Statement of Latanya Sweeney, PhD
Associate Professor of Computer Science, Technology and Policy
Director, Data Privacy Laboratory
Carnegie Mellon University

before the Privacy and Integrity Advisory Committee
of the Department of Homeland Security ("DHS")

"Privacy Technologies for Homeland Security"

June 15, 2005

Respected Chairman and the Members of the Board, thank you for the opportunity to testify today on emerging technologies that are impacting privacy.

As the founder and director of the Data Privacy Lab at Carnegie Mellon University, I have had an unusual opportunity to work on real-world privacy technologies that relate to national security. I would like to share some of this work with you.

I will talk today about projects related to: (1) face de-identification; (2) bioterrorism surveillance; (3) identity theft; and, (4) webcam surveillance. Attached at the end of my testimony are one-page appendices that further describe each of these projects. More information on each, including related papers, can also be found at privacy.cs.cmu.edu.

Following the events of September 11, there is a common false belief that in order for America to be safe, the public must give up its privacy. This is not necessary. Our work suggests that ubiquitous technologies, such as data mining, location tracking, and sensor networks can be deployed while maintaining privacy. We show that society can enjoy both safety and privacy.

Similarly, many who develop ubiquitous technology believe in order for collected data to be useful it must be privacy-invasive. They also tend to believe that data that is not privacy invasive is useless. These are also false beliefs I want to address.

From my perspective, the question is not "what is the legal basis for deploying privacy invasive technology?", but rather, "how do we construct privacy enhancing technology that performs the same function as the privacy invasive technology, but does so with provable privacy protection?" This is the basis of the work I would like to share with you today.

One problem is that people don't understand what makes data unique or identifiable. For example, in 1997 I was able to show how medical information that had all explicit identifiers, such as name, address and Social Security number removed could be re-identified using publicly available population registers (e.g., a voter list). In this

particular example, I was able to show how the medical record of William Weld, the governor of Massachusetts of the time could be re-identified using only his date of birth, gender and ZIP. In fact, 87% of the population of the United States is uniquely identified by date of birth (e.g., month, day and year), gender, and their 5-digit ZIP codes. The point is that data that may look anonymous is not necessarily anonymous. Scientific assessment is needed. [For more information, see < privacy.cs.cmu.edu/dataprivacy/projects/explosion/index.html>.]

A goal of work in the Data Privacy Lab is to show how useful data can be shared freely while providing provable privacy protection. Here are some examples.

Face De-identification

After 9/11, law enforcement sought more access to video recordings, and many became concerned that with steady improvements in face recognition, open access to video data by law-enforcement without a search warrant can lead to massive tracking of all the people. The question became:

"Can we share video with law-enforcement such that no matter how good face recognition software might become, people cannot be re-identified without due process?"

In the context of sharing video surveillance data, a significant threat to privacy is face recognition software, which can automatically identify known people, such as from a database of drivers' license photos, and thereby track people regardless of suspicion. We introduced an algorithm to protect the privacy of individuals in video surveillance data by de-identifying faces such that many facial characteristics remain but the face cannot be reliably recognized. A trivial solution to de-identifying faces involves blacking out each face, in part or whole. This thwarts any possible face recognition, but because all facial details are obscured, the result is of limited use to law enforcement. Many ad hoc attempts, such as covering eyes, fail to thwart face recognition because of the robustness of face recognition methods. We provided a solution that creates new faces by averaging image components, which may be the original image pixels (k-Same-Pixel) or eigenvectors (k-Same-Eigen) so that each image is guaranteed to match at least k individuals ambiguously. The end result is data that can be shared with law enforcement from which vast populations cannot be tracked. If suspicious activity is found, a search warrant can be obtained to reveal the original faces. [For more information, see < privacy.cs.cmu.edu/dataprivacy/projects/video/index.html>.]

Privacy-Preserving Bio-terrorism Surveillance

The sooner officials can determine whether a biological agent, such as anthrax, has been released into the environment, the more lives that can be saved. The problem is that the symptoms of many of these biological agents closely resemble the flu, making early detection difficult. One way to combat this problem is to engage in on-going surveillance

of daily medical information to determine whether an unusual number of people are experiencing respiratory distress. This is termed early detection bio-terrorism surveillance using secondary sources (or simply "bio-terrorism surveillance" in this discussion).

In order to conduct bio-terrorism surveillance, medical information would have to be shared with public health agencies daily.

Without revision, public health laws are reporting laws –that is, they generally tend to require one of a set of known diagnoses be identified before identifiable medical information is required to be forwarded to a public health agency. But in this case, the flu is not a reportable disease and the bulk of most cases will not have an anthrax (or equivalent) diagnosis.

The solution we developed, termed "selective revelation," provides data with a sliding scale of identifiability, where the level of anonymity matches the scientifically derived need based on suspicious occurrences appearing within the data.

Bio-terrorism surveillance begins with data sufficiently de-identified in accordance to HIPAA. We were able to show that the data were provably anonymous. We were able to show that the data remained useful for CDC's anomaly detection algorithms. In general operation, bio-terrorism surveillance used provably anonymous data.

If an unusual number of cases are found in the sufficiently anonymous data, a "drill-down" occurs, providing increasing more identifiable data in accordance to public health law. Selective revelation is a technical instantiation of the probable cause predicate used by human judges to evaluate requests for search warrants. In this case, technology plays the role of the human judge in making data release decisions.

For example, if there are unusual number of cases appearing in the sufficiently anonymous data, but more evidence is needed as to whether the cases are geographically related, then more identifiable (but not explicitly identified) data is needed. If a geographical relation is found, then explicitly identified data can be shared under public health law. [For more information, see < privacy.cs.cmu.edu/dataprivacy/projects/bioterror/index.html >.]

[Note: The safe harbor provision of HIPAA would require the loss of significant ZIP information, thereby making the results anonymous, but not useful. For our purposes, we used the scientific standard provision of HIPAA, allowing the sharing of 5-digit ZIP codes. The Privacert Risk Assessment Server (www.privacert.com) was used to demonstrate that the data were sufficiently anonymous under the scientific standard of HIPAA. This technology was created by me, but has been subsequently transitioned into commercial use.]

Identify Theft

When a large number of citizens are at risk to identity theft, national security and economic prosperity are threatened. This work shows that thousands of Americans are at such risk, and introduces technology, named "Identity Angel," to help. Identity Angel's goal is to crawl through information available on the World Wide Web ("Web") and notify people for whom information, freely available on the Web, can be combined sufficiently to impersonate them in financial or credentialing transactions. This is an ambitious goal due to the various types of available information and the many inferences that relate disparate pieces of data. Therefore, the work discussed today focuses specifically on acquiring information sufficient to fraudulently acquire a new credit card using on-line resumes. An imposter needs to learn the {name, Social Security Number, address, date of birth} of a subject. Results show how resumes containing needed information can automatically be found and values harvested, and how many subjects removed such information from the Web once notified. [For more information, see < privacy.cs.cmu.edu/dataprivacy/projects/idangel/index.html >.]

Mining Images in Publicly Available Webcams

A dramatic increase or decrease in the number of people appearing at a location can be an indicator that something has happened that may be of interest to law-enforcement, public health, or national security. This work demonstrates how low quality camera images can be used to automatically alert when an unusual number of people are absent or present at a location. We report on experiments using publicly available, inexpensive cameras already operational over the Web. A "historical database" (H) for each camera is constructed by capturing images at regular time intervals and applying a face detection algorithm to store the number of faces appearing in each image ("face count"). Later, given an image X having timestamp t, if the face count of X is significantly higher or lower than the expectation inferred from H for time t, an unusual number of people are considered to be present in the image.

With almost 6,000 publicly-available web cams already showing public spaces around the USA, this approach offers a low-cost national surveillance system. But it offers it to anyone in the world. Not only is that a concern of personal privacy, but perhaps also of national security. On the other hand, there may be many worthy uses for this network of webcams. For this reason, we introduce the notion of "smart cameras" in which high-level information from webcams, and not actual images, is shared. We envision that smart cameras will include programs within, such as this work, and would provide reports, updates and alerts, and not images. [For more information, see < privacy.cs.cmu.edu/dataprivacy/projects/videocount/index.html >.]

In summary, the first two examples demonstrate how data can be rendered anonymous and shared freely for surveillance purposes.

The third example, Identify Angel, shows how not helping the public protect privacy leads to national security vulnerabilities.

The final example shows how surveillance can already be conducted using publicly available cameras, and using smart cameras can increase effectiveness and reduce privacy concerns.

This is a sample of the work done in the Data Privacy Lab and an idea of the kind of work that can be done. Of course, one problem is the lack of funding in this area to pursue this kind of work. Almost all of the work described was donated by members of the Lab. No government funds were received. No opportunity for government funding has been forthcoming. Imagine what can be done with government funding.

For example, the final appendix discusses the "Watchlist problem," for which there is currently no acceptable technical solution. I would welcome the opportunity to be commissioned (receive funds) to solve this problem. [For more information, see < privacy.cs.cmu.edu/dataprivacy/projects/watchlist/index.html >.]

Recommendation #1

Funds should be made available to pursue the development of privacy enhanced technologies for DHS purposes. No traditional government funds are available for this kind of work.

Recommendation #2

DHS should nurture the existence of a new field for creating privacy enhanced technologies, establishing their scientific basis, and demonstrating their real-world applicability. This work lies outside traditional areas like computer security and cryptography, and even outside newer emerging areas like privacy-preserving data mining. This work is situated at the confluence of law-enforcement, intelligence, computer science, and law.

Thank you.

Latanya Sweeney, PhD
Associate Professor of Computer Science, Technology and Policy
Carnegie Mellon University
Voice: 412-268-4484
Email: latanya@privacy.cs.cmu.edu
Web: privacy.cs.cmu.edu/people/sweeney/index.html