Cybercasing the Joint: On the Privacy Implications of Geo-Tagging

Gerald Friedland [†] and Robin Sommer *[†]

TR-10-005

May 3, 2010

Abstract

This article aims to raise awareness of a rapidly emerging privacy threat that we term *cybercasing*: leveraging geo-tagged information available online to mount real-world attacks. While users typically realize that sharing locations has some implications for their privacy, we provide evidence that many *(i)* are unaware of the full scope of the threat they face when doing so, and *(ii)* often do not even realize *when* they publish such information. The threat is elevated by recent developments that make systematic search for geo-located data and inference from multiple sources easier than ever before. In this paper, we summarize the state of geo-tagging; estimate the amount of geo-information available on several major sites, including YouTube, Twitter, and Craigslist; and examine its programmatic accessibility through public APIs. We then present a set of scenarios demonstrating how easy it is to correlate geo-tagged data with corresponding publicly-available information for compromising a victim's privacy. We were, e.g., able to find private addresses of celebrities as well as the origins of otherwise anonymized Craigslist postings. We argue that our community needs to shape the further course of geo-location technology for better protecting users from such consequences.

† International Computer Science Institute, Berkeley, California
* Lawrence Berkeley National Laboratory, Berkeley, California

1 Introduction

Location-based services are rapidly gaining traction in the online world. With the big players heavily invested in the space already, it is not surprising that GPS and WIFI triangulation are becoming standard functionality for mobile devices: starting with Apple's iPhone, all the major smartphone makers are now offering models allowing to instantaneously upload *geo-tagged* photos, videos, and even text messages to sites such Flickr, YouTube, and Twitter. Likewise, numerous start-ups are basing their business models on the expectation that users will install applications on their mobiles continuously reporting their current location to company servers.

Clearly, many users realize that sharing location information has implications for their privacy, and thus device makers and online services typically offer different levels of protection for controlling whether, and sometimes with whom, one wants to share this knowledge. Sites like *pleaserobme.com* (see § 2) have started campaigns to raise the awareness of privacy issues caused by intentionally publishing location data. Often, however, users do not even *realize* that their files contain location information. For example, Apple's iPhone embeds high-precision geo-coordinates with all photos and videos taken with the internal camera unless explicitly switched off in the global settings. Their accuracy even exceeds that of GPS as the device determines its position in combination with cell tower triangulation and thus regularly reaches resolutions of +/- 1 m in good conditions and still postal-address accuracy indoors.

More crucial, however, is to realize that publishing geo-location (knowingly or not) is only *one* part of the problem. The threat is elevated to a new level by the *combination* of three related recent developments: *(i)* the sheer amount of images and videos online that make even a small relative percentage of location data sufficient for mounting systematic privacy attacks; *(ii)* the availability of large-scale easy-to-use *location-based search* capabilities, enabling everyone to sift through large volumes of geo-tagged data without much effort; and *(iii)* the availability of so many *other* location-based services and annotated maps, including Google's Street View, allowing to correlate findings across diverse independent sources.

In this article, we present several scenarios demonstrating the surprising power of combining publicly available geo-information resources for what we term *cybercasing*. According to the Urban Dictionary, *to case a joint* is "to check out the details to, and make speculations about, a home, car, store or other location for means of getting in undetected and/or removing material from said location." Consequently, we define *cybercasing* as to use online tools to check out the details, infer from related data, and make speculations about a location in the real world for questionable purposes.

The primary objective of this paper is to raise our community's awareness to the scope of the problem at a time when we still have an opportunity to steer the further course. While geo-tagging clearly has the potential for enabling a new generation of highly useful personalized services, we deem it crucial to discuss an appropriate trade-off between the benefits of location-awareness and the protection of everybody's privacy.

We structure our discussion as follows. We begin with briefly reviewing geo-location technology and related work in § 2. In § 3 we examine the degree to which we already find geo-tagged data "in the wild". In § 4 we demonstrate the privacy implications of combining geotags with other services using a set of example cybercasing scenarios. We discuss preliminary thoughts on mitigating privacy implications in § 5 and conclude in § 6.

2 Geo-Tagging Today

We start the discussion by introducing the technological background of geo-tagging along with related work in locational privacy.

Geo-location Services. An extensive and rapidly growing set of online services is collecting, providing, and analyzing geo-information. Besides major players like Google and Yahoo!, there are many smaller startups in the space as well. The main driving force behind these services is the enabling of a very personalized experience. *Foursquare* for example encourages its users to constantly "check-in" their current position, which they then propagate on to friends; *Yowza!!* provides an iPhone application that automatically locates discount coupons for stores in the user's current geographical area; and *SimpleGeo* aims at being a one-stop aggregator for location data, making it particularly easy for others to find and combine information from different sources.

In a parallel development, a growing number of sites now provide public APIs for structured access to their content, and many of these already come with geolocation functionality. Flickr, YouTube, and Twitter all allow queries for results originating at a certain location. Many 3rd-parties provide services on top of these APIs. *PicFog* for example provides real-time location-based search of images posted on Twitter. Such APIs have also already been used for research purposes, in particular for automated content analysis such as work by Crandall et al. [1], who crawled Flickr for automatically identifying landmarks.

Locational Privacy. Location-based services take different approaches to privacy. While it is common to provide users with a choice of privacy settings, the

set of options as well as their defaults tend to differ. YouTube for example extracts geo-information from uploaded videos per default, while Flickr requires explicit opt-in. Likewise, defaults differ across devices: Apple's iPhone geo-tags all photos/videos taken with the internal camera unless specifically disabled; with Android-based phones, the user needs to turn that functionality on.

The privacy implications of recording locations have seen attention particularly in the blogosphere. However, these discussions are mostly anecdotally and rarely consider how the ease of searching and correlating information elevates the risk. From a more general perspective, the *EFF* published a thoughtful white-paper on *Locational Privacy* [2], discussing implications of secretly recording peoples' activity in public spaces.

`pleaserobme.com` [8] was probably the first effort that demonstrated the malicious potential of systematic location-based search: the authors leveraged Foursquare's "check-ins" to identify users who are currently not at their homes. However, here locations were deliberately provided by the potential victims, rather than being implicitly attached to files they upload. Also, geotagging is a hazard to a much wider audience, as even people who consciously opt-out of any reporting of their geo-location might still be a victim when their location becomes public through a third party publishing photos or videos, e.g., from a private party.

To mitigate the privacy implications, researchers have started to apply privacy-preserving approaches from the cryptographic community to geo-information. Zhong et al. [3] present protocols for securely learning a friend's position if and only if that person is actually nearby, and without any service provider needing to be aware of the users' locations. Olumofin et al. [6] discuss a technique that allows a user to retrieve point-of-interest information from a database server without needing to disclose the exact location. Poolsappasit et al. [7] present a system for location-based services that allows to specify and enforce specific privacy preferences. From a different perspective, Krumm [4] offers a survey of ways that computation can be used to both protect and compromise geometrical location data.

GPS and WIFI Triangulation. Currently, it is mostly high-end cameras that either come with GPS functionality or allow separate GPS receivers to be inserted into their Flash connector or the so-called "hot shoe". Likewise, it is mostly the high-end smartphones today that have GPS built in, including the iPhone, Android-based devices, and the newer Nokia N-series. An alternative (or additional) method for determining the current location is WIFI-hotspot or cell-tower triangulation: correlating signal strengths with known locations allows to compute a device's coordinates with high precision, as we demonstrate in § 4. If a device does not directly geo-tag media itself, such information can also be added in post-processing, either by correlating recorded timestamps with a corresponding log from a hand-held GPS receiver; or manually using a map or mapping software.

Metadata. For our discussion, we are primarily concerned with *geo-tagging*, i.e., the process of adding location information to documents later uploaded online. The main motivation behind geo-tagging is the capability for personalized organizing and searching. For example, by including current location and time with a series of vacation photos, it becomes easy to later group them automatically, as well as to find further photos online from others who visited the same place.

The most common mechanism to associate locations with photos are *EXIF* records, which were originally introduced by the *Japan Electronic Industry Development Association* for attaching metadata to images such as exposure time and color space. Since then EXIF has been extended to also cover geographical coordinates in the form of latitude and longitude. Currently, EXIF is used only with JPEG & TIFF (image) and WAV (audio) files. However, most other multimedia formats can contain metadata as well, often including geo-tags. In addition, most camera manufacturers specify proprietary metadata formats. For videos, these "maker notes" are the most common form for storing locations. We note that all these formats are easy to parse with the help of standard tools and libraries, including browser plugins for revealing metadata and Apple's *Preview* program that offers a convenient *Locate* button for geo-tags taking one directly to Google Maps/Street View.

As metadata is however typically not immediately apparent when opening a document, it can pose a privacy risk by storing unexpected information. Murdoch and Dornseif [5] demonstrated that editing software may leave thumbnails of the original image behind in EXIF data, and similar problems have been found with other formats, such as Word and PDF documents.

It turns out, however, that many automatic image manipulation tools—especially those used in content management systems—"accidentally" discard metadata during processing. For that reason, we find that images on many of the larger Web sites do *not* have any metadata attached. As we discuss in § 4, we did however find EXIF data (and thus also locations) on private homepages and blogs as well as on sites such as *Craigslist* and *TwitPic*.

3 Prevalence of Geo-Tagged Data

In this section a number of experiments are presented that aim at understanding to which degree image and video data is geo-tagged today.

Flickr. Both Flickr and YouTube have comprehensively integrated geo-location into their infrastructure,

and they provide powerful APIs for localized queries. Leveraging these APIs, we can estimate the number of public geo-tagged photos/videos they offer.

Flickr's API allows to directly query for the number of images that are, or are not, geo-tagged during a certain time interval. Examining all 158 million images uploaded during the first four months of 2010, we found that about 4.3% are geo-tagged. When looking at the development over the past years, we see—somewhat counter-intuitively—a declining trend: while there was steep increase in geo-tagging from 2004 to about the end of 2006 (when it peaked at 9.3%), its share has been declining to the current level since then. We speculate that Flickr's opt-in privacy policy is the reason for this development: as more users are joining Flickr, the number of those explicitly enabling EXIF import is likely shrinking, and thus the number of images not geo-tagged is rising much more quickly than those which do. This would then indicate that such an opt-in policy is indeed a suitable mechanism to contain the amount of geo-tagging.

We also examined the brands of cameras used for taking the photos that have geo-information, derived from their EXIF records which can be retrieved via Flickr's API as well. Doing so however requires one API request per image, and hence we resorted to randomly sampling a 5% set of all geo-tagged images uploaded in 2010. We found that the top-five brands were Canon (31%), Nikon (20%), Apple (6%), Sony (6%), and Panasonic (5%). A closer look at the individual models confirms our observation in § 2 that today mostly devices at the higher end of the price scale are geo-tagging.

YouTube. With YouTube, due to restrictions of the API, it is not possible to directly determine the number of geo-tagged videos, as we could with Flickr. YouTube restricts the maximum number of responses per query to 1,000; and while it also returns an (estimated) number of total results, that figure is also capped at 1,000,000. Furthermore, the granularity for time-based queries is coarse: YouTube only allows to specify the attributes `all_time`, `this_month`, `this_week`, and `today`. Still, we believe we can estimate the number of geo-tagged videos in the following way: We submitted an unconstrained query, which results in an estimation of 1,000,000 results. The query was then refined by filtering for all videos that contain geo-location. Repeating the experiment a number of times resulted in total result estimates ranging from about 30,000 to 33,000 videos. In other words, out of what we assume to be a random sample of 1,000,000 YouTube videos, roughly 3 % have geo-location. While this number is clearly just an estimate, it matches with what we derived for Flickr. [1]

[1] YouTube's API distinguishes between videos *without* location, with *coarse* location (usually manually added, e.g. "Berlin"), and with *exact* location. For our experiments, we only considered the latter.

#	*Model*	#	*Model*
414	iPhone 3G	6	Canon PowerShot SD780
287	iPhone 3GS	3	MB200
98	iPhone	2	LG LOTUS
32	Droid	2	HERO200
26	SGH-T929	2	BlackBerry 9530
20	Nexus One	1	RAPH800
9	SPH-M900	1	N96
9	RDC-i700	1	DMC-ZS7
6	T-Mobile G1	1	BlackBerry 9630

Table 1: *Devices geo-tagging photos found on Craigslist.*

Craigslist. The virtual flea market *Craigslist* allows users to include photos with their postings by inserting external HTML `IMG` links. If such a link points to the original image as taken by the poster's camera, it may still have its EXIF records intact. [2]

To estimate the number of geo-tagged photos on *Craigslist*, we examined all postings to the San Francisco Bay Area's *For Sale* section over a period of four days (including a weekend). While *Craigslist* does not provide a dedicated query API, it offers RSS feeds that include the postings' full content. Consequently, during our measurement interval, we queried a suitably customized RSS feed every 10 minutes for the most recent 500 postings having images, each time downloading all linked JPEGs that we had not yet seen. In total, we collected 68,729 images, of which about 48% had EXIF information. 914 images were tagged with GPS coordinates, i.e., about 1.3% of the total.[3] We presume that this number is lower than what we found for Flickr and YouTube because many photos on *Craigslist* are edited before posting. Still, already with a cursory look we found several cases where precise locations had the potential to compromise privacy, as we discuss in § 4.

We also examined the camera models used to take the 914 geo-tagged *Craigslist* photos; see Table 3. While the three iPhone models are clearly ahead of all other models, it is interesting to see a wide range of other devices.

4 Cybercasing Scenarios

In this section we take the perspective of a potential attacker to investigate examples of *cybercasing*. We focus on four different scenarios: one on Craigslist, one on Twitter and two on YouTube.

Craigslist. In our first scenario, we manually inspected a random sample of Craigslist postings contain-

[2] Linking to external files is only one way to include photos with a posting. A user can also upload images directly to the *Craigslist* server, in which case they are recoded and loose any metadata in the process.

[3] Postings often contain multiple photos and thus the number of geo-tagged images is larger than the number of postings having any. The relative share should however map to the number of postings as well.

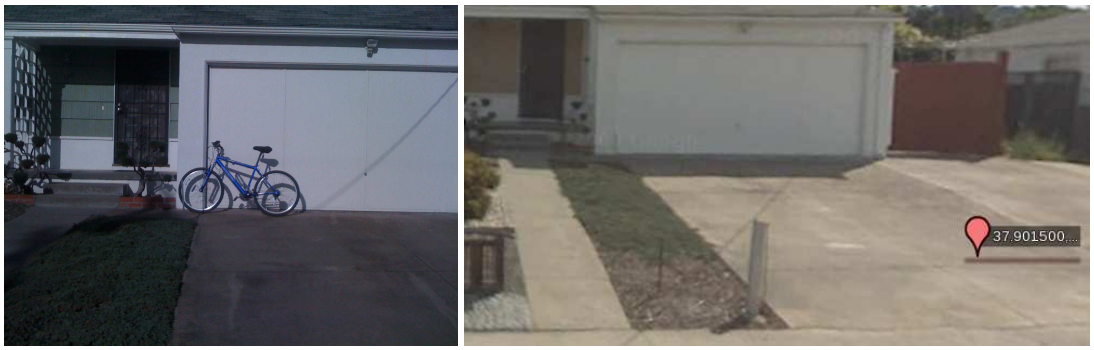

Figure 1: Photo of a bike taken with an iPhone 3G and a corresponding Google Street View image based on the stored geo-coordinates. The accuracy of the camera location (marked) in front of the garage is about $+/-1\,m$. Many classified advertisement sites contain photos describing objects for sale taken at home that automatically contain geo-tagging.

ing geo-located images, collected as described in § 3. One typical situation we found was cars being offered for sale with images showing them mostly parked in private parking lots. Most of the time it was straight-forward for us to verify a photo's geo-location by comparing the image with what we saw on Google Street View.

A fair amount of postings with geo-located images also offered other high-valued goods, such as diamonds, obviously photographed at home; making them potential targets for burglars. In addition, many offered specifics about when and how the owner wants to be contacted ("please call Sunday after 3pm"), which allows for speculation about when somebody might be at home. Since many postings published more than one image, and some locations were the origin of more than one offer, a more accurate estimation of the postal address would have been possible through averaging the geo-tags.

While we did not further verify addresses, we set up an experiment to assess GPS accuracy in a typical Craigslist setting: We first photographed a bike in front of a garage with an iPhone 3G, as if we wanted to offer it for sale (see Figure 1). When we then entered the geo-coordinates that the phone embedded into the picture into Street View, Google was able to locate the photo's position within $+/-1\,m$. Such accuracy is much higher than what we believe most people would expect.

Among the Craigslist postings with geo-information we also found a significant number where the poster chose to not specify a home address, phone number, or e-mail account but opted for Craigslist's anonymous emailing option. We take this as an indication that many posters were not aware that their images were geo-tagged and thus leaked their location.

We note that while we only performed experiments with Craigslist's "For Sale" category, it is not hard to imagine what consequences unintended geo-tagging might have in "Personals" or "Adult Services".

Twitter. Blogging has become a common tool for celebrities to provide their fans with updates on their lives, and most of such blogs contain images. Likewise, many celebrities now also use public Twitter feeds that, besides potentially being geo-tagged themselves, may also link to external images they took. Our second scenario therefore involved tracking one of the co-author's favorite reality-show host who is currently very active on Twitter. The celebrity's show is broadcast on US national TV and has been exported into various foreign countries. In the latest episodes, the TV channel even started advertising that the host is maintaining a Twitter feed. It turns out that most images posted to that feed—including photos taken at the host's studio, places where he walks his dog, and of his home–were taken with an iPhone 3GS and are hosted on *TwitPic*, which conserves EXIF data. In addition, we noticed that the host is also commonly tweeting while he is on travel or meeting other well-known people away from home. Using the Firefox plug-in *Exif Viewer*, a right-click on any of the Twitter images suffices to reveal these locations using an Internet map service of one's choice. Again, averaging geo-tags from multiple images taken at the same location would increase the confidence in the results further.

Geo-location can also be exploited for the taking the opposite route: finding a celebrity's non-advertised Twitter feed intended only for private purposes (e.g., for exchanging messages with their personal friends[4]). Doing so becomes possible by sites such as *Picfog*, which allows anybody to search all images appearing on Twitter by keyword and geo-location in real-time. Having a rough idea of where a person lives thus allows to tailor queries accordingly until the right photo shows up. As an experiment, we succeeded to find a non-advertised but publicly-accessible Twitter-feed of a celebrity having a residence in Beverly Hills, CA.

[4]By default, Tweets are public and do not require authentication.

Note that while our scenario focused on celebrities, similar strategies obviously apply to tracking any person publishing geo-locations in similar ways.

YouTube. In our final setup, we examined whether one can semi-automatically identify the home addresses of people who normally live in a certain area but are currently on vacation. Such knowledge offers opportunities for burglars to break into their unoccupied houses. We wrote a script using the YouTube API that, given a home location, a radius, and a keyword, finds a set of matching videos. For all the videos found, the script then gathers the associated YouTube user names and downloads all videos that are a certain *vacation distance* away and have been uploaded the same week.

In our first experiment, we set the home location to be in Berkeley, CA, downtown and the radius to 100 km. As the keyword to search for we picked "kids" since many people publish home videos of their children. The vacation distance was 1000 miles. Our script reported 1000 hits (the maximum amount) for the initial set of matching videos. These then expanded to about 50,000 total videos in the second step identifying all other videos from the corresponding users. 106 of these turned out to have been taken more than 1000 miles away and uploaded the same week. Sifting quickly through the titles of these videos, we found about a dozen that looked promising for successful cybercasing. However, we stopped after the first real hit: a video uploaded by a user who was currently traveling in the Caribbean, as could clearly be seen by content, geo-tagging, and the date displayed in the video (one day before our search). The title of the video was similar to "first day on the beach". Also, comments posted along with the video on YouTube indicated that the user had posted multiple vacation videos and is usually timely in doing so. When he is not on vacation, he seemed to live with his kids near Albany, CA (close to Berkeley) as several videos were posted from his home, with the kids playing. Although the geo-location of each of the videos seemed not to directly point to a single house due to GPS inaccuracies indoors, the user had posted his real name in the YouTube profile, which would likely make it easy to find the exact location using social engineering. In addition, averaging locations across his home videos seemed to converge to one particular house.

We also performed a second search with the same parameters, except for the keyword which we now set to "home". This time, we found a person who seemed to have moved out of the city. While many of that person's videos had been geo-tagged with coordinates at a specific address in San Francisco, the most recent video was at a place in New Jersey for which Google Maps offered a real-estate ad including a price of $ 399,999. The person had specified age and real name in his YouTube entry.

Finally, we note that using our 204-line Python script we were able to gather all the data for these two experiments within about 15 minutes each.

5 Improving Locational Privacy

The key to avoid wide-spread misuse of location information is clearly educating users about its misuse potential. However, it is our experience that currently even many tech-savvy users find it difficult to accurately assess the risk they face. We thus believe that the security and privacy community needs to take a more active role in shaping the deployment of this rapidly emerging technology. We deem it crucial to ensure that users *(i)* are put into a position where they can make informed decisions; and *(ii)* are sufficiently protected unless they explicitly opt-in to potentially risky exposure. In the following, we frame preliminary suggestions towards this end.

On a general level, we encourage our community to aim for a consensus on what constitutes an acceptable privacy policy for location-based services. As we discuss, current approaches differ widely across devices and services, and we believe that establishing a unified strategy across providers would go a long way towards avoiding user confusion and thus unnecessary exposure. More specifically, we claim that a global opt-out approach to sharing high accuracy location-information is almost always inappropriate. Rather, users should need to acknowledge usage at least on a per-application basis. Apple's iPhone takes a step into the right direction by requesting permission each time a new application wants to access the GPS sensor. However, its user interface still has two short-comings: *(i)* it does not apply that policy to photos/videos taken with the internal camera; and *(ii)* for each application, it is an all-or-nothing decision. Regarding the latter, it seems that often simply reducing a location's resolution might already be a suitable trade-off. As an experiment on how such an approach could be supported, Figure 2 shows a mockup we did for extending the iPhone's standard permission request dialog with a slider allowing to choose an acceptable resolution in intuitive terms. According to the choice, the device would then strip off a corresponding number of the least-significant digits of any coordinates.[5] Not only would this give users more control, but it would also explicitly point out that house-level accuracy is in the cards.

It is also stimulating to think about how APIs such as offered by Flickr and YouTube can offer a higher level of privacy without restricting geo-technology in its capabilities. One way would be for *them* to reduce the resolution; that however would limit some services significantly. Conceptually more interesting is to leverage ap-

[5] For jail-broken iPhones, there are already 3rd-party tools available that spoof the location information visible to other applications.

Figure 2: Our mockup of a mobile-phone dialog to give users more control over the geo-location embedded in their photos.

proaches from related fields such as privacy-preserving data mining. As discussed § 2, researchers have started to examine such approaches, yet we are not aware of any real-world API that incorporates these ideas.

6 Final Remarks

This article makes a case for the emerging privacy issue caused by wide-spread adaptation of location-enabled photo and video capturing devices, allowing potential attackers to easily "case out joints" in cyberspace. Several factors aggravate the problem. First, many people are either unaware of the fact that photos and videos taken with their cell phones contain geo-location, especially with such accuracy; or what consequences publishing the information may have. Second, even experts tend to neglect the easy search capabilities of today's online APIs and the resulting inference possibilities. Third, the fact that only a small percentage of all data is currently geo-tagged must not mislead us to ignore the problem because *(i)* with all the commercial pressure, the number seems poised to rise; and *(ii)* our preliminary experiments demonstrate that even a seemingly small fraction like 1% can already translate into several hundred relevant cases within just a small geographical area.

Finally, we want to emphasize that we are not advocating a ban of geo-location in general or geo-tagging specifically. It is a wonderful technology that drives innovation in many areas. However, we feel there is a clear need for education as well as for research how to best design systems to be location-aware while at the same time offer maximum protection against privacy infringement.

References

[1] David Crandall, Lars Backstrom, Dan Huttenlocher, and Jon Kleinberg. Mapping the Worlds PhotosMapping the Worlds Photos. In *Proc. WWW*, 2009.

[2] EFF. On Locational Privacy, and How to Avoid Losing it Forever. `http://www.eff.org/wp/locational-privacy`.

[3] Ian Goldberg Ge Zhong and Urs Hengartner. Louis, Lester and Pierre: Three Protocols for Location Privacy. In *Proc. Privacy Enhancing Technologies Symposium*, 2007.

[4] John Krumm. A Survey of Computational Location Privacy. *Personal and Ubiquitous Computing*, 13(6):391–399, 2009.

[5] Steven J. Murdoch and Maximillian Dornseif. Far More Than You Ever Wanted To Tell Hidden Data in Internet. `http://md.hudora.de/presentations/forensics/HiddenData-21c3.pdf`.

[6] Femi Olumofin, Piotr K. Tysowski, Ian Goldberg, and Urs Hengartner. Achieving Efficient Query Privacy for Location Based Services. In *Proc. Privacy Enhancing Technologies*, 2010.

[7] Nayot Poolsappasit and Indrakshi Ray. Towards Achieving Personalized Privacy for Location-Based Services. *Transactions on Data Privacy*, 2(1):77–99, 2009.

[8] Please Rob Me - Raising Awareness About Over-sharing. `http://pleaserobme.com`.