Geo-Privacy in Data-Rich Social Media Environments

Marc P. Armstrong
Department of Geographical and Sustainability Sciences
The University of Iowa
Iowa City, IA 52242
marc-armstrong@uiowa.edu

Introduction

Concerns about privacy are deeply ingrained in the psyche of most US citizens. Such concerns are particularly manifested on the heels of news reports about data breeches by hackers or an accidental release of information normally presumed to be confidential. Yet many of the same individuals who express concerns about data privacy are either willfully ignorant or display a wanton disregard for matters related to geo-privacy in social media transactions. The purpose of this brief paper is to sketch out some basic ideas about how locational information can be gleaned, directly or inferred, from social media and how such information can be transformed to yield space-time traces of human activity patterns at different levels of spatial and temporal resolution. Such traces can be used by a determined data spy to violate the privacy of individuals and groups of individuals.

Geo-privacy

Despite it being a fluid concept that may change drastically depending on culture, or other contexts, all humans need, and expect, some measure of privacy (Dash *et al*., 1959; Armstrong and Ruggles, 2005). While a right to privacy is not explicitly guaranteed in the US Constitution, it is an important element in tort law and figures prominently in many state-level statues and federal acts (e.g., HIPPA and FERPA). Moreover, privacy is a foundational tenant of the Universal Declaration of Human Rights[1] (Article 12). Geo-privacy is a relatively new construct that has arisen as a consequence of the emergence of new technologies (GIS, GPS, smart phones and social media) that are able to capture and map individual-level movement in space. Having geo-privacy means that an individual is assured that they are secure from any unwanted observation or tracking of their activities. Amassed time-stamped location data allows for the creation of a detailed profile of individual behavior, including inferred habits, preferences, and routines—private information that could be exploited and cause harm: where you go implies what you do. However, many people leave persistent and discoverable digital trails behind them as they engage in their everyday activities. Some of this trail-like information has value to an individual (e.g., location for an E911 emergency), while other data is unvalued and usually unnoticed by social media users.

Types of Locational Information Relevant to Social Media

1. Direct specification. Several types of media provide an explicit way to attach locational tags to transactions. Some applications are explicitly location-dependent (e.g., Foursquare) while others have location-providing options (Facebook).
2. Place name. While not explicitly locational, names can be transformed using several resources such as telephone directories and gazetteers to yield an address or location that can be processed to yield an explicit location. However, location might be provided in multiple formats with

[1] http://www.un.org/en/documents/udhr/

different levels of locational specificity. For example, New York can refer to either a city or state and New York City is far less specific than Brooklyn.

3. Postal address. Addresses are a commonly employed locational identifier (e.g., pizza delivery). They can be transformed into coordinates for mapping using several different methods.
4. Coordinates. Most smart phones have GPS receivers that provide accurate coordinates which can be attached to tags.
5. Wi-Fi locations. In 2011, it became public knowledge that Apple, Google and other companies had been collecting locational information about WiFi hotspots and cell towers to track users and improve their networks. Google, for example, drove down a very high percentage of US streets and geocoded Wi-Fi access points as they collected images for Street View (they announced that this practice was halted when it became public).

Locational Transformations

Raw locational data typically must be transformed to make it more useful in any application, benign or nefarious. Two common transformations are referred to as geocoding and inverse geocoding—attaching locational identifiers to addresses and finding addresses from mapped data.

Geocoding

The geocoding transformation requires two data sources, 1) an input file containing addresses to be transformed through the addition of a coordinate and 2) a geographic base file, which may consist of a street centerline file with address ranges (e.g., US Census TIGER, used for interpolated geocoding) or a parcel map with addresses and coordinates for either a parcel or building footprint centroid (used for direct geocoding).

There are several steps in a typical interpolated geocoding process (single address). First, select an address to geocode: 1147 Maple Street, Iowa City, IA 52245. Parse it into constituent elements (SOUNDEX (NARA) may help). To reduce search in the geographic base file, it may be helpful to divide and conquer by restricting search to IA and ZIP (Iowa City is not relevant except for error check). Then, for all streets in ZIP=52245, match to "Maple" "Street", and for all chains in Maple Street, contains 1147 in range L-H= True.

Once the correct street segment and geometrical chain is found, required quantities are solved by proportion: compute address (1147) as a proportion of the address range (noting odd-even parity) of the containing segment (e.g., 1101-1199). Apply the same proportion using geometry to get location along the street segment centerline. Two additional steps are also often employed to make realistic maps: offset the geocoded coordinate from the street centerline (e.g. 10 meters), again using odd-even parity, and squeeze in from ends (e.g., 20 meters) to disambiguate assignment of a corner address to an incorrect street.

Reverse Geocoding

As the name suggests, this inverts the geocoding process to recover addresses from an "anonymous" dot map. First, start with a map of geocoded locations and precisely register it in a coordinate system (this may require some guesswork, but for small areas this is not a big problem). Then for each "dot", find the street segment in the geographic base file that is closest and snap to it. Calculate the geometrical proportion of the dot along the length of that segment (between street intersections) and use that proportion to calculate the address proportion from the address range and use pre-snap parity (L-R). This

will yield the "best guess" of an address for a dot on a map, subject to errors or distortions that might be introduced as part of the mapping process, intentionally or not (Armstrong, *et al*., 1999)

Cross-linking Information to Compromise Privacy

With the ability to transform between addresses and coordinates, it then becomes a relatively straightforward effort to cross-link with other types of digital information (NRC, 2007). In addition to addresses, other identifiers can then be gleaned which enable further linkages to be established either directly (e.g., via a relation "join" operation), probabilistically, or by ecological inference (a home residence in a particular census block group with certain socio-economic indicators). As data quantity and types continue to increase in this era of "big data", such linkages are often at the fore of discussions related to privacy (e.g., Craig and Ludloff, 2011).

Activity Spaces

An activity space refers to the collection of locations (and the paths between them) that an individual has direct contact with on a daily, weekly, or other cycle (see, e.g., Horton and Reynolds, 1971a; 1971b). At an individual level, information about activity spaces can be used to construct behavioral profiles that detail journey to work, as well as the location of day care, place of worship, social clubs, dry cleaner, grocery and liquor stores and other places of business (see, e.g., Sherman *et al*., 2005). While it is an enormous stretch to suggest that individuals would "tweet" with a geo-tag every time they visited a particular grocery store, if they post information only occasionally, over time, a profile can be inferred.

Visualization and Methodologies

A key area of research that can be employed for theoretical and methodological constructs is generally referred to as time geography. The original work in this area can be traced to the Swedish geographer Torsten Hägerstrand (1970) and his students. Several main themes of Hägerstrand's work have been placed in a modern research context and extended by, for example, Miller (1991) and Kwan (1998).

References

Armstrong, M.P. and Tiwari, C. 2007. Geocoding methods, materials, and first steps toward a geocoding error budget. Rushton, G., Armstrong, M.P., Gittler, J., Greene, B.R., Pavlik, C.E., West, M.M., and Zimmerman, D.L. (editors) *Geocoding Health Data*. Boca Raton, FL: CRC Press, pp. 11-35.

Armstrong, M.P and Ruggles, A. 2005. Geographic information technologies and personal privacy. *Cartographica*, 40(4):63-73.

Armstrong, M.P., Rushton, G. and Zimmerman, D.L. 1999. Geographically masking health data to preserve confidentiality. *Statistics in Medicine,* 18 (5): 497-525.

Craig, T and Ludloff, M.E. 2011. *Privacy and Big Data*. Sebastopol, CA: O'Reilly Media Inc.

Dash, S., Schwartz, R.F. and Knowlton, R.E. 1959. *The Eavesdroppers*. New Brunswick, NJ: Rutgers University Press.

Hägerstrand, T. 1970. What about people in regional science? *Papers of the Regional Science Association* 24 (1): 6–21.

Horton, F.E. and Reynolds, D.R. 1971a. Effects of urban spatial structure on individual behavior. *Economic Geography*, 47 (1): 36-48.

Horton, F.E. and Reynolds, D.R. 1971b. Action-space differentials in cities. McConnell, H. and Yaseen, D.W. (editors) *Perspectives in Geography 1: Models of Spatial Variation*. Dekalb, IL: Northern Illinois University Press, pp. 84-102.

Kwan M.P. 1998.Space-time and integral measures of individual accessibility: A comparative analysis using a point-based framework. *Geographical Analysis*, 30:191-217.

Miller, H. J. 1991. Modelling accessibility using space-time prism concepts within geographical information systems. *International Journal of Geographical Information Systems* 5(3): 287–302.

National Research Council. (Committee on Confidentiality of Linked Social-Spatial Information: Gutmann, M., Stern, P., Armstrong, M.P., Balk, D., Green, K., Levine, F., Onsrud, H., Reiter, J., and Rindfuss, R.) 2007. *Putting People on the Map: Protecting Confidentiality with Linked Social-Spatial Data*. Washington, DC: The National Academies Press.

Sherman, J.E., Spencer, J., Preisser, J.S., Gesler, W.M. and Arcury, T.A. 2005. A suite of methods for representing activity space in a healthcare accessibility study. *International Journal of Health Geographics*, 4:24 (unpaged) http://www.ij-healthgeographics.com/content/4/1/24