Anonymous Usage of Location-Based Services Through Spatial and Temporal Cloaking

Marco Gruteser and Dirk Grunwald
Department of Computer Science
University of Colorado at Boulder
Boulder, CO 80309
{gruteser,grunwald}@cs.colorado.edu

Abstract

Advances in sensing and tracking technology enable location-based applications but they also create significant privacy risks. *Anonymity* can provide a high degree of privacy, save service users from dealing with service providers' privacy policies, and reduce the service providers' requirements for safeguarding private information. However, guaranteeing anonymous usage of location-based services requires that the precise location information transmitted by a user cannot be easily used to re-identify the subject. This paper presents a middleware architecture and algorithms that can be used by a centralized location broker service. The adaptive algorithms adjust the resolution of location information along spatial or temporal dimensions to meet specified anonymity constraints based on the entities who *may* be using location services within a given area. Using a model based on automotive traffic counts and cartographic material, we estimate the realistically expected spatial resolution for different anonymity constraints. The median resolution generated by our algorithms is 125 meters. Thus, anonymous location-based requests for urban areas would have the same accuracy currently needed for E-911 services; this would provide sufficient resolution for wayfinding, automated bus routing services and similar location-dependent services.

1 Introduction

Improvements in sensor and wireless communication technology enable accurate, automated determination and dissemination of a user's or object's position [1, 2]. There is an immense interest in exploiting this positional data through location-based services (LBS) [3, 4, 5, 6]. For instance, LBSs could tailor their functionality to the user's current location, or vehicle movement data would improve traffic forecasting and road planning.

However, without safeguards, extensive deployment of these technologies endangers users' location privacy and exhibits significant potential for abuse [7, 8, 9]. Common privacy principles demand, among others, user consent, purpose binding,[1] and adequate data protection for collection and usage of personal information [10]. Complying with these principles generally requires notifying users (data subjects) about the data collection and the purpose through privacy policies; it also entails implementing security measures to ensure that collected data is only accessed for the agreed-upon purpose.

This paper investigates a complimentary approach that concentrates on the principle of minimal collection. In this approach location-based services collect and use only de-personalized data—that is, *practically anonymous* data [11]. This approach promises benefits for both parties. For the service provider, practically anonymous data causes less overhead. It can be collected, processed, and distributed to third parties without user consent. For data subjects, it removes the need to evaluate potentially complex service provider privacy policies.

Practical anonymity requires that the subject cannot be reidentified (with reasonable efforts) from the location data. Consider a message to a road map service that comprises a network address, a user ID, and coordinates of the current location. Identifiers like the user ID and the network address are obvious candidates for reidentification attempts. For anonymous service usage, the user ID can be omitted and the network address problem is addressed by mechanisms such as Crowds [12] or Onion Routing [13], which provide sender anonymity.

However, revealing accurate positional information can pose even more serious problems. Consider a bus wayfinding application that overlays bus route and arrival information, such as that marketed by NextBus [14]. The Global Positioning System (GPS) typically provides 10–30 foot accuracy, and this accuracy can be increased using enhancement techniques, such as differential GPS. A location-based service could query a bus transit server and return information about buses in the current vicinity and when they will arrive at various stops. By issuing such a query, the location-based service has learned information about the application user, including her location and some network identity information. This location information can be correlated with public knowledge to reidentify a user or vehicle. For example, when the service is used while still parked in the garage or on the driveway, the location coordinates can be mapped to the address and the owner of the residence. If queries

[1] When seeking user consent, data collectors need to explain the specific purpose for which the data will be used. Subsequent use for other purposes is prohibited without additional user approval.

are sufficiently frequent, they can be used to track an individual. Note that this tracking uses publicly available information as opposed to the identity behind network addresses. The privacy problems are magnified if location information is recorded and distributed continuously as envisioned in telematics applications such as "pay as you drive" insurance, traffic monitoring, or fleet management. In this case an adversary not only learns about network services that a subject uses but also can track the subjects movements and thus receives real-world information such as frequent visits to a medical doctor, nightclub, or political organizations.

Anonymity in LBSs must be addressed at multiple levels in the network stack depending on what entities can be trusted. This paper approaches the problem of anonymity at the application layer by giving service providers access to anonymous location information; that is, information that is sufficiently altered to prevent re-identification. It contributes the following key ideas:

- a formal metric for location anonymity
- an adaptive quadtree-based algorithm that decreases the spatial resolution of location information to meet a specified anonymity constraint
- an algorithm that yields higher spatial resolution through decreasing temporal resolution for the same anonymity constraint
- an evaluation of the expected resolution for these algorithms based on traffic models comprised of cartographic material and automotive traffic counts

The structure of the paper is as follows: First we review related work in the areas of location privacy, anonymous communication, and privacy-aware databases. In Section 3 we describe location-based service scenarios from the telematics domain and discuss their data accuracy requirements. Section 4 then analyzes privacy threats caused by the location information used in LBSs. We continue by developing the concept of k-anonymous location information and an algorithm for cloaking too precise information in section 5. After that, we describe our implementation and evaluation based on automotive traffic models and present the corresponding results. Finally, we discuss the usefulness of the cloaking algorithms as well as security and anonymity properties of the system.

2 Related Work

Prior work on privacy aspects of telematics and location-based applications has mostly focused on a policy-based approach [15, 16]. Data subjects need to evaluate and choose privacy policies offered by the service provider. These policies serve as a contractual agreement about which data can be collected, for what purpose the data can be used, and how it can be distributed. Typically, the data subject has to trust the service provider that private data is adequately protected. In contrast, the anonymity-based approach de-personalizes data before collection, thus detailed privacy-policies and safeguards for data are not critical.

Specifically, the IETF Geopriv working group [15] is addressing privacy and security issues regarding the transfer of high resolution location information to external services and the storage at location servers. It concentrates on the design of protocols and APIs that enable devices to communicate their location in a confidential and integrity-preserving manner to a location server. The location server can reduce the data's resolution or transform it to different data formats, which can be accessed by external services if the data subject's privacy policy permits. The working group is also interested in enabling unidentified or pseudonymous transfer of location information to the server and access from the server. However, it does not claim that this provides a sufficient degree of anonymity.

The Mist routing project for mobile users [17] combines location privacy with communication aspects. It addresses the problem of routing messages to a subject's location while keeping the location private from the routers and the sender. To this end, the system comprises a set of mist routers organized in a hierarchical structure. The leaf nodes have knowledge of user locations but not their identities. They refer to them through handles (or pseudonyms). Each user selects a higher-level node in the tree, which acts as a semi-trusted proxy. It knows the identity of the user but not his exact location. The paper then presents a cryptographic protocol to establish connections between users and their semi-trusted proxies and mechanisms to connect to communication partners through their proxies. The paper does not address the problem of sending anonymous messages to external location-based services.

Location privacy has also been studied in position sensor systems systems. The Cricket system [1] places location sensors on the mobile device as opposed to the building infrastructure. Thus, location information is not disclosed during the position determination process and the data subject can choose the parties to which the information should be transmitted. Smailagic and Kogan describe a similar approach for a wireless LAN based location system [18]. However, these solutions do not provide for anonymity when location information is intentionally revealed.

Anonymous communication in packet-switching networks and web browsing has received a fair amount of attention. The fundamental concept of a mix has been proposed by Chaum [19] for email communications that are untraceable even for eavesdroppers and intermediary routers. A mix is a message router that forwards messages with the objective that an adversary cannot match incoming messages to outgoing messages. In particular, such Chaum-mixes have the following properties: messages are padded to equal size, incoming and outgoing

messages are encrypted with different keys, messages are batched and reordered, and replay of incoming messages is prevented. Pfitzmann and colleagues [20] extend this mechanism to communication channels with continuous, delay-sensitive voice traffic.

Onion Routing [21] implements this anonymization protocol for an IP network layer and is applicable to both connection-based and connectionless protocols. In an initialization phase, the sender determines a route through a series of onion routers. The sender then repeatedly adds routing information to the payload and encrypts it using the onion routers public key. The result is an onion consisting of several layers of encryption that are stripped off while the packet passes through the router. Since the onion routers act as mix routers, it is difficult to trace the path of a data packet through the network.

Crowds [12] adapts a rerouting system for anonymous web browsing. This system focuses on protecting against individual adversaries, such as the web server, or a number of compromised routers. It does not require encryption techniques, because it relies on the jondos (mix routers) to be set up in different administrative domains. Thus no party has a global network view over all jondos. The Anonymizer service [22] has a similar goal, whereby users need to trust the single service provider. Finally, Hordes [23] reduced the performance overhead inherent in such rerouting systems by exploiting multicast communications and Guan *et al*. [24] contributed an analysis of anonymity properties of these systems using the probabilistic method.

In the database community, a large amount of literature exists on security control in statistical databases, which is covered by Adam and Wortmann's survey [25]. This research addresses the problem wherein a database should grant access to compute statistical functions (sum, count, average, etc.) on the data records only under the condition that the results do not reveal any specific data record. Approaches fall into the categories conceptual, input data perturbation, query restriction, and output perturbation; the solution we propose in this paper is similar to input data perturbation.

Instead of statistical point estimates, Agrawal and Srikant [26] describe how to obtain estimates of the distribution of values in confidential fields, which are suitable for data-mining algorithms. Confidential values are perturbed by adding a uniformly distributed random variable. The distribution of the original values can then be estimated through a Bayesian reconstruction procedure. An improved reconstruction procedure is described in [27].

Samarati and Sweeney [28] have developed generalization and suppression techniques for values of database tables that safeguard the anonymity of individuals. While this research is similar in goal, our work differs in that we protect dynamic data delivered from sensors as opposed to static database tables.

3 Accuracy Requirements of Location-Based Telematics Services

A key question for developing an anonymous LBS is: How accurate does a location based service need to be in order to provide useful information? It proves difficult to determine minimum accuracy requirements, since, from the service provider's perspective, more accurate information is generally more useful. However, we attempt to convince the reader that more general information is still sufficient for a large class of services by reviewing example services and the E-911 requirements on mobile phone carriers.

In October 1996, the United States Federal Communications Commission mandated the implementation of position systems for wireless 911 emergency callers (E-911) [29]. This service is designed to provide emergency rescue and response teams with the location of a cell phone emergency call, comparable to the traditional "911" service for regular phones. In the final phase, wireless carriers are required to estimate the caller's position with an accuracy of 125 m (RMS) in 67 percent of cases. The details have subsequently been subject to debate, but this initial requirement gives an indication of the expected accuracy. The location systems developed for the E-911 requirement have been widely regarded as an enabling technology for location-based services; therefore, we will regard this level of accuracy as useful.

3.1 System Assumptions

We assume that clients communicate position information to a location server with very high precision; in other words, the network client actually provides an accurate location to the location server. Position determination can be implemented either on the client itself (e.g., GPS) or by the wireless service provider, for example through triangulation of the wireless signal (hybrid approaches are also possible). To our knowledge, mobile phone operators in the United States found it challenging to meet the E-911 accuracy requirements through the latter approach. Thus, GPS information is likely far more accurate and privacy sensitive. Location-based service providers access location information through the location server. The full system comprises a location information source, a wireless network, location servers, and LBS servers. In a typical system, location information is determined by a location information source such as a GPS receiver in a vehicle. It is then periodically transmitted through a cellular or wireless network to the location server. When a vehicle sends a message or request to an LBS, the service accesses the vehicle's current location information from the location server, which acts as a proxy or middleware agent.

Finally, this paper focuses on services that do not require the user to log in and or present any kind of identifying information at the application layer. We believe that such LBSs will become available analogous to free services over the Internet. However, it would be inter-

esting to extend this research to pseudonymous LBSs, which would allow tailoring services to individual users, for example.

3.2 Scenarios

To illustrate different accuracy requirements of location-based services, we provide three typical automotive telematics scenarios: Driving Conditions Monitoring, Road Hazard Detection, and a Road Map. Services are differentiated along the following dimensions:

- Frequency of Access
- Time-accuracy / Delay sensitivity
- Position accuracy

Table 1 presents a summary of the resulting requirements.

Driving Conditions Monitoring

Modern vehicles carry a variety of sensors that can determine weather and road conditions. Instead of deploying an expensive array of fixed sensors alongside highways, highway operators could obtain this information from the in-vehicle sensors. For example, the rain sensor built into high-end windshield wipers detects rainfall; additionally, traction control systems can report slippery or icy road conditions. The operator might respond to this information by dynamically adjusting speed limits on the highway.

Weather phenomena and corresponding road conditions typically cover larger areas. In addition, most warnings and speed limits must be given well ahead of the hazardous conditions. Thus, highly accurate position information is not necessary; about 100m road segments should be a suitable resolution for most cases. Conditions also do not change very abruptly, thus updates with a few minutes delay can be tolerated. In order to detect a change in driving conditions the external application needs quasi-continuous access to location information.

Road Hazard Detection

Dangerous, near-accident situations could be inferred from braking or electronic stability systems. Additionally, crash sensors for airbag deployment detect severe accidents. This information could be exploited to automatically generate statistics about the accident risk at intersections and road segments. These statistics are valuable for deciding on accident prevention measures.

Since this application collects longer-term statistics, information delay is not important and time accuracy requirements are low. For example, it would be useful to distinguish night and daylight situations or rush hour from mid-day traffic but not to collect information with second-resolution. Precise location information is crucial, however, to pinpoint dangerous spots such as intersections or pedestrian crossings.

Road Map

Drivers might request information related to their current location from LBSs. For example, the driver can ask for an area map or nearby hotels. The current location can be automatically obtained from the GPS sensor of a vehicle navigation system.

Response times of these services are important, thus, this application requires high time accuracy. The location, however, can be transmitted with medium accuracy; about 100m accuracy should be sufficient for obtaining point-of-interest information and area maps. Location is revealed only sporadically, when the driver issues requests. If such systems are used for navigation, the location can be revealed much more frequently.

4 Privacy Threats Through Location Information

We assume that an adversary seeking to violate anonymity may be able to intercept wireless and wired communications, may obtain data from the service provider's systems, and may have prior knowledge about a subject, whose messages he seeks to identify.

Our main concern is to prevent an accumulation of identifiable location information in service providers systems. LBS providers, without any malicious intent, will likely log service requests, similar to a web server that logs requested URLs and source IP addresses of the requester. Logs that include location information would open the door for subpoenas in court (e.g., divorce) proceedings, or individual adversaries who obtain a subject's location information under a pretext. Moreover, a less conscientious service provider might seek to identify subjects for marketing purposes or sell location records to third parties. In these cases, an adversary targets a large number of subjects, or seeks to obtain a location history for a particular subject from the records of a service provider.

A different type of adversary seeks to track future movements of a particular subject. However, such location information can also be obtained through traditional investigative methods such as shadowing a subject or mounting a location transmitter to a vehicle. These methods are related to the LBS problem in that they define a currently accepted level of protection. We consider the protection of anonymous LBSs sufficient if location tracking requires effort comparable to the traditional methods.

4.1 Threats

We distinguish two classes of privacy threats related to location-based services: communication privacy threats and location privacy threats. In the communication privacy domain, this paper concentrates on sender anonymity, meaning that eavesdroppers on the network and LBS providers cannot determine the originator of a message. Compared to non-LBS web services, the location information is the key problem: an adversary can

Service	Position Accuracy	Time Accuracy	Frequency of Access
Driving Conditions Monitoring	100 meters	minutes	continuous
Road Hazard Detection	10 meters	$>$ days	sporadic
Road Map	100 meters	sub-second	sporadic

Table 1: Approximate accuracy requirements of telematics services

reidentify the sender of an otherwise anonymous message by correlating the location information with prior knowledge or observations about a subject's location.

Consider the case where a subject reveals her location L in a message M to a location-based service and an adversary A has access to this information. Then, sender anonymity and location privacy is threatened by location information in the following ways:

Restricted Space Identification. If A knows that space L exclusively belongs to subject S then A learns that S is in L and S has sent M. For example, when the owner of a suburban house sends a message from his garage or driveway, the coordinates can be correlated with a database of geocoded postal addresses (e.g., [30]) to identify the residence. An address lookup in phone or property listings then reveals the owner and likely originator of the message.

Observation Identification. If A has observed the current location L of subject S and finds a message M from L then A learns that S has sent M. For example, the subject has revealed its identity and location in a previous message and then wants to send an anonymous message. The later message can be linked to the previous one through the location information.

Location Tracking. If A has identified subject S at location L_i and can link series of location updates $L_1, L_2, \ldots, L_i, \ldots, L_n$ to the subject, then A learns that S visited all locations in the series.

Location privacy threats describe the risk that an adversary learns the locations that a subject visited (and corresponding times). Through these locations, the adversary receives clues about private information such as political affiliations, alternative lifestyles, or medical problems. Assuming that a subject does not disclose her identity at such a private location, an adversary could still gain this information through location tracking. If the subject transmits her location with high frequency, the adversary can, at least in less populated areas, link subsequent location updates to the same subject. If at any point the subject is identified, her complete movements are also known.

5 Anonymizing Location Information

In our system model, the mobile nodes communicate with external services through a central anonymity server that is part of the trusted computing base. In an initialization phase, the nodes will set up an authenticated and encrypted connection with the anonymity server. When a mobile node sends position and time information to an external service, the anonymity server decrypts the message, removes any identifiers such as network addresses, and perturbs the position data according to the following cloaking algorithms to reduce the reidentification risk. Moreover, the anonymity server acts as a mix-router [19], which randomly reorders messages from several mobile nodes, to prevent an adversary from linking ingoing and outgoing messages at the anonymity server. Finally, the anonymity server forwards the message to the external service.

For designing the perturbation algorithms, we start with the assumption that the anonymity server knows the current position of all subjects. The subject's mobile nodes could periodically update their position information with the anonymizer.

5.1 k-Anonymous Location Information

While *anonymity* is etymologically defined as "being nameless" or "of unknown authorship" [31], information privacy researchers interpret it in a stronger sense. According to Pfitzmann and Koehntopp, "anonymity is the state of being not identifiable within a set of subjects, the anonymity set"[11]. Inspired by Samarati and Sweeney [28], we consider a subject as k-*anonymous* with respect to location information, if and only if the location information presented is indistinguishable from the location information of at least $k-1$ other subjects.

Unless otherwise stated, we assume that location information includes temporal information (i.e., when the subject was present at the location). More specifically, location information is represented by a tuple containing three intervals $([x_1, x_2], [y_1, y_2], [t_1, t_2])$. The intervals $[x_1, x_2]$ and $[y_1, y_2]$ describe a two dimensional area where the subject is located. $[t_1, t_2]$ describes a time period during which the subject was present in the area. Note that the intervals represent uncertainty ranges; we only know that at some point in time within the temporal interval the subject was present at some point of the area given by the spatial intervals. Thus, a location tuple for a subject is k-anonymous, when it describes not only the location of the subject, but also the locations of $k-1$ other subjects. In other words, $k-1$ other subjects also must have been present in the area and the time period described by the tuple. Generally speaking, the larger the anonymity set k is, the higher is the degree of anonymity. Thus, we will measure the degree of anonymity as the

size of the anonymity set.

5.2 Adaptive-Interval Cloaking Algorithms

The key idea underlying this algorithm is that a given degree of anonymity can be maintained in any location—regardless of population density—by decreasing the accuracy of the revealed spatial data. To this end, the algorithm chooses a sufficiently large area, so that enough other subjects inhabit the area to satisfy the anonymity constraint.

The desired degree of anonymity is specified by the parameter k_{min}, the minimum acceptable anonymity set size. Furthermore, the algorithm takes as inputs the current position of the requester, the coordinates of the area covered by the anonymity server, and the current positions of all other vehicles/subjects in the area.

The spatial discretization algorithm that identifies a sufficiently large area for a given k_{min} is described in more detail in Table 2. In summary, the algorithm is inspired by quadtree algorithms [32]. It subdivides the area around the subject's position until the number of subjects in the area falls below the constraint k_{min}. The previous quadrant, which still meets the constraint, is then returned.

An orthogonal approach to spatial cloaking is temporal cloaking. This method can reveal spatial coordinates with more accuracy, while reducing the accuracy in time. The key idea is to delay the request until k_{min} vehicles have visited the area chosen for the requestor. The spatial cloaking algorithm is modified to take an additional spatial resolution parameter as input. It then determines the monitoring area by dividing the space until the specified resolution is reached. The algorithm monitors vehicle movements across this area. When k_{min} different vehicles have visited the area, a time interval $[t_1, t_2]$ is computed as follows: t_2 is set to the current time, and t_1 is set to the time of request minus a random cloaking factor. The area and the time interval are then returned.

6 Implementation

To be effective, the location anonymizer requires location-based services that are used with precise position information by a large user base. We anticipate that such services will soon be available based on telematics, mobile phone, or wireless community network platforms. To our knowledge, no such suitable testbed exists to date. Therefore, we implemented the anonymization algorithms on a Java server platform and evaluated them using automotive traffic simulations based on US geological survey (USGS) cartographic material.

The USGS publishes detailed transportation network information at the city level in the Spatial Data Transfer Standard [33]. We extracted vector coordinates of primary, secondary, and minor roads from the transportation layer of the $1 : 24,000$ scale Digital Line Graph [34] data files. The data has a resolution of 0.61m. Specifically, we selected 2000x2000m areas from the city of Denver, Colorado, where we had access to traffic count statistics. Figure 1 shows maps of selected areas. The thickest lines indicate expressways, the medium lines arterials, and the thin lines collector streets. Two maps (area 1, area 2) included predominantly expressways, the other maps (area 3, area 4) mostly collector streets. Coordinates are given in meters in zone 13 of the Universal Transverse Mercator (UTM) projection using the North American 83 geodetic datum.

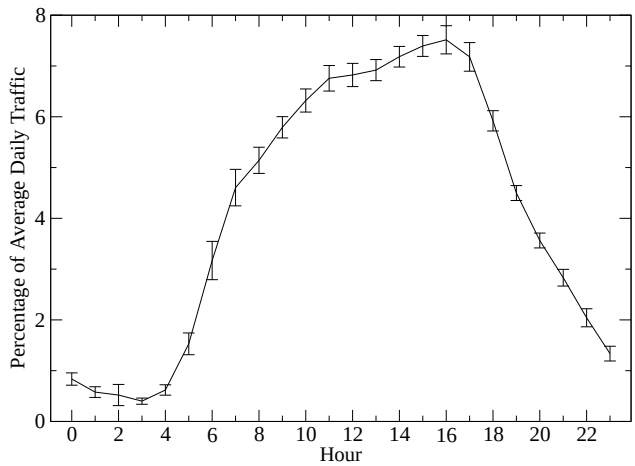


Figure 2: Hourly traffic volume relative to daily traffic volume during a typical August 2002 day

A traffic study [35] reports the 24 hour traffic volume at specific points along roads. We averaged the counts for different urban road types and mapped them onto the USGS road classes as shown in Table 3. Traffic volume was computed as the average 24 hour bi-directional traffic count.

USGS Class	Road Type	Traffic Volume
1	Expressway	70000
2	Arterial	22000
3	Collector	6000

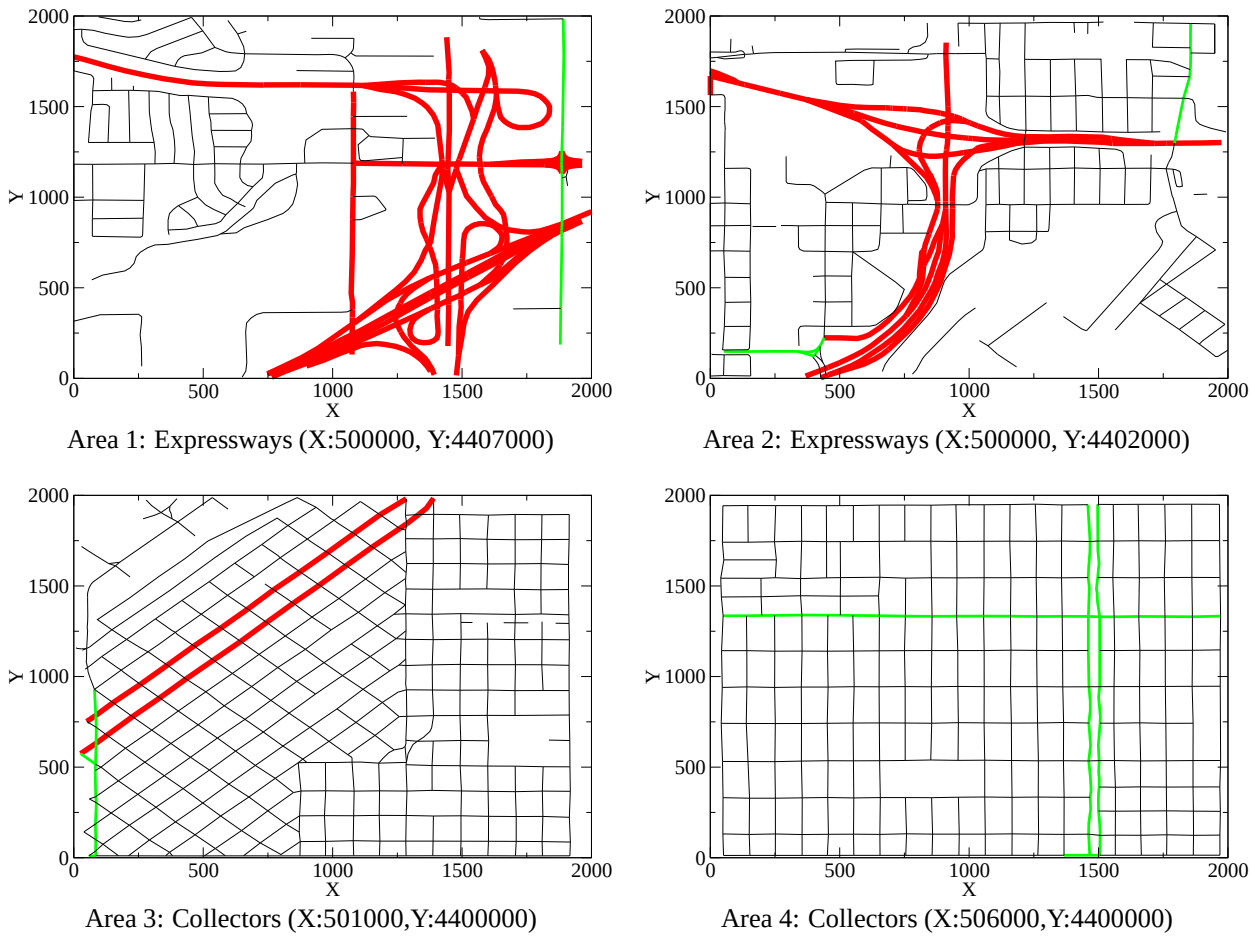
Table 3: Mapping of Traffic Study volumes to road classes from USGS data

The algorithms are evaluated at different times of day, because traffic volume changes heavily between peak and night hours. An adjustment factor for each hour was derived as follows. The Colorado Department of Transportation maintains continuous traffic counters along several highways. For each hour of a day, we calculated the percentage of total daily traffic present during this hour from the mean August 2002 traffic counts for 25 highways [36]. Figure 2 shows the results marked with 95% confidence intervals.

To create a traffic snapshot for a given hour, we place vehicles on the road segments according to a uniform stochastic process. The number of vehicles on a road

1.	Initialize the quadrants q and q_{prev} as the total area covered by the anonymizer
2.	Initialize a traffic vector with the current positions of all known vehicles
3.	Initialize p as the position of requestor vehicle
4.	If the number of vehicles in traffic vector $< k_{min}$, then return the previous quadrant q_{prev}
5.	Divide q into quadrants of equal size
6.	Set q_{prev} to q
7.	Set q to the quadrant that includes p
8.	Remove all vehicles outside q from the traffic vector
9.	Repeat from Step 2

Table 2: Adaptive-interval cloaking algorithm. The algorithm computes an area (quadrant) that includes the actual requester and enough potential requesters to satisfy the anonymity constraint k_{min}.



Area 1: Expressways (X:500000, Y:4407000)

Area 2: Expressways (X:500000, Y:4402000)

Area 3: Collectors (X:501000,Y:4400000)

Area 4: Collectors (X:506000,Y:4400000)

Figure 1: Selected 2000x2000m evaluation areas. The thickest lines indicate expressways, the medium lines arterials, and the thin lines collector streets. Two maps (area 1, area 2) included predominantly expressways, the other maps (area 3, area 4) mostly collector streets. Coordinates are in meters based on UTM Zone 13.

segment n is determined by

$$n = \frac{l \times c \times h}{v}$$

with traffic count c, hour adjustment h, vehicle velocity v, length of a road segment l.

For all experiments we assume an average velocity v of 10m/s and report mean results for a 24-hour period, that is, one snapshot for each hour of day. Unless otherwise stated, the anonymity constraint k_{min} was set to 5, which we intuitively judge as a fair level of anonymity.

7 Accuracy Results

Figure 3 presents an overview of our results. It illustrates the dependency of the resulting spatial resolution on road characteristics and traffic densities. For each of the selected maps and corresponding traffic densities, the median spatial resolution of 10000 simulated LBS requests is shown. In addition, the mean anonymity k, which represents the average number of subjects in the chosen area, is plotted against the second scale (right) on the y-axis.

The median resolution decreases as collector street mileage increases over highway mileage. For the highway areas with their high density of vehicles, the median accuracy is 30 and 65 meters. For the collector areas the resolution decreases to 125 and 250 meters. Interestingly, across all areas the spatial intervals selected by the adaptive algorithms not only have the same anonymity bound (5 subjects), but also a similar mean anonymity at approximately 10 subjects.

Figure 4 and Figure 5 provide more detail on the spatial resolution results by showing the relative distribution of resolutions in the form of histograms for a highway (1) and a collector area (4), respectively. While in the highway area less than 10% of requests reach a resolution lower than 125 meters, it is about 60% for the collector street area. Figure 5 also illustrates the relationship between anonymity and resolution in a single area. For lower resolutions (bigger areas) the mean anonymity does not stay near the minimum, but increases to more than triple the k_{min} constraint of five. When the algorithm is forced to choose a lower resolution, it has to quadruple the area and thereby includes more vehicles than necessary.

Figure 6 illustrates the tradeoff between the degree of anonymity and resolution, showing median resolution and mean anonymity k for different anonymity constraints k_{min}. The results stem from area number 3 with predominantly collector streets. Resolution is negatively correlated to the anonymity constraint. Also note that mean anonymity is approximately double the anonymity constraint. Again, this suggests that an improved discretization algorithm could yield better resolution with lower mean anonymity (i.e., closer to the minimum constraint).

Spatial resolution can also be improved through reductions in temporal accuracy. Figure 7 shows the mean reduction in temporal resolution (and delay of messages) required to reach a specified spatial resolution. Results are reported for a highway area (2) and a collector street area (3). The anonymity constraint is also varied between five and ten. As expected, the temporal accuracy decreases for higher anonymity constraints, more collector streets, and higher spatial resolution. For highways the temporal resolution stays below 30s for resolutions up to 15m. On collector streets the resolution decreases to about 70s at this level of spatial accuracy.

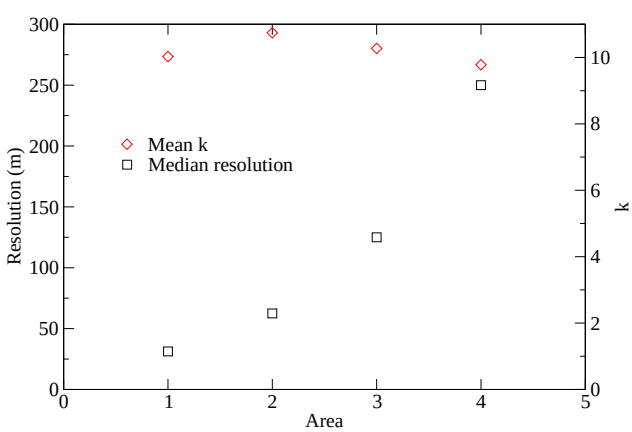


Figure 3: Dependency of spatial resolution and mean anonymity on area characteristics. For each evaluation area, the figure shows the median resolution from a large number of requests (left y-axis scale) and the mean *actual* anonymity—the number of subjects indistinguishable from the requestor (right y-axis scale).

8 Discussion

The analysis concentrates on interpreting the accuracy results and identifying anonymity and security limitations. We define security problems as adversarial attempts to obtain more accurate or extra data from the system that violates the anonymity constraint. Anonymity problems involve identification based on the data allowed by the anonymity constraint.

8.1 Accuracy

The results are encouraging when compared against the E-911 requirements introduced in Section 3 as a yardstick for useful location information. However, they vary widely across different types of areas. The highway areas yield better than required accuracy (less than 10% over

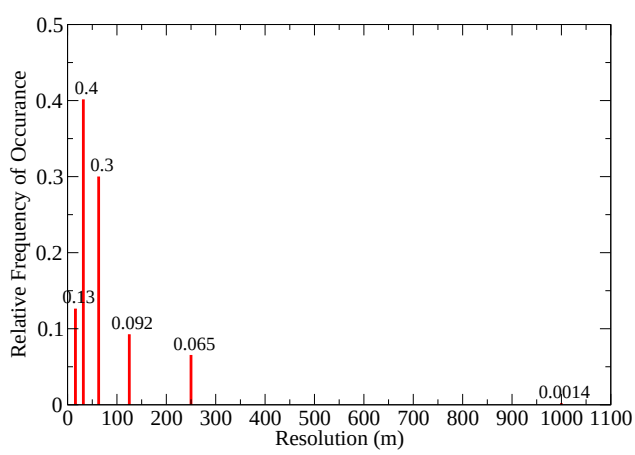


Figure 4: Relative frequency of spatial resolution for highway area (1). This figure illustrates the distribution of resolutions over a large number of simulated requests.

125m), collector street area (3) with a median of 125m is close to the requirements, and collector street area (4) clearly does not meet the requirements.

The results also show that the spatial accuracy can be adjusted through reductions in temporal resolution. The inherent delay of this approach makes it unsuitable for services that require a quick response. However, a large class of monitoring services along the scenarios of driving conditions monitoring and road hazard detection are well served by this approach. Brief delays, comparable to the delays experienced in web browsing over slower Internet connections, might also be acceptable for more interactive LBSs such as a road map service. These delays would also ensure at least 125m resolutions for the collector street area (4). Furthermore, delaying requests becomes unnecessary if the system can precompute temporal and spatial resolutions before the requests are issued. We believe that an investigation of this approach would be a worthwhile continuation of this work.

8.2 Security Analysis

From a security perspective, the wireless carriers or eavesdroppers can attempt to circumvent the location anonymizer and accurately locate a subject using the wireless channel. Authentication and encryption between the location client and the anonymity server effectively prevents them from listening to the exact coordinates or impersonating anonymity servers. The timestamp in the location beacons ensures that the bitstring of subsequent encrypted packets from the same location differs and also protects from replay attacks.

More difficult to prevent are attempts to estimate the location of a transmitter based on physical layer properties of the network. Several cooperating receivers can triangulate the position of a transmitter through methods such as time of arrival (TOA)[29]. Judging from the technical difficulties encountered in implementing the E-911

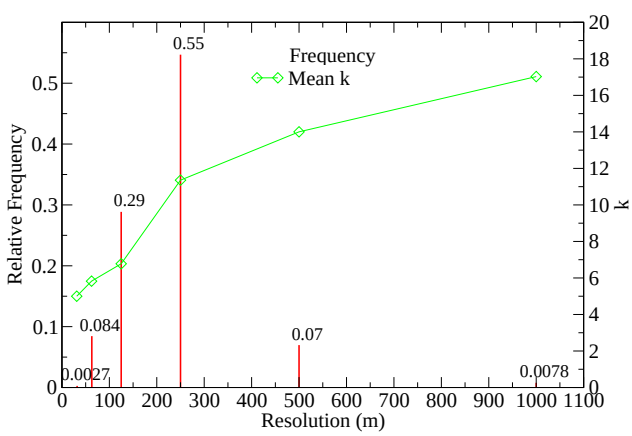


Figure 5: Relative frequency of spatial resolution for collector road area (4). In addition to the distribution of resolutions (left y-axis scale), the figure shows the mean actual anonymity at each resolution (right y-axis scale).

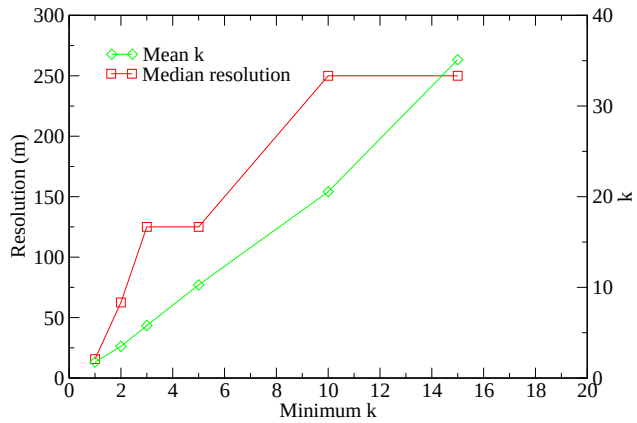


Figure 6: Dependency of spatial resolution and mean anonymity on anonymity constraint. This figure illustrates how spatial resolution (left scale) and mean actual anonymity (right scale) vary with different anonymity constraints (x-axis).

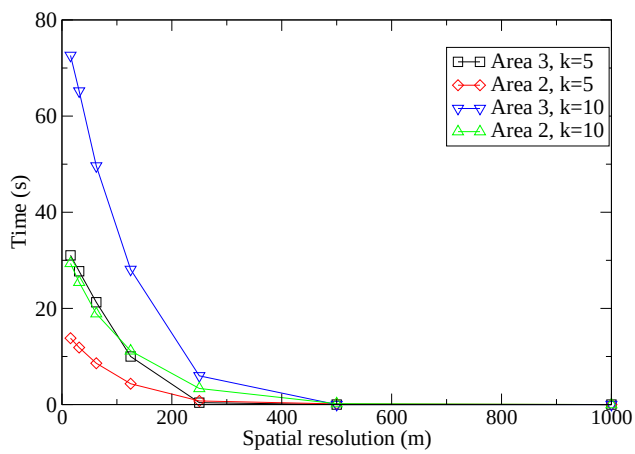


Figure 7: Tradeoff between temporal and spatial resolution. The figure shows the mean reduction in temporal resolution necessary to reach a specified spatial resolution. The tradeoff is shown for a highway (2) and a collector (3) area at different anonymity constraints.

requirement for mobile phones, location information obtained through these mechanisms would likely be about 1-2 orders of magnitude less accurate than information reported by in-vehicle GPS receivers. Thus, anonymity constraints would rarely be violated.

Another potential attack seeks to trick the location server into releasing too accurate data. An adversary could spoof a number of additional *virtual* vehicles or have real vehicles report incorrect location information. If the location server includes these nonexistent vehicles in its computation, the released location information would likely not meet the anonymity constraints. However the location server only accepts one location beacon from each authenticated vehicle. This means, the adver-

sary would need to acquire a potentially large number of authentication keys. Therefore, authentication keys require adequate protection, such as storage in secure hardware.

8.3 Anonymity

k-Anonymity reduces the privacy threats outlined in section 4. If a location tuple is k-anonymous, the adversary cannot uniquely identify the originator of a message through space identification or observation identification, since the tuple matches $k-1$ other subjects as well. Given no other information the reidentification risk is therefore $\frac{1}{k}$. Similarly, location tracking faces obstacles when attempting to link subsequent location updates to a subject. Since $k-1$ other subjects are in the area, it is not clear whether the location update actually originated from the same subject. In other words, if multiple subjects are using a LBS through the anonymity service, it is difficult for the LBS to generate movement paths of subjects even if they provide location updates with high frequency.[2]

At this point, it is difficult to gauge which size of k is minimally necessary or sufficient. Fundamentally, it depends on the resources of the potential adversary. A minimum of 2 is obviously required in this particular algorithm to yield any protection. In practice, the parameter will likely be determined through user preferences.

While the basic algorithm ensures k-anonymity for individual location requests, problems can arise when requests for multiple vehicles are issued. Consider the following location tuples obtained from 4 different vehicles:

$$1 : ([0,1],[0,1],[t_1,t_2])$$

$$2 : ([1,2],[0,1],[t_1,t_2])$$

$$3 : ([0,1],[1,2],[t_1,t_2])$$

$$4 : ([0,2],[0,2],[t_1,t_2])$$

These tuples are overlapping in time and space. The first three tuples specify adjacent quadrants, while the fourth one specifies a larger quadrant that covers the three others; this scenario is also illustrated in Figure 8. For simplicity, assume that all tuples were processed with the same k_{min} parameter, say 3, and the time interval is too small for vehicles to significantly move. Then an adversary can conclude that request number 4 must have originated from quadrant $([1,2],[1,2])$, because otherwise the algorithm would have chosen a smaller quadrant. This inference violates the anonymity constraint; it illustrates that an adversary gains information from tuples overlapping in time and space.

Furthermore, sophisticated adversaries may mount an identification attack if they can link multiple requests to the same subject and can repeatedly obtain the subject's location information from other sources. Consider an unpopular LBS that is rarely accessed. If this service is requested repeatedly from approximately the same spatial region, an adversary could conclude that the requests stem with high probability from the same subject (i.e., link the requests). If, in addition, the adversary knows that a certain subject was present in each of the spatial areas specified in the LBS requests, the adversary can determine with high probability that this subject originated the requests. This inference is based on the assumption that it is unlikely that other subjects from the anonymity set traveled along the same path, given the path is long enough (enough request were observed) relative to the size of the anonymity set. However, such an attack requires a large effort in determining a subject's actual position for a sufficient number of requests.

[2]Recall that we assume subjects do not transmit an identifier or pseudonym such as user ID to the LBS that would allow for trivial linking of subsequent location updates

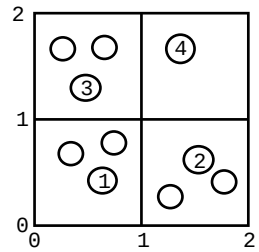


Figure 8: Compromised anonymity through overlapping requests. The circles and squares represent subjects and the quadrants computed by the cloaking algorithm, respectively. If each numbered subject requests a service simultaneously, subject 4 could be identified.

Although the anonymity constraint is not met in such cases, further research is needed to determine how serious these issues are. In practice, not every overlapping request allows such straightforward inferences and the probability of overlaps depends on the frequency of requests issued by subjects. To ensure meeting the anonymity constraint, disclosure control extensions to the cloaking algorithm could keep track of and prevent overlapping requests. Similarly, the algorithm could take into account the popularity of the accessed LBS to prevent linking of unusual requests.

Finally, it is important to realize that k-anonymity is only provided with respect to location information. Other service-specific information contained inside a message to a LBS could still identify the subject. This is analogous to anonymous communication services, which reduce reidentification risks of network addresses, but do not address other message content. Location information, however, will likely pose more serious risks than other typical message content.

9 Conclusions and Future Work

This paper analyzed the technical feasibility of anonymous usage of location-based services. It showed that location data introduces new and potentially more se-

vere privacy risks than network addresses pose in conventional services. Both the reidentification and the location tracking risk can be reduced through k-anonymous data. A system model and a quadtree-based algorithm were introduced to guarantee k-anonymous location information through reductions in location resolution. The main question we addressed was whether the resulting data accuracy is adequate for location-based services. Since the accuracy is dependent on traffic conditions, the algorithm was empirically evaluated using a traffic distribution model derived from traffic counts and cartographic material. Specifically, we draw the following conclusions:

- The quadtree-based algorithm reached accuracy levels comparable to the phase II E-911 requirements, and thus should be suitable for many location-based services.
- In areas with major highways the median accuracy is approximately 30m and increases to 250m for city areas with large block sizes. These results were obtained with an anonymity constraint of 5, yielding a mean anonymity level of approximately 10 people who may have issued a particular request.
- Spatial resolution can be significantly improved through a several seconds reduction in temporal resolution. Because of the imposed delay, this method is most applicable to noninteractive services.

9.1 Future Work

There are three directions for future work. The first avenue attempts to improve upon the resolution of the anonymizer. We plan to study clustering algorithms that can more intelligently pick minimally sized areas with sufficient traffic. The mean traffic volume in the areas identified by the current algorithms is approximately double the anonymity constraint, which leaves ample room for improvements. Furthermore, the algorithms should be able to operate with incomplete location information, where the position of subjects is periodically sampled rather than continuously updated.

The more difficult issue is decoupling the anonymizer from the current client-server architecture. For individual users to remain anonymous, the location server must have sufficient users within a geographic locale; unless the different users subscribe to the same location service, the reduced sample population available to any given location server may not suffice to anonymize queries for a given area. The algorithms we have used are efficient, and could execute on a wireless device. However, they require location information from different devices in the local area in order to judge the density of devices. Thus, at first sight, a "peer-to-peer" location anonymizing system requires access to the same information that it is attempting to cloak.

Lastly, we plan to deploy this anonymity system in a wireless LAN community network. Such community networks use high-speed wireless networking to provide Internet access; one example are the wireless access points common at coffee shops. These wireless networks have a limited range of 300–1500 feet, meaning that coarse location information can be determined simply by associating with a specific access point. In these networks, location based cloaking must occur at the application, network and physical layers.

Acknowledgments

Paul Chou and our colleagues at the IBM T.J. Watson Research Center encouraged us to research location privacy challenges. The anonymous referees and our paper shepherd, Maria Ebling, provided useful comments on a draft of this paper. This work was funded by award #9988548 from the Division of Computer-Communications Research of the National Science Foundation. All views and opinions stated reflect those of the authors, not those of the National Science Foundation.

References

[1] Nissanka B. Priyantha, Anit Chakraborty, and Hari Balakrishnan. The cricket location-support system. In *Proceedings of the sixth annual international conference on Mobile computing and networking*, pages 32–43. ACM Press, 2000.

[2] I. Getting. The global positioning system. *IEEE Spectrum*, 30(12):36–47, December 1993.

[3] Mike Spreitzer and Marvin Theimer. Providing location information in a ubiquitous computing environment (panel session). In *Proceedings of the fourteenth ACM symposium on Operating systems principles*, pages 270–283. ACM Press, 1993.

[4] Andy Harter, Andy Hopper, Pete Steggles, Andy Ward, and Paul Webster. The anatomy of a context-aware application. In *Mobile Computing and Networking*, pages 59–68, 1999.

[5] Rui Jose and Nigel Davies. Scalable and flexible location-based services for ubiquitous information access. In *Proceedings of First International Symposium on Handheld and Ubiquitous Computing, HUC'99*, pages 52–66. Springer Verlag, 1999.

[6] C. Bisdikian, J. Christensen, J. Davis II, M. Ebling, G. Hunt, W. Jerome, H. Lei, and S. Maes. Enabling location-based applications. In *1st Workshop on Mobile commerce*, 2001.

[7] P. A. Karger and Y. Frankel. Security and privacy threats to ITS. In *Proceedings of the Second World Congress on Intelligent Transport Systems*, volume 5, Yokohama, Japan, Nov 1995.

[8] Roy Want, Andy Hopper, Veronica Falco, and Jonathan Gibbons. The active badge location system. *ACM Transactions on Information Systems (TOIS)*, 10(1):91–102, 1992.

[9] Philip E. Agre. Transport informatics and the new landscape of privacy issues. *Computer Professionals for Social Responsibility (CPSR) Newsletter*, 13(3), 1995.

[10] Marc Langheinrich. Privacy by design – principles of privacy-aware ubiquitous systems. In G.D. Abowd, B. Brumitt, and S. Shafer, editors, *Ubicomp 2001 Proceedings*, volume 2201 of *Lecture Notes in Computer Science*, pages 273–291. Springer, 2001.

[11] Andreas Pfitzmann and Marit Koehntopp. Anonymity, unobservability, and pseudonymity —a proposal for terminology. In Hannes Federrath, editor, *Designing Privacy Enhancing Technologies — Proceedings of the International Workshop on Design Issues in Anonymity and Unobservability*, volume 2009 of *LNCS*. Springer, 2000.

[12] Michael K. Reiter and Aviel D. Rubin. Crowds: anonymity for Web transactions. *ACM Transactions on Information and System Security*, 1(1):66–92, 1998.

[13] D. Goldschlag, M. Reed, and P. Syverson. Onion routing for anonymous and private internet connections. *Communications of the ACM (USA)*, 42(2):39–41, 1999.

[14] NextBus Information Systems. Nextbus website. 1321 67th Street, Emeryville, CA 94608, USA, `http://www.nextbus.com`, Oct 2002.

[15] J. Cuellar, J. Morris, and D. Mulligan. Internet engineering task force geopriv requirements. http://www.ietf.org/html.charters/geopriv-charter.html, Oct 2002.

[16] 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 *Proceedings of the second international workshop on Mobile commerce*, pages 25–32. ACM Press, 2002.

[17] Jalal Al-Muhtadi, Roy Campbell, Apu Kapadia, M. Dennis Mickunas, and Seung Yi. Routing through the mist: Privacy preserving communication in ubiquitous computing environments. In *Proceedigns of IEEE International Conference of Distributed Computing Systems (ICDCS)*, pages 65–74, Vienna, Austria, Jul 2002.

[18] Asim Smailagic and David Kogan. Location sensing and privacy in a context-aware computing environment. *IEEE Wireless Communications*, 9(5):10–17, Oct 2002.

[19] David L. Chaum. Untraceable electronic mail, return addresses, and digital pseudonyms. *Communications of the ACM*, 24(2):84–90, 1981.

[20] A. Pfitzmann, B. Pfitzmann, and M. Waidner. Isdnmixes: Untraceable communication with very small bandwidth overhead. In Wolfgang Effelsberg, Hans Werner Meuer, and Günter Müller, editors, *Proceedings of Kommunikation in Verteilten Systemen, Grundlagen, Anwendungen, Betrieb, GI/ITG-Fachtagung*, volume 267 of *Informatik-Fachberichte*, Mannheim, Germany, Feb 1991. Springer.

[21] David Goldschlag, Michael Reed, and Paul Syverson. Onion routing. *Communications of the ACM*, 42(2):39–41, 1999.

[22] Anonymizer. Anonymizer website. 5694 Mission Center Road #426, San Diego, CA 92108-4380, `http://www.anonymizer.com`, 2000.

[23] Clay Shields and Brian Neil Levine. A protocol for anonymous communication over the internet. In *Proceedings of the 7th ACM conference on Computer and communications security*, pages 33–42. ACM Press, 2000.

[24] Yong Guan, Xinwen Fu, Riccardo Bettati, and Wei Zhao. A quantitative analysis of anonymous communications. *IEEE Transactions on Reliability*, (to appear).

[25] Nabil R. Adam and John C. Worthmann. Security-control methods for statistical databases: a comparative study. *ACM Computing Surveys (CSUR)*, 21(4):515–556, 1989.

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

[27] Dakshi Agrawal and Charu C. Aggarwal. On the design and quantification of privacy preserving data mining algorithms. In *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 247–255. ACM Press, May 2001.

[28] P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical Report SRI-CSL-98-04, Computer Science Laboratory, SRI International, 1998.

[29] J. Reed, K. Krizman, B. Woerner, and T. Rappaport. An overview of the challenges and progress in meeting the e-911 requirement for location service. *IEEE Personal Communications Magazine*, 5(3):30–37, April 1998.

[30] Tele Atlas North America, Inc. Geocode website. 1605 Adams Drive, Menlo Park, CA 94025, `http://www.geocode.com/`, Oct 2002.

[31] J.A. Simpson and E.S.C. Weiner, editors. *Oxford English Dictionary, Second Edition*. Clarendon Press, 1989.

[32] Hanan Samet. *The Design and Analysis of Spatial Data Structures*. Addison-Wesley, Reading, MA, 1990.

[33] U.S. Geological Survey (USGS). Spatial data transfer standard. 12201 Sunrise Valley Drive, Reston, VA 20192, USA, `http://mcmcweb.er.usgs.gov/sdts/`, 1995.

[34] U.S. Geological Survey (USGS). Digital line graph data. 12201 Sunrise Valley Drive, Reston, VA 20192, USA, `http://edc.usgs.gov/geodata/`, Oct 2002.

[35] DCROG. Denver regional council of governments: Denver regional travel behavior inventory, 2001. 2480 W. 26th Avenue, Suite 200B, Denver, CO 80211-5580.

[36] Colorado Department of Transportation. Traffic statistics & data. Public Relations Office, 4201 E Arkansas Ave, Denver, CO 80222, `http://www.dot.state.co.us/`, Oct 2002.