PERSONAL INFORMATION MANAGEMENT

EDITED BY

WILLIAM JONES + JAIME TEEVAN

16 Privacy and Public Records

Michael Shamos

Much of the concern about data privacy centers on the surreptitious collection and sale of personal information outside the view or control of the data subject. Even if such dealings are prohibited by law, they are difficult or impossible to prevent because by its very nature the activity remains hidden. It is a transgression of an entirely different character when government itself at all levels voluntarily becomes complicit in the massive disclosure of data about its citizens. When public records are made available on a grand scale through free, publicly available Internet databases, we are forced to reconsider the very notion of what a public record is or ought to be.

To grasp the problem in a concrete way, please visit http://www2.county.allegheny.pa.us/RealEstate/Search.asp. This is a Web site maintained legally and officially by the government of Allegheny County, Pennsylvania, which paid more than $20 million to create it. Allegheny is the county in which I reside and own a house. Near the bottom of the Web page is a search box labeled "Owner's Name." Enter the name Shamos and click the "Search" button directly below the box. You will be presented with an array of information about me and my house. You'll learn my wife's name, how much we paid for the house, its assessed value, how many bathrooms it has, that we have central heating and air conditioning, how much we pay in real estate taxes, whether we were ever delinquent in paying, how much we were assessed in penalties, and a lot more data you didn't imagine the county even knew. You will also be treated to a photo of my house and its floor plan. Many people are horrified when they see how much is disclosed.

Here's the crux of the problem: there are very important and legitimate reasons for all of the above information about me to be publicly available. In general, records are designated as public to achieve some policy purpose. In this case, it has to do with the fairness of property assessments. Each homeowner needs to be able to verify that his assessment is equitable. The only way he can do that is to obtain data about nearby properties and learn their assessed values. For this reason, the decision has been made to allow the records to be viewed by the general public.

There's a rub, however. In the days before the Internet, the records were still public but difficult to access. They were stored in the county office building and a personal visit was required to see them. There was often a charge for copying, and the fact that physical files and binders had to be handled and the time needed to do that set a natural bound on the amount of effort, and the cost, that could be expended. When the very same records are digitized and made available with powerful search engines, the cost and pain of access dwindles to nearly nothing.

There are a multitude of unintended uses that can be made of the property database, ranging from the innocuous to the frightening. Of course the records have tremendous value for marketing purposes. Every home that has central heating or air conditioning is listed as such, and any company that provides services for such systems now has a free mailing list. It is the easiest chore to spider the entire property database and obtain your own copy of it. The set of people who have neither air conditioning nor heating has suddenly become a target for people selling those items, some legitimate, some not.

Are you interested in the complete set of properties your competitor owns so you can preview his expansion plans? The database will be happy to tell you. Planning a crime? The photos and floor plans will enable you to avoid casing the house, which might attract undue attention. Do you have some need for a complete list of the residents on a particular street, neatly sorted by house number? That's easy. Just type in the street name and you'll get more than you need. Might it be embarrassing which of your neighbors is delinquent in paying property taxes? The "tax information" tab will provide as much blackmail material as you could want.

When viewed in this light, the property database seems to be a huge intrusion into the private lives of land owners. There is no comparable resource that lists apartment renters, so the situation seems asymmetric. If you want to keep your data away from the public, don't buy any property. Yet it is difficult to argue that any single field of data ought to be kept from the citizen who has a legitimate question about the fairness of assessments. We will argue later that the source of the trouble is not the existence of the database, but the ability of anyone in the world to use for purposes *other* than checking the fairness of assessments.

The database has had unintended consequences, a common phenomenon that infects technological innovations. It has altered the social fabric of Pittsburgh to some degree. After meeting someone at a cocktail party, it is now commonplace to go to the Web, find out whether they are married, learn how valuable their house is, and view a picture of it. This is the sort of information that

was not readily available previously. One may argue that it keeps people honest, knowing that if they lie about such things they will be found out in short order. But suppose they wish to keep mum? That is no longer a viable alternative.

The Allegheny County assessment database is just a small example of the sort of information published on the Internet. To see the magnitude of the issue, please visit www.searchsystems.net, a Web site that lists legal public records depositories available on the Web. As of the beginning of February 2006, it contained links to over 39,000 such databases. They range from the innocuous (such as lists of certified public accountants) to the highly controversial (sexual predator databases). They vary from the surprising (data on prisoners released from incarceration in Florida) to the simply macabre (last meals served to executed murderers in Texas). What is striking is the breadth and scope of information available, from which it is possible, through automated means, to develop whole dossiers on individuals, built purely by using data legitimately obtained.

Each individual fact about every person listed in any of these sources has been determined by a government authority to be a public record. Therefore, as a matter of policy the subject has by law lost the ability to conceal the data or control its dissemination and use. A thief who wants to steal guns in Florida can look up licensed gun owners in any part of the state and then, through a different database, find an aerial photograph of the owner's house and details of his property. Clearly this information is invaluable in planning crimes, a result surely not intended by the Florida officials who decided to build these databases and release them for public use.

How, then, can we ensure that databases are only used for their intended purposes and not others? Even if there were a way to do so, the problem would not be solved, though it might be alleviated to some degree. The reason is that all sorts of "public" information can be accumulated legally without ever accessing these official databases. Let us distinguish "public record" from "public information." A *public record* is data maintained by a government agency that, by policy, is made available to members of the public. *Public information* is information that may be obtained lawfully by all persons, with or without the help of the government. Whether a record is public or not is determined by duly authorized officials. The decision can be changed by executive or legislative action. Whether information is public is not so constrained.

Generally, all "public records" are "public information," but the latter class is much broader. It includes, for example, anything that a person sees on the street. Thus, if I see you walk into the Main Street Bank at 12:32 p.m. on Friday and walk out again at 12:54, that fact is public information. You have no legal right to prevent anyone from noticing the fact, writing it down, putting it in

a database, or publishing it on a Web site. (We exclude from this discussion activities that could be considered threatening or stalking. I can't legitimately learn the fact that you were at Main Street Bank, for example, by following you around all day at a distance of three feet.) Suppose I am able to watch John Doe from a considerable distance, and only when I am in a public place and he is also. I can't follow him into his house, or go into the bathroom with him, or invade his office. But I can stay at a discrete distance and key into my PDA a brief description of where he is every five minutes throughout a 24-hour period. Logically, we must agree that each single individual fact about his whereabouts while he is observable from a public place must constitute public information. He has no capability to shield himself from view while in public.

Suppose now that my PDA is wireless and I create a publicly accessible John Doe Web site to which his location is uploaded every five minutes for the rest of his life. That is, at any time anyone in the world can tell not only where Doe is now but where he was at all times in the past. Most Americans would consider such a Web site to be a gross violation of Doe's "privacy." Indeed, when I conduct such a poll in my privacy lectures, almost all Americans indicate just that. The results are quite different for audiences in different countries. I have never seen a Chinese class in which even half of the people felt there would be anything at all wrong about such a Web site. Even the Americans, when pressed, are unable to articulate any ethical or legal reason that such a site could not be operated. And they all agree that Doe has no privacy interest in any of the individual facts listed on the Web site. Why should he be concerned if people know he visited the library at 2:32 p.m. on March 23?

He might not be worried about that, but by being observed constantly he has lost an important facet of privacy that he thought he had. When Doe engages in an act on the public street, he has the opportunity to inspect his surroundings to see whether anyone is watching. If not, he may do things that he would never consider doing while being observed. If he is spied upon surreptitiously, he has given up the ability to make this choice. He certainly expects to be able to tell whether anyone who knows him is watching, and he may make decisions based upon assumed anonymity. The Doe Web site forfeits his anonymity without his permission, yet under U.S. law he has no right to anonymity while in public.

There are two major problems with the Doe database. The first is in moving from the individual facts to a comprehensive collection (and then storing the collection for the subject's lifetime and beyond). The second is in making the collection freely available at essentially no cost to anyone and everyone, including anonymous and possibly malevolent users. There is a feeling, difficult

to make precise, that at some point the database of Doe location data is just too intrusive to be acceptable. However, having just two lines of data would be innocuous. So where is the boundary beyond which the aggregation becomes objectionable?

To understand where the line might be drawn, consider how public records were accessed before the Internet. Generally they were maintained as paper files or bound volumes that could be examined by members of the public. To do that, they would have to go to the county hall of records, park, make a personal visit to the appropriate office, learn how to use the record indexes, locate the records they wanted, and pay to have copies made—if copying were even permissible. The time, effort and money expended acted as a natural limit on the volume of access by any individual.

At least it limited those who lacked resources. A single person could never accumulate a county-wide property database, for lack of both time and money. However, a corporation intent on selling the data could easily afford to station a representative at the county office building to make a copy of any new data that was recorded, and indeed companies did so and made the databases available online, but for a fee. Someone intent on building a dossier would still have to incur charges to do so and spend time searching through large numbers of databases. Nevertheless, hunting while sitting at one's own computer is vastly more efficient than making physical visits to government offices. The aggregation therefore permits not only the assembling of large amounts of data on one individual, but also allows invasion of many different lives all from the convenience of a desk chair.

What answer might there be? The European approach is to require registration and regulation of data gatherers, known as *data controllers* in EU parlance. Under a rather strict regimen, data controllers are highly constrained in taking data directly from subjects themselves. The controller must inform the subject of the purpose for which the information is being collected and will be used, must obtain the subject's permission, and cannot pass the information on to any party that has not also agreed to the same conditions. The penalty for violation of these provisions is to be stricken from the register of authorized data controllers, essentially making it impossible to remain in business. Thus, the privacy commissioner, who maintains the register, has substantial enforcement power.

Such a structure is unlikely to be adopted in the United States, where data is considered to be a commodity readily available for sale. Various state governments have enacted a variety of sunshine and freedom of information laws that require the governments to make data public. And neither the U.S. structure nor the European one prevents creation of the Doe database. Even in

the EU, if data is to be collected on a subject from sources other than the subject himself, it is only necessary to inform the subject in advance that data is going to be gathered and to identify the gatherer. In the U.S., of course, no notice at all is required.

One approach that deserves further exploration is to arm the subject with a right to learn not only who is collecting data, but who is accessing it—the same sort of protection a person has in a public place, namely the ability to observe who is in the vicinity.

Suppose anyone collecting data on a subject were required not only to notify the subject and give him access to, and the ability to correct, the database, but also—and this is the key to symmetry—notify the user each time any data about him was accessed and identify the party who made the access. The "notification" can be nonintrusive; for example, the subject may be allowed to visit a Web site or perform a query to retrieve the promised information. It is not necessary to interrupt him constantly throughout the day every time he has been the target of an access.

A completely different approach is to restore the natural dampening role that cost played in traditional paper-based systems. If there were a mandatory cost to accessing personal information, its incidence would certainly decline. To avoid making public databases too expensive to examine, each citizen could be given a quota of free accesses each year at a level deemed sufficient to promote open-access policies but not to foster large-scale invasions of privacy. A similar proposal has been made to reduce the volume of spam emails. If a small charge akin to postage were made for every email, individual senders would suffer very little, but bulk spammers, who may send millions of messages at a time, would find the cost prohibitive. And so it would be for information access.

A third possibility is to restrict the use of public records to the purposes for which they are being maintained. If the reason for keeping a property ownership database is to allow verification of the fairness of assessments and confirmation of the chain of title to land, then using the database for those purposes ought to be free or so inexpensive that the rights of the public are not crippled. Other uses, such as marketing air-conditioning services to those unlucky enough to live in a county having an extensive home database, could be made expensive or even prohibited altogether. Whether a particular use is permissible or not can be determined by a simple administrative procedure with an associated appeals process of the usual kind.

The problem of tracking onward transfers of information remains, however. Suppose the data gatherer passes data about you to another party, and properly informs you of the transfer. If the transferee is outside the jurisdiction of the legal

system, then you will not expect to see any further reports on the movement of your data. A way to diminish this possibility is to make it illegal to furnish data to a party that has not agreed to follow the same set of regulations, a policy similar to the EU approach. If there is a violation and you learn of it, presumably a remedy could be fashioned against the transferor.

The cost of Internet access and especially the cost of digital storage continues to decline precipitously. The retail cost of a gigabyte of disk in 2006 is under $0.50 in small quantities, and much cheaper in bulk. It is estimated that this cost will be under $0.02 in 2010. This means that $1,000 will be able to buy 50 terabytes of disk, enough to store 200K bytes of information on every person in the United States. It is obvious that, if the present trends continue unchecked, the desire and ability to accumulate vast amounts of personal information, once only the domain of large corporations, will drop easily into the hands of ordinary citizens. Long before that happens we need to face the potential consequences.