairs with error $\leq$10%. More high accuracy connections (red) indicate a higher accuracy in differentially private counts of OD travel

ity analysis that can be conducted. Based on an error threshold of $\leq$10%, the CDP mechanism retains a significant proportion of the OD pairs in the original network (Fig. 5a), while the CMS mechanism only retains the central, high-frequency network connections around major urban centers (Fig. 5b). It is important to note however, that the low accuracy of OD counts in the CMS network presented in Fig. 5b is closely connected to the privacy parameters defining the mechanism, the magnitude of the original OD travel counts, and the units of spatial and temporal aggregation defining the travel network. We examine the role of each of these factors in defining the accuracy of OD counts with LDP in the following sections.

### 3.3 Impact of parameter choices on data utility with local differential privacy

Although Fig. 5c shows extremely low utility of OD data produced by the CMS mechanism, this error is a function of the chosen privacy budget $\varepsilon_l$, which defines a given implementation of the CMS mechanism, as well as the other privacy-related parameters $w$, $d$, $u$. Each of these parameters plays a role in the accuracy of federated OD counts privatized with LDP. To understand the relative importance of each parameter for defining the error distribution of the CMS mechanism, we compute the partial $R^2$ for a range of parameter

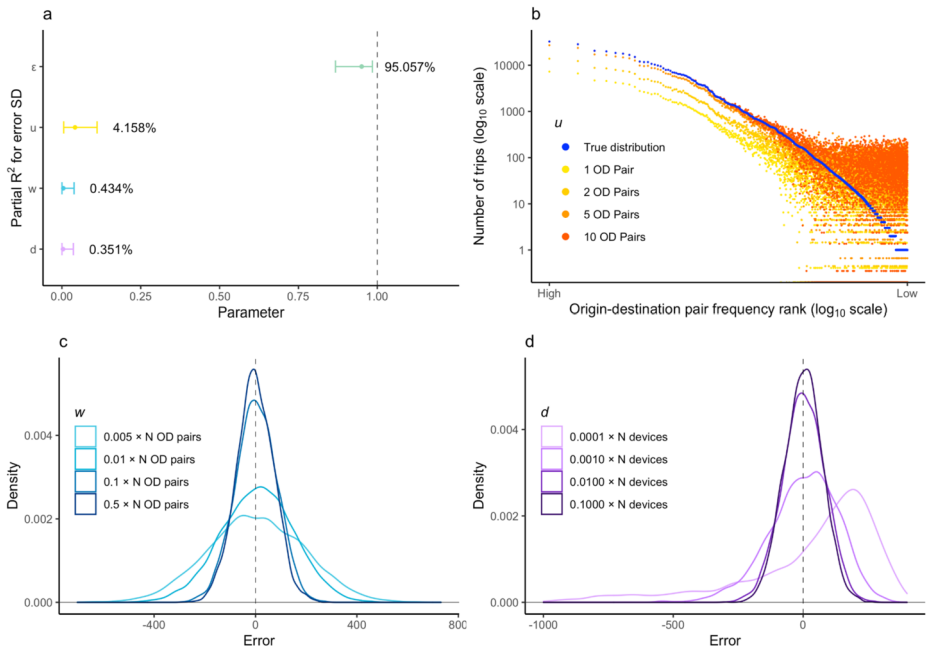


**Figure 6** *The impact of CMS parameters on data accuracy*. a) Partial $R^2$ showing the relative importance of different CMS parameters for defining the standard deviation of the error distribution of privatized values relative to true values. b) Systematic underestimation of travel volumes is introduced by low $u$, which defines the number of OD pairs shared by devices with the server. Other parameters: $\varepsilon_l = 10$, $w/N_{OD} = 0.01$, and $d/N_{clients} = 0.1$. c) The error distribution produced by different values of $w$, where low $w$ increases the probability of hash collisions, resulting in higher noise. Other parameters: $\varepsilon_l = 10$, $u = 10$, and $d/N_{clients} = 0.1$. d) The error distribution produced by different values of $d$. Low values of $d$ reduce the number of hash functions used by devices to privatize OD pairs, resulting in biased estimates of travel volume. Other parameters: $\varepsilon_l = 10$, $u = 10$, and $w/N_{OD} = 0.01$

values based on a multiple regression describing the relationship between different parameter combinations and the standard deviation of the empirical error distribution. This shows the dominant role played by $\varepsilon_l$ in defining the error distribution of CMS-privatized counts, where $\varepsilon_l$ accounts for 95% of the variation in $R^2$ for a model of the standard deviation of the error distribution given a selection of CMS parameter combinations (Fig. 6a).

While $\varepsilon_l$ defines the magnitude of the noise introduced by the CMS algorithm, the other parameters defining the mechanism each introduce different classes of uncertainty. For example, variations in $u$, the total number of OD pairs shared by mobile devices, result in systematic underestimation of the true OD travel volumes for low $u$ (Fig. 6b). By contrast, variations in $w$, which defines the size of the representation into which OD pairs are encoded before being privatized by individual devices, can markedly increase the variance of errors for $w$ (Fig. 6c). This is because of an increase in the probability of hash collisions, where different OD pairs produce the same encoded representation. Finally, $d$, which defines the number of hash functions selected by mobile devices to privatize their OD pairs, effectively represents the number of groups from which an unbiased estimate of OD pair frequency is computed. Systematic error can be introduced for low values of $d$, because of the higher probability of hash collisions occurring for a smaller number of hash functions (Fig. 6d). We confirm the directional effect of CMS parameters in a sensitivity analysis using a real-world trajectory dataset from Japan, which shows a similar

effect of privacy budget $\varepsilon_l$, and user reports $u$ on error distributions produced by LDP (Supplemental Sect. 4).

### 3.4 Impact of spatial and temporal aggregation

A final consideration in the design of a federated system for generating OD matrices, independent from the choice of privacy mechanism and associated parameters, is the scale of spatial and temporal units used for aggregation. In a K-anonymous system, privacy guarantees are closely related to the choice of the spatial and temporal scale of aggregation. In such a system, records of OD travel over a sufficient time duration with high spatial granularity could reveal information about individuals with unique travel behaviors. By contrast, in a differentially private system, data utility, instead of data privacy, is directly related to the choice of spatial and temporal units of aggregation. Choosing highly granular units can result in a decrease in data utility caused by the statistical noise required to guarantee individual privacy.

We compare the relationship between data utility and spatiotemporal scales of aggregation with a comparison of OD matrices aggregated into different spatial and temporal units (Fig. 7a, Supplemental Sect. 5). We compare the Absolute Percentage Error between true counts and counts privatized with the CMS mechanism, for OD matrices produced with increasingly coarse spatial tessellations, and time periods of increasing length (Fig. 7b). For comparison across spatial and temporal scales, we select OD pairs in the top 5% of

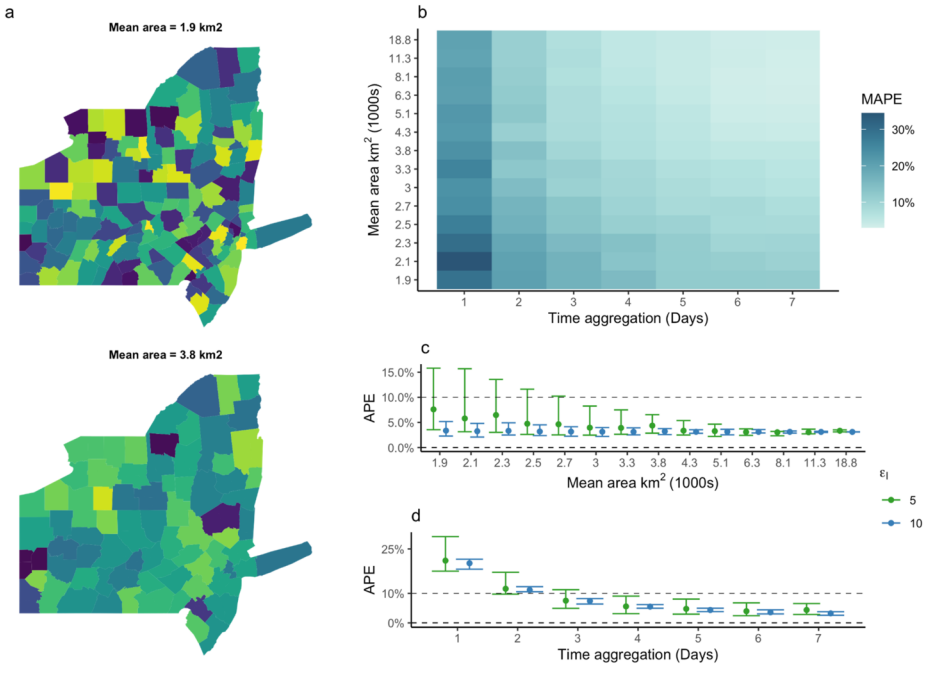


**Figure 7** *Impact of spatial and temporal aggregation on data utility.* a) Example of hierarchically aggregated spatial units exemplifying different scales of spatial aggregation. b) MAPE for OD matrices privatized with the CMS mechanism (with $\varepsilon_l = 5$, $d/N_{clients} = 0.1$, and $w/N_{OD} = 0.5$). c) Distribution of absolute percentage error for spatial tessellations of varying granularity, with a time aggregation of 7 days and $\varepsilon_l = 5, 10$. d) Distribution of absolute percentage error for different time windows for spatial tessellation with mean area = 3800 km$^2$ and $\varepsilon_l = 5, 10$. Plots show the top 5% of OD pairs, ranked by true travel frequency, for all spatial and temporal aggregations

travel volume across all spatial and temporal aggregations. Comparison of Mean Absolute Percentage Error (MAPE) for counts privatized with the CMS algorithm (with $\varepsilon_l = 5$, $d/N_{clients} = 0.1$, and $w/N_{OD} = 0.5$) shows a decrease in error caused by increasing the size of either spatial or temporal units of aggregation. Changes in spatiotemporal scale increase the total volume of travel along a given network connection, thereby reducing the error introduced by privacy-preserving noise. While this shows the role played by spatiotemporal aggregation, data utility also depends on the other privacy parameters defining a system. The effect of changing values of $\varepsilon_l$, for example, have a greater effect on data utility than any alteration of spatial resolution (with time resolution of 7 days), or temporal resolution (with average spatial resolution of regions equal to 3800 $km^2$) (Fig. 7c, d).

## 4 Discussion

This paper explores the feasibility of a federated system with LDP for generating accurate privacy-preserving OD matrices. Our study provides a unique opportunity to empirically quantify the error introduced by different choices in the design of a federated system for generating OD matrices, from the choice of privacy mechanism and associated parameters, to the more subtle impact of choices regarding the spatial and temporal units of aggregation into which underlying location data is aggregated. Overall, our study highlights the challenge of achieving privacy protection using LDP, due to the noise required to privatize individual OD pairs shared by mobile devices with a central server. We have shown how an LDP mechanism, CMS (with parameters $w = 2340$, $d = 205$, $\varepsilon_l = 5$, $u = 10$), produces error significantly above a theoretically "acceptable" level of accuracy of $\leq$10% error for most OD pairs in the travel network.

While our chosen definition of the specific "acceptable" level of error is contingent on a given application of mobility data, our choice reflects a permissive upper bounds on privacy-related error introduced by mechanisms in the DP literature [58, 59]. It is also important to note that, while measuring error provides a general indication of the degree of bias introduced by LDP, future research could explore the implications of LDP on specific applications such as infectious disease or transit modelling. The impact of LDP will vary depending on the application of mobility data. In infectious disease modelling, for example, significant error in low frequency network edges may alter predictions of infection dynamics for transmission across long distance edges. By contrast, in transit modelling, error in high frequency edges could significantly impact predicted estimates of network traffic as these edges are responsible for a disproportionate volume of network traffic.

In our analysis, we show that the relative error introduced by LDP for a specific network edge is dependent on the distribution of trips in the specific dataset we analyze. However, our simulation procedure, relying on the well-established d-EPR model of individual mobility [44] informed by a Gravity Model parameterized using real-world mobility networks [45], reproduces common patterns of subnational mobility networks including a heavy-tailed distribution of edge weights and distance-related decay of mobility flows for long distance edges [60]. These are well established features of mobility networks across spatial contexts and increase our confidence that the results of this analysis, namely that only OD pairs for the highest frequency travel network connections around major urban areas have acceptable data accuracy, will be generalizable to other sources of mobility data. To ensure that our findings generalize beyond the specific simulated dataset presented in our paper, we conduct a sensitivity analysis using real-world trajectory data aggregated

to a 500 m $\times$ 500 m grid which shows a similar frequency distribution observed in the simulated dataset, and confirms the role played by different CMS parameters on noise distributions required to achieve LDP (Supplemental Sect. 4).

Although our initial comparison of LDP with other privacy mechanisms underscores the challenge of achieving privacy-preserving OD matrices with LDP, we have also identified a number of opportunities to increase the accuracy of generated mobility data. The key insight regarding frequency estimation with CMS is that the noise added by a differential privacy mechanism is independent of the magnitude of the value being privatized. This points directly towards opportunities to increase the accuracy of mobility data by decreasing the noise introduced by CMS *relative to the size of the values being privatized*. This insight has encouraged research into adaptive noise mechanisms which unevenly relax privacy requirements based on context [39]. In OD matrices, adaptive LDP mechanisms could provide greater utility by emphasizing specific network connections, such as low-frequency edges, with higher accuracy. Researchers have also developed Robust LDP mechanisms which guarantee formal $\varepsilon$-privacy over a set of plausible data distributions estimated from publicly available information [38]. Although there are no widely accepted mechanisms for Robust LDP at scale, this approach is especially promising for mobility networks given the substantial quantity availability of public mobility data which could inform domain specific privacy mechanisms.

In this paper, we have also highlighted alternative approaches to improving data utility for a federated system with LDP without altering the inherent privacy provided by the CMS mechanism. Designers of a federated system for generating mobility data can make careful choices regarding the units of spatial and temporal aggregation defining the generated mobility data, leading to higher values with lower relative noise introduced by LDP. Like changes to the privacy budget $\varepsilon$, a system could theoretically maximize the volume of OD travel by choosing extremely large spatial or temporal units of analysis, thereby sacrificing data utility. In practice, there are opportunities to choose intermediate spatial and temporal units which achieve higher data accuracy while still permitting useful applications of the resulting mobility data.

Because data distributions interact with a given privacy mechanism and parameter set to define the utility of data generated by a federated system using LDP, choosing parameters for an LDP mechanism is challenging in practice. Moreover, there is no clear means to optimize parameters such as the privacy budget $\varepsilon$, which must be set according to normative definitions of privacy. Typically, parameters are chosen heuristically to manage computational cost while maintaining a given privacy budget. In this study, we perform a grid search over CMS parameters where specific parameter values are chosen as a function of the size of the simulated mobility dataset. This offers clear intuition regarding the impact of parameters $u$, $w$, and $d$ which have been tightly calibrated to the number of OD pairs and number of devices in the simulated mobility network. In practice, privacy parameters can be chosen according to intuition regarding the size of the data domain and number of users, an approach which is supported in existing literature [36, 37]. We show, for example, that a choice of $w$ that is 0.5 times the domain size (number of OD pairs) provides low variance estimates of mobility frequency, and similarly, that $d$ approaching 0.1 times the number of devices offers unbiased estimates of mobility frequency. We note that there are also diminishing returns, where increases in $w$ and $d$ do not significantly increase accuracy, but will still increase computational costs. Using the approach presented

in this study, practitioners using data from a federated system with LDP can understand the empirical impact of parameter choices, and how these choices interact with the underlying distribution of location data within chosen units of aggregation. For designers of a federated system with LDP, this type of simulation analysis can provide insight into the minimum acceptable values of privacy parameters like $w$ and $d$ required to maximize data utility, and permits a systematic way to assess different units of spatiotemporal aggregation while avoiding the need to rely on sensitive individual-level location data.

We have also shown the importance of parameters besides privacy budget $\varepsilon_l$ for improving the accuracy of estimates of travel in a federated system with LDP. We have assessed the impact of parameters which define the CMS mechanism, $u$, $w$, $d$, but note that this mechanism has been chosen as a representative example of a class of LDP algorithms which aim to achieve privacy-preserving frequency estimation with LDP [61]. In the supplement, we provide a sensitivity analysis comparing CMS with HR, which shows similar error distributions required to achieve LDP. Although CMS is implemented in commercial systems applying LDP at scale, recent research has shown that CMS does not provide comprehensive privacy guarantees and is subject to attacks which may infer sensitive characteristics of individuals [49]. More broadly, researchers have demonstrated that LDP mechanisms such as CMS are vulnerable to 'poisoning' attacks, where an attacker may manipulate individual data reports to alter aggregate statistics [62].

As the trend towards the decentralization of mobile phone location data continues to develop, there will be a pressing need to audit the accuracy and limitations of novel forms of mobility data [63]. The societal benefit provided by mobility data in applications ranging from pandemic response, transit modelling, international development, and natural disaster management will continue to drive demand for empirical insights derived from mobility data. In the use of mobility data for these applications, questions regarding the appropriate balance between the accuracy of mobility data insights and individual rights to privacy will remain, and will inform the adoption and specific calibration of technical privacy mechanisms. Technical privacy solutions like LDP, while they may help to ensure compliance with privacy regulations, do not guarantee ethical use of mobility data. Individual comfort participating with a given privacy-preserving system, and specific legislation governing the appropriate processing of individual data will be key to informing the design of future systems for mobility data collection which rely on LDP.

Federation will also raise new challenges in the use of mobile phone location data, as the requirement that analytics are computed on mobile devices shifts computational burden onto "edge" devices with fewer computational resources. LDP mechanisms are also more computationally intensive compared to k-anonymity or CDP which essentially entail an aggregation and filtering operation (as well as the application of random noise for CDP), which are standard data analytic procedures with trivial cost. By contrast, a federated system with LDP requires computation by edge devices to produce private reports, network communication to transmit these reports to a central server, and server-side computation to recover approximate frequencies from the private reports. Though this is significantly more computationally intensive than K-anonymity or CDP approaches, the existence of LDP systems operating at scale in commercial applications, like the CMS algorithm used in this study, indicates that the computational cost is not prohibitive. In more complex domains, on-device computational resources can be a key limitation reducing the number of devices eligible for participation in a federated system. This is a particular concern in

the related domain of federated learning, where model complexity can reduce the pool of devices available for model training and evaluation [63].

In this paper, we have addressed the accuracy of mobility data generated with different privacy mechanisms, rather than the privacy provided by a given mechanism and associated parameters. Although there are well-established techniques to establish the probability of a privacy failure for a given privacy model, there remains significant uncertainty regarding the normative definition of "acceptable privacy." Such a definition involves translating personal opinions, legal requirements, and commercial interest in preventing data misuse into specific parameters which will define a chosen privacy model. There is ongoing academic debate over the appropriate parameter values defining current implementations of central differential privacy [21, 29, 64], and in a future federated system for generating mobility data, the choice of privacy parameters must be transparent and interpretable by the individuals contributing location data to the system.

We have quantified data utility primarily in terms of the deviation of differentially-private analytics from true values in analytics produced without privacy protection. While this approach indicates areas of relative error, it is important to note that the level of acceptable error in mobility analytics is closely related to the chosen domain in which these analytics will be used. There is significant opportunity for future research to better understand the implications of approximate data produced with differential privacy on different domains of public health emergency response.

## 5 Conclusions

This paper has provided a detailed exploration of the feasibility of a federated system for producing accurate human mobility analytics with a selection of privacy mechanisms. Moreover, our analysis is fully reproducible, relying on simulated mobility data informed by publicly-available mobility networks. Our findings show that CDP provides a scalable and accurate means of producing OD matrices from decentralized mobility data, in line with existing implementations. We have also found that mechanisms which provide strong privacy protection through LDP are scalable, but their utility for the generation of OD matrices is restricted to a relatively small, high frequency portion of the aggregate travel network. This study contributes to our understanding of the empirical error introduced by a LDP mechanism in a realistic mobility network, and provides a general mechanism to assess the data utility in future federated systems for mobility data generation with LDP.

## Supplementary information

**Supplementary information** accompanies this paper at https://doi.org/10.1140/epjds/s13688-025-00611-4.

**Additional file 1.** (DOCX 5.1 MB)

**Author contributions**
HG and MM developed the methodology. HG Implemented the analysis and wrote the manuscript. MM, JC, and RME reviewed and edited the manuscript and provided support throughout the research.

**Funding information**
The authors acknowledge the following funding sources: ESRC UBEL Doctoral Training Partnership (UKRI ESRC: ES/P000592/1) (HG). Geographic Data Service (UKRI ESRC: ES/Z504464/1) (JC).

**Data availability**
This study relies on publicly available data. The datasets used in this study are referenced in the bibliography. Software used to conduct the analysis is available in an open source repository which includes instructions for end-to-end reproducibility: https://github.com/hamishgibbs/federated_analytics_paper.

## Declarations

**Ethics approval and consent to participate**
This research has been approved by the University College London Research Ethics Service (ref: 21813/001).

**Competing interests**
The authors declare no competing interests.

**Author details**
[1]Network Science Institute, Northeastern University, Boston, USA. [2]Department of Computer Science, University College London, London, UK. [3]Department of Computer Science and Engineering, University of Bologna, Bologna, Italy. [4]Department of Geography, University College London, London, UK. [5]Department of Infectious Disease Epidemiology, London School of Hygiene & Tropical Medicine, London, UK.

Received: 18 January 2025 Accepted: 12 December 2025 Published online: 16 January 2026

**References**

1. Cinnamon J, Jones SK, Adger WN (2016) Evidence and future potential of mobile phone data for disease disaster management. Geoforum 75:253–264
2. Oliver N, Lepri B, Sterly H, Lambiotte R, Deletaille S, Nadai MD, et al (2020) Mobile phone data for informing public health actions across the COVID-19 pandemic life cycle. Sci Adv 6(23):eabc0764
3. Yabe T, Jones NKW, Rao PSC, Gonzalez MC, Ukkusuri SV (2022) Mobile phone location data for disasters: a review from natural hazards and epidemics. Comput Environ Urban Syst 94:101777
4. Hilbert M (2016) Big data for development: a review of promises and challenges. Dev Policy Rev 34(1):135–174
5. Ratti C, Frenchman D, Pulselli RM, Mobile WS (2006) Landscapes: using location data from cell phones for urban analysis. Environ Plan B, Plan Des 33(5):727–748
6. Wang R, Zhang X, Li N (2022) Zooming into mobility to understand cities: a review of mobility-driven urban studies. Cities 130:103939
7. Tizzoni M, Bajardi P, Decuyper A, King GKK, Schneider CM, Blondel V, et al (2014) On the use of human mobility proxies for modeling epidemics. PLoS Comput Biol 10(7):e1003716
8. Lu X, Bengtsson L, Holme P (2012) Predictability of population displacement after the 2010 Haiti earthquake. Proc Natl Acad Sci USA 109(29):11576–11581
9. Kissler SM, Kishore N, Prabhu M, Goffman D, Beilin Y, Landau R, et al (2020) Reductions in commuting mobility correlate with geographic differences in SARS-CoV-2 prevalence in New York City. Nat Commun 11(1):4674
10. Moro E, Calacci D, Dong X, Pentland A (2021) Mobility patterns are associated with experienced income segregation in large US cities. Nat Commun 12(1):4633
11. Grantz KH, Meredith HR, Cummings DAT, Metcalf CJE, Grenfell BT, Giles JR, et al (2020) The use of mobile phone data to inform analysis of COVID-19 pandemic epidemiology. Nat Commun 11(1):4961
12. Kishore N, Kiang MV, Engø-Monsen K, Vembar N, Schroeder A, Balsari S, et al (2020) Measuring mobility to monitor travel and physical distancing interventions: a common framework for mobile phone data analysis. Lancet Digit Health 2(11):e622–e628
13. Nouvellet P, Bhatia S, Cori A, Ainslie KEC, Baguelin M, Bhatt S, et al (2021) Reduction in mobility and COVID-19 transmission. Nat Commun 12(1):1090
14. Schlosser F, Maier BF, Jack O, Hinrichs D, Zachariae A, Brockmann D (2020) COVID-19 lockdown induces disease-mitigating structural changes in mobility networks. Proc Natl Acad Sci USA 117(52):32883–32890
15. Jeffrey B, Walters CE, Ainslie KEC, Eales O, Ciavarella C, Bhatia S, et al (2020) Anonymised and aggregated crowd level mobility data from mobile phones suggests that initial compliance with COVID-19 social distancing interventions was high and geographically consistent across the UK. Wellcome Open Res 5:170
16. de Montjoye YA, Hidalgo CA, Verleysen M, Blondel VD (2013) Unique in the crowd: the privacy bounds of human mobility. Sci Rep 3(1):1376
17. Valentino-DeVries J, Singer N, Keller MH, Krolik A (2018) Your apps know where you were last night, and they're not keeping it secret. NY Times. https://www.nytimes.com/interactive/2018/12/10/business/location-data-privacy-apps.html
18. Thompson SA, Warzel C (2019) Twelve million phones, one dataset, zero privacy. NY Times. https://www.nytimes.com/interactive/2019/12/19/opinion/location-tracking-cell-phone.html
19. Cox J (2022) Data broker is selling location data of people who visit abortion clinics. Vice. https://www.vice.com/en/article/m7vzjb/location-data-abortion-clinics-safegraph-planned-parenthood
20. Basu A, Monreale A, Trasarti R, Corena JC, Giannotti F, Pedreschi D, et al (2015) A risk model for privacy in trajectory data. J Trust Manag 2(1):9
21. Houssiau F, Rocher L, de Montjoye YA (2022) On the difficulty of achieving differential privacy in practice: user-level guarantees in aggregate location data. Nat Commun 13(1):29
22. Pyrgelis A, Troncoso C, De Cristofaro E, (2018) Knock knock, who's there? Membership inference on aggregate location data. Netw Distrib Syst Secur Symp
23. Razaghpanah A, Nithyanand R, Vallina-Rodriguez N, Sundaresan S, Allman M, Kreibich C, et al (2018) Apps, trackers, privacy, and regulators: a global study of the mobile tracking ecosystem. In: Proceedings 2018 network and distributed system security symposium. Internet Society, San Diego. https://www.ndss-symposium.org/wp-content/uploads/2018/02/ndss2018_05B-3_Razaghpanah_paper.pdf
24. Federal Trade Commission (2024) FTC order prohibits data broker X-mode social and outlogic from selling sensitive location data. https://www.ftc.gov/news-events/news/press-releases/2024/01/ftc-order-prohibits-data-broker-x-mode-social-outlogic-selling-sensitive-location-data
25. Bradford L, Aboy M, Liddell K (2020) COVID-19 contact tracing apps: a stress test for privacy, the GDPR, and data protection regimes. J Law Biosci 7(1):lsaa034

26. Bampoulidis A, Bruni A, Helminger L, Kales D, Rechberger C, Walch R (2020) Privately connecting mobility to infectious diseases via applied cryptography. https://eprint.iacr.org/2020/522
27. Dwork C, McSherry F, Nissim K, Smith A (2006) Calibrating noise to sensitivity in private data analysis. In: Halevi S, Rabin T (eds) Theory of cryptography. Lecture notes in computer science. Springer, Berlin, pp 265–284
28. Dwork C, Roth A (2014) The algorithmic foundations of differential privacy. Found Trends Theor Comput Sci 9(3–4):211–407
29. Bassolas A, Barbosa-Filho H, Dickinson B, Dotiwalla X, Eastham P, Gallotti R, et al (2019) Hierarchical organization of urban mobility and its connection with city livability. Nat Commun 10(1):4817
30. Sadilek A, Dotiwalla X New insights into human mobility with privacy preserving aggregation. Google AI Blog. http://ai.googleblog.com/2019/11/new-insights-into-human-mobility-with.html
31. Google (2023) Understand the data - community mobility reports help. https://support.google.com/covid19-mobility/answer/9825414?hl=en&ref_topic=9822927
32. Aktay A, Bavadekar S, Cossoul G, Davis J, Desfontaines D, Fabrikant A, et al (2020) Google COVID-19 community mobility reports: anonymization process description (version 1.1). http://arxiv.org/abs/2004.04145
33. Herdagdelen A, Dow A, State B, Mohassel P, Pompe A, Meta Research (2020) Protecting privacy in Facebook mobility data during the COVID-19 response - meta research. https://research.facebook.com/blog/2020/06/protecting-privacy-in-facebook-mobility-data-during-the-covid-19-response/
34. Ramage D, Federated MS (2020) Analytics: collaborative data science without data collection. https://blog.research.google/2020/05/federated-analytics-collaborative-data.html
35. Elkordy A, Ezzeldin YH, Han S, Sharma S, He C, Mehrotra S, et al (2023) Federated analytics: a survey. APSIPA Trans Signal Inf Process
36. Cormode G, Maddock S, Maple C (2021) Frequency estimation under local differential privacy. Proc VLDB Endow 14(11):2046–2058
37. Differential Privacy Team, Apple (2017) Apple machine learning research. Learning with privacy at scale. https://machinelearning.apple.com/research/learning-with-privacy-at-scale
38. Lopuhaä-Zwakenberg M, Goseling J (2024) Mechanisms for robust local differential privacy. Entropy 26(3):233
39. Hong Y, Li J, Lin Y, Hu Q, Li X (2024) Trajectory-aware privacy-preserving method with local differential privacy in crowdsourcing. EURASIP J Inf Secur 2024(1):28
40. McGriff M, Google (2023) Updates to location history and new controls coming soon to maps. https://blog.google/products/maps/updates-to-location-history-and-new-controls-coming-soon-to-maps/
41. Wang H, Li Y, Gao C, Wang G, Tao X, Jin D (2021) Anonymization and de-anonymization of mobility trajectories: dissecting the gaps between theory and practice. IEEE Trans Mob Comput 20(3):796–815
42. Buckee CO, Balsari S, Chan J, Crosas M, Dominici F, Gasser U, et al (2020) Aggregated mobility data could help fight COVID-19. Science 368(6487):145–146
43. Kang Y, Gao S, Liang Y, Li M, Rao J, Kruse J (2020) Multiscale dynamic human mobility flow dataset in the U.S. during the COVID-19 epidemic. Sci Data 7(1):390
44. Pappalardo L, Simini F, Rinzivillo S, Pedreschi D, Giannotti F, Barabási AL (2015) Returners and explorers dichotomy in human mobility. Nat Commun 6(1):8166
45. Pappalardo L, Rinzivillo S, Human SF (2016) Mobility modelling: exploration and preferential return meet the gravity model. Proc Comput Sci 83:934–939
46. US Census Bureau. Census.gov. Geographic levels. https://www.census.gov/programs-surveys/economic-census/guidance-geographies/levels.html
47. Abowd JM, Adams T, Ashmead R, Darais D, Dey S, Garfinkel SL, et al (2023) The 2010 census confidentiality protections failed, here's how and why. National Bureau of Economic Research (Working Paper Series). https://www.nber.org/papers/w31995
48. Wang T, Zhang X, Feng J, Yang X (2020) A comprehensive survey on local differential privacy toward data statistics and analysis. Sensors 20(24):7030
49. Gadotti A, Houssiau F, Annamalai MSMS de Montjoye YA (2022) Pool inference attacks on local differential privacy: quantifying the privacy guarantees of Apple's count mean sketch in practice. In: 31st USENIX security symposium, pp 501–518. https://www.usenix.org/conference/usenixsecurity22/presentation/gadotti
50. Zhang M, Lin S, Yin L (2023) Local differentially private frequency estimation based on learned sketches. Inf Sci 649:119667
51. Acharya J, Sun Z, Hadamard ZH (2019) Response: estimating distributions privately, efficiently, and with little communication. In: Proceedings of the twenty-second international conference on artificial intelligence and statistics. PMLR, pp 1120–1129. https://proceedings.mlr.press/v89/acharya19a.html
52. Gallotti R, Maniscalco D, Barthelemy M, De Domenico M (2024) Distorted insights from human mobility data. Commun Phys 7(1):1–10
53. Ouadrhiri AE, Abdelhadi A (2022) Differential privacy for deep and federated learning: a survey. IEEE Access 10:22359–22380
54. Steinke T, Ullman J (2016) Between pure and approximate differential privacy. J Priv Confid 7(2). https://journalprivacyconfidentiality.org/index.php/jpc/article/view/648
55. Yabe T, Tsubouchi K, Shimizu T, Sekimoto Y, Sezaki K, Moro E, et al (2024) YJMob100K: city-scale and longitudinal dataset of anonymized human mobility trajectories. Sci Data 11(1):397
56. Ward JH (1963) Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58(301):236–244
57. Murtagh F, Legendre P (2014) Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? J Classif 31(3):274–295
58. Kohli N, Aiken E, Blumenstock JE (2024) Privacy guarantees for personal mobility data in humanitarian response. Sci Rep 14(1):28565
59. Ioannou G, Marchioro T, Nicolaides C, Pallis G, Markatos E (2024) Evaluating the utility of human mobility data under local differential privacy. In: 2024 25th IEEE international conference on mobile data management (MDM), pp 67–76. https://ieeexplore.ieee.org/document/10591493
60. González MC, Hidalgo CA, Barabási AL (2008) Understanding individual human mobility patterns. Nature 453(7196):779–782

61. Ye Q, Hu H (2020) Local differential privacy: tools, challenges, and opportunities. In: U LH, Yang J, Cai Y, Karlapalem K, Liu A, Huang X (eds) Web information systems engineering. Communications in computer and information science. Springer, Singapore, pp 13–23
62. Cheu A, Smith A, Ullman J (2021) Manipulation attacks in local differential privacy. In: 2021 IEEE symposium on security and privacy (SP), pp 883–900. https://ieeexplore.ieee.org/document/9519418
63. Gecer M, Garbinato B (2024) Federated learning for mobility applications. ACM Comput Surv 56(5):133:1–133:28
64. Bassolas A, Barbosa-Filho H, Dickinson B, Dotiwalla X, Eastham P, Gallotti R, et al (2022) Reply to: On the difficulty of achieving differential privacy in practice: user-level guarantees in aggregate location data. Nat Commun 13(1):30

## Publisher's note