Chapter 9
Location Privacy

9.1 Introduction

Location-based services exploit location information to provide a variety of fancy applications. For instance, E-911 in the USA (correspondingly E-112 in Europe) tries to help the caller of an emergency call as soon as possible by locating him or her through GPS. Besides, applications that notify users the nearby places of interest (such as the nearest hospital, restaurant, and store.) can facilitate daily life. While amount of attractive quality of life enhancing applications are presented by location-based services, new threats are also brought in. Among these threats, perhaps the most important one is the intrusion of location privacy.

To clarify the meaning of the term "location privacy," we use Alan Westin's commonly quoted definitions of information privacy [112]. Location privacy can be defined as a special type of information privacy:

> *Location privacy is the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent location information about them is communicated to others.*

In other words, location privacy mainly concerns user's ability of controlling location information.

Most people have not paid enough attention on their location privacy. They tend to underestimate the harm of location leaking for possible two reasons. First, they do not fully understand the negative consequence of privacy divulging. Along with the proliferation of pervasive and mobile computing, however, location disclosure not only leaks location information, but also leads to the implications of personal information. For example, by tracking the history of one's movement, it is possible for attackers to reveal some personal information, such as who he is, where he usually goes shopping, what company he is working for, and how often he does exercise.

Second, protecting location privacy usually to some extent sacrifices the quality of services. Therefore, when we study location privacy, there is a key question throughout: How much protection on location privacy is effective and acceptable? Although the answer to this question is actually application and user dependent, the public has a common belief that a good service design should take both the quality and the privacy concerns into account.

Y. Liu and Z. Yang, *Location, Localization, and Localizability: Location-awareness Technology for Wireless Networks*, DOI 10.1007/978-1-4419-7371-9_9,

In Sect. 9.2, we discuss the threats on location privacy. Section 9.3 discusses the four classes of privacy protection strategies. In Sect. 9.4, we concentrate on location anonymity, which involves most of recent research works. Section 9.5 provides several directions of ongoing research on location privacy.

9.2 Threats

To illustrate the threats of location privacy, we focus on two questions: How can adversary obtain the location information of others? What if location information is leaked?

9.2.1 How Can the Adversary Obtain Location Information of Others?

Most users only desire to release their location information to certain service providers. A straightforward question is how a third-party adversary can get access to the location information.

There are several possible ways. For example, an adversary can intercept the communications between the user and the service provider, or crack data from the service provider directly, if the service provider does not protect user data well. What is worse, some service providers might be camouflaged and malicious, so they intentionally collect user information and sell them to hostile parties.

9.2.2 What Is the Negative Consequence of a Location Leak?

The second question is what the consequence of location leaking is. A direct negative effect is that personal well-being and safety are influenced. The leakage of location information not only yields the uncomfortable creepiness of being watched, but also leads to physical harms to individuals.

Another negative effect is the unwanted revelation of user activities. For most people, it might be embarrassing to be seen at places such as abortion clinic and AIDS clinic. It might also be unwilling for a staff if the proximity to a business competitor is revealed to the boss. Generally speaking, location information consists of three explicit or implicit factors: time, location, and personal identity. Therefore, a large amount of personal information, such as political affiliations, religious beliefs, lifestyles, and medical status, can be inferred by gathering location information.

Here we use the term “gathering” because rather than the presence at certain locations, the pattern of movement can be acquired by tracking an

individual's location for a period and help the adversary to read the meaning of the individual's routes.

The following example shows how one's home address can be inferred from one's pattern of everyday movement. Assume his location is recorded by an attacker every 5 min. Then all these location information can be segmented into discrete trips. Observing these trips long enough (say, at least 1 km long), the adversary can gather many clues in order to infer the location of his home. First, if the last trip always ends in a same place everyday, this place has a high probability of being his home. Second, if the subject spends much more time in a same place than in other places, then this place may be his home. Third, considering the place of his stay between 6 p.m. and 8 a.m., if there is a place that occupies a high percentage, that place is probably his home.

9.3 Protection Strategies

In [113], existing location privacy protection strategies fall into four categories: regulatory, privacy policies, anonymity, and obfuscation. Regulatory strategies try to govern the use of personal information by legislation. Privacy policies provide flexible privacy protection in order to meet the different requirement of users. Anonymity approaches aim at disassociating the location information from the real identity of a user. Obfuscation protects privacy by degrading the resolution of location information provided by service providers. The former two strategies mainly aim at preventing the attacker from obtaining the location information of others through political efforts of mechanism designs. The latter two, on the contrary, aim to preserve location privacy technically.

9.3.1 Regulatory Approaches

The most fundamental privacy protection strategy is to govern fair use of personal location information by developing related regulations. Existing regulations are quite different from one another since they are drafted by different organizations and nations based on their own requirements. These regulations can be mainly summarized by the five core principles proposed in Fair Information Practice Principles [114]:

1. *Notice/awareness*. Individuals must be aware of the identification of the entity collecting the data and the purpose of data collecting.
2. *Choice/consent*. Individuals must be able to decide how any personal information collected from them may be used.
3. *Access/participation*. Individuals must be able to access data about themselves and to contest the data accuracy and completeness.

4. *Integrity/security*. Collectors must ensure the accuracy of personal data and protect these data from disclosure.
5. *Enforcement/redress*. Collectors must be accountable for any violation of above principles.

Although legislation provides a powerful way of protecting privacy, it also brings about troubles. Privacy laws vary from nation to nation, so that location-based services abide by the laws of a particular nation might violate privacy rules of another nation. This issue makes it difficult for service providers to extend their business in different nations without changing the services.

Another issue is that regulations only ensure the mechanisms of enforcement and accountability when a violation of location privacy is detected. They cannot prevent invasions of privacy afore. Moreover, regulation legislating always lags behind the development of new technologies.

9.3.2 Privacy Policies

Regulation provides global or group-based protection of privacy, while it lacks flexibility. Different individuals may have different concerns about their location privacy. A super star might be very sensitive about the disclosure of his location, but for ordinary people, most of their location information is less interesting to the public.

Privacy policies aim at providing flexible privacy protection by adopting individual requirements. They are trust-based mechanisms. The term "trust based" means that the system must be trusted by the users. Policy-based approaches cannot provide privacy if the system betrays.

PIDF (presence information data format) [115] is a location privacy policy scheme adopted by the IETF (Internet engineering task force). A user specifies his acceptable usage of location information, such as whether retransmission of the data is allowed, at what time the data expire and should be discarded, etc. Personal preference of privacy policy is then attached to the location information to be submitted. Both location information and privacy policy are encapsulated into a location object and digitally signed (in order to prevent separating the location information from privacy policy) before sending out.

P3P (privacy preferences project) [116] is a Web-based privacy protection mechanism developed by W3C (World-Wide Web consortium). Unlike PIDF, P3P focuses on the service providers rather than the users. Service providers can publish their data practices, including the purpose of data collecting, how long would these data be held, and whom might these data be shared with. And it leaves for the users proscribing a particular service to decide whether its data practices violate their own privacy requirement. P3P does not explicitly address location privacy issues, while its mechanism can be extended for location awareness context.

There are other policy-based mechanisms for location privacy protection, such as PDRM (personal digital rights management) [117] and IBM's EPAL (enterprise

privacy authorization language) [118]. All these policy-based initiatives only provide a partial solution to privacy. The practicality of these policies under location-aware environment, which involves frequent and dynamic location information, is not yet proved. Unlike the regulatory approaches, privacy policies provide no enforcement, but rely on economic, social, and regulatory pressures.

9.3.3 Anonymity

As mentioned in Sect. 9.2, adversary inference mainly counts on the three factors: time, location, and personal identity. A direct thought is that if we can hide the personal identity, i.e., make the released location information anonymous, we can avoid being affected by the disclosure of location information, because even some inferences are successfully obtained, an attacker still has no idea about the identity of the subject.

Anonymity is a technical countermeasure that dissociates information about an individual from his identity. Its goal is to use location-based services without revealing user identity. Unlike the trust-based mechanisms, anonymity-based approaches always suspect every service provider. A service intermediary is introduced for anonymity-based scheme, which is trusted and might help users hiding their identities. In such a scheme, users do not communicate with service providers directly. Instead, they communicate with the intermediary first, and then the intermediary would fetch data from the service providers and send the data back to the users. The design of a service intermediary is important for both service providers and users.

Notice that, it is clear that some location-based services, such as "when I am at home, let my family know where I am" cannot work without the identity of the user. The anonymity-based approaches mainly focus on other types of services that can work in the absence of real identities, such as "when I walk into a restaurant, show me the menu." In Sect. 9.4, we discuss anonymity-based approaches in detail.

There are drawbacks for anonymity-based approaches. First, anonymity-based approaches usually rely on the design and deployment of the intermediary. Second, anonymity barriers authentication and personalization, and thus prevents some customized applications.

9.3.4 Obfuscation

Obfuscation deliberately degrades the resolution of location information in order to protect privacy while allowing user identities to be revealed. There are three types of imperfection in the literature that can be introduced into the location information: *inaccuracy*, *imprecision*, and *vagueness*. In location awareness context, inaccuracy means telling a location differs from the real location; imprecision means telling a

region including the real location instead of the real location; and vagueness means involving linguistic terms like "near" or "far from" in the conveyed location. Many researches on obfuscation concentrate on the use of imprecision.

Some anonymity-based approaches also use imprecision. The difference of anonymity and obfuscation is that anonymity aims to make an individual indiscernible to a number of other individuals, while obfuscation aims to make the location of an individual indiscernible to a number of other locations.

Commonly used in location-based services, proximity query typically asks about the life facilities close to a user's location, e.g., "where is the nearest restaurant?". In [119], an algorithmic approach is proposed to obfuscating proximity queries. An individual reports a set O of locations instead of his real location. The service provider then tries to find the position of interest for each location in O. If all locations in O have the same result, the provider can return this result to the user. Otherwise, it asks whether the user agrees to refine his location. If the user agrees to do so, the algorithm reiterates. If the user refuses, the provider returns the best estimate approximation according to the coarse-grained information provided by the user.

Obfuscation does not rely on any intermediary, and users can communicate with service providers directly. As a result, the architecture is lightweight and distributed. Also, it enables the applications that require authentication or personalization, which might be blocked for the anonymity-based approaches. Even though researchers claim that most location-based services can work with imprecise location, the loss of quality of service is left open for study.

9.4 Anonymity-Based Approaches

Releasing location information anonymously (i.e., using a pseudonym instead of an actual identity) can prevent attackers from linking the location information to an individual. However, hiding the name is not enough. It is possible for attackers to reidentify an individual from the location information of a pseudonym. For example, certain regions of a space, such as desk location in an office, can be closely associated with certain identities, and hence can be used to deanonymize the users. Therefore, by tracking a pseudonym and gathering related clues (for example, where the pseudonym spends most of its time and whether the pseudonym spends more time at a certain desk than anyone else), the adversary can easily find out the user identity, although the pseudonym is used.

To relieve the threat of linking attack, anonymity-based approaches need to make a pseudonym indiscernible with a number of other pseudonyms. To achieve this, most approaches introduce a trusted intermediary to coordinate users and to provide a large enough anonymity set. In this section we discuss four anonymity-based countermeasures in detail, and at the end of this section, we present a brief comparison of these works.

9.4.1 k-Anonymity

The concept of k-anonymity is originally proposed in [120] in order to provide protection for linking attack. A released data set is considered to be k-anonymous if every element in it is indistinct with at least k-1 other elements. In other words, every combination of values of attributes can be indistinctly matched to at least k elements.

Gruteser and Grunwald [121] extend the k-anonymous concept to the scope of location information. A subject is considered as k-anonymous if and only if the location of the subject is indistinguishable from the locations of at least k-1 other subjects. If a k-anonymous individual reports his location, attackers cannot tell which of the k subjects actually locates at the reported location.

Now the problem turns to be how to achieve k-anonymity. The location information can be represented by a tuple of three intervals ($[x_1, x_2]$, $[y_1, y_2]$, $[t_1, t_2]$). $[x_1, x_2]$ and $[y_1, y_2]$ describe a region in two-dimensional space where the subject is located at a time span $[t_1, t_2]$. Basically, a set of tuples that dissatisfies the k-anonymity requirement can be converted to a k-anonymous set by generalization. Generalization is similar to the degrading techniques used for obfuscation, which decreases the precision of the revealed information. For example, two distinct intervals [12, 23] and [24, 37] can be generalized to [12, 37] and becomes indistinguishable. Since the location information contains both spatial and temporal information, generalization can be applied spatially and/or temporally.

The basic idea of spatial cloaking is to choose a sufficiently large area so that enough number of subjects inhabit this region. Obviously, a larger region means less precision and lower quality of services. Therefore, the challenge is to report spatial information as precise as possible while satisfying the k-anonymity constraint. The algorithm in [121] uses the quadtree to achieve this objective. It keeps dividing an area into quadrants of equal size, until further dividing would create a quadrant with less than k subjects, as illustrated in Fig. 9.1. Each subject reports its host quadrant as its spatial information.

Temporal cloaking, the orthogonal approach to spatial cloaking, tends to reveal more precise spatial coordinates while reducing the precision in time dimension. The idea is to delay a service request containing location information until k individuals have visited the same area of the requestor. Temporal cloaking can be combined with spatial cloaking to make a balance between spatial and temporal resolution.

Certainly, a trusted intermediary is necessary for this approach, since it requires a global knowledge of the distribution of users. If the k-anonymity constraint is satisfied, an attacker only has a probability of $1/k$ at the most to figure out the identity of a user.

Nonetheless, Bettini et al. [122] point out that simple k-anonymity might be insufficient since an attacker can track the historical location information of a pseudonymous user and analyze the movement pattern (e.g., the commuter route of a pseudonym). To mitigate this type of attack, they introduce the notion of

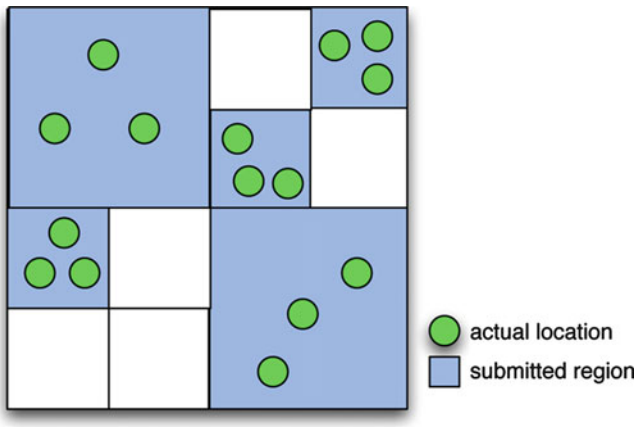


Fig. 9.1 An example of achieving three-anonymity by quadrants dividing

"historical *k*-anonymity," which concerns that the personal history of location received by a service provider cannot be distinguished from *k*-1 other sets of personal history of location received by the same service provider.

Generally speaking, spatial and temporal cloaking provides a limited protection for location privacy. Tracking the path of a user can break the protection easily. Also, this approach sacrifices spatial and temporal resolution of location information as well as the quality of service.

9.4.2 Mix Zone

The method of "mix zones" [123] introduces a trusted middleware. A user registers a list of location-based applications that he is interested in with the middleware. An application receives event callbacks about the user from the middleware when the user enters or exits the areas related to this application. The middleware updates user location periodically and issues callbacks to applications when necessary. When communicating with service providers, the middleware uses pseudonyms instead of identities so as to protect privacy.

Mix zone is designed to solve two main drawbacks of anonymity-based approaches. First, it is obvious that the longer a user keeps using a same pseudonym, the weaker the anonymity becomes. The anonymity would be invalidated if the identity of a subject one gets revealed at any location on its path. For example, if a user divulges the identity and location (probably due to the imprudence) in some messages caught previously, then the user appoints a new anonymous message to the middleware. Unfortunately, this measurement does not work. The attacker can link the later message with the previous ones.

Second, the history of location information provides clues that can help attackers figure out the identity of a subject. Suppose an attacker knows that a pseudonym's

home and office are in regions A and B, respectively. But the attacker fails to figure out the identity of the pseudonym because there are at least k different pseudonyms in each region. If considering the two clues simultaneously, the attacker might be able to reidentify the pseudonym, since the individual satisfying both constraints (home in region A and office in region B) might be unique.

A direct countermeasure is to change user pseudonym frequently. However, it brings out two new problems. First, some applications might not work properly with fast-changing pseudonyms. Second, if the spatial and temporal resolution provided by the middleware is sufficiently high, attackers can still link the old and new pseudonyms.

To solve the two problems, the concept of "mix zone" is proposed. A mix zone for a group of users is defined as a connected spatial region of maximum size in which none of these users have registered any application callback. The areas where some users have registered for callbacks are called application zones. Users keep using same pseudonyms within the same application zone. When users are inside a mix zone, applications would not receive any location information about them. The following measurement makes the user identities "mixed." When a user enters an application zone from a mix zone (or enters a mix zone from an application zone), the user is assigned with a new, unused pseudonym. As a result, when appears in a mix zone, a user cannot be distinguished from others inside the mix zone at the same time. Also, it is difficult to link a user coming out of a mix zone with any user who enters the mix zone previously.

Figure 9.2 shows an example of this procedure. Suppose there are two users who have registered services in airport, bank, and coffee shop. At some time, one user is in the airport and the other is in the coffee shop. Their presence might be aware by all three service providers since the providers can communicate with each other. Afterwards both users have entered the mix zone and have their pseudonyms changed. When one of the two users enters the bank zone, the service providers only see a new pseudonym appears, but they cannot know which previously appeared pseudonym should be linked to this new pseudonym, since it could be either one of the two users.

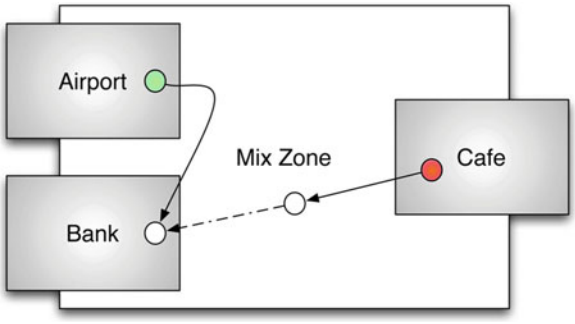


Fig. 9.2 Mix zone example

However, how to divide users into different groups and set the mix zone for these groups is complex. A large mix zone would reduce the security due to the relevance in spatial and temporal coordinates. A user entering a coffee shop in downtown cannot be the one who just appears in the airport one minute ago. A small mix zone increases the difficulty of pseudonym mixing, since it requires the diversity of pseudonyms inside a mix zone.

9.4.3 Using Dummies

Kido et al. propose a way to fool attackers by using dummies [124]. When a user sends position information to a service provider, the report is attached with a set of fake position data which are called "dummies," as illustrated in Fig. 9.3. From the view of the service provider, it looks like there are several different user requests. The provider answers these requests by sending back a message (which contains all the responding to these positions) to the user. The user only selects the necessary data corresponding to his location.

However, if the dummies are generated randomly, observers can easily tell apart the true location and the dummies, because the distance that a subject can move in a fixed time interval is limited. To avoid this, the dummy behavior should be related to the user. Two dummy generalization algorithms are presented in [124]: moving in a neighborhood and moving in a limited neighborhood.

Compared to the *k*-anonymity approaches, using dummies have several advantages. First, it is difficult for attackers to find out the true pattern of movement of an individual. Second, users can report precise location information with high spatial and temporal resolution, so that little quality of services would be lost. This approach has a drawback that it increases the cost of communication. Users need to report additional dummy location information to service providers, and service providers need to return additional service data for the dummies. Only a small fraction of the communications is useful and all dummy-related communications are overheads.

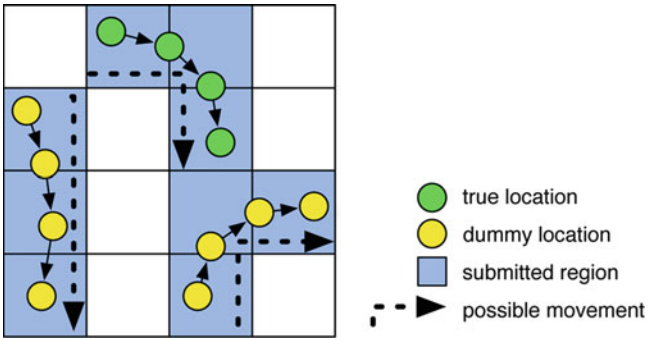


Fig. 9.3 Dummy generation. Attackers cannot determine the true movement

9.4.4 Path Confusion

The *k*-anonymity approaches and the mix zone have a common weakness: they all rely on the density of individuals. If the density is not sufficient, the *k*-anonymity approaches deserve poor quality of services due to imprecise location information, while the mix zone might provide poor anonymity since attackers can easily link pseudonyms by temporal and spatial relevance.

Path confusion is proposed to preserve privacy in GPS traces, which can guarantee a certain level of location privacy even for users in low-density regions [125]. The idea is similar to temporal cloaking but it works on paths. The intermediary would delay releasing the user's location, until it finds out the user's path intersects with another user's. Then the intermediary reveals all locations on the two paths altogether, as illustrated in Fig. 9.4. Attackers can only see a bundle of locations on the two paths occurring at the same time. The attacker can tell neither which path the target being tracked is on, nor which direction on the path the target is heading for. Therefore, the target being tracked is confused with other individuals. To provide better anonymity, the intermediary can simply wait longer until more paths are intersected.

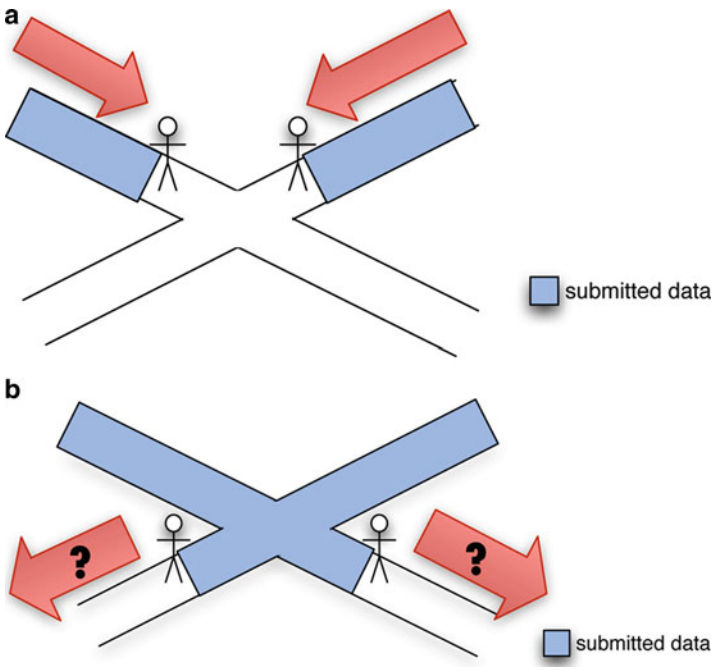
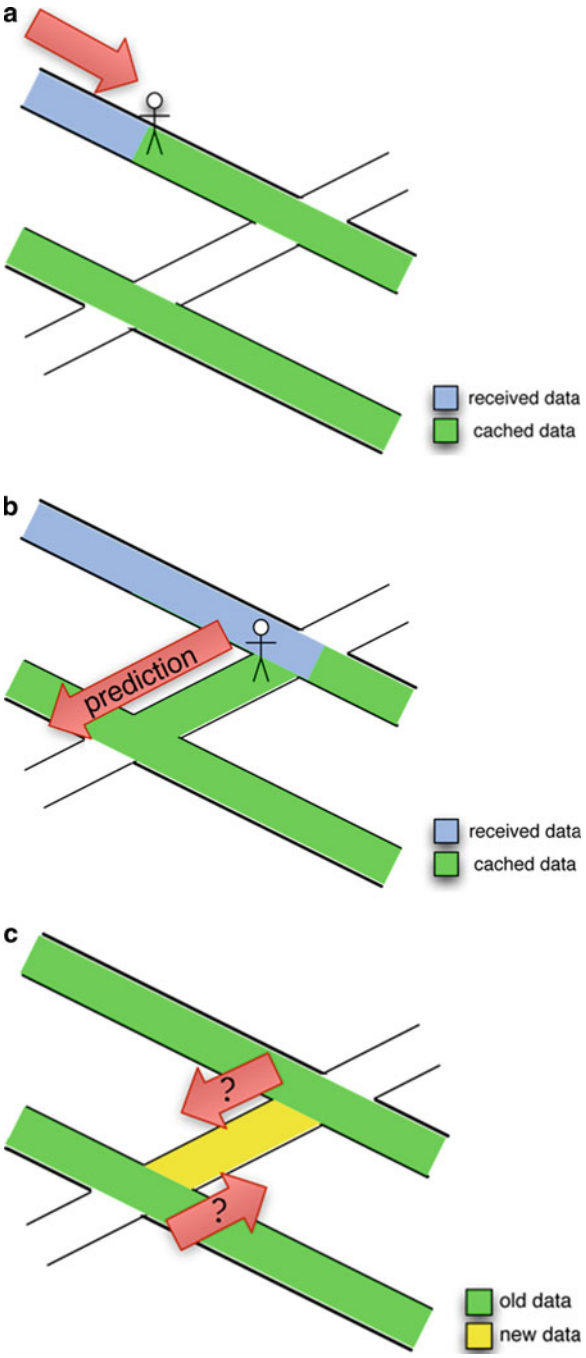


Fig. 9.4 Path confusion. (**a**) At $t = t_0$, an attacker can track the two users according to previously revealed locations. and (**b**) At $t = t_0 + \varepsilon$, since the two users have coincided in space and time, the attacker cannot say whether they turn or go straight


a
received data
cached data
b
prediction
received data
cached data
c
?
?
old data
new data

Releasing precise location information, path confusion can keep the quality of services. The main drawback is that, similar to temporal cloaking, it sacrifices real-time services due to the delay of requests.

Meyerowitz and Choudhury develop cache cloak based on the idea of path confusion [126]. Rather than posteriori analysis of a user's path, cache cloak prefers using mobility prediction to do a prospective form of path confusion. It keeps a spatial cache which contains data for a set of position points. If a user submits a position point that hits the cache, then the intermediary returns the cached data for that location directly, without fetching data from the service provider. If a user submits a position that is not in the cache, which means a cache miss, cache cloak would generate a predicted path for the user. The predicted path is extrapolated until it reaches another path that exists in the cache. (i.e., the predicted path is connected on both ends to other cached paths) The entire predicted path is then submitted to the service provider and all responses for locations along the path are retrieved and cached. Moving along a path, a user gets serviced directly from the cache until deviates from the predicted.

From the attackers' view, each location release contains a bunch of locations on a path. Each newly released path connects two paths released previously, say, path A and path B, as illustrated in Fig. 9.5. There are three possible cases that will trigger a new query: the user on path A turns toward path B; the user on path B turns toward path A; and a new user on the newly released path begins to use the service. Attackers cannot tell apart the three possibilities and accordingly fail to track users.

Cache cloak does not degrade the spatial or temporal resolution as the dummy-based approach. Moreover, a predicted path can be viewed as a dummy (which confuses attackers), and probably this kind of dummies acts more "reasonable" than the dummies generated by the two algorithms, moving in a neighborhood and moving in a limited neighborhood. For the cost of communication, cache cloak does not increase any unnecessary communication between users and the intermediary, although it brings about unnecessary communications between the intermediary and service providers. This overhead can be low if the cache cloak intermediary and service providers are connected by wired networks.

9.4.5 Comparison

Before comparing anonymity-based algorithms, we need to answer the problem that how we can tell if an algorithm is better than any other? The level of location privacy can be reflected by the size of anonymity set, but the definition of anonymity set varies

Fig. 9.5 An example for cache cloak. (**a**) A user is moving along a previously cached path. He retrieves data from the intermediary directly. (**b**) A user deviates his path, which triggers a cache miss. New path is predicted and service data along the predicted path are requested from the service providers by the intermediary, and all the retrieved data are stored in the cache. (**c**) An attacker cannot determine what triggers the new data queries. It could be users turning in from the upper street (path A), or from the lower street (path B)

Table 9.1 Anonymity-based approaches

	Loss of QoS	Antitracking	Cost of comm.	Intermediary
Spatial & temporal cloaking	Degraded	Not capable	No increase	Necessary
Mix zone	Degraded	Capable	No increase	Necessary
Dummy	Not affected	Capable	Increased (wireless)	Not necessary
Cache cloak	Not affected	Capable	Increased (wired)	Necessary

among different approaches. In addition, the size of anonymity is usually a parameter that can be set flexibly if necessary. Besides anonymity, we try to characterize an anonymity approach by the following factors:

1. *Loss of quality of service (QoS)*. Approaches, such as spatial and temporal cloaking, which degrade the resolution of location information would certainly sacrifice the quality of service. Mix zone breaks the continuity of services, which might degrade the quality of services as well.
2. *Antitracking ability*. We have shown that the historical location data, a.k.a. the pattern of movement, would lead to privacy leaks. Approaches like spatial and temporal cloaking cannot curb an attacker from extracting information through tracking, while some approaches like path confusion can deal well with the attacks based on tracking.
3. *Cost of communication*. Approaches like using dummies and cache cloak increase the cost of communication. The cost of communication on wired network is much cheaper than that on wireless network.
4. *Intermediary dependence*. Most approaches require a trusted intermediary. However, the deployment of an intermediary is expensive, and the communication between users and an intermediary needs to be protected from being interrupted; otherwise, all the efforts would be meaningless.

At last, we summarize the anonymity-based approaches in Table 9.1.

9.5 Summary

Although a lot of approaches have been proposed, a number of issues remain open.

Distributed anonymity. Most anonymity-based approaches require a trusted intermediary, but what if an intermediary cannot be trusted? Or the communication between users and an intermediary is not secure? Dummy-based approach gives a solution without an intermediary, but it increases the cost of communication. Can users cooperate without an intermediary? These questions are still unanswered.

Other types of attacks. Anonymity-based approaches only solve linking attack problem, but are vulnerable for other types of attacks, such as homogeneity attack. Taking k-anonymity as an example, the lack of diversity inside the anonymity set might leak user privacy. For instance, if a location region is inside

an abortion clinic or AIDS clinic, in spite of several indistinguishable subjects inside the region, an attacker can still infer the activity of a victim as long as the victim is among these subjects. How to protect privacy from other types of attacks? These problems are worth researching.

Hybrid schemes. No approach can solve the privacy problem perfectly and a combination of privacy strategies might be more effective. How to make different strategies working together is need to be studied.

Pervasive and mobile computing changes the scale of the privacy issue. Future privacy protection approaches are expected to deal with a large number of users, a flood of service requests, and highly frequent data updates. In summary, the privacy issue must be fully addressed before the real proliferation of pervasive computing and the Internet of things (IoT).