Generating Privacy-Preserving Artificial Call Detail Records (CDRs) From Mobile Phone Subscriber Profiles

Viren Dias
LIRNEasia
Colombo, Sri Lanka
viren@lirneasia.net

Lasantha Fernando
LIRNEasia
Colombo, Sri Lanka
lasantha@lirneasia.net

***Abstract*—Call Detail Records (CDRs) are primarily generated for billing purposes by mobile network operators. They provide a history of network-related transactions for an individual subscriber. In recent years, due to the rich individual social and spatiotemporal traces generated by such datasets, pseudonymized mobile network CDRs have been increasingly used in studying various aspects of human mobility, social connectivity and economic activity. However, contemporary studies have demonstrated that a significant portion of subscribers can be reidentified given a number of spatiotemporal observations, despite pseudonymization. This violates commonly accepted notions of individual privacy and restricts the shareability of these datasets, impairing both replicability and wider academic uses of such data. In this study, we propose a novel probabilistic methodology to generate privacy-preserving artificial data by combining selected characteristics of each individual in the dataset with random noise. We demonstrate this methodology using a CDR dataset and show that it substantially reduces the reidentifiability of each subscriber in the generated dataset, while maintaining the utility of the dataset by preserving each subscriber's mobility characteristics. In addition, we show that the reidentifiability of the subscriber is resistant to an increase in the number of spatiotemporal observations.**

***Index Terms*—Data mining, data sharing, knowledge and data engineering tools and techniques, location-dependent and sensitive, privacy**

I. Introduction

Mobile network operators maintain a record of network related transactions for each individual subscriber, primarily for billing purposes. These records are commonly referred to as Call Detail Records (CDRs). Each record comprises the subscriber's identity, the identity of the corresponding subscriber, the direction of the call, the identity of the base transceiver station (BTS) the subscriber connected to, a timestamp, and the duration of the call.

Given that a mobile device preferentially connects to the nearest BTS and that the locations of these BTSs are known, CDRs inadvertently provide a mobility trace for each subscriber. In conjunction with the evident social network profile it provides, CDR datasets present a valuable resource in studying various aspects of human mobility, social connectivity and economic activity.

This work has been funded through a grant from the International Development Research Centre (IDRC) of Canada.

In the interest of safeguarding the privacy of their subscribers, mobile network operators typically pseudonymize their CDRs prior to disclosure for research. Each subscriber is given a pseudonym that is consistent throughout the dataset. However, a study by Montjoye et al. [1] discovered that human mobility traces were surprisingly distinct and that even a modest number of spatiotemporal observations of an subscriber was sufficient to successfully reidentify a sizeable proportion of subscribers in the dataset. This significantly impairs the shareability of CDR datasets.

Traditionally, reidentification in CDR datasets had been mitigated via the use of generalization. Introduced by Samarati and Sweeney [2], this technique involved temporal and spatial aggregation to reduce the resolution of each observation in the dataset. This evidently came at the cost of diminished utility of the dataset.

However, Montjoye et al. [1] explored this solution and concluded that it contributes little to safeguarding privacy. The marginal increase in privacy was not worth the rapidly-diminishing utility of the dataset, and it could be easily overcome with just a few additional spatiotemporal observations. Hence, novel methods that transform such datasets to safeguard privacy, whilst preserving the desired characteristics of the original dataset, are of significant interest.

In this study, we propose a methodology that constructs a profile for each individual in a dataset, with the intent of preserving selected characteristics in Section III. We then offer an algorithm that couples these profiles with random noise and generates artificial data for each individual. In addition, we propose a mechanism by which to evaluate the preservation of privacy in artificial datasets. In Section IV we present the results of this evaluation, which shows a substantial improvement over conventional methods and a strong resilience to additional spatiotemporal observations. We demonstrate the preservation of utility in the generated dataset by presenting a strong correlation between corresponding characteristics in the original and generated datasets. In Section V, we discuss some of the limitations of our methodology and posit possible solu-

TABLE I
SAMPLE OBSERVATIONS

CALL_DIRECTION	DEVICE_NAME	SUBSCRIBER_ID	OTHER_ID	BTS_ID	TIMESTAMP	DURATION
IN	E-TELT10	A8129421	G7515498	C5210	2012/11/10 06:35:37	20
OUT	BLUDEEJAY II	D2575246	J8547632	L2509	2013/11/10 20:07:55	35

tions to these limitations. We consider possible applications of artificial datasets generated by our methodology in research. Finally, we propose possible extensions to our methodology to generate artificial datasets that are applicable to a wider range of research topics.

II. RELATED WORK

Samarati and Sweeney [2] addressed the issue of reidentification of individuals in any pseudonymized dataset by introducing the concept of k-anonymity, whereby the combined data for each individual in a dataset cannot be distinguished from at least k-1 individuals, and proposing two approaches by which k-anonymity could be achieved; generalization and suppression.

However, a study by Aggarwal [3] showed that for high dimensional datasets k-anonymity cannot be achieved without an unacceptable loss of utility of the dataset. Narayanan and Shmatikov [4] demonstrated this by presenting and applying a new class of statistical de-anonymization attacks on the Netflix Prize dataset.

Montjoye et al. [1] illustrated the inadequacy of k-anonymity in pseudonymized CDR datasets by conducting a study on 15 months of data for 1.5 million subscribers in a small European country. The authors determined the uniqueness of a trace based on a varying number of randomly selected data points for each subscriber at varying temporal and spatial resolutions. They concluded that at a temporal resolution of one hour and a spatial resolution equal to the region defined by the Voronoi tessellation of the network operator's BTS locations, four spatiotemporal observations were sufficient to uniquely identify 95% of the individuals in the dataset.

Buczak et al. [5] developed an approach for creating synthetic electronic medical records (EMRs) based on authentic EMRs in order to safeguard privacy. Isaacman et al. [6] applied the same concept to CDRs and developed a method which profiles a population as a whole and uses the temporal and spatial probability distributions to produce synthetic subscribers and corresponding CDRs. Mir et al. [7] improved upon this model by adding controlled noise to satisfy differential privacy.

Plausible deniability [8] is another theoretical notion of privacy that had been proposed as a criterion that can be used to provide a formal privacy guarantee for synthetic generated datasets. The generative technique introduced by Bindschaedler, et al. in [9] had been validated on large-scale input datasets to generate datasets satisfying plausible deniability.

More recent studies have indicated a growing trend in exploring machine learning based methods to generate synthetic datasets that satisfy differential privacy [10], [11]. These studies have utilized techniques such as Generative Adversarial Networks (GANs) [12] and Autoencoders [13] to generate synthetic datasets for multiple domains from healthcare to credit scores to optical digit recognition.

III. METHODS

The methodology described in this section was tailored to the specific use-case of generating artificial data from a CDR dataset with the intent of preserving the mobility characteristics of each subscriber.

A. Dataset Description

For the purposes of this work, we used a random sample of 1 million active subscribers from a dataset comprising one month of pseudonymized CDRs generated from voice calls for approximately 4.8 million subscribers from mobile network operators in Sri Lanka. An active subscriber was defined as having on average at least two calls per day over the one month time period.

Each observation comprised the unique identifier of the subscriber, the unique identifier of the corresponding subscriber, the direction of the call, the identifier of the BTS that the subscriber was connected to, the timestamp of the call precise to a second and the duration of the call precise to a second. A sample observation is provided in Table. I.

B. Subscriber Profiles

The methodology for profiling a subscriber is contingent on the characteristics required to be preserved. For this study, we chose to preserve just mobility characteristics. As such, a subscriber's profile comprised: a temporal probability distribution, a spatial probability distribution and a total number of observations.

The temporal probability distribution was based on a set of temporal categories. The temporal categories were defined by initially identifying whether a particular day was a working or non-working day and then dividing the day into 8 contiguous 3-hour segments, which we termed as 'octants'. The resulting 16 temporal categories are shown in Table. II. Non-working days were defined as weekends and national holidays, and the remaining days were considered to be working days. Each observation was categorized into one of the temporal categories. The probability of an observation within each temporal category was determined by calculating the proportion of total observations that lie within that temporal category.

TABLE II
TEMPORAL CATEGORIES

Day Type	Day Octant	ID
Working	00:00 - 03:00	1
	03:00 - 06:00	2
	06:00 - 09:00	3
	09:00 - 12:00	4
	12:00 - 15:00	5
	15:00 - 18:00	6
	18:00 - 21:00	7
	21:00 - 00:00	8
Non-Working	12:00 - 03:00	9
	03:00 - 06:00	10
	06:00 - 09:00	11
	09:00 - 12:00	12
	12:00 - 15:00	13
	15:00 - 18:00	14
	18:00 - 21:00	15
	21:00 - 00:00	16

The spatial probability distribution was based on a set of spatial categories, which were in turn based on a 1km by 1km grid. The placement of the grid was fixed by minimizing the error term, defined as the cumulative distance between each BTS and the center of the grid cell to which it was assigned. Given a temporal category, the probability of an observation within each spatial category was determined by calculating the proportion of total observations that lie within that temporal and spatial category.

C. Algorithm Outline

Given a subscriber's profile, privacy-preserving artificial CDRs were generated for each subscriber as follows:

1) Blank observations amounting to the total observations of the subscriber's profile, with an error of up to 5%, were created to initiate the process.
2) Each blank observation was assigned a temporal category in accordance with the temporal probability distribution of the subscriber as well as an arbitrary timestamp conforming to the temporal category.
3) The observations were then ordered chronologically, and the first observation was assigned a spatial category in accordance with the spatial probability distribution of the subscriber, with a predefined 5% probability of being substituted with a randomly selected neighboring spatial category.
4) The next observation was then assigned a spatial category in accordance with the spatial probability distribution, with a predefined 5% probability of being substituted with a randomly selected neighboring spatial category.
5) The speed between the current and preceding observations was constrained to 100 km/h. The speed limit of 100 km/h was chosen as this was the highest speed limit for any road in Sri Lanka.

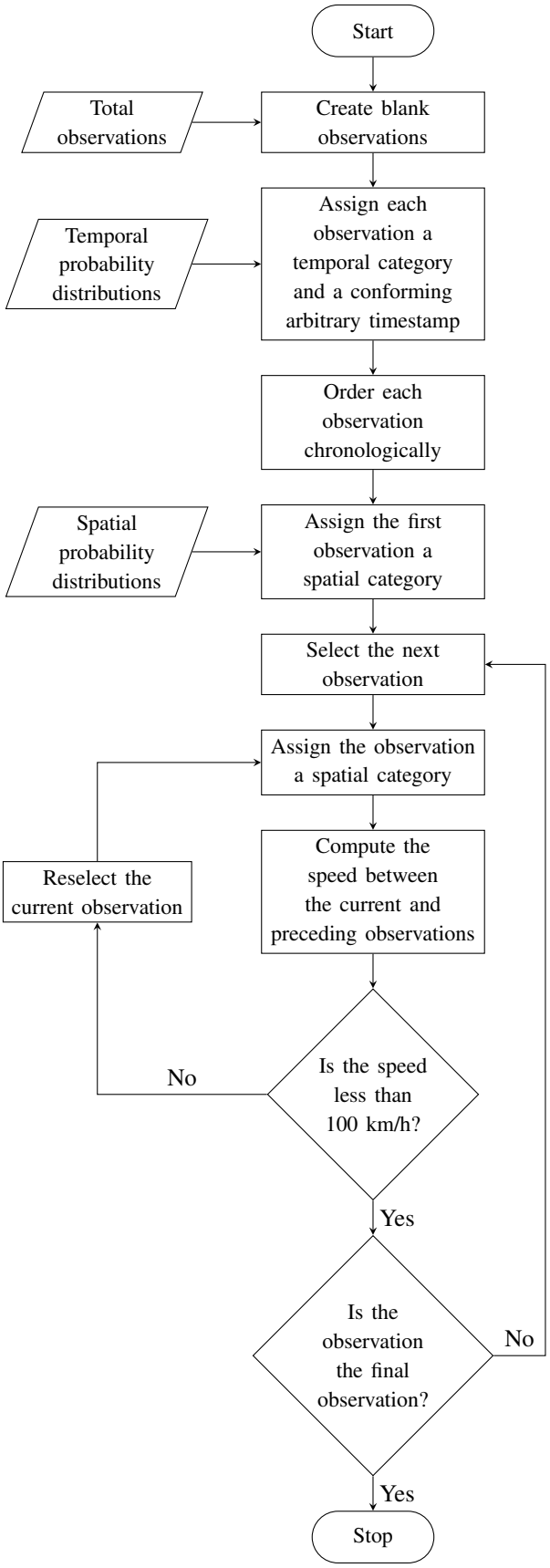

Fig. 1. Flowchart for the artificial data generation algorithm.

a) If the speed was within the limit of 100 km/h, the current observation was deemed plausible and Step 4 was repeated for the following observation.
b) If the speed exceeded the limit of 100 km/h, the current observation was deemed highly improbable and Step 4 was repeated for the current observation.

If Step 4 was attempted more than 100 times for the same observation, all the artificial CDRs associated with that subscribers profile were scrapped, and the CDR generation process was restarted from Step 1. If the number of restarts exceeded 100, the subscriber's profile was discarded. A flowchart for the algorithm is provided in Fig. 1.

D. Privacy Preservation Evaluation

The efficacy of our Profile-Based Data Generation methodology DG_p in preserving the privacy of subscribers in the dataset was evaluated by emulating an inference attack, whereby an attacker would use external information to infer protected information in a pseudonymized dataset. We gauged the vulnerability of datasets generated by our methodology to such attacks by the following measures:

- Uniqueness U - a measure borrowed from Montjoye et al. [1]. Given a number of real world spatiotemporal observations, it is the proportion of subscribers that can be unambiguously reidentified in a dataset.
- Mean probability of reidentification $\overline{R}$ - a measure adapted from uniqueness. Given a number of real world spatiotemporal observations, it is the probability that any given subscriber can be reidentified from a dataset.

To accomplish the above, we considered a dataset of subscribers S, where $S = \{s_1, s_2, s_3, \ldots, s_k\}$ and k is the total number of subscribers in the dataset. Each subscriber s had a set of spatiotemporal observations o_s. The set of all spatiotemporal observations was denoted by O, where $O = \{o_{s_1}, o_{s_2}, o_{s_3}, \ldots, o_{s_k}\}$. For a subscriber s_1, a predefined number n of samples were taken without replacement to form a set of spatiotemporal samples $o_{s_1 n}$. This set of samples $o_{s_1 n}$ was then compared with the set of all sets of spatiotemporal observations O. A subscriber was said to match if $o_{s_1 n} \subseteq o_s$. Since $o_{s_1 n} \subseteq o_{s_1}$ at least one match was guaranteed, and the set of matching subscribers $M_{s_1 n}$ satisfied the inequality $M_{s_1 n} \geq 1$. Therefore, for any subscriber s the probability of reidentification R_{sn} via random selection was defined as

$$R_{sn} = \frac{1}{|M_{sn}|}\,.$$

The sample of spatiotemporal observations $o_{s_1 n}$ was intended to represent the external information an attacker would use. This is a reasonable representation as they are real-world observations that would appear externally. However, this does not hold true in the case of an artificial dataset.

In light of this, we then considered an original dataset and a corresponding artificial dataset of subscribers S, where $S = \{s_1, s_2, s_3, \ldots, s_k\}$ and k is the total number of subscribers in both datasets. Each subscriber s had a set of spatiotemporal observations o'_s and a set of corresponding artificial spatiotemporal observations o_s. The set of all spatiotemporal observations was denoted by O', where $O' = \{o'_{s_1}, o'_{s_2}, o'_{s_3}, \ldots, o'_{s_k}\}$, and the set of all artificial spatiotemporal observations was denoted by O, where $O = \{o_{s_1}, o_{s_2}, o_{s_3}, \ldots, o_{s_k}\}$. For a subscriber s_1, a predefined number n of samples were taken without replacement to form a set of original spatiotemporal samples $o'_{s_1 n}$. This set of samples $o'_{s_1 n}$ was then compared with the set of all sets of artificial spatiotemporal observations O. A subscriber was said to match if $o'_{s_1 n} \subseteq o_s$. Since the expression $o'_{s_1 n} \subseteq o_{s_1}$ was not necessarily true, a match was not guaranteed, and the set of matching subscribers $M_{s_1 n}$ satisfied the inequality $M_{s_1 n} \geq 0$. For the case where $|M_{s_1 n}| = 0$, R_{sn} was undefined. Considering this special case, for any subscriber s the probability of reidentification R_{sn} via random selection was redefined as

$$R_{sn} = \begin{cases} \frac{1}{|M_{sn}|} & s \in M_{sn} \\ 0 & s \notin M_{sn}\,. \end{cases}$$

Having defined R, the mean probability of reidentification $\overline{R}$ for any subscriber s in the dataset was defined as

$$\overline{R}_{sn} = \frac{1}{|S|} \sum_{s \in S} R_{sn}\,.$$

In the case where $|M_{sn}| = 1$ and $s \in M_{sn}$, the probability of reidentification $R_{sn} = 1$, and the subscriber s was unambiguously reidentified and was therefore said to be unique. The set of all unique subscribers in the dataset was denoted by Q. Thus, the uniqueness U of the dataset was defined as

$$U_n = \frac{|Q|}{|S|}\,.$$

The results of these measures were then benchmarked against differing levels of generalization, as introduced by Samarati and Sweeney [2]:

- Generalization $G_{t_1 s_1}$ - aggregation at a temporal resolution t_1 of one hour and a spatial resolution s_1 equal to the region defined by the Voronoi tessellation of the network operator's BTS locations. This level of aggregation was selected for parity with the results of Montjoye et al. [1].
- Generalization $G_{t_2 s_2}$ - aggregation at a temporal resolution t_2 equal to the temporal categories and a spatial resolution s_2 equal to the spatial categories, as defined in Section III-B. This level of aggregation was selected for parity with our method for profiling a subscriber.

E. Utility Preservation Evaluation

The ability of our Profile-Based Data Generation methodology DG_p to preserve the utility of the dataset was evaluated via a number of mobility characteristics. These characteristics were computed for the original and generated datasets as follows:

- Home and work locations - for each subscriber, the home and work locations were determined by the location with the most tower days. As introduced in detail by Isaacman

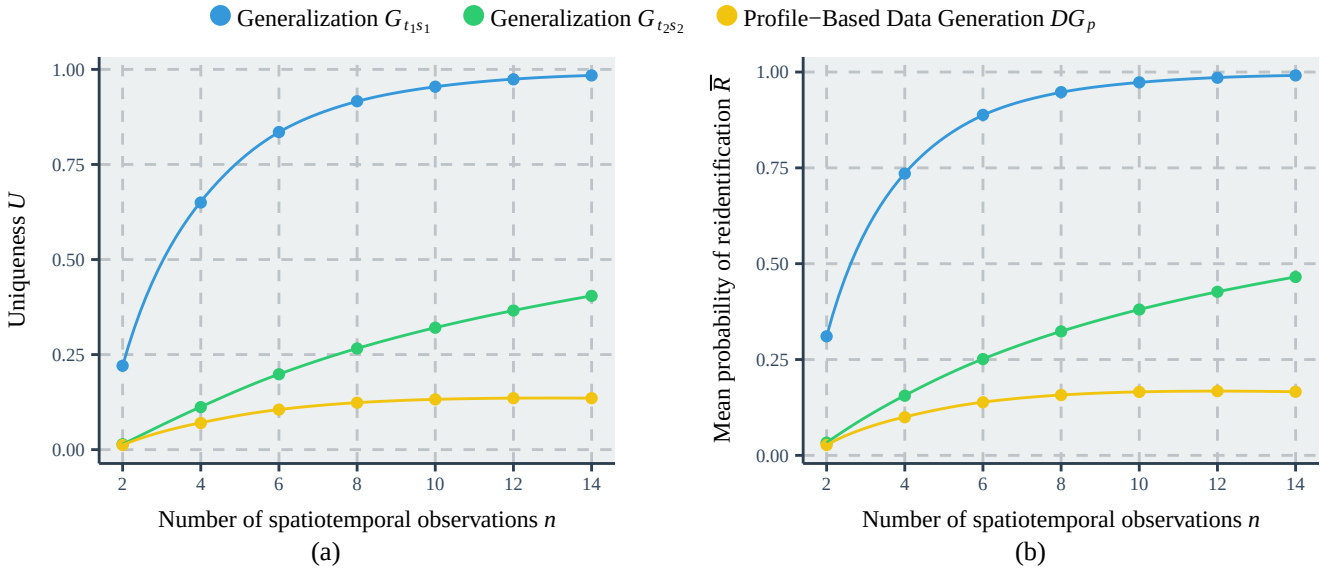

Fig. 2. Results of the privacy preservation evaluation. (a) Uniqueness of a dataset against number of spatiotemporal observations with respect to several privacy preservation methods. (b) Mean probability of reidentification of subscribers in a dataset against number of spatiotemporal observations with respect to several privacy preservation methods.

et al. [14], tower days is defined as the number of days in which a subscriber is observed at a particular location. For the derivation of home locations, only observations with a timestamp between 9pm and 5am were considered. And for the derivation of work locations, only observations with a timestamp between 10am and 3pm on working days were considered.

- Unique location count - for each subscriber, the number of distinct locations visited were computed. To facilitate a direct comparison, the locations for both the original and generated datasets were defined by the spatial categories as described in Section III-B.
- Radius of gyration - for a subscriber s with a total number N_s of spatiotemporal observations o_s and haversine distance d_{o_s} between each observation and the subscriber's center of gravity, where the center of gravity is the average location of all observations, the radius of gyration ROG_s is given by

$$ROG_s = \sqrt{\frac{1}{N_s}\sum_{o_s=1}^{N_s} d_{o_s}^2}\,.$$

A more comprehensive definition is provided by González et al. [15]. Radius of gyration is regularly used in studies of human mobility to represent the typical distance traveled by an individual [15]–[17].

- Mobility entropy - for a subscriber s with spatiotemporal observations o_s and a total number N_s of spatiotemporal observations, the probability p_{sl} of visiting a location l was calculated by dividing the number of observations at location l by the total number N_s of spatiotemporal observations. The set of all locations is denoted by L. As defined by Song et al. [16], the temporal-uncorrelated mobility entropy E_s^{unc} is given by

$$E_s^{unc} = -\sum_{l\in L} p_{sl} log_2(p_{sl})\,.$$

Mobility entropy has been used as a measure of diversity in contemporary human mobility studies [18]–[20].

IV. Results

A. Privacy Preservation

The privacy preservation evaluation was conducted, as described in Section III-D, for a varying number n of spatiotemporal observations. The resulting uniqueness U and mean probability of reidentification $\overline{R}$ values were plotted against the number n of spatiotemporal observations for each of the methodologies: Generalization $G_{t_1s_1}$, Generalization $G_{t_2s_2}$ and Profile-Based Data Generation DG_p. A logarithmic curve was fitted through each set of values, and the resulting plots are shown in Fig. 2. The following observations were made:

- $G_{t_1s_1}$ showed a sharp increase in both U and $\overline{R}$ with respect to n, with $U_6 = 0.84$ and $\overline{R}_6 = 0.89$.
- $G_{t_2s_2}$ showed a significant improvement over $G_{t_1s_1}$ with $U_6 = 0.20$ and $\overline{R}_6 = 0.25$. However, U and $\overline{R}$ continued to increase at a consistent rate and did not appear to stabilize by $n = 14$.
- DG_p initially showed a substantial improvement over $G_{t_1s_1}$ and a notable improvement over $G_{t_2s_2}$ with $U_6 = 0.11$ and $\overline{R}_6 = 0.14$. As n increased, DG_p swiftly stabilized at $U = 0.14$ and $\overline{R} = 0.17$, resulting in a substantial improvement over both $G_{t_1s_1}$ and $G_{t_2s_2}$. Beyond $n = 8$, additional spatiotemporal observations had little to no effect: $U_{10} = 0.13$, $U_{12} = 0.14$, $U_{14} = 0.14$, $\overline{R}_{10} = 0.17$, $\overline{R}_{12} = 0.17$ and $\overline{R}_{14} = 0.17$.

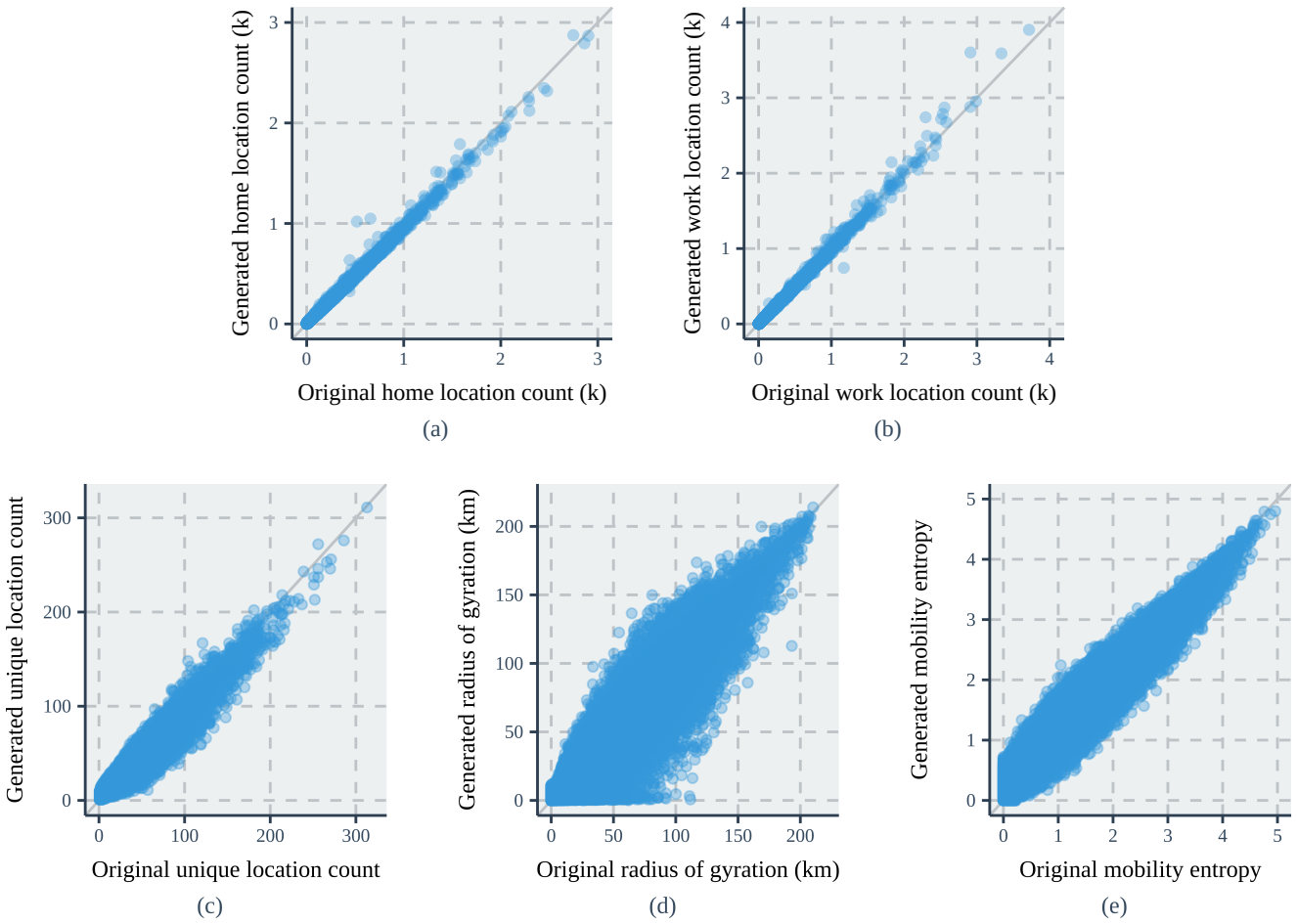

Fig. 3. Results of the utility preservation evaluation. (a) Home location counts aggregated by spatial category of the generated dataset against the original dataset. (b) Work location counts aggregated by spatial category of the generated dataset against the original dataset. (c) Unique location count of each subscriber of the generated dataset against the original dataset. (d) Radius of gyration of each subscriber of the generated dataset against the original dataset. (e) Mobility entropy of each subscriber of the generated dataset against the original dataset.

B. Utility Preservation

The mobility characteristics of the original and generated datasets were computed as described in Section III-E. The number of home and work locations were aggregated by spatial category, as defined in Section III-B. The home and work location counts of the generated dataset were plotted against the corresponding counts in the original dataset for each spatial category, as shown in Fig. 3(a) and Fig. 3(b) respectively. The unique location counts, radius of gyration and mobility entropy for each subscriber in the generated dataset were plotted against the corresponding values in the original dataset, as shown in Fig. 3(c), Fig. 3(d), and Fig. 3(e) respectively. An identity line was plotted on each of the plots to aid in their evaluation. The following observations were made:

- Fig. 3(a) shows an excellent correlation between home location counts with a negligible number of outliers.
- Fig. 3(b) shows an excellent correlation between work location counts with a negligible number of outliers.
- Fig. 3(c) shows an excellent correlation between unique location counts. It has a slight tendency toward larger original unique location counts.
- Fig. 3(d) shows a good correlation between radius of gyration. Some subscribers having a significantly larger generated radius of gyration and vice versa. However, a noteworthy portion of subscribers have a substantially smaller radius of gyration in the generated dataset.
- Fig. 3(e) shows an excellent correlation between mobility entropy. Subscribers tend to have a higher mobility entropy in the generated dataset.

V. Discussion

As evidenced in Section IV-A, our methodology is a substantial improvement over conventional methods of privacy preservation. An observation of particular interest is its resistance to additional spatiotemporal observations beyond a certain point, whereas the additional privacy afforded by a higher level of generalization can be easily counteracted by additional spatiotemporal observations.

A non-apparent facet in favor of our methodology is that in the case where the cardinality of the set of matching sub-

scribers $|M_{sn}| = 1$ and the subscriber $s \notin M_{sn}$, the subscriber s would appear to have been unambiguously reidentified, when in actual fact the subscriber has been reidentified incorrectly. Therefore, the attacker can never know for certain the success of the attack.

It should be noted that we used the random sample of 1 million active subscribers as opposed to the entire dataset of 4.8 million subscribers in order to better illustrate the efficacy of our methodology in preserving privacy. The restrictive time span of one month presented a low temporal dimensionality, and in conjunction with the sheer number of subscribers, resulted in an inherently homogeneous dataset. As such, it has a intrinsically low uniqueness. The geographical density of the population of Sri Lanka was of no concern as it was accompanied by a dense network of BTSs, which increased the spatial dimensionality to compensate.

Whilst the results of all the computed characteristics in Section IV-B present a strong correlation, the radius of gyration precipitates a point of minor concern. A portion of subscribers have a substantially smaller radius of gyration in the generated dataset.

We speculate that anomalous long-distance trips are underproduced as their generation necessitates the assignment of multiple improbable spatial categories in succession. This is corroborated by unique location counts, which depict a slight tendency toward larger numbers in the original dataset.

The profiling method described in Section III-B makes the assumption that subscribers' activities are consistent between working days, and between non-working days. Within these day classifications, it makes a further assumption that subscribers' activities are consistent between corresponding times of the day. As such, it is only designed to capture regular trips.

Whilst the limitation of the profiling method accounts for the absence of one-off long-distance trips, a single trip cannot justify such a large discrepancy in a subscriber's radius of gyration. Consequently, we deduce that uncommon, yet regular, long-distance trips are underproduced as well. This is a limitation of the algorithm outlined in Section III-C.

A possible enhancement to the algorithm in order to alleviate the limitation identified above would be to have the spatial probability distribution temporarily influenced by the spatial category assigned to the preceding observation. The spatial probabilities would be adjusted as a function of the distance to the preceding observation. Nearby spatial categories would be impacted positively, whilst those further away would be impacted negatively. The magnitude of the impact would also be governed by the time elapsed since the preceding observation, with a shorter time period resulting in a larger impact.

A. Similar Methodologies

WHERE [6] and its successor DP-WHERE [7], present a similar methodology to our own. DP-WHERE profiles a population as a whole and synthesizes subscribers based on population-level attributes such as distributions of home and work locations and commute distances. Mir et al. show that synthetic datasets generated by their model retain much of its utility at a population scale. However, our methodology preserves the characteristics of each individual subscriber, which arguably improves the accuracy at a population scale, and allows for studies that require a finer resolution of data.

As discussed briefly in Section II, there are a number of recent studies that utilize machine learning to synthesize datasets. Despite positive results, we believe that such methods will be plagued by the same shortcomings of machine learning in decision-making - interpretability. As such, we speculate that research based on these datasets will be met with a level of skepticism. Conversely, we offer a simple interpretable methodology that achieves the same objectives.

B. Potential Applications

The results in Section IV-B demonstrate the capacity of our methodology to preserve the utility of the dataset in terms of selected mobility characteristics. As such an artificial CDR dataset generated by our methodology can be used to study human mobility.

It should be noted that an artificial dataset generated based on the exact profiling method described in Section III-B would only be viable for mobility studies for time periods of the order of a month. The temporal categories would be unsuitable for capturing mobility characteristics for larger time frames. However, the temporal categories could be effortlessly modified to capture long-term patterns with the simple addition of a monthly or seasonal dimension.

The profiling method can additionally be revised to include social characteristics, by way of a social probability distribution. The social probability distribution of a subscriber would define the probability to contact a specific subscriber given a temporal and spatial category. This would enable the use of an artificially generated dataset in studying social connectivity in addition to mobility.

VI. Conclusion

In this study, we outlined the issue of reidentification in CDR datasets and discussed its implications on academic uses of such datasets. We discussed traditional methods of combating reidentification and their limitations. We then proposed a methodology where individuals in a dataset would be profiled to capture selected characteristics, and these profiles would then be coupled with random noise to generate artificial data.

After having discussed the methodology and mechanisms to evaluate its efficacy, we demonstrated the effectiveness of our methodology in protecting privacy whilst preserving utility via its application to a CDR dataset. We also showed that the generated dataset is resilient to additional information in the event of an inference attack.

Furthermore, we discussed some of the limitations of our methodology and potential improvements. We then explored similar methodologies in recent literature and speculated what competencies our methodology would have over them. We concluded by outlining the potential applications of datasets generated by our methodology and proposed perturbations to make such datasets viable for a wider range of research.

ACKNOWLEDGMENT

The authors would like to thank Sriganesh Lokanathan for his valuable insights and comments, and Yashothara Shanmugarasa for her technical contributions. The authors would also like to thank Yudhanjaya Wijeratne and Merl Chandana for their constructive feedback.

REFERENCES

[1] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Scientific Reports*, vol. 3, no. 1, Mar. 2013. [Online]. Available: https://doi.org/10.1038/srep01376

[2] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression," technical report, SRI International, Tech. Rep., 1998.

[3] C. C. Aggarwal, "On k-anonymity and the curse of dimensionality," in *Proceedings of the 31st International Conference on Very Large Data Bases*, ser. VLDB '05. VLDB Endowment, 2005, p. 901–909.

[4] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *2008 IEEE Symposium on Security and Privacy (sp 2008)*. IEEE, May 2008. [Online]. Available: https://doi.org/10.1109/sp.2008.33

[5] A. L. Buczak, S. Babin, and L. Moniz, "Data-driven approach for creating synthetic electronic medical records," *BMC Medical Informatics and Decision Making*, vol. 10, no. 1, Oct. 2010. [Online]. Available: https://doi.org/10.1186/1472-6947-10-59

[6] S. Isaacman, R. Becker, R. Cáceres, M. Martonosi, J. Rowland, A. Varshavsky, and W. Willinger, "Human mobility modeling at metropolitan scales," in *Proceedings of the 10th international conference on Mobile systems, applications, and services - MobiSys '12*. ACM Press, 2012. [Online]. Available: https://doi.org/10.1145/2307636.2307659

[7] D. J. Mir, S. Isaacman, R. Caceres, M. Martonosi, and R. N. Wright, "DP-WHERE: Differentially private modeling of human mobility," in *2013 IEEE International Conference on Big Data*. IEEE, Oct. 2013. [Online]. Available: https://doi.org/10.1109/bigdata.2013.6691626

[8] V. Bindschaedler and R. Shokri, "Synthesizing plausible privacy-preserving location traces," in *2016 IEEE Symposium on Security and Privacy (SP)*. IEEE, May 2016. [Online]. Available: https://doi.org/10.1109/sp.2016.39

[9] V. Bindschaedler, R. Shokri, and C. A. Gunter, "Plausible deniability for privacy-preserving data synthesis," *Proceedings of the VLDB Endowment*, vol. 10, no. 5, pp. 481–492, Jan. 2017. [Online]. Available: https://doi.org/10.14778/3055540.3055542

[10] B. K. Beaulieu-Jones, Z. S. Wu, C. Williams, R. Lee, S. P. Bhavnani, J. B. Byrd, and C. S. Greene, "Privacy-preserving generative deep neural networks support clinical data sharing," *Circulation: Cardiovascular Quality and Outcomes*, vol. 12, no. 7, Jul. 2019. [Online]. Available: https://doi.org/10.1161/circoutcomes.118.005122

[11] A. Yale, S. Dash, R. Dutta, I. Guyon, A. Pavao, and K. P. Bennett, "Privacy Preserving Synthetic Health Data," in *ESANN 2019 - European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, Bruges, Belgium, Apr. 2019. [Online]. Available: https://hal.inria.fr/hal-02160496

[12] R. Torkzadehmahani, P. Kairouz, and B. Paten, "Dp-cgan: Differentially private synthetic data and label generation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[13] N. C. Abay, Y. Zhou, M. Kantarcioglu, B. Thuraisingham, and L. Sweeney, "Privacy preserving synthetic data release using deep learning," in *Machine Learning and Knowledge Discovery in Databases*. Springer International Publishing, 2019, pp. 510–526. [Online]. Available: https://doi.org/10.1007/978-3-030-10925-7_31

[14] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky, "Identifying important places in people's lives from cellular network data," in *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2011, pp. 133–151. [Online]. Available: https://doi.org/10.1007/978-3-642-21726-5_9

[15] M. C. González, C. A. Hidalgo, and A.-L. Barabási, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, Jun. 2008. [Online]. Available: https://doi.org/10.1038/nature06958

[16] C. Song, Z. Qu, N. Blumm, and A.-L. Barabasi, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, Feb. 2010. [Online]. Available: https://doi.org/10.1126/science.1177170

[17] J. E. Blumenstock, "Inferring patterns of internal migration from mobile phone call records: evidence from rwanda," *Information Technology for Development*, vol. 18, no. 2, pp. 107–125, Feb. 2012. [Online]. Available: https://doi.org/10.1080/02681102.2011.643209

[18] P. Bajardi, M. Delfino, A. Panisson, G. Petri, and M. Tizzoni, "Unveiling patterns of international communities in a global city using mobile phone data," *EPJ Data Science*, vol. 4, no. 1, Apr. 2015. [Online]. Available: https://doi.org/10.1140/epjds/s13688-015-0041-5

[19] L. Pappalardo and F. Simini, "Modelling individual routines and spatio-temporal trajectories in human mobility," *CoRR*, vol. abs/1607.05952, 2016. [Online]. Available: http://arxiv.org/abs/1607.05952

[20] ——, "Data-driven generation of spatio-temporal routines in human mobility," *Data Mining and Knowledge Discovery*, vol. 32, no. 3, pp. 787–829, Dec. 2017. [Online]. Available: https://doi.org/10.1007/s10618-017-0548-4