Policy of the Department of Homeland Security on Data Aggregation

Current Policy

The attacks on September 11, 2001 turned our country's concept of security upside-down. On this day, our country realized that we are not completely safe nor are we exempt from terrorist attention. However, the attacks could have been prevented – and hundreds of lives saved – had our nation's intelligence and police agencies been working together more efficiently. Shortly after 9/11, the President formed the Department of Homeland Security (“the Department”) in order to streamline the intelligence collection and dissemination process.

The government has always collected information on suspected individuals and has detected terrorist plots in time to prevent attacks on numerous occasions. However, in this time of heightened threat, its previous methods will not suffice. As illustrated on 9/11 when various agencies each held different pieces of intelligence on the hijackers, individually gathered information is nearly useless without consolidation. It is therefore imperative that national intelligence agencies (ie. CIA, FBI, NSA) and state and regional law enforcement pool their intelligence; information must be shared in order to prevent future terrorist attacks.

Five years ago, this concept of information sharing was one of the foundations of the Department’s Strategy for Homeland Security. Its aim was to build a “system of systems” that would bring “complete and common awareness of threats” to all levels of government, private industry, and citizens, while still protecting sensitive and classified information.1 (p. 56) To this end, it laid out five principles to guide the development of information systems:

1. It will balance our homeland security requirements with citizens' privacy.
2. The homeland security community will view the federal, state, and local governments as one entity – not from the point of view of any agency or level of government.
3. Information will be captured once at the source and used many times to support multiple requirements.
4. It will create databases of record, which will be trusted sources of information.
5. The homeland security information architecture will be a dynamic tool, recognizing that the use of information technology to combat terrorism will continually evolve to stay ahead of the ability of terrorists to exploit its systems.

Justification for the Current Policy

1. The current policy provides for the efficient use of collected information

The Department for Homeland Security has directed much of its focus on the aggregation of data from disparate government agencies. This aggregation has not been toward a centralized database of citizen records, but toward the facilitation of easy data sharing.

The initial diverse information systems of these agencies made the efficient exchange of information difficult, if not nearly impossible. This has severe implications for homeland security efforts. For instance, a suspect on a US State Department “most wanted” list may be watched by the CIA – but the US Border Patrol may not be aware of him. The state police mostly likely would not be. The focus of the Department is to increase the awareness and knowledge of these agencies, across levels and geographic regions.

One way it has pursued information access is through the Homeland Security Information Network. The HSIN provides not only secure communication channels, but also allows agencies, bundled into portals by area (law enforcement, State, etc.), to upload documents and information. Each agency can communicate within its own portal and with the counter-terrorism portal, which is connected to each. This is one example of the central data access model of data aggregation.

The Department also argues that it may need to partner with business groups to develop a deeper profile of potential terror suspects. The combination of governmental information and records with private information can shed light on the activities and associates of terror suspects without direct surveillance. Naturally, the access of commercial data requires a good deal of oversight, and the Department believes that it has created strong frameworks for assessing appropriate access. These frameworks will be discussed below.

2. The current policy includes safeguards sufficient for protecting individual privacy

Although the Department’s ability to collect and aggregate information is extremely important for the aforementioned reasons, it is also important that the Department not invade citizens’ privacy unnecessarily or allow false positives to result in the persecution of innocent individuals. To this end, the Department has several frameworks that should dictate how and when data is collected and aggregated.

First, the Department requires that Privacy Impact Assessments (PIA) and Memoranda of Understanding be created whenever a new data mining project is undertaken. These are fairly traditional security measures, but because PIAs are made available to Congress and to the general public, they allow for oversight of the Department’s activities. That said, the PIAs and Memoranda, created within the Department, may not be unbiased and further security and privacy protections are necessary.

In a recent report by the Department to Congress\(^2\), the Department recognized the need to ask and answer the following questions prior to starting any new data mining project:

What is the purpose of the project? Why is data being collected?
Does the agency collecting the data have the authority to do so?
The Department suggests that their own Privacy and Integrity Board be charged with asking and answering these questions. While the Department is correct in recognizing the need for oversight of this kind, the suggestion that this review be carried out by an internal board is not the best way to maintain reliable or transparent review. While the Board consists of a number of individuals appointed from outside the Department, transparency might be improved if the Board itself remained beyond Departmental influence. These recommendations also make no mention of the types of answers that should cause the intended project to be disbanded. In fact, there is no suggestion that any answers to these questions would cause such steps to be taken. The Department simply recommends that these questions be asked. Asking, in and of itself, is not a sufficient privacy protection. In order to protect the privacy of citizens who are the subjects of data mining activities, the Department and Congress should together create guidelines that state what purposes make data collection acceptable as well as what kind of authority is required for an entity to pursue such a project. If these guidelines are not met, the project should be modified or disbanded. And oversight should remain independent of the Department.

In addition to these questions asked prior to beginning a project, the Department recognizes that the following principles should be adhered to when collecting data:

- Recognizing that the existence of patterns is not proof of cause or effect, no one should make automatic decisions on the basis of the results of a data mining project.
- When possible, projects should use anonymized data.
- Quality standards should be ensured.
- Models should be validated and standards documented.
- Data that identifies individuals for further investigation should be subject to review and redress and a person independent of the data-gathering process should confirm individuals’ identities.
- Audits should be conducted randomly to maintain transparency.

These recommendations, made internally by the Department’s own privacy office, are significant and demonstrate an understanding of the risks associated with data mining projects. However, there is often a difference between publicly stated policy and actual behavior. In order to ensure that these principles are in fact taken into account, the same independent oversight committee mentioned above should be charged with conducting the audits.

**Assessment of the Defensibility of the Current Policy**

1. **Handling uncertainties in the quality of information collected**

The Department recognizes that there is always the possibility of error when collecting data. Errors can take the form of incorrect information in individual cells, sometimes easily recognizable (i.e. Age = 200) and other times, not. However, even when the information collected is correct, there is the danger of creating false information through the poor design of algorithms that, in turn, elicit false positives and as a result, the Department might waste time and resources in addition to invading the privacy of an innocent citizen. In addition to incorrect data, there might be missing data, which can also threaten the ability of a model to be used
effectively. Data labels can be ambiguous is words (i.e. address) can have more than one meaning. In order to guard against these risks, the Department recommends that the following techniques be used with caution:

- Association or link analysis
- Sequence or path analysis
- Classification
- Clustering
- Forecasting

Ensuring that the entities collecting the data have authority to do so, that data remains linked to its original sources, that data retention periods are limited (and/or that data is updated), that uses of the data are consistent with originally stated aims, and that profiling not be based on improper fields will all help to protect against errors. These actions will also help the Department to comply with the Privacy Act of 1974. However, these actions will only help if they reflect actual behavior and if they are done forthrightly and with transparency. For this reason, the independent oversight board previously referenced is extremely important. The Department seems to have a very strong understanding of the risks associated with data collection and the safeguards that should be in place. Now, it is a matter or ensuring that these safeguards are actually followed.

2. Retaining responsibility of information gathered

Although the Department is specific about the data mining projects and initiatives it is undertaking and how this information is being analyzed, less explicit information has been provided on how this information will be distributed and how it will be safeguarded in this respect. This is understandable; the Department needs to be able to maintain complete secrecy on the specifics of its operations to safeguard its effectiveness against terrorist operations. However, this also exposes a weakness in the extent to which the Department can be held accountable for its actions, and the defensibility of the current policy. The Department should be very careful about how its data mining projects are conducted, and place enough checks and balances necessary to ensure integrity in how this information is handled. Specific steps are outlined below.

Recommendations

As with any government initiative, the execution of even the most well-conceived plans may result in issues. We applaud the Department’s effort in conceptualizing their data aggregation motives, thinking through the privacy challenges, and taking the initiative to minimize privacy invasions. As these changes become systemized, though, we strongly recommend that the Department maintain its vigilance in these areas:

Thoroughness
The Department must make sure that it keeps to high standards of privacy protection. For instance, PIA filings for commercial data access should be thorough and accurate. At scheduled times, the types of information held to account should be reviewed and updated if necessary.

Transparency
The Department must engage in conversation with citizens and advocacy groups about their actions. While we understand that, for security reasons, it cannot disclose the specifics of its searches, it should at least publicize the rationale behind its searches. We find that the frameworks, while public, are not widely known, which may cause unnecessary fear or confusion.

**Security**
How will the Department secure information that it does actively collect? And what will it do with it two, five, twenty years later? From what we have seen, these policies are less developed.

**Human intelligence**
While data aggregation and computer algorithms are invaluable aids, they are rendered worthless if the people using them are not prepared. Two years and $300 million later, the Homeland Security Information Network was found to be not fully effective. The reason, according to the Department’s Office of the Inspector General, was that people were not trained in the technology. Rather than using its communication and documentation sharing capabilities, people were reported to still be using ad hoc methods to gather information, individually calling people in other departments.