Opportunistic multi-party shuffling for data reporting privacy

Marios Fanourakis
CUI, Quality of Life Lab
University of Geneva, Switzerland
marios.fanourakis@unige.ch

Abstract

An important feature of data collection frameworks, in which voluntary participants are involved, is that of privacy. Besides data encryption, which protects the data from third parties in case the communication channel is compromised, there are schemes to obfuscate the data and thus provide some anonymity in the data itself, as well as schemes that 'mix' the data to prevent tracing the data back to the source by using network identifiers. This mixing is usually implemented by utilizing special mix networks in the data collection framework. In this paper we focus on schemes for mixing the data where the participants do not need to trust the mix network or the data collector with hiding the source of the data so that we can evaluate the efficacy of peer to peer mixing strategies in the real world. To achieve this, we present a simple opportunistic multi-party shuffling scheme to mix the data and effectively obfuscate the source of the data. We successfully simulate 3 cases with artificial parameters and then use the real-world Mobile Data Challenge (MDC) data to simulate an additional 2 scenarios with realistic parameters. Our results show that such approaches can be effective depending on the time constraints of the data collection and we conclude with design implications for the implementation of the proposed data collection scheme in real life deployments.

1 Introduction

Mobile crowd sensing leverages the number of user-companioned devices, including mobile phones, wearable devices, and smart vehicles, and their inherent mobility to collect information such as location, personal and surrounding context, noise level, and more [1]. The users, acting as sensors, have a certain expectation of privacy about the data they might be sharing and often do not trust that is it possible to hide their identity while at the same time provide usable data [2]. Providing data privacy in crowd sensing or other participatory data collection context has been an important task that ensures that the participants privacy is protected (ex., data cannot be traced to the individual) while the data is being collected at large scales without bias stemming from privacy-aspects (ex. participants switching off their phone in certain contexts). There are several elements of the data collection process that can be exploited to reveal sensitive information about the participants: the data communication channel, the reporting of the data, and the data itself.

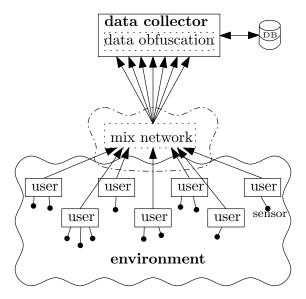


Figure 1: A diagram of a generic data collection scheme where users collect data in some environment and then send the data through an optional mix network that can either be a geographical zone in their environment or a separate network. The data is eventually communicated to the data collector who may chose to use data obfuscation techniques to provide privacy to the users. The communication channel is represented by arrows.

Securing the communication channel from third parties that might want to intercept the data can be achieved using data encryption techniques. However, unless some additional steps are introduced, the entity collecting the data (a researcher or a company), from here on referred to as the *data collector*, can link the data to its source. Giving pseudonyms to the participants can help mitigate this but it is still not completely safe. The data collector will still know that a certain batch of data belongs to a certain pseudonym which can be compromising depending on the content of the data. Even one piece of identifiable data will allow the data collector to know that all the data in that batch with the same pseudonym belongs to the same user. For this reason, mix networks were introduced in the data collection process. These networks mix the batches of data from each participant and send it to the data collector which then has no way to directly trace the source of a batch of data by essentially unlinking the data from all original network identifiers.

Note that mixing alone is not sufficient to guarantee privacy, the data itself can still be used to identify a participant. For this, there has been a lot of research in data obfuscation which provides some level of anonymity (k-anonymity, l-diversity, t-closeness, and others) which should be used in addition to mix networks. Other privacy and security related requirements are outlined by Giannetsos et al. [3], He et al. [4], and Christin et al. [5] however, these are out of the scope of this paper. These include privacy-preserving resilient incentive mechanisms and fairness (users should receive credits and rewards for their participation without associating themselves with the data or the tasks they contributed), communication integrity confidentiality and authentication (all entities should be authenticated and their communications should be protected from any alteration and disclosure to unauthorized parties), authorization and access control (participating users should act according to the policies specified by the sensing task), data-centric trust (Mechanisms must be in place to assess the trustworthiness and the validity of user submitted data), and accountability (entities should be held accountable for actions that

could disrupt the system operation or harm users).

As we will see in section 2 the concept of mix networks is not new, however, there are no studies that evaluate the performance of peer to peer opportunistic mixing using real world mobility data of the participants. In this paper, our goal is **not** to introduce a novel data mixing strategy but rather to focus specifically on evaluating the efficacy of mixing the data, as part of a slicing and mixing strategy, in a fully opportunistic way with the goal of achieving a uniform distribution of the data among all participants. We assume that participants are mobile and generate data by using sensors or answering surveys and that they regularly cross paths with at least one other participant in order to exchange data in a peer to peer manner. Furthermore, we require that all participants together form a connected graph with respect to who they meet. The mixing strategy we use is very basic so as to provide baseline results that can later be used to evaluate more complex strategies.

In section 2 we present some related work on privacy techniques and data mixing in the context of participatory sensing, in section 3 we describe a data mixing scheme, in section 4 we verify the mixing scheme in a simulated environment. Finally, in section 5 we summarize the results and provide design implications.

2 Related Work

There are several measures of anonymity for data content: k-anonymity is achieved when each data value occurs at least k times, l-diversity is a stronger privacy indicator and ensures that there are at least l well represented values for sensitive attributes based on entropy among other metrics, other stronger privacy indicators commonly used is t-closeness [6][7][8]. As we mentioned, these metrics are about the data content as a whole and are not relevant to this work which focuses on how to ensure privacy during data reporting at the source.

If only aggregate information is needed from a set of data, privacy-preserving data aggregation schemes have been proposed in order to safely provide aggregate information such as value averages, minimums, and maximums about the underlying data using homomorphic encryption or other techniques [9, 10, 11, 12, 13]. These aggregate values cannot be used by a researcher or other entity to train machine learning models like neural networks or SVMs which most often require the data to not be aggregated. It is advantageous to do the least amount of manipulation on the data in order to retain as much utility as possible.

A common practice when performing studies is to use pseudonyms for the participants while keeping the data at its pure form. However it has been shown overwhelmingly that this does not necessarily guarantee the privacy of the participants [14, 15, 16]. Other approaches use mix networks or mix zones to reassign pseudonyms to the participants or to mix data. These methods have been shown to be an effective way to protect the participant's identity by decoupling the data from the user who collected it [17, 18, 19]. Mix networks are well studied and quite robust at what they are designed to do, which is to shuffle the batches of data (i.e. permuting the order of the batches with respect to the pseudonyms). The limitation of this approach is that the participants who generate the data often have to trust the mix network and additionally, for many mix network designs, individual data entries remain in the original batch so that when the identity of the participant is discovered for one piece of data from a batch then the rest of the batch can be assumed to belong to that same participant. Mix zones are fixed in space and require that participants enter these zones to satisfy certain privacy aspects like k-anonymity by guaranteeing that the data is mixed among k participants making them unsuitable for opportunistic settings.

A novel approach to data privacy is slicing and mixing. First developed for wireless sen-

sor networks, it partitions the data horizontally and then mixes it before aggregating values (SMART) [20]. It was adapted for privacy in data publishing for human-generated data by partitioning the data both vertically and horizontally where in the former, care is taken to group highly correlated attributes together. Then these *slices* are permuted in order to break the linking between different columns [21]. Many works have extended slicing to be used in participatory sensing scenarios for privacy-preserving data aggregation[9] but only a few have looked into how the mixing might perform in real world environment with real people (ex. Qiu et al. [22] use taxi traces to simulate participants) and none to our knowledge do an analysis of the effectiveness of mixing when it comes to opportunistic peer to peer (P2P) mixing scenarios. Christin et al. [23, 24] also used P2P techniques to obfuscate location information however, their evaluation does not generalize to other types of data.

3 Privacy-conscious Data Shuffling

From here on we will refer to the participants (users in Figure 1) who are generating the data as $sensor\ nodes$ or just nodes. Our simple opportunistic shuffling scheme consists of the following steps: Once a node i meets another node j, they randomly select some of their data in order to swap it with each other. The nodes follow their regular mobility patterns and are exchanging data with each other when they come into direct communication range of each other until certain stopping criteria relating to the state of the shuffle are met. As we mentioned earlier, we require that the nodes form a connected graph so that the entirety of the data can be uniformly shuffled. If there are any disconnected subsets of nodes the data will have no way to be communicated between those subsets, only within them.

3.1 Data exchange

When two nodes come within communication range of each other, each node randomly selects half of the data they have in their possession, M, to exchange with the other. This value can be adjusted individually on each node to adjust their data exposure and either reduce or increase the potential amount of personal data that they might share at each transaction. The specifics of this adjustment are not explored in this paper and we keep the amount of data that each node exchanges at $\frac{M}{2}$ as it is optimal for reaching a uniform distribution of data in the least number of shuffles. This fact can be easily verified by looking at the number of ways there are to choose k data from M given by the binomial coefficient which can be calculated using the equation below:

$$\binom{k}{M} = \frac{M!}{k! (M-k)!} \tag{1}$$

Each shuffle becomes more random as the binomial coefficient increases in value. If we set k to be some fraction $x \in [0,1]$ of the total data M, k = xM, then we can find that the value of x which maximizes the result in the equation above is $\frac{1}{2}$.

3.2 Stopping criteria

There are two different stopping criteria that can be used to signify that the data has been sufficiently shuffled (uniformly distributed) and that it is safe to send it to the data collector.

The first one is based on each nodes perception of how well the data is mixed. Each node can keep track of the data that they come into contact with and measure the probability that they encounter some specific piece of data. Since the data may be encrypted, the nodes must keep track of the encrypted data or a shorter hashed version of the encrypted data which can

come paired up with the encrypted data. Once the probability is close to being uniform across all data, then they can stop the shuffling process since this indicates a near uniform mix. This might work well when there are not many nodes, but as the number of nodes increases, the time it takes to verify the uniformity of the mix also increases.

The second set of stopping criteria is based on the properties of the graph like closeness centrality. If, in addition to data, the nodes exchanged information about their graph connections (nodes that they have previously encountered) or there is prior knowledge about the graph, then they can estimate the number of exchanges that they need to perform before the data is near uniformly mixed.

Closeness centrality to estimate stopping criteria. Closeness centrality is a measure of the degree to which an individual node is near all other nodes in the network. In order for each node to calculate its closeness centrality it needs to know its distance to all other nodes. This is trivial when there is global knowledge about the graph, however, it may not always be the case, especially when there is no trusted party to provide this information. When there is no prior graph knowledge, each node i needs to communicate its personal adjacency matrix A^i in addition to the data at each exchange. A^i should initially indicate which nodes are directly connected (one hop) along with the edge weight as it is calculated by the node (in this case, edge weight is equivalent to the number of times that the node has encountered each of its one hop neighbors). Then the node can update A^i by combining all the personal adjacency matrices it has acquired $(A^j, A^k,$ etc.) from other nodes. To update A^i when the node receives another node's personal adjacency matrix A^j we perform the following operation:

Algorithm 1: Procedure to combine A^i with A^j

```
Data: A^i, A^j
Result: updated A^i

1 if A^i_{k,l} = \emptyset and A^j_{k,l} \neq \emptyset then
2 A^i_{k,l} \leftarrow A^j_{k,l}
3 end
4 if A^i_{k,l} \neq \emptyset and A^j_{k,l} \neq \emptyset then
5 A^i_{k,l} \leftarrow min\left(A^i_{k,l}, A^j_{k,l}\right)
6 end
```

Finally, performing a shortest path algorithm such as the Floyd-Warshall or Dijkstra algorithm can reduce the redundancies and update the paths in A^i . The process is illustrated in figure 2

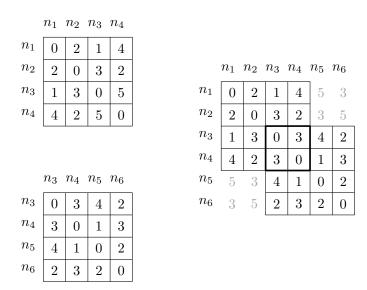


Figure 2: On the left, two personal adjacency matrices. On the right, the combined outcome where overlapping weights are assigned the minimum of the two values. Grey values indicate values calculated by the shortest path algorithm that have not been directly measured otherwise.

The closeness centrality can then be calculated for each node and by each node using the information in their respective adjacency matrix.

The stopping criteria for the number of exchanges necessary to sufficiently mix the data can then be estimated from the adjacency matrix and closeness centrality information using empirical data which we will show in section 4.

It is important to note that if the graph of the nodes is known to everyone, encrypting the communication channel becomes even more vital for the protection of the security and privacy of nodes against malicious nodes.

4 Experiment and Results

4.1 Experimental setup

To verify our data mixing scheme, we perform simulations using artificial parameters as well as simulations using real mobility data from the MDC dataset. The data mixing occurs in shuffling rounds that consist of either a group of markov chain state transitions (representing data exchanges) based on the transition matrix or a full day (24 hours) of proximity events in the real mobility data simulations.

At each time t, a node i will exchange data with a node j either with a probability based on the $A_{i,j}$ element of the transition matrix A for the artificial parameter simulations or based on the proximity of the two nodes in the real mobility data simulations. At each shuffle we take note of what data each node has.

In order to get representative probability distributions of the data, we run 30000 trials of the simulation with each of the parameter sets (a parameter set consists of the following: the number of nodes, data size per node, and the transition matrix or proximity events). This number was selected because it gives us a confidence level of 99% based on the equation $n \geq \frac{log(a)}{log(1-p)}$ to calculate the number of trials necessary given the probability of the occurrence of an event p and

the confidence level 1-a that we require. In our case we seek to be confident of events that occur with a probability of at least p=0.00015, that is to say that our probability distributions in our results will have a granularity of 0.00015. We choose 1-a=0.99 which is equivalent to being 99% confident in our results.

For our experiments, the total number of nodes, N, and the amount of data items, M, that they start with is selected as described in section 4.1.3. To keep track on how the data flows throughout the network we make sure that each node's initial data is uniquely identifiable by labeling them with integers. For example, if we set the number of data to 6, then node 1's data will consist of the numbers from 1 to 6, node 2's data will consist of the numbers from 7 to 12, etc.. In this way we can easily check how uniformly the data has been distributed by evaluating the probability distribution of each number being in any specific node at the end of the shuffle.

4.1.1 Artificial Parameter Simulations Setup

Initially, we performed simulated experiments with artificial parameters to illustrate and validate the data shuffling procedure. Node mobility is artificially simulated by using three different Markov models where each one is defined by a transition matrix A. The three models consist of a best case scenario transition matrix (equivalent to a group of co-workers or students enrolled in the same course), an intermediate case scenario (equivalent to shift-based co-workers), and a worst case scenario transition matrix (equivalent to otherwise unrelated commuters crossing paths on their way to their individual workplaces). The specifics of the transition matrices are described in section 4.1.3.

4.1.2 Real Mobility Data (MDC) Simulations Setup

We use real user mobility traces from the data of the Mobile Data Challenge (MDC) which include GPS traces of real mobile users [25, 26]. The data we used included 191 users and spanned over a year of data for some of the users. To normalize and be able to compare with the artificial cases, we define one shuffling round as a single day and we analyze up to 100 contiguous days (i.e. 100 shuffling rounds) of GPS traces for each trial.

For some users, we might have less than 100 days of data, when we reach the end of the data without having completed the 100 shuffling rounds we cycle from the beginning until we reach the desired number. For example, if a user set only has 50 days worth of data, we will go though his GPS data twice to complete the 100 day trial.

In most cases we have more than 100 days of data for each user set. In this case, since we limit our simulations to 100 shuffling rounds consisting of 100 contiguous days, we make sure that we select 100 contiguous days when the user set is sufficiently active based on two criteria in order of priority: the median of the number of proximity events between all pairs in the user set, and the total number of proximity events.

We assume an exchange of data between two users can be performed under the following conditions:

- They are within 50 meters or less of each other. We call this a *proximity event* since they are within direct communication range of each other.
- They have not exchanged data between each other in the past 30 minutes.

In these experiments we do not consider the bandwidth or throughput of the data transmission and assume that it can be instantaneously exchanged when two users are within communication range.

For each of the 30000 trials for the MDC data simulations, a random subset of N users was selected from the 191 in such a way that they formed a connected graph with a minimum edge weight of 10 (in this case, edge weight indicates the number of exchanges between two users during the entire duration of the study). With this random selection, when we use N=10 the average number of hops between the two most remote users was 9 and the median number of hops between any two users was 3 (which resembles a line topology). The user set was unique in each trial of our simulations, i.e. no two trials had the same set of 10 random users.

We ran an additional simulation with the MDC dataset in which instead of selecting N random users, we selected N users that formed a clique (i.e. a fully connected topology with maximum distance of one hop between any two users). Again, we limited the edge weight to be equal to or above 10. During our experiments we discovered that there were not enough cliques of size $\geq N$ in the dataset to justify doing 30000 trials. The total number of maximal cliques (cliques that are not subsets of larger cliques) in the dataset is 890 and the number of cliques with size of at least N is often much smaller than that. It is redundant to perform more trials than there are number of cases because this means that the same case will need to be repeated several times to reach the desired amount of trials. However, to get meaningful statistics it was necessary to do a much larger number of trials than there were number of cliques. To remedy this, we relaxed the requirement for the cliques and allowed ourselves to combine cliques to form a set of N users. The exact procedure by which we combined the cliques is described in Algorithm 2. The **if** statement on lines 4-6 is optional and serves to limit the minimum size of the cliques that form the user set thus ensuring a better connected user set. This algorithm allowed us to

Algorithm 2: Procedure to combine cliques

```
Data: The user cliques
   Result: A well connected user set
 1 userSet \leftarrow \emptyset
   while size(userSet) < N do
 3
       cliq \leftarrow random \, clique
       if size(cliq) < 0.5 (N - size(userSet)) then
 4
           go to line 3
5
       \mathbf{end}
 6
       userSet \leftarrow userSet \cup cliq
 7
8
       if size(userSet) > N then
           userSet \leftarrow select \ N \ users \in userSet
9
       end
10
11 end
12 if userSet not connected then
       go to line 1
13
14 end
```

generate much more than 30000 different user sets as evidenced by the results in our simulations where for N=10 there were no two trials with the same user set in all of the 30000 trials. Furthermore, N=10 resulted in user sets with the average number of hops between the two most remote users at 4 and a median number of hops between any two users of 1.

4.1.3 Parameter Sets

The number of nodes and number of data items for the MDC data simulations was selected after the artificially simulated cases where we verified that the number of data items did not

Sim. type	Transition Matrix	{# of Nodes, # of Data}
best case	fully connected topology w/ prob of no transaction 0 and prob of transation with each of the other nodes $\frac{1}{N-1}$	{10,6} {30,6} {10,20}
intermediate case	line topology w/ prob of no transaction 0.5 for edge nodes and 0 for all other nodes	{10,6} {30,6} {10,20}
worst case	line topology w/ prob of no transaction 0.8 for edge nodes and 0.6 for all other nodes	{10,6} {30,6} {10,20}
MDC data random	GPS traces of a random selection of users	{10,6}
MDC data cliques	GPS traces of cliques of users	{10,6}

Table 1: Simulations performed showing total nodes and total amount of data per node for each simulation. There are 11 simulations in total.

significantly affect the number of shuffles needed since we always exchange half of a node's total data (as per the protocol discussed in section 3.1). We chose to simulate only 2 representative cases with the MDC dataset: choosing a connected set of random users, or choosing users that form cliques in the adjacency matrix. Other cases would be redundant since we already show the effects of changing the number of nodes and data items with the artificial parameter simulations. All simulations are summarized in Table 1.

4.2 Performance Criteria

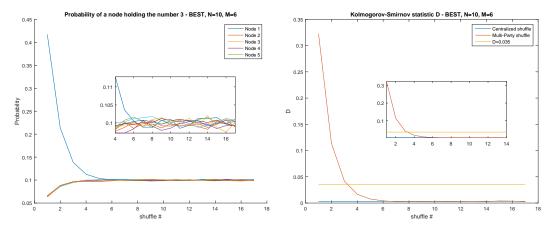
The performance criteria that we mainly use is the Kolmogorov-Smirnov test with a uniform distribution of $\frac{1}{N}$ as the reference distribution. With this test, we measure the absolute error between the distribution of the data in our experiment and the ideal uniform distribution. As a result of our experimental setup we are able to perform this test after every shuffling round in our experiment allowing us to see the exact number of shuffles needed to achieve a near uniform mix.

For illustrative purposes we first take a look at the probability of holding a specific data item for each node at each shuffling round.

4.3 Results

4.3.1 Results Using Artificial Parameters

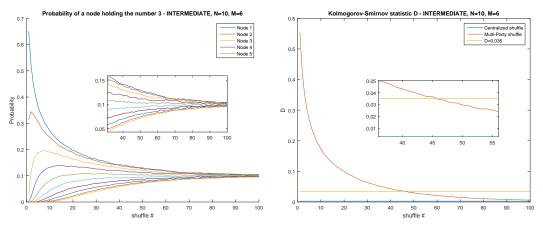
Intuitively, the more shuffles we do then the more uniform the distribution of the data should be. This intuition is verified in figure 3a where we clearly see that the probability of holding a specific data item (as an example we use the data item with number 3) approaches an ideal probability with amplitude $\frac{1}{N}$ as the number of shuffles increases, where N is the number of nodes. Since node 1 is the initial holder of the data item with number 3, it starts with the highest probability in the initial stages and as it shares data with all the other nodes the probability evens out.



(a) The probability distribution of the number 3 (b) The results of the Kolmogorov-Smirnov test at being at different nodes at different number of different number of shuffles (subplot is of a magnified region) fied region)

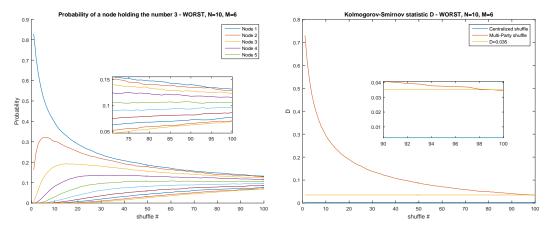
Figure 3: Results for the best case scenario for N=10 and M=6

The same is true for the intermediate and worst case of the line topology as we can see in figures 4a and 5a, although, in this case it takes more than 40 shuffling rounds to reach the same ideal probability for each of the cases. Similarly to the best case, we notice that node 1 starts out with higher probability of holding the data with number 3, and then we notice a sharp increase on the probability of node 2 holding it since it is the only node that is connected to node 1 (recall that intermediate and worst case scenarios have the line topology of nodes).



(a) The probability distribution of the number 3 (b) The results of the Kolmogorov-Smirnov test at being at different nodes at different number of different number of shuffles (subplot is of a magnified region) fied region)

Figure 4: Results for the intermediate case scenario for ${\cal N}=10$ and ${\cal M}=6$

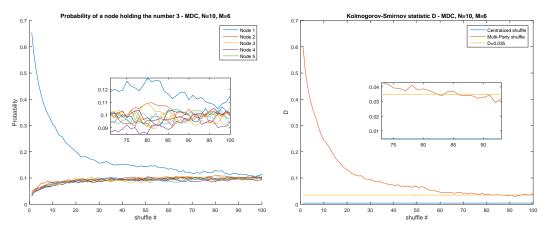


(a) The probability distribution of the number 3 (b) The results of the Kolmogorov-Smirnov test at being at different nodes at different number of different number of shuffles (subplot is of a magnified region) fied region)

Figure 5: Results for the worst case scenario for N=10 and M=6

4.3.2 Results using MDC Dataset

In figures 6a and 7a we see the results of the MDC dataset simulations. For the MDC simulation with the random selection of users, although the connectivity resembles that of line topology, we cannot see it in figure 6a (like in figures 4a and 5a) because the users are not ideally ordered at each trial to reveal the same pattern as the artificial parameter simulation with line topology (recall that they were randomly selected).

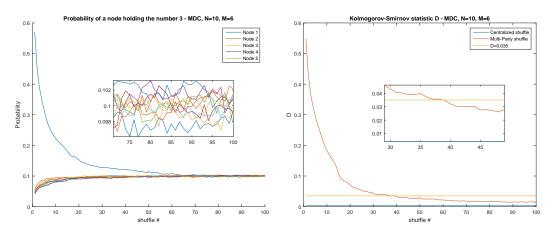


(a) The probability distribution of the number 3 (b) The results of the Kolmogorov-Smirnov test at being at different nodes at different number of different number of shuffles (subplot is of a magnified region) fied region)

Figure 6: Results for the MDC data with random user selection for N=10 and M=6

Simulation type	# of shuffles for $D < 0.035$
Best case	4
Intermediate case	46
Worst case	100
MDC data random	85
MDC data cliques	40

Table 2: Summary of Kolmogorov-Smirnov test results for N = 10 M = 6.



(a) The probability distribution of the number 3 (b) The results of the Kolmogorov-Smirnov test at being at different nodes at different number of different number of shuffles (subplot is of a magnified region) fied region)

Figure 7: Results for the MDC data with clique user selection for N=10 and M=6

4.3.3 Kolmogorov-Smirnov test of the experiment

As we can see in figures 3b, 4b, 5b, 6b, and 7b, the error decreases as the number of shuffles increases. For our experimental setup, 4 shuffles is sufficient to adequately shuffle the data in the best case scenario while more than 60 shuffling rounds are needed in order to reach a similar distribution of the data for the worst case scenario. The MDC dataset simulation with random user selection is comparable to the worst case scenario while with clique based user selection it is better than the intermediate case but worse than the best case scenario.

4.3.4 Effects of varying number of nodes and number of data items

For the fully connected topology (best case), varying the number of nodes or number of data items does not seem to have an effect on performance. For the line topology (intermediate and worst case), increasing the number of nodes also increases the number of shuffles necessary. However increasing the number of data items does not have a noticeable effect for those cases. These conclusions can be verified in the figures 8.9, and 10 which show the Kolmogorov-Smirnov statistic as function of the shuffles for different selection of total nodes N and total data items M.

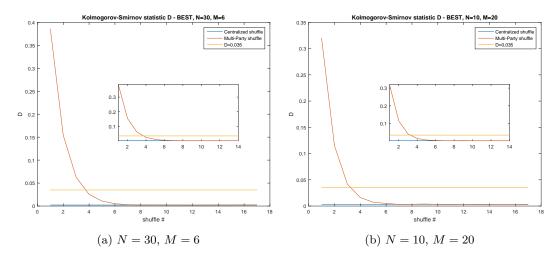


Figure 8: Kolmogorov-Smirnov test for different selection of N and M of the best case scenario

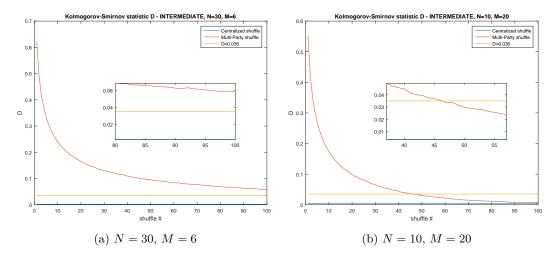


Figure 9: Kolmogorov-Smirnov test for different selection of N and M of the intermediate case scenario

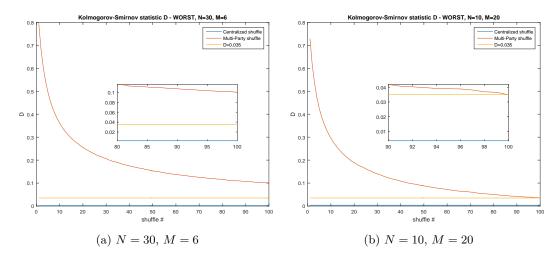


Figure 10: Kolmogorov-Smirnov test for different selection of N and M of the worst case scenario

5 Conclusion and Future Work

In this paper, we evaluated a basic peer to peer opportunistic mixing strategy in order to generate a baseline of results which can later be used to compare different strategies. We did this by opportunistically shuffling the data among the participants and showed that the number of shuffles is dependent on the properties of the graph that represents the participant interconnections. A fully connected topology requires only 4 shuffling rounds. On the other hand, a line topology required significantly more shuffling rounds; 46 rounds for the intermediate case and 100 shuffling rounds for the worst case. Using real user GPS traces from the MDC dataset we saw that the number of shuffling rounds did not exceed the worst case when selecting random nodes from the population but at 85 rounds it was significantly higher than the intermediate case. Carefully selecting nodes from the population in the MDC dataset to form a more connected topology made a significant difference in the efficiency of the shuffling (only 40 shuffling rounds) and was significantly better than the intermediate case.

These results can be used to define the stopping criteria for a near uniform shuffle based on the topology of the nodes. A set of 10 nodes in a fully connected topology (having an average closeness centrality of 1) would require at least 4 shuffles. Whereas a set of 10 nodes in a line topology (having an average closeness centrality of 0.3430), would require 100 shuffles. For the MDC dataset the closeness centrality ranged from 0.3430 to 1 and from 0.6 to 1 for the random user selection and the clique-based user selection respectively.

Opportunistic peer to peer mixing, as part of a slicing and mixing strategy, can therefore reasonably mix the data so as to protect the identity of the source in the context of the data routing. However, the data content itself should be further obfuscated in order to protect the identity of the source which might be revealed from analyzing the data content. Such techniques require the manipulation of data entries and can reduce the quality of the data but it is often necessary to do so for the protection of the participants.

Based on the positive results of these simulations, we plan to implement this in a real study to verify the results with mobility data that is collected in a different city.

References

- [1] B. Guo, Z. Yu, X. Zhou, and D. Zhang, "From participatory sensing to Mobile Crowd Sensing," in 2014 IEEE International Conference on Pervasive Computing and Communication Workshops (PERCOM WORKSHOPS), pp. 593–598, IEEE, mar 2014.
- [2] M. Gustarini, K. Wac, and A. K. Dey, "Anonymous smartphone data collection: factors influencing the users' acceptance in mobile crowd sensing," *Personal and Ubiquitous Computing*, vol. 20, no. 1, pp. 65–82, 2016.
- [3] T. Giannetsos, S. Gisdakis, and P. Papadimitratos, "Trustworthy People-Centric Sensing: Privacy, security and user incentives road-map," in 2014 13th Annual Mediterranean Ad Hoc Networking Workshop (MED-HOC-NET), pp. 39–46, IEEE, jun 2014.
- [4] D. He, S. Chan, and M. Guizani, "User privacy and data trustworthiness in mobile crowd sensing," *IEEE Wireless Communications*, vol. 22, pp. 28–34, feb 2015.
- [5] D. Christin, "Privacy in mobile participatory sensing: Current trends and future challenges," *Journal of Systems and Software*, vol. 116, pp. 57–68, jun 2016.
- [6] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, pp. 557–570, oct 2002.
- [7] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "L-diversity: privacy beyond k-anonymity," in 22nd International Conference on Data Engineering (ICDE'06), vol. 2006, pp. 24–24, IEEE, 2006.
- [8] N. Li, T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity," in 2007 IEEE 23rd International Conference on Data Engineering, no. 3, pp. 106–115, IEEE, apr 2007.
- [9] J. Shi, R. Zhang, Y. Liu, and Y. Zhang, "PriSense: Privacy-Preserving Data Aggregation in People-Centric Urban Sensing Systems," in 2010 Proceedings IEEE INFOCOM, pp. 1–9, IEEE, mar 2010.
- [10] Z. Wei, B. Zhao, and J. Su, "PDA: A Novel Privacy-Preserving Robust Data Aggregation Scheme in People-Centric Sensing System," *International Journal of Distributed Sensor Networks*, vol. 9, p. 147839, nov 2013.
- [11] Rui Zhang, Jing Shi, Yanchao Zhang, and Chi Zhang, "Verifiable Privacy-Preserving Aggregation in People-Centric Urban Sensing Systems," *IEEE Journal on Selected Areas in Communications*, vol. 31, pp. 268–278, sep 2013.
- [12] S. M. Erfani, S. Karunasekera, C. Leckie, and U. Parampalli, "Privacy-preserving data aggregation in Participatory Sensing Networks," in 2013 IEEE Eighth International Conference on Intelligent Sensors, Sensor Networks and Information Processing, vol. 1, pp. 165–170, IEEE, apr 2013.
- [13] Y. Zhang, Q. Chen, and S. Zhong, "Privacy-Preserving Data Aggregation in Mobile Phone Sensing," *IEEE Transactions on Information Forensics and Security*, vol. 11, pp. 980–992, may 2016.
- [14] M. P. Scipioni and M. Langheinrich, "I' m Here! Privacy Challenges in Mobile Location Sharing," *IWSSI/SPMU*, 2010.

- [15] D. Christin, A. Reinhardt, S. S. Kanhere, and M. Hollick, "A survey on privacy in mobile participatory sensing applications," *Journal of Systems and Software*, vol. 84, pp. 1928–1946, nov 2011.
- [16] M. Shin, C. Cornelius, A. Kapadia, N. Triandopoulos, and D. Kotz, "Location Privacy for Mobile Crowd Sensing through Population Mapping," Sensors, vol. 15, pp. 15285–15310, jun 2015.
- [17] D. L. Chaum, "Untraceable electronic mail, return addresses, and digital pseudonyms," Communications of the ACM, vol. 24, pp. 84–90, feb 1981.
- [18] C. Cornelius, A. Kapadia, D. Kotz, D. Peebles, M. Shin, and N. Triandopoulos, "Anony-sense," in *Proceeding of the 6th international conference on Mobile systems, applications, and services MobiSys '08*, (New York, New York, USA), p. 211, ACM Press, 2008.
- [19] C. A. Neff, "A verifiable secret shuffle and its application to e-voting," in *Proceedings of the 8th ACM conference on Computer and Communications Security CCS '01*, (New York, New York, USA), p. 116, ACM Press, 2001.
- [20] W. He, X. Liu, H. Nguyen, K. Nahrstedt, and T. Abdelzaher, "PDA: Privacy-Preserving Data Aggregation in Wireless Sensor Networks," in *IEEE INFOCOM 2007 - 26th IEEE International Conference on Computer Communications*, pp. 2045–2053, IEEE, 2007.
- [21] T. Li, N. Li, J. Zhang, and I. Molloy, "Slicing: A New Approach for Privacy Preserving Data Publishing," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, pp. 561–574, mar 2012.
- [22] F. Qiu, F. Wu, and G. Chen, "SLICER: A Slicing-Based K-Anonymous Privacy Preserving Scheme for Participatory Sensing," in 2013 IEEE 10th International Conference on Mobile Ad-Hoc and Sensor Systems, pp. 113–121, IEEE, oct 2013.
- [23] D. Christin, J. Guillemet, A. Reinhardt, M. Hollick, and S. S. Kanhere, "Privacy-Preserving Collaborative Path Hiding for Participatory Sensing Applications," in 2011 IEEE Eighth International Conference on Mobile Ad-Hoc and Sensor Systems, pp. 341–350, IEEE, oct 2011.
- [24] D. Christin, A. Reinhardt, and M. Hollick, "On the efficiency of privacy-preserving path hiding for mobile sensing applications," in 38th Annual IEEE Conference on Local Computer Networks, pp. 818–826, IEEE, oct 2013.
- [25] N. Kiukkonen, J. Blom, O. Dousse, D. Gatica-Perez, and J. Laurila, "Towards rich mobile phone datasets: Lausanne data collection campaign," *Proceedings ACM International Conference on Pervasive Services (ICPS)*, vol. Berlin, 2010.
- [26] J. K. Laurila, D. Gatica-Perez, I. Aad, J. Blom, O. Bornet, T.-M.-T. Do, O. Dousse, J. Eberle, and M. Miettinen, "The mobile data challenge: Big data for mobile computing research," Proceedings of the Workshop on the Nokia Mobile Data Challenge, in Conjunction with the 10th International Conference on Pervasive Computing, pp. 1–8, 2012.