Delphine Reinhardt, Daniel M. Bub, Andrey Moerov, Saffija Kasem-Madani: *A Distributed Privacy-Preserving Mechanism for Mobile Urban Sensing Applications*. In: Proceedings of the 10th IEEE Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), pp. 1-6, 2015.

A Distributed Privacy-Preserving Mechanism for Mobile Urban Sensing Applications

Delphine Christin*†, Daniel M. Bub‡, Andrey Moerov*, Saffija Kasem-Madani*
* Computer Science IV, University of Bonn, Bonn, Germany
† Fraunhofer Institute for Communication, Information Processing and Ergonomics, Wachtberg, Germany
‡ Secure Mobile Networking Lab, Technische Universität Darmstadt, Darmstadt, Germany
Emails: christin@cs.uni-bonn.de, daniel.bub@seemoo.tu-darmstadt.de, moerov@informatik.uni-bonn.de, kasem@cs.uni-bonn.de

***Abstract*—In urban sensing applications, participants carry mobile devices that collect sensor readings annotated with spatiotemporal information. However, such annotations put the participants' privacy at stake, as they can reveal their whereabouts and habits to the urban sensing campaign administrators. A solution to protect the participants' privacy is to apply the concept of k-anonymity. In this approach, the reported participants' locations are modified such that at least $k-1$ other participants appear to share the same location, and hence become indistinguishable from each other. In existing implementations of k-anonymity, the participants need to reveal their precise locations to either a third party or other participants in order to find $k-1$ other participants. As a result, the participants' location privacy may still be endangered in case of ill-intentioned third-party administrators and/or participants. We tackle this challenge by proposing a novel approach that supports the participants in their search for other participants without disclosing their exact locations to any other parties. To evaluate our approach, we conduct a threat analysis and study its feasibility by means of extensive simulations using a real-world dataset.**

I. Introduction

Urban sensing applications leverage mobile devices, such as mobile phones, to collect sensor data. Due to the ubiquity of these devices, urban phenomena, such as noise pollution [1] or traffic conditions [2], can be monitored in unprecedented detail. In comparison to existing static sensing stations, the data collection is improved in both quality and quantity, but in turn, the privacy of participants carrying mobile devices is put at stake. In fact, most of the collected sensor readings are usually tagged with time and location information [3]. Consequently, urban sensing applications do not solely record the urban phenomena of interest, but also the participants' whereabouts. This raises two key risks for the participants' privacy. First, their visited locations are disclosed to the campaign administrators, hence threatening their location privacy. Secondly, a further analysis of their visited locations (such as frequency, duration, time of the day) can reveal further sensitive insights about the participants, such as their medical condition or political views [4]. In presence of such risks, users may prefer renouncing their participation rather than putting their privacy at stake. Protecting the participants' privacy is therefore mandatory to ensure an application's viability, as it primarily depends on the participants' willingness to contribute data.

To protect the participants' location privacy, most existing methods tailored to urban sensing applications are based on the concept of k-anonymity [5]. Its key idea is to protect the participants' location privacy against campaign administrators by replacing their actual location by one shared with at least $k-1$ other participants. To find these participants and compute the corresponding common location, several approaches have been proposed in the context of urban sensing applications. However, they require that the participants disclose their original locations to either a third party or other participants. The participants therefore need to trust these parties not to breach their location privacy. In this paper, we tackle this issue by making the following contributions:

1) We first present a novel solution that enables participants to find the remaining $k-1$ participants in a distributed fashion without revealing their original locations to any other parties involved in the urban sensing applications. In addition to protect the participants' location privacy against malicious administrators and participants, our scheme prevents the campaign administrators from inferring which participants have collected the reported sensor readings.
2) We next conduct a detailed threat analysis showing that the participants do not disclose their original locations to neither other participants nor the third-party administrators nor the campaign administrators. In case of an active attack, our approach ensures that malicious participants can at most gain access to the cloaked locations of honest participants.
3) We finally evaluate the feasibility of our approach by means of extensive simulations based on the GeoLife dataset [6], [7], [8]. To this end, we evaluate the impact of different parameters on the number of groups of k participants that can be built using our approach.

The paper is organised as follows. In Section II, we detail related work. We introduce our system and threat models in Section III. We present the architecture of our approach in Section IV. We conduct a threat analysis in Section V and present our evaluation results in Section VI. We finally make concluding remarks and discuss future work in Section VII.

II. Related Work

Different approaches based on k-anonymity [5] have been proposed to protect the participants' location privacy. The idea is to build sets of k participants sharing the same location so that these participants become indistinguishable for the campaign administrators. Most solutions, however, differ in the methods applied to build the different participants' sets. For example, centralised approaches rely on a central trusted third party (TTP) to which participants report their encrypted precise location. The TTP then computes the smallest region containing at least $k-1$ other participants to fulfil the k-anonymity property. In [9], the TTP directly replaces the participant's location with the computed region in the participant's message before forwarding it to its final destination. Alternatively, it returns the computed region to each participant, who uses it in lieu of his actual location in future messages as proposed in [10]. Building on this concept, different optimisations have been introduced. For example, [11] maintains the locality of the computed region as close as possible to the actual participants' location. Depending on the participants' spatial distribution, fewer than k participants may be located in the same area. Instead of waiting for the remaining participants and hence increasing the reporting latency, [12] proposes to introduce a timeout, which determines when the region is computed independently on the number of participants. In all schemes, the participants hence need to trust the TTP not to breach their location privacy by, e.g., disclosing their locations to unauthorised parties.

To overcome this issue, several decentralised approaches leveraging ad hoc communication between participants' devices have been presented. In the case of location-based services, [13] proposes to locally cache the results of the requests answered by the server. By doing so, the participants' location is only disclosed to the server when the requested information is not already available from nearby devices. Instead of caching information, devices can search for nearby participants such that the k-anonymity requirement could be fulfilled [14], [15]. The area containing all k devices is then locally computed, and used as a cloaked location in the request to the server. Instead of ad hoc-based communication, [16] uses peer-to-peer communication for the potential identification of k nearby peers. However, in all these solutions, participants reveal their locations to other participants and hence need to trust their peers not to misuse the disclosed information. We tackle this issue by proposing a novel scheme that shares the more similarities with [16] but kept the participants' locations hidden from both other participants and third parties.

III. Assumptions

We make the following assumptions regarding our system model and the related threats to privacy.

A. System Model

We assume an urban sensing system including multiple clients (e.g., smartphones) carried by the participants and an application server managed by the campaign administrators. We further assume that all parties own a public/private key pair. To fulfil sensing tasks, the clients collect sensor readings in form of triplets. Each triplet T has a unique identifier ID_T and consists of the time of the measurement, its location noted l, and the measured sensor data. Additional processing, such as noise filtering or feature extraction, can be locally applied on s before the triplets are reported to the application server. The application server then processes the received triplets to, e.g., build maps and compute statistics. The results are finally made available to the end users who can be the participants themselves or any interested parties depending on the application scenario.

B. Threat Model

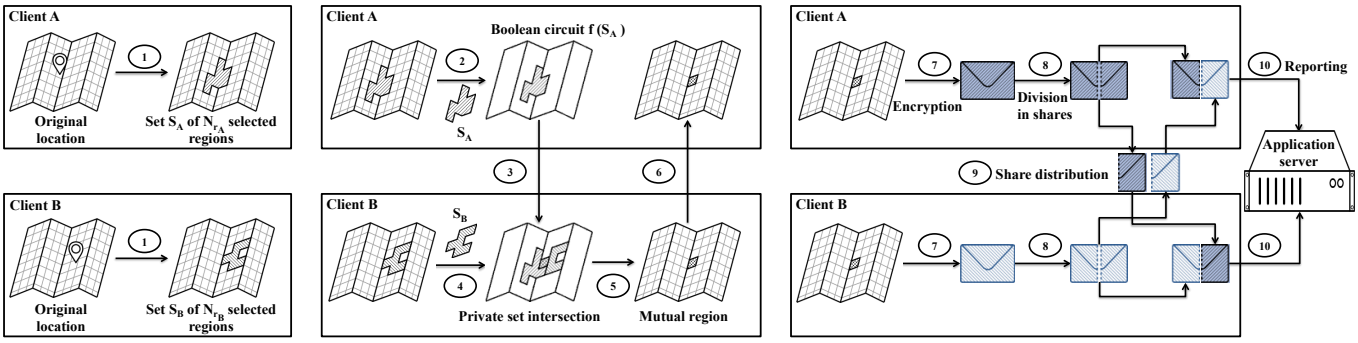
We assume a honest-but-curious adversary model, in which the campaign administrators attempt to passively breach the privacy of the participants, but runs the system normally and faithfully. This means that the campaign administrators focus on the data reported by the participants to the application server in order to breach their privacy. They, however, do not launch active attacks (such as collusions with malicious participants) to obtain further information. As an artefact of our approach, participants as well as the administrators of the third party introduced in our solution can also be interested in inferring the location of (other) participants. Like for the campaign administrators, we assume that the third party administrators are honest-but-curious. We further consider malicious participants who can launch active attacks against peers.

IV. Architecture

Our approach follows the three main steps illustrated in Fig. 1 and detailed in the following sections.

1) Partitioned Spatial Cloaking: We first assume a client A. It first cloaks its location l using the common spatial partition as depicted in Fig. 1(a). This means that it selects a set $S_A = \{r_{A_0}, r_{A_1}, ...r_{A_{(N_{r_A}-1)}}\}$ of N_{r_A} regions r_{A_i} (with $0 \leq i \leq N_{r_A}$ and $N_{r_A} \geq 1$). The number of selected regions N_{AN} depends on the participants' privacy preferences. In the case of $N_{r_A} = 1$, the cloaked location only corresponds to the region including l, while it includes additional adjacent regions for $N_{r_A} > 1$. Choosing multiple regions reduces the accuracy of l, and hence the precision of the application results. Simultaneously, it increases both the probability to find other participants sharing the same region(s) and the participants' privacy protection. The region selection can be done either randomly or by the participants, who can manually discard sensitive regions. A second client B does the same.

2) Private Location Matching: To protect their location k-anonymity, the clients then look for $k-1$ other participants that may share the same region(s) as depicted in Fig. 1(b). A third party in form of a publish-subscribe server supports this search by enabling direct communication between clients. Each pair of clients apply a private set intersection mechanism, such as [17]. The goal is hence to compare their respective set of regions s_A and s_B using a function f that returns the



(a) Partitioned spatial cloaking (b) Private location matching (c) Joint information disclosure

Fig. 1. Three key steps of our architecture

regions common to both A and B without disclosing their sets to each other. To this end, A converts s_A into a boolean circuit $f(s_A)$ using Yao's garbled circuit technique [18] and transfers this circuit to B. Note that A does not reveal s_A to B as it is only included in the circuit. B then applies the received circuit to s_B and only sends the result to A. The result is the mutual region(s) shared by both A and B. In other words, neither A nor B gain access to unshared regions at any steps of the protocol. After having run the private matching scheme, two cases are possible: (1) $n < k-1$ or (2) $n \geq k-1$, with n being the number of successful matches. In the first case, the client can wait until $n \geq k-1$, but this increases the latency until the sensor readings can be reported to the application server. Alternatively, the client can directly report T using its cloaked location obtained in the above step (see Sec. IV-1) or even drop it to protect its privacy. This would reduce the reporting latency, but both the participant's location k-anonymity and the originator's k-anonymity would not be guaranteed. If $n \geq k-1$, the matching process is completed and the matched clients replace each l by the intersecting region(s) in their respective T.

3) Joint Information Disclosure: To further ensure that the application administrators cannot infer who has collected the sensor readings, each client encrypts its sensor reading T into $m \geq k$ shares according to the secret sharing algorithm introduced in [19]. Then, it encrypts each share along with ID_T (i.e., the unique identifier of T) using the application server's public key. Each client keeps one share and distributes the $m-1$ remaining ones to the other group members as illustrated in Fig. 1(c). Since the shares are encrypted with the application server's public key, the shares are hidden from the other group members. Then, each client reports its own share and $m-1$ shares from the other group members to the application server. According to [19], the application server is only able to decrypt each T if at least $p \leq m$ shares have been reported. In our case, we especially choose $p \geq k$. This first ensures that at least k clients have reported shares to the application server, i.e., at least k-anonymity is guaranteed. Additionally, this supports missing shares, e.g., if clients opt out before the mechanism completion. Once T is decrypted, the application server can further process it as usual and compute result summaries, for example.

V. Threat Analysis

We consider the threat model presented in Section III-B and argue that our solution is resilient against the following threats.

A. Malicious Participants

Malicious participants can launch attacks in two steps of our approach. In the first step, they can start a brute-force attack by proposing all region combinations during the private location matching process. Since the participants first cloak their location before applying the private set intersection mechanism, malicious participants would have only gain access to the cloaked location of honest participants. Moreover, a brute-force attack can be easily detected by analysing the number of submitted region combinations. In the normal case, only participants sharing the same cloaked region(s) learn about them.

In the second step, malicious participants may drop the shares received from other participants and hence not report them to the application server. By doing so, the malicious participants would not gain access to the cloaked location of honest participants, as it is encrypted using the application server's public key. But, it would prevent the application server from decrypting the corresponding triplets and hence disturb the application function as soon as the number of reported shares p is lower than k. As a direct consequence, the application server will also have no access to the location information included in the triplets. An incentive system could be introduced to encourage and reward participants that report shares. This is however considered as future work.

B. Honest-but-Curious Administrators

Based on our threat model (see Sec. III-B), the administrators of both the sensing campaign and the third party are honest-but-curious. Assuming that the private set intersection algorithm used in the location matching process remains unbroken, these administrators will not gain access to the regions submitted by each participant. Depending on the result of the matching, the campaign administrators will have access at most to the participants' cloaked location. This happens when the number of matches n is below k and the participants decide to report their triplets using their cloaked location. Alternatively, the participants can choose to drop them in

(a) Squares (b) Rectangles (c) Triangles (d) Circles

Fig. 2. Considered forms for the partition regions

order to protect their privacy. If $n \geq k$, the corresponding n participants share the same region(s) and hence become indistinguishable for the campaign administrators. By splitting the triplets into shares, the campaign administrators cannot identify who has originally collected the triplets. Assuming that not all k participants collaborate and report the shares, this would reveal neither the origin of the triplets nor the participants' cloaked location, as the application server would be unable to decrypt the triplets.

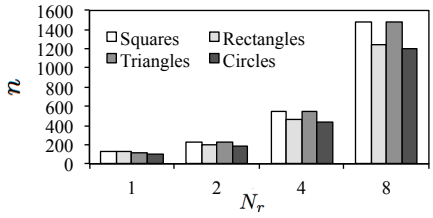
In summary, our solution ensures that the participants' precise location is not known from any other entities contributing to our system model. In the worst case, the participants' cloaked location might be disclosed but the cloaking still protects the exact locations from malicious participants.

VI. Evaluation

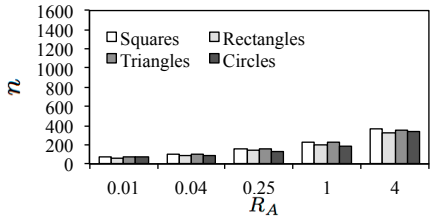
Our evaluation is based on the GPS traces from the GeoLife project ([6], [7], [8]). In this real-world deployment, 182 participants carried GPS-enabled devices to monitor their location. For our evaluations, we have focused on a square of edge length 100 km located in Beijing that contains most GPS traces, and hence removed all entries outside this area. To simulate an urban sensing application, we have further subsampled the GPS traces to obtain a data collection period of 5 min. We have created different types of partitions using different geometric shapes. We have concentrated on GPS traces collected between March, the 1st and the 7th, 2009, as the maximum number of users (i.e., 35) were simultaneously active in that area during this week. While this clearly presents a best case scenario, this number still remains low as compared to the 182 participants having contributed to the dataset as well as the whole population of Beijing. We have repeated each simulation 100 times and present the corresponding results in the following sections.

A. Form of the Partition Regions

We first analyse the impact of different region forms presented in Fig. 2 on the number of matches n, i.e., the number of participants sharing at least one region together. Note that in the case of circles (see Fig. 2(d)), we assign the closest circle to the participants whose original location is not covered by any existing circles. The greater n, the better for the participants' privacy protection as more participants



(a) Impact of N_r on the mean value of n for the selected forms ($R_A = 1km^2$)



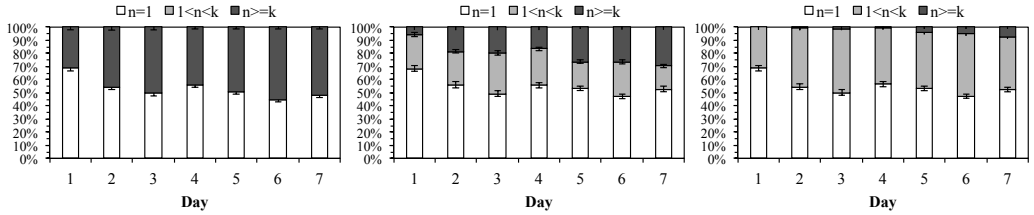
(b) Impact of R_A (in km^2) on the mean value of n for the selected forms ($N_r = 2$)

Fig. 3. Impact of the selected forms for the partition regions on the mean value of n depending on N_r and R_A (k=2)

share the same region(s). For all selected forms, Fig. 3(a) shows the evolution of n depending on the number of regions N_r selected by each participant, while Fig. 3(b) shows its evolution depending on the region area R_A. In both figures, we can observe that the smaller N_r or R_A, the lower the impact of the region form on n. As expected, n increases with both N_r and R_A, as the probability that two participants share common region(s) increases. As compared to the other forms, we see that both squares and triangles lead to the greatest values for n. For the remaining simulations, we hence select squares as baseline for the partition common to all participants.

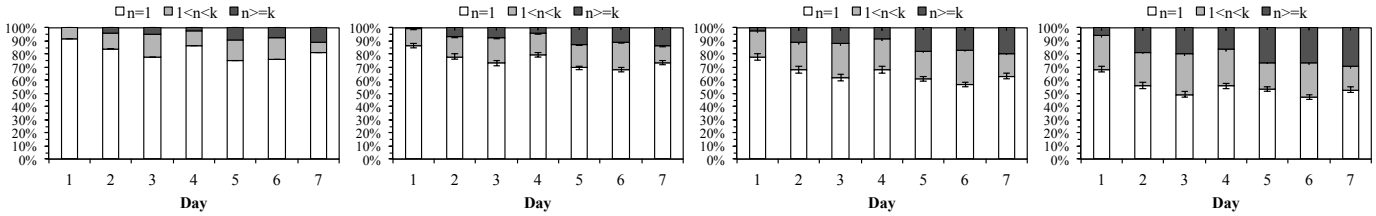
B. Value of k

In the previous simulations, we have chosen $k = 2$. This means that only two clients should share at least one common region to have a successful match and guarantee 2-anonymity to the corresponding participants. In Fig. 4, we next compare the ratio of successful matches (i.e., $n \geq k$) , incomplete matches ($1 < n < k$) and no matches ($n = 1$) depending on chosen values for k. Recall that k is the number of matches required to consider the matching as completed and achieve the targeted k-anonymity. As expected, the number of successful



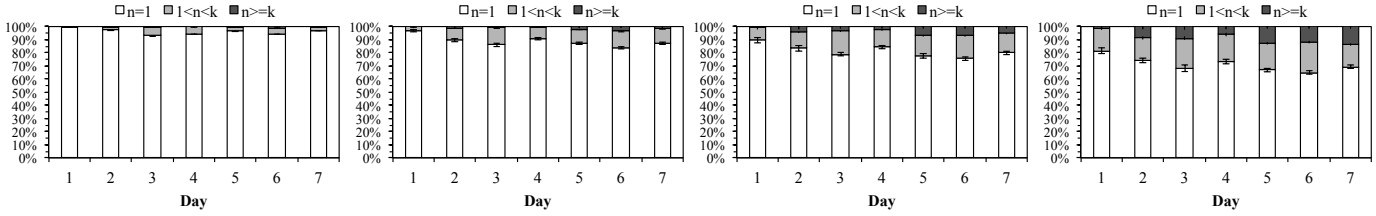
(a) $k = 2$ (b) $k = 3$ (c) $k = 5$

Fig. 4. Impact of the value of k on the median ratio and standard deviation of successful, incomplete, and no matches ($R_A = 25km^2$, $N_r = 8$)



(a) N_r=1 (b) N_r=2 (c) N_r=4 (d) N_r=8

Fig. 5. Influence of number of regions N_r on the median ratio and standard deviation of successful, incomplete, and no matches ($R_A = 25km^2$, $k = 3$)



(a) $R_A = 0.01km^2$ (b) $R_A = 0.25km^2$ (c) $R_A = 1km^2$ (d) $R_A = 4km^2$

Fig. 6. Influence of the region area R_A on the median ratio and standard deviation of successful, incomplete, and no matches ($N_r = 8$, $k = 3$)

matches decreases when k increases, as more participants are required to achieve a successful match. In our case, we have around 50% of successful matches for $k = 2$, 25% for $k = 3$, and 5% for $k = 5$. This means that adding one more participant to reach 3-anonymity almost halves the number of successful matches as compared to $k = 2$. The obtained results hence illustrate the existing tradeoff between k and hence the possible number of successful matches. The absolute values of these ratios are relatively low especially when $R_A = 25km^2$ and $N_r = 8$ in these simulations. This is mainly due the number of participants (35) considered in our evaluation. Assuming a real-world application, the degree of granularity provided for the location may be insufficient. Note that we are mainly interested in comparing the evolution of the number of matches depending on different parameters, instead of gauging their absolute values.

C. Number of Regions N_r

We next study how the number of regions selected by the participants can influence the number of matches. Recall that the participants can choose N_r regions including the region in which they are currently located. Fig. 5 shows that the ratio of successful matches almost increases as N_r increases. For example, doubling N_r leads to doubling the median of the successful matches for all participants. By selecting $N_r > 1$, the area covered by the N_r regions increases and hence, the probability of sharing the same region(s) increases. In particular cases, the combination of multiple regions can correspond to a square with a greater area, i.e., R_A. While the number of matches is the same in these cases, choosing multiple regions benefits to the application accuracy in the remaining cases. Additionally, we have chosen a predefined value for N_r for these simulations. In a real-world scenario, the clients could adapt N_r to the population density of the visited regions to optimise the precision of the cloaked information while still protecting their privacy.

D. Region area R_A

We finally vary the region area R_A and display the corresponding results in Fig. 6. Based on the 35 participants active during the studied week, we can easily observe that only few successful matches are possible (despite $N_r = 8$). Around 5% of the matches are successful, i.e., at lest 3 participants are in the region, when the squares have an edge length of 1 km. Shorter edges lead to even fewer successful matches, while this ratio increases to 10% for edge length of 2 km. Again, applying our solution with such limited number of

participants greatly impacts the accuracy of the results sent to the application server.

VII. Conclusions and Future Work

In this paper, we have presented a novel approach to protect the privacy of participants in urban sensing applications. By applying it, participants can find $k-1$ other participants in a distributed fashion to build groups of k indistinguishable users sharing the same region(s), and hence implement the concept of k-anonymity. Our solution improves existing work as the participants do not need to reveal their precise location to neither a third party nor other participants. Instead, each participant first cloaks her location by selecting regions around her current location based on a partition common to all participants. A private set intersection algorithm next allows the participants to identify shared regions without disclosing them to each others. Moreover, our scheme protects the anonymity of the participants having collected the sensor readings, as these are divided into shares and distributed within the previously built groups of k users before being reported by each participant to the application server. The results show that increasing k from 2 to 3 halves the number of possible shared regions between participants, while doubling the number of regions N_r selected by the participants leads to doubling the median of the successful matches for all participants.

In the future, we plan to adapt the different studied parameters (i.e., k, N_r, and R_A) to the density of active participants in the same area. By doing so, we aim at dynamically increasing the number of successful region matches. Moreover, we will investigate additional real-world datasets in order to evaluate our scheme with more than 35 participants active during the same week as it is the case in the GeoLife dataset, and hence refine our feasibility study.

Acknowledgment

This work was partially supported by the Secure Mobile Networking Lab at Technische Universität Darmstadt and CASED (www.cased.de). Our thanks go to M. Hollick for his support and A. Reinhardt for his feedback.

References

[1] M. Bilandzic, M. Banholzer, D. Peev, V. Georgiev, F. Balagtas-Fernandez, and A. De Luca, "Laermometer: A Mobile Noise Mapping Application," in *Proceedings of the 5th ACM Nordic Conference on Human-Computer Interaction (NordiCHI)*, 2008, pp. 415–418.

[2] P. Mohan, V. N. Padmanabhan, and R. Ramjee, "Nericell: Rich Monitoring of Road and Traffic Conditions using Mobile Smartphones," in *Proceedings of the 6th ACM Conference on Embedded Network Sensor Systems (SenSys)*, 2008, pp. 323–336.

[3] 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, no. 11, pp. 1928–1946, 2011.

[4] K. Shilton, "Four Billion Little Brothers?: Privacy, Mobile Phones, and Ubiquitous Data Collection," *Communications of the ACM*, vol. 52, no. 11, 2009.

[5] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.

[6] Y. Zheng, Q. Li, Y. Chen, X. Xie, and W.-Y. Ma, "Understanding Mobility Based on GPS Data," in *Proceedings of the 10th International Conference on Ubiquitous Computing (UbiComp)*, 2008, pp. 312–321.

[7] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma, "Mining Interesting Locations and Travel Sequences from GPS Trajectories," in *Proceedings of the 18th International Conference on World Wide Web (WWW)*, 2009, pp. 791–800.

[8] Y. Zheng, X. Xie, and W.-Y. Ma, "GeoLife: A Collaborative Social Networking Service among User, Location and Trajectory," *IEEE Data Engineering Bulletin*, vol. 33, no. 2, pp. 32–39, 2010.

[9] M. Gruteser and D. Grunwald, "Anonymous Usage of Location-Based Services Through Spatial and Temporal Cloaking," in *Proceedings of the 1st International Conference on Mobile Systems, Applications and Services (MobiSys)*, 2003, pp. 31–42.

[10] K. L. Huang, S. S. Kanhere, and W. Hu, "Preserving Privacy in Participatory Sensing Systems," *Computer Communications*, vol. 33, no. 11, pp. 1266–1280, 2010.

[11] K. Vu, R. Zheng, and J. Gao, "Efficient Algorithms for K-anonymous Location Privacy in Participatory Sensing," in *Proceedings of the 31th IEEE Conference on Computer Communications (INFOCOM)*, 2012, pp. 2399–2407.

[12] I. Rodhe, C. Rohner, and E. C.-H. Ngai, "On Location Privacy and Quality of Information in Participatory Sensing," in *Proceedings of the 8th ACM Symposium on QoS and Security for Wireless and Mobile Networks (Q2SWinet)*, 2012, pp. 55–62.

[13] R. Shokri, G. Theodorakopoulos, P. Papadimitratos, E. Kazemi, and J.-P. Hubaux, "Hiding in the Mobile Crowd: Location Privacy through Collaboration," *IEEE Transactions on Dependable and Secure Computing*, vol. 11, no. 3, pp. 266–279, 2014.

[14] T. Hashem and L. Kulik, "Safeguarding Location Privacy in Wireless Ad-Hoc Networks," in *UbiComp 2007: Ubiquitous Computing*, ser. Lecture Notes in Computer Science, J. Krumm, G. D. Abowd, A. Seneviratne, and T. Strang, Eds. Springer Berlin Heidelberg, 2007, vol. 4717, pp. 372–390.

[15] W. Li and C. Liu, "A Decentralized Location-Query-Sensitive Cloaking Algorithm for LBS," in *Proceedings of the 8th International Wireless Communications and Mobile Computing Conference (IWCMC)*, 2012, pp. 1040–1045.

[16] L. Kazemi and C. Shahabi, "A Privacy-aware Framework for Participatory Sensing," *SIGKDD Explor. Newsl.*, vol. 13, no. 1, pp. 43–51, 2011.

[17] Y. Huang, D. Evans, and J. Katz, "Private Set Intersection: Are Garbled Circuits Better than Custom Protocols," in *Proceedings of the 19th Network and Distributed Security Symposium (NDSS)*, 2012, pp. 1–15.

[18] A. C.-C. Yao, "How to Generate and Exchange Secrets," in *Proceedings of the 27th IEEE Annual Symposium on Foundations of Computer Science (FOCS)*, 1986, pp. 162–167.

[19] A. Shamir, "How to Share a Secret," *Communications of the ACM*, vol. 22, no. 11, pp. 612–613, 1979.