Introduction

The primary concern of the EFF is the use of data aggregation by the government to assemble personal profiles of American citizens and identify potential terrorists. In projects like Total Information Awareness, the government has proposed to use combinations of medical, financial, educational, employment, travel, and telephone records to search for patterns of transactions that are suggestive of terrorist activity. While the EFF understands the need for national security, TIA and similar projects are unlikely to be effective in practice and will raise too many security and privacy concerns in the process.

Although the most dangerous use of data aggregation is that of the government, data aggregation by private companies is also cause for concern. This is because companies can sell to the government information that it could not legally obtain on its own. Even companies that refuse to do business with the government can be subpoenaed. For these reasons, data aggregation by private companies poses the same risks that data aggregation by the government poses.

The risks that have been acknowledged by the EFF fall into the areas of security, privacy, economics, and personal violations. Although all of these risks are worth considering, the extent to which they can be justified and defended varies. Certain risks that have been used by the EFF to argue against data aggregation may be of little consequence or unrealistic. Others provide good fodder for continuing to defend freedom in the digital world.

Security Risks

Stance

Data security technology is not (and may possibly never be) mature enough to be safe enough to hold large amounts of PII. In addition, technology is not the only player in information security. Any person with database access is a potential attacker, along with the database administrators who configure the permissions. No amount of technology can change the basic human nature of individuals who may be motivated by greed, blackmail, or other personal intentions to take advantage of their access to sensitive information.
Storing any PII is dangerous, and should be avoided. Any database access should be restricted to only the essential users.

**Justifications**

A massive case of identity theft involved a computer help-desk employee who abused his access to sensitive passwords from banks and credit companies. By collecting the personal information on over 30,000 people over a period of three years, the employee had a bounty of information that he sold to scam artists [3].

Data in aggregation, without personally identifiable information, is normally sufficient to track usage patterns for administering online services. Any online service should be able to define a short period of time after which individual logs will no longer be needed for troubleshooting. [1]

Encryption is not a solution for securing the storage of PII. Any subpoena order can still force online service providers to turn over the encryption keys along with the encrypted data. [1]

Large databases are targets for malicious users, criminals, and terrorists. It is unlikely that even the most sophisticated database security technology with infinite funding will be immune to attack, especially as the data inside the databases are aggregated from more sources (and therefore of more value) [3].

**Defensibility**

Various governmental bodies may retort that they are not trying to aggregate the data from smaller public and privately held databases, and therefore are not inducing more security risks. However, creating any interface for accessing the data in these databases exposes each one to a new pool of possible risk, let alone the new risk of being able to correlate the data from each smaller pool. [1]

Corporations may retort that it is necessary for them to retain PII about their customers in order to provide a high quality of service, or to protect their business. The EFF believes that PII is not necessary in order to provide a high quality of service, and that online service providers should obfuscate their individual transactions (by removing detail, such as the exact page from a website http request) and retain only aggregate usage data.

**Economic Risks**

**Stance**
Electronic commerce in the U.S. will be threatened by the privacy issues that arise from data aggregation and that are not well handled by security technology or legislature. This threat may cause consumers to back away from using e-commerce.

There will be a threat on the U.S. economy as a whole by excluding U.S. companies from export opportunities and preventing trading or cooperating with U.S. e-commerce systems because of data privacy conflicts.

The cost of identity theft to the companies, government and victims may also increase as a result of data aggregation. Moreover, there is also the cost of “poisoning” other databases such as the TIA database with stolen identities.

**Justifications**

One justification is based on envisioning a European Union subsidiary of a U.S. based e-commerce company. And due to the EU’s Data Privacy Directive (because of conflicts with the TIA’s surveillance system) the subsidiary might be forbidden from running the company's systems in the EU.[3]

Another justification is based on the fact provided by the national bank regulators which estimates the number of identity theft cases that occur yearly to be half a million. And the cost due to this crime estimated to be billions of dollars. [3]

**Defensibility**

It appears that the threat of the decrease in internet sales or e-commerce is not backed up by hard evidence supporting the claim. According to U.S. Census Bureau News, it states that ‘The first quarter 2006 e-commerce estimate increased 25.6 percent (±5.9%) from the first quarter of 2005’[4]. Also, research firm comScore Networks calculated that consumers spent up to 102.1 billion dollars through e-commerce in 2006 that represented an increase of 24 percent over 2005.[5] According to these numbers the expectations are that they will keep growing and fewer consumers will drawback from e-commerce. But it’s important to point out that consumers are beginning to be aware of data breaches on the web and in an article by Matthew Mogul and Michael Doan on how data breaches can affect e-commerce, they state that ‘buyers will become choosier about where they e-shop and about what kinds of personal information they're willing to give out.’[6]

The claim that U.S companies would be excluded from foreign business is also not based on hard evidence. Frankly there have not yet been any statistics or numbers coming out that relate to this. And as mentioned previously, e-commerce is expected to grow in the coming years and so inevitably foreign companies will try to find opportunities.

However, it is true that the costs that arise from identity theft as a result of data aggregation are very high. In an article by the government technology forum about privacy and security breaches, it mentions that over 100 million data records of U.S. residents have been exposed to security breaches since February 2005. And recent
breaches include ChoicePoint, The Bank of America, LexisNexis, Time Warner, CardSystems, Boeing, Ford, FedEx, MasterCard, AT&T, General Motors, General Electric, Starbucks, Ernst & Young, Deloitte and Touche, YMCA, Internal Revenue Service, and the Department of Defense[7]. Moreover, the AOL history breach that occurred in August 2006 resulted in disclosing 650,000 users’ search records that can be easily linked to their identities through unique identification numbers. [8]

Privacy Risks

Stance

Any form of surveillance, by definition, introduces an invasion of privacy. Therefore, by gathering information on Americans, mining this data to generate new information, and combing the data for patterns, the government is already denying citizens their right to privacy.

Worse still, there is no accountability for what appears in these vast databases. Combining databases is algorithmically complex and not unlikely to result in error. However, individuals whose information appears in the database have no way to view, dispute, or correct their database entries. This is especially true in the case of government databases, where security concerns necessitate that citizens neither know nor have the power to change the information held about them. If a citizen is tagged as a potential terrorist, he may never know, and he therefore will not have the opportunity to clear his name.

Justifications

The right of Americans to access and change information that is held about them is guaranteed by the Code of Fair Information Practices. According to this set of guidelines, there must be some accountability for information that is gathered about individuals. Therefore, projects like TIA would directly violate the Code. Again, there is no way to avoid this violation because of the level of security required for the project.

Defensibility

Although errors in data are a current problem, proponents of data aggregation may argue that technology can be improved to eliminate this danger. This is an accurate statement, and it should be kept it mind when evaluating the privacy risk. At the same time, however, the government should be reminded that we are not at that point today, and so it cannot yet safely make judgments based on aggregated data. Terrorist detection algorithms have already been used on real data about real people in Guantanamo Bay, Cuba! [9] This is a violation of privacy because the subjects of the research were not given the opportunity to view, verify, or correct the data held about them. While it is true that data aggregation algorithms may some day be advanced enough to avoid errors, it is not safe to begin using them until we have reached that point. Government officials without a technical background are unable to evaluate an algorithm’s degree of accuracy
and may be quick to exploit “science” to justify wrongful detainment and surveillance of innocent Americans. This is exemplified by their use of terrorist detection algorithms in Guantanamo Bay, where reasons for detainment were fuzzy to begin with.

Perhaps the most common argument against the privacy risks of data aggregation is that all of the data being collected is already out there, and could be compiled by any interested individual. This is simply not the case. Much of the data being sold to companies like Choicepoint and Cogito Inc. consists of personal records known only to individuals and, for instance, their banks, or their insurance companies. This data is then being sold to other organizations who did not previously have access to the data without the knowledge or consent of the individuals it concerns. Most disturbingly, the government is now buying access to data it could not legally obtain on its own. [10] Regardless of any claims it might make about the data already existing prior to aggregation, the government is indeed invading the privacy of Americans when it purchases access to their medical, financial, and other records.

Personal Risks

**Stance**

Any aggregated data that is either owned by or otherwise accessible to the government represents extreme personal risk for average Americans. In any screening or profiling process of this magnitude, false positives are inevitable. In a national security context, each false positive means an ordinary citizen will be flagged for additional scrutiny or surveillance. This is unconstitutional. Data aggregation has great potential to harm Americans, and an unclear (though probably negligible) potential to catch terrorists.

**Justifications**

Even an unrealistically low rate of false positives (0.1%) could result in as many as 3 million American citizens being mistakenly flagged for additional surveillance. [3] Public knowledge of such false positives will create a chilling effect on speech in the country. People will feel compelled to avoid doing anything that deviates from average behavior so as to avoid being noticed by the data aggregation system or called "Un-American" by their peers.

Matching criteria would presumably be kept secret by the government, which would add to the uncertainty fueling above chilling effect.

The Fourth Amendment enumerates the right against unreasonable search and seizure, which has been interpreted to encompass digital searches as well as physical searches (Kyllo v. United States, 2001). Any surveillance or even increased scrutiny of individuals that resulted from a false positive match in aggregated data could be considered an infringement of Fourth Amendment rights.
Defensibility

Government groups have disputed the idea that false positives will cause such far-reaching problems. They point out that persons flagged as suspicious by their data mining tools will not be immediately arrested, but simply queued for human review. They contend that this human review is harmless. However, the EFF can argue that it constitutes direct scrutiny of a personal dossier. Additionally, the EFF can argue that without oversight in place, the government could place surveillance upon all suspect persons churned out by their data mining system.

Conclusion

The aggregation of data by both the government and private corporations poses grave risks to American citizens on a range of issues. Most directly, it is a clear and unavoidable invasion of Americans’ privacy. It also introduces the possibilities of security risks, by putting sensitive information in danger of theft, as well as personal violations, by giving the government justification for future unconstitutional actions. Economic costs are at this time less evident, but may be a growing trend as consumers become more aware of the risks of data aggregation. For these reasons the EFF continues to oppose data aggregation.

Sources

[1] Review of May 20 Report on T1
http://www.eff.org/Privacy/TIA/20030523_tia_report_review.php
http://www.eff.org/osp/
[3] Letter from ACM Senate Committee on Armed Service Leaders
http://www.eff.org/Privacy/TIA/acm-letter.php
http://www.census.gov/mrts/www/data/html/06Q1.html
[5] Digital lifestyles
http://www.kiplinger.com/businessresource/forecast/archive/data_breaches__how_big_a_threat_to_e_commerce___html
http://ebusinessforum.com/index.asp?layout=rich_story&channelId=5&categoryId=18&title=The+lowdown+on+privacy+and+security+threats&doc_id=10339
http://www.washingtonpost.com/wp-dyn/content/article/2006/06/14/AR200606140206_3_2.html