Adam Shostack

14 July 2007

Mr David Wood
General Accounting Office
(by email)

Dear Mr. Wood,

I am writing to you today to comment on your recent report, "Personal Information Data Breaches are Frequent, but Evidence of Resulting Identity Theft Is Limited..."

I am an information security professional with 15 years of experience. I am currently working for a large software vendor's security development policy and strategy group, but the comments here are my own professional opinion, and do not reflect those of my employer.

Since the Choicepoint incident, I have been publicly analyzing and commenting on data breaches. I have spoken at several conferences on the subject, and my blog, Emergent Chaos, is referred to by your main data sources as an important source of analysis.

For reasons that will become clear, I found GAO's report and its implied recommendations to be disappointing, and not representative of the usual high quality of GAO reports. This is, as you note, a difficult and challenging field in which to do research. As such, I am hesitant to criticize, and do so because of the esteem in which GAO reports are generally held. For ease of writing, I shall refer to GAO as "you."

My concerns can be summarized as your analysis of the data fails to pursue important and possibly revelatory data to which the public does not yet have access, your selection of data sources lacks justification, you failed to consider (or discuss) alternate methodologies which may have resulted in different results, you make unjustified assumptions that companies can provide data, and you fail to identify systemic sources of bias in comments on which you rely. I will explain each of these concerns in order.

Failure to pursue important questions

You were charged with "identifying what is known about the incidence and circumstances of breaches of sensitive personal information" (page 3). In a set of paragraphs from page 12 through 17, you list incidence and circumstance, and fail to analyze commonalities between your data sources. You even fail to bring them together to draw attention to how disparate they are:

Source	Dates	# of Incidents
FBI	Unclear	1,300 under investigation
Secret Service	2006	327
House Government Reform Committee	Jan 1, 2003-July 10, 2006	788
US CERT	FY 2006	477
5 banking regulators	"past few years"	"several hundred"
FDIC	May 2005-Dec 2006	194 at regulated, 14 third party
Office of thrift supervision	April 2005-Dec 2006	56 at regulated, 72 third party
New York State	Dec 7, 2005-Oct 5, 2006	225
North Carolina state	Dec 2005-Dec 2006	91 affecting > 1000 people
Educase survey	2005	127.4 (26% of 490)
American Hospital Assoc. survey of 46 institutions	2006	13 hospitals reported 17 breaches
Attrition	Not listed in GAO	500+
Privacy Rights Clearinghouse	Not listed in GAO	300+
All except attrition, PRCH	Jan 2003-Dec 2006	3688.4* (Not rigorous)

This chart, in conjunction with your chart on page 26, indicates a several facts, which I believe are critical to the answer to the question of "identifying what is known about the incidence and circumstances of breaches of sensitive personal information." In particular:

1. There is no authoritative central source of data. The best available data is kept by private research and advocacy organizations. There is no central clearinghouse to which data must be reported.
2. The data varies widely between sources. As Chris Walsh pointed out in his paper on "Data on Data Breaches," what you find depends strongly on where you look.
3. Collection of this data is a substantial burden and effort, and distracts from the analysis phase of research. Much of the data is not available to the public, inhibiting analysis.
4. GAO has an opportunity to analyze commonalities in the data and show us a standardized and normalized representation of the data. In my summation, I've added the numbers reported, excluding the "several hundred" reported by the five regulators, to get 3688. It seems likely that there is overlap between the reported breaches. We can also note that both the House Government Reform committee (HGRC) list and the FBI list are each larger than the publicly known information. However, we don't know if the HGRC number is 311 larger than the CERT number because the HGRC period is longer, or if there are FY 2006 breaches reported to one but not the other.

To state critique #4 differently, we don't know if the breaches covered in the reports are heavily overlapping or not. Do they more closely resemble possibility 1 or possibility 2?

Possibility 1:
There is little similarity
between breaches

Possibility 2:
There is a lot of similarity
between breaches

We don't know. Many people instinctively believe in #2. What we do know is the one time the experiment has been done (New York vs. the University of Washington dataset, derived from Attrition) the data looked a lot more like possibility 1. To effectively answer Congress's question, we need the answer, and GAO has not provided it.

I believe that a fair answer to the question would have pointed out these issues.

Unjustified data selection.

Starting from the highlights, you state that you examine the 24 largest breaches reported in the media from January 2000 through June 2005. You do not justify this selection. We have reason to believe that the largest breaches are not all reported in the media. (Analysis by Chris Walsh showed that 3 of the 5 largest breaches reported to the State of New York were not in the attrition or Privacy Rights databases on which you relied.)

You do not justify your selection of the largest breaches. We have no reason to believe that the largest incidents have the same likelihood of identity theft, and there are reasons to believe that they will show a lower incidence. In particular, several of the largest incidents involve loss of backup tapes, which are likely in Iron Mountain and UPS warehouses. Some of the others may have been "trophy hunting" by hackers, where, rather than taking the data for profit, they were attacking for reasons of prestige.

A more reasonable methodology might have been to randomly select incidents from the data sets, or to investigate the largest and a random sample, in order to identify if biases (perhaps accounted for by the hypothesis above) were present.

You do not justify the size of your sample set. As you identified, there were at least 572 publicly reported incidents in your time sample (page 11). You examined 4.2% of these, and have no comment on how your sample size was selected.

Alternate methodologies possible

You fail to justify your decision to start from data breaches. An alternative investigative methodology would have been to select a set of victims reporting ID theft to the FTC, FBI, or other source of criminal data, and trace those reports back to their source as best as could be done. This has a challenge in that (as you note) many of the victims of identity fraud do not know how they were victimized. GAO could have presented a list of known breaches to these individuals, and looked for correlations, or considered only the known cases.

Unjustified assumption that companies can supply data

There is an assumption that breached organizations are notified of identity theft by their customers. This assumption shows strongly on page 5, where you write, "available data and interviews with researchers, law enforcement officials and industry representatives indicated that most breaches have not resulted in detected incidents of identity theft." However, there are several assumptions here. First is that a company who has suffered a breach would be told by a consumer that that consumer has suffered identity theft. Consumers have little motivation to do so, and so looking to companies as a source of data is, at best, a partial answer.

Failure to identify commenter biases

Even if a company's call center representative was told that, the call center computers likely have no way to record that information. Modern call centers are expensive to run, and are often run from `scripts' and 'trees.' If these trees have not been updated, even a company that had been notified of issues might not have captured and analyzed that information. Even if a company has captured and analyzed that information, it is likely being treated as highly sensitive in conversations with attorneys in order to contain liability. It is unlikely to be shared with industry association representatives, at conferences, etc. The information is likely to be kept close to the chest. Finally, even if the representatives with whom you spoke were aware of fraud, they might be biased against sharing that with you. They are likely aware that Congress is considering further regulations, and may be eager to sweep evidence of the breadth of the problem under the rug, to avoid further regulation.

As a final note before I conclude, you imply that notifications are expensive and complex, and seem to endorse a 'reasonable likelihood of harm' standard (although you do not come out and say so). Before you endorse such a standard, I would urge you to pay close attention to the difficulty that that would cause banks (as you cover on page 35). Absent more and better data on the relationship between breaches and fraud, it will be hard to figure the odds of fraud. The best way to get information on the relationship is to expand the datasets available to all researchers to allow and encourage research. A 'reasonable likelihood of harm' standard will prevent us from crawling out of the mess that we're in today.

In conclusion, your failure to pursue important questions about the nature of the data, your failure to justify your data selection or sample sizes, your failure to explain your choice of methodologies in the presence of alternatives, and your assumptions that companies have the data you wanted, and would, unbiased, provide it, cause this report to be deeply flawed, and create a worrisome possibility that anyone relying on it would come to erroneous conclusions.

I would urge you to update your research to take these concems into account. In the future, I would be happy to work with you on this subject, which I believe to be of considerable import.

Thank you for your attention to these matters,

Adam Shostack