9 From Data Privacy to Location Privacy

Ting Wang and Ling Liu[1]

Abstract: Over the past decade, the research on data privacy has achieved considerable advancement in the following two aspects: First, a variety of privacy threat models and privacy principles have been proposed, aiming at providing sufficient protection against different types of inference attacks; Second, a plethora of algorithms and methods have been developed to implement the proposed privacy principles, while attempting to optimize the utility of the resulting data. The first part of the chapter presents an overview of data privacy research by taking a close examination at the achievements from the above two aspects, with the objective of pinpointing individual research efforts on the grand map of data privacy protection. As a special form of data privacy, location privacy possesses its unique characteristics. In the second part of the chapter, we examine the research challenges and opportunities of location privacy protection, in a perspective analogous to data privacy. Our discussion attempts to answer the following three questions: (1) Is it sufficient to apply the data privacy models and algorithms developed to date for protecting location privacy? (2) What is the current state of the research on location privacy? (3) What are the open issues and technical challenges that demand further investigation? Through answering these questions, we intend to provide a comprehensive review of the state of the art in location privacy research.

9.1 Introduction

Recent years have witnessed increasing concerns about the privacy of personal information in various data management and dissemination applications. Typically, such data is stored in a relational data model, and each record consists of three categories of information: identity attributes, quasi-identity attributes, and sensitive attributes. Individuals intend to protect their sensitive information from exposure to unauthorized parties via direct disclosure or indirect inferences [20]. Concretely, releasing of microdata with identity attributes removed may still disclose sensitive information about individuals with high probability, due to the inference attacks that link quasi-identity attributes

[1] Distributed Data Intensive System Lab, College of Computing, Georgia Tech

J.J.P. Tsai and P.S. Yu (eds.), *Machine Learning in Cyber Trust: Security, Privacy, and Reliability*, DOI: 10.1007/978-0-387-88735-7_9,

to some external knowledge. One of the most well known examples of such attacks was given in the original *k*-anonymity paper [49].

In order to alleviate such concerns, various privacy protection mechanisms have been proposed to perform transformation over the raw data before publishing. The literatures of data privacy protection techniques can be grouped into two categories: the first category of work aims at proposing privacy models and principles, which serve as criteria for measuring if the publication of a raw dataset provides sufficient privacy protection. The second category of research explores data transformation techniques that can meet the proposed privacy principles, while maximizing the utility of the resulting data.

The first part of the chapter is devoted to a close examination on the advancement of the data privacy research in the past decade, from the aforementioned two aspects. Specifically, we review the proposed data privacy principles and models, according to the types of privacy breaches identified, and analyze the implicit relationships among different types of privacy breaches. We classify the existing data anonymization algorithms and methods, based on their underlying privacy models and implementation techniques, and compare their strengths and weaknesses accordingly. In addition, we discuss various data utility optimization frameworks. Through the survey, we aim at pinpointing the existing independent research efforts on the grand map of data privacy protection.

With the ubiquitous wireless connectivity and the continued advance in mobile positioning technology (e.g., cellular phones, GPS-like devices), follows an explosive growth of location-based services (LBS). Examples include location-based store finders ("*Where is the nearest gas station to my current location?*"}), traffic condition tracking ("*What is the traffic condition on Highway 85 North?*"), and spatial alarm ("*Remind me to drop off a letter when I am near a post office.*"}). The mobile users obtain such services by issuing requests together with their location information to the service providers. While offering great convenience and business opportunities, LBS also opens the door for misuse of users' private location information. For example, the collected location information can be exploited to spam users with unwanted advertisements; Personal medical conditions, alternative lifestyles, unpopular political or religious views can be inferred by knowing users' visit to specific locations; GPS devices have even been used in physical stalking [22, 54].

Such concerns have spurred intensive research on location privacy protection recently [40]. As a special form of data privacy, location privacy features its unique characteristics and research challenges. Specifically, location privacy requirements are inherently personalized and context-sensitive. The level of privacy protection is intricately related to the quality of service delivered to the customers. Furthermore, location data is highly dynamic, and subjected to frequent updates. Such uniqueness makes it inadequate to directly apply the data privacy research results and techniques to the problem of location privacy.

In the second part of the chapter, we present a brief survey of the existing literatures of location privacy research, and examine the research challenges and opportunities, in a perspective analogous to data privacy. Within the survey, we intend to address the following three key questions: (1) Is it sufficient to apply the data privacy models and algorithms developed to date for protecting location privacy? (2) What is the current state of the research on location privacy? (3) What are the open issues and technical challenges that are worth further investigation? Specifically, we summarize the proposed location privacy principles and models, and their implicit relationships; We analyze various system architectures for implementing location anonymization; We classify the existing location anonymization tools based on their underlying architecture models and implementation techniques, and compare their strengths and weaknesses; We then discuss open issues and technical challenges for constructing complete solutions for location privacy protection. Through this survey, we intend to provide a comprehensive review of the state of the art in location privacy research.

9.2 Data Privacy

While the scope of data privacy broadly covers the topics of privacy-preserving data publication, data mining, information retrieval, etc, and involves techniques such as cryptography, perturbation, auditing, etc, this survey particularly focuses on the problem of privacy protection in data publication, using the techniques of data perturbation, as is most relevant to the location privacy protection.

The problem of controlling information disclosure in data dissemination for public-use has been studied extensively in the framework of statistical databases. Motivated by the need of publishing census data, the statistics literatures focus mainly on identifying and protecting the privacy of sensitive data entries in contingency tables, or tables of counts corresponding to cross-classification of the data.

A number of disclosure limitation techniques have been proposed [19], which can be broadly classified into *query restriction* and *data perturbation*. The query restriction family includes controlling the size of query results [21], restricting the overlap between the answers of successive queries [16], suppressing the cells of small size [13], and auditing queries to check privacy compromises [11]; The data perturbation family includes sampling data [15], swapping data entries between different cells [14], and adding noises to the data [53] or the query results [15]. Nevertheless, as shown in [1], these proposed techniques can not meet the requirement of providing high precision statistics, and meanwhile preventing exact or partial disclosure of personal in-

formation. Moreover, the perturbation-based techniques generally compromise the data integrity of the table.

Motivated by the need of publishing non-aggregated *microdata* involving personal sensitive information, e.g., medical data [49, 51], the data privacy researcher have mainly focused on ensuring that no adversary can accurately infer the sensitive information of an individual in the data, based on the published data and her background knowledge.

Typically, the microdata is stored in a relational data model, and each record in the microdata corresponds to an individual, which can be divided into three sub-categories: (1) *identifier* attribute, e.g., social security number, which can explicitly identify an individual, and therefore is usually removed from the microdata for publication; (2) *quasi-identifier* (QI) attributes, e.g., zip-code, gender and birth-date, whose values in combination can potentially identify an individual, and are usually available from other sources (e.g., voter registration list); (3) *sensitive* (SA) attribute, e.g., disease, which is the private information to be protected for the individuals.

Within this setting, a majority of the research efforts focus on addressing the *linking attacks*: the adversary possesses the exact QI-attribute values of the victim individual, and attempts to infer his/her sensitive value from the published data. Aiming at providing protection against linking attacks while preserving the information truthfulness, a set of *group-based anonymization* techniques have been proposed, which guarantees that each individual is hidden within certain group, called *QI-group* with respect to their QI-attributes values.

Two main methodologies have been proposed for achieving such group-based anonymization: *suppression* and *generalization* [51]. Specifically, for a given QI-attribute value, the suppression operation involves not releasing the value at all, while the generalization operation replaces it with a less specific, more general value that is faithful to the original. Clearly, suppression can be considered as a special form of generalization, where the QI-attribute value is replaced by a non-informative wildcard symbol '*'.

9.2.1 Models and Principles

Essentially, in linking attacks, the adversary infers the sensitive value of a victim individual, by leveraging the association between the QI attribute values of the victim, and the corresponding sensitive value (QI-SA association), as appearing in the microdata. Group-based anonymization weakens such associations by reducing the granularity of the representation of the QI-attributes values, and the protection is adequate if the weakened associations are not informative enough for the adversary to infer individuals' sensitive values with high confidence. Aiming at providing sufficient protection, various anonymi-

zation principles and models have been proposed, which can be classified, according to the type of QI-SA associations that they are designed to address, and the background knowledge that the adversary is assumed to have.

9.2.1.1 QI-SA Association

The associations between the QI attribute values and sensitive values can be categorized as *exact association* and *proximate association*. The former refers to the link between QI-attributes values and *specific* sensitive values, while the latter refers to the link between QI-attributes values and *a set of proximate* sensitive values.

Exact QI-SA Association. The exact QI-SA association is particularly meaningful for publishing categorical sensitive attribute, where different values have no sense of proximity, and it is desired to prevent the adversary from linking an individual to a specific sensitive value in the published data with high confidence. A number of principles have been proposed, with the objective of avoiding exact QI-SA association re-construction, including k-anonymity [49], l-diversity [41] and its variant (α, k)-anonymity [59].

- k-anonymity [49]. As a pioneering work on privacy-preserving data publication, Sweeney introduced the k-anonymity model to preserve information truthfulness in microdata, which paved the way for the development of group-based anonymization methodology. Intuitively, k-anonymity requires that each QI-group contains at least k tuples, therefore each individual is indistinguishable from a group of more than (k-1) others, with respect to their QI-attribute values. Clearly, k-anonymity is effective in preventing identification of an individual record.
- l-diversity [41]. It is recognized that k-anonymity ignores the characteristics of the sensitive attribute, which therefore does not prevent the disclosure of sensitive attribute values. For example, after identifying an individual belongs to a QI-group G, and 3 out of 4 tuples in G share an identical sensitive value, without further information, the adversary can infer that this sensitive value belongs to the victim individual with probability 75%. To address such "homogeneity attacks", Machanavajjhala et al. introduced the principle of l-diversity, which demands that each QI-group contains at least l "well represented"' sensitive values, such that no single sensitive value is dominant in the group. A number of instantiations of l-diversity were proposed in [41], e.g., entropy l-diversity, recursive (c, l)-diversity.
- (α, k)-anonymity [59]. As a variant of l-diversity, (α, k)-anonymity essentially combines k-anonymity and l-diversity, which requires that (1) each

QI-group has size at least k, and (2) at most α percent of the tuples in each QI-group share identical sensitive values.

Proximate QI-SA Association. The principles above aim at preventing the reconstruction of exact sensitive values from the anonymized data, which is fairly reasonable for publishing categorical sensitive data. However, when publishing quantitative sensitive data is taken into consideration, the proximate QI-SA association arises as another important privacy concern, which refers to the link between QI-attributes values and a set of proximate sensitive values. Specifically, by studying the published data, the adversary may conclude with high confidence that the sensitive value of a victim individual falls in a short interval, even though with low confidence about the exact value.

To remedy this problem, several anonymization principles have been proposed recently, including (k, e)-anonymity [65], t-closeness [39], variance control [36] and (ε, m)-anonymity [38].

- (k, e)-anonymity [65]. Intuitively, (k, e)-anonymity dictates that every QI-group should be of size at least k, and in each group, the difference between the maximum and minimum sensitive values should be at least e. Essentially, (k, e)-anonymity counters the adversary from narrowing down the sensitive value of an individual to a small interval, by specifying constraints on the extreme values of each QI-group. However, it does not prevent the case that in a QI-group, a majority of tuples have nearly identical sensitive values, and the remaining few carry faraway values [38]. In such cases, the adversary can still link the victim with the set of proximate sensitive values with high confidence.
- t-closeness [39]. Li and Li separate the information gain an adversary can get from the released data into two parts: that about all the population in the data, and that about specific individuals. To limit the second kind of information gain, they proposed t-closeness, which is a further enhancement of the concept of l-diversity. Intuitively, it demands that the distribution of the sensitive values in every QI-group should not deviate from the distribution in the overall table more than a threshold t. In the paper, the EMD (earth mover distance) metric is used to measure the distance of probability distributions, which however fails to capture the probability scaling characteristics of the distribution itself.
- Variance control [37]. A natural measurement of the "concentration" of a set of numeric values is their variance. The variance control model imposes a threshold t on the variance of sensitive values in each QI-group. Nevertheless, it is proved in [38] that large variance does not ensure sufficient protection against attacks based on proximate QI-SA associations.
- (ε, m)-anonymity [38] In a recent work, Li et al. proposed the principle (ε, m)-anonymity, designed to address the proximate QI-SA-association based attacks, in publishing numeric sensitive data. Intuitively, (ε, m)-

anonymity requires that for a given QI-group, each sensitive value is similar to at most $1/m$ of all the sensitive values of the group, and the similarity is controlled by ε. One shortcoming of this principle is that since m is an integer, it only allows adjusting the level of privacy protection in a harmonic sequence manner, i.e., 1/2, 1/3, etc.

9.2.1.2 Background Knowledge

The aforementioned principles and models all assume that the adversary possesses full identification information [42], which includes (i) the identifier of the individuals in the microdata table, and (ii) their exact QI-attribute values. For scenarios with alternative background knowledge assumptions, a number of anonymization principles have been proposed, including δ-presence [46], (c, k)-safety [42], privacy skyline [8], m-invariance [62], and sequential anonymization [55].

Less External Knowledge. An implicit assumption made in the anonymization principles discussed so far is that the adversary already knows that the victim individual is definitely in the microdata. In the scenario where she has no prior knowledge regarding the presence of the individuals in the microdata, the privacy protection can be achieved from another perspective.

- δ-presence [46]. Assuming that the adversary has no prior knowledge regarding whether the victim appears in the microdata, δ-presence achieves privacy protection by preventing the adversary from inferring the presence of the individual in the data with probability no more than δ. Clearly, in this setting, if the adversary is only δ (percent) sure that the victim is in the microdata, any specific sensitive value belongs to the victim with probability no more than δ.

More External Knowledge. The problem of privacy-preserving data publishing is further complicated by the fact that in addition to the published data, the adversary may have access to other external knowledge, e.g., public records and social network relating individuals. To enforce the privacy protection against such external knowledge-armed adversary, several stricter privacy criteria have been proposed.

- (c, k)-safety [42]. Martin et al. considered the case that the adversary possesses the *implicational knowledge*, as modeled as a set of rules that if the individual o_1 has sensitive value s_1 then the individual o_2\$ has sensitive value s_2. (c, k)-safety guarantees that even if the adversary possesses k pieces of such knowledge, the probability that she can infer the exact sensitive value of an individual is no more than c.

- Privacy skyline [8]. Besides the implicational knowledge, Chen et al. considered a set of other types of background knowledge, including that about the target individual, that about others (implicational knowledge), and that about the family of individuals sharing identical sensitive value. They proposed a multidimensional approach to quantifying an adversary's background knowledge. This model enables the publishing organization to investigate various types of privacy threats, and tune the amount of privacy protection against each type of adversarial knowledge.

Multiple Releases. The centralized-publication models discussed so far focus on the scenario of "one-time" release, while there are scenarios where the microdata is subjected to insertion and deletion operations, and need to be released multiple times. It is clear that in such cases, the adversary may potentially leverage previous releases of microdata to infer sensitive information in the current release.

- *m*-invariance [62]. To remedy this problem, Xiao and Tao proposed the principle *m*-invariance for sequential releases of microdata, which prevents the adversary from using multiple releases to infer the sensitive information of individuals. Intuitively, *m*-invariance can be considered as a stringent version of *l*-diversity, which dictates that at each release, each QI-group contains at least *m* tuples, and all of them have different sensitive values. An efficient algorithm was proposed in [62] for fulfilling *m*-invariance, via inserting counterfeit tuples.
- Sequential anonymization [55]. Wang and Fung considered a different scenario of re-publication from *m*-invariance. They assume a static microdata table that contains a large number of QI-attributes. In the first release, a subset of QI-attributes, together with the sensitive attribute are published. Later, the publisher is requested to publish an additional subset of QI-attributes. They intend to prevent the adversary from inferring the sensitive information by joining multiple publications. In [55], a technique based on lossy joins is proposed to counter such table-joining attacks. The intuition behind the approach is: if the join is lossy enough so that the adversary has low confidence in relating previous releases with the current one, then she is effectively prevented from discovering the identities of the records.

9.2.2 Techniques

While the first category of literatures on data privacy aims at proposing anonymization principles that guarantee adequate protection against the privacy attacks in question, the second category explores the possibility of fulfill-

ing the proposed privacy principles, while preserving the utility of the resulting data to the maximum extent. They either analyze the hardness of implementing the anonymization principles, or propose efficient algorithms for computing the anonymized table under a given principle, and meanwhile minimizing the information loss. Following, we summarize the relevant literatures from these according to these two categories.

9.2.2.1 Negative Results

The techniques of fulfilling a generalization principle are usually limited by two inherent difficulties: the *hardness of optimal generalization* with minimum information loss and the *curse of dimensionality* for high-dimensional microdata.

Specifically, while there are numerous ways of generalizing the microdata table to satisfy the given privacy principle, one is usually interested in the optimal one which incurs the minimum amount of information loss, in order to preserve as much data utility as possible. However, finding such optimal generalization in general can be typically modeled as a search problem over a high-dimensional space, which is inherently difficult.

Meanwhile, when one has to deal with high-dimensional microdata, the curse of dimensionality becomes an important concern for privacy-preserving data publication. Intuitively, the combination of a large number of attribute values is so sparsely populated, that even to achieve 2-anonymity, one needs to suppress a large number of attribute values, resulting in almost useless anonymous data. Following, we present a brief survey of the existing literatures from these two perspectives respectively.

Hardness of Optimal Generalization. The first algorithm for computing generalization under the *k*-anonymity principle was proposed in [50]. It employs the domain generalization hierarchies of the QI-attributes to construct *k*-anonymous table. To minimize the information loss, the concept of *k*-minimal generalization was proposed, which requests for the minimum level of generalization, in order to maintain as much data utility as possible, for given level of anonymity. However, it is proved in [3], with the generalization height on the hierarchies as the metric of information loss, achieving the optimal *k*-anonymity with minimum information loss is NP-hard.

Meyerson et al. [43] studied the hardness of the optimal *k*-anonymity problem under the model of suppression, where the only allowed operation is to replace a QI-attribute value with a wildcard symbol '*'. As expected, it is theoretically proved that finding the optimal *k*-anonymous table with minimum number of suppressed cells is also NP-hard.

LeFevre et al. [36] proposed a multidimensional model for QI-attributes, and modeled the problem of constructing *k*-anonymous table as finding a par-

tition of the multidimensional space. They used the *discernability measure* of attribute values [6] as the metric of information loss, and proved that finding the optimal *k*-anonymous partition with minimum incurred information loss is NP-hard.

Curse of Dimensionality. In [2], Aggarwal analyzed the behavior of the suppression approach for implementing *k*-anonymity, in the presence of increasing dimensionality. It was observed that for high-dimensional microdata, in order to meet the anonymity requirement, a large number of attribute values need to be suppressed. The situation is even worse when the adversary has access to considerable background information, which is usually the case in practice.

As a result, in anonymizing high-dimensional microdata, the boundary between QI-attributes and sensitive attribute is blurred. On one hand, it becomes extremely hard to fulfill the anonymization principle; On the other hand, a large number of attribute values are generalized to wide ranges, resulting in significant loss of data utility. So far there is no formal analysis of the curse of dimensionality regarding generalization principles other than *k*-anonymity. However as suggested in [41], it tends to become increasingly infeasible to implement *l*-diversity as the dimension of the microdata grows.

9.2.2.2 Positive Results

Despite the hardness results of finding the optimal generalized relation, it is shown that with certain constraints on the resulting generalization [35], it is usually feasible to enumerate all the possible generalizations, and find the optimal one based on a set of heuristics [6, 35]. Meanwhile, efficient greedy-manner solutions naturally are also good candidates for solving NP-hard generalization problems. To this end, extensive research has been conducted on devising generalization algorithms based on these heuristic principles [36, 56, 23].

Meanwhile, though simple to implement and efficient to operate, such heuristic methods inherently can not guarantee the quality of the resulting generalization. Recognizing this drawback of heuristic methods, another line of research has been dedicated to developing approximation algorithms [43, 3, 47], in which the quality of the solution found is guaranteed to be within a certain constant factor of that by the optimal one. Following, we summarize the existing literatures on heuristic methods and approximation methods respectively.

Heuristic Methods. Intuitively, the first set of generalization algorithms base themselves on the following principle: enumerating all possible generaliza-

tions, and using heuristic rules to effectively prune non-promising solutions along the traversal process.

In [6], Bayardo and Agrawal proposed a k-anonymization algorithm based on the technique of set enumeration tree. Specifically, they impose a total order over all the QI-attribute domains, and the values in each domain are also ordered, which are preserved by the total order. Under this model, a generalization can be unambiguously represented as a union of the generalization sets for each QI-attribute, and finding the optimal k-anonymization involves searching through the powerset of all domains of QI-attributes for the generalization with the lowest cost. While it is impossible to construct the entire set enumeration tree, they apply a systematic set-enumeration-search strategy with dynamic tree rearrangement and cost based pruning. In particular, a tree node can be pruned when it is determined that no descendent of it could be optimal.

In [35], Lefevre et al. proposed the Incognito framework for computing optimal k-anonymization, based on bottom-up aggregation along domain generalization hierarchies. Specifically, they apply a bottom-up breadth-first search strategy over the domain hierarchies: At the ith iteration, it computes the i-dimensional generalization candidates from the (i-1)-dimensional generalizations, and removes all those generalizations which violate k-anonymity, i.e., in a similar spirit of Apriori algorithm [4]. Such search continues until no further candidates can be constructed, or all the domains of QI-attributes have been exhausted.

Aiming at finding suboptimal solutions, the greedy-manner solutions usually outperform the enumeration-based methods in terms of time and space complexity. Motivated by their superior computation efficiency, a set of generalization algorithms based on the greedy heuristic have been proposed [36, 56, 23].

In [36], a multidimensional generalization model was proposed for fulfilling k-anonymity, which could also be extended to support other generalization principles. Specifically, it constructs a generalization by partitioning the multidimensional space spanned by the QI-attributes, using a kd-tree structure. At each iteration, a partitioning dimension is selected with the minimum impact with respect to the quality metric in use, and its median value is chosen to partition the space. This process continues until no further partition which obeys k-anonymity can be constructed.

In [56, 23], two complementary anonymization frameworks have been proposed, based on bottom-up generalization and top-down specialization over the domain generalization hierarchies respectively. In [56] a bottom-up heuristic is applied: starting from an initial k-anonymity state, it greedily hill-climbs on improving an information-privacy metric which intuitively measures the information loss for a given level of k-anonymity. The approach is scalable in the sense that it examines at most one generalization at each iteration for each QI-attribute. However, like all other local search methods, it may get stuck at

some local optimum, if the initial configuration is not proper. Fung et al. [23] presented a complementary top-down specialization strategy, which starts from a general solution, and specializes certain QI-attributes, so as to minimize the information loss, without violating k-anonymity.

As mentioned above, the problem of finding high-quality generalization can be modeled as a search problem over a high-dimensional space, therefore a set of general heuristic tools, e.g., genetic algorithms and simulated annealing can be readily applied. In [31], Iyengar defines a utility metric in terms of classification and regression modeling, and applies genetic algorithm to optimizing the utility metric, for the given level of k-anonymity. Unfortunately, at the cost of large computational complexity, this solution offers no guarantees on the solution quality.

Approximation Methods. Along another line of research efforts, researchers have been seeking the techniques which provide guarantees on the solution quality, as measured by its deviation from that of the optimal one. A number of approximation generalization algorithms have been proposed [43, 3, 47], which guarantees the quality of the found solution to be within a constant factor or that found by the optimal one.

In [43], Meyerson and Williams present a polynomial time algorithm for optimal k-anonymity, which achieves an approximation ratio of O(k log k), independent of the size of the dataset. However, the runtime is exponential in terms of k. They further remove this constraint by proposing an O(k log m)-approximation, where m is the degree of the relation.

Aggarwal et al. [3] improved the result of [43] by showing an algorithm with an approximation ratio of O(k). They also provided improved positive results for specific values of k, including a 1.5-approximation algorithm for k = 2, and a 2-approximation algorithm for k = 3.

In a recent work [47], Park and Shim present an algorithm which achieves an approximation ratio of O(log k). Furthermore, by explicitly modeling the trade-off between solution quality and execution efficiency, they proposed an O(β log k)-approximate algorithm, which allows the users to trade the approximation ratio for the running time of the algorithm by adjusting the parameter β.

9.2.2.3 Utility Optimization

Generally, the protection for the sensitive personal information is achieved by performing certain transformation over the microdata, which necessarily leads to the loss of the data utility for the purpose of data analysis and mining. Because of this inherent conflicts between privacy protection and data utility, it is imperative to take account of information loss in the privacy preservation process, and optimize the data utility under the required privacy protection.

This is especially important when dealing with high-dimensional microdata, as shown in [2], for high-dimensional microdata, a large number of attribute values need to be generalized, even to achieve a modest level of anonymity, due to the cure of dimensionality.

To this end, extensive research has been conducted on optimizing the utility of the anonymized data, without violating the hard privacy requirement. The existing literatures can be summarized as three subcategories: The first one attempts to devise general-purpose metrics of data quality, and incorporates them in the anonymization process [6, 41, 31, 56]; The second one focuses on alleviating the impact of the anonymization operation over the data utility [33, 61, 65]; The third one targets specific applications, and tailors the anonymization to preserve maximum data utility with respect to the particular applications [56, 37].

Data Utility Metrics. Various measurements of the information loss caused by the anonymization operation have been proposed in the literatures:

- Generalization height [6]. It is the level of the generalized QI-attribute value on the domain generalization hierarchy. The problem with this notion is that not all generalization levels are of equal importance, and a generalization step on one attribute may include more tuples into an anonymous group than a generalization step on another one [41].
- Average size of anonymous groups [41] and discernability measure of attribute values [6]. Both metrics take account of the QI-group size. Specifically, discernability assigns a cost to each tuple based on how many tuples are indistinguishable from it, and depending on whether the tuple is suppressed, the cost is either the size of the microdata table, or the size of the QI-group. However, neither of the two metrics take consideration of the underlying distribution of the data, which may reveal important information.
- Classification metric [31] and information-gain-privacy-loss-ratio [56]. Both metrics take account of the distribution of the underlying microdata. The classification metric is designed specifically for the purpose of training classifier over the data, therefore may not be appropriate for measuring general-purpose information loss. The information-gain-privacy-loss-ratio is a local heuristic used to determine the next generalization step, similar in spirit to the information gain metric for deciding the splitting point in a decision tree. It is also not clear how to apply it to measure general-purpose information loss.

Utility-Based Anonymization. This category of work focuses on improving the anonymization operation itself, in order to alleviate its impact over the quality of the resulting data, without compromising the privacy protection.

Instead of anonymizing the microdata table as a whole, Kifer and Gehrke [33] advocated publishing the marginals, each of which anonymizes the projection of the microdata table on a subset of QI-attributes and sensitive attributes, in order to ameliorate the effect of the curse of dimensionality. It is shown that this approach can preserve considerable utility of the microdata, without violating hard privacy requirements.

Xiao and Tao [61] proposed a simple alternative anonymization framework, called *anatomy*, which essentially publishes QI-attributes and sensitive attributes in two separate tables. This way, the QI-attributes values need not to be generalized, since the separation already provides the same amount of protection as generalization, in the case that the adversary already knows the presence of the individuals in the microdata table. It is shown that this approach brings considerable improvement over the data quality in answering aggregation queries. A similar idea was also proposed by Zhang et al. [65].

Motivated by the fact that different attributes tend to have different utility with respect to the applications in question, a utility-based anonymization method using local-recording has been proposed in [63], which takes into consideration of the different utility weights of the QI-attributes, in performing the generalization operation.

Application-Specific Utility. Another direction of preserving data utility in the privacy protection process is to minimize the information loss in an application specific or workload-specific manner. In such cases, the utility metrics are defined depending on the underlying application or workload.

In [56], using information-gain-privacy-loss-ratio as a local heuristic, designed specifically for the classification task, a bottom-up generalization algorithm has been proposed. In [65], the anonymization algorithm is optimized for the purpose of answering aggregation queries.

An interesting workload-aware generalization algorithm has been proposed by Lefevre et al. [37]. Specifically, it differentiates the subsets of microdata according to their frequency of being requested by the users, and performs less generalization over those frequently requested subsets than others, thus achieving considerable information saving, while providing the same amount of privacy protection.

9.3 Location Privacy

Along with the great convenience of the location-based services, follow their potential threats to the users' privacy and security: the location information disclosed by the mobile users could be abused in malicious ways. Location privacy protection concerns about preventing the exposure of individuals' private location information (associated with identities) to unauthorized party.

In general, such exposures can be loosely categorized as three classes: the first one happens through direct communication, i.e., the communication channel is eavesdropped by the adversary; the second one occurs through direct observation, e.g., the LBS service provider is not trusted, therefore having direct observation over the received location and identity information; the third one takes place through indirect inference of location information combined with other properties of individual, e.g., by tracking the moving pattern of a victim individual to infer his/her identity.

As a special form of data privacy, location privacy possesses its unique characteristics and research challenges:

- First, location privacy requirements are inherently personalized, and context-specific. Different users tend to have various location privacy requirements, e.g., some users regard their positions as extremely private information, while others may care much more about the quality of the service delivered. Moreover, users tend to have different requirements with respect to the context, e.g., users may have stricter privacy requirement during night time, to reduce the risk of being stalked.
- Second, location privacy is intricately related to the service quality. The service provider processes a request based on its understanding regarding the customer' position, and more precise location information leads to higher quality of the service delivered, while ambiguous or fake location information may result in the degradation of the service quality. Therefore, there exists an implicit trade-off between location privacy and location utility.
- Third, location data is extremely dynamic, and subjected to frequent updates. Unlike that in protecting ordinary data privacy, where the microdata is typically stored in databases, in preserving location privacy, the location data is usually processed in an on-line stream manner, in order to meet the strict requirement of response time. Meanwhile, customers' location data is usually subjected to frequent updates, which opens the door for the adversary to combine multiple “snapshots” to infer individuals' current location.

In general, based the assumptions regarding the trustfulness of the LBS service providers, customers' location privacy requirements can be fulfilled using two methods: Under the model of trusted service providers where the service providers faithfully act according to its agreement with the customers, users' location privacy can be preserved using a policy-based approach: the customers specify their privacy requirements in the form of consents with the service providers. However, with the explosive growth the LBS services, it becomes extremely difficult to fulfill this model in practice.

Under the model of untrusted service providers, which reflects the current main trend, it is imperative to provide technical countermeasure against poten-

tial privacy violation by the service providers. One simple solution is that instead of using their true identities, the mobile users provide pseudonyms to request for services. However, this solution is generally insufficient for two main reasons: (i) A set of applications require to verify customers' true identifies in order to provide the corresponding services, e.g., credit-card related services. (ii) A user's identity can be potentially inferred from his/her location information. Several types of inferences are possible, as shown in [28, 7]: in precise location tracking, successive position updates can be linked together to form moving patterns, which can be used to reveal the users' identity; in observation identification, external observation is available to link a position update to an identity; in restricted space identification, a known location owned by identity relationship can link an update to an identity.

Hence, a set of location hiding techniques have been proposed to address users' location privacy concerns, including reducing the granularity of the representation of users' location information (spatial/temporal cloaking, location blurring), reporting the nearest landmark, sending false dummy locations, location obfuscation, etc. Essentially, location hiding provides protection for the location privacy of a mobile user by guaranteeing that no adversary can pinpoint this particular user to a location with high precision and confidence, through inference attacks. This system property is termed location anonymization. The rest of the section is devoted to a brief survey of the current state-of-the-art location anonymization techniques.

9.3.1 Models and Principles

Analogous to preserving data privacy, in order to provide adequate protection for mobile users' location privacy, various location anonymization principles have been proposed, including location k-anonymity [28], location l-diversity [5], and minimum spatial resolution [44]. Meanwhile, in order to guarantee the quality of the service delivered, several QoS metrics have been proposed, including maximum tolerable spatial and temporal resolutions [24, 25]. In the first part of the section, we describe in detail the location privacy profile, which captures mobile users' privacy and QoS requirements.

The existing location anonymization techniques can be roughly classified into three main categories based on their system architectures, concretely centralized trusted third party model, client-based non-cooperative model, and decentralized cooperative mobility group model. Each system architecture model is associated with its unique assumptions about the privacy threat model, and supports a different set of location anonymization mechanisms. In the second part of the section, we discuss the strengths and weaknesses of these architecture models respectively.

9.3.1.1 Location Privacy Profile Model

In location privacy profiles, mobile users specify their privacy protection and quality of service requirements. As have been mentioned earlier, location privacy protection is inherently personalized, and context-specific, which implies that the location privacy profiles need to be specified on a per-user, per message basis, and the system should support users to change their privacy profiles in an on-line manner.

Location Privacy Metrics. Borrowing from ordinary data privacy the idea of *k*-anonymity [51], the principle of *location k-anonymity* has been introduced in [28] to preserve the location privacy of mobile users through the use of location *k*-anonymization. Intuitively, location *k*-anonymity ensures that at a given time instance, for each LBS service request, there are at least (*k*-1) other messages with the same location information, each associated with a different (pseudo) identity. It guarantees that without further knowledge, the adversary can not differentiate at least *k* participants with respect to their location information. However, the definition of location *k*-anonymity in [28] is limited to a system-supplied uniform *k* for all users in a given LBS system. Gedik and Liu [24, 25] revised the initial definition by introducing the concept of personalized location *k*-anonymity, allowing variable *k* for different users and for different service requests of the same user. Most of the subsequent research on location privacy has adopted the concept of personalized location *k*-anonymity [44, 45, 5].

In [41], *k*-anonymity is proved insufficient for protecting data privacy in the sense that it ignores the characteristics of the sensitive attribute, and opens the door for the inference of sensitive attribute values. Interestingly, an analogous situation also happens for location privacy protection.

The first weakness of location *k*-anonymity is that it ignores the granularity of the reported location, i.e., under location *k*-anonymity, a location of fairly small area is considered as providing enough protection, as long as it contains more than *k* active mobile users, even though the adversary can pinpoint the users with high precision. The concept of *minimum spatial resolution* [44] has been introduced to address this problem. It allows each user to specify the minimum spatial area of his/her released location.

The second weakness of location *k*-anonymity is its ignorance of the number of symbolic locations or static objects associated with the reported location. Intuitively, if a location *k*-anonymous spatial region is associated with only one static object, e.g., church or doctor's office, then with this region as the reported location of a user, the adversary may associate the user with the specific location object with high probability. This qualifies as severe privacy threat, considering a simple example: If *Bob*'s reported location is an area containing a specific clinic as the only static object, an adversary may infer that "*Bob must be visiting the clinic with high probability*". The concept of loca-

tion *l*-diversity [5, 40] has been introduced into the framework of location anonymization to address this weakness. Intuitively, it demands that for each LBS request, in addition to user-level *k*-anonymity, the released location should also be associated with at least *l* different symbolic objects. Mobile users can specify desired *k* (*k*-anonymity) and *l* (*l*-diversity) values in their location privacy profile [40].

Note that in order to fulfill location *l*-diversity, the component responsible for location anonymization is expected to support efficient access to the databases of public location objects, which makes it difficult to fulfill this principle under the client-based anonymization architecture, as will be discussed in Section 9.3.1.2.

Quality of Service Metrics. Meanwhile, in addition to the location privacy requirements, the users may also desire to specify their demands for the quality of service delivered to them. Two commonly used QoS metrics are *maximum tolerable spatial* and *temporal resolutions* [24, 25].

Specifically, maximum tolerable spatial resolution represents the threshold on the maximum area of the anonymized location. As have been mentioned above, depending largely on the location information provided by the client, the quality of a LBS service tends to degrade as the level of anonymization increases. Consider the example of querying the nearest gas station: if the client reports his/her location as a wide area, a large number of candidate results will be returned, which not only incurs heavy computation overhead of filtering false positive information, but also wastes the precious wireless bandwidth. The maximum tolerable resolution bounds the level of anonymization, therefore indirectly guaranteeing the quality of the delivered service.

Another important QoS metric is the response time of the service request. As will be discussed in Section 9.3.2, under the centralized location anonymization architecture, caching the service requests temporarily can improve the throughput of the location anonymization component, at the cost of delaying the processing of the requests. Therefore, mobile users are inclined to specify threshold for the allowed delay time as the maximum tolerable temporal resolution.

9.3.1.2 Location Anonymization: Alternative System Architectures

Three alternative system architectures have been proposed for implementing location anonymization [40]: centralized architecture with trusted third-party, non-corporative client-based architecture, and decentralized peer-to-peer corporative architecture. In this section, we briefly describe each of these system architecture models, including their privacy threat models, as well as their advantages and drawbacks. We defer the discussion of the concrete techniques developed under each architecture model to the next section.

Centralized Architecture. In a centralized location anonymization architecture, a trusted third party, called *location anonymizer component*, acts as a proxy (middleware) for all communications between mobiles users and LBS service providers.

The overall communication between the clients, the location anonymizer, and the service provider can be loosely divided into four phases: (1) The location anonymizer receives the service request, plus the position information from the mobile client, and performs spatial and temporal anonymization over the location information, based on the user-specified privacy profile. Meanwhile, according to the performed transformation, it generates the filtering condition that will be used later to prune the false positive information from the candidate results as returned by the service provider, to produce the exact result to the original request; (2) The location anonymizer then relays the request associated with anonymized location information to the corresponding service provider. Upon receiving the anonymized request, the service provider invokes the anonymous request processing module, which produces the set of candidate results; (3) After receiving the candidate results from the service provider, the location anonymizer performs the filtering operation to remove the false positive results, by applying the corresponding filtering condition; (4) Finally, the exact results to the original request are delivered to the mobile user who issues it.

Note that under this architecture, the location anonymizer is responsible for both anonymizing the LBS service requests, and filtering false positive information from the candidate results, which could potentially become the bottleneck of the system. An alternative design suggested in [5] is to perform the filtering at the client side: for each service request received at the location anonymizer, the anonymized request will be relayed to the service provider, and the filtering condition is returned to the mobile user who issues the request. The service provider then directly passes the candidate results to be filtered at the client side. This design however introduces additional communication and processing overhead for the mobile clients.

The privacy threat model commonly addressed by this centralized location anonymization architecture can be specified as follows: (i) The true identity of the mobile user is hidden from the service provider, as specified by the security policy, and the LBS application is assumed to accept pseudo-identity; (ii) The LBS service provider is considered as hostile observer, therefore setting constraints on what information can or cannot be revealed from one service provider to anther is usually insufficient; (iii) The location sensing infrastructure (e.g., GPS, WiFi, Cricket) and the location anonymizer are trusted by the mobile users.

Since the location anonymizer gathers the location information from all the active users, and is usually able to provide efficient access to the database of public location objects, the strength of this architecture is its being able to

support all the location anonymization requirements as mentioned in Section 9.3.1.1 (strong privacy protection). Meanwhile, since all the communications pass through the location anonymizer component, which could potentially become the bottleneck of the whole architecture.

Non-Cooperative Architecture. In a non-cooperative architecture, the mobile users maintain their location privacy based on their knowledge only, without involving any centralized trusted authority.

In this client-based location anonymization model, the communications between the mobile client and the LBS service providers follow the model as: (1) the client obfuscates its location information, which is sent, in conjunction with the service request to the service provider. Meanwhile, the client generates the filtering condition according to the location perturbation performed; (2) Upon receiving the anonymized request, the service provider invokes the anonymous request processing module, which produces the set of candidate results, to be delivered to the mobile client; (3) After receiving the candidate results from the service provider, the client obtains the exact answer to the original request by filtering out the false positive information from the candidate results.

The privacy threat model commonly addressed in the client-based non-cooperative location anonymization model is as follows: (i) The true identities of the users are hidden from the service provider, and the LBS application is considered to accept pseudonyms; (ii) No third authority (e.g., LBS service provider, location anonymizer) are trusted, who are interested in intruding mobile users' location privacy; (iii) The location sensing infrastructure (e.g., GPS, WiFi, Cricket) and the location anonymizer resided on mobile clients are trusted.

Under this model, all the communications are inherently distributed, leading to its strength of high throughput and fault-tolerance, and no centralized trusted authority is needed. Meanwhile, since the location perturbation is performed at the mobile clients, without knowing the location information of other clients, and it is usually not affordable for the clients to directly access the databases of public location objects (one of the main LBS applications), this model does not support location k-anonymity or location l-anonymity, i.e., weak privacy protection.

Peer-to-Peer Cooperative Architecture In a peer-to-peer location anonymization framework, a group of mobile users collaborate with each other to provide location privacy protection for each single user, without the interleaving of a centralized trusted entity. Specifically, within this architecture, the communications between the mobile clients and the service providers follow the model as: (1) the request-issuing mobile client communicate with neighboring clients to collect the location information of peers, and performs informed location perturbation to meet its privacy requirement. It (or a peer) then sends

the service request, plus the anonymized location information to the LBS service provider, meanwhile generating the filtering condition; (2) upon receiving the anonymized request, the service provider invokes the anonymous request processing module, and produces the set of candidate results, to be delivered to the mobile client who sends the request; (3) The results are routed to the client who issues the request (if the issuer and the sender are different), which then obtains the exact results by filtering the false positive information from the candidate results.

Commonly, the peer-to-peer cooperative location anonymization architecture is applied to address the privacy threat model as follows: (i) the true identity of the mobile user is hidden from the service provider, and the LBS service provider is considered to accept pseudonyms; (ii) all third parties (e.g., LBS service providers, location anonymizer) are considered as hostile observers, who are interested in intruding mobile users' privacy; (iii) the location sensing infrastructure, the location anonymizer resided on mobile clients, and the peers in the mobility group are trusted.

This architecture requires no centralized trusted authority, therefore can scale to a large number of mobile clients. However, it is shown in [27] that it is naturally computationally expensive to support location anonymity under this architecture, since each client has to communicate with each other to construct anonymous location. Also, to support this model, a mobile peer-to-peer communication infrastructure is needed. Clearly, this architecture offers support for location *k*-anonymity; nevertheless, it is inherently difficult to support location *l*-diversity since it is usually not feasible for mobile clients to directly access the databases of public location objects. In conclusion, this architecture provides protection stronger than the client-based non-cooperative architecture, but weaker than the centralized architecture.

9.3.2 Location Anonymization Techniques

We have given an overview of location privacy research, with focus on privacy models and system architectures. In this section, we present a brief survey of the representative techniques developed for aforementioned system architectures, and analyze their advantages and drawbacks.

9.3.2.1 Centralized Architecture

As a pioneering work on location k-anonymization, Grusteser and Grunwald [28] introduced a quadtree-based spatial cloaking method, under the centralized location anonymization model. Intuitively, for a given set of service requests, it recursively divide entire geographical space of interest into quad-

rants until a quadrant has less than k users, and then returns the previous quadrant, i.e., the minimum quadrant that meets k-anonymity, as the anonymous location for the mobile users within it. An example is shown in Fig.9.1, where the dashed box is the minimum quadrant that contains more than k = 4 users.

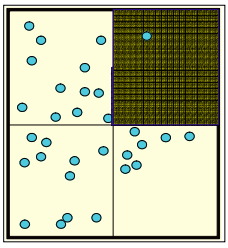

Fig. 9.1 Quadtree based spatial cloaking

Using a quadtree-based spatial partitioning technique, this method provides efficient support for *universal location k-anonymity*, where a system-supplied uniform *k* is used and thus the same level of *k*-anonymity is applied to all the users. However, as a straight-forward use of the spatial partitioning techniques, it is inherently difficult to support personalized location privacy in this framework.

As mentioned earlier, motivated by the need to support personalized location privacy requirements, Gedik and Liu [24, 25] introduced the CliqueCloak framework. It allows each message to specify a different k value based on its specific privacy requirement, and maximum spatial and temporal tolerance values based on its QoS requirements. The service requests are processed in an on-line stream manner: a constraint graph is constructed according to the maximum spatial and temporal tolerance settings of the requests, and a set of requests are perturbed if they form a clique in the constraint graph. An example is shown in Fig.9.2, where three messages m_1, m_2 and m_4 form a clique on the constraint graph, and are anonymized together. However, identifying cliques in a graph is an expensive operation, which severely limits the scalability of this framework. It has been empirically shown that this approach can only support fairly small *k*.

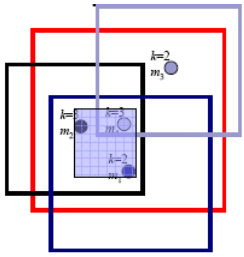

Fig. 9.2 Constraint graph based spatial cloaking.

Both frameworks above focus solely on devising efficient and effective solutions for the location anonymizer component. In [44, 45], Mokbel et al. proposed Casper, a complete location anonymization framework. It employs a hierarchical indexing structure: the universe of disclosure is represented as a complete pyramid structure, which is divided into grids at different resolution levels and each cell maintains the number of mobile users inside. It supports personalized location *k*-anonymization: on anonymizing a request, it traverses the pyramid structure bottom-up until a cell satisfying the user's privacy profile is found. They further improved this basic paradigm by introducing adaptive searching strategy, at the cost of maintaining the adaptive pyramid structure. A unique contribution of the Casper framework is its support for anonymous query processing, which extends existing spatial query processing and optimization techniques to support *k*-anonymous nearest neighbor and range queries.

It is observed in [5] that the CliqueCloak method models requests in a constraint graph and achieves anonymization by finding cliques, resulting in its low scalability. Meanwhile, the Casper method adopts a hierarchical pyramid structure, and supports anonymization by using quadrant spatial expansion. It is therefore inherently not optimal in the sense that the cloaked location is usually much larger than the minimum area satisfying *k*-anonymity. The PrivacyGrid framework developed at the DiSL group of Georgia Tech [5] addresses the drawbacks of these two frameworks, by employing a dynamic grid-based spatial index, powered with fast bottom-up, top-down and hybrid cloaking algorithms. In addition to supporting personalized location privacy and QoS requirements as defined in [24], PrivacyGrid enhances the location anonymization semantics by introducing location *l*-diversity. By adopting a flat grid index structure, PrivacyGrid achieves better efficiency than the CliqueCloak approach in terms of supporting large *k*, meanwhile higher success rate of request anonymization, and superior quality of anonymized location than the Casper approach.

9.3.2.2 Non-Cooperative Architecture

Under the non-cooperative client-based architecture, the location privacy of mobile users is usually achieved by injecting uncertainty to their location information, called location obfuscation, based solely on their own knowledge.

In [29], Hong and Landay proposed to use the nearest public landmark, instead of his/her current exact position, as a mobile user's locations in constructing a LBS request. Specifically, based on a set of public landmarks, a Voronoi diagram is constructed over the universe of disclosure, which is em-

ployed by the mobile users to identify their closest landmarks. Clearly, within this framework, the evaluation of the request is also based on the selected landmark; hence the distance between the selected landmark and the user's exact location controls the trade-off between privacy protection and service quality.

Kido et al. [32] proposed to use a set of false dummies to protect user's exact location: the reported location information consists of m locations, among which only one is the true position while other (m-1) false dummies. Only the user who issues the request knows the true location, and the service provider replies with a service for each received location. It is clear that within this model, the privacy protection is achieved at the cost of the query evaluation overhead at the server, and the communication bandwidth, since the server essentially has to evaluate m queries in order to answer one original request.

In [18], Duckham and Kulik proposed a location obfuscation approach based on a graph abstraction of the universe of disclosure. Specifically, all locations are represented as vertices in a graph with edges corresponding to the distance between two locations. A user obfuscates her location as a set of vertices, and the query is then evaluated at the server side based on the distance to each vertex in the imprecise location. Clearly, it suffers from the same problem as the false dummies approach.

In a recent work [64], Yiu et al. proposed the SpaceTwist framework to support kNN queries based on the incremental NN query processing techniques. Specifically, the mobile user reports a location different from (but near to) his/her actual position, and the nearest neighbor objects of the reported position are incrementally retrieved until the query is correctly answered by the service provider. Though simple to implement, the SpaceTwist framework incurs additional overhead on the communication bandwidth, for the protocol involves multiple exchanges of packets to ensure that all the actual results have been retrieved. Moreover, this approach offers no explicit modeling of the quality of privacy protection, which makes it difficult for users to quantitatively specify privacy requirements.

Recently, transformation-based matching techniques have been proposed to enable location privacy, which however do not offer query accuracy guarantees. In [30], a theoretical study on a client-server protocol for deriving the nearest neighbor of a query is reported. In [34], Khoshgozaran et al. defined a specific Hilbert ordering based on a key, whose value is known only by the client and a trusted entity. It is shown that without the key value, it is impossible to decode a Hilbert value into a location correctly. However, a Hilbert curve does not completely preserve spatial proximity, so the reported result can be far from the actual result. To improve the accuracy, they proposed to use two keys with orthogonal Hilbert curve.

9.3.2.3 Peer-to-Peer Cooperative Architecture

In [26, 27], Ghinita et al. proposed PRIVE, a representative framework of the peer-to-peer cooperative location anonymization model. The main idea of PRIVE is that whenever a mobile user desires to issue a LBS request, it broadcasts a group formation request to its neighbors, and a member of the group is randomly selected to act as the query sender. PRIVE provides two modes of group formation: on-demand mode and proactive mode. In on-demand mode, a mobile user invokes the group formation only when necessary; In proactive mode, mobile users periodically execute the on-demand approach to maintain their anonymous groups. Clearly, the two modes represent the two ends of the spectrum of the trade-off between maintenance cost and response time of anonymous group formation.

Specifically, in PRIVE, an anonymous location is obtained in three phases: (1) Peer searching. The user who intends for a LBS service broadcasts a multi-hop request until at least (k-1) peers are found; (2) Location adjustment. The initial anonymized location is adjusted according to the velocities of group members; (3) Spatial cloaking. The anonymized location is cloaked into a region aligned to a grid covering the (k-1) nearest peers.

9.3.3 Open Issues and Challenges

While a plethora of work has been done on preserving the location privacy of mobile users in LBSs, the techniques developed to date are far from being adequate in providing comprehensive solutions to the location privacy problem in general. In this section, we discuss open issues and possible research directions towards constructing a complete end-to-end solution for location privacy protection.

Enhancement of Location Anonymization Techniques. As a special form of data privacy, location privacy exhibits its unique characteristics and challenges. Thus it is inadequate to directly apply the extensive collection of data privacy techniques proposed to date for location privacy protection. To the best of our knowledge, the successful extensions of data privacy preserving techniques to location privacy are still limited to *k*-anonymity [49, 51], *l*-diversity [41] and distance-preserving space transformation techniques [9, 10]. One possible direction is to further extend the available data privacy techniques for enhancing existing location anonymization tools.

For example, the *m*-invariance [62] principle has been proposed for sequential releases of microdata, which can be possibly adapted to protecting location privacy for continuous location-based queries. For such continuous location updates/queries, an alternative solution can possibly be the extension of

the output perturbation techniques [57], originally developed for countering inferences over the mining results of consecutive stream windows.

The aforementioned three location anonymization architectures possess their own strengths and weaknesses. Though the centralized architecture provides the strongest privacy protection, and extensive research efforts so far have been focused on developing anonymization techniques under this architecture, with the explosive growth of the scale of mobile clients and LBS services, the client-based and decentralized architectures are anticipated to become the main trends. Therefore it is imperative to put forward the research effort on remedying their weaknesses by introducing stronger privacy guarantees.

Potential Attacks and Countermeasures. A variety of inference attack models have been studied in the area of ordinary data privacy, which show that by leveraging certain external knowledge and her understanding regarding the anonymization techniques, an adversary can potentially penetrate the protection mechanism and infer the sensitive information of individuals. For example, in [62, 55], it is shown that in the scenario of multiple releases of microdata, the adversary may potentially exploit previous releases to infer sensitive information in the current one. In [60], it is shown that knowledge of the mechanism or algorithm of anonymization can also lead to extra information that assists the adversary and jeopardizes individual privacy.

Most existing location anonymization tools have not taken full account of the possible attacks to the anonymized solutions, resulting in their weak resilience to inference attacks.

For example, analogous to the sequential releases of microdata, continuous updates/queries in LBSs can be exploited to infer sensitive location or identity information of individuals. In [17], a mobility model based attack is considered: The adversary utilizes the knowledge about mobile users' motion parameters such as maximum velocity, known trajectory, frequent travel path to perform inference attacks. For instance, if the adversary knows the maximum velocity of the user, and obtains its consecutive cloaked location updates, by computing the maximum movement boundary of the same identity (pseudonym), she can potentially locate the user at the intersection of two cloaked spatial regions. In [12], an overlapping window based inference attack model is considered. Intuitively, if the adversary knows some locations of the targeted victim, even different pseudonyms are used in different updates/queries, by analyzing the overlapping spatial or temporal windows of two consecutive cloaked location updates/queries, she can infer the linkage of location with the targeted victim.

As another example, analogous to the minimality attack in publishing microdata [60], it is shown in [58] that knowing the principles applied in the location anonymization algorithms, e.g., optimality of service quality for given privacy requirement, and the underlying background of users' movement, e.g.,

road network, the adversary can pinpoint a victim individual with high precision.

Therefore, a promising future direction is to study such potential attack models, and to devise effective countermeasures to inject attack resilience into current location anonymization tools.

General Framework for Anonymous Query Processing The anonymous location query processing modules in literatures, e.g, [44], are designed to support only limited types of spatial queries, e.g., range queries, or *k*NN queries, and are constructed over the top of existing spatial query processing techniques. As mobile users' location privacy becomes a paramount concern in numerous LBS services, general location query processing modules designed specially for anonymous spatial queries are expected to bring significant impact over the industries. One future research direction is to extend the existing spatial query processing techniques, e.g.,[52, 48], to support more types of queries, e.g., continuous *k*NN queries.

9.4 Summary

We have presented an overview of the advances in the area of data privacy research. We discussed a variety of privacy principles and models, and analyzed their implicit relationships. We also summarized the representative algorithms and methods for implementing the proposed principles, and discussed their fundamental limitations. In a perspective analogous to data privacy, we examined the state of the art of the location privacy research, and compared the strengths and weaknesses of the proposed models, architectures, and algorithms. We concluded the chapter with a discussion of open issues and promising research directions towards providing comprehensive end-to-end location privacy solutions.

Acknowledgments This work is partially supported by grants from NSF CyberTrust program, an AFOSR grant, and an IBM SUR grant.

References

[1]. N. Adam, and J. Wortman. Security-control methods for statistical databases. *ACM Computing Surveys*, 21(4), 1989.
[2]. C. Aggarwal. On *k*-anonymity and the curse of dimensionality. In *VLDB*, 2005.
[3]. G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigraphy, D. Thomas, and A. Zhu. Anonymizing tables. In *ICDT*, 2005.
[4]. R. Agrawal, and R. Srikant. Fast algorithms for mining association rules. In *VLDB*, 1994.

[5]. B. Bamba, L. Liu, P. Pesti, and T. Wang. Supporting anonymous location queries in mobile environments with PrivacyGrid. In *WWW*, 2008.
[6]. R. Bayardo, and R. Agrawal. Data privacy through optimal *k*-anonymization. In *ICDE*, 2005.
[7]. A. Beresford. Location privacy in ubiquitous computing. PhD thesis, University of Cambridge, 2005.
[8]. B. Chen, R. Ramakrishnan, and K. LeFevre. Privacy skyline: privacy with multidimensional adversial knowledge". In *VLDB*, 2007.
[9]. K. Chen, and L. Liu. A random rotation perturbation approach to privacy preserving data classification. In *ICDM*, 2005.
[10]. K. Chen, and L.Liu. Towards attack-resilient geometric data perturbation. In *SDM*, 2007.
[11]. F. Chin, and G. Ozsoyoglu. Auditing and inference control in statistical databases. *IEEE Trans. Softw. Eng.*, SE-8(6), 1982.
[12]. C. Chow, and M. Mokbel. Enabling private continuous queries for revealed user locations. In *SSTD*, 2007.
[13]. L. Cox. Suppression methodology and statistical disclosure control. *J. Am. Stat. Assoc.*, 75(370), 1980.
[14]. T. Dalenius, and S. Reisss. Data swapping: a technique for disclosure control. *J. Stat. Plan. Infer.*, 6, 1982.
[15]. D. Denning. Secure statistical databases with random sample queries. *ACM TODS*, 5(3), 1980.
[16]. D. Dobkin, A. Jones, and R. Lipton. Secure databases: Protection against user influence". *ACM TODS*, 4(1), 1979.
[17]. J. Du, J. Xu, X. Tang, and H. Hu. iPDA: enabling privacy-preserving location-based services". In *MDM*, 2007.
[18]. M. Duckham, and L. Kulik. A formal model of obfuscation and negotiation for location privacy. In *Pervasive*, 2005.
[19]. G. Duncan, S. Fienberg, R. Krishnan, R. Padman, and S. Roehrig. Disclosure limitation methods and information loss for tabular data. *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, pp 135-166, Elsevier, 2001.
[20]. C. Farkas, and S. Jajodia. The inference problem: a survey. *SIGKDD Explor. Newsl.*, 4(2), 2002.
[21]. I. Fellegi. On the question of statistical confidentiality. *J. Am. Stat. Assoc.*, 67(337), 1972.
[22]. Foxs News. Man accused of stalking ex-grilfriend with gps. http://www.foxnews.com /story/0293313148700.html.
[23]. B. Fung, K. Wang, and P. Yu. Top-down specialization for information and privacy preservation. In *ICDE*, 2005.
[24]. B. Gedik, and L. Liu. Location privacy in mobile systems: a personalized anonymization model". In *ICDCS*, 2005.
[25]. B. Gedik, and L. Liu. Protecting location privacy with personalized *k*-anonymity architecture and algorithms. *IEEE Transactions on Mobile Computing*.
[26]. G. Ghinita, P. Kalnis, and S. Skiadopoulos. MOBIHIDE: a mobile peer-to-peer system for anonymous location-based queries. In *SSTD*, 2007.
[27]. G. Ghinita, P. Kalnis, and S. Skiadopoulos. PRIVE: anonymous location based queries in distributed mobile systems. In *WWW*, 2007.
[28]. M. Gruteser, and D. Grunwald. Anonymous usage of location-based services through spatial and temporal cloaking. In *MobiSys*, 2003.
[29]. J. Hong, and J. Landay. An architecture for privacy-sensitive ubiquitous computing. In *MobiSys*, 2004.
[30]. P. Indyk, and D. Woodruff, Polylogarithmic private approximations and efficient matching. In *TCC*, 2006.
[31]. V. Iyengar. Transforming data to satisfy privacy constraints. In *KDD*, 2002.

[32]. H. Kido, Y. Yanagisawa, and T. Satoh. An anonymous communication technique using dummies for location-based Services. In *ICPS*, 2005.
[33]. D. Kifer, and J. Gehrke. Injecting utility into anonymization databases. In *SIGMOD*, 2006.
[34]. A. Khoshgozaran, and C. Shahabi. Blind evaluation of nearest neighbor queries using space transformation to preserve location privacy. In *SSTD*, 2007.
[35]. K. LeFevre, D. Dewitt, and R. Ramakrishnan. Incognito: efficient full-domain *k*-anonymity. In *SIGMOD*, 2005.
[36]. K. LeFevre, D. DeWitt, and R. Ramakrishnan. Mondrian multidimensional *k*-anonymity. In *ICDE*, 2006.
[37]. K. LeFevre, D. DeWitt, and R. Ramakrishnan. Workload aware anonymization. In *SIGKDD*, 2006.
[38]. J. Li, Y. Tao, and X. Xiao. Preservation of proximity privacy in publishing numerical sensitive data. In *SIGMOD*, 2008.
[39]. N. Li, T. Li, and S. Venkatasubramanian. *t*-closeness: privacy beyond *k*-anonymity and *l*-diversity. In *ICDE*, 2007.
[40]. L. Liu. From data privacy to location privacy. In *VLDB*, 2007.
[41]. A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. *l*-diversity: privacy beyond *k*-anonymity. In *ICDE*, 2006.
[42]. D. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Halpern. Worst-case background knowledge in privacy. In *ICDE*, 2007.
[43]. A. Meyerson, and R. Williams. On the complexity of optimal *k*-anonymity. In *PODS*, 2004.
[44]. M. Mokbel, C. Chow, and W. Aref. The new casper: query processing for location services without compromising privacy. In *VLDB*, 2006.
[45]. M. Mokbel. Privacy in location-based services: state of art and research directions. In *MDM,* 2007.
[46]. M. Nergiz, M. Atzori, and C. Clifton. Hiding the presence of individuals from shared databases. In *SIGMOD*, 2007.
[47]. H. Park, and K. Shim. Approximate algorithm for *k*-anonymity. In *SIGMOD*, 2007.
[48]. S. Saltenis, C. Jensen, S. Leutenegger, and M. Lopez. Indexing the positions of continuously moving objects. In *SIGMOD*, 2000.
[49]. P. Samarati, and L. Sweeney. Protecting privacy when disclosing information: *k*-anonymity and its enforcement through generalization and suppression. Technical Report SRI-CSL-98-04, SRI Computer Science Laboratory, 1998.
[50]. P. Samarati. Protecting respondents' identities in microdata release. *IEEE Trans. Knowl. Data Eng.*, 13(6), 2001.
[51]. L. Sweeney. *K*-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzz.*, 10(5), 2002.
[52]. Y. Tao, D. Papadias, and Q. Shen. Continuous nearest neighbor search. In *VLDB*, 2002.
[53]. J. Traub, Y. Yemini, and H. Woznaikowski. The statistical security of a statistical database.. *ACM TODS*, 9(4), 1984.
[54]. [54] USA Today. Authorities: Gps systems used to stalk woman. http://www.usatoday.com /tech/news/2002-12-30-gps-stalker_x.htm.
[55]. K. Wang, and B. Fung. Anonymizing sequential releases. In *KDD*, 2006.
[56]. K. Wang, P. Yu, and S. Chakraborty. Bottom-up generalization: a data mining solution to privacy protection". In *ICDM*, 2004.
[57]. T. Wang, and L. Liu. Butterfly: protecting output privacy in stream mining. In *ICDE*, 2008.
[58]. T. Wang, and L. Liu. Location privacy protection for road network basedmobile computing system. CS Technical Report, Georgia Tech, 2008.
[59]. R. Wong, J. Li, A. Fu, and K. Wang. (alpha, *k*)-anonymity: an enhanced *k*-anonymity model for privacy preserving data publishing. In *SIGKDD*, 2006.
[60]. R. Wong, A. Fu, K. Wang, and J. Pei. Minimality attack in privacy preserving data publishing. In *VLDB*, 2007.

[61]. X. Xiao, and Y. Tao. Anatomy: simple and effective privacy preservation. In *VLDB*, 2006.
[62]. X. Xiao, and Y. Tao. *m*-invariance: towards privacy preserving re-publication of dynamic datasets. In *SIGMOD*, 2007.
[63]. J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A. Fu. Utility based anonymization using local recording. In *KDD*, 2006.
[64]. M. Yiu, C. Jensen, X. Huang, and H. Lu. SpaceTwist: managing the trade-offs among location privacy, query performance, and query accuracy in mobile services. In *ICDE*, 2008.
[65]. Q. Zhang, N. Koudas, D. Srivastava, and T. Yu. Aggregate query answering on anonymized tables. In *ICDE*, 2007.