Location Privacy in Pervasive Computing

As location-aware applications begin to track our movements in the name of convenience, how can we protect our privacy? This article introduces the mix zone—a new construction inspired by anonymous communication techniques—together with metrics for assessing user anonymity.

Alastair R. Beresford and Frank Stajano
University of Cambridge

Many countries recognize privacy as a right and have attempted to codify it in law. The first known piece of privacy legislation was England's 1361 *Justices of the Peace Act,* which legislated for the arrest of eavesdroppers and stalkers. The Fourth Amendment to the US Constitution proclaims citizens' right to privacy, and in 1890 US Supreme Court Justice Louis Brandeis stated that "the right to be left alone" is one of the fundamental rights of a democracy.[1] The 1948 *Universal Declaration of Human Rights*[2] declares that everyone has a right to privacy at home, with family, and in correspondence. Other pieces of more recent legislation follow this principle. Although many people clearly consider their privacy a fundamental right, comparatively few can give a precise definition of the term. The Global Internet Liberty Campaign[3] has produced an extensive report that discusses personal privacy at length and identifies four broad categories: information privacy, bodily privacy, privacy of communications, and territorial privacy.

This article concentrates on *location privacy,* a particular type of information privacy that we define as *the ability to prevent other parties from learning one's current or past location*. Until recently, the very concept of location privacy was unknown: people did not usually have access to reliable and timely information about the exact location of others, and therefore most people could see no privacy implications in revealing their location, except in special circumstances. With pervasive computing, though, the scale of the problem changes completely. You probably do not care if someone finds out where you were yesterday at 4:30 p.m., but if this someone could inspect the history of all your past movements, recorded every second with submeter accuracy, you might start to see things differently. A change of scale of several orders of magnitude is often qualitative as well as quantitative—a recurring problem in pervasive computing.[4]

We shall focus on the privacy aspects of using location information in pervasive computing applications. When location systems track users automatically on an ongoing basis, they generate an enormous amount of potentially sensitive information. Privacy of location information is about controlling access to this information. We do not necessarily want to stop all access—because some applications can use this information to provide useful services—but we want to be in control.

Some goals are clearly mutually exclusive and cannot be simultaneously satisfied: for example, wanting to keep our position secret and yet wanting colleagues to be able to locate us. Despite this, there is still a spectrum of useful combinations to be explored.

Our approach to this tension is a privacy-protecting framework based on frequently changing pseudonyms so users avoid being identified by the locations they visit. We further develop this framework by introducing the concept of *mix zones* and showing

Published by the IEEE CS and IEEE Communications Society ■ 1536-1268/03/$17.00 © 2003 IEEE

Related Work: Location Awareness and Privacy

The first wide-scale outdoor location system, GPS,[1] lets users calculate their own position, but the flow of information is unidirectional; because there is no back-channel from the GPS receiver to the satellites, the system cannot determine the user's computed location, and does not even know whether anyone is accessing the service. At the level of raw sensing, GPS implicitly and automatically gives its users location privacy.

In contrast, the first indoor location system, the Active Badge,[2] detects the location of each user and broadcasts the information to everyone in the building. The system as originally deployed assumes anyone in the building is trustworthy; it therefore provides no mechanisms to limit the dissemination of individuals' location information. Ian W. Jackson[3] modified the Active Badge system to address this issue. In his version, a badge does not reveal its identity to the sensor detecting its position but only to a trusted personal computer at the network edge. The system uses encrypted and anonymized communication, so observation of the traffic does not reveal which computer a given badge trusts. The badge's owner can then use traditional access control methods to allow or disallow other entities to query the badge's location.

More recent location systems, such as Spirit[4] and QoSDream,[5] have provided applications with a middleware event model through which entities entering or exiting a predefined region of space generate events. Applications register their interest in a particular set of locations and locatables and receive callbacks when the corresponding events occur. Current location-aware middleware provides open access to all location events, but it would be possible to augment this architecture to let users control the dissemination of their own location information.

So far, few location systems have considered privacy as an initial design criterion. The Cricket location system[6] is a notable exception: location data is delivered to a personal digital assistant under the sole control of the user.

Commercial wide-area location-based services will initially appear in mobile cellular systems such as GSM. BTexact's Erica system[7] delivers sensitive customer information to third-party applications. Erica provides an API for third-party software to access customer billing information, micropayment systems, customer preferences, and location information. In this context, individual privacy becomes much more of a concern. Users of location-aware applications in this scenario could potentially have all their daily movements traced.

In a work-in-progress Internet draft that appeared after we submitted this article for publication, Jorge R. Cuellar and colleagues also explore location privacy in the context of mobile cellular systems.[8] As we do, they suggest using pseudonyms to protect location privacy. Unlike us, however, they focus on policies and rules—they do not consider attacks that might break the unlinkability that pseudonyms offer.

REFERENCES

1. I. Getting, "The Global Positioning System," *IEEE Spectrum,* vol. 30, no. 12, Dec. 1993, pp. 36–47.
2. R. Want et al., "The Active Badge Location System," *ACM Trans. Information Systems,* vol. 10, no. 1, Jan. 1992, pp. 91–102.
3. I.W. Jackson, "Anonymous Addresses and Confidentiality of Location," *Information Hiding: First Int'l Workshop,* R. Anderson, ed., LNCS, vol. 1174, Springer-Verlag, Berlin, 1996, pp. 115–120.
4. N. Adly, P. Steggles, and A. Harter, "SPIRIT: A Resource Database for Mobile Users," *Proc. ACM CHI'97 Workshop on Ubiquitous Computing,* ACM Press, New York, 1997.
5. H. Naguib, G. Coulouris, and S. Mitchell, "Middleware Support for Context-Aware Multimedia Applications," *Proc. 3rd Int'l Conf. Distributed Applications and Interoperable Systems,* Kluwer Academic Publishers, Norwell, Mass., 2001, pp. 9–22.
6. N.B. Priyantha, A. Chakraborty, and H. Balakrishnan, "The Cricket Location-Support System," *Proc. 6th Int'l Conf. Mobile Computing and Networking* (Mobicom 2000), ACM Press, New York, 2000, pp. 32–43.
7. P. Smyth, *Project Erica,* 2002, www.btexact.com/erica.
8. J.R. Cuellar, J.B. Morris, and D.K. Mulligan, *Geopriv Requirements,* Internet draft, Nov. 2002, www.ietf.org/internet-drafts/draft-ietf-geopriv-reqs-01.txt.

how to map the problem of location privacy onto that of anonymous communication. This gives us access to a growing body of theoretical tools from the information-hiding community. In this context, we describe two metrics that we have developed for measuring location privacy, one based on *anonymity sets* and the other based on *entropy*. Finally, we move from theory to practice by applying our methods to a corpus of more than three million location sample points obtained from the Active Bat installation at AT&T Labs Cambridge.[5]

Problem, threat model, and application framework

In the pervasive computing scenario, location-based applications track people's movements so they can offer various useful services. Users who do not want such services can trivially maintain location privacy by refusing to be tracked—assuming they have the choice. This has always been the case for our Active Badge (see the "Related Work" sidebar) and Active Bat systems but might not be true for, say, a nationwide network of face-recognizing CCTV cameras—an Orwellian dystopia now dangerously close to reality. The more challenging problem we explore in this article is to develop techniques that let users benefit from location-based applications while at the same time retaining their location privacy.

To protect the privacy of our location information while taking advantage of location-aware services, we wish to hide our true identity from the applications receiving our location; at a very high level, this can be taken as a statement of our security policy.

Users of location-based services will not,

in general, want information required by one application to be revealed to another. For example, patients at an AIDS testing clinic might not want their movements (or even evidence of a visit) revealed to the location-aware applications in their workplace or bank. We can achieve this compartmentalization by letting location services register for location callbacks only at the same site where the service is provided. But what happens if the clinic and workplace talk about us behind our back? Because we do not trust these applications and have no control over them, we must assume they might collude against us to discover the information we aim to hide. We therefore regard all applications as one global hostile observer.

Some location-based services, such as "when I am inside the office building, let my colleagues find out where I am," cannot work without the user's identity. Others, such as "when I walk past a coffee shop, alert me with the price of coffee," can operate completely anonymously. Between those extremes are applications that cannot be accessed anonymously—but do not require the user's true identity either. An example of this third category might be "when I walk past a computer screen, let me teleport my desktop to it." Here, the application must know whose desktop to teleport, but it could conceivably do this using an internal pseudonym rather than the user's actual name.

Clearly, we cannot use applications requiring our true identity without violating our security policy. Therefore, in this article, we concentrate on the class of location-aware applications that accept pseudonyms, and our aim will be the anonymization of location information.

In our threat model, the individual does not trust the location-aware application but does trust the raw location system (the sensing infrastructure that can position locatables). This trust is certainly appropriate for systems such as GPS and Cricket, in which only the user can compute his or her own location. Its applicability to other systems, such as the Active Bat, depends on the trust relationships between the user and the entity providing the location service.

We assume a system with the typical architecture of modern location-based services, which are based on a shared event-driven middleware system such as Spirit or QoSDream (see the "Related Work" sidebar). In our model, the middleware, like the sensing infrastructure, is trusted and might help users hide their identity. Users register interest in particular applications with the middleware; applications receive event callbacks from the middleware when the user enters, moves within, or exits certain application-defined areas. The middleware evaluates the user's location at regular periodic intervals, called *update periods*, to determine whether any events have occurred, and issues callbacks to applications when appropriate.

Users cannot communicate with applications directly—otherwise they would reveal their identity straight away. We therefore require an anonymizing proxy for all communication between users and applications. The proxy lets applications receive and reply to anonymous (or, more correctly, *pseudonymous*) messages from the users. The middleware system is ideally placed to perform this function, passing user input and output between the application and the user. For example, to benefit from the application-level service of "when I walk past a coffee shop, alert me with the price of coffee," the user requests the service, perhaps from the coffee shops' trade-association Web site, but using her location system middleware as the intermediary so as not to reveal her identity. The coffee shop application registers with the user's location system and requests event callbacks for positive containment in the areas in front of each coffee shop. When the user steps into a coffee shop event callback area, the coffee shop application receives a callback from the user's location system on the next location update period. The coffee shop service then sends the current price of coffee as an event message to the registered user via the middleware without having to know the user's real identity or address.

Using a long-term pseudonym for each user does not provide much privacy—even if the same user gives out different pseudonyms to different applications to avoid collusion. This is because certain regions of space, such as user desk locations in an office, act as "homes" and, as such, are strongly associated with certain identities. Applications could identify users by following the "footsteps" of a pseudonym to or from such a "home" area. At an earlier stage in this research, we tried this kind of attack on real location data from the Active Bat and found we could correctly deanonymize all users by correlating two simple checks: First, where does any given pseudonym spend most of its time? Second, who spends more time than anyone else at any given desk? Therefore, long-term pseudonyms cannot offer sufficient protection of location privacy.

Our countermeasure is to have users change pseudonyms frequently, even while they are being tracked: users adopt a series of new, unused pseudonyms for each application with which they interact. In this scenario, the purpose of using pseudonyms is not to establish and preserve reputation (if we could work completely anonymously instead, we probably would) but to provide a return address. So the problem, in this context, is not whether changing pseudonyms causes a loss of reputation but more basically whether the application will still work when the user changes under its feet. Our preliminary analysis convinced us that many existing applications could be made to work within this framework by judicious use of anonymizing proxies. However, we decided to postpone a full investigation of this issue (see the "Directions for Further Research" sidebar) and concentrate instead on whether this approach, assuming it did not break applications, could improve location privacy. If it could not, adapting the applications would serve no purpose.

Users can therefore change pseudonyms while applications track them; but then, if the system's spatial and temporal resolution were sufficiently high (as in the Active Bat), applications could easily link the old and new pseudonyms, defeating the purpose of the change. To address this point, we introduce the mix zone.

Mix zones

Most theoretical models of anonymity

Directions for Further Research

During our research on location privacy, several interesting open problems emerged.

Managing application use of pseudonyms

A user who does not wish to be tracked by an application will want to use different pseudonyms on each visit. Many applications can offer better services if they retain per-user state, such as personal preferences; but if a set of preferences were accessed by several different pseudonyms, the application would easily guess that these pseudonyms map to the same user. We therefore need to ensure that the user state for each pseudonym looks, to the application, different from that of any other pseudonym.

There are two main difficulties. First, the state for a given user (common to all that user's pseudonyms) must be stored elsewhere and then supplied to the application in an anonymized fashion. Second, the application must not be able to determine that two sets of preferences map to the same user, so we might have to add small, random variations. However, insignificant variations might be recognizable to a hostile observer, whereas significant ones could adversely affect semantics and therefore functionality.

Reacting to insufficient anonymity

You can use the methods described in this article to compute a quantitative measure of anonymity. How should a user react to a warning that measured anonymity is falling below a certain threshold? The simple-minded solution would be for the middleware to stop divulging location information, but this causes applications to become unresponsive.

An alternative might be to reduce the size or number of the application zones in which a user has registered. This is hard to automate: either users will be bombarded with warnings, or they will believe they are still receiving service when they are not. In addition, removing an application zone does not instantaneously increase user anonymity; it just increases the size or number of available mix zones. The user must still visit them and mix with others before anonymity increases.

Reconciling privacy with application functionality is still, in general, an open problem.

Improving the models

It would be interesting to define anonymity sets and entropy measurements from the perspective of what a hostile observer can see and deduce, instead of counting the users in the mix zone. The formalization, which we have attempted, is rather more complicated than what is presented in this article, but it might turn out to provide a more accurate location privacy metric.

Dummy users

We could increase location privacy by introducing dummy users, similar to the way cover traffic is used in mix networks, but there are side effects when we apply this technique to mix zones. Serving dummy users, like processing dummy messages, is a waste of resources. Although the overhead might be acceptable in the realm of bits, the cost might be too high with real-world services such as those found in a pervasive computing environment. Dummy users might have to control physical objects—opening and closing doors, for example—or purchase services with electronic cash. Furthermore, realistic dummy user movements are much more difficult to construct than dummy messages.

Granularity

The effectiveness of mixing depends not only on the user population and on the mix zone's geometry, but also on the sensing system's update rate and spatial resolution. At very low spatial resolutions, any "location" will be a region as opposed to a point and so might be seen as something similar to a mix zone. What are the requirements for various location-based applications? How does their variety affect the design of the privacy protection system? Several location-based applications would not work at all if they got updates only on a scale of hours and kilometers, as opposed to seconds and meters.

Scalability

Our results are based on experimental data from our indoor location system, covering a relatively small area and user population. It would be interesting to apply the same techniques to a wider area and larger population—for example, on the scale of a city—where we expect to find better anonymity. We are currently trying to acquire access to location data from cellular telephony operators to conduct further experiments.

and pseudonymity originate from the work of David Chaum, whose pioneering contributions include two constructions for anonymous communications (the *mix network*[6] and the *dining cryptographers algorithm*[7]) and the notion of *anonymity sets*.[7]

In this article, we introduce a new entity for location systems, the mix zone, which is analogous to a mix node in communication systems. Using the mix zone to model our spatiotemporal problem lets us adopt useful techniques from the anonymous communications field.

Mix networks and mix nodes

A mix network is a store-and-forward network that offers anonymous communication facilities. The network contains normal message-routing nodes alongside special mix nodes. Even hostile observers who can monitor all the links in the network cannot trace a message from its source to its destination without the collusion of the mix nodes.

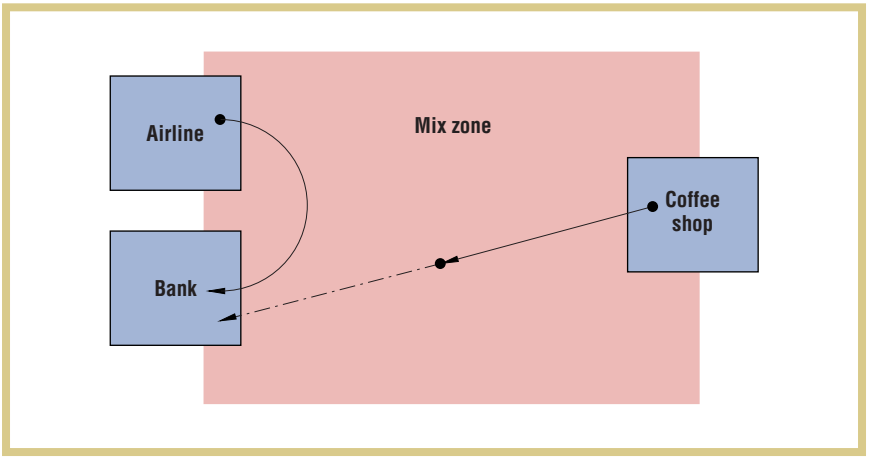


Figure 1. A sample mix zone arrangement with three application zones. The airline agency (A) is much closer to the bank (B) than the coffee shop (C). Users leaving A and C at the same time might be distinguishable on arrival at B.

In its simplest form, a mix node collects *n* equal-length packets as input and reorders them by some metric (for example, lexicographically or randomly) before forwarding them, thus providing unlinkability between incoming and outgoing messages. For brevity, we must omit some essential details about layered encryption of packets; interested readers should consult Chaum's original work. The number of distinct senders in the batch provides a measure of the unlinkability between the messages coming in and going out of the mix.

Chaum later abstracted this last observation into the concept of an *anonymity set*. As paraphrased by Andreas Pfitzmann and Marit Köhntopp, whose recent survey generalizes and formalizes the terminology on anonymity and related topics, the anonymity set is "the set of all possible subjects who might cause an action."[8] In anonymous communications, the action is usually that of sending or receiving a message. The larger the anonymity set's size, the greater the anonymity offered. Conversely, when the anonymity set reduces to a singleton—the anonymity set cannot become empty for an action that was actually performed—the subject is completely exposed and loses all anonymity.

Mix zones

We define a mix zone for a group of users as a connected spatial region of maximum size in which none of these users has registered any application callback; for a given group of users there might be several distinct mix zones. In contrast, we define an *application zone* as an area where a user has registered for a callback. The middleware system can define the mix zones a priori or calculate them separately for each group of users as the spatial areas currently not in any application zone.

Because applications do not receive any location information when users are in a mix zone, the identities are "mixed." Assuming users change to a new, unused pseudonym whenever they enter a mix zone, applications that see a user emerging from the mix zone cannot distinguish that user from any other who was in the mix zone at the same time and cannot link people going into the mix zone with those coming out of it.

If a mix zone has a diameter much larger than the distance the user can cover during one location update period, it might not mix users adequately. For example, Figure 1 provides a plan view of a single mix zone with three application zones around the edge: an airline agency (A), a bank (B), and a coffee shop (C). Zone A is much closer to B than C, so if two users leave A and C at the same time and a user reaches B at the next update period, an observer will know the user emerging from the mix zone at B is not the one who entered the mix zone at C. Furthermore, if nobody else was in the mix zone at the time, the user can only be the one from A. If the maximum size of the mix zone exceeds the distance a user covers in one period, mixing will be incomplete. The amount to which the mix zone anonymizes users is therefore smaller than one might believe by looking at the anonymity set size. The two "Results" sections later in this article discuss this issue in more detail.

A quantifiable anonymity measure: The anonymity set

For each mix zone that user *u* visits during time period *t*, we define the anonymity set as the group of people visiting the mix zone during the same time period.

The anonymity set's size is a first measure of the level of location privacy available in the mix zone at that time. For example, a user might decide a total anonymity set size of at least 20 people provides sufficient assurance of the unlinkability of their pseudonyms from one application zone to another. Users might refuse to provide location updates to an application until the mix zone offers a minimum level of anonymity.

Knowing the average size of a mix zone's anonymity set, as opposed to its instantaneous value, is also useful. When a user registers for a new location-aware service, the middleware can calculate from historical data the average anonymity set size of neighboring mix zones and therefore estimate the level of location privacy available. The middleware can present users with this information before they accept the services of a new location-aware application.

There is an additional subtlety here: introducing a new application could result in more users moving to the new application zone. Therefore, a new application's average anonymity set might be an underestimate of the anonymity available because users who have not previously visited the new application zone's geographical area might venture there to gain access to the service.

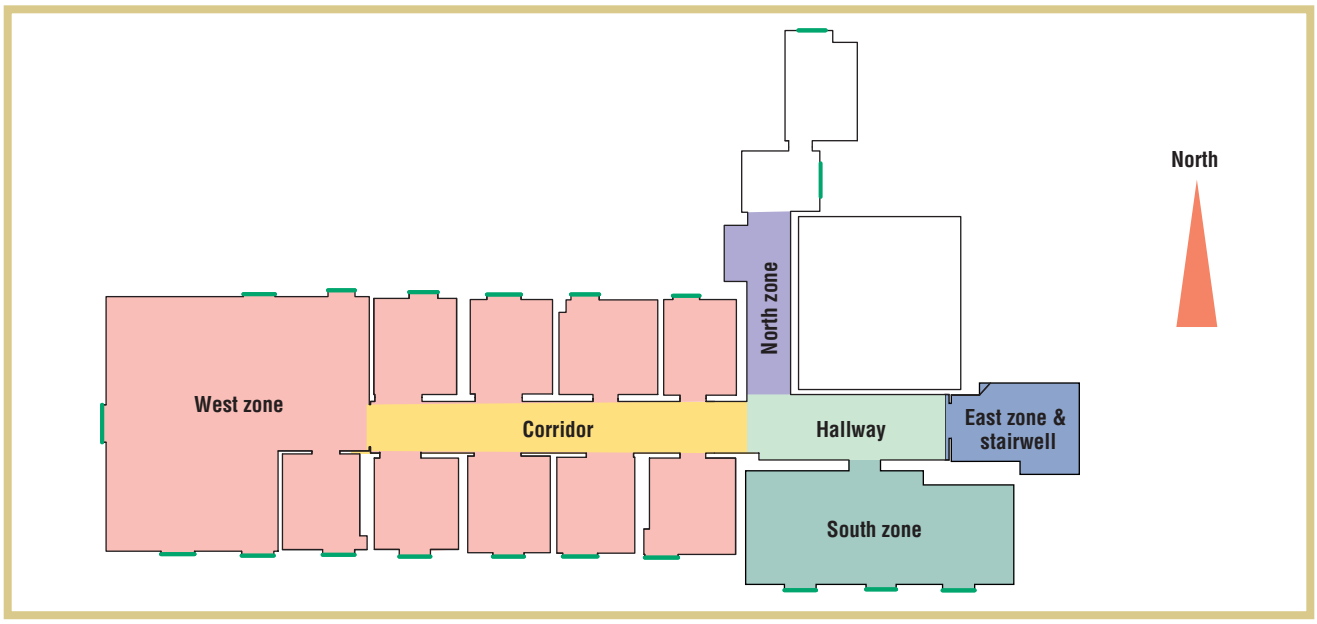


Figure 2. Floor plan layout of the laboratory. Combinations of the labeled areas form the mix zones z_1, z_2, and z_3, described in the main text.

Experimental data

We applied the anonymity set technique to location data collected from the Active Bat system[5] installed at AT&T Labs Cambridge, where one of us was sponsored and the other employed. Our location privacy experiments predate that lab's closure, so the data we present here refers to the original Active Bat installation at AT&T. The Laboratory for Communications Engineering at the University of Cambridge, our current affiliation, has since redeployed the system.

The system locates the position of *bats:* small mobile devices each containing an ultrasonic transmitter and a radio transceiver. The laboratory ceilings house a matrix of ultrasonic receivers and a radio network; the timing difference between ultrasound reception at different receivers lets the system locate bats with less than 3 cm error 95 percent of the time. While in use, typical update rates per bat are between one and 10 updates per second.

Almost every researcher at AT&T wore a bat to provide position information to location-aware applications and colleagues located within the laboratory. For this article, we examined location sightings of all the researchers, recorded over a two-week period between 9 a.m. and 5 p.m.—a total of more than 3.4 million samples. All the location sightings used in the experiment come from researchers wearing bats in normal use. Figure 2 provides a floor plan view of the building's first floor. The second- and third-floor layouts are similar.

Results

Using the bat location data, we simulated longer location update periods and measured the size of the anonymity sets for three hypothetical mix zones:

- z_1: first-floor hallway
- z_2: first-floor hallway and main corridor
- z_3: hallway, main corridor, and stairwell on all three floors

We chose to exclude from our measure of anonymity set sizes of zero occupancy, for the following reason. In general, a high number of people in the mix zone means greater anonymity and a lower number means less. However, this is only true for values down to one, which is the least-anonymous case (with the user completely exposed); it is not true for zero, when no users are present. While it is sensible to average anonymity set size values from one, it is rather meaningless to average values that also include zero.

Figure 3 plots the size of the anonymity set for the hallway mix zone, z_1, for update periods of one second to one hour. A location update period of at least eight minutes is required to provide an anonymity set size of two. Expanding z_1 to include the corridor yields z_2, but the results (not plotted) for this larger zone do not show any significant increase in the anonymity level provided.

Mix zone z_3, encompassing the main corridors and hallways of all three floors as well as the stairwell, considerably improves the anonymity set's cardinality. Figure 4 plots the number of people in z_3 for various location update periods. The mean anonymity set size reaches two for a location update period of 15 seconds. This value is much better than the one obtained using mix zone z_1, although it is still poor in absolute terms.

The anonymity set measurements tell us that our pseudonymity technique cannot give users adequate location privacy in this particular experimental situation (because of user population, resolution of the location system, geometry of the mix zones, and so on). However, two positive results counterbalance this negative one: First, we

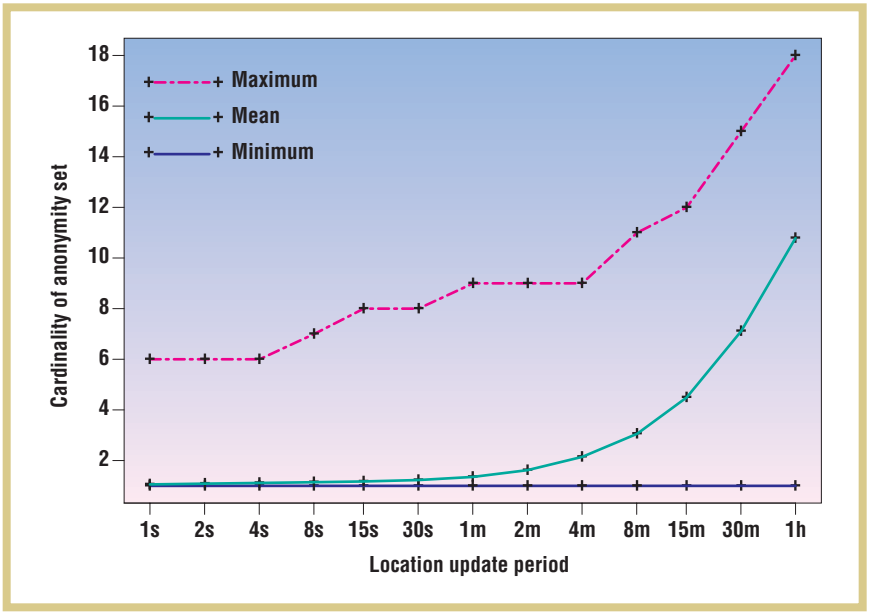


Figure 3. Anonymity set size for z_1.

have a quantitative measurement technique that lets us perform this assessment. Second, the same protection technique that did not work in the AT&T setting might fare better in another context, such as wide-area tracking of cellular phones.

Apart from its low anonymity set size, mix zone z_3 presents other difficulties. The maximum walking speed in this area, calculated from bat movement data, is 1.2 meters per second. To move from one extreme of z_3 to another, users require approximately 50 seconds. For update periods shorter than this value, we cannot consider users to be "mixed." Intuitively, it seems much more likely that a user entering the mix zone from an office on the third floor will remain on the third floor rather than move to an office on the first. Analysis of the data reveals this to be the case: in more than half the movements in this mix zone, users travel less than 10 meters. This means that the users in the mix zone are not all equal from the hostile observer's point of view: If I am in the mix zone with 20 other people, I might consider myself well protected. But if all the others are on the third floor while I am on the second, when I go in and out of the mix zone the observer will strongly suspect that those lonely pseudonyms seen on the second floor one at a time actually belong to the same individual. This discovery motivated the work we describe next.

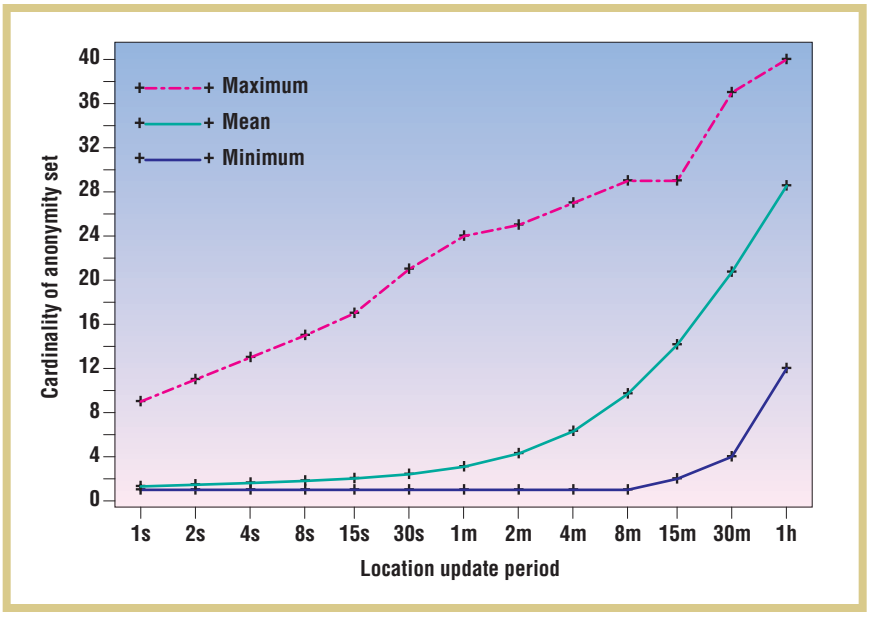


Figure 4. Anonymity set size for z_3.

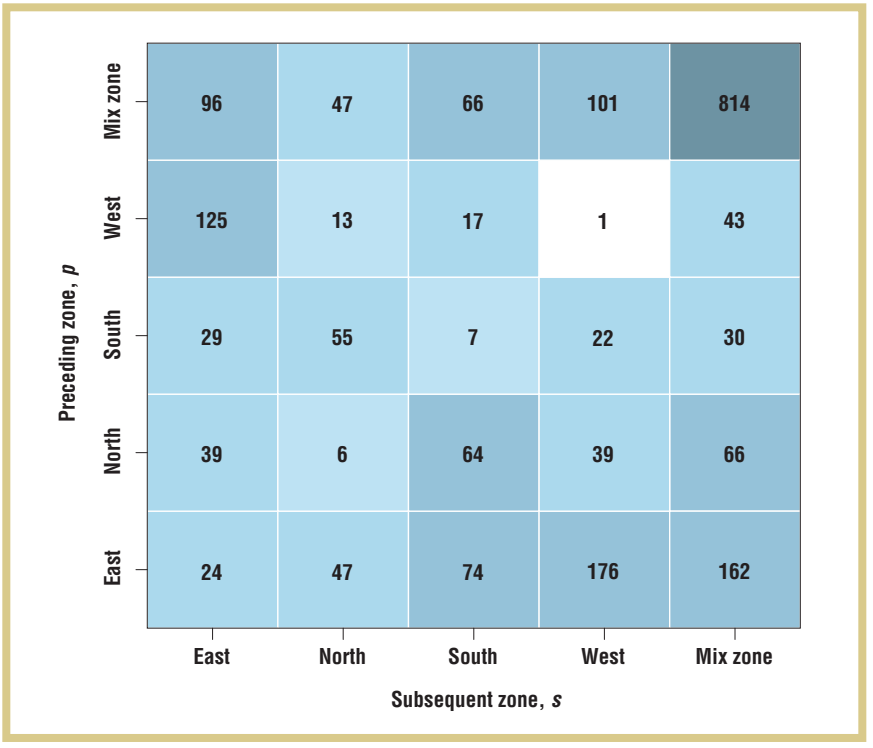
Accounting for user movement

So far we have implicitly assumed the location of entry into a mix zone is independent of the location of exit. This in general is not the case, and there is a strong correlation between ingress position and egress position, which is a function of the mix zone's geography. For example, consider two people walking into a mix zone from opposite directions: in most cases people will continue walking in the same direction.

Anonymity sets do not model user entry and exit motion. Andrei Serjantov and George Danezis[9] propose an information-theoretic approach to consider the varying probabilities of users sending and receiving messages through a network of mix nodes. We apply the same principle here to mix zones.

The crucial observation is that the anonymity set's size is only a good measure of anonymity when all the members of the set are equally likely to be the one of interest to the observer; this, according to Shannon's definition,[10] is the case of maximal entropy. For a set of size n, if all elements are equiprobable, we need $\log_2 n$ bits of information to identify one of them. If they are not equiprobable, the entropy will be lower, and fewer bits will suffice. We can precisely compute how many in the following way.

Figure 5. The movement matrix *M* records the frequency of movements through the hallway mix zone. Each element represents the frequency of movements from the preceding zone *p*, at time $t-1$, through zone *z*, at time *t*, to the subsequent zone *s*, at time $t+1$.



Consider user movements through a mix zone *z*. For users traveling through *z* at time *t*, we can record the preceding zone, *p*, visited at time $t-1$, and the subsequent zone, *s*, visited at time $t+1$. The user doesn't necessarily change zones on every location update period: the preceding and subsequent zones might all refer to the same mix zone *z*. Using historical data, we can calculate, for all users visiting zone *z*, the relative frequency of each pair of preceding and subsequent points (p, s) and record it in a movement matrix *M*. *M* records, for all possible (p, s) pairs, the number of times a person who was in *z* at *t* was in *p* at $t-1$ and in *s* at $t+1$. The entries of *M* are proportional to the joint probabilities, which we can obtain by normalization:

$$P(prec = p, subs = s) = \frac{M(p,s)}{\sum_{i,j} M(i,j)}.$$

The conditional probability of coming out through zone *s*, having gone in through zone *p*, follows from the product rule:

$$P(subs = s \mid prec = p) = \frac{M(p,s)}{\sum_{j} M(p,j)}.$$

We can now apply Shannon's classic measure of entropy[10] to our problem:

$$h = -\sum_{i} p_i \cdot \log p_i.$$

This gives us the information content, in bits, associated with a set of possible outcomes with probabilities p_i. The higher it is, the more uncertain a hostile observer will be about the true answer, and therefore the higher our anonymity will be.

Results

We applied this principle to the same set of Active Bat movement data described earlier. We consider the hallway mix zone, z_1, and define four hypothetical application zones surrounding it: east, west, north, and south (see Figure 2). Walls prevent user movement in any other direction. Figure 5 shows *M*, the frequency of all possible outcomes for users entering the hallway mix zone z_1, for a location update period of five seconds. From Figure 5 we observe that users do not often return to the same mix zone they came from. In other words, it is unlikely that the preceding zone, *p*, is the same as the subsequent zone, *s*, in every case except a prolonged stay in the mix zone itself. It is quite common for users to remain in z_1 for more than one location update period; the presence of a public computer in the hallway might explain this.

We can use the data from Figure 5 to determine how modeling user movements affects the anonymity level that the mix zone offers. As an example, consider two users walking in opposite directions—one east to west and the other west to east—passing through the initially empty hallway mix zone z_1 in the same location update period. East and west are application zones, so the hostile observer will know that two users went into z_1, one from east and another from west, and that later those users came out of z_1, under new pseudonyms, one of them into east and the other into west. What the observer wants is to link the pseudonyms—that is, to find out whether they both went straight ahead or each made a U-turn. Without any a priori knowledge, the information content of this alternative is one full bit. We will now show with a numerical example how the knowledge encoded in movement matrix *M* lets the observer "guess" the answer with a chance of success that is better than random, formalizing the intuitive notion that the U-turn is the less likely option.

We are observing two users, each moving from a preceding zone *p* through the mix zone z_1 to a subsequent zone *s*. If *p* and *s* are limited to {E, W}, we can reduce *M* to a 2×2 matrix M'. There are 16 possible cases, each representable by a four-letter string of Es and Ws: EEEE, EEEW, …, WWWW. The four characters in the string represent respectively the preceding and

subsequent zone of the first user and the preceding and subsequent zone of the second user.

We identify two events of interest. The first, *observed*, corresponds to the situation we observed: two users coming into z_1 from opposite directions and then again coming out of it in opposite directions. We have

$$observed = \text{EEWW} \vee \text{EWWE} \vee \text{WEEW} \vee \text{WWEE},$$

where $\vee$ means logical OR. The second event, *uturn*, corresponds to both users doing a U-turn:

$$uturn = \text{EEWW} \vee \text{WWEE}.$$

We see that, in fact, the first event includes the second. We can easily compute the probabilities of these two events by summing the probabilities of the respective four-letter subcases. We obtain the probability of a four-letter subcase by multiplying together the probability of the paths taken by the two users. For example, we obtain the probability of EWWE by calculating $P(\text{EW}) \times P(\text{WE})$. We obtain these terms, in turn, from the reduced movement matrix $\boldsymbol{M}'$ by normalization. Numerically, $P(observed) = 0.414$. Similarly, $P(uturn) = 0.0005$.

As the hostile observer, what we want to know is who was who, which is formally expressible as the conditional probability of whether each did a U-turn given what we saw: $P(uturn \mid observed)$. From the product rule, this is equal to $P(uturn \wedge observed) / P(observed)$. This, because the second event is included in the first, reduces to $P(uturn) / P(observed) = 0.001$.

So the observer now knows that they either did a U-turn, with probability 0.1 percent, or they went straight, with probability 99.9 percent. The U-turn is extremely unlikely, so the information content of the outcome is not one bit, as it would be if the choices were equiprobable, but much less—because even before hearing the true answer we are almost sure that it will be "they both went straight ahead." Numerically, the entropy of this choice is 0.012 bits.

This value is much lower than 1 because the choices are not equiprobable, and a hostile observer can therefore unlink pseudonyms with much greater success than the mere anonymity set size would suggest. The entropy therefore gives us a tighter and more accurate estimate of the available anonymity.[9]

Technologies for locating and tracking individuals are becoming increasingly commonplace, and they will become more pervasive and ubiquitous in the future. Location-aware applications will have the potential to follow your every move, from leaving your house to visiting the doctor's office, recording everything from the shelves you view in a supermarket to the time you spend beside the coffee machine at work. We must address the issue of protecting location information before the widespread deployment of the sensing infrastructure.

Applications can be built or modified to use pseudonyms rather than true user identities, and this is one route toward greater location privacy. Because different applications can collude and share information about user sightings, users should adopt different pseudonyms for different applications. Furthermore, to thwart more sophisticated attacks, users should change pseudonyms frequently, even while being tracked.

Drawing on the methods developed for anonymous communication, we use the conceptual tools of mix zones and anonymity sets to analyze location privacy. Although anonymity sets provide a first quantitative measure of location privacy, this measure is only an upper-bound estimate. A better, more accurate metric, developed using an information-theoretic approach, uses entropy, taking into account the a priori knowledge that an observer can derive from historical data. The entropy measurement shows a more pessimistic picture—in which the user has less privacy—because it models a more powerful adversary who might use historical data to de-anonymize pseudonyms more accurately.

Applying the techniques we presented in this article to data from our Active Bat system demonstrated that, because the temporal and spatial resolution of the location data generated by the bats is high, location privacy is low, even with a relatively large mix zone. However, we anticipate that the same techniques will show a much higher degree of unlinkability between pseudonyms over a larger and more populated area, such as a city center, in the context of locating people through their cellular telephones. P

ACKNOWLEDGMENTS

We thank our colleagues Andrei Serjantov, Dave Scott, Mike Hazas, Tom Kelly, and Gray Girling, as well as Tim Kindberg and the other IEEE Computer Society referees, for their thoughtful advice and comments. We also thank all of our former colleagues at AT&T Laboratories Cambridge for providing their movement data, as well as EPSRC and AT&T Laboratories Cambridge for financial support.

REFERENCES

1. S. Warren and L. Brandeis, "The Right to Privacy," *Harvard Law Rev.*, vol. 4, no. 5, Dec. 1890, pp. 193–200.
2. United Nations, *Universal Declaration of Human Rights*, General Assembly Resolution 217 A (III), 1948; www.un.org/Overview.
3. D. Banisar and S. Davies, *Privacy and Human Rights*, www.gilc.org/privacy/survey.
4. F. Stajano, *Security for Ubiquitous Computing*, John Wiley & Sons, New York, 2002.
5. A. Ward, A. Jones, and A. Hopper, "A New Location Technique for the Active Office," *IEEE Personal Communications*, vol. 4, no. 5, Oct. 1997, pp. 42–47.
6. D. Chaum, "Untraceable Electronic Mail, Return Addresses and Digital Pseudonyms," *Comm. ACM*, vol. 24, no. 2, 1981, pp. 84–88.
7. D. Chaum, "The Dining Cryptographers Problem: Unconditional Sender and Recipient Untraceability," *J. Cryptology*, vol. 1, no. 1, 1988, pp. 66–75.
8. A. Pfitzmann and M. Köhntopp, "Anonymity, Unobservability and Pseudonymity—A Proposal

for Terminology," *Designing Privacy Enhancing Technologies: Proc. Int'l Workshop Design Issues in Anonymity and Observability*, LNCS, vol. 2009, Springer-Verlag, Berlin, 2000, pp. 1–9.

9. A. Serjantov and G. Danezis, "Towards an Information Theoretic Metric for Anonymity," *Proc. Workshop on Privacy Enhancing Technologies*, 2002; www.cl.cam.ac.uk/users/aas23/papers_aas/set.ps.

10. C.E. Shannon, "A Mathematical Theory of Communication," *Bell System Tech. J.*, vol. 27, July 1948, pp. 379–423, and Oct. 1948, pp. 623–656.

the AUTHORS

Alastair Beresford is a PhD student at the University of Cambridge. His research interests include ubiquitous systems, computer security, and networking. After completing a BA in computer science at the University of Cambridge, he was a researcher at BT Labs, returning to study for a PhD in the Laboratory for Communications Engineering in October 2000. Contact him at the Laboratory for Communication Engineering, Univ. of Cambridge, William Gates building, 15, JJ Thomson Avenue, Cambridge, CB3 0FD, UK; arb33@cam.ac.uk.

Frank Stajano is a faculty member at the University of Cambridge's Laboratory for Communication Engineering, where he holds the ARM Lectureship in Ubiquitous Computing. His research interests include security, mobility, ubicomp, scripting languages, and middleware. He consults for industry on computer security and he is the author of *Security for Ubiquitous Computing* (Wiley, 2002). Contact him at the Laboratory for Communication Engineering, Univ. of Cambridge, William Gates building, 15, JJ Thomson Avenue, Cambridge, CB3 0FD, UK; fms27@cam.ac.uk; www-lce.eng.cam.ac.uk/~fms27.