Path Privacy in Location-aware Computing

Marco Gruteser†, Jonathan Bredin‡, Dirk Grunwald†

†Department of Computer Science   ‡Department of Mathematics
University of Colorado at Boulder   Colorado College
Boulder, CO 80309   Colorado Springs, CO 80903
gruteser,grunwald@cs.colorado.edu   jbredin@coloradocollege.edu

Abstract

Context-aware applications often require sharing private data with a service provider. Most contextual information, such as location, changes over time. Privacy mechanisms that can safeguard information when shared with less trusted organizations, however, remain more suitable for static or point-in-time information. We make the case for developing privacy mechanisms that adequately address time-series, such as locational path information, and discuss our work in-progress in path segmentation and minutiae suppression.

1 Introduction

Some location-based applications must continually monitor a user's movements. For example, a navigation application that provides turn-by-turn directions typically synchronizes the presentation of each path leg with a user's actual movements. An application might also update directions when a user misses a turn or chooses an alternate path.

Such location tracking applications are increasingly designed for resource limited devices, where location information must be processed on an external server. This is in stark contrast to the well-known automotive navigation systems in luxury vehicles, that process all location information on the in-car system; thus it creates immense privacy concerns as Barkhuus and Dey's user study [1] highlights.

One thread of research in privacy for context-aware systems [2, 3, 4] develops technologies to make users

aware of a service provider's data collection practices. It also allows them to easily express preferences that govern under what circumstances private data can be shared. Lederer and colleagues [5] found that the identity of the requester typically is the most significant factor in users' privacy decisions. These mechanisms allow sharing information with trusted parties, while blocking unwanted intrusions from untrusted ones. When sharing data with semi-trusted parties, such as a little-known service provider, users have to make a decision. It may not always be obvious what the revealed information implies, however.

Data perturbation or resolution control mechanisms offer users additional options when releasing data to semi-trusted service providers—a compromise between the extremes of foregoing a service altogether and entrusting a service provider with sensitive data. Such mechanisms can perturb location information to protect user privacy, but the service providers can still glean useful information from the data. There are two general approaches—one that renders users unidentifiable [6, 7] and one that hides individual observations but preserves aggregate distributions [8]. Both approaches, however, aim at protecting static sensitive values from each user. Thus they perform well when users reveal their location sporadically, but cannot protect a continuous path.

We proceed by discussing some location-aware applications and how they distribute user information in section 2. In section 3, we look at related work, including a fielded application to examine its privacy policy and discuss the value of extending the policy. Section 4 describes several attacks an adversary might mount on
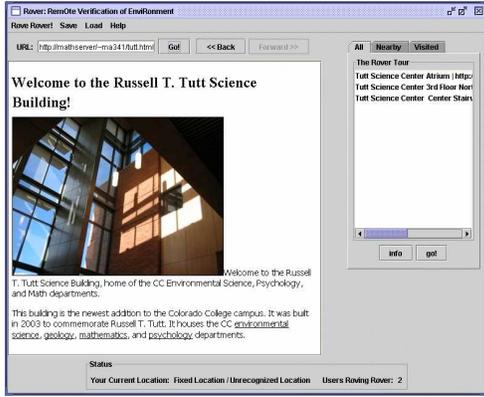
Figure 1: A screen shot of our students' location-aware browser informing the user of the various academic departments located nearby.

location-based application users and introduces the notion of anonymous paths. The implementations of such attacks are open research efforts. To further address them, we are gathering experimental results on the travels of active urban denizens that we preview in section 5.

2 Application Requirements

The path protection problem is motivated by location-based campus applications that we are developing. As part of an undergraduate course project, our students developed a location-aware browser, Rover, that determines a user's location from the available access-point service identifiers and associated signal strengths. The browser compares the signal-strength signature to a list of known locations' signatures to determine the best match. It then displays a page that describes the inferred environment. The browser also searches for users who report to be in nearby locations, and can establish chat connections to the local users. Figure 2 shows an example location inside the browser window. Our students are working on extending the browser to implement a more complete tour-guide similar to [9]; a disc-golf caddy; and digital-graffiti. These applications require frequent location updates and thus collect path information about users.

We observe a trend towards more complex location-based applications that require frequent location updates

also among commercial service offerings. Two examples are cell phone navigation and automotive traffic management.

Cell phone navigation [10] uses the phone to determine a user's position and to deliver turn-by-turn directions. The phone collects GPS data and sends it to a navigation server, that the service provider operates. The navigation server then computes an updated route to the intended destination and displays directions on the user's cell phone.

The automotive industry is conducting experiments to infer traffic conditions from data collected in vehicles [11]. Selected vehicles will periodically send their locations, speeds, road temperatures, windshield wiper status, and other information to a traffic monitoring facility. This data reveals the length of traffic jams (through speed and position), weather conditions such as rain (through windshield wiper), and slick road conditions (through frequent anti-lock breaking). Using vehicles as mobile sensing platforms promises dramatic cost reductions over deploying specialized roadside sensors.

These applications have different data quality requirements. The navigation application needs to track a particular user's movements from origin to destination of the route. Location updates need to be accurate enough for the service provider to determine the street on which a user is traveling.

The traffic management application also requires accurate location updates, but does not require the complete routes that users travel. Instead, it processes aggregate data from many users.

3 Related Work

Most location-based applications have privacy mechanisms that allow a user to control the publication of personal information. In this section, we look at a fielded application, and review related research work.

AT&T Wireless offers the mMode Find Friends service to broadcast a user's location to friends.[1] To address privacy issues, AT&T provides a simple access control list

---

[1] http://www.attwireless.com/personal/features/ organization/findfriendsprivacy.jhtml;dsessionid= GXYG2MHESFAFTB4R0EHCFEY Bell Mobility and TeliaSonera offer similar services to mobile-phone users.

(ACL) that allows only specified people to find user.

While Lederer and colleagues find that users' biggest concern is who requests location [5], further restrictions may allay users' privacy concerns. For example, time-based or location-based modifiers could support the ACL. A user may wish to cloak his location to friends during business hours, or cloak location to his employer off business hours. The privacy policy may include the presence of other users, say to hide a user's location in the presence of his girlfriend. As the policy becomes more specific (e.g., [2]), cognitive issues arise. A more complicated policy can obfuscate the actual level of privacy and security.

Privacy-aware data mining [8] seeks to enable a database owner to generate accurate *aggregate statistics* without possessing the accurate information for individuals. To this end, data is perturbed by adding a random offset. If a large number of individuals submit perturbed data, the perturbation can be filtered out to estimate the original distribution of data values, while each *individual* datum remains perturbed. An adversary could recover values for an individual, however, if each individual submits a large number of samples from a continuous function, such as physical movement [12]. Thus, current mechanisms only offer good protection, when subjects sporadically submit a single location, not when they submit frequent location updates (path information).

In prior work [7], we described a mechanism to strengthen anonymity of location-based queries. The mechanism reduces the resolution of revealed location information, so that a potential adversary cannot link the position to an individual user. It also can only guarantee anonymity for single positions. If it is known that multiple queries belong to the same user, the adversary can identify the user through intersection attacks.

To our knowledge, only Beresford and Stajano [13] have conducted research on anonymizing path information. They propose mix zones, which are geographic regions in which users are not visible to location-based services (i.e., location updates are suppressed). If sufficient users simultaneously pass through these zones, an adversary cannot determine which path segments leading into and out of a zone belong to the same users. This mechanisms relies on statically defined zones; therefore, it cannot guarantee protection in low-traffic areas.

4 Protecting Path Privacy

There are many potential attacks involving location-based data. For this discussion we will restrict our attention to passive inference based on the data that a user knowingly transmits on to an external service provider. We assume that an adversary gains access to this data—for example, through an insider attack or accidental data disclosure from the service provider. We discuss anonymizing paths by associating only a user's pseudonym with location information and allow the user to periodically change a pseudonym. Additionally, we note that the frequency of reporting visits to certain locations, such as an office cubicle or home, encroach privacy and we believe that some such observations should be suppressed.

4.1 Anonymous Paths

Obvious identifiers such as user names attached to collected path information can easily be removed or be replaced by pseudonyms. It is difficult to ensure that the characteristics of the path information do not give away the user's identity, however. For example, a longer stay at a particular home or workplace can identify the user behind an anonymous path.

From the adversary's perspective, a path is a collection of location/time pairs that a single user has visited (e.g., GPS data). Some service providers may allow the use of pseudonyms. More precisely, the path will contain all data from a user if the user always communicates under the same name or pseudonym with the service provider. Whenever a user switches his pseudonym, the following location information appears as a different path to the adversary. In the extreme case where the user transmits every location/time pair separately and anonymously, paths only comprise a single point.

Strong anonymity requires that a larger group of potential service users travel along the same path at the same time. An adversary could try to identify the user who generated an anonymous path by linking other available information to the path. For example, the adversary may know a user's home or office location and when that user passed through certain automatic toll booths. If a path matches the home, office, and toll booth locations, the adversary has identified the user, unless there are several other potential users who have visited the same locations at the

same time.

It is unlikely that many users simultaneously travel the same path. Therefore, we begin by looking at two forms of weaker anonymity protections: path segmentation and minutiae suppression. Weak anonymity implies that the data is distinctive enough so that it could be linked to an individual, however it requires much more effort.

Path segmentation truncates longer paths into several segments, whereby the goal is that an adversary cannot determine that two segments were generated by the same user. Thus, the adversary may be able to identify the user of a segment, but does not learn about the locations that the user visited in other segments.

We envision a proxy that adaptively controls the release of location updates that are revealed to the service provider. Such a mechanism could be implemented with two time parameters: segment duration and mean pause. Whenever the segment duration time passes, location updates are suppressed until the end of the pause period. The pause period is the mean pause offset by a random value.

The adversary, however, can often concatenate multiple segments (assuming changing pseudonyms) into a longer path because most users' movements are relatively predictable in small areas—they typically follow streets and often move at known speeds (near the speed limit, walking speed, etc.) When the paths of multiple users cross, however, linking segments becomes more difficult. The adversary's certainty is higher when precise position and time information is available in areas with low user density. Thus, the parameters must be chosen with care and adapted to the current situation. In the next section, we derive a model to calculate the likelihood that of connecting two paths and we will empirically derive the range of acceptable intervals at which applications can safely publish data.

Minutiae Suppression attempts to hide the most distinctive characteristics of a path. While any combination of points may give away a user's identity, some characteristics are much easier to exploit than others. For example, Beresford and Stajano found that all anonymous location traces from an office environment were correctly identified by heuristics such as checking at which desk subjects spend most of their time [13].

5   First Experiments

To better understand the synergism that data amass towards identifying a user, we are currently running experiments to collect paths to drive an analytic model to determine the likelihood that multiple observations relate to the same subject. The model shows the strength of path segmentation in the experimental study. We have two mechanisms to collect the data, global positioning (GPS) devices and network signal-strength analysis, with which we study the impact on privacy of path crossover, user pseudonym change, and location publication suppression.

5.1   Privacy Model

In this section we describe a model to join path segments. To determine the likelihood that two segments came from the same subject, we compute the conditional probability that two observations relate to the same subject given that subjects were observed at a location pair,

$$P(\text{identity}|\text{observed pair}) = \frac{P(\text{identity} \wedge \text{pair})}{P(\text{pair})}. \quad (1)$$

If "first" and "second" are the events comprising "pair," and if the locations are reasonably frequented, then

$$P(\text{pair}) = P(\text{first})P(\text{second}), \quad (2)$$

where $P(\text{first})$ and $P(\text{second})$ can be computed by measuring the frequency with which the locations are occupied. The conjunction in equation 1 is simply the product

$$P(\text{identity} \wedge \text{pair}) = P(\text{first})P(\text{second}|\text{first}). \quad (3)$$

Linking the above calculations together, we arrive at

$$P(\text{identity}|\text{observed pair}) = \frac{P(\text{second}|\text{first})}{P(\text{second})}. \quad (4)$$

In the next section, we collect data to calculate the conditional probability $P(\text{second}|\text{first})$ from looking at the population's average mobility.

4

5.2 Path Collection

Our current experiment asks Colorado College students to carry GPS devices over the course of their daily activities. Colorado College is a small, urban, residential college situated near downtown Colorado Springs. We believe that with a small sample size, we can quickly collect frequently-intersecting paths that well model a larger society. The data for our experiment have no real identities associated to them, but for training and evaluation purposes, we will have all data linked to a corresponding pseudonym for the subject who generated the datum.

Figure 2 plots segments of five paths from our study that were collected over the course of several days. The plot shows several interesting characteristics. First, the green and white paths show significant similarities and a cluster of samples in a residential area (upper left corner). Thus, they likely stem from the same user, who lives in the house at the center of the cluster. The yellow and blue paths also contain unique clusters of points on campus, where users stayed for a longer time. These clusters may help an adversary identifying the users behind these paths and support use of a minutiae suppression technique to protect privacy.

We also observe that even this small sample of users shows areas where different users' paths overlap. In these areas the privacy algorithm could create enough uncertainty so that the paths become indistinguishable and a potential adversary looses track of a user.

From the tracks, we calculate the probability of observed displacement over a fixed time conditioned on the subject's heading and position. Qualitatively, we see that users tend to continue traveling in the same direction, and we get a distribution of distances traveled and direction changed over the time period. The distribution allows us to compute $P(\text{second}|\text{first})$ for equation 4.

Additionally, how the probability distribution changes over time tells us how much information is disclosed with a second observation linked to the first. Figure 3 plots how the entropy in displacement changes over time. Initially, knowing the distribution of all users' movements helps in predicting future positions. After about five minutes, the predictions using only average displacement lose most of their predictive power. The plot determines how densely observations must be made for location-aware applications to accurately function.
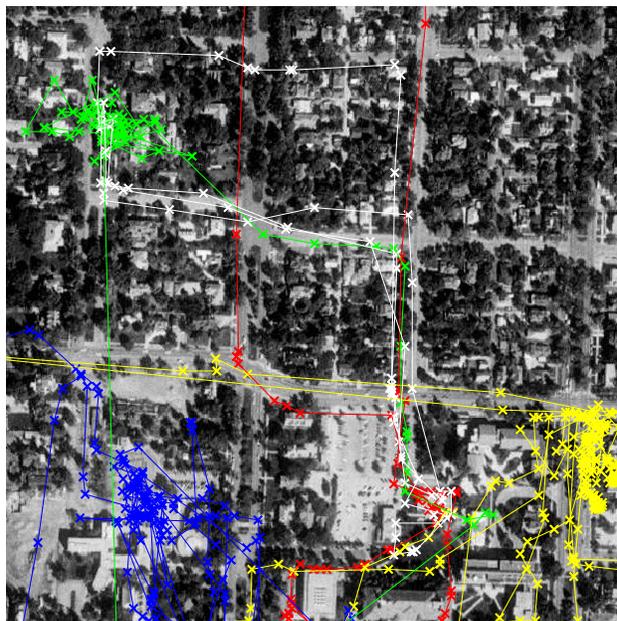

Figure 2: Five GPS paths shown over a satellite image of the area. The paths contain several clusters, where users stayed for an extended time. There are also several areas where different users' paths overlap.

6   Summary and Future Work

Realistically, many location-aware applications must run from resource-poor devices and hence rely on remote, possibly untrusted, devices to store and compute path information. We believe that perturbing location information does not adequately protect a user's privacy. We promote associating only pseudonyms with users' location data; the ability to periodically change pseudonym; and to suppress geographic minutiae, such as observations of a user's path around heavily frequented locations, such as an office. These two mechanisms require an adversary to reconstruct paths from disjoint observations. To investigate the likelihood of reconstruction, we are examining paths collected by subjects, and derive a simple model to calculate the likelihood that two observations associate to the same identity.

Our interests lie in calculating how frequently and from where users can submit location updates. The result likely
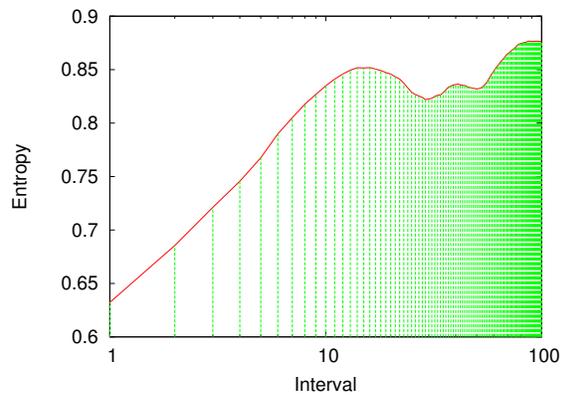
5

Figure 3: The measured entropy of subjects' displacement calculated every 15 seconds. All measurements are normalized by the maximum possible entropy.

depends on user density. Additionally, we are investigating the protection afforded by clusters of users traveling along paths. Perhaps by perturbing every cluster member's position, we can better cloak each user's identity.

On the horizon, we also plan to compare GPS based results to those from wireless LAN positioning. GPS systems only function under an open sky and we would also like to know how privacy can be supported or exploited indoors as well. We have developed tools to associate wireless-networked devices with discrete locations, similar to methods employed by the Place Lab [14]. The tools calculate the signal strengths of all available access points and compare them to a list of known signal-strength/access-point pairs to determine the best fit. Our current implementation uses a least-squares method to associate signal-strength signatures and locations.

References

[1] Louise Barkhuus and Anind Dey. Location-based services for mobile telephony: a study of users' privacy concerns. In *9th Internation Conference on Human-Computer Interaction (INTERACT)*, 2003.

[2] Ginger Myles, Adrian Friday, and Nigel Davies. Preserving privacy in environments with location-based applications. *IEEE Pervasive Computing*, 2(1):56–64, 2003.

[3] Marc Langheinrich. A privacy awareness system for ubiquitous computing environments. In *4th International Conference on Ubiquitous Computing*, 2002.

[4] Sastry Duri, Marco Gruteser, Xuan Liu, Paul Moskowitz, Ronald Perez, Moninder Singh, and Jung-Mu Tang. Framework for security and privacy in automotive telematics. In *2nd ACM International Workshpop on Mobile Commerce*, 2002.

[5] Scott Lederer, Jennifer Mankoff, and Anind Dey. Who wants to know what when? privacy preference determinants in ubiquitous computing. In *Extended Abstracts of Conference on Human Factors in Computing Systems (CHI)*, pages 724–725, 2003.

[6] Latanya Sweeney. Achieving k-Anonymity Privacy Protection Using Generalization and Suppression. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5):571–588, 2002.

[7] Marco Gruteser and Dirk Grunwald. Anonymous usage of location-based services through spatial and temporal cloaking. In *Proceedings of the First International Conference on Mobile Systems, Applications, and Services*, 2003.

[8] Rakesh Agrawal and Ramakrishnan Srikant. Privacy-preserving data mining. In *Proc. of the ACM SIGMOD Conference on Management of Data*, pages 439–450. ACM Press, May 2000.

[9] Keith Cheverst, Nigel Davies, Keith Mitchell, and Adrian Friday. Experiences of developing and deploying a context-aware tourist guide: the GUIDE project. In *Proceedings of the sixth annual international conference on Mobile computing and networking*, pages 20–31. ACM Press, August 2000.

[10] Telenav. `http://www.telenav.net/`, Jan 2004.

[11] Rajiv Vyas. Ford device intended to unclog roads. `http://www.freep.com/money/autonews/ford27_20040227.htm`, Feb 2004.

[12] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar. Random data perturbation techniques and privacy preserving data mining. In *IEEE International Conference on Data Mining*. IEEE Press, 2003.

[13] Alastair Beresford and Frank Stajano. Location privacy in pervasive computing. *IEEE Pervasive Computing*, 2(1):46–55, 2003.

[14] Jason Hong, Geatano Boriello, James Landay, David McDonald, Bill Schilit, and J.D. Tygar. Privacy and security in the location-enhanced world wide web. In *Workshop on ubicomp communities: privacy as boundary negotiation (held at Ubicomp)*, Oct 2003.

6