ISSN:0975-9646

Secure Multi-party Computation Protocols For Collaborative Data Publishing With m-Privacy

K. Prathyusha[1]
M.Tech Student,
CSE, KMMITS , JNTU-A, TIRUPATHI,AP

Sakshi Siva Ramakrishna[2]
Assistant proffesor,
Dept of CSE, KMMITS,JNTU-A,TIRUPATHI, AP

Abstract: ***In this paper collaborative data publishing setting with horizontally partitioned data across multiple data providers, in additional bag round knowledge of each contributing a subset of records . As a special case, a data provider could be the data owner itself who is contributing its own records. This is a very common scenario in social networking and recommendation systems. In this paper we introduce a priory algorithm and genetic algorithms are to publish an anonymized view of the integrated data such that a data recipient including the data providers will not be able to compromise the privacy of the individual records provided by other parties transferring SMC protocol from the forwarding and benefaction the backward of multiple data records to providing m- privacy .***

***Key words:** Data publisher, Recipient, Data records, ananymizying algorithm, SMC protocol, and m-privacy.*

I INTRODUCTION

Data mining is the process of extracting useful, interesting, and previously unknown information from large data sets. The success of data mining relies on the availability of high quality data and effective information sharing. The collection of digital information by governments, corporations, and individuals has created an environment that facilitates large-scale data mining and data analysis. Moreover, driven by mutual benefits, or by regulations that require certain data to be published, there is a demand for sharing data among various parties. For example, licensed hospitals in California are required to submit specific demographic data on every patient discharged from their facility [3].

Nowadays, the terms “information sharing” and “data publishing” not only refer to the traditional one-to-one model, but also the more general models with multiple data holders and data recipients. Recent standardization of information sharing protocols, such as eXtensible Markup Language (XML), Simple Object Access Protocol (SOAP), and Web Services Description Language (WSDL) are catalysts for the recent development of information sharing technology.

Detailed data in its original form often contain sensitive information about individuals, and sharing such data could potentially violate individual privacy.

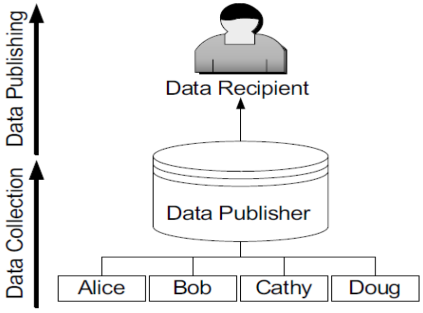


Figure:1.1 Data Collection and Publishing

Data collection and publishing is described in Figure 1.1. In the data collection phase, the data holder collects data from record owners (e.g., Alice and Bob). In the data publishing phase, the data holder releases the collected data to a data miner or the public, called the data recipient, who will then conduct data mining on the published data. data mining has a broad sense, not necessarily restricted to pattern mining or model building. For example, a hospital collects data from patients and publishes the patient records to an external medical center. In this example, the hospital is the data holder, patients are record owners, and the medical center is the data recipient. The data mining conducted at the medical center could be any analysis task from a simple count of the number of men with diabetes to a sophisticated cluster analysis. There are two models of data holders [8]. In the un trusted model, the data holder is not trusted and may attempt to identify sensitive information from record owners. Various cryptographic solutions [15], anonymous communications [4, 9], and statistical methods [13] were proposed to collect records anonymously from their owners without revealing the owners’ identity. In the trusted model, the data holder is trustworthy and record owners are willing to provide their personal information to the data holder; however, the trust is not transitive to the data recipient. **privacy-preserving data publishing (PPDP)**, the data holder has a table of the form *D(Explicit Identifier, Quasi Identifier, Sensitive Attributes, Non-Sensitive Attributes),* where Explicit Identifier is a set of attributes, such as name and social security number (SSN), containing information that explicitly identifies record owners; Quasi Identifier is a set of attributes that could potentially identify record owners; Sensitive Attributes consist of sensitive person-specific information such as disease, salary, and disability status; and Non-Sensitive Attributes contains all attributes that do not fall into the previous three categories [3]. Most works assume that the four sets of attributes are disjoint. Most works assume that each record in the table represents a distinct record owner.

Anonymization [6, 7] refers to the PPDP approach that seeks to hide the identity and/or the sensitive data of record owners, assuming that sensitive data must be retained for data analysis. Clearly, explicit identifiers of record owners must be removed.

II EXISTING SYSTEM

A single data provider setting and considered the data recipient as an attacker. A large body of literature assumes limited background knowledge of the attacker, and defines privacy using relaxed *adversarial* notion by considering specific types of attacks. Representative principles include *k*-anonymity, *l*diversity, and *t*-closeness. A few recent works have modeled the instance level background knowledge as corruption, and studied perturbation techniques under these syntactic privacy notions

Disadvantages Of Existing System

1. Collaborative data publishing can be considered as a multi-party computation problem, in which multiple providers wish to compute an anonymized view of their data without disclosing any private and sensitive information
2. The problem of inferring information from anonymized data has been widely studied in a single data provider setting. A data recipient that is an attacker, e.g., *P*0, attempts to infer additional information about data records using the published data, *T*∗, and background knowledge, *BK.*

III PROPOSED SYSTEM

We consider the collaborative data publishing setting with horizontally partitioned data across multiple data providers, each contributing a subset of records Ti. As a special case, a data provider could be the data owner itself who is contributing its own records. This is a very common scenario in social networking and recommendation systems. Our goal is to publish an anonymized view of the integrated data such that a data recipient including the data providers will not be able to compromise the privacy of the individual records provided by other parties.

ADVANTAGES OF PROPOSED SYSTEM

Compared to our preliminary version, our new contributions extend above results. First, we adapt privacy verification and anonymization mechanisms to work for *m*-privacy with respect to any privacy constraint, including nonmonotonic ones. We list all necessary privacy checks and prove that no fewer checks are enough to confirm *m*-privacy. Second, we propose SMC protocols for secure *m*-privacy verification and anonymization. For all protocols we prove their security, complexity and experimentally confirm their efficiency.

IV IMPLEMENTATION

1. Dataset Collection
2. Attacks by External Data Recipient Using Anonymized Data
3. Attacks by Data Providers Using Anonymized Data and Their Own Data
4. Doctor Login
5. Secure *m*-Privacy Verification

Dataset Collection :

In this if patients have to take treatment, he/she should register their details like Name, Age, and Disease they get affected, Email etc. These details are maintained in a Database by the Hospital management. Only Doctors can see all their details. Patient can only see his own record. When the data are distributed among multiple data providers or data owners, two main settings are used for anonymization. One approach is for each provider to anonymize the data independently (anonymize-and-aggregate), which results in potential loss of integrated data utility. A more desirable approach is collaborative data publishing which anonymize data from all Providers as if they would come from one source (aggregate-and-anonymize), using either a trusted third-party(TTP) or Secure Multi-party Computation (SMC) protocols to do computations .

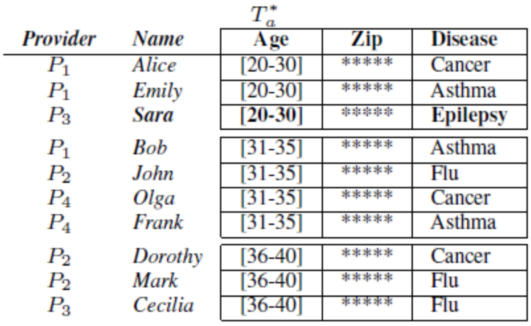
Attacks by External Data Recipient Using Anonymized Data:

A data recipient, e.g. P0, could be an attacker and attempts to infer additional information about the records using the published data (T∗) and some background knowledge (BK) such as publicly available external data.

Attacks by Data Providers Using Anonymized Data and Their Own Data:

Each data provider, such as P1 in Table 1, can also use anonymized data T∗ and his own data (T1) to infer additional information (Age,Zip,Disease) about other records. Compared to the attack by the external recipient20-30 years in the first attack scenario, each provider has additional data knowledge of their own records, which can help with the attack. This issue can be further worsened when multiple data providers collude with each other.

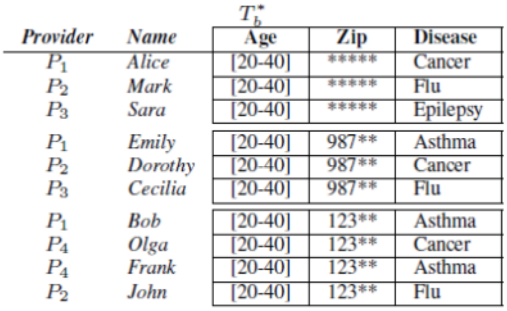
Table1



T_a^*

Provider	*Name*	Age	Zip	Disease
P_1	*Alice*	[20-30]	*****	Cancer
P_1	*Emily*	[20-30]	*****	Asthma
P_3	***Sara***	**[20-30]**	*****	**Epilepsy**
P_1	*Bob*	[31-35]	*****	Asthma
P_2	*John*	[31-35]	*****	Flu
P_4	*Olga*	[31-35]	*****	Cancer
P_4	*Frank*	[31-35]	*****	Asthma
P_2	*Dorothy*	[36-40]	*****	Cancer
P_2	*Mark*	[36-40]	*****	Flu
P_3	*Cecilia*	[36-40]	*****	Flu

Table: 2



T_b^*

Provider	*Name*	Age	Zip	Disease
P_1	*Alice*	[20-40]	*****	Cancer
P_2	*Mark*	[20-40]	*****	Flu
P_3	*Sara*	[20-40]	*****	Epilepsy
P_1	*Emily*	[20-40]	987**	Asthma
P_2	*Dorothy*	[20-40]	987**	Cancer
P_3	*Cecilia*	[20-40]	987**	Flu
P_1	*Bob*	[20-40]	123**	Asthma
P_4	*Olga*	[20-40]	123**	Cancer
P_4	*Frank*	[20-40]	123**	Asthma
P_2	*John*	[20-40]	123**	Flu

Doctor can see all the patients details and will get the background knowledge(BK),by the chance he will see horizontally partitioned data20-40 of distributed data base of the group of hospitals and can see how many patients are affected without knowing of individual records20-30 and 20-40 of the patients and sensitive information about the individuals.

Benefaction:

We define address and Quasi ID new type of “insider Attack” by data providers in this papers. In general Define an m-adversary as a coalition of m colluding data providers or data owners, and attempts to infer data records benefaction by other providers. Note that 0, l l –Adversary models the multiple recipients, who has only access to multiple bag round knowledge(BF). an anonymization satisfies *m*-privacy with respect to *l*-diversity if the records in each equivalence group excluding ones from any *m*-adversary still satisfy *l*-diversity. In our example in Table I, $T* b$ is an anonymization that satisfies *m*-privacy ($m = 1$) with respect to *k*-anonymity and *l*- diversity ($k = 3$, $l = 2$).

Second, to address the challenges of checking a combinatorial number of potential *m*-adversaries, we present heuristic algorithms for efficiently verifying *m*-privacy given a set of records , complexity and Experimental conformation of SMC protocol.

Suppose a data holder has released multiple views of the same underlying raw data data. Even if the data holder releases one view to each data recipient based on their information needs, it is difficult to prevent them from colluding with each other behind the scene. Thus, some recipient may have access to multiple or even all views. In particular, an adversary can combine attributes from the two views to form a sharper QID that contains attributes from both views.

Checking Violations of k-Anonymity on Multiple Views:

We first illustrate violations of k-anonymity in the data publishing scenario

where data in a raw data table T are being released in the form of a view set. A view set is a pair (V, v), where V is a list of selection-projection queries (q1, . . . , qn) on T , and v is a list of relations (r1, . . . , rn) without duplicate records [15]. Then, we also consider the privacy threats caused by functional dependency as prior knowledge, followed by a discussion on the violations detection methods.

Name	Job	Age	Disease
Alice	Cook	40	Flu
Bob	Engineer	50	Diabetes
Alvin	Lower	60	Malaria

Table3

Verification of m- privacy

The data holder previously collected a set of records T1 time stamped t1, and published a k-anonym zed version of T1, denoted by release R1. Then the data holder collects a new set of records T2 time stamped t2 and wants to publish a k-anonym zed version of all records collected so far, T1^{U} T2, denoted by release R2. Note, Ti contains the “events” that happened at time T$_{i}$. An event, once occurred, becomes part of the history, therefore, cannot be deleted. This publishing scenario is different from update scenario in standard data management where deletion of records can occur. Ri simply publishes the “history,” i.e., the events that happened up to time ti. A real-life Anonymizing Incrementally Updated Data Records 1000 example can be found in below show figure 2. where the hospitals are required to submit specific demographic data of all discharged patients every six months.

$$InfoGain(v) = E(T'[\perp_j]) - \frac{|T'[v]|}{|T'[\perp_j]|}E(T'[v]) - \frac{|T^{*'}[\perp_j]|}{|T'[\perp_j]|}E(T^{*'}[\perp_j]).$$

Algorithm:Anonymization algorithm

```
--------------------------------------------------------
Input:    T1, T2 a    m-privacy requorement, a taxonomy tree for
each categorical attribute in $x_n$.
Output:a generalized T2 satiisfying the privacy require ment.
    1. Generalize entry value of Ai to ANYwhere $A_i$€$X_i$
    2. While there is a valid candidate in $^U$cut, do
    3. Find the paire of highest diseases ($x_i$ )from Úcut.
    4.  Specialized or on t2  and remove $X_i$from Úcut.
    5. Replace new (xi) and the valid status of xi  for all in
       Úcut.
    6. Out put the generalized T2 and Úcut.
--------------------------------------------------------
```

*Continuous data publishing***.** Publishing the release R2 for T1^{U}T2 would permit an analysis on the data over the combined time period of t1 and t2. It also takes the advantage of data abundance over a longer period of time to reduce data distortion required by anonymization.

Multi-purpose publishing. With T2 being empty, R1 and R2 can be two releases of T1 anonym zed differently to serve different information needs, such as correlation analysis vs. clustering analysis, or different recipients, such as a medical research team vs. a health insurance company. These recipients may collude together by sharing their received data. We first describe the publishing model with two releases and then show the extension beyond two releases and beyond k-anonymity [10, 11], we assume that each individual has at most one record in T1 ^{U}T2. This assumption holds in many real-life databases. For example, in a normalized customer data table, each customer has only one profile. In the case that an individual has a record in both T1 and T2, there will be two duplicates in T1 ^{U}T2 and one of them can be removed in a preprocessing.

Example:

The data holder (e.g., a hospital) published the 5-anonymized R1 for 5 records a1-a5 collected in the previous month (i.e., timestamp t1). The anonymization was done by generalizing UK and France into Europe; the original values in the brackets are not released. In the current month (i.e., timestamp t2), the data holder collects 5 new records (i.e., b6-b10) and publishes the 5-anonymized R2 for all 10 records collected so far. Records are shuffled to prevent mapping between R1 and R2 by their order. The recipients know that every record in R1 has a “corresponding record” in R2 because R2 is a release for T1UT2. Suppose that one recipient, the adversary, tries to identify his neighbor Alice’s record from R1 or R2, knowing that Alice was admitted to the hospital, as well as Alice’s QID and time stamp.

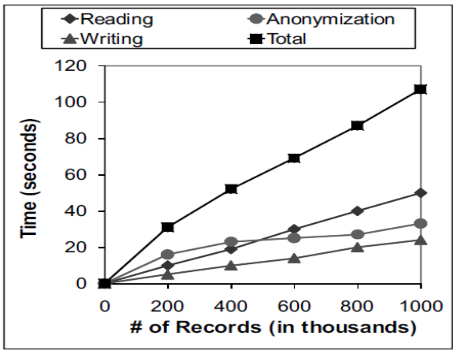


Figure2

Forward-attack, denoted by F-attack(R1,R2). P has timestamp t1 and the adversary tries to identify P's record in the cracking release R1 using the background release R2. Since P has a record in R1 and a record in R2, if a matching record r1 in R1 represents P, there must be a corresponding record in R2 that matches P's QID and agrees with r1 on the sensitive attribute. If r1 fails to have such a corresponding record in R2, then r1 does not originate from P's QID, and therefore, r1 can be excluded from the possibility of P's record.

Cross-attack, Denoted by C-attack(R1,R2). P has timestamp t1 and the adversary tries to identify P's record in the cracking release R2 using the background release R1. Similar to F-attack, if a matching record r2 in R2 represents P, there must be a corresponding record in R1 that matches P's

QID and agrees with r2 on the sensitive attribute. If r2 fails to have such a corresponding record in R1, then r2 either has timestamp t2 or does not originate from P's QID, and therefore, r2 can be excluded from the possibility of P's record.

Backward-attack, denoted by B-attack (R1,R2). P has timestamp t2 and the adversary tries to identify P's record in the cracking release R2 using the background release R1. In this case, P has a record in R2, but not in R1. Therefore, if a matching record r2 in R2 has to be the corresponding record of some record in R1, then r2 has timestamp t1, and therefore, r2 can be excluded from the possibility of P's record.
Note that it is impossible to single out the matching records in R2 that have time stamp t2 but do not originate from P's QID since all records at t2 have no corresponding record in R1.

Genetic Algorithm:

The pioneer to address the anonymization problem for classification analysis and proposed a genetic algorithmic solution to achieve the traditional k-anonymity with the goal of preserving the data utility.

Secure m-Privacy Verification

In this module Admin acts as Trusted Third Party (TTP).He can see all individual records and their sensitive information among the overall hospital distributed data base. Anonymation can be done by this people. He/She collected information's from various hospitals and grouped into each other and make them as an anonymized data.

Algorithm :Secure fitness protocol

```
Input: T-thresholds from all constraints, data records T.
Results: Share of the minimal fitness value.
    1.  lcm=1 leaset _common_multiple(T)
    2.  For each I belongs to {0,………,w) do
    3.  Securely compute $¥_I$ measured value for $C_{I,}$ and  T
    4.  [$F_i$ =multiplicate ([$¥_i$],lcm/$T_i$)
    5.  Return reconstruct(min([F1]……[Fw]))/lcm
```

V.Experiment Work:

The experiments confirm that the specification of the multi-QID anonymity requirement helps avoid unnecessary masking and, therefore, preserves more of the cluster structure. However, if the data recipient and the data holder employ different clustering algorithms, then there is no guarantee that the encoded raw cluster structure can be extracted. Thus, in practice, it is important for the data holder to validate the cluster quality, using the evaluation methods proposed, before releasing the data. Finally, experiments suggest that the proposed anonymization approach is highly efficient and scalable for multi QID.

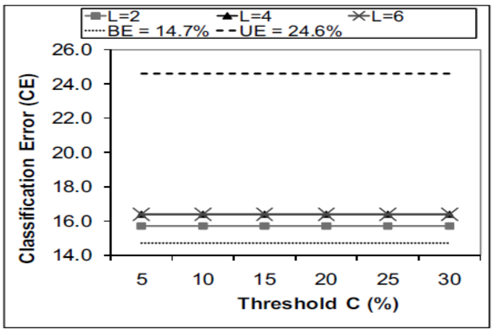


Figure 3

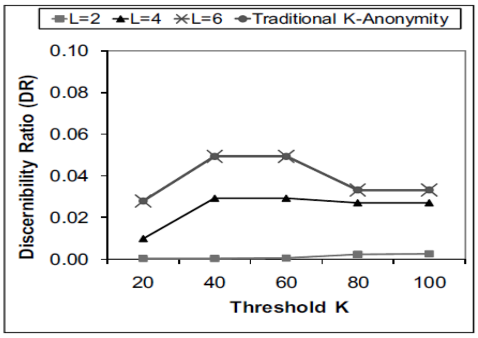


Figure 4

VI Related Work

Most of the work multiple data public has an increased sense of privacy loss. Since data mining is often a key component of information systems, homeland security systems [12], and monitoring and surveillance systems [7], it

gives a wrong impression that data mining is a technique for privacy intrusion.

This lack of trust has become an obstacle to the benefit of the technology. For example, the potentially beneficial data mining research project, Terrorism Information Awareness (TIA), was terminated by the government due to its controversial procedures of collecting, sharing, and analyzing the trails left by individuals [12]. Motivated by the privacy concerns on data mining tools, a research area called privacy-preserving data mining (PPDM) emerged in 2000 [2, 6]. The Initial idea of PPDM was to extend traditional data mining techniques to work with the data modified to mask sensitive information.

The key issues were how to modify the data and how to recover the data mining result from the modified data. The solutions were often tightly coupled with the data mining algorithms under consideration. In contrast, privacy-preserving data publishing (PPDP) may not necessarily tie to a specific data mining task, and the data mining task is sometimes unknown at the time of data publishing. Furthermore, some PPDP solutions emphasize preserving the data truthfulness at the record level as discussed earlier, but PPDM solutions often do not preserve such property.

VII CONCLUSION

In this paper we considered a new type of potential attackers in collaborative data publishing – a coalition of data providers, called *m*-adversary. Privacy threats introduced by *m*-adversaries are modeled by a new privacy notion, *m*-privacy, and use adaptive ordering techniques for higher efficiency. We also presented a *provider-aware* anonymization algorithm with an adaptive verification strategy to ensure high utility and *m*-privacy of anonymized data. Experimental results confirmed that our heuristics perform better or comparable with existing algorithms in terms of efficiency and utility. All algorithms have been implemented in distributed settings with a TTP and as SMC protocols. All protocols have been presented in details and their security and complexity has been carefully analyzed. Implementations of algorithms for the TTP setting is available on-line for further development and deployments3. There are many potential research directions. For example, it remains a question to model and address the data knowledge of data providers when data are distributed in a vertical or ad-hoc fashion. It would be also interesting to investigate if our methods can be generalized to other kinds of data such as set-valued data.

FEATURE ENHANCEMENT

The solution presented above focuses on preventing the privacy threats caused by record linkages, but the framework is extendable to thwart attributes linkages by adopting different anonymization algorithms and achieving other privacy models, such as ℓ-diversity and the extension requires modification of the Score or cost functions in these algorithms to bias on refinements or masking's that can distinguish class labels. The framework can also adopt other evaluation methods, such as entropy , or any ad-hoc methods defined by the data holder

REFERENCES

[1] C. C. Aggarwal and P. S. Yu. A framework for condensation-based anonymization of string data. Data Mining and Knowledge Discovery (DMKD), 13(3):251–275, February 2008.

[2] R. Agrawal and R. Srikant. Privacy reserving data mining. In Proc. of ACM International Conference on Management of Data (SIGMOD), pages 439–450, Dallas, Texas, May 2000.

[3] D. M. Carlisle, M. L. Rodrian, and C. L. Diamond. California inpatient data reporting manual, medical information reporting for california, 5th edition. Technical report, Office of Statewide Health Planning and Development, July 2007.

[4] D. Chaum. Untraceable electronic mail, return addresses, and digital pseudonyms. Communications of the ACM, 24(2):84–88, 1981.

[5] S. Chawla, C. Dwork, F. McSherry, A. Smith, and H. Wee. Toward privacy in public databases. In Proc. of Theory of Cryptography Conference (TCC), pages 363–385, Cambridge, MA, February 2005.

[6] L. H. Cox. Suppression methodology and statistical disclosure control. Journal of the American Statistical Association, 75(370):377–385, June 1980.

[7] T. Dalenius. Finding a needle in a haystack - or identifying anonymous census record. Journal of Official Statistics, 2(3):329–336, 1986.

[8] J. Gehrke. Models and methods for privacy-preserving data publishing and analysis. In Tutorial at the 12th ACM Internationalconference on Knowledge Discovery and Data Mining (SIGKDD), Philadelphia, PA,August 2006.

[9] M. Jakobsson, A. Juels, and R. L. Rivest. Making mix nets robust for electronic voting by randomized partial checking. In Proc. of the 11th USENIX Security Symposium, pages 339–353, 2002.

[10] P. Samarati. Protecting respondents' identities in microdata release. IEEE Transactions on Knowledge and Data Engineering (TKDE),13(6):1010–1027, 2001.

[11] L. Sweeney. k-Anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-based Systems,10(5):557–570, 2002.

[12] J. W. Seifert. Data mining and homeland security: An overview. CRS Report for Congress, (RL31798), January 2006.http://www.fas.org/sgp/crs/intel/RL31798.pdf.

[13] S. L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. Journal of the American Statistical Association, 60(309):63–69, 1965.

[14] X. Xiao, Y. Tao, and M. Chen. Optimal random perturbation at multiple privacy levels. In Proc. of the 35th Very Large Data Bases (VLDB),pages 814–825, 2009.

[15] Z. Yang, S. Zhong, and R. N.Wright. Anonymity-preserving data collection. In Proc. of the 11th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), pages 334–343, 2005.

[16] T.-H. You, W.-C. Peng, and W.-C. Lee. Protect moving trajectories with dummies. In Proc. of the International Workshop on Privacy-Aware Location-based Mobile Services (PALMS), pages 278–282, May 2007.