A Novel Methodology to Generate Privacy-Preserving Artificial Call Detail Records (CDRs) from Mobile Phone Subscriber Profiles

Viren Dias, Lasantha Fernando
LIRNEasia, 12 Balcombe Place, Colombo, Sri Lanka
(viren, lasantha)@lirneasia.net

LIRNE*asia* is a pro-poor, pro-market think tank whose mission is *Catalyzing policy change through research to improve people"s lives in the emerging Asia Pacific by facilitating their use of hard and soft infrastructures through the use of knowledge, information and technology*.

Contact: 12 Balcombe Place, Colombo 00800, Sri Lanka. +94 11 267 1160. info@lirneasia.net
www.lirneasia.net

This work was carried out with the aid of a grant from the International Development Research Centre (IDRC), Canada.

A Novel Methodology to Generate Privacy-Preserving Artificial Call Detail Records (CDRs) from Mobile Phone Subscriber Profiles

Viren Dias
LIRNE*asia*
Colombo, Sri Lanka
viren@lirneasia.net

Lasantha Fernando
LIRNE*asia*
Colombo, Sri Lanka
lasantha@lirneasia.net

***Abstract*—Call Detail Records (CDRs) are primarily generated for billing purposes by telecommunication network operators. They provide a history of network-related transactions for an individual subscriber. In recent years, due to the rich individual spatiotemporal traces generated by such datasets, pseudonymized mobile network CDRs have been increasingly used in studying various aspects of human mobility.**

However, contemporary studies have demonstrated that a significant portion of subscribers can be re-identified given sufficient matching spatiotemporal observations, despite pseudonymization [1]. 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 methodology to generate privacy-preserving artificial CDRs by combining the spatiotemporal characteristics of an individual mobile phone subscriber with random noise. We demonstrate that our methodology substantially reduces the reidentifiability of an individual subscriber in the generated dataset, while maintaining the utility of the dataset by preserving the mobility characteristics of the original dataset.

I. INTRODUCTION

Mobile network operators maintain a record of network related transactions for each individual subscriber, commonly referred to as Call Detail Records (CDRs), primarily for billing purposes. Each record comprises the subscriber's pseudonymized identity, the identity of the base transceiver station (BTS) the subscriber connected to, and a timestamp. Given that a mobile telecommunications device preferentially connects to the nearest BTS and that the locations of these BTSs are known, CDRs inadvertently provide a digital footprint for each subscriber. Consequently, such datasets are a valuable resource in studying various aspects of human mobility, social connectivity and economic activity.

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, Montjoye et al. (2013) discovered that human mobility traces are surprisingly distinct and that even modest knowledge of an individuals whereabouts is sufficient to successfully re-identity the individual in the dataset. This significantly impairs the shareability of CDR datasets.

Re-identification can be mitigated to a certain degree by decreasing the temporal and spatial resolution of each record, at the cost of diminished utility of the dataset. Montjoye et al. [1] explores this solution and concludes that it contributes little to safeguarding privacy. Hence, novel methods that transform such datasets to safeguard privacy, whilst preserving the desired statistical characteristics of the original dataset, are of significant interest.

In this work, we explore a methodology that profiles the mobility characteristics of each subscriber in the authentic dataset and then generates a synthetic dataset based on these profiles coupled with random noise. This bolsters privacy by adding a secondary layer of obscurity. We then verify that the mobility characteristics of the population are preserved using several measures, as well as demonstrate the synthetic dataset's resilience to re-identification of the original subscriber.

II. RELATED WORK

Samarati and Sweeney [2] addressed the issue of re-identification 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 demonstrated this by presenting and applying a new class of statistical de-anonymization attacks on the Netflix Prize dataset [4].

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

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

TABLE I
SAMPLE VOICE RECORDS

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

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 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.

III. METHODOLOGY

A. Dataset description

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

Each record 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 record is provided in table I.

B. Subscriber profiles

In order to preserve the individual mobility characteristics as much as possible, we generated a profile for each mobile phone subscriber in the original dataset. The following features were extracted to generate a mobility profile for a given subscriber:

- *Temporal probability distribution:* Each observation was categorized into one of the temporal categories. The temporal categories were defined by initially identifying whether a particular day was a working or non-working day and then diving the day into 8 contiguous 3-hour segments, which we termed as 'octants'. The resulting 16 temporal categories are shown in table II. The probability of an observation within each temporal category was determined by calculating the proportion of total observations that lie within that temporal category.
- *Spatial probability distribution:* The spatial categories were 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.
- *Total observations:* The total number of observations were calculated.

TABLE II
TEMPORAL CATEGORIES

Day Type	Day Octant	ID
Working	12 am - 3 am	1
	3 am - 6 am	2
	6 am - 9 am	3
	9 am - 12 pm	4
	12 pm - 3 pm	5
	3 pm - 6 pm	6
	6 pm - 9 pm	7
	9 pm - 12 am	8
Non-Working	12 am - 3 am	9
	3 am - 6 am	10
	6 am - 9 am	11
	9 am - 12 pm	12
	12 pm - 3 pm	13
	3 pm - 6 pm	14
	6 pm - 9 pm	15
	9 pm - 12 am	16

C. Algorithm outline

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

1) Blank records amounting to the total observations, with an error of up to 5%, were created to initiate the process.
2) Each blank record was assigned a temporal category in accordance with the temporal probability distribution as well as an arbitrary timestamp conforming to the temporal category.
3) The records were then ordered chronologically, and the first record was 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.

4) The next record 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 records was computed for constraint validation. A speed limit of 100 km/h was chosen as this was the highest speed limit for any road in Sri Lanka.
    a) If the speed was within the limit of 100km/h, the current record was deemed plausible and step 4 was repeated for the following record.
    b) If the speed exceeded the limit of 100km/h, the current record was deemed highly improbable and step 4 was repeated for the current record.

If the number of attempts for assigning spatial categories for any record exceeded 100, the artificial CDRs associated with the subscriber's profile were regenerated from scratch. If the number of attempts for regeneration exceeded 100, the subscriber's profile was discarded. The pseudocode for the algorithm is provided in 1

IV. RESULTS

A. Privacy preservation

The efficacy of our methodology in preserving privacy was evaluated by calculating two measures - the uniqueness and the mean probability of reidentification. The results were then benchmarked against a commonly used transformation - a simple reduction in the resolution of the original dataset via differing levels of temporal and spatial aggregation.

For clarity, the transformations we compared are as follows:

- Aggregation A_m at a temporal resolution of one hour and a spatial resolution equal to the Voronoi tessellation of the network operator's BTSs. This level of aggregation was selected for parity with the results of Montjoye et al. [1].
- Aggregation A_p at a temporal resolution equal to the temporal categories and a spatial resolution equal to the spatial categories, as defined in III-B. This level of aggregation was selected for parity with our method for profiling a subscriber's mobility.
- Profile-based CDR generation G_p as described in section III.

Given the original dataset D_o and a modified dataset D_t after having applied one of the above methods, the measures used in this evaluation are defined as follows:

- *Uniqueness U:* This measure was borrowed from Montjoye et al. [1]. For each subscriber S_o, a predetermined number n of spatiotemporal observations were randomly sampled without replacement from D_o to form a mobility trace $T_{n,o}$. U is defined as the proportion of the transformed mobility traces $T_{n,t}$ that are unique.
- *Mean probability of reidentification $\overline{P_r}$:* For each subscriber S_o, a predetermined number n of spatiotemporal observations were randomly sampled without replacement from D_o to form a mobility trace $T_{n,o}$. For a given S_{ori} and corresponding T_n, P_r is defined as the reciprocal of the number of subscribers with a matching T_n.

Algorithm 1 Generate artificial CDRs based on a subscriber profile

```
$profile_{orig} \leftarrow$ array[1..n][1..m]
$n_{obs} \leftarrow$ array[1..n]
$generated \leftarrow$ array[1..n]
$th_{loc} \leftarrow$ threshold for spatial attempts
$th_{time} \leftarrow$ threshold for temporal attempts
for $i \leftarrow 1$ to $n$ do
  $n_{gen} \leftarrow n_{obs}[i] \pm round(n_{obs}[i] * 0.05)$
  $n_{attempts} \leftarrow 0$
  $cdrs_{gen} \leftarrow array[1..n_{gen}]$
  repeat
    $n_{attempts}$ += $1$
    for $j \leftarrow 1$ to $n_{gen}$ do
      $r_{time} \leftarrow rand(0,1)$
      $p_{time} \leftarrow 0.0f$
      $t_{cat} \leftarrow 0$
      for $k \leftarrow 1$ to $m$ do
        $p_{time}$ += $profile_{orig}[i][k].p_{time}$
        if $p_{time} > r_{time}$ then
          $t_{cat} \leftarrow k$
          $break$
        end if
      end for
      $cdrs_{gen}[j].ts \leftarrow generate\_ts(t_{cat})$
      $data_{loc} \leftarrow profile_{orig}[i][t_{cat}].data_{loc}$
      $r_{loc} \leftarrow rand(0,1)$
      $p_{loc} \leftarrow 0.0f$
      for $l \leftarrow 1$ to $data_{loc}.length$ do
        $p_{loc}$ += $data_{loc}[l].prob$
        if $p_{loc} > r_{loc}$ then
          $cdrs_{gen}[j].loc \leftarrow data_{loc}[l].loc$
          $break$
        end if
      end for
    end for
    $generated[i] \leftarrow recalibrate(cdrs_{gen}, th_{loc}, data_{loc})$
  until ($n_{attempts} < th_{time}$ and $generated[i] == null$)
end for
```

B. Utility preservation

We evaluated the utility of the generated dataset by comparing home and work location counts aggregated by spatial category, the radius of gyration of each subscriber and the mobility entropy of each subscriber, with those of the original dataset.

We determined home and work locations by calculating the "tower days" of each spatial category for each subscriber. We borrowed the concept of "tower days" from Isaacman et al. [8] and it is defined as the number of days in which a subscriber appears in a specific spatial category. We We considered only records with a timestamp between 9pm and 5am for home

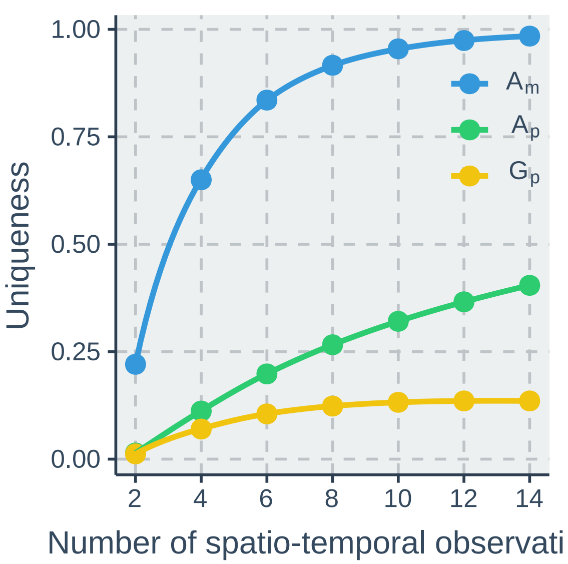

Fig. 1. The proportion of unique traces of subscribers against the number of spatiotemporal observations with respect to several privacy preservation methods.

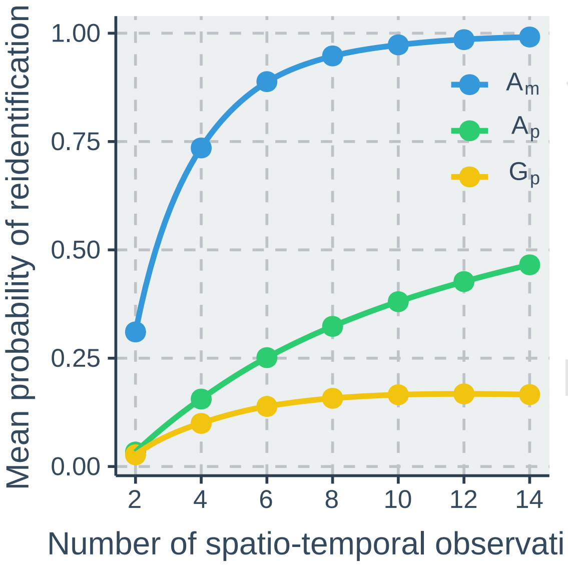

Fig. 2. The probability of reidentification of each subscriber against the number of spatiotemporal observations with respect to several privacy preservation methods.

locations, and records with a timestamp between 10am and 3pm on working days for work locations. We assigned the spatial categories with the most home and work tower days for each subscriber as their home and work locations respectively.

The mobility entropy for each subscriber was derived by calculating the temporal-uncorrelated mobility entropy as defined by Song et al. [9].

Figs. 3, 4, 5 and 6 exhibit a high correlation between the corresponding features of generated and original datasets. However a noteworthy portion of subscribers have a much lower radius of gyration in the generated dataset when compared with the original.

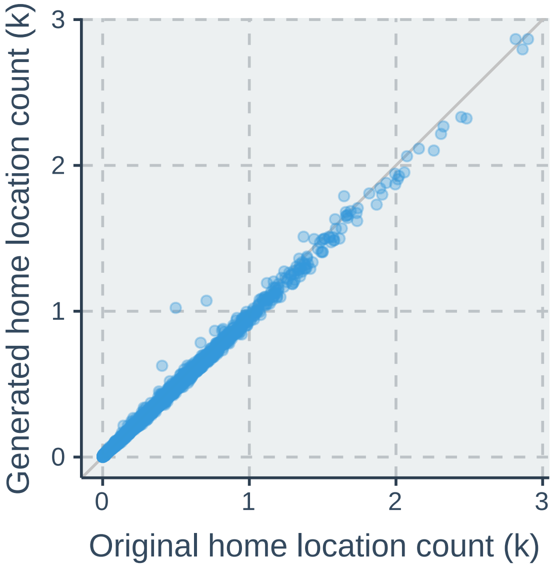

Fig. 3. The home location counts aggregated by spatial category of the generated dataset against the original dataset.

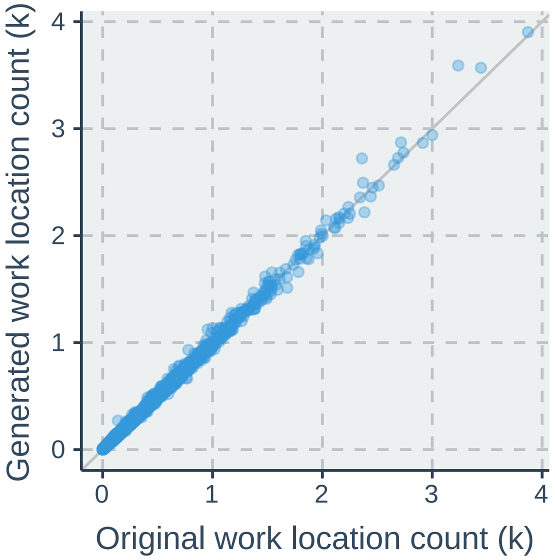

Fig. 4. The work location counts aggregated by spatial category of the generated dataset against the original dataset.

V. DISCUSSION

Several assumptions were made in developing the methodology that needs to be taken into consideration. One of the key assumptions in this study is that subscribers' activities were consistent between weekdays, as well as between weekends and national holidays. Within these day classifications, a further assumption was made that subscribers' activities were consistent between corresponding times of the day. We speculate that infrequent, long-distance trips are underpro-

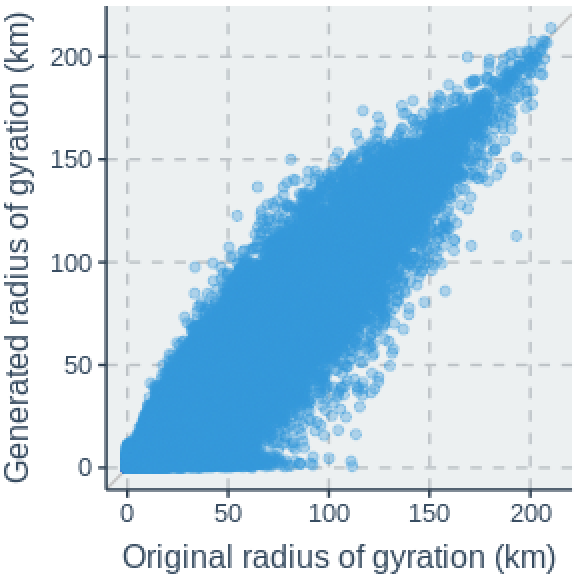

Fig. 5. The radius of gyration of each subscriber of the generated dataset against the original dataset.

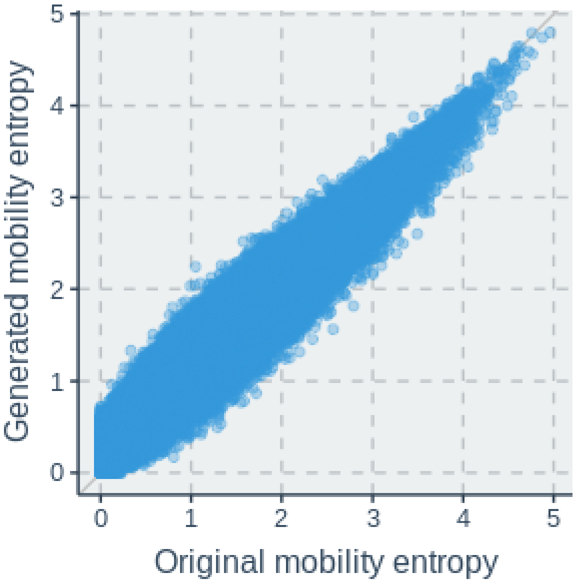

Fig. 6. The mobility entropy of each subscriber of the generated dataset against the original dataset.

duced as their generation necessitates multiple improbable events in succession, as corroborated by the radius of gyration comparison in Fig. 5.

VI. CONCLUSION

Our methodology generates a synthetic CDR dataset that preserves the core characteristics of the authentic CDR dataset, whilst safeguarding privacy. Extending this work to larger time spans would simply require the addition of a seasonal dimension to the temporal categories. Social characteristics could be preserved with the inclusion of the probability distribution of social contacts in a subscriber's profile.

REFERENCES

[1] Yves-Alexandre De Montjoye et al. "Unique in the crowd: The privacy bounds of human mobility". In: *Scientific reports* 3 (2013), p. 1376.

[2] Pierangela Samarati and Latanya Sweeney. *Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression*. Tech. rep. technical report, SRI International, 1998.

[3] Charu C Aggarwal. "On k-anonymity and the curse of dimensionality". In: *Proceedings of the 31st international conference on Very large data bases*. VLDB Endowment. 2005, pp. 901–909.

[4] Arvind Narayanan and Vitaly Shmatikov. "Robust de-anonymization of large datasets (how to break anonymity of the Netflix prize dataset)". In: *University of Texas at Austin* (2008).

[5] Anna L Buczak, Steven Babin, and Linda Moniz. "Data-driven approach for creating synthetic electronic medical records". In: *BMC medical informatics and decision making* 10.1 (2010), p. 59.

[6] Sibren Isaacman et al. "Human mobility modeling at metropolitan scales". In: *Proceedings of the 10th international conference on Mobile systems, applications, and services*. ACM. 2012, pp. 239–252.

[7] Darakhshan J Mir et al. "Dp-where: Differentially private modeling of human mobility". In: *2013 IEEE international conference on big data*. IEEE. 2013, pp. 580–588.

[8] Sibren Isaacman et al. "Identifying important places in people's lives from cellular network data". In: *International Conference on Pervasive Computing*. Springer. 2011, pp. 133–151.

[9] Chaoming Song et al. "Limits of Predictability in Human Mobility". In: *Science* 327.5968 (2010), pp. 1018–1021. ISSN: 0036-8075, 1095-9203. eprint: 0307014 (cond-mat). URL: http://www.sciencemag.org/cgi/content/full/327/5968/1018.