Privacy and Data

Parvez Anandam
Brenda Hernandez
Jessica Miller
Santeri Voutilainen
Vitaliy Zavesov

December 7, 2005

University of Washington CSE P 590TU
University of California Berkeley PP 190/290-009
University of California San Diego CSE 291 (C00)

Authors

Chapter 1: Jessica Miller

Chapter 2, Sections 2.1 – 2.7: Parvez Anandam

Chapter 2, Sections 2.8 – 2.12: Vitaliy Zavesov

Chapter 3: Brenda Hernandez

Chapter 4: Santeri Voutilainen

Chapter 5: Parvez Anandam & Vitaliy Zavesov

Contents

1 The Complexities of Privacy

With the help of technologies such as the Internet, video cameras, cell phones with cameras, and most recently sensor technologies like RFID, recording of personal information is occurring now more than any other time in human history. Alongside this influx of massive amounts of recording is a gap in policy and public knowledge (e.g., many Americans refer to their "right to privacy" without knowing the word 'privacy' doesn't even appear in the U.S. Constitution or Bill of Rights). Given this technical capability for massive recording and lack of coherent, uniform policy (in the United States and also globally), many key questions start coming to the fore. Namely, to what extent should one be able to assert control over their personal information that is collected by various organizations? What role should technology play in protecting privacy? What role should policy play? The complex interaction between policy and technology has become a research hotbed for the fields of law, economics, computer science and information systems, psychology, and human-computer interaction over the last decade. In this paper, we survey literature from these fields to understand the interaction between policy and technology with respect to personal privacy. Working from this survey, we come to our own conclusions on how we (four computer science students and one policy student) think technology and policy should move forward in personal privacy landscape.

In chapter 1 we investigate why personal privacy is of such a concern today. Particularly, we discuss several key complexities of privacy and how technology has pushed on and

caused peoples' attitudes of privacy to evolve. Next, in Chapter 2, we will explore the human organizations that want to collect/analyze personal information, the technologies they are using (or will use) that collect, analyze and store personal information, and the benefit/cost to the public of giving up their personal privacy to these organizations. Next, in Chapter 3, we will explore governmental and corporate policies (or lack thereof) that have been developed to protect privacy. We will look at laws in both the United States as well as internationally. We will also try to look for measures of success in these policies. Since personal privacy is such a concern, much research is being done in the technological domain about how to provide awareness without forsaking personal privacy. In Chapter 4, we will investigate current research that exposes data while maintaining privacy. This chapter will conclude with policy recommendations, with justifications, on why or why not these current methods are sufficient to adequately protect privacy while providing access to the data. Our last chapter concludes by summarizing what we have learned and where we would like to see both technology and policy go from here.

1.1 Defining privacy

One of the complexities of privacy is that it is a very tough concept to define. In this section we present several definitions of privacy and select one definition that we will refer to throughout the paper. Entire books have been written trying to explore what exactly privacy is[1]. Clarke[2] defines privacy as "the interest that individuals have in sustaining a 'personal space', free from interference by other people and organizations" and then applies privacy to several dimensions: bodily privacy, privacy of personal behavior, privacy of personal communications, and privacy of personal data. Since the

1980's the close coupling between computing (personal data storage) and communications has created a new privacy dimension, "information privacy"[2,3]. Clarke[2] defines information privacy as "the interest an individual has in controlling, or at least significantly influencing, the handling of data about themselves." This is the dimension of privacy we will be referring to in this paper. However, we prefer to use a slightly different definition given by Culnan[4] that not only takes personal information usage into account but also personal information collection: privacy is the ability to "control the terms under which their personal information is acquired and used" (p. 20). We also borrow Culnan's definition[4] of personal information: "Personal information is information that can be associated with an identifiable individual" (p. 20).

Although, we refer to the above definition of privacy throughout the rest of this paper, we recognize privacy is highly subjective and is socially, contextually, and culturally situated. Depending on the context one is in and how one thinks about privacy, very different notions of privacy can apply. For example, information shared with family members is likely to be considered private in the context of sharing the same information with one's bank. Yet, many families consider financial information to be private and are not shared among family members, but it is quite accepted for banks to know this information. Palen and Dourish[5] borrow from Irwin Altman to characterize privacy as a dialectic and dynamic process,

> While traditional approaches understand privacy as a state of social withdrawal, Altman sees it as a dialectic and dynamic boundary regulation process. As a dialectic process, privacy regulation is conditioned by our own expectations and experiences, and by those of others with whom we interact. As a dynamic process, privacy is understood to be

> under continuous negotiation and management, with the boundary that distinguishes privacy and publicity refined according to circumstance. Privacy management is a process of give and take between and among technical and social entities—from individuals to groups to institutions—in ever-present and natural tension with the simultaneous need for publicity. (p. 129)

1.2 Privacy concerns and levels of concern

In addition to privacy being a very difficult notion to pin down, there are also a number of different concerns about privacy with respect to technology as well as levels of concern. Culnan and Armstrong[6] identify two kinds of information privacy concerns: (1) the concern of unauthorized access to personal information as a result of a security breach or absence of internal controls, and (2) the concern about the risk that personal information provided for one purpose may be re-used for unrelated purposes without the individual's knowledge or consent (i.e. concern about secondary use) (pp. 105-106). The second concern includes sharing personal information with third parties that were not involved in the original transaction. Smith, Milberg, and Burke[7] identify two additional concerns. The first is general anxiety about personal data collection. The second concern mentioned is about the ability to correct errors in one's personal information.

Alan Westin, who has been researching information privacy for over 50 years, is perhaps most well known for his distinctions in different levels of concern[8]. Westin distributed a survey about privacy attitudes to a group of American consumers on three occasions, in 1995, in 2000, and again in 2003. In each of these distributions he found three groupings of privacy concerns: the privacy unconcerned, the privacy fundamentalists, and the pragmatic majority. These groups differ significantly in

their privacy preferences, regulatory philosophy and attitudes about privacy (see Table 1).

Privacy unconcerned	**Privacy pragmatists**	**Privacy fundamentalists**
• don't know what privacy "fuss" is about • eager to get discounts by giving up personal data • generally trust of business • worried about government action to protect privacy	• look to see benefits offered • want to know privacy risks • want to know privacy safeguards promised • will decide if they trust company/industry • if worried, will seek independent verification that privacy promises are followed • if still distrustful, will support government action	• generally reject offered consumer benefits in exchange for personal information • assume business will misuse consumer information or violate promises • reject self-regulation or industry guidelines • want privacy legislation, enforcement, and consumer right to sue

Table 1: Characteristics of privacy clusters of American consumers (Westin, 2003)

1.3 Evolution of privacy and privacy policy with technology

As if differing notions of privacy, different concerns about privacy and varying levels of concern about privacy weren't enough to muddle the privacy landscape, privacy and privacy policy also evolves with technology. Agre[9] explains that, "As new technologies are adopted and incorporated into the routines of daily life, new wrongs can occur, and thee wrongs are often found to invalidate the tacit presuppositions on which ideas about privacy had formerly been based" (p. 7). The first well-documented instance of this occurring came with the advances photographic technology. Early in American history privacy existed primarily in relation to physical property. Consequently, there were

protections for physical property and against battery to one's person. Liberty meant "freedom from actual restraint" (Warren & Brandeis[10]). As photographic technology advanced, photographs could be produced without sitting for them – therefore allowing one to take pictures surreptitiously. In reaction, Warren and Brandeis wrote a landmark article in 1890 that was published in *Harvard Law Review* in which they call for a new sort of right – the right "to be let alone". This is an excellent first example of how technology pushes on our notions of privacy and in response policy protections are evolved.

In the late 1960s and early 1970s large, centralized databases full of personal information caused another push on notions of privacy and resulted in an enforceable code of practices that became to be known as "data protection"[9]. In United States, this notion was embodied in the Code of Information Practices that stated that organizations collecting personal information about individuals had certain responsibilities and individuals had rights against organizations in possession of personal information (ibid). We will expand on the notion and tenants of "data protection" in Chapter 3.

Then came the power of networks in the 1980s and 1990s. The Internet connects databases (which have grown exponentially in size and variety), allows for new communication media, and does it all in a digital, easily-capturable form[9]. On top of this add new techniques for inferring information from the field of data-mining and other sorts of networks that provide an infrastructure for tracking the movement of people and things (e.g. the cellular network). At the same time, privacy-enhancing technologies

(such as public-key cryptography and digital cash) have made architectural choices quite rich for protecting privacy (see Chapter 4). Networking technology has also given way to a new public sphere, "Privacy activists and concerned technologists have used the Internet to organize themselves, broadcast information, and circulate software instantaneously without regard to jurisdictional boundaries" (ibid, p. 4).

All of this is continually pushing on privacy notions and policy. An evolution in the way people are conceiving of privacy is shown through Westin's three deployments of his survey about consumer privacy attitudes (see Table 2). Notice that since 1995 the group of privacy unconcerned has shrunk while the group of privacy fundamentalists has grown while a majority remain privacy pragmatists. This data does not isolate the effects of technology on privacy attitudes alone. Surely the fact that new national privacy-invasive security policies that began to be proposed and enacted in after 9/11 (2001) has contributed to the 2003 survey results that show 1/3 of the American population grew to be privacy fundamentalists. This data highlights the tension that exists between to the values of security and privacy.

	Privacy unconcerned	**Privacy pragmatists**	**Privacy fundamentalists**
Westin 1995	20%	55%	25%
Westin 2000	12%	63%	25%
Westin 2003	11%	52%	37%

Table 2: American population estimates by privacy cluster (Westin, 2003)

With considerable interest in ubiquitous and pervasive computing, the challenge of address privacy concerns is likely to only grow in the future. These systems propose the use of hundreds or even thousands of sensors and other computational devices spread throughout a room, a building, a city, or other environments. These architectures have people wearing sensors, carrying sensors or even have them embedded. Recognizing that privacy concerns may trump adoption and acceptance of this technology, the Ubicomp community has begun to think of various models, requirements, and architectures for privacy in ubiquitous environments[11,12,13].

Privacy is a notion that is hard to define, is subjective and contextually dependent, and is a moving target. In the past few decades, notions of privacy have increasingly been challenged by fast-paced development of new technology. What can technologists do to help preserve privacy? What can policy makers due to preserve privacy? How can technology and policy work together to preserve privacy? With a better understanding of the slippery concept of privacy in hand, the following chapters will try to answer these questions by looking a closer look at current invasive technologies and the organizations that use them, policies that have been enacted in both the United States and globally, and privacy-preserving technologies.

2 Personal Privacy – Organizational threats and associated technologies

In this chapter, we explore the invasions of privacy by both the public and the private sector. These invasions are typically portrayed as a trade-off: one must give up a certain amount of privacy to obtain something else. For governments, that something is a citizen's physical security. For corporations, that something is typically the safeguard of the corporation itself, so that it may continue to employ individuals.

As noted in Chapter 1, privacy has vastly divergent meaning for different people and that a normative definition must involve the concept of control. Beate Rössler says[14]: "Something counts as private if one can oneself control the access to this 'something'". It is in this context that we survey the current privacy landscape. An exhaustive list of threats would, and does, fill up books; we therefore concentrate on topics that are currently the focus of controversy.

2.1 The power of correlating pieces of public data

Before we look at new invasions of privacy, it is interesting to look at data that has been public for a long time, the retrieval of which has experienced a new dynamic with the advent of the World Wide Web. The tremendous facility with which one can correlate seemingly unrelated pieces of information is unprecedented. An example is most beneficial in driving the point home.

Suppose we wish to roughly estimate the real estate equity an individual has, just by knowing the person's name and the city of residence. This equity is often a substantial portion of an individual's net worth and therefore gives some indication of the latter. For this example, we will assume the person lives in the Seattle area.

First, we need to find out where exactly the person lives. Online phone lists make that easy. Dex Online[15] promptly gives us the phone number and address of the person.

Next, we find out the price paid for the house. The King County parcel viewer[16] makes that equally easy. Additionally we obtain the King County estimates of the value of the house for property tax collection purposes. A quick check of the market value of houses sold nearby[17] provides an independent check of whether the property tax appraisal is in the right ballpark.

This is the beginning of a profile that would be of interest to many companies and to some individuals. It takes a matter of minutes to compile manually. This can also easily be automated. Further, if the number of databases is augmented, the profile becomes increasingly detailed.

While the data may have been publicly available hitherto, it was much harder for an individual to obtain it. Certainly there was a big barrier to discovery to those who were not in close proximity to the subject: someone in Shanghai would have had a hard time gathering this information about someone in Seattle. What makes this a privacy matter,

by our definition, is the fact that the subject of the search has no control over who accesses this information.

Another point of note is that although corporations sometimes offer ways of keeping personal information to themselves (however lax they may be in practice), there is no "opt-out" policy provided by governments. While the richest man in the world can force the telephone company to not publish his address, he cannot force King County to keep private a description of his house[18] if one happens to know that he lives at 1835 73rd Ave NE, Medina, WA 98039[19].

Even the mere fact of compiling several tidbits of information, all of which are publicly available, into a convenient report is deemed an invasion of privacy by some. A highly publicized recent event is the July 14th CNET compilation[20] of such a report about Google's CEO Eric Schmidt using, ironically, Google to perform searches on him. Google shot back by black-listing CNET reporters for a year[21].

2.2 RFID Passports

In the wake of the September 11, 2001 terrorist attacks, the Department of Homeland security made a strong push for more stringent checks at ports of entry. One of the initiatives currently under development is to enable passports to store a large amount of digital data. Not only would U.S. passports need to possess this technology, but so also would passports from all 27 countries whose citizens are currently able to visit the U.S. without a visa. The U.S. threatens to revoke this no-visa access to the countries that do

not comply. This is the result of the Enhanced Border Security and Visa Reform Act of 2002[22].

The technology initially proposed was a RFID (Radio Frequency IDentification) chip[23] without any form of access control. RFID chips can be read from a distance. Passport services originally claimed that these chips could only be read at a distance of a few centimeters and therefore illicit access was not an issue. Security experts, however, demonstrated that these chips could in fact be read at distances a hundred times greater than that.

In an International Herald Tribune article, Bruce Schneier explained[24] the privacy implications of the first version of the technology: "It means that passport holders are continuously broadcasting their name, nationality, age, address and whatever else is on the RFID chip. It means that anyone with a reader can learn that information, without the passport holder's knowledge or consent." Since the holder of the passport has no control over who can access her personal information, our operating definition makes it clear that is a violation of her privacy, not just by governments, but by anyone with the right technology.

The outcry from the public and security experts prompted passport services to reconsider their initial attempt and to propose a way of addressing the privacy concerns. The current solution is Basic Access Control (BAC)[25,26]. The official would have to first manually scan the optically-readable data in the passport to obtain a key that can be used to

authenticate the reader to the RFID chip. The subsequent communication between reader and chip is encrypted.

If the BAC system works as advertised (as usual, the devil is in the details), the owner of the passport has control over when and to whom she provides her personal information. This greatly diminishes the privacy concerns.

This is not to say that there are no privacy concerns at all, however. There is nothing stopping governments from maintaining a database of the authentication keys and using them for police surveillance, for example.

2.3 Real ID

US citizens have traditionally been opposed to a national ID card. In a federation, the states are free to choose what kind of ID system they wish to have. The American people are going to get a national ID card soon, however: the Real ID Act of 2005[27] makes sure of that.

The technology of choice that the Department of Homeland Security will dictate to the states is still in the works. The CAGW (Citizens Against Government Waste) says[28]: "Currently, two main forms of protection are being considered: using magnetic stripes or two-dimensional (2-D) barcodes, or embedding contactless integrated circuits such as radio frequency identification (RFID) chips into driver's licenses."

No matter what the technology is, Bruce Schneier believes that this will make theft of private information easier. “Assume that this information will be collected by bars and other businesses, and that it will be resold to companies like ChoicePoint and Acxiom. It actually doesn't matter how well the states and federal government protect the data on driver's licenses, as there will be parallel commercial databases with the same information.”[29]

2.4 Erosion of Privacy by the USA PATRIOT Act

Forty five days after the September 11, 2001 terrorist attacks, legislation with significant impact on individual privacy was passed: the USA PATRIOT (Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism) Act[30].

The passing of this act has had a polarizing effect on American society. To those who perceived it as an effective measure to monitor terrorist activity, it was hailed as long overdue. To those who felt they had to give up too much of their privacy for little improvement in personal safety and security, the price was not worth paying. Hartzel and Deegan observe: “… what makes us individual also makes us different in terms of what we value and what we fear. The strong and diverse reaction to the USA PATRIOT Act is a prime example of the lack of consensus in our society as far as what is a fair and just response to terrorist attacks.”[31]

Section 216 of the Act broadens the scope of surveillance by law enforcement officials from just obtaining a list of telephone numbers called by a suspect to "…routing, addressing, and signaling information". The concept of a pen register now encompasses list of email addresses, web URLs, etc. Privacy experts argue that the web browsing history of an individual, for example, provides far more information than the list of phone numbers that individual dialed; this is one of the reasons that the act is deemed so invasive[32]. While a court order is required for installing a tracing device, the authorities do not have to demonstrate why they suspect the individual in order to obtain one. The Sunset clause of the Act, which ends several provisions on December 31, 2005, does not apply to the increased reach of the pen register, trap and trace capabilities.

The American Civil Liberties Union (ACLU) has taken a strong stand against the USA PATRIOT Act, arguing that it erodes the privacy of individuals[33]. To counter some of the accusations of the ACLU, the Department of Justice has a web site called "Preserving Life & Liberty"[34].

2.5 Adware and Spyware

Adware is installed software (usually bundled with a shareware program of interest to the user) that continuously displays advertisements on the user's computer. Spyware is typically adware that gathers and records user information without the user's knowledge or consent. Gator from Claria Corporation[35] is the most well known of spyware programs.

The SPY ACT[36], passed in 2004, should help curb the spread of spyware, at least those spread by U.S. entities.

2.6 Targeted Advertising

Targeted advertising has naturally held great appeal for advertisers. Since the advent of advertising, certain groups have been exposed to certain ads. What the Web makes possible is the ability to very inexpensively target individuals by finding out their tastes and preferences. Whether this is of value to society or whether this is an invasion of privacy is a topic of debate.

It is easy to label certain practices of advertisers (such as the erstwhile practices of DoubleClick) as invasions of privacy. Maintaining databases of user data, collected without the user's content (through cookies, or worse, through spyware) is not easily condoned.

Google's AdWords program[37] is an example of a program that many people would not typically term invasive. It targets and displays ads based on keywords that are interesting to the user; many would call that a useful feature. However, Gmail, Google's mail service, also uses AdWords to analyze the content in all the emails that someone sends or receives. This includes emails sent to the recipient by unsuspecting third-parties[38]. Since the ad targeting system is proprietary, individuals have no recourse but to trust that Google is not making use of the data in nefarious ways. Stopping from using Google completely is becoming less and less of an option.

2.7 Supermarket loyalty cards

At first glance, supermarket loyalty cards should have little to do with privacy. They are marketed as a way to pass on savings to loyal customers. Whether the savings are real or not does not concern us here; what does is the vast amount of personal data that is gathered by enticing customers to use the cards. There is no way for shoppers to not use the cards without paying a "privacy tax"[39].

This data is used for marketing and for pricing decisions. Often loyalty card membership requires one to give out a driver's license number or social security number. Having, for example, the number and kind of contraceptives bought during the year stored permanently in a database is not something one envisions while signing up for the cards. That this information can sometimes be made public for all to see is even more worrisome.

Stop & Shop, the largest grocery chain in New England, partnered with Smartmouth.com in 2001 to provide customers with nutritional information on their purchases. The problem was that all one needed to access someone's information the first time was their loyalty card number; this number was printed on Stop & Shop sales receipts.[40]

2.8 Data and Information Privacy at the Workplace and at Home

In the previous sections, we have seen ways in which privacy has been encroached upon by government organizations in search of national security and criminals trying to steal information for personal gain. However, this does not comprise the bulk of the privacy invasions one is subjected to.

Most people's lives are usually centered on their work and their home. They often get comfortable in these environments, letting their guard down in conversations with friends, coworkers, spouses, or boyfriends/girlfriends. People expect to be able to browse the Web freely and use their credit cards to purchase items online. What most people do not realize though is that their actions could be monitored at any time, and quite often legally. The monitoring could be done by their employers trying to see how good of a job a particular employee is performing. The monitoring could also be done by vendors trying to identify buying trends. It could be done by jealous spouses or parents enforcing their children's curfew.

2.9 Video and Audio Monitoring

While thieves and hackers are trying to bypass the law often at their peril, large corporations can track the actions of employees and customers legally. Moreover, such companies have entire teams of professional engineers backed with sufficient funding, whose primary responsibility is to develop and install tracking and monitoring devices. The simplest of these are omnipresent video cameras installed in offices, businesses, and

elevators. These are placed with the goal of providing security to both the employees themselves and the company. Businesses often use cameras to prevent theft. It is also common for employers to use video cameras to track their employee's whereabouts and actions to make sure they are doing good work. The cameras could be displayed in the open, or could be cleverly concealed. In any case, they are clearly encroaching on people's right to privacy.

Security video cameras and audio "bugs" have been in use since mid-20th century. Today these are considered old technologies. Their implication on personal freedoms troubled generations of prominent writers, journalists, and politicians. Both George Orwell in "1984" and Ray Bradbury in "Fahrenheit 451" have described societies where people are monitored 24 hours a day by the "big brother" (the phrase was coined by Orwell), and how diminishing and inhumane such monitoring is. These writers' main concern was that such video and audio monitoring is rarely done in the interests of the person being monitored, but more so in the interests of the person who does monitoring, whatever these interests might be. If fact, this monitoring could be used as a powerful tool of control, or it could be used for the purpose of extracting personal or sensitive information which the person monitored would otherwise be unwilling to divulge.

Not surprisingly, there are a number of laws which regulate video and audio monitoring passed in 1950's and expanded with time. Audio monitoring is both federally and state regulated, but always requires consent of either one of the people engaged in the conversation (such as in the state of New York) or consent of all the people being

monitored (such as in California)[41]. These laws are rigorously enforced – breaking them could lead to serious consequences. Video recording on the other hand is almost always allowed except for places of expected privacy, such as bathrooms, dressing rooms, and adult bedrooms[42]. This is surprising because it is possible to "read lips" from video recordings to extract information that the ban on audio recording is aimed to protect. Audio recording is more intrusive than video however, because it could be done through walls and solid objects and at much greater distances (over 800 meters in the open)[43]. Video monitoring is also regulated by property laws which make it illegal to install cameras on another persons' private property.

Video monitoring might seem intrusive, but it has its benefits. It is critical in preventing and resolving criminal activities, especially monitoring of public areas in businesses, unsafe areas such as elevators, and bank ATMs. In our opinion, with strict regulations governing the use of audio and video monitoring devices, their benefits outweigh any inconveniences they cause.

2.10 Phone Call Monitoring

Given that phone calls, including cellular phone calls, are by far the most common way to communicate between two people, a great deal of information could be extracted by the way of wiretapping, or as it often called call monitoring. Similarly to video and audio recording, wiretapping is an old technique well known from spy movies. Also similarly to video and audio monitoring, there are numerous laws that regulate it.

Under the federal law, the employers might monitor business-related phone calls. As soon as the employer realizes that the phone call is private however, he/she must immediately terminate eavesdropping (unless the employee was previously clearly told that he/she is not allowed to use company phones for personal communication). California law goes even farther to require any phone monitoring to be reported to the parties involved[44]. In the private sector, this means that the parent cannot record their children phone calls without notifying them. Of course, as with all monitoring, law enforcement agencies are an exception if they have a court order. Outside law enforcement and employer monitoring of company phones however, wiretapping is strictly illegal. Phone companies and cell phone service providers can listen in on conversations for technical maintenance reasons only with one of the involved parties' consent. If someone suspects that his/her phone is being tapped, this person can contact the phone company which can and is required by law to detect if it is in fact the case[44].

Contrary to common logic, cell phone conversations are no easier to tap then wired phone conversations. Although the signal from a cellular phone can travel up to 20 miles, it is digital and is usually encoded. It is very difficult to intercept and interpret cell phone signals without using law enforcement-grade scanners. Few people know however that even carrying a cell phone might cause a privacy risk. The Federal Communications Committee has mandated that cellular phone providers need to be able to locate 911 calls within 100 feet. This is done either through triangulation of the signal or by using GPS chips embedded in cell phones[44]. Recently a new industry started to emerge that uses this feature (called E-911) to manufacture monitoring devices. Such devices could be used by

parents to monitor their children's location (given that their children are using cell phones) or could be used to deliver targeted advertisement (i.e. to people in a certain location). The security of cellular phone signals does not spread to their wireless counterparts which use analog radio signals to communicate with the base. These signals could be picked up by radio scanners (although only at short distances). Older phones are especially vulnerable because they use lower frequencies which are easier to detect. Newer phones also have scramblers and other security features. On the legal front, the Counterfeit Access Device Law prohibits manufacturing or importing devices that allow intercepting cellular, cordless phone, or other wireless device signals.

Voice over IP communication allows transferring phone signals over the Internet (or other networks). This is hailed as an emerging technology providing affordable alternatives to expensive long distance calls teleconferencing. The fact that the signal is transferred over the Internet however is making phone conversations susceptible to hacking. More importantly however, the laws protecting phone conversation privacy do not always apply to VOIP. Employers and other network administrators are allowed to intercept and record these conversations if their computer systems are used.

As a final thought, the laws and regulations governing phone call monitoring apparently do not prohibit selling of wiretapping devices (other than the ones intercepting wireless signals). Companies like Voice Print[45] offer high quality phone tapping devices, apparently to be used by employers and/or in countries where phone monitoring is allowed.

2.11 Computer Monitoring

Since computer communication is more recent than video, audio, or phone monitoring, the law is generally less specific on what constitutes a violation of privacy punishable by law. The federal Electronic Communication Privacy Act prohibits interception of the content of wired or wireless electronic communication by a third party without the consent of the people monitored. However there is ambiguity on what constitutes the "content" of communication and who is considered the "third party". Although Internet browsing is generally covered by this law, only the content of the web pages visited is protected. It is allowed to record the addresses of the visited pages, which is routinely done by the Internet service providers. Similarly, although the content of email messages is protected under the law, the addresses of the message recipients are not protected. [44]

Even then, the Electronic Communication Privacy Act does not usually apply at a workplace. Employee computers are usually owned by the company they work for, which allows it to monitor all computer activity. In fact, according to the American Management Associations, over 63% of all companies monitor employee Internet connections[46]. An interesting case study of what constitutes legal or illegal monitoring of computer activity could be done with online shopping. When an employee enters his/her credit card information when shopping at Amazon.com using a company computer (which employees are often allowed in their free time), this private credit card information gets recorded by the system/network administrator. Beside ethical and legal reasons for why this should not happen, this could also be dangerous. More people now

have access to this private credit card information and potentially can use it for personal gain or leak it to the public where it could be picked up by common crooks. Network administrator logs become a clear target for hackers and a single point of failure. Recent news of break-ins into corporate databases is a major reason for concern.

There is a multitude of technologies and techniques available for monitoring computer use. These include primitive hardware devices which are used with keyboard and/or mouse adapters which plug into computer ports. These are often easy to detect and disable. The devices also include a number of software applications, such as eBlaster[47], which either record keystrokes and mouse clicks or periodically do screen captures. This information could be sent to the monitoring party via email or be stored on a shared network drive.[46] Such software applications are also easy to detect and disable with the help of anti-virus software, although employees often do not have sufficient system privilege to install it. In general, these applications target people who either do not have sufficient privilege to install anti-virus programs, or less-computer savvy individuals who don't know how to do this. Monitoring devices also provide a serious security loophole, since the captured data might not be properly secured and could be intercepted by third-party criminal elements. Nevertheless, keystroke monitors and screenshot monitors are widely available for use to both large corporations and individuals. Some of these devices are very inexpensive and have free trial periods.

2.12 Radio Frequency Identification (RFID)

RFID is an old, but rapidly developing technology which allows identifying objects. These IDs are made of a tag, which is a microchip with an antenna mounted on a substrate. The IDs are attached to the tracked objects, and could be scanned by a reader device which emits radio waves and receives information back from the tag. Such information usually contains unique identification, but could also include additional data up to several kilobytes in size used to characterize the object. The reader device then passes this information on to a computer that could make it available to be accessed over the Internet or another network[48]. There are two broad categories of RFID technologies: active which transmit the signal using energy supplied by a power source, and passive which simply reflect radio waves back to the reader device. Active RFID systems have a much greater range, up to 300 feet[48]. Such systems are used to track larger objects which need to be scanned from afar, like railroad cars. Passive systems have a much smaller range, and are typically used to track merchandise in stores, such as to prevent theft or for inventorying.

RFID technology is commonly used in many areas. Besides tracking merchandise in stores, production lines, or supply chains, RFID could be used to track food produce expiration dates, or to be used by banks as means of personal identification. Many banks are releasing "smart cards" with RFID in them which allow uniquely identifying user accounts. Correspondingly, there are many privacy issues this might cause. For example, criminals might be able to scan RFID tags in passers-by's jewelry to find targets for attacks, or scan RFID tags in bank "smart cards" to steal account information.

Similarly to using loyalty cards, stores might be able to use RFID technology to identify customers and their purchases for directional advertisement. In extreme cases, the tags could be used to identify people themselves (by tracking what they are wearing for example) and monitor their whereabouts. This of course could be used by companies tracking their employees or by the government tracking specific individuals. Surprisingly, there are no laws which specifically target the use of RFID technology, leaving its regulation to "catch all" privacy laws[49]. This void in applicable regulations would become more obvious as the use of RFIDs spreads, which could in turn cause passing of new laws.

All in all, RFID technology provides just another tool to legally monitor customers and/or employees. This technology has many proponents and critics. The proponents point out its tremendous value to the industry where RFID tags help to cut down inefficiency. The many opponents are concerned with how RFID tags are being used when products leave the supply chain and become private property. Similar concerns were expressed when bar coding technology first appeared in the 1950s. As with any new technology, the lack of specific laws and regulations governing its use causes public concern. This is no reason to abandon the technology altogether though. The public concerns would at least be partially alleviated as new laws are passed into effect.

In conclusion, there are many corporate, government, or even non-profit entities that are interested in installing monitoring devices to tack individuals. Often, such monitoring is allowed under the law. It is ultimately the responsibility of the person monitored to learn

the laws governing the issues of privacy to preserve sensitive information. There are also clear gaps in the law, especially with regard to new or developing technologies, such as RFID. It is a classic example of technology marching forward at a pace not anticipated by government and public organizations. While bringing the promise to make things better, these technologies often provide more ways to encroach on people's privacy. It is the responsibility of the government to streamline the process of investigating these issues in a timely fashion and implementing measures to combat them.

3 Current Policy to Protect Privacy

3.1 Privacy not a Constitutional Right and its Opposition to Free Speech

In the U.S privacy protection through policy is an issue of debate for many as some believe that the right to privacy is an unassailable human right while others think of it as dependent on the individual's choice. Yet, the U.S legal system treats privacy as a personal property right that may be disposed of as one sees best, rather than an unassailable human right. Constitutionally there is no explicit right to privacy. Nevertheless, the Supreme Court has ruled that there is an implicit, limited constitutional right of privacy based on a number of provisions in the Bill of Rights, such as the right to privacy from the government surveillance into an area where a person has a 'reasonable expectation of privacy' under the Fourth Amendment.[50] This protection is not a general one, however. Information held by third parties, such as telephone calling records, is generally not protected unless safeguarded by a specific statute. The implications of the right to privacy not being protected within the Constitution are that citizens have to protect specific rights through specific policies after legislators realize that certain violations are being committed and restrictions need to be placed through the enactment of policies.

As stated before the right to privacy in the U.S is a sensitive issue that many believe should ultimately be decided by the individual. Therefore,

privacy law has an important role in protecting individual self-determination and democratic liberation. By providing access to one's personal data, information practices, the law seeks to structure the terms on which individuals confront the information demands of the community, private bureaucratic entities, and the State. [51]

Further arguments regarding regulations do arise specifically in trying to distinguish between the juxtaposition that many people see with freedom of speech in opposition to privacy. One way it has been recommended to deal with this problem is by following the fair information practices which generally require: (1) the creation of a statutory fabric that defines obligations with respect to the use of personal information; (2) the maintenance of processing systems that are understandable to the concerned individual (transparency); (3) the assignment of limited procedural and substantive rights to the individual; (4) the establishment of effective oversight of data use, whether through individual litigation (self-help), a government role (external oversight), or some combination of these approaches.[51] Fair information practices can best be thought of as fulfilling two roles regarding communicative discourse. First, these rules help maintain the boundary between public discourse and the other realms of communication. This role is largely fulfilled by the nondisclosure subset of fair information practice. Second, standards of fair information practices serve to safeguard deliberative democracy by shaping the terms of individual participation in social and political life. [51]

Therefore, as argued by Swarchtz, in concrete and specific ways, "upholding certain kinds of information privacy speech restrictions could affect the protection of other

speech." If the legal system accepts the propriety of laws mandating 'fair information practices,' people may become more sympathetic to legal mandates of, for instance, fair news, reporting practices or fair political debate practices. Second, in place of privacy law, Swarchtz argues that it is preferable to protect information privacy through privacy-enhancing techniques such as technological self-protection, market pressures, restraints on government collection and revelation of information, and recourse to social norms. [51] In a sense, it is easier to protect privacy with these measures rather than trying to define the validity of specific information using the freedom of speech argument. By ensuring 'fair information practices' and privacy-enhancing techniques, in fact, the right to freedom of speech is also being protected.

3.2 Electronic Surveillance Laws

Electronic surveillance laws are often argued to be laws infringing on personal privacy. Electronic surveillance involves the traditional laws on wiretapping—any interception of a telephone transmission by accessing the telephone signal itself—and eavesdropping—listening in on conversations without the consent of the parties. There are many of these electronic surveillance laws and each one has specific regulations and motives to be requested and utilized. These laws include The Wire and Electronic Communications Interception and Interception of Oral Communications Act. This law typically requires a court order issued by a judge who must decide that there is a probable cause to believe that a crime has been, is being, or is about to be committed. The government can wiretap in advance of a crime being perpetrated and judges seldom deny government requests for wiretap orders.[52] Another one is the Foreign Intelligence and Surveillance Act 1978

targeted for U.S citizens and permanent resident aliens and there must be a probable cause to believe that the person is engaged in activities that "may" involve a criminal violation. Suspicion of illegal activity is not required in the case of aliens who are not permanent residents. Also, no legislative limits on U.S government electronic eavesdropping is carried out overseas.[52]

More electronic laws include The Electronic Communication Act of 1986 which sets standards to access cell phones, e-mail and other electronic communications and to transactional records (subscriber identifying information, logs, toll records). The pen registers and trap and trace device statute which governs real-time interception of the numbers dialed or otherwise transmitted on the telephone line to which such a device is attached. The Communications Assistance for Law Enforcement Act of 1994 is a digitally telephony law. CALEA was intended to preserve law enforcement wiretapping capabilities by requiring telephone companies to design their systems to ensure a basic level of government access. [52]

Perhaps the most controversial electronic surveillance law is one that was enacted after the September 11th attacks and is known as the H.R 3162 (the USA Patriot Act) because it significantly broadened the scope of federal electronic surveillance laws. This law adds terrorism offenses, computer fraud, and abuse offenses to the list of predicates for obtaining Title III wiretaps. It also permits roving wiretaps under the Foreign Intelligence Surveillance Act of 1978 (FISA) in the same manner as they are permitted under Title III wiretaps. Pursuant to H.R 3162 intelligence information obtained from wiretaps may be

shared with the law enforcement, intelligence, immigration and national security personnel. Recipients can use the information only in the conduct of their duties and are subject to the limitations in current law of unauthorized disclosure of wiretap information.

As discussed in Chapter 2, H.R 3162 also expands the use of traditional pen register or trap and trace devices (captures the telephone numbers of incoming callers) so that they apply not just to telephones, but also to the Internet communications so long as they exclude "content." These devices may now also be used under FISA without having to show that the telephone covered was used in communications with someone involved in terrorism or intelligence activities that may violate U.S criminal laws. Multi-jurisdictional warrants may be obtained for wiretapping purposes, making it easier to track criminals across borders. [52]

As a result of the enactment of the USA PATRIOT Act a strong need was seen to create a counterbalance that would help protect fundamental rights of Americans. Protecting the Rights of Individuals Act (PRI) (S. 1552) seeks to place reasonable limits on the powers granted to law enforcement and intelligence agencies under the USA PATRIOT Act. PRI would amend many of the Patriot's most troublesome provisions, reasserting traditional checks and balances on the Executive Branch to ensure the proper balance between law enforcement authority and Americans' fundamental liberties. Specifically, PRI: limits the use of secret "sneak and peek" searches to terrorism investigations. It also protects 1st Amendment rights by narrowing the definition of "domestic terrorism." It shields Americans' sensitive personal information from government access without some

specific suspicion and prevents the government from accessing library records without judicial approval. [52]

PRI further ensures that the government cannot monitor what Americans read on the Internet without probable cause. Also, it forbids government data mining without prior congressional approval and requires that the Attorney General provide Congress with basic information about foreign intelligence surveillance. It also reinstates longstanding discovery procedures for the use of foreign intelligence evidence in criminal proceedings and restores the requirement that foreign intelligence must be the primary purpose of surveillance conducted under FISA.

Lastly, it prevents government access to education records without specific facts showing why those records are needed. [52]

Although Americans are aware of the strong need of enacting surveillance laws that can be utilized to help prevent crimes and prosecute criminals; the desire of living a "normal life" in which they are not questioned and tracked for harmless actions is very important. Experts have a difficult time in defining the balance and that is why specific laws were created to help combat crime but also to guarantee Americans of their fundamental rights. The balance created between the issue of protecting privacy and the government's job of ensuring homeland security for all Americans is best represented through the USA

PATRIOT Act and the PRI which create checks and balances that ensures both parties rights are secured.

3.3 Protecting Personal Health Information

Another sector where protection of personal information is very important and strictly monitored is the health sector. The department of Health and Human Services has been careful with its regulations of the HIPAA Privacy Rule whose main goal is to protect the protected health information (PHI) of individuals from covered entities. The HIPAA Privacy Rule regulates the way certain health care groups, organizations, or businesses, called covered entities under the Rule, handle the individually identifiable health information transmitted by electronic media, or transmitted or maintained in any other form or medium, known as protected health information (PHI). Researchers should be aware of the Privacy Rule because it establishes the conditions under which covered entities can use or disclose PHI for many purposes, including research. [53] The Privacy Rule establishes minimum Federal standards for protecting the privacy of individually identifiable health information. The HIPAA Privacy Rule confers certain rights on individuals, including rights to access and amend their health information and to obtain a record of when and why their PHI has been shared with others for certain purposes.[53]

In addition to the Privacy Rule, State and other Federal laws and regulations, such as HHS regulations for protecting human subjects, continue to govern research when applicable. Also, the FDA Protection of Human Subjects Regulations is intended to protect the rights, safety, and welfare of participants involved in studies subject to FDA

jurisdiction. [53] Altogether these organizations and regulations ensure that the rights and privacy of participants are ensured and protected.

3.4 International Privacy Protection through Policy

Two recent developments have increased fears of loss of personal data privacy. First, information and communication technologies used to communicate, store, and manipulate data have dramatically increased the level of information generated and exchanged on each individual. Second, the globalization of trade and finance has meant that states find it increasingly difficult to monitor and control the activities of transnational corporations that move data across national jurisdictions. Yet, this is not a new concern since for decades now countries around the world have been working at increasing their measures to try to protect privacy. For example, in the 1970s, in response to the increased call for privacy, countries throughout the world began to enact data protection legislation. Two crucial international instruments evolved from this movement: the Council of Europe's (COE's) *Convention for the Protection of Individuals with Regard to the Automatic Processing of Personal Data*, and the *Organization for Economic Co-operation and Development's (OECD's) Guidelines Governing the Protection of Privacy and Transborder Data Flows of Personal Data.* These international guidelines have had a profound effect on the enactment of privacy laws around the world. Since then, more than twenty countries have adopted the COE convention, and the OECD guidelines are widely incorporated in national legislation both within and outside the OECD.[50]

The many efforts have resulted in new policies as well as what are considered models of data protection which each country chooses to adopt as they see best fit. In recent years, four models of data protection have emerged: (1) comprehensive laws; (2) sectoral laws; (3) industry self-regulation; and (4) reliance on privacy-enhancing technologies. Comprehensive laws are omnibus legislation establishing broad standards seeking to provide the best legal protection governing the collection, use, and dissemination of personal information by both the public and private sectors. In most cases, there is an official agency that oversees enforcement of legislation. Sectoral laws, in contrast, are more narrowly defined, and regulate specific sectors of government, business, or civic activity. The third model, self-regulation, allows companies and industry bodies freer rein to establish their own codes of practice. Advocates of self-regulation argue that this model of privacy protection is less costly and more flexible in meeting individual preferences and needs for privacy. Finally, some countries may choose to rely on commercially available privacy-enhancing technologies such as encryption and digital cash to protect personal information. [50]

3.5 European Union Directive Approach on Data Privacy

Unlike Americans who tend to be more trusting of the private sector and the free market to protect personal privacy - fearing more the invasion of privacy from the state not the market - the European approach toward data protection is grounded in the concept of privacy as a fundamental human right. The state intercedes between organizations and individuals to create parity, and guarantees this fundamental right to personal privacy through prophylactic protection, including: (1) creating norms for collecting and

processing personal information, (2) establishing avenues for individuals to review information collected about themselves and reviewing the controller's information practices, (3) providing special protection for sensitive data, such as that pertaining to ethnic origins, political affiliation, and religion, and enforcing and overseeing the systems of protection. [50]

The EU Directive is the most influential data protection policy to emerge in the last decade. It was designed to prevent the use of disparate national data protection laws as a barrier to trade within the EU while providing comprehensive data privacy for the citizens of member states. The EU Directive mandates that private sectors in member countries adhere to six key provisions: (1) employees and customers must be adequately informed about how their personal information will be used; (2) companies cannot use such personal data for any activities other than what is made known to the proprietors of the data; (3) individuals must have the right of a review, and means to correct errors; (4) companies must give notice before providing the data to third parties for direct marketing; (5) employees and customers must be given the choice to opt out of a data collection scheme without having to incur any costs; and (6) enforcement provisions must be put in place. [50]

The EU Directive has been crucial in maintaining the European ideal of treating privacy as a fundamental human right and that it is protected as such for all citizens of the member states. The U.S, on the other hand, has compromised with the EU to guarantee access and success in business transactions with the member states but has had trouble

with the provisions mainly because of the difference in definition of privacy for America and the European Union. Both states agree in the importance of privacy for their citizens but they differ in the approaches to be taken as individuals and at times as leaders of a nation.

3.6 Other International Privacy Related Laws

In 2004, in Australia, Federal Parliament passed a Bill that completely removed undelivered email, SMS and voice mail messages stored on service provider's equipment from the protections of the Telecommunications Act of 1979.[54] Police and many other state government agencies have thereby been given new powers to intercept undelivered communications, including when investigating very minor suspected offenses which was not previously permitted. This law has caused great controversy as the right to privacy seems to be at risk but since previous laws seemed to give criminals an advantage because government officials did not have these newly obtained powers, it was deemed as necessary and now the federal government is working at ensuring that it is correctly enforced by government agencies.

3.7 Recommendations on Policy to Protect Personal Data

Protection of privacy rights in all sectors has its costs and benefits all around the world. The use of policies to protect privacy rights enables the individual or agency to a right of 'checks and balances' and to having a voice if those rights are being violated under a specific policy. The cost of privacy through such policies are not only understanding that

one has to abide by these rules or else there are consequences but also that to enjoy the rights of such "privacy" other rights will have to be sacrificed including freedom. Yet, as the analysis has proven there is always a balance to ensure that individual entities as well as individual citizens are protected. Under any system whether it is here in the U.S or under the European Union individuals will gain privileges and lose privileges but have the comfort that there is a system of accountability for all parties under a system of protection with policy as opposed to a system with no policy protection. A system without policy protection for personal data protection is one that grants more freedom and flexible but has no system of accountability and there is no responsible party for anything that may happen. In the end, it is up to the individual to decide what system works best for them one with protection through policy or one without any protection as their voice has a direct effect on the government's decision of how to respond to personal privacy rights issues. The state will cater to the response of their citizens while trying to ensure that the nation is secure from unexpected threats. Privacy rights are vulnerable to an individual's interpretation, thus, governmental policies are the only safe way to ensure that the individuals' rights are protected while still protecting the other agencies and all involved parties are under a system of accountability.

Currently, although there are several mechanisms utilized to help protect an individual's privacy, some may argue that these mechanisms are not sufficient as many privacy rights are being violated. The main problem is not that there are not enough laws and corporate involvement but that individuals are not aware of their privacy rights and the laws protecting them. When encountered with a privacy problem individuals do not know

how they can do something to solve it. Instead, many citizens are paranoid and just let some of these issues go unresolved. A good solution to this problem, aside from continuing to implement laws that protect individuals' privacy rights, is to advertise to individuals their rights and the laws protecting those rights. That should help minimize violations of privacy significantly.

4 Public yet Private: Analysis of Privacy Preserving Data Mining Techniques

As covered in previous chapters, private data is being collected for a wide range of uses by both private and public parties. This collection causes friction between an individual's wish for privacy and the benefits of information collection, such prevention of terrorist attacks by identifying suspicious behavior. While data on individuals can be collected in various manners, it is stored in a database. From the database it can then be further shared to other parties via access to the actual database or published subsets of the database. The dangers of public access to private data through these databases are demonstrated by Latanay Sweeney's ability to determine former Massachusetts Governor William Weld's medical records through two publicly available datasets: A list of medical records with all explicit identifiers removed, and a voter registration list [4]. The extremes of data privacy with respect to data collection – that of no collection and unfettered access – are both clearly undesirable as they preclude all the benefits of the other. Any solution for preserving privacy while providing access to collections of data of private nature must balance the needs of privacy and usefulness of the exposed data.

In this chapter we will discuss several privacy preserving data mining techniques and evaluate their applicability and performance. We will not discuss technological measures to prevent the collection of data in the first place, but rather how collected data can be shared with third parties without violating the privacy of individuals whose data is shared.

For the purposes of this chapter we assume that the data collector has the individual's permission to collect the information. We will not discuss how to prevent collection of data in the first place. We will consider the data mining techniques with respect to two data usage scenarios: Behavior or trend analysis, such as early outbreak detection or web usage analysis, and investigative uses, such as law enforcement. Both scenarios are relevant to homeland security efforts to prevent and respond to attacks. We will first define applicable terms and concepts, then define several privacy preserving data mining techniques followed by consideration of their applicability and effectiveness, and conclude with recommendations on their usage.

4.1 Definition of terms

As mentioned, data that is collected is stored in a database. A database consists of multiple datasets each of which consists of one or more records each with one or more attributes. Attributes can be of any type such as name, date of birth, zip code, disease, price or time. The identity of a record is the individual to whom it pertains, thus, the identity of a medical record is the patient. The attributes of a record can be divided into three categories based on the extent to which they identify a record[55]:

1. Explicit identifiers, such as names and social security numbers, are attributes that directly reveal the identity of the record.

2. Quasi identifiers, such as zip code and date of birth, are attributes that do not directly expose the identity of a record. However, groups of quasi identifiers may be unique to an individual and thus may allow the identification of the individual

when joined with other datasets containing both the quasi identifiers and explicit identifiers. The set of zip code, date of birth and gender is unique for 87% of the US population, and even country, date of birth and gender are unique for 18% of the population[56].

3. Non-identifiers are attributes that do not provide identifying information.

Although we will refer to datasets and records as pertaining to individuals, the concepts generalize to any data.

For discussing these techniques, we define data privacy as the inability to associate, with high confidence, a record in a dataset with a single individual. The level of confidence is subjective. Is sufficient privacy preserved if a record can match only one of two individuals? How about one of a thousand? Although we don't propose a definite value of sufficient privacy, we assume a high level, and our discussion does not depend on any specific value. The choice of a specific value is a policy decision. Not all datasets require the application of privacy preserving techniques such as those whose attributes are all considered public information. Voter registration lists are such a dataset.

It is important to note the difference between data privacy and data security. Data privacy refers solely to the ability to determine the identity of a record. Data security concerns itself with the preservation and access control of data regardless of whether the data is considered private. Complete data privacy solutions must also handle data security as security vulnerabilities can expose identifiable data.

4.2 Privacy Preserving Data Mining Techniques

The common element among the various privacy preserving data mining techniques is that they aim to anonymize datasets or queries over datasets so that specific records can not be identified with a specific individual. While some techniques retain full control over the databases, most techniques anonymize the dataset so that the full dataset can be made publicly available.

The most commonly known technique is aggregation, which generally produces datasets with counts of occurrences of a value in some set of attributes.

While most techniques exclude any explicit identifiers from datasets containing private data, the following ones allow for publication of explicit identifiers. This is achieved by modifying the datasets in such a manner that the explicit identifiers can not be matched to the actual attribute values. Randomization modifies the value of an attribute by a random amount. Due the nature of random numbers, the attribute's distribution remains statistically equivalent to that of the non-randomized dataset[57]. A related technique preserves the original attribute values but scrambles the values among all the records thus an attribute's value does not necessarily match the record in which it appears. As opposed to randomization, this technique allows accurate existence checks. However, neither technique allows inferences to be drawn between two attributes in a record as the values are either not accurate or may correspond to different identities.

When exposing datasets where a handful of records contain attributes with unique values that could allow identification, it is also possible to block these attribute values by replacing them, not with random values, but a special value to indicate that the value is unknown.

While the preceding techniques modify the source of the data, output permutation modifies the output of query. Once the results of a query have been calculated using the real data, techniques similar to the ones described above are applied to the result dataset before being made public.

The remaining privacy preservation techniques retain consistency of attributes within a record and thus allow inferences to be drawn between the attributes. They do not, however, allow for the inclusion of explicit identifiers.

Query restriction anonymizes results by placing limits on the queries that can be posed against a dataset. Direct access to the dataset is not allowed; instead users submit queries to a query controller. The controller approves or rejects the query based on whether it could violate data privacy. The query controller maintains, a log of all previous queries in order to use them when making approval decisions because appropriately designed non-identifying queries can violate privacy when joined.

Non-identifying unique ids can also be used to anonymize records. This technique replaces any explicit identifiers with unique ids. These ids can be random or

deterministic, based on the explicit identifier. When deterministic ids are used, accurate inferences can be drawn between records based on matching ids.

Multiparty secure computing has also been proposed as a technique for protecting data privacy. Multiparty secure computing uses cryptographic techniques to allow parties to share datasets and perform computations while not having access to anything but their own datasets and the results of the computations. This technique works similarly to attribute value randomization in that the actual attribute values are not available. However, due to cryptographic properties, in this technique the encryption of the values does not impact the result of the computation.

A technique that has gained focus in recent years is the anonymization of datasets by guaranteeing the existence of at least *k* records with identical identifiers. The identifiers are generalized so that they correspond to at least *k* individuals. For example, rather than include the records {Peter Jones, 98122, cancer}, {Peter Johansen, 98122, healthy} the identifiers would be changed so the table contains {Peter Jo*, 98122, cancer}, {Peter Jo*, 98122, healthy}.

4.3 Effectiveness and Usability

While all of these techniques can protect privacy, none of them is a silver bullet and each has faults; some are not applicable or practical to every usage scenario. Each also provides differing levels of privacy protection.

While techniques that provide aggregate or statistically accurate data may be acceptable to many behavior or trend analyses, they are not useful for law enforcement applications that require the ability to draw accurate inferences among records in order to detect suspicious behaviors among a small number of individuals.

Query restriction uses raw, non-anonymized, datasets and thus it can draw accurate inferences among records in the dataset. However, it is not a practical technique for any large scale use. The query controller must, by definition, track all queries that have been issued against the dataset in order to eliminate the possibility of determining the exact contents of a dataset's subset through the issuance of multiple queries. The task verifying that data privacy is maintained grows progressively difficult as the number of previously issued unique queries increases, eventually making it impossible to determine the acceptability of the query within an acceptable time period. In fact, the problem has been shown to be NP-hard[58]. Query restriction also has a saturation point at which it is not possible to issue any new queries without violating privacy. Interestingly, the act of approving or rejecting a query can itself breach privacy protection[59]. Consider a query which asks "How many US Senators have ever had a sexually transmitted disease?" If the answer is ten, then this query is acceptable because it does not reveal the identity of any US Senator with an STD. However, if the answer is a hundred or zero, then the query must be rejected because in either case the answer to "Has Senator X ever had a sexually transmitted disease?" is clear and data privacy has been breached. However, the mere act of rejecting the query limits the possible states to two.

Although randomization and scrambling appear effective at protecting privacy they do have vulnerabilities in edge cases which lower the practicability of the techniques. Although randomization can not expose exact original values, it can reveal limits on the ranges. If the randomization value is chosen from the range [-100, 100], then a randomized attribute with value 180 reveals that the original value must be at least 80. With certain attributes, such as age, this may be considered a breach of privacy. In order to preserve the statistical equivalence of the distributions, it is not practical to use the range $[-\infty, \infty]$; there is a non-zero probability that lower or upper bound can be determined for an attribute. In datasets where the diversity of an attribute's values is limited, scrambling does not effectively protect privacy because the value for a particular individual can be determined with high confidence.

While proxy identifiers, such as unique identifiers, are effective replacements for explicit identifiers, they are not sufficient by themselves. They do not provide protection from identification through quasi identifiers, which, as evident from Latanay Sweeney's exposure of Governor Weld's medical records, are sufficient to identify an individual when joined with other datasets.

Although anonymization techniques such as *k*-anonymity handle the issue of quasi identifiers, they too are susceptible to data privacy violations[60]. The choice of how to achieve *k*-anonymity by generalizing the quasi-identifiers is the main source of data privacy violations. If a dataset is *k*-anonymized using two different generalizations, then those two datasets can be joined on common attributes to reveal a dataset that is not *k*-anonymous. Thus one of two restrictions must be placed on subsequent public releases of *k*-anonymous datasets: Either a secondary release must be based on the initial k-

anonymous dataset, or all attributes of previously released *k*-anonymous datasets must be considered quasi-identifiers and therefore *k*-anonymity must be achieved across all attributes, not just the original attributes[60]. In the former option the first *k*-anonymous dataset effectively replaces the original dataset as the privately held anonymized dataset.

Most techniques also suffer from privacy vulnerabilities related to the addition or deletion of records in the dataset. Consider a dataset maintained by an entity which allows external parties to submit new entries and then subsequently submit queries over the database. Law enforcement databases are an example of this type of database. If the unique identifier technique is used on this dataset, an external party could submit a new record with identity X to the dataset. With a carefully crafted record, the same external party could then submit a query to retrieve the unique identifier. The unique identifier can then be used to find other records for the individual. For example, a detective might enter new record of form {former-lovers-name, unique-salary-amount} and then issue a query asking for all video rental records for individuals with that specific salary. If the detective chose a salary value that is unique in the dataset then the results would with 100% confidence be the video rental records for the detective's former lover. Note that the detective does not even need to determine the identifier. As mentioned, most of the techniques, including aggregation and *k*-anonymity, suffer from this vulnerability.

Another vulnerability that affects many of the techniques is trail matching. Trail matching is a special case of inferring identifiable data by joining multiple data sets. Malin describes a hypothetical situation where hospitals release patient identities as one

dataset and their DNA sequences as a separate dataset[55]. There are no common attributes between the two datasets. If similar datasets are available from multiple hospitals then it might be possible to analyze them and determine that John was patient at hospitals 1, 2 and 3; Brad at 1 and 3; and Bob at 1 and 2. DNA sequence Dx exists in the datasets for hospitals 1 and 2; Dy at 1, 2, and 3; and Dz at 1 and 3. Thus Dx must be Bob, Dy John, and Dz Brad. [55]

The existence of unknown numbers of datasets and the possibility for inferring identifiable information by joining these datasets poses a significant challenge to any privacy protecting data mining technique. Trail matching and the *k*-anonymity vulnerabilities are examples of this general class of vulnerability. The challenge of determining whether privacy can be violated using some combination of existing or proposed datasets is equivalent to that faced by query restriction in determining which queries can be accepted and which must be rejected. This implies that in general, it is not possible to expose datasets containing private information that is guaranteed not to be re-identifiable. The publishers of datasets can not be realistically expected to have complete knowledge of all datasets in existence. This situation is made harder by the ever increasing number of data collectors, each of whom acts independently in releasing their datasets.

4.4 Recommendations

Although it is our conclusion that it is not possible to use privacy preserving data mining techniques to guarantee complete non-identifiability of published data, usage of these techniques does raise the bar for re-identification and thus should be required for all data collectors. Although incomplete, they do provide a basic level of protection. The risk of private data exposure can also be decreased by restricting the entities that can collect individually identifiable information. As discussed in chapter three, this approach is in use in Canada and the European Union countries where legislation or privacy advocates limit the type of information that can be stored and who can store it. This is vastly different in the United States where there are far fewer restrictions on what information can be stored by whom.

Usage of the techniques we discussed should be accompanied by regulations that address their shortcomings. Some of the regulations or legislation required should cover approved usages of private data, limitations on distribution, and penalties for circumvention of protections.

For behavioral and trend analyses we recommend data aggregation with carefully chosen quasi identifiers, data randomization, and scrambling of attributes among records in the dataset. These provide the greatest level of privacy, are simpler to implement than anonymization based techniques, and are compatible with the requirements of these uses. Our recommendation for law enforcement applications is the usage of anonymization techniques such as *k*-anonymity. These techniques combined with processes that allow

for controlled re-identification allow for both privacy and access to data. They allow for a sliding scale of privacy preservation and thus are suited for law enforcement applications where it may be necessary to determine an identity with 100% confidence. Indiscriminate identification can be prevented using policy and technological measures that would require court approval re-identification to take place. One such system, using Selective Revelation and based on *k*-anonymity, has been proposed by Sweeney[61]. Systems Research and Development's (now IBM) ANNA system provides similar functionality using unique identifiers based on hashes of explicit identifiers[62].

In general we recommend limiting the publication of datasets that can be used to join quasi-identifiers with explicit identifiers so as to reduce the problem of inference. The benefits and privacy cost of publishing these datasets should be carefully weighed. In some cases, such as voter registration lists, it may not be possible to avoid publication, but unnecessary publications should be avoided so as to make re-identification more difficult.

5 Conclusion

What does the future hold for the privacy of individuals? Will we all live within a digital Panopticon? Or will we have complete control over who accesses our personal data? The answer will, of course, lie between these two extremes.

With each new disruptive communication mechanism, there is an initial period of chaos where the privacy pendulum swings wildly. The confidentiality of letters took some time to establish; now we trust that letters will not be opened in transit. The telegraph, the telephone, etc. brought about privacy concerns that were dealt with to the majority's satisfaction. It is likely that we are in the chaotic period of the Internet, where privacy concerns have not had time to be addressed. One hopes that they too will be ironed out.

Even if we manage to control the technology, does the looming threat of terrorism change the privacy landscape irrevocably? On the basis of history, there is no reason to believe that. Critics of the current excesses in surveillance often hark back to another era, to the fight against another –ism[63]: the McCarthy witch-hunts for communists targeted certain groups of people with the best surveillance technology of the day. We are ironically reassured by such comparisons: just as the former era ended, the current one should.

That will take hard work, of course, but we seem to be headed in the right direction. The reauthorization of sunset clauses of the Patriot Act is generating healthy debate: six senators voiced their opinion against some of the provisions on November 17, 2005[64].

The emergence of new technologies, such as the Internet, RFID, or cellular phones, has brought new ways to infringe personal privacy. Some of this infringement is pursued by government agencies trying to combat real or presumed threats to national security. Communism was the biggest threat in the 50's, terrorism is the biggest threat today, and, unfortunately, there will emerge new threats in the future. There are also thieves and criminal organizations which seek to infringe on personal privacy for fraud and extortion. And, finally, there are many law abiding individuals and companies that would like to spy on their wives, children, employees, customers for a variety of reasons. To combat these threats to privacy, local and federal governments put laws into effect which regulate what may or may not be monitored. Yet, there is always a lag between the time new technologies emerge and the laws which govern them, simply because some time needs to be taken to learn what the new threats are.

Although it seems that this mechanism of passing new laws into effect works to some degree, the effects of globalization bring additional challenges. Now, through the use of the Internet, criminals might be able to steal personal information across the globe, and across national boundaries. The laws regulating the use of new technologies are local to a specific country by definition, and might not apply to thieves living in other countries. The laws between countries often differ in such areas as the freedom of speech, gambling, and others, which makes things even more difficult. To effectively protect individuals' rights, new and more effective international legal structures are needed. As a minimum, there needs to be a unified set of laws, similar to maritime laws applicable in international

waters, which govern Internet behavior. The United Nations might play an important role here, although it has to fight the will of some regimes on many issues, such as the freedom of religion and speech. Nevertheless, these laws are critical to the future of the individuals' right to privacy.

The Association for Computing Machinery makes that clear in this excerpt from a December 4 draft letter of the ACM to Congress[65]:

> We at USACM […] are acutely aware of the risks to individuals posed by unprotected or poorly protected personal information. For that reason, we're writing to increase your awareness of some of the technical community's concerns in this area. Considering the staggering number of data breaches that have come to light this year, the rapid growth in identity theft, and growing public concern about the safety of their personal and consumer information, the time is indeed right for Congress to carefully consider increasing protections for personal information.

The "year of the data breach"[65] – 2005 – will also hopefully be remembered as the year that serious legislative measures to protect privacy and personal data were considered.[66]

Individuals, corporations and governments make trade-off decisions on privacy on a frequent basis. While many of these trade-offs are for the public good, the definition of public good has widened in scope. Loss of a certain level of privacy in order to better protect against terrorism is generally considered in the "public good", but care must be taken not to take this to an extreme by incurring severe violations of privacy. Movement in this direction can already been seen in sections of the USA PATRIOT Act. Many

trade-off decisions that are made in order to gain access to services, such as phone service or credit cards, are in the immediate scope well intentioned but further erosion of privacy may no longer fit into the original benefits and are not in the public interest. An example is reducing credit card fraud by tracking spending patterns on an individual card basis. This in itself benefits all credit card users and would be considered a public good. However, if this information is further sold by the credit card company, it can then lead to further erosion of privacy and issues, such as identity theft, that are clearly not beneficial. Thus although the immediate trade-offs on privacy are generally for the public good, care must be taken to avoid further erosion.

While prophetic statements about the death of privacy are currently popular in the literature, we do not share those views. It is certainly true that there is a significant lag – years – between advances in technology and the passing of legislation, but eventually the deliberate, legal machinery appears to always address the more egregious breaches of privacy. Since we live in the present, surrounded by threatening technology, it is sometimes difficult to assess the progress made. In the past decade, laws such as the HIPAA Privacy Rule and the California Security Breach Notification Law have changed the privacy landscape irreversibly, for the better. In the coming decade, we believe that the U.S. will move closer to the European Union's model of explicit privacy laws, rather than an over-reliance on industry self-regulation. We also believe that precise legislation will provide increased protection against unwarranted surveillance by the counter-terrorism agencies. The reason for our optimism is the heightened public awareness of

privacy issues, as pointed out in Chapter 1. In due time, this awareness and concern invariably translate into action, in a democracy.

6 References

[1] Schoeman, F. D., ed. (1984). Philosophical Dimensions of Privacy: An Anthology. Cambridge: Cambridge University Press.

[2] Clarke, R. (1997). Introduction to Dataveillance and Information Privacy, and Definitions of Terms. Retrieved November 26, 2005, from http://www.anu.edu.au/people/Roger.Clarke/DV/Intro.html

[3] Turkington, R. C. (1990). Legacy of the Warren and Brandeis article: The emerging unencumbered Constitutional right to informational privacy. *Northern Illinois Law Review* 10(3), 479-520.

[4] Culnan, M. J. (2000). Protecting privacy online: Is self-regulation working? *Journal of Public Policy and Marketing*, 19(1), 20-26.

[5] Palen, L, & Dourish, P. (2003). Unpacking 'privacy' for a networked world. *Proceedings of the ACM Conference on Human Factors in Computer Systems (CHI)*, Ft. Lauderdale, FL, USA, 5-10 April 2003, 129-136. New York: ACM.

[6] Culnan, M. J. (1999). Information privacy concerns, procedural fairness and impersonal trust: An empirical investigation. *Organization Science* 10(1), 104-115.

[7] Smith, J. H., Milberg, S. J. & Burke, S. J. (1996). Information privacy: measuring individuals' concerns about organizational practices. *MIS Quarterly* 20(2), 167-196.

[8] Westin, A. (2003). Retrieved November 26, 2005, from http://www.harrisinteractive.com/advantages/pubs/DNC_AlanWestinConsumersPrivacyandSurveyResearch.pdf

[9] Agre, P. E. (2001). Introduction. In P. E. Agre & M. Rotenberg (Eds.), *Technology and Privacy: The New Landscape* (pp. 1-28), Cambridge, MA: The MIT Press.

[10] Warren S. D. & Brandeis, L. D. (1890). The right to privacy. *Harvard Law Review* IV(5).

[11] Langreich, M. (2001). Privacy by design – principles of privacy-aware ubiquitous systems. *Proceedings of Ubicomp 2001*, Atlanta, GA, USA, 30 September – 2 October 2001, 273-291.

[12] Hong, J. I., Ng, J. D., Lederer, S., & Landay, J. A. (2004). Privacy risk models for designing privacy-sensitive ubiquitous computing systems. *Proceedings of the ACM Conference on Designing Interactive Systems (DIS 2004)*, Cambridge, MA, USA, 1-4 August 2004, 91-100.

[13] Hong, J. I. & Landay, J. A. (2004). An architecture for privacy-sensitive ubiquitous computing. *Proceedings of the 2nd International Conference on Mobile Systems, Applications, and Services (MobiSys 2004)*, Boston, MA, USA, 6-9 June 2004, 177-189.

[14] Rössler, Beate. 2005. *The value of Privacy*.

[15] http://www.dexonline.com

[16] http://www.metrokc.gov/gis/mapportal

[17] http://realestate.yahoo.com/re/homevalues

[18] http://www5.metrokc.gov/reports/property_report.asp?PIN=9208900079

[19] http://papertoys.com/gates.htm

[20] CNET. 2005. Google balances privacy, reach. http://news.com.com/Google+balances+privacy,+reach/2100-1032_3-5787483.html

[21] Money. CNET: We've been blackballed by Google. http://money.cnn.com/2005/08/05/technology/google_cnet/

[22] http://thomas.loc.gov/cgi-bin/query/D?c107:5:./temp/~c1074rKqij::

[23] http://www.usatoday.com/travel/news/2005-08-08-electronic-passports_x.htm

[24] http://www.iht.com/articles/2004/10/04/edschneier_ed3_.php

[25] http://www.icao.int/mrtd/download/documents/TR-PKI%20mrtds%20ICC%20read-only%20access%20v1_1.pdf

[26] http://www.wired.com/news/privacy/0,1848,67333,00.html

[27] http://thomas.loc.gov/cgi-bin/bdquery/z?d109:h.r.00418:

[28] http://www.cagw.org/site/DocServer/Real_ID_FINAL_with_cover.pdf?docID=1281

[29] http://www.schneier.com/blog/archives/2005/05/real_id.html

[30] http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=107_cong_public_laws&docid=f:publ056.107.pdf

[31] Hartzel, Kathleen and Patrick Deegan. 2005. Balancing individual privacy rights and intelligence needs: procedural-based vs. distributive-based justice perspectives on the

PATRIOT act. In *Information ethics: privacy and intellectual property*, ed. Lee Freeman and Graham Peace. Hershey: Information Science.

[32] http://www.epic.org/privacy/terrorism/usapatriot/

[33] http://www.aclu.org/safefree/resources/17343res20031114.html

[34] http://www.lifeandliberty.gov/subs/u_myths.htm

[35] http://www.claria.com

[36] http://thomas.loc.gov/cgi-bin/bdquery/z?d108:HR02929:@@@L&summ2=m&

[37] https://adwords.google.com/select/ct_faq.html

[38] http://www.epic.org/reports/decadedisappoint.pdf

[39] http://www.amadorbooks.com/nocardsh.htm

[40] http://www.findarticles.com/p/articles/mi_qa3872/is_200103/ai_n8941466

[41] The information on the laws regulating video and audio monitoring is provided by the PalmVid security cameras manufacturer on their website www.palmvid.com

[42] www.palmvid.com

[43] This information is picked up from the website of Endoacustica audio monitoring devices manufacturer www.endoacustica.com

[44] This information is provided by the Privacy Rights Clearinghouse non-profit organization on their web site www.privacyrights.org

[45] www.voiceprintonline.com

[46] "Watching Me Watching You" PC Magazine (www.pcmag.com)

[47] www.eblaster.com

[48] "Basic Info", RFID Journal, www.rfidjournal.com

[49] "FAQ", RFID Journal

[50] Long, William J. and Pang Quek, Marc. Journal of European Policy: Personal Data Privacy Protection in an Age of Globalization: the US-EU Safe Harbor Compromise.

[51] P.1564.Swarchtz, Paul M. Free Speech vs. Information Privacy.

[52] National Conference of State Legislatures. http://www.ncsl.org.

[53] Protecting Personal Health Information in Research: Understanding the HIPAA Privacy Rule. http://privacy and research.nih.gov

[54] Electronic Frontiers, Australia. http://www.efa.org/au

[55] Malin, B. An Evaluation of the Current State of Genomic Data Privacy Protection Technology and a Roadmap for the Future. Journal of the American Medical Informatics Associations. 2005; 12 (1): 28-34.

[56] Sweeney, L. Uniqueness of Simple Demographics in the US Population. Technical Report LIDAP-WP4. Data Privacy Laboratory, Carnegie Mellon University, Pittsburgh, PA, 2000.

[57] Agrawl, R. and Srikant, R. Privacy preserving data mining. In Proceedings of the 19th ACM SIGMOD Conference on Management of Data, Dallas, Texas, USA, May 2000.

[58] J.H. Kleinberg, C.H. Papadimitriou and P. Raghavan. Auditing Boolean Attributes. PODS 2000: 86-91.

[59] I. Dinue, K. Nissim. Revealing Information while Preserving Privacy. PODS 2003: 202-210.

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

[61] Sweeney, L. Privacy-Preserving Surveillance Using Selective Revelation. Carnegie Mellon University, LIDAP Working Paper 15, February 2005.

[62] Levy, S. (2004, March 22). Geek war on terror. *Newsweek*. Retrieved December 5, 2005, from http://msnbc.msn.com/id/4486823/

[63] http://www.guardian.co.uk/comment/story/0,3604,922542,00.html

[64] http://www.epic.org/privacy/terrorism/usapatriot/senateletter_111705.pdf

[65] December 4, 2005 email from the ACM US Public Policy Committee

[66] http://www.acm.org/usacm/PDF/privacy_bills3.pdf