Control of Personal Location Data*

(Extended Abstract)

Sudarshan S. Chawathe
Computer Science Department, University of Maryland, College Park, MD 20742
chaw@cs.umd.edu

Introduction Several recent technological developments have contributed to the ability of individuals and organizations to monitor people at a level of detail previously reserved for science-fiction movies. Global-positioning system (GPS) devices, video cameras, RFID tags, sensors [19], and other devices that permit fine-grained monitoring of objects and people are now produced with mass-market efficiencies and very low prices. Concurrent improvements in the capabilities and prices of networking and mass storage devices have made the collection and perpetual storage of detailed monitoring data not only feasible, but quite inexpensive. In fact, in most cases there are few, if any, *technological* reasons for ever destroying data collected in this manner.

Benefits In several scenarios, the advantages of these technologies are quite obvious. For example, much has been written about the potential of RFID tags to revolutionize the entire distribution and retail business by permitting accurate and up-to-the-minute monitoring and control of inventory. Similarly, GPS devices are invaluable tools for navigation and rescue. One is likely to be thankful for the location-advertising features of cell phones in an emergency situation. Inexpensive sensors placed at remote locations or attached to animals provide a wealth of scientific data that would otherwise be very difficult or impossible to obtain.

These technologies are also useful in several less obvious scenarios. For example, the ability to monitor visitors' locations accurately may permit a national-park ranger to allow access to areas that would otherwise be completely off-limits due to infeasibility or expense of ensuring safety. Further, monitoring the routes of visitors may help guide decisions on improvements to park facilities.

*This work was supported by the National Science Foundation with grants IIS-9984296 (CAREER) and IIS-0081860 (ITR).

Dangers It is already quite easy to monitor the location of a person over time. For example, we may infer a person's location using his or her cell phone (using GPS or signal-strength data), Internet activity (service-provider logs), cameras in various public and private locations (parking lots, convenience stores), and so on. These information channels are likely to become more numerous and more accurate over time. Further, while each source of data is quite powerful on its own, the increasing interconnection of such sources and resulting correlation of data (e.g., credit card information from convenience store and RFID tag from bookstore) increases the power manifold.

For example, an RFID tag encodes and announces a unique identifier that permits tracking of individual items at very low cost. Unlike UPC bar-codes, RFID tags permit the tracking of individual items (e.g., razor blade package number 736) and not just product types (Gillette razor model G4). Further, reading an RFID tag requires only moderate proximity and radio-frequency communication, and can therefore be accomplished without alerting others. It is not clear whether it will be easy to remove RFID tags from products after purchase; however, it is likely that such removal will be difficult or impossible, since many RFID tags are also designed to serve as theft-prevention devices. When such tags are embedded in clothing, books, and other everyday articles, a simple RFID scan may reveal an embarrassing wealth of personal data.

Technology We now address the following specific problem: How should location data be managed in order to permit a balance between privacy and utility? We will focus on mechanisms, and not policies,

for privacy-sensitive use of data. The goal is to provide mechanisms that permit a wide range of privacy policies to be implemented (and enforced).

A key decision in managing location data (or any data where privacy is a major concern) is the choice of a storage location [20]. In general, storing data at a location fully under the control of the person it describes is likely to provide privacy benefits. For example, Alice's GPS receiver may store very detailed information about her whereabouts (as time-coded track-points). Since the GPS receiver remains under Alice's control, one may assume that there are no significant privacy concerns. However, even in this simple situation, care must be exercised. If Alice connects her receiver to her computer and then browses some Web site, the track data could very easily be compromised (e.g., using scripting or simple social engineering).

Even if Alice is careful not to reveal sensitive data from her GPS receiver in the obvious ways (e.g., by turning scripting off and filling out Web forms carefully), a Web site operator may be able to infer significant information based on Alice's access patterns. For example, if Alice uses a Web site for downloading maps in order to plot the course of her morning jog, anyone with access to the Web server logs may be able to reconstruct substantial parts of the route, especially by making use of supplemental information and assumptions (e.g., jogging is most likely to occur on a well-established trail or path). The general problem of accessing a database (map server in our example) in a manner that does not divulge sensitive information to the database owner has been studied as the problem of Private Information Retrieval (PIR) [13, 3]. Unfortunately, the current techniques often make several restrictive assumptions and require considerable computing (processing and network) resources.

Storing data on a device under direct control of the person it describes also creates other problems. In some cases, the data may describe more than one person and may be difficult to divide into per-person partitions. Further, querying and manipulating such data is difficult. For example, consider the geocoded data generated by several devices on board an ambulance (multiple cell phones, GPS devices, sensors). In order to plan and monitor the activities of a fleet of ambulances, it is necessary to access such location data from a central site. If the data remains on the on-board units, such querying poses a difficult information-integration problem. Further, loss of data (due to device failure or operator error) is much more likely when data is stored in a disparate and poorly managed collection of devices.

We need methods that permit location data to be transmitted and stored by a number of different computers without compromising privacy. An especially interesting case is when the data must traverse a set of computers in a peer-to-peer [25], sensor [19], or distributed upload [14] network (e.g., the method of privacy homomorphisms [7]). This need gives rise to several sub-problems. When data is shared across databases (belonging to different individuals or organizations), we would like to ensure that unnecessary data is not revealed to the other party. Recently proposed methods for privacy-preserving database operations [18] may be used for this purpose. However, the work to date has focused on very simple operations (e.g., set intersections) and further work is needed to support other operations.

When location data is transmitted and stored by a third party, ensuring the authenticity and accuracy of the data is an important concern. Recent work on authentic third-party data publication and authentic data structures and trust management is relevant here [28, 11, 1, 6]. While it is trivial to sign all data using cryptographically strong schemes, such a method does not permit effective querying or generation of derived data. Methods for evaluating SQL queries on relational data may be adapted to location data [15]. These methods permit a trade-off between the granularity at which the server is aware of the client's location and the efficiency of queries. Unfortunately, this trade-off cannot be made dynamically.

It is important to prevent not only direct exposure of sensitive data, but also exposure through more circuitous routes, such as inference based on a knowledge of the database or domain. This so-called inference problem has been extensively studied in the context of standard relational databases [12, 27]. However, the spatio-temporal nature of location data poses some additional challenges. For example, a query that can only narrow down a person's location to a 20 square-mile area may be considered acceptable from a privacy point of view. However, if the area in question is in remote Nevada with only one highway through it, the result is a much more precise determination of location.

Location data that is stored or referenced on the Web raises additional concerns. Web sites can correlate information using cookies and other features, posing additional privacy concerns [23, 29]. For ex-

ample, a Web site may be able to access sensitive information stored in a cookie by another Web site, leading to identity theft [5]. Initiatives such as P3P are promising developments in this arena [8]. Similarly, methods that provide secure access control for XML, which is likely to become the lingua franca of the Internet, can be used to support a fine-grained access control policy [10].

Education and Public Policy In the light of the formidable technical challenges posed by needs of privacy-sensitive management of location data, it is perhaps easy to lose sight of the broader issues of education and public policy. This point is perhaps best made using a topical example which, while not specific to location data, serves well to illustrate how easily privacy can be eroded.[1] The past few weeks have witnessed a flurry of articles in the mainstream press addressing the privacy concerns raised by Google's new *GMail* service. (Templeton's article provides a good summary [26].) In a nutshell, GMail provides its subscribers free email service, along with *enough storage to store all their email forever.*[2] The controversial issues center around how Google, as a profit-seeking company may use the wealth of information that is easily available from a person's email history. For example, advertising links on Google's search pages may be tailored to the perceived interests of the viewer, as gleaned from his or her email messages. It is not hard to imagine embarrassing, dangerous, or even sinister consequences of this setup.

While the privacy concerns raised by GMail are certainly worth serious consideration, more interesting is the fact that many of these concerns are almost equally applicable to a variety of other service providers (email, Internet, news, and so on). Yet, these concerns had received very little attention prior to the GMail story. For example, an email service provider, such as Yahoo or Hotmail, already has access to a large collection of email messages. Although the limited storage provided by these services may result in a quantitatively different problem, it is likely to be qualitatively quite similar. (For example, much can be inferred about a person's interests and recent activities by examining only a week's worth of messages.) Even more sensitive information is available to an Internet service provider, who can easily monitor the Web sites frequented by a customer.

Few email users are aware of the ease with which their activities can be monitored and of the laws (or lack thereof) governing their messages (e.g., the Stored Communications Act [17]). This situation is more troubling given the relative technical ease with which many of the problems can be addressed (e.g., by using encryption, storing messages on foreign servers for no longer than 180 days, etc.). When presented with the facts on the ease with which their privacy may be compromised [24], people often respond by assuming that no one is likely to have enough of an interest in their lives to warrant such snooping. In effect, many are still mentally in a world in which privacy is the default and publicity requires effort while the reality is closer to the opposite, a panopticon society [16]. This situation underscores the importance and difficulty of educating a population that is being thrust into using technologies that can seriously erode their privacy at an ever-increasing rate.

The appropriate response in terms of public policy is certainly not clear and the policy decisions are at least as difficult as the technical ones. Geographical information has an interesting legal history (e.g., fair-housing laws). However, the availability of a large amount of very detailed geocoded data that can be correlated with other databases raises concerns that were not previously critical [9, 4]. Ideas such as treating privacy as intellectual property are intriguing and deserve further debate [22]. While intellectual property rights may enhance privacy in this manner, used elsewhere (e.g., to prevent analysis of software used for managing ones data) they may erode privacy because the user cannot verify how his or her data is being used [21, 2].

[1] To see a more direct connection to location privacy, it suffices to replace GMail with a service such as `geocaching.com`.

[2] More precisely, it currently provides a gigabyte of storage, which permits most casual email users to store all their messages for a very long time.

References

[1] A. Anagnostopoulos, M. Goodrich, and R. Tamassia. Persistent authenticated dictionaries and their applications. In *Proceedings of the Information Security Conference*, 2001.

[2] R. Anderson. Security in open versus closed systems—the dance of Boltzmann, Coase and Moore. In *Proceedings of the Conference on Open Source Software: Economics, Law and Policy*, Toulouse, France, June 2002. Institut d'Economie Industrielle. Available at `http://www.idei.asso.fr/ossconf.html`.

[3] A. Beimel and Y. Stahl. Robust information-theoretic private information retrieval. In *Proceed-*

ings of the Conference on Security in Communication Networks, 2002.

[4] C. J. Bennett and P. M. Regan. Surveillance and mobilities. *Surveillance and Society*, 1(4):449–455, 2004. `http://surveillance-and-society.org/`.

[5] H. Berghel. Identity theft, social security numbers, and the Web. *Communications of the ACM*, 43(2):17–21, Feb. 2000.

[6] M. Blaze, J. Feigenbaum, and A. Keromytis. The role of trust management in distributed systems security. In *Secure Internet Programming*, 1999.

[7] C. Boyens and M. Fischmann. Profiting from untrusted parties in Web-based applications. In *Proceedings of the 4th International Conference on E-Commerce and Web Technologies*, volume 2738 of *Lecture Notes in Computer Science*, pages 216–226, Prague, Czech Republic, Sept. 2003. Springer-Verlag Heidelberg.

[8] L. Cranor, M. Langheinrich, M. Marchiori, and J. Reagle. The platform for privacy preferences 1.0 (P3P1.0) specification. W3C Recommendation. `http://www.w3.org/TR/P3P/`, Apr. 2002.

[9] M. R. Curry. In plain and open view: Geographic information systems and the problem of privacy. In *Proceedings of the Conference on Law and Information Policy for Spatial Databases*, Santa Barbara, California, 1996. National Center for Geographic Information and Analysis.

[10] E. Damiani, S. De Capitani di Vimercati, S. Paraboschi, and P. Samarati. A fine-grained access control system for XML documents. *ACM Transactions on Information Sytems Security (TOISS)*, 5(2):169–202, 2002.

[11] P. Devanbu, M. Gertz, C. Martel, and S. Stubblebine. Authentic third-party data publication. In *14th IFIP Conf. on Database Security*, 2000.

[12] C. Farkas and S. Jajodia. The inference problem: A survey. *SIGKDD Explorations*, 4(2):6–11, 2002.

[13] Y. Gertner, Y. Ishai, E. Kushilevitz, and T. Malkin. Protecting data privacy in private information retrieval schemes. In *Proceedings of the ACM Symposium on Theory of Computing*, 1998.

[14] L. Golubchik, W. C. Cheng, C.-F. Chou, S. Khuller, H. Samet, and C. J. Wan. Bistro: a scalable and secure data transfer service for digital government applications. *Communications of the ACM*, 46(1):50–51, 2003.

[15] H. Hacigumus, B. Iyer, C. Li, and S. Mehrotra. Executing SQL over encrypted data in the database-service-provider model. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD)*, pages 216–227, 2002.

[16] D. G. Johnson and K. Miller. Anonymity, pseudonymity, or inescapable identity on the Net. In *Proceedings of the ACM Policy Conference*, pages 37–38, Washington, DC, May 1998.

[17] O. S. Kerr. A user's guide to the stored communications act—and a legislator's guide to ammending it. The George Washington University Law School. Public Law and Legal Theory Working Paper No. 68, 2004. `http://ssm.com/abstract=421860`.

[18] G. Liang and S. S. Chawathe. Privacy-preserving inter-database operations. In *Proceedings of the Symposium on Intelligence and Security Informatics*, Tucson, Arizona, June 2004. To appear. `http://www.cs.umd.edu/~chaw/`.

[19] S. R. Madden, M. J. Franklin, J. M. Hellerstein, and W. Hong. TAG: a Tiny AGgregation service for ad-hoc sensor networks. In *Proceedings of the Symposium on Operating System Design and Implementation*, Dec. 2002.

[20] D. Mulligan and A. Schwartz. Your place or mine?: Privacy concerns and solutions for server and client-side storage of personal information. `http://secinf.net/`, Apr. 2000.

[21] F. Olsen. Control issues. *The Chronicle of Higher Education*, Feb. 2003.

[22] P. Samuelson. Privacy as intellectual property? *Stanford Law Review*, 52:1125, 2000.

[23] E. Sit and K. Fu. Web cookies: Not just a privacy risk. *Communications of the ACM*, 44(9):120, Sept. 2001.

[24] S. Staniford, V. Paxson, and N. Weaver. How to own the Internet in your spare time. In *Proceedings of the USENIX Security Symposium*, San Francisco, California, Aug. 2002.

[25] I. Stoica, R. Morris, D. Karger, M. F. Kaashoek, and H. Balakrishnan. Chord: A scalable peer-to-peer lookup service for Internet applications. In *Proceedings of the ACM SIGCOMM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication (SIGCOMM)*, pages 149–160, San Diego, California, Aug. 2001.

[26] B. Templeton. The GMail saga. `http://www.templetons.com/brad/gmail.html`, Apr. 2004.

[27] B. Thuraisingham. Security checking in relational database management systems augmented with inferene engines. *Computers and Security*, 6:479–492, 1987.

[28] B. Thuraisingham. Selective and authentic third-party distribution of XML documents. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 2004. To appear. Personal communication.

[29] L. Weinstein and P. G. Neumann. Internet risks. *Communications of the ACM*, 43(5):144, May 2000.