The Risks and Regulation of Location

Dirk Grunwald, Aaron Beach, Kevin Bauer, Qin Lv, and Douglas Sicker

Department of Computer Science

University of Colorado

Abstract

In the last ten years, concerns about location privacy have evolved from an academic topic that struggled to justify concerns about security to a mainstream issue that is affecting consumers, businesses and the legal system. Much of this proliferation of concerns arises from telecommunication and mobile computing platforms. Smart phones and GPS-assisted devices play an increasing role in people's lives, and the technology of precise and easily obtained location information has imbued mobile and social media with location in advance of the public knowing how that information will be used and fully grasping the implication of pervasive location information. Furthermore, social media sites such as Twitter, Facebook, Foursquare, Buzz and many others have adopted location as a key part of the information they communicate. While social networks are not critical parts of the communication landscape, they are used by hundreds of millions of people who are only beginning to understand the potential problems with providing easily accessible location information to an industry with a checkered history of transparent privacy policies. At the same time, telecommunication networks use GPS, assisted GPS and other location technologies to enhance localization as a necessary part of the Emergency-911 services. On an increasing basis, consumers' location information is being distributed with and without their knowledge. To what extent are users of these new technologies exposing themselves to identity attacks through sharing location information?

How can consumers understand and control how and when their information is being used or distributed and who has access to it?

In this paper, we briefly describe the technologies that underlie location-based services, their access and the notion of how location can be linked or inferred from multiple sources. We then survey common visible and hidden uses of location services, including social networks and emergency services. The technical community has developed a number of methods to hide or mask locations to provide a degree of anonymity while still preserving the benefit of location services. We briefly survey those methods and the "threat models" they seek to counter. We then describe threat models, or disclosures of location information, not commonly considered by the research community and their implications.

We lastly turn our attention to the policy implications of these technologies and the potential concerns that they present in terms of user privacy and safety. The technical community has long used automated policy descriptions to inform users about how their data will be used, but these mechanisms do not address location privacy. We theorize that it may be possible to enhance these methods to support location services based on some of the threat models that researchers have specified, but that broader concerns about location privacy can not rely on technical solutions and will need to rely on both education, good corporate citizenship and regulation.

1 Introduction

With the proliferation of personal wireless devices such as laptop computers, so-called "smart phones," and portable media players, a wide variety of applications and services have been developed for mobile users. These users can access the Internet from any place where they can obtain WiFi or cellular service, enabling immediate and ubiquitous access to e-mail, instant messaging communications, news, or any other traditional Internet-enabled applications. While personal wireless devices allow users to access these traditional applications as they move from place to place, they also drive an entirely new class of services that utilize users'

physical locations as input to applications that perform some function of utility with their location, i.e., location-aware or location-based services. For example, such location-aware applications include navigation and mapping tools (*e.g.* Google Maps), recommendation services that offer reviews of nearby businesses (*e.g.* Yelp), and location-based social networking services like Foursquare or Google Latitude that connect users with their nearby friends.

Traditionally, the location of your person has been assumed to be a "private matter" - part of the "right to be left alone" covered by the Warren-Bandeis articulation of privacy. The absolute disclosure of location is a penalty, inflicted on criminals either through physical constraint (as in prisons) or the more ephemeral penalty of ankle bracelets and tracking devices. When told that government officials have access to the location of their every move, most citizens would object, particularly if warrants for this information were not needed. The acquisition of location information from cellphones is common, particularly since such information is already collected by telecommunications companies. In 2009, US Attorney Chris Christie authorized such tracking [17], raising concern during the New Jersey gubernatorial campaign. The Department of Justice has responded that such disclosures are covered by the "pen register/trap and trace" statue and the stored-communications act [24] and stated that

> *"In enacting CALEA, Congress struck a balance in which although an individual may well prefer to keep his whereabouts confidential when he uses a cellular telephone, he has no right to privacy with respect to that information because he voluntarily discloses it to his service provider in the course of placing or receiving cellular calls."*

At the same time, recent court rulings indicate that warrant-less "GPS-tracking" violates the Fourth Amendment [11]. Would the common user of cellular be able to distinguish between one technology and another?

How do new technologies influence the expectations of location privacy? At the same time that existing technology provide surreptitious methods for monitoring location, millions of citizens are voluntarily disclosing their location, either explicitly or implicitly, using a new host of technologies. These technologies and services provide considerable benefit, but do

users have expectations of location privacy?

Can users of these systems determine for themselves when, how and to what extend information about them is communicated to others? We feel that existing systems provide a mixed record of such control, in part because of the accidental disclosure of information originally not felt to be identifying [12]. There are three "actors" in location privacy - governmental, corporate and private individuals. Location privacy for governmental agencies is currently being debated in court cases such as the "GPS-tagging" case [11]; the primary distinction is when location tracking becomes more than casual observance and beyond an extension of normal human sensory tracking (*e.g.* the distinction between automated tracking and watching someone on a "stake out") and when warrant-less tracking is authorized. Corporations and individuals have fewer regulations concerning invading the privacy of others. For all these cases, when users present information in a public forum, they release some of their right to privacy of that information. However, Solove's argument [33] that information disclosure is less about *1984* and more about Kafka's *Trial*, where information is used to make decisions rather than control, applies – people will want to control what is disclosed and to whom in order to maintain some identity and control over their life.

The primary concern of this paper - the accidental disclosure and aggregation of excessive information that abrogates the location privacy of individuals - is addressed through identifying direct disclosure, determining the possibility of "re-identification" using disparate data and providing means for consumers being informed when and where such information will be used. We assume that people have an interest in maintaining "location privacy" - knowledge of their every day motions and associations. We assume that the "threats" posed by loss of location privacy are more severe in the aggregate than the individual. Given sufficient resources, any one individual can be tracked or followed, but knowing the movement or location of a large number of people allows us to infer associations, group memberships and behaviors beyond the single actions of being in a given place at a given time. Such group disclosure is likely if individuals are unaware how information can be used and re-combined to "re-identify" individuals.

We start with describing the many ways in which location information is gathered. We then narrow our concern to how that information is *disseminated* rather than gathered. We believe that the largest "disclosure risk" for most location information is through social networking and commercial location-based services software. These systems often well articulated privacy policies for traditional personally identifying information, but have an opaque policy and controls for location-based information. Following this, we discuss how location data can be used to "re-identify" individuals and the steps that can be taken to thwart such re-identification. We close by asking how companies and individuals can cooperate to given users of such services better controls over their location privacy now.

2 Localization Background

There is no question that location-awareness can enable a vast array of interesting and practical services. However, to reap the benefits of such services, users must ostensibly reveal their locations to third-party applications, often without knowing how their location data will be handled or retained beyond their interaction with the service. The arbitrary disclosure of a user's physical location has potentially dangerous implications for the user's personal privacy. Before we can discuss the nature of these consequences in detail, we first offer a brief overview of the underlying technologies that enable location-aware services and applications.

There are many misconceptions of the how location information can be collected, how accurate that information is and what parties can access that information. These issues are important, because without regulation, the point of disclosure of information is the last point of control.

There are many "dimensions" to describing different location or localization technologies. What party receives the information? How accurate is the information? Does an individual need to take steps to determine location? Can the information be gathered without their knowledge? To date, most location based information is based on *data networks* and *telecommunication networks*. Increasingly, information is being gathered in other forms.

2.1 Location Determination

The Global Positioning System (GPS) is perhaps the most well-known and oldest technology for determining location. In short, GPS works by precisely timing signals transmitted by multiple in-orbit GPS satellites, calculating the distance to each satellite using simple speed-of-light calculations, and applying trilateration to estimate the receiver's approximate position. In order to receive GPS signals, it is necessary to use specialized GPS receiver hardware, which may be installed on select smart phones or other personal devices. While under typical conditions, GPS offers highly precise location estimates (with accuracy typically within one meter), it requires special GPS receiver hardware and may not reliably receive signals for devices that move indoors.

Many people believe that GPS "tracks" individuals; this perception is emphasized by both a lack of understanding of the underlying technology, media reporting and the combination of GPS "tracking devices" with radio transmitters, either in phones or "tagging" devices used by police.

Other devices, such as LoJack, actually "track" users. The LoJack system tracks stolen vehicles by having a small radio transmitter embedded in a vehicle rather than by using GPS information. When activated, the unit transmits radio signals that can be used to "triangulate" to the stolen vehicle. Such a signal doesn't require the cooperation of the user (who is presumably a thief if the system has been activated).

Other signals can be "triangulated" - this is the common method for providing location information for cellular phones. Phones can estimate their location using signal strength or time-of-flight readings from multiple cellular towers. The accuracy of such systems is fairly low, but can be enhanced with "dummy towers" that provide a additional signals for location. Cellular systems focus on meeting E911 requirements for 300 meter accuracy. Radio networks that have shorter range than cellular service allow more precise localization, even in the absence of GPS signals, because the limited communication range more precisely circumscribes the possible location. As 802.11 networks have become ubiquitous for providing convenient Internet connectivity at hot-spot locations, nearly all personal devices developed today have

integrated WiFi (or 802.11) hardware. Beyond providing Internet connectivity, the 802.11 wireless infrastructure can be leveraged to infer location information.

WiFi networks are typically identified by a human-readable network name, or *Service Set Identifier* (SSID) and also by a special unique hardware address called a *MAC address* to identify a physical access point. To announce the existence of a network, 802.11 access points broadcast their human-readable network name, MAC address, and other information as *beacon* messages to nearby wireless clients who may be interested in using the wireless service. The physical range of these broadcast beacon messages is typically less than a few hundred meters; thus, if a client can hear the beacon for a particular wireless network, then they can reasonably assume that the network is physically nearby. Also, since wireless networks tend to be stationary, it is possible to generate a mapping between a set of observed wireless networks and their physical locations. With the aid of an online database that stores these mappings, a wireless client can infer its approximate location by simply identifying the nearby wireless networks' SSIDs or MAC addresses and obtaining their respective physical locations as reported by the database. While this localization technique does not typically offer location estimation with the high precision of GPS, it does supply reliable estimates, typically within one city block of the user's true location. This granularity is more than sufficient to facilitate a wide variety of location-based services. This general method for localization was first developed by Place Lab [19] and has since been commercialized by SkyHook [1], which is now an integral localization feature of many personal wireless devices including the Apple iPhone, iPod Touch, and others.

GPS systems provide location information to the user of a device. Systems like LoJack provide location information to any users of a receiver, but aren't intended to provide precise location. Triangulation methods, whether cellular or based on system such as WiFi, provide location information to *both* the network operator and the user of a device. In almost all wireless networks, devices have a unique identification mechanism (similar to the 802.11 MAC address) that identifies a mobile device. Telecommunications companies maintain logs of the approximate location of cellular devices; this is the information in dispute in warrant-less

tracking cases. Local wireless networks, such as WiFi, can be used to "track" wireless devices, and companies such as Cisco sell WiFi location-tracking systems that use methods similar to cellular network triangulation; however, there is no current "federated" information repository to gather information from all WiFi access points to allow wide-area tracking of WiFi or other localized networks, although such systems have been proposed and analyzed [16].

Thus, in summary, there are two forms of location information generally available - the less accurate location information collected by cellular network providers and the more accurate information available to a device (possibly combining cellular, GPS and WiFi or other localization information). This latter information is what is typically used for "location based services" and other applications that use location. That location information is usually provided to an application or a website using a programming interface or to websites using a web standard such as HTML5.

There are a number of other "accidental" sources of location information. One example is the "metadata" associated with photographs; this information can be used to automatically organize photos by their location, but the information can also be extracted for other purposes. In a recent study, between 1-4% of photos sampled on site such as Flickr, Facebook and Craigslist included location information in photo metadata [12]. Other information includes "content" provided by a user. For example, someone posting to Twitter might indicate they are visiting a specific restaurant, and text matching could be used to infer a location.

3 Social Networks

This section discusses how location information is stored, used, and shared by Social Networks. Modern Social Networks have grown far beyond their original role of connecting users with each other into websites (such as Facebook) that model themselves (and their public APIs) as graphs linking anything and everything. This trend toward social networks as a sort of "Internet of Things" has made describing what happens within them a very complicated endeavor.

This transition from simple social website to a site "linking" all things "social," has coincided with an explosion in technologies and applications that integrate and rely on location information. The wide-spread adoption of smart phones, of particular note the iPhone and its use of SkyHook, has created a mobile development community which assumes access to location information and this community is integrating this information into applications of every kind, one of the most prominent examples being "social applications."

Location information is actively being gathered, used, and shared by social networks such as Facebook. The automatic linking and notification functionalities of social networks support the viral spread of social phenomena throughout these networks. This viral linking functionality is now spreading location information attached to a greater number and diversity of social phenomena (e.g., location tagged photos, tweets, comments, "check-ins," ads, etc...). This section attempts to detail how location information is and might be used in the context of social networks, emphasizing unique ways in which location information might be "leaked" or shared unintentionally or automatically.

3.1 Sources of location information

Social network user profiles often include general location information such as hometown, current city, or geographic networks to which the user is affiliated. In most cases this information is available as a string which must be understood to infer location. However, as Facebook has begun transitioning from user-entered text strings to explicit links between objects, these location strings have become standardized and linked to maps and information associated with the location. As such, this information has become very easy to integrate into applications and share in many different ways. Furthermore, Facebook considers basic location information such as hometown public by default along with basic preference information. This set of information often forms a unique quasi-identifier for users, Section 4 will discuss how mixing location information with other types of information may increase the "linkability" or ease of identifying an identity in an anonymous database using external information.

Less explicit forms of location are also scattered through social networks. Users include

location references in comments, tweets, photos, and videos. This information comes in many different forms requiring different levels of sophistication to be automatically understood. For instance, a user may comment as to where they are eating dinner one evening - the name of a restaurant and city of residence are then sufficient to locate the user around dinner time. A more sophisticated example would be a user being tagged in a photo in which the photo background implies a location and possibly even a time at which the user was located there (such as a picture of user with the statue of liberty and fireworks in the background). This type of location inference presents a unique challenge for defending against linkability and will be discussed in Section 4.

Many social network items or items linked to from social network users are time-stamped and/or tagged with location information. This information is often very precise and can give the exact time and location coordinates associated with the item. Examples of such tagged items include tweets, photos, events, check-ins (mobile applications), and most activities associated with certain types of augmented reality mobile applications. In fact, there is an increasing trend toward time and location stamping across many different types of objects and applications. In particular, mobile applications are leading the way allowing for automatic location and time-stamping of nearly anything. Furthermore, social networks are increasing support for mobile uploading of these location/time-stamped items, often automatically linking the items with a users profile and identity.

Finally, the prevalence of third party applications using social network data means that assumptions cannot be made about the limitation of how location information might be used or linked to information outside the social network. Facebook's recent redefinition of preferences as "links" led to preferences along with general user information being considered by default public.[1] This expansion of users' public information set has been accompanied by an additional clause in Facebook's privacy policy which refers to automatic sharing of "general" information with "pre-approved third parties." Allowance for automatic sharing of preference and location information with third-parties has incentive motivation for Facebook as it may

[1]Facebook refers to this public information set as a user's PAI or Publicly Available Information

increase the value of Facebook's primary capital, user information.

This section has largely focused on online social networks and Facebook in particular due to the trend in social networks linking of everything. However, other types of social sites and applications such as Twitter, Gowalla, Foursquare, Yelp, Meetup, and Brightkite are growing sources of location information and increasingly linking this information to their users and all things associated with those users. Twitter has added location tagging to tweets. Foursquare, Gowalla, and Brightkite are inherently location-based, integrating location information with all messages and actions made by their users. Online recommendation sites such as Yelp and event sites such as Meetup have also integrated location information and Yelp has a popular mobile application which automatically uses location information to support location-based recommendations.

3.2 How location information is accessed and shared

Traditionally, social networks were accessed through a web browser. However, social networking now takes many different forms as third-party applications that access social network information through a web interface. Often these applications use the social networking information in a very different way than the traditional social networking behavior of browsing user profiles. A growing percentage of social network information now passes through a web API to a third-party before reaching the social network user. It is through these third-party applications that location information often enters the social network, gathered automatically through any of the technologies discussed in Section 2.

Social networks often include rules as to the usage of data accessed from their API in the terms of service agreement with the third-party application developer. However, when one considers the magnitude of data and its usage along with its growth rate - management of how such data is used seems improbable. For example, Facebook is the largest provider and linker of social network information. There are over half a billion Facebook users (roughly one out of every two people on Earth that use the Internet). These half-a-billion users share more than 30 billion pieces of content every month (notes, photos, etc...) which are "linked"

to over 900 million non user objects or entities (groups, events, communities, etc.). These 30 billion pieces of content and their associated "links" are accessed by over one million external websites (outside of Facebook) and half a million applications every month. These applications may be running on home desktop computers, mobile phones, or even foreign server farms performing massive data mining. This is equivalent to creating, indexing, and sharing the entire library of congress across the Internet every two hours. Obviously, traditional expectations for monitoring and responding to violations of Facebook's terms of service are not practical at any granular level.

Considering how location information is integrated with social networking information whether embedded in the semantics of a message or tagged in the meta-data of a picture, location information is often automatically generated and flows along with the information through the social networks API. The trend is toward location information being integrated into a greater number and diversity of content items and this trend largely driven by the types of applications that are using social network data. The next section discusses how these applications use social network information.

3.3 How location information is used

The most basic location information shared through social networks, such as hometown or residence, is shared and accessed through the users profile as a string. Other users can view this information when viewing their friends profiles. This usage model is what most people think of when they think of people accessing their social network information. While there are privacy issues associated with this basic location information on users profiles, this section focuses on how location information that is tied to social network photos, Twitter tweets and other content items might be used.

Many photo applications and sites such as Flickr support location tagging. Applications such as Flickr's map application allows users to view photos on a map showing the location (and possibly time) that a photo was taken. Applications such as Bing Map's Streetside Photos take this a step further, using image processing the application can map location tagged photos

onto 3-dimensional representations of cities throughout the world, together with timestamps these temporally and spatially mapped photos are beginning to piece together a detailed historical catalog of everything that happens near a camera, whether that camera be on a phone or a stop light. These photos are often automatically tagged and integrated (linked) into social applications connecting them with the social networks vast web of interconnecting items, people, places, and events.

4 Linkability & Re-identification Attacks

Large amounts of information regarding individuals are being captured and shared by service providers, organizations, and government agencies for various purposes such as advertisement, scientific research, or national security. Although such data are typically "anonymized" or "de-identified" (i.e., removing personally identifiable information such as social security number) before being released, it has been demonstrated in many scenarios that re-identification is still possible (and surprisingly easy) in such anonymized data sets, linking data records back to individual identities and leaking sensitive personal information. For example, the study by Sweeney using 1990 US census data reported that 87% of the US population can be uniquely identified via the combination of gender, date of birth, and 5-digit zip code information [34]. In 2006, AOL published a data set containing Web searches by anonymized individuals, and the content of the searches allowed others to identify the specific individuals who made those searches. In 2007, using the anonymized user movie ratings published by Netflix and very little auxiliary information (e.g., public user-posted movie ratings at the Internet Movie Database (IMDb)), Narayanan and Shmatikov were able to identify specific users in the Netflix anonymized data set [27]. As the authors pointed out in their recent article [28]: "*Any information that distinguishes one person from another can be used for re-identifying anonymous data*." Ohm offers a detailed analysis of these and other high profile re-identification cases [29] – his analysis focused on an adversarial model where identifying information is intentionally anonymized and the failures of such methods. He doesn't address

the problem facing many people – the accidental self-disclosure of information with no thought to anonymization because of a lack of concern about the use of that information at the time of posting. Location information, which is closely related to individual person's behavior and activities, makes the situation even worse as large amounts of personalized precise location information become increasingly available through online social networks and location-based services.

In this section, we first explain the key terminologies and challenges with regard to data anonymization, linkability, and re-identifiably. We then describe techniques and general practices to defend against re-identification attacks. We also highlight the limitations of current approaches and discuss directions for further improvements.

4.1 What is "Linkability" or "Anonymity"?

Given a data set containing personal records, such as census data, medical records, or Web search/browsing activities, each person in the data set may be associated with a set of attributes. Some attributes are *explicit identifiers*, which can uniquely identify individual users. Such explicit identifiers include social security numbers, driver's license numbers, address, etc. Other attributes, when used in combination, can potentially identify unique users. For example, a large percentage of the US population can be uniquely identified by the combination of gender, date of birth, and 5-digit zip code [13, 34]. Such attributes are referred to as *quasi identifiers*. There are also *sensitive attributes* such as income or disease that users do not wish to disclose or be associated with. Therefore, when anonymizing a data set, the goal is to ensure that individual users cannot be re-identified in the anonymous data set, nor can they be associated with sensitive attributes.

The problem of "linkability" or how "linkable" an item is considers how much effort it takes to link a piece or set of information to its source or associated identity. The flip-side of this problem is that of measuring "linkability" by measuring the ambiguity or commonness of a piece of data in light of a particular set of known facts. For instance learning that a person with blond hair exists in Sweden does not tell you much about the identity of any particular

person in the world – having blond hair does not imply that one exists in Sweden nor that existing in Sweden implies that you have blond hair. However, if only one or a few people in the world had blue hair, and one was to learn that a person with blue hair fled the scene of a particular crime, it gives the learner of this information a very small set of identities which may be investigated as to whether or not they where the ones that fled the scene of the crime. While an example of someone's blue hair showing up in a database may seem only hypothetical, we must not assume that data in actual public releases of data is any different. In fact, it is wrong assumptions or lack of consideration that has lead to many real-world privacy violations involving anonymized databases.

The previous example of how a database containing a very common attribute value (blond hair in Sweden) and a very unique attribute value (blue hair) highlights the importance of not including very-unique quasi-identifiable attribute values in anonymized databases (those values that are not intended as unique identifiers but none the less may uniquely identify a person). However, removing these unique values or guaranteeing that many copies or shared value entries exist in a database does not solve this problem completely. For instance, guaranteeing that Swedish databases only include blond hair attribute values by removing blue hair attribute values presents a new problem if having blond hair is something people prefer to keep private. Consider the case of a hospital releasing anonymized medical records on its cancer patients, guaranteeing that all cancer cases referred to in the database are not too unique to identify the user uniquely. This anonymization may result in only a few types of cancer cases being included in the database. This presents a new kind of problem – whereas in the case of blue hair the attribute value was so unique it made re-identification very easy, in this case the regularity and uniformity of attribute values presents the problem. If a person's medical record was known to be in the database due to some external information such as the fact that they participated in a study on a certain date, it would imply that the person must be matched to some attribute value in the database. In the case of the cancer patient medical records, a privacy violation may occur without matching any patient to their particular record since simply being associated with the database is enough to associate the patient with at least some kind

of cancer, in this case a relatively few types one of which is associated with every user in this database.

To measure the linkability or anonymity of a data set that has been de-identified and anonymized, a formal definition of "k-anonymity" has been proposed [31, 35, 36]. An anonymous data set satisfies the k-anonymity property if each person in the data set cannot be distinguished from at least k-1 other individuals in the same data set. In other words, these k or more individuals are indistinguishable within this group (the *anonymity set*). Generally, a larger k means bigger anonymity sets and lower linkability, therefore better privacy protection. While k-anonymity prevents the disclosure of user identity, it does not protect sensitive attribute from being released. For instance, if all k individuals have the same disease, such sensitive information may be learned by an adversary, even though the adversary cannot identify the individuals in this anonymity set. To address this issue, new metrics such as "l-diversity" and "t-closeness" have also been proposed [21, 22]. The former considers the diversity of attribute values within each anonymity set, while the latter ensures that the overall data set and the anonymity set have similar distributions of sensitive attribute values.

Location information is now integrated (tagged) with so many different types of information that the linkability of location information depends on the linkability of the associated data. When considering "anonymous" data and location information one must focus on the accuracy of the location information itself, the linkability of the information to which the location information is associated, and to answer the linkability of any one object, we must be able to understand all that is implied by the content of that object. These three questions imply their own unique types of anonymity measures and problems. As discussed in Section 2, SkyHook is releasing "anonymous" information about how many people were in particular locations at particular times. When information such as this is released one might not only ask the question of "whether or not" the information is anonymous, but rather "how" or "to what degree" the information is anonymous. For example, the study by Golle and Partridge [14] shows that learning the home and work location pairs of an individual from a location trace can significantly reduce the size of the individual's anonymity set, making it much easier to

re-identify individual in anonymous data set. Specifically, the median anonymity set size in the U.S. working population is only 1, 21, and 34,980 for locations at the granularity of census block, census tract and county, respectively. If an individual works and lives in different regions, re-identifying this individual is even easier.

These results are surprising for many people and make it difficult for users of location based services to correctly assess any "threat" based on information disclosure. The obvious solution is for location based services to simply not collect or retain identifying information; in practice, this may not be possible because users perceive value in that associated information.

4.2 How to defend against Re-identification attacks

Various types of re-identification attacks have been identified. Earlier approaches exploit quasi-identifiers, which are sets of attributes, each set when used in combination can uniquely identify individual persons. Other approaches supplement a given anonymous data set with other data sets (either anonymous or public identified) and certain auxiliary information that may be available from the same data source or other data sources. For example, a de-identified medical records may be coupled with a de-identified DNA database or with a public identified voter registration list. Multiple randomized copies of the same data source may also be obtained and more accurate estimates of the anonymized values may be obtained. As shown by Cassa et al., spatial locations anonymized by a non-deterministic blurring algorithm may not prevent re-identification of home addresses [4]. If multiple anonymized copies may be obtained from the same original data set, each de-identified using a Gaussian skew that shifts the geocoded values, the error of the estimated home addresses can be reduced from 0.7km to 0.2km with 10 copies, and to 0.1km with 50 copies. This is illustrated schematically in Figure 1 – the multiple overlaping circles represent "anonymized" locations for the "true" location indicated by the black dot. With a single anonymized location, there is little opportunity to identify the true location; with an increasing record of reported anonymized locations, the true location clearly becomes the intersection of the reported regions.

In addition, when multiple anonymized data sets can be obtained from different locations

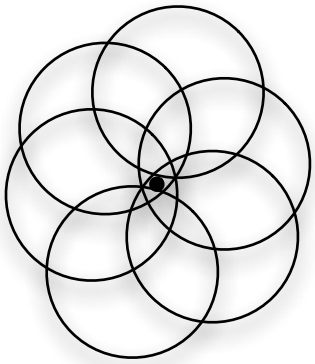

Figure 1: Sufficient overlapping "anonymized" locations can indicate true location

(e.g., medical records from multiple hospitals), such trail-based location patterns lead to another type of trail re-identification [23], making privacy protection even more challenging.

To robustly defend against re-identification attacks, in addition to removing explicit identifiers from the data, further processing and anonymization is needed. Common techniques include generalization, suppression, aggregation, non-deterministic blurring, etc. Given the threat of de-anonymization with ancillary information or side channels, these ad-hoc data perturbation techniques have been formalized into a new definition of data privacy known as *differential privacy* [6–10]. Differential privacy offers a solution to the statistical disclosure problem, or the ability to reveal accurate statistical information about a data set in a manner such that it still has utility for analytical purposes but without revealing any private or identifiable information about individual records contained within the data set. This framework has the promise to offer stronger privacy guarantees for PII data (including location) that is retained by third-party service providers than has been possible to date.

One important question is how much location obfuscation (degradation of location quality) is required in order to achieve a certain level of anonymity. The study by Golle and Partridge on the anonymity of home and work location pairs provides useful information with regard to this question. They considered the size of the anonymity set at different location granular-

ities, including census block, census tract, and county [14]. Other related issues in location privacy protection includes user privacy controls, spatial and temporal cloaking, and hybrid location obfuscation mechanisms [2, 15, 20, 26, 32]. Another type of location obfuscation was proposed by Hoh et al., which uses a "time-to-confusion" metric and periodically withholds location information such that a location trace may be confused with many others [18]. Query-based location services that protect location privacy have also been proposed [5, 25]. In these designs, a location anonymizer first blurs the location information using its own obfuscation mechanism; the anonymized location data can then be queried by a privacy-aware processor.

Location information comes in many different forms and granularities. As discussed in Section 3, location information may consist of latitude and longitude coordinates, regions (e.g., census block, county) or may take on many more complicated forms such as being inferred from a photo or message between social network users. When location is represented as coordinates it can be generalized or "fuzzed" by reducing its accuracy, this reduces the uniqueness or specificity of the location. In some cases, this kind of generalization reduces the utility of the location information as well. When location information is represented as the name of a place or an address, the location can be generalized to refer to a more general place. For instance, a reference to one's hometown could be generalized to refer to the state or country of origin, but not the city. These types of anonymization or anonymity measurement are generally simple and straightforward. Inferred location, whether from a message or photo is more complicated. In the cases in which location was inferred automatically using computer algorithms, those same computer algorithms can often also be used to measure the ambiguity of location information in content items. In cases in which human intelligence is used to infer location information, computer algorithms cannot necessarily be relied upon to fully measure the location information contained in a content item. For these types of location inferences, it is very hard to make any clear guarantees about the anonymity of a content item - however, it should be noted that while certain information may be too complicated, subtle, or unique to human senses to support computer guarantees on its anonymity, the same limitations apply to any computer automated re-identification or "re-linking" attack. As such, when a data

provider refers to a piece of content as "anonymous" one should consider whether the data provider means the content: does not have any unique identifiers such as social security numbers; is un-linkable by any computer algorithm; is un-linkable in the face of any intelligence including any human's.

5 Policy Implications

Location privacy threats can be broader divided into three classes of threats, according to the actors involved.

The first class involves unwanted government surveillance; in practice, it is difficult for individual users to take actions that counter this threat other than through regulation or legislation because government entities can *e.g.* subpoena cellular tracking (and other) information that is not under the control of technology users.

The second class are personal threats due to disclosure of location information. This may take the form of having location exposed to abusive or violent individuals [3] or provide socially awkward interactions. This disclosure is akin to revealing other identifying information, such as bank account numbers or the like. Privacy in these cases is typically mediated through social networking and other venues where "location" is typically treated in a manner similar to other information. It is likely that aggregation methods, such as reporting only a town, state or region rather than a specific location, may suffice for this threat model because the user can judge the threat and measure the response to that threat. However, as detailed by Boyd [3], changes in privacy policies need to be transparent and clearly communicated to users prior to their implementation.

However, the threat model from commercial use of information is likely because commercial applications typically have repeated queries (limiting the effectiveness of k-anonymity methods). Porter [30] has discussed the legal implications of anonymized data and the risk of re-identification of anonymous data, as well as the consequences of violating privacy policies. Commercial use of location information would best be served by retaining information for as

short a period of time as possible. Given that k-anonymity metrics and similar mechanisms for cloaking data have typically been susceptible to re-identification, simply aggregating data is unlikely to provide sufficient consumer protection. This is particularly true since consumers are unlikely to know the k in the k-anonymity – without reliable estimates of the number of unique people in an area using a service, it is difficult to estimate the what level of granularity is needed. As described, possible solutions to this involve "broker" architectures.

The most promising privacy outcome is adopting modern anonymization strategies such as differential privacy [7], coupled with the use of privacy policy mechanisms such as P3P. However, the existing use of systems such as P3P has shown that intent rarely matches actions – most policy specifications are insufficient or incomplete. Tools such as differential privacy mechanism can not be implemented by an end-user – they require the retention of aggregate information in order to be useful. We believe this will require the adoption of "broker architectures" that can provide location information on behalf of users.

6 Summary & Conclusion

This paper has discussed how new technologies have made the risks and regulation of location information more urgent. This paper discusses how new wireless technologies and location services provide location information to mobile phones, laptops, or social networks. The difficulty and unique nature of anonymizing social network data is discussed. The size of social networks, the scattered and highly interlinked nature of their data, and the unique nature of much social network data makes traditional regulation expectations rather impractical. These difficulties emphasize the importance and role of strong access control and non-technological protection methods. Furthermore, policies related to anonymization and public release of private data must be aware of the complications involved in anonymizing not only location, but all data that may be linked to location.

References

[1] Skyhook wireless. http://www.skyhookwireless.com.

[2] C. A. Ardagna, M. Cremonini, E. Damiani, S. De Capitani di Vimercati, and P. Samarati. Location privacy protection through obfuscation-based techniques. In *Proceedings of the 21st annual IFIP WG 11.3 working conference on Data and applications security*, pages 47–60, Berlin, Heidelberg, 2007. Springer-Verlag.

[3] Danah Boyd. Making sense of privacy and publicity. Keynote address, SXSW Conference, Austin, Tx, March 2010. Transcript available as `http://www.danah.org/papers/talks/2010/SXSW2010.html`.

[4] Christopher A Cassa, Shannon C Wieland, and Kenneth D Mandl. Re-identification of home addresses from spatial locations anonymized by gaussian skew. *Int J Health Geogr*, 7(45), 2008.

[5] Chi-Yin Chow, Mohamed F. Mokbel, and Walid G. Aref. Casper*: Query processing for location services without compromising privacy. *ACM Trans. Database Syst.*, 34(4):1–48, 2009.

[6] Cynthia Dwork. Differential privacy. In *ICALP*, pages 1–12. Springer, 2006.

[7] Cynthia Dwork. Differential privacy: A survey of results. In Manindra Agrawal, Ding-Zhu Du, Zhenhua Duan, and Angsheng Li, editors, *TAMC*, volume 4978 of *Lecture Notes in Computer Science*, pages 1–19. Springer, 2008.

[8] Cynthia Dwork. A firm foundation for private data analysis. *Commun. ACM (to appear)*, 2010.

[9] Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In Michael Mitzenmacher, editor, *STOC*, pages 371–380. ACM, 2009.

[10] Cynthia Dwork, Moni Naor, Toniann Pitassi, and Guy N. Rothblum. Differential privacy under continual observation. In Leonard J. Schulman, editor, *STOC*, pages 715–724. ACM, 2010.

[11] Electronic Frontier Foundation. Court rejects warrantless gps tracking. `https://www.eff.org/press/archives/2010/08/06-0`, August 2010.

[12] G. Friedland and R. Sommer. Cybercasing the joint: On the privacy implications of geotagging. In *Proceedings of Fifth USENIX Workshop on Hot Topics in Security (HotSec 10)*, 2010.

[13] Philippe Golle. Revisiting the uniqueness of simple demographics in the us population. In *WPES '06: Proceedings of the 5th ACM workshop on Privacy in electronic society*, pages 77–80, 2006.

[14] Philippe Golle and Kurt Partridge. On the anonymity of home/work location pairs. In *Pervasive '09: Proceedings of the 7th International Conference on Pervasive Computing*, pages 390–397, 2009.

[15] Marco Gruteser and Dirk Grunwald. Anonymous usage of location-based services through spatial and temporal cloaking. In *MobiSys '03: Proceedings of the 1st international conference on Mobile systems, applications and services*, pages 31–42, New York, NY, USA, 2003. ACM.

[16] Marco Gruteser and Dirk Grunwald. Enhancing location privacy in wireless lan through disposable interface identifiers: a quantitative analysis. *Mob. Netw. Appl.*, 10(3):315–325, 2005.

[17] Claire Heininger. Aclu says chris christie authroized warrantless cellphone tracking. `http://www.nj.com/news/index.ssf/2009/04/aclu_says_chris_christie_autho.html`, Apr 2009.

[18] B. Hoh, M. Gruteser, H. Xiong, and A. Alrabady. Preserving privacy in gps traces via density-aware path cloaking. In *Proc. of ACM Conference on Computer and Communications Security (CCS)*, 2007.

[19] Anthony LaMarca, Yatin Chawathe, Sunny Consolvo, effrey Hightower, Ian Smith, James Scott, Tim Sohn, James Howard, Jeff Hughes, Fred Potter, Jason Tabert, Pauline Powledge, Gaetano Borriello, and Bill Schilit. Place lab: Device positioning using radio beacons in the wild. In *Proceedings of the Third International Conference on Pervasive Computing*, Lecture Notes in Computer Science. Springer-Verlag, May 2005.

[20] Mingyan Li, Krishna Sampigethaya, Leping Huang, and Radha Poovendran. Swing & swap: user-centric approaches towards maximizing location privacy. In *WPES '06: Proceedings of the 5th ACM workshop on Privacy in electronic society*, pages 19–28, New York, NY, USA, 2006. ACM.

[21] Ninghui Li and Tiancheng Li. t-closeness: Privacy beyond k-anonymity and -diversity. In *In Proc. of IEEE 23rd Intl Conf. on Data Engineering (ICDE07)*, pages 106–115, 2007.

[22] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data*, 1(1):3, 2007.

[23] Bradley Malin. A computational model to protect patient data from location-based re-identification. *Artif. Intell. Med.*, 40(3):223–239, 2007.

[24] Roslynn Mauskoph. In re application for pen register and trap and trace device with cell site location authority. `http://www.eff.org/files/filenode/celltracking/celltracking_govt_reply.pdf`, Oct 2005.

[25] Mohamed F. Mokbel, Chi-Yin Chow, and Walid G. Aref. The new casper: query processing for location services without compromising privacy. In *VLDB '06: Proceedings of the 32nd international conference on Very large data bases*, pages 763–774. VLDB Endowment, 2006.

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

[27] A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In *Proceedings of the 2008 IEEE Symposium on Security and Privacy*.

[28] A. Narayanan and V. Shmatikov. Myths and fallacies of "personally identifiable information". *Communications of the ACM (CACM)*, 53(6), 2010.

[29] Pau Ohm. Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. *SSRN eLibrary*, 2009.

[30] C. Christine Porter. De-identified data and third party data mining: The risk of re-identification of personal information. *Shidler Journal of Law, Commerce + Technology*, 5(3), September 2008.

[31] P. Samarati. Protecting respondents' identities in microdata release. *IEEE Trans. on Knowl. and Data Eng.*, 13(6):1010–1027, 2001.

[32] Mohsen Sharifi and Leila Naghavian. Providing location privacy in pervasive computing through a hybrid mechanism. *Int. J. Internet Technol. Secur. Syst.*, 2(1/2):160–173, 2010.

[33] Daniel J. Solove. 'I've Got Nothing to Hide' and Other Misunderstandings of Privacy. *San Diego Law Review, Vol. 44, p. 745, 2007*.

[34] L. Sweeney. Uniqueness of simple demographics in the u.s. population. LIDAPWP4. Carnegie Mellon University, Laboratory for International Data Privacy, Pittsburgh, PA, 2000.

[35] L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal on uncertainty, Fuzziness, and Knowledge-Based Systems*, 10(5):571–588, 2002.

[36] L. Sweeney. k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5):557–570, 2002.