Digital Forensics Association

The Leaking Vault

Five Years of Data Breaches

A study conducted by Suzanne Widup
Published by the Digital Forensics Association
Publication Date: July 2010

The Leaking Vault

Five Years of Data Breaches

Presented by the
Digital Forensics Association

Author: Suzanne Widup

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Findings

2,807 incidents

Data breaches are a concern for every organization. Until now, studies have been based on data that must either be kept confidential or have a small number of data points. The Leaking Vault study presents data on 2,807 data breach incidents—and is the largest study of its kind to date. Information was gleaned from the organizations that track these breaches, government sites, and media reports. These incidents represent data from 28 countries, including the United States.

721.9 million records

This study covers breaches from 2005 through 2009, and includes over 721.9 million known records disclosed. On average, these organizations lost 395,362 people's data per day, every day from January 1, 2005 through December 31, 2009. With just over 300 million U.S. citizens, (and with U.S. record losses at 630.5 million), this means that each U.S. citizen's data could have been exposed more than twice on average [24].

Attack Vectors

The Laptop vector was the leader for loss incidents, with 49% of all breaches. This vector was only responsible for 6% of the record loss, however. Laptops were stolen 95% of the time. In 33% of cases, they were stolen from office; 28% from vehicles. The loss leader was the Hack vector with 327 million records, or 45% of all records disclosed. Hacking accounted for only 16% of the incidents, but had an average loss of 716,925 records per incident.

Insiders

When an incident involved Insiders, it was more than twice as likely to have been an accident. Insiders were responsible for 205.9 million records disclosed. Incidents by Insiders were fewer overall than those involving Outsiders. Insiders were responsible for only 29% of the incidents, while Outsiders were responsible for 48%. Outsiders also led the record losses, at 357.6 million disclosed.

Third Parties

When Third Party facilitated breaches occur (16% of the cases), the median number of records disclosed per incident is almost twice the Outsider figure—and more than 10 times that of an Insider. As with the rest of the study, the leading vector for Third Parties is Laptop. Their lead vector for records disclosed was Hacking. Third party partners facilitated the disclosure of over 111 million records.

Data Types

Social Security Numbers (SSNs) are the most frequent data element reported. The Business sector led in the number of incidents and records disclosed, and was also the leader in disclosing SSNs. Actually, the Business sector was the loss leader in Credit Card Numbers as well. Between these two types of data, they account for 70% of breach incidents.

Customers

The relationship between the disclosing organization and the data subject was explored, and Customer's data was the most frequently exposed. The Business sector lost Customer data 71% of the time. They lost their SSN data in 40% of the cases. Only 27% of the cases lost Customer credit card data. Yet, in only 60% of cases was credit card monitoring offered to these victims.

$139 billion

Using figures from the recent Cost of a Data Breach study, a figure of $139 billion was calculated as the estimated cost over the five years of the study [20]. This includes only the cost suffered by the disclosing organizations, not the downstream/upstream costs nor the costs to the data subjects in time spent trying to repair their records. Data subjects are estimated to spend an average of 68 hours fixing the damage if an existing account is compromised. If new accounts are opened, the average jumps to 141 hours [8].

Recommendations

Data Lifecycle

Organizations should ensure that their data lifecycle is managed end-to-end whether the data is on paper or in electronic form. If proper lifecycle management of sensitive paper documents is not part of the organization's culture, the most stringent of controls on the data in electronic form won't stop a breach if someone is careless with their printed material.

Portable Data

Organizations that rely on the login password to keep the data safe on a laptop that has been lost or stolen are operating under an inaccurate risk assumption. Additional measures should be taken to protect sensitive data from the loss of physical control of the electronic device. This includes data stored on portable storage devices, as well as laptops, portable devices and smart phones. Encryption is an example of a control that will accomplish this goal.

Third Parties

Security requirements for third-party partners must be included in contracts from the beginning. They should include the processes required to secure the data, as well as lifecycle management. They should also include provisions for data destruction or return should the partnership end.

Internet Facing

Internet-facing systems should be scanned regularly for both the presence of sensitive data that should not be stored there, and for code vulnerabilities that put data at risk. Code review for internal application development groups should be instituted to ensure common flaws such as SQL injection and buffer overflows have not been introduced. Input validation on all forms should be tested for unexpected behavior.

INTRODUCTION

Prior to the passage of the current data breach disclosure laws, companies were under no obligation to notify customers when their data was exposed. Victims of these data thefts were left unable to determine which organization—entrusted with their data—had failed in their duty to safeguard its confidentiality, and thus have no recourse for damages. What has not changed since 2005 is that the victims are left to deal with the financial and emotional consequences of a situation beyond their control to prevent [17]. What has changed with time (and legislation) is that organizations are increasingly required to publicly report incidents of disclosure of the data they have a duty to safeguard.

> A data breach is generally considered an "unauthorized acquisition of computerized or other electronic data, or any equipment or device storing such data, that compromises the security, confidentiality, or integrity of personal information [22]."

Currently, there are laws requiring notification in 46 states, plus the District of Columbia, Puerto Rico and the Virgin Islands [23]. However, these laws have disparate and sometimes conflicting requirements for what must be reported and to whom. There is no central reporting agency, nor single source tracking all breaches. In many cases, the victims are quietly notified by mail, and the media never learns of the event. The stories that are covered tend to be the most shocking, with the largest number of records exposed, giving a skewed picture of the problem [10]. It should be noted, however, that these incidents primarily came to light because of a legal requirement for the organizations to report them (typically either to the victims or a government agency). Before these data breach laws, public coverage of data loss was rare. Since not all data disclosure incidents include data covered by a specific law, this data set represents a subset of data breach incidents.

By examining the incidents, breach vectors and record loss figures, this study provides an in-depth analysis from empirical data, with the hype removed from the equation.

METHODOLOGY

While there is no single data source for all data breach incidents, there are a number of organizations attempting to track them. These incidents are drawn from media reports, sourced from state government websites via Freedom of Information Act (FOIA) requests, or reported by the victims after receiving disclosure notifications from the organization experiencing the data loss event. While many of these incidents have been disclosed in the media, the focus of this study is the collective data set and the insights that can be drawn from focusing on data breaches as a whole.

For this study, incident reports from the Open Security Foundation (OSF) [1, 16], the Privacy Rights Clearinghouse [21], Sound Assurance [15] and the Identity Theft Resource Center [12] were combined and normalized for the time period of January 2005 through December 2009. The final data set contained 2,807 incidents from these sources. When originally queried, the OSF database listed 2,317 incidents for this time period. These events formed the base data set, and unique incidents that were identified from the other sources were added. Where possible, those cases where original source documents were available (typically in the form of notification letters sent from the company to either government agencies or the data subjects), further analysis was conducted.
In calculations where a subset of the data was used, notation is made to indicate the number of applicable records.

RELATED WORK

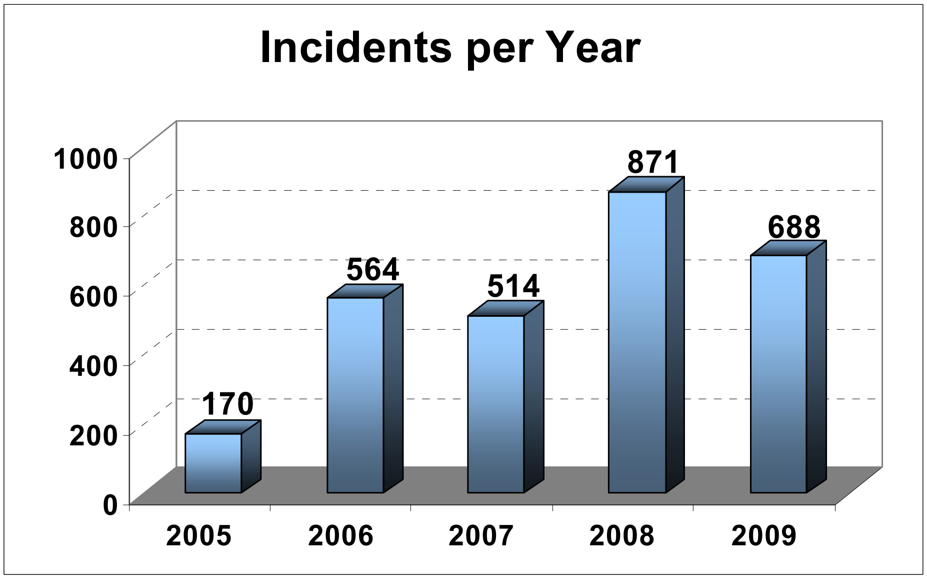
Many researchers have studied the various aspects of the data disclosure problem. This research was based on the initial work of Hasan & Yurcik who analyzed incidents from some of the same sources from 2003 to 2005 [10]. Campbell, Gordon, Loeb & Zhou studied the problem from the impact of a data breach on the stock market value of a company [4]. Cavusoglu, Mishra & Raghunathan studied the effects from a capital markets perspective [6]. Foley & Gordon focused on the impact to the consumers whose data was compromised [8, 9]. Hoofnagle focused on identity theft from top banks as a subset of incidents [11]. Romanosky, Telang & Acquisti focused on the laws and whether they reduce incidences of identity theft [22]. Finally, Baker, Hylender & Valentine produced an in-depth analysis of 500 incidents that Verizon Business investigated in 2008 and produced an updated analysis in 2009 with a larger team [2, 3].

ANALYSIS AND FINDINGS

The data was analyzed from several perspectives—frequency of incidents, number of records disclosed, breach vectors, geography, organizational and data types. Also explored was the relationship between the data subjects and the organizations and how the data subject victims were treated by the disclosing organizations. Finally, cost estimates are provided based on the work of others combined with the findings here.

Frequency of Incidents

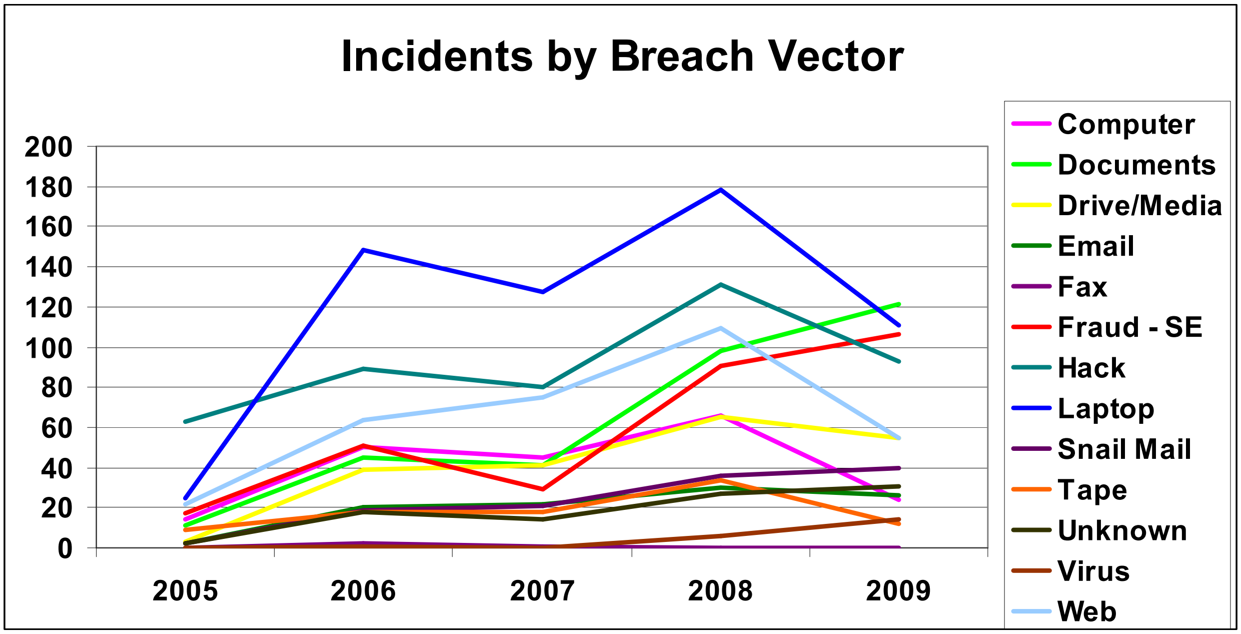
The Incidents per Year graph illustrates the trend in events over the past five years. There is a spike in the number of breach incidents in 2008, followed by a comparative drop off in the number of incidents since then. While the cause is a topic of some discussion among those organizations tracking these cases, no definitive cause has been identified [16]. In time, it may be determined that the spike in the 2008 total is the anomaly, and a steady increase is the overall trend, but that would require significantly more years of data to definitively conclude.



For some time, the OSF has been posting data from FOIA requests, including replicas of the original documents (submitted by the companies to government agencies) from some of these cases. This was extremely helpful, as viewing the source documents revealed additional metrics that could be captured. Where source documents were available, additional data was gathered—specifically the metrics dealing with stolen equipment (e.g., where they were stolen from), whether credit monitoring services were offered to the victims, and the nature of the relationship between the data subject and the disclosing organization. Without access to the source documents, some of this information would have been impossible to obtain.

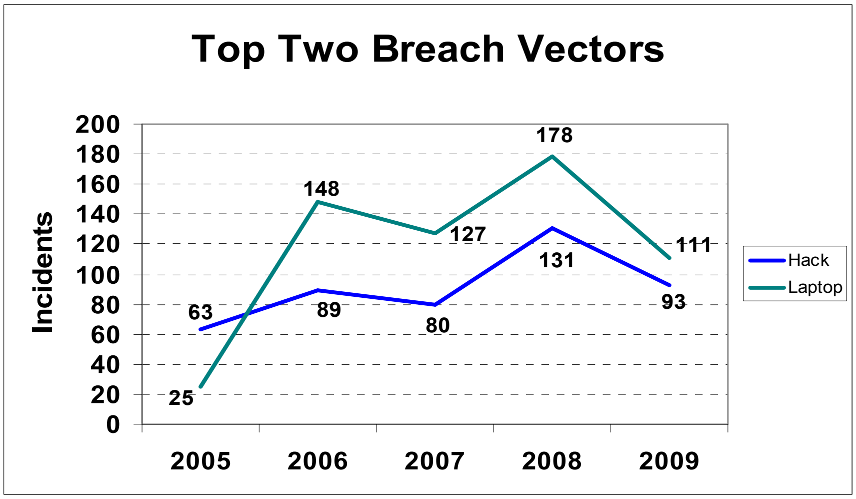
The OSF's approach to gaining access to these sources through FOIA requests has also provided access to breaches that the media did not report. Over time, this should increase the accuracy of the data set. It also illustrates the need for this kind of access to the source documents for researchers. In looking at these source documents, a frequent clause in them was found to be requests to keep the incident confidential to the agency where it was reported. Indeed, without these requests, these incidents still would not have come to light in the majority of cases.

While the increased media focus on these types of events since the passage of the breach disclosure laws assists with tracking, the lack of centralized reporting requirements makes it likely that these figures are a fraction of the actual number of incidents in the given time period. This has been borne out by the FOIA data that has resulted in numerous past unpublicized incidents being revealed [16].



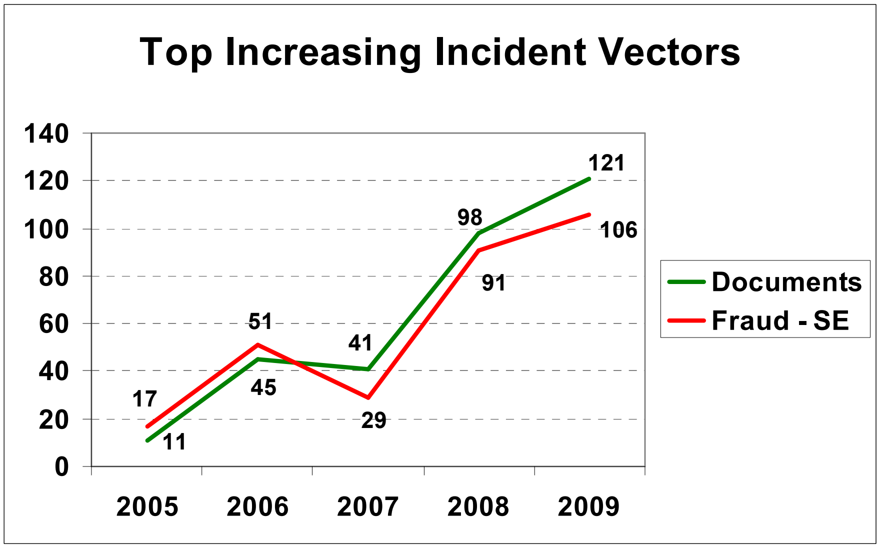
The Incidents by Breach Vector graph shows the incidents by vector over the course of the study. A definition of each vector can be found in Appendix A. Note the dominance of the Laptop vector, high above the others. Laptops were the leading vector for breaches across the study by a significant margin, and accounted for 21% of breach incidents. In fact, they were the leading incident vector for three out of the five years. As shown in the figure above, in 2005, the Hacking vector was the incident leader. By 2006, Laptops had taken the lead position and retained it until 2009, when the Documents vector took the lead.

Because laptops are both powerful and portable, there is a high likelihood that a company's sensitive data may reside on those systems as the percentage of their employees that use them as their sole computer platform increases. However, their portability also makes them more vulnerable to loss or theft. In this study, laptops were stolen (as opposed to lost) 95% of the time—most commonly from an office (33% of cases), followed closely by vehicles (at 28%). Their intrinsic value as easily fenced electronics makes them an attractive target for thieves, yet the company may face disclosing a data breach whether the target was the data or the electronics.



The Top Two Breach Vectors graph gives a closer view into the vectors that together accounted for 37% of the incidents in the study. Like most of the vectors, they show a general increase until 2009, when there is an overall decrease in incidents for the year.

The exceptions to this trend were the Documents and Fraud-Social Engineering (Fraud-SE) vectors. Both show a pointed increase from 2007 to 2009, despite the decrease in incidents in 2009. Looking more closely at these two, we can see the Documents vector took the lead position for the year for the first time. Documents are most frequently a vector for data breach in the disposal phase of their lifecycle. In 62% of the incidents, the disclosure was a result of careless disposal of sensitive data. This is in contrast with the 23% that are stolen and 14% that are lost.

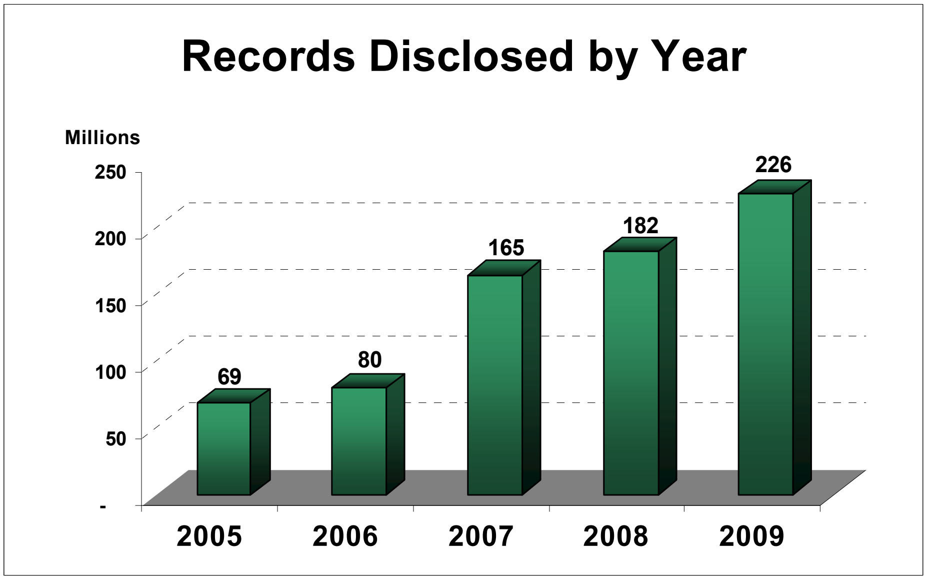


Those entrusted with sensitive information must not neglect how their paper documents are managed, particularly since security controls applied to electronic documents are disabled once they have been printed. If proper lifecycle management of sensitive paper documents is not part of the organization’s culture, the most stringent of controls on the data in electronic form won’t stop a breach if someone is careless with their printed material.

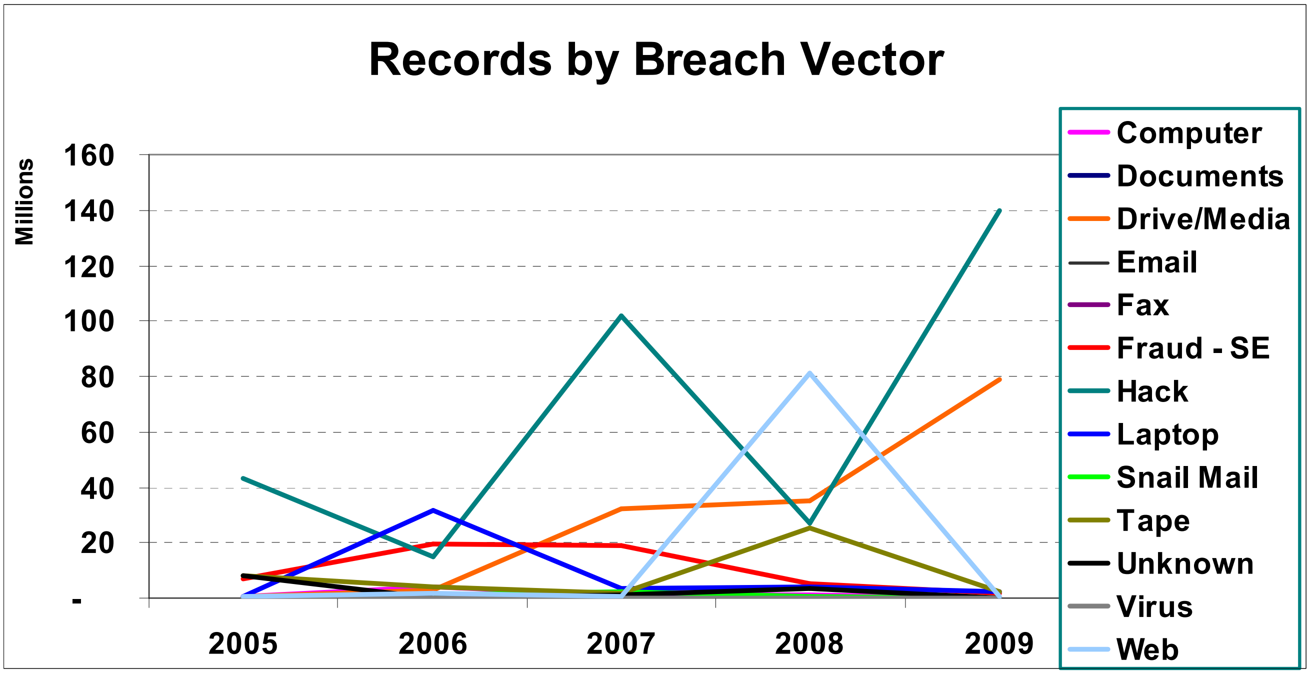
The Fraud-SE vector is most frequently an insider incident (69% of the time). These are the people with malicious intent, inside the network perimeter protections, with approved access to data. Only a small percentage of these are outsiders (15%), while third party partners (14%) are the least common in the Fraud-SE vector incidents. This vector is also the most likely to show evidence of subsequent criminal use of the data. This was the case in 62% (182) of the Fraud-SE incidents in the study. In contrast, the Hack vector only has 8% of the incidents designated as having confirmed criminal activity.

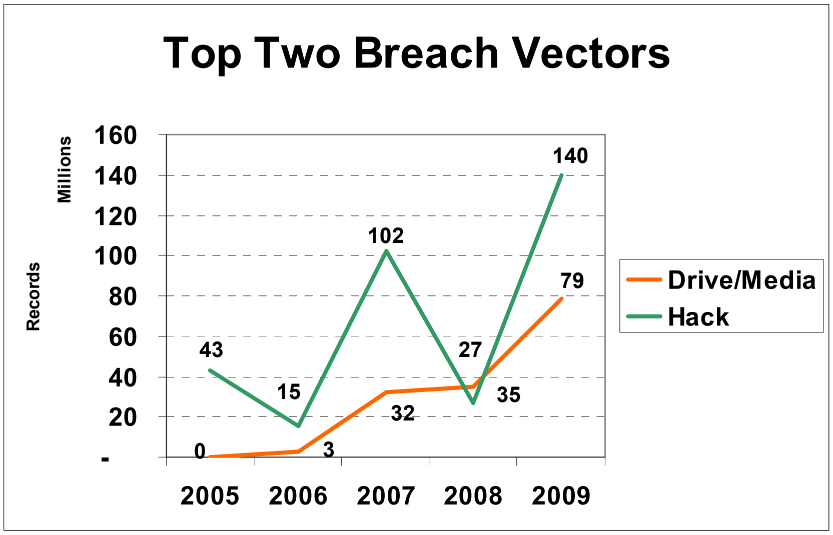
Records Disclosed

Tracking the number of incidents only gives a partial picture of the data breach impact. Arguably, it is the number of records disclosed that is the real figure of interest when looking at this type of data, as it represents the number of people affected by the breach. As shown here, the number of records disclosed per year is increasing while the number of incidents fluctuates. The highest number of records was in 2009, while the highest number of incidents per year was in 2008.



The Records by Breach Vector graph displays the risk by vector and breach size instead of by incidents. The largest were the Hack vector with 326.9 million records disclosed, and the Drive/Media vector with 148.6 million disclosed. The Web vector rounds out the top three with 84.2 million records. Between these three, they account for 78% of the records disclosed overall.

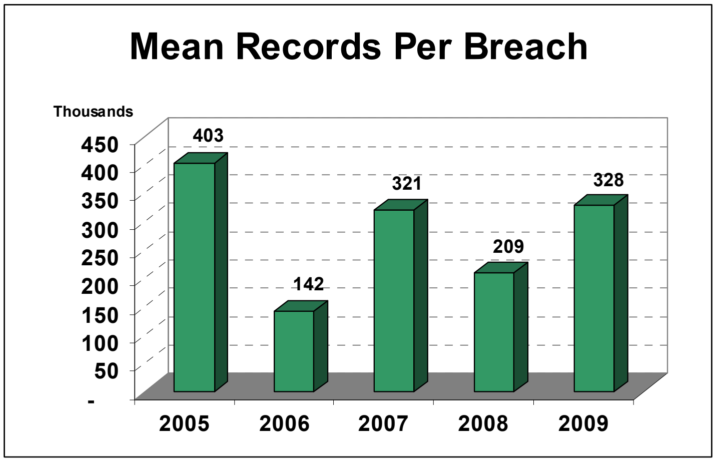




The Top Two Breach Vectors graph gives closer detail on these two vectors. Note that these are the only two vectors that are increasing between 2008 and 2009 despite the considerable drop in incidents between the two years. While the Drive/Media vector shows a steady increase (more on that later), the Hack vector has a strong variation year over year. Despite this oscillation, it retained the lead position for four of the five years as the top vector for records lost. In fact, despite the increasing nature of the Drive/Media vector, the number of records lost to the Hack vector in just one year (2009) nearly eclipsed the entire five year Drive/Media vector's total.

Statistical Measures

Looking at the increasing records figure despite the inconsistency in the number of incidents per year—sometimes increasing, sometimes decreasing—indicated a closer look was needed at the statistical measures of this data set. The first inclination would be to assume that the continual increase in the records disclosed indicates the incidents are simply getting larger year over year, and that is what is driving up the records disclosed figure, despite the fluctuations in the number of incidents.



To test this, first look at the Mean Records per Breach graph, which shows the variation in the average number of records disclosed per incident each year. In 2005, for example, with relatively few incidents, we see a high average record loss per incident. Conversely, the 2008 data showed a sharp increase in the number of incidents without a corresponding increase in the number of records relative to other years, so the average per incident figure went down.

Further statistics were calculated to help clarify how this data is distributed. Table T-1 provides statistics around the number of records in the data set. As shown below, the standard deviation is a large figure, indicating that the reported number of records varies by a wide margin over the data set. This is to be expected given minimum values as low as 1 record and maximum values as high as 130 million. The median figure is for the entire study dataset, and more granular data is provided for specific vectors in subsequent sections of the report.

Table T-1: Statistics on Number of Records Disclosed*

	2005 - 2009
Mean	387,926
Median	2,100
Min	1
Max	130,000,000
Standard Deviation	4,764,314

*For those incidents with finite numbers reported (1,861 total)

We will also look closer at the median records per year figures later when we get into the actors disclosing the records and try to identify where the greatest risk resides. In the meantime, when discussing total records lost in the hundreds of millions over the course of the study, it helps to get some perspective on what that means in terms of people affected.

Table T-2: Number of Records Disclosed/Year*

Year	**Records Disclosed**
2005	68,555,563
2006	80,363,058
2007	164,749,413
2008	182,414,761
2009	225,847,364
Total	**721,930,159**

*For those incidents with finite numbers reported (1,861 total)

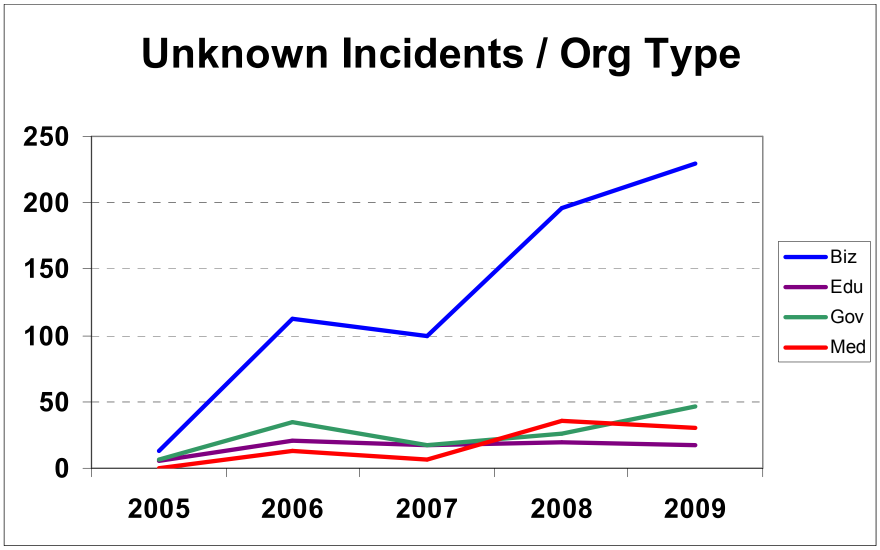
As shown in Table T-2, the total number of known records breached is over 721.9 million. Put another way, assuming 1,826 days between 1/1/2005 and 12/31/2009, these organizations lost an average of 395,362 people's data per day for every single day of those five years. It is interesting to note that as of July 2010, there were just over 300 million U.S. residents [24]. With the U.S. accounting for 91% of the incidents and 87% of the records disclosed (630.5 million), this means that each U.S. citizen's data would have been exposed more than twice in the past five years on average. Clearly this has not been the case, since that level of exposure would have created even more public outcry than that which triggered the passage of the current data breach laws. This means that there are unfortunate data subjects whose data has been compromised numerous times (and potentially different elements of their data) during the course of this study.

While it is evident that the number of records disclosed per year is increasing over the course of the study, despite the variations in the number of incidents per year, it is important to note that a significant number of breaches do not report actual numbers of records affected. They list the number exposed as “unknown” which makes estimating more accurate figures a challenge.

Table T-3: Unknown Number of Records Disclosed per Year

	# Incidents	# Listing Unknown	Total % Unknown
2005	170	25	15%
2006	564	182	32%
2007	514	141	27%
2008	871	277	32%
2009	688	322	47%
Total	**2,807**	**947**	**34%**

As shown in Table T-3, over one third of these organizations provided no finite number of records disclosed for their incident. As these numbers are not defined, they could range from miniscule to enormous. In fact, when the Heartland Payments Systems breach (the largest in the study at 130 million records) was reported, it fell into this category. It was several months later and after much media speculation that a number was finally determined. This fact should be kept in mind when reading the statistics about the records disclosed in each incident. The calculations show the scale of the situation, given the available data, but this is not a complete picture. In fact, the level of uncertainty indicates that these numbers underestimate the actual figures, since “unknowns” are counted as a zero records disclosed value.



The number of incidents listing unknown values varied significantly from year to year, as shown above, although the figure overall was 34%. This trend of increasingly refraining from disclosing the size of breaches has increased over the years. It started out at its lowest in 2005 with just 15%, and by 2009 was the highest with 47%, with only minor fluctuations (±5%) between the intervening years.

The problem of the unknown values varied across organizational types as well, as shown below. The Business sector was the most frequent to have undefined record disclosure figures by a wide margin. The other three sectors put together made up a fraction of the Business total.

As indicated above, given that so many of the incidents reported do not indicate how many records were disclosed, the actual loss figure is potentially considerably higher. To get an estimate of how much higher, calculations were made based on number of records per breach in those that reported finite numbers.

Table T-4: Mean and Median Number of Records/Breach*

Year	Known # Records Disclosed	Mean Records/ Breach	Median Records/ Breach
2005	68,555,563	403,268	9,450
2006	80,363,058	142,488	3,128
2007	164,749,413	320,524	3,250
2008	182,414,761	209,431	1,151
2009	225,847,364	328,267	1,076

*For those incidents with finite numbers reported (1,861 total)

Table T-4 shows the mean and median number of records per incident. This takes into account only those incidents where the number of records disclosed is a known quantity. The median figure, however, is the more accurate of the two, since the number of records per breach varies so widely in this data set. The median record figure is decreasing over time, potentially as more data points with finite numbers of known disclosed records are identified; the data is becoming more accurate. The increase in the number of records with lower record loss per incident helps to negate the impact of those sensational, media-favored, high record loss incidents. Many of these critical incidents were found through the previously mentioned FOIA requests [16].

To get a granular estimate of the scope of the underreporting, the median records disclosed figure was calculated on a yearly basis per breach vector. This will allow for more accurate estimates of the total records disclosed if the records marked as "unknown" are calculated using these median figures, since we know the year and vector for each incident. This is the formula used:

known records disclosed per vector and year	+	median records per breach vector per year	X	number of incidents where records lost is "unknown"

While this is an estimate, it is the best data we have given the high percentage of uncertainty in self-reporting the number of records disclosed. For example, the Drive/Media vector in 2008 was responsible for the known disclosure of 32,017,656 records. The 2008 median figure per incident was 3,000, and 11 incidents were listed with the number of records disclosed as unknown. To get the total of how many records were disclosed including the estimate, we use the median figure for those "unknown" values:

32,017,656	+	3,000	X	11	=	32,050,656
known records		median records		number of incidents		New Records Disclosed Total

Table T-5 shows the results of applying this calculation across the breach vectors for each year.

Table T-5: Estimate of Records Disclosed

Year	# of Unknown Incidents	Known Records disclosed	Estimated Additional Records Disclosed	Estimated Records Disclosed Totals
2005	25	68,555,563	3,634,950	72,190,513
2006	182	80,363,058	954,614	81,317,672
2007	141	164,749,413	1,382,681	166,132,094
2008	277	182,414,761	695,882	183,110,643
2009	322	225,847,364	930,641	226,778,005
Total	**947**	**721,930,159**	**7,598,768**	**729,528,927**

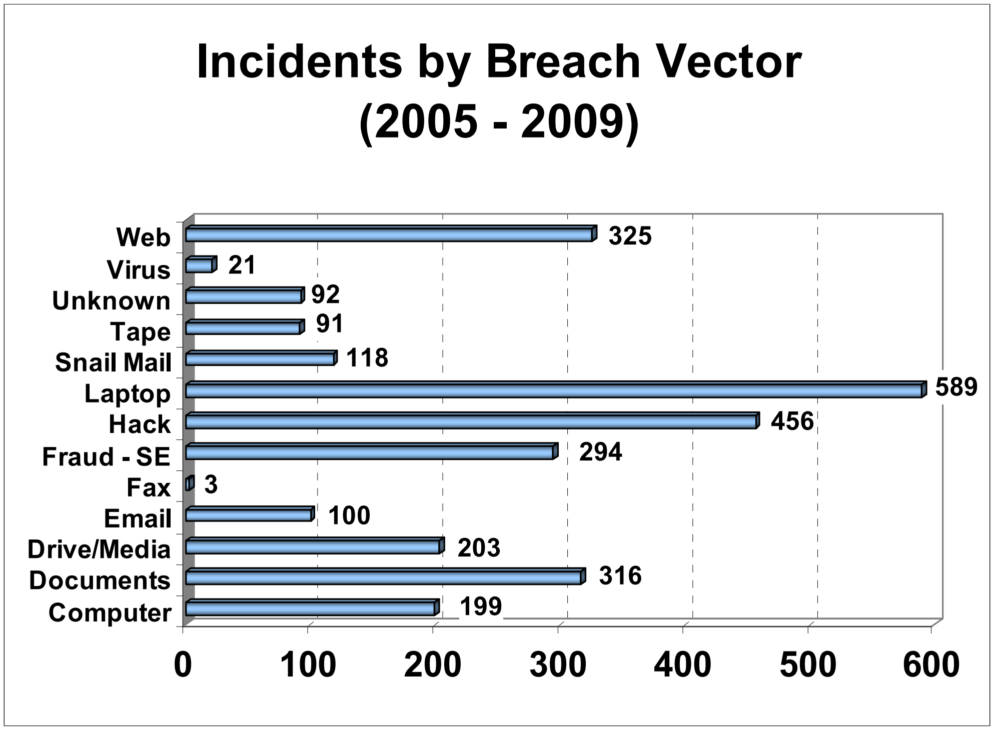
In this manner, we can estimate that an additional 7.5 million records were exposed over what was previously reported, bringing the estimated count to over 729.5 million. By increasing the number of data points over time, the high and low values should no longer have such a disproportionate impact on the median figure, and this estimate should become more accurate. Certainly, there are sectors and specific categories within them that have higher and lower incidence of the values reported as "unknown" for number of records disclosed. As indicated earlier, the Business sector has the highest number of incidence of this type of under reporting. Certain breach vectors seems particularly likely to be underreported as well—for instance, while the overall rate is 34%, the Document vector lists 53% of the incidents as "unknown". The Fraud-SE vector is in second place with 43%. The Drive/Media vector had the lowest number of incidents reporting "unknown" records disclosed at 21%.

What is not indicated in the incident reports is if, when an organization lists the number of records disclosed as "unknown", this means that they do not attempt to notify the data subjects. Consequently, the above estimate of 7.5 million records may represent people whose data has been disclosed, but who have not been notified that they are at increased risk for criminal use of their information. This would seem to defeat the intent behind the requirement that the breach be disclosed. In calls to have a single Federal breach disclosure law, a requirement to at least estimate the number of records disclosed should be a part of the reporting. This would help researchers to get better data on the phenomenon as well as serve the public in ensuring efforts are made to scope the breach appropriately and notify the affected.

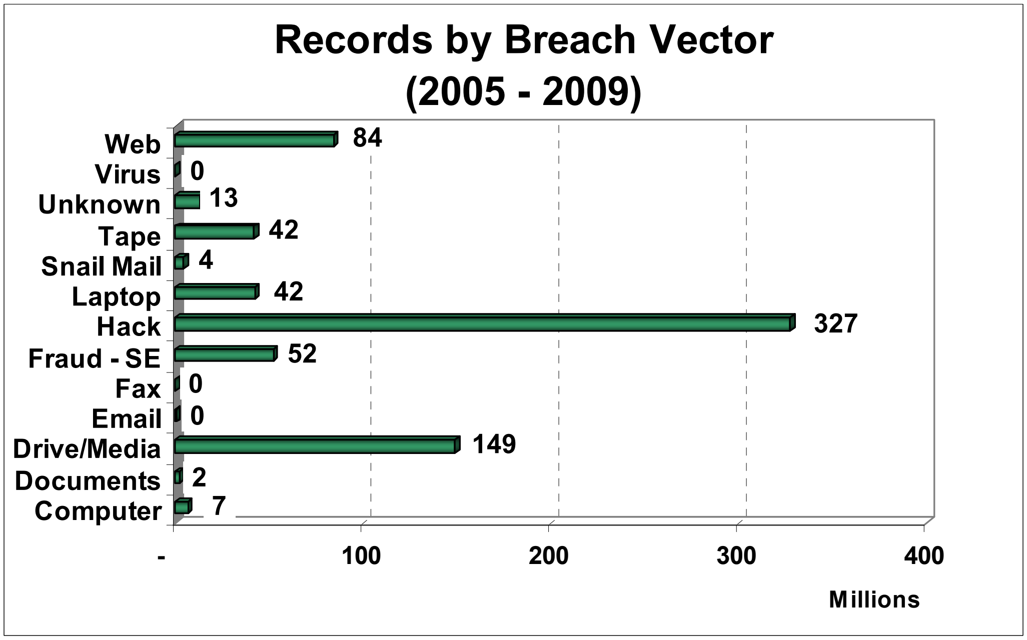
With this many records released, the next logical question is “How are these records being disclosed?” To answer that, the data breach vectors were studied.

Breach Vectors

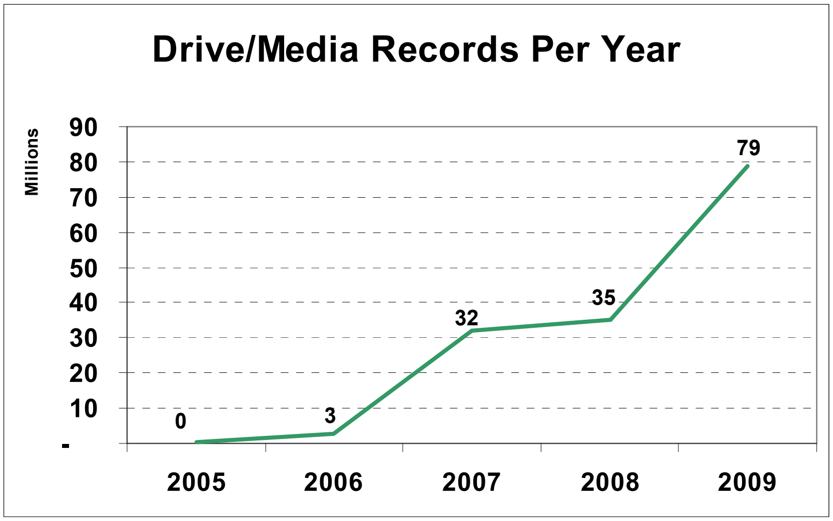
The high number of records exposed demonstrates that involuntary data disclosure is a very large problem. For more insight, the vector data was studied to determine where the greatest risk resides. As the Incidents by Breach Vector graph indicates, the leading incident vector overall is Laptop, followed by Hack and Web. Between these three categories, they accounted for 49% of all incidents.



For comparison, the Records by Breach Vector graph illustrates the records disclosed by vector for the length of the study. In that case, the Hack vector is the loss leader for records, accounting for 326.9 million records divulged. While laptops were the leader in incidents, they accounted for only 6% of the records disclosed, at just above 42 million records.



In contrast, the second highest method for record loss is the Drive/Media vector, responsible for 148.6 million records disclosed. In fact, this breach vector saw the fastest growth over the study. From a relatively slow start in the first two years of the study, the Drive/Media experienced significant growth as shown in the Drive/Media Records per Year graph.



This category includes portable hard drives, USB thumb drives and other types of portable media. The growth seen in this area is likely due to:

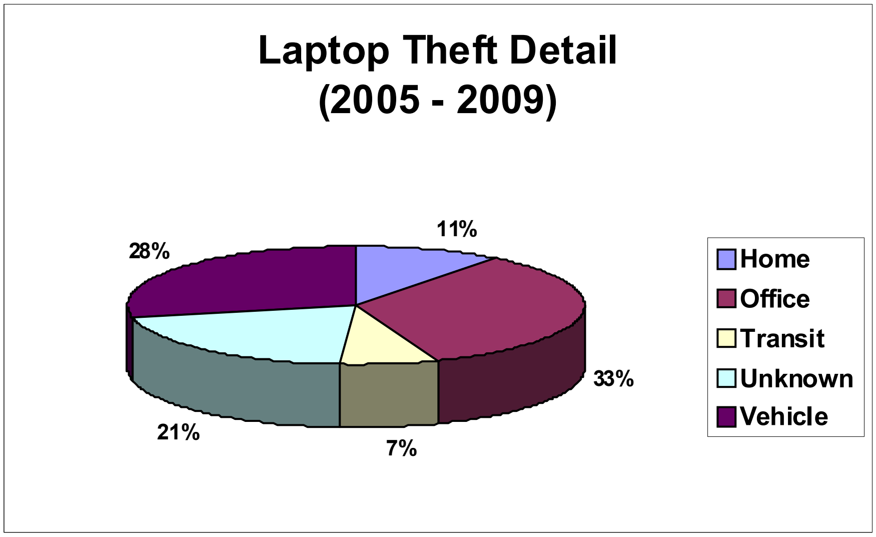
- the increased popularity in these types of media
- how easy it is to lose track of them
- the capacity on these devices has increased over the past five years
- the cost has dropped dramatically

All of the above are contributing to the growth in this vector. In 2005, the Drive/Media vector was responsible for 310,790 records. In 2009, it was responsible for a record loss of 78,666,741.

The Drive/Media vector is a good candidate for data at rest encryption as a control to provide safe harbor in the event of the loss of the media. If the data must be stored on portable storage media, it should be encrypted so that the loss event doesn't necessarily expose the information.

The Laptop Vector

As shown previously, the laptop vector is the leader in incidents by a significant margin. Laptops being stolen accounted for 95% of the incidents in this vector, with the remaining 5% representing lost laptops. As indicated in the Laptop Theft Detail graph, 28% of the time, laptops are stolen from vehicles, compared to 33% from the place of business. The Unknown category represents those records where the report did not indicate where the laptop was stolen from. The Transit category represents cases where the laptop was in the possession of the person it is assigned to, but they were traveling, or the laptop was being transported by a third party.



As this vector illustrates, organizations must understand just how portable their data has become. If they are not cognizant of the location of their data, they cannot hope to keep it secure from the simple loss of such a convenient electronic device.

A 2008 study by the Ponemon Institute focused on airport laptop loss incidents, and it serves to illustrate how underreported these cases may be. That study determined that over 12,000 laptops are lost per week at airports alone. At 12,000 per week, the number of laptops reported that contained sensitive data should have been much higher than the 589 reported over the course of this study. It is clear that most of these incidents never made it into the tracking database, and thus the dataset of this study. Given an average figure of 12,000 lost per week, the total laptops lost should approach 2.5 million laptops over the course of this study. Instead, only 39 laptops are listed in the "In Transit" category, which encompasses travel related incidents. Even assuming that only 1% of these laptops contained sensitive data, there should be closer to 21,000 incidents in this vector alone [19].

> Given the high number of laptops lost in airports, companies should be looking at methods to protect the data on these systems should they become lost or stolen [18]. Awareness programs should stress the value that thieves perceive when they encounter an unprotected laptop. The FTC recommends people treat laptops like cash. "Like a wad of money, a laptop in public view—like the backseat of the car or at the airport—could attract unwanted attention" [7].

If a laptop must be left in a vehicle, the time to put it in the trunk is before leaving the office, not once the person reaches their interim destination. Putting something in the trunk for "safe keeping" at the destination alerts watching thieves exactly where the valuables are located. This is why it is important to secure the valuables, and then move the car from the premises. Also, the organization's awareness program should stress that employees not leave laptops in vehicles overnight. They should be taken inside when the employee arrives home. When a laptop is in the place of business, it should be locked to prevent theft. These basic physical security measures can prevent this vector from causing a breach. The less time spent unattended and unprotected, the better.

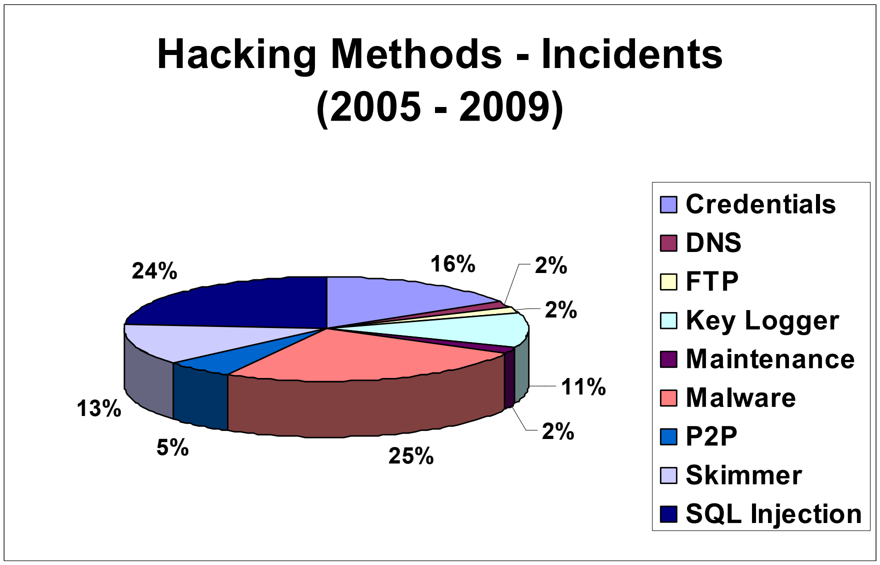
One of the assurances mentioned time and again in the letters sent to the data subjects about the incident was that the laptop, stolen while containing their sensitive data, was password protected [16]. This sounds reassuring, but in testing it took only a couple of minutes (the time it took to boot from a CD-ROM or USB device) to bypass this control. There are a number of tools freely available on the internet to perform this function. In fact, a quick internet search turned up approximately 948,000 results for the search term "bypass windows login". There are even tutorial videos on popular media hosting sites.

In testing with Kon-Boot (one of the many choices), access to the non-encrypted, password protected laptop took a few short minutes, and the data was freely accessible [5]. When tested against a laptop that employed encryption, however, while the system booted and the file names were visible, access to the content of the files was not successful.

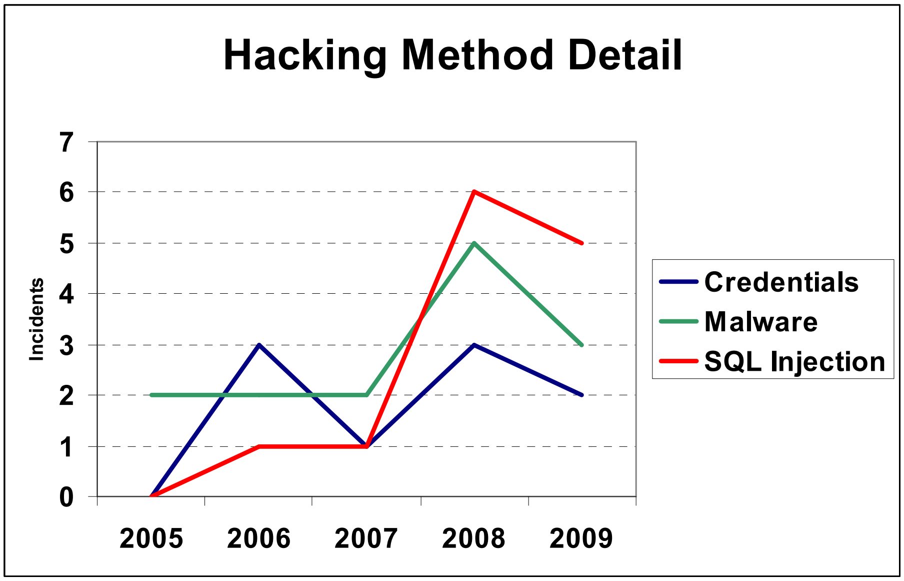
Companies relying on login password protection from the operating system or central directory services need to realize that once physical custody of the device is lost, it is trivial to obtain access to the data when other controls are not in place.

The Hacking Vector

The Hacking vector was the record loss leader by a significant margin. For the purpose of this study, this vector refers to someone actively trying to gain unauthorized entry into an organization's systems. When successful, these incidents result in the highest number of disclosed records. Despite only 456 incidents in this vector, it was responsible for an average of 716,925 records per incident. Put another way, the hacking vector, which accounted for only 16% of the incidents accounted for 45% of total records breached over the course of the study. Thus, while the most frequent incident is a laptop loss, the highest risk by number of records disclosed involves people attacking an organization's systems directly.



Most of the incident reports do not give much detail as to what method was used to hack into the systems. For the 12% of the Hack vector incidents that did provide this information, the Hacking Methods (Incidents) graph illustrates the frequency of hacking methodology choices. The top vectors were Malware, SQL Injection and compromising login credentials. SQL Injection started out slow the first three years and then accelerated in 2008 and 2009. In fact, in other studies, this has been found to be the method of choice when the easier routes are not available. Since the weakness is in the input validation of the applications, this vector would benefit from a program targeting the organizations coding practices [3]. The Hacking Method Detail graph illustrates the top three choices in terms of incidents where the method was known.

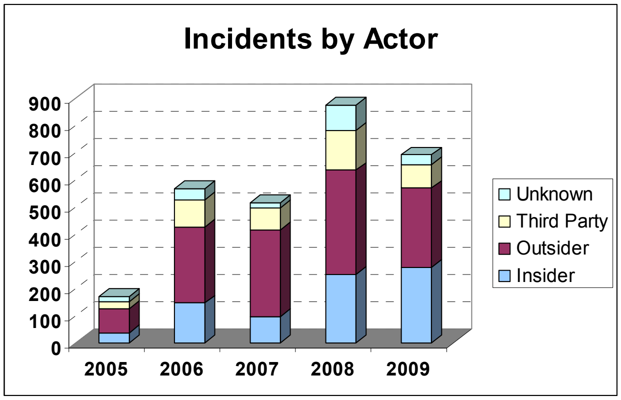


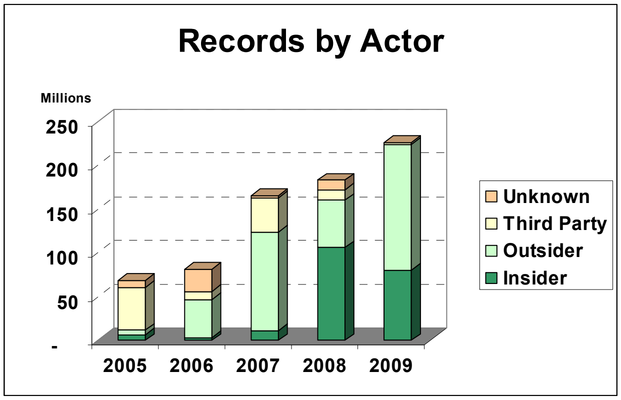
Most commonly, however, the method is not mentioned in the source documents or reports on the incidents. The data provided is from the 98 incidents that listed a specific methodology used. This accounts for just 3% of the hack vector incidents, so until more detail is available, the information is presented here just to give an idea of trend. This lack of information illustrates a weakness in the current self-disclosure process—that no standardized metrics for collection exist between the differing breach notification laws. When researchers must rely on news reports, the details frequently become scarcer. In 53% of the cases in the above chart, source documents were available and listed the actual cause from the reporting organization. The remaining 47% of known causes were gleaned from media reports of the incidents.

Insider, Outsiders and Third Party Partners

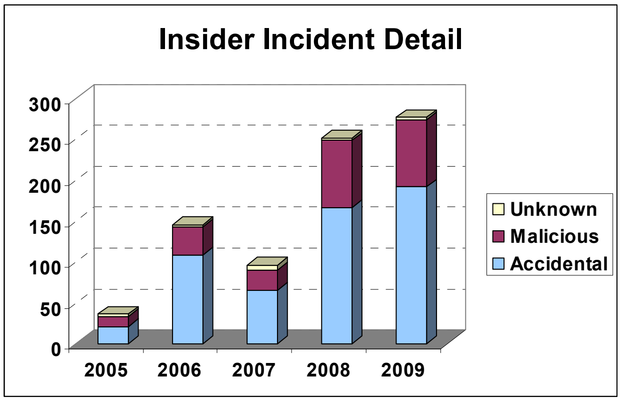
In this category, the incidents were classified by the actors involved—whether the breach was the result of an organizational Insider—someone with permission to have access to the system from within the company; an Outsider—someone who did not have permission; or a Third Party Partner—someone entrusted by the organization with the (temporary or permanent) custody of the data.

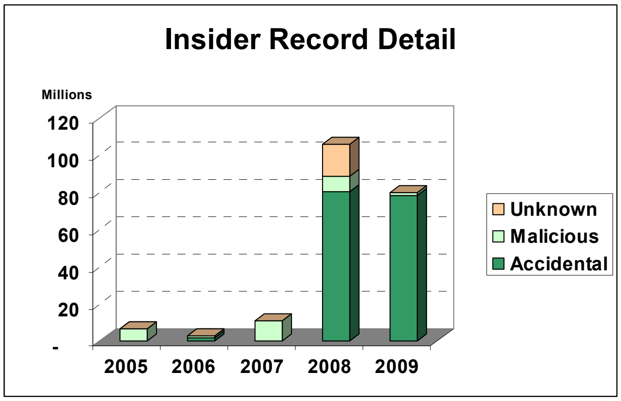
The Incidents by Actor graph illustrates the risk on a per-incident basis. Incidents by Insiders were less overall than those by Outsiders, and Third Party Partners were fewer still. The Records by Actor shows the differences in who is responsible versus the size of the breach for each Actor.





Outsiders were responsible for 48% of the cases and 50% of the records disclosed. Attacks from within were reported less frequently as the vector for a data breach, accounting only 29% incidents, and responsible for 29% the records exposed.





As the above graphs show, when the incident involves Insiders, it is more than twice as likely to be accidental in nature as someone behaving maliciously. As the Insider Record Detail graph shows, Insider's mistakes caused far less damage in the first three years of the study than they did in 2008 and 2009, when they caused millions of records to be disclosed.

While these percentages are compelling, the actual records tell the story. As Table T-6 below indicates, the number of records disclosed by Outsiders is considerably higher than those by Insiders.

Table T-6: Insiders, Outsiders & Third Party Partners Records Detail

	2005 - 2009
Insider	205,950,745
Outsider	357,601,885
Third Party	111,038,481
Unknown	47,339,048
Total	**721,930,159**

This finding is in contrast with the general assumption that those with inside information are in a better position to wreak havoc than those who are not [2]. Looking at the sources over time in the above graphs, you can see that Outsiders are consistently the largest source of breach incidents for all five years of this study. Table T-6 shows that Third Party Partners posed just under half the risk of records disclosed over the course of the study as Insiders. That trend started strongest in the early years of the study, and by the end it had been eclipsed by the Outsider threat.

Table T-7 shows the breakdown of the Insider category. Insider accidental data loss incidents disclosed more than three times the record loss of the other two causes combined.

Table T-7: Insiders Records Detail

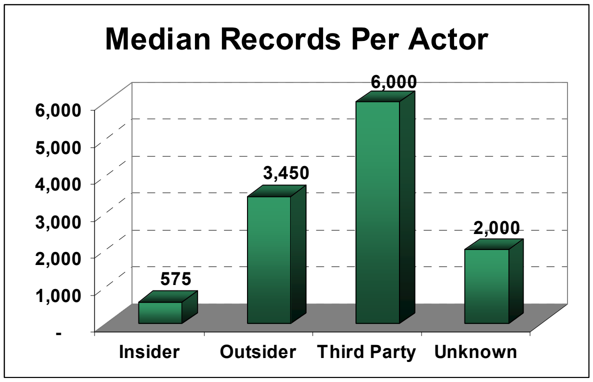
	2005 - 2009
Insider	
Accidental	160,359,601
Malicious	28,358,365
Unknown	17,232,779
Total	**205,950,745**

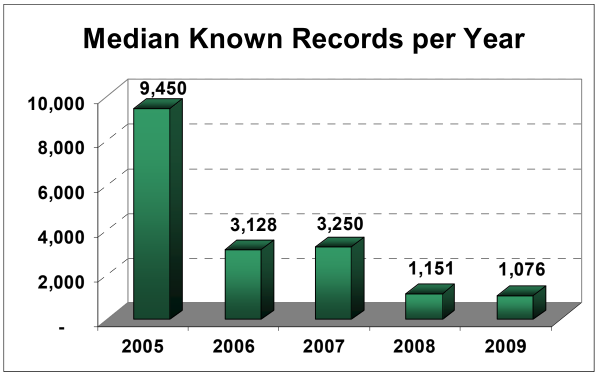
Companies are increasingly outsourcing non-core competencies. This means they are also outsourcing the security for that data that goes with the work, whether they acknowledge and plan for it or not.

“A third-party breach is defined as a case where a third party (such as professional services, outsourcers, vendors, business partners) was in the possession of the data and responsible for its protection [18].” When performing initial calculations for return on investment in decisions involving partner access, security concerns must be at the forefront lest companies find themselves in the same predicament that 16% of these organizations did. Over the course of the study, third party facilitated breaches were responsible for over 111 million records disclosed.

The median size of a data breach involving an Outsider is significantly larger than for an Insider as shown in the Median Records per Actor graph. However, the median size of a breach involving a Third Party Partner exceeds even that of an Outsider. This illustrates the increased risk of companies outsourcing the processing of their data to third parties.

When engaging a Partner, the company assumes the risk of the Partner's systems and processes in addition to their own where that data is concerned. When performing the cost/benefit analysis of using third parties, these incremental risks must be taken into consideration, or the company is operating under an inaccurate risk picture [18].





The data was reviewed to determine the most common vector in third party facilitated breaches. Unsurprisingly, the highest vector for third party incidents is Laptop with 27% of the cases. Of those cases where the details about the laptop theft were disclosed, the laptops were stolen from offices 30% of the time and vehicles 26% of the time. The highest vector for records disclosed in third-party cases (37%) was the Hack vector, responsible for over 40.5 million records. In some cases, the partner can become a conduit for the outsider to gain access to your infrastructure:

> Partner-side information assets and connections were compromised and used by an external entity to attack the victim's systems in 57 percent of breaches involving a business partner. Though not a willing accomplice, the partner's lax security practices—often outside the victim's control—undeniably allow such attacks to take place. Exacerbating this situation, the victim organization frequently lacks measures to provide accountability for partner-facing systems [2].

The security requirements for partners should be "baked in" to the contracts prior to the start of the relationship. Penalties should be spelled out and the required controls for data security (both detective and preventive) should be clearly defined. Companies must hold their partners accountable; since they will be the ones suffering the consequences of a third party facilitated data breach. In addition, provisions for the reclamation of a company's data and the sanitization of the third party's systems should be included in the contract in the event that the partnership is terminated [18].

In looking at the data, the majority of the Partner breaches did not name the company responsible. For those that did name them, the top five are listed here:

Table T-8: Most Frequent Third Party Partner Breach Sources (Incidents)

Partner Organization	Total Incidents
United Parcel Service (UPS)	14
Iron Mountain	7
Deloitte & Touche	5
Affiliated Computer Services (ACS)	5
Electronic Data Systems (EDS)	4

*With 378 incidents listing the name of the third party involved

Some of this is self explanatory—UPS ships millions of packages, and the amount that go astray are statistically few. However, each time an organization's data is in transit—whether through a courier or a network—the consideration should be protecting it from disclosure. Most of the laws have some provision for encrypted data [23]. The Iron Mountain entry illustrates this point as well—these were all tapes being transported, and the number of records lost in these cases were particularly high (1.9 million) due to the volume of data stored on a typical backup tape.

Table T-9: Largest Third Party Facilitated Breaches (Records)

Partner Organization	Total Records
CardSystems	40,000,000
TNT	25,000,000
Certegy Check Services, Inc.	8,500,000
United Parcel Service (UPS)	4,649,628
Archive Systems Inc.	4,500,000

For the largest third party facilitated breaches, the lost data was primarily customer information. The top three were credit card data, with the addition of financial data in the Certegy breach. The Archive Systems breach exposed customer SSNs, Names and Addresses.

Criminal Use

An interesting pattern came out of the data in the records where the data was confirmed to have been used for fraudulent purposes subsequent to the disclosure. First, the top two vectors for this were the Fraud-Social Engineering and the Hack vectors, responsible for 219 of the confirmed 251 criminal use cases. These cases where subsequent criminal activity was confirmed involved over 203 million records. While the full number of records was disclosed, the cases do not automatically indicate that 100% were directly used. In many instances, new accounts were created in a subset of the victim's names. However, the nature of data theft is that the data is resold multiple times, and the total number of compromises may not have surfaced yet. With the increasing involvement of organized crime into the compromise of financial data, and the drop in the cost per record for compromised data, the likelihood is high that the number of records compromised versus the number actively used will rise [3].

In looking at the records where both evidence of subsequent criminal use of the data was present, and where the question of whether credit monitoring was offered was answered, the data was broken into the relationship between the victim organization and the data subject whose information was used. It should be noted that for 176 of the 251 cases, both of these pieces of information were not available. However, for 25% of the remaining cases where all data was present, customer data was lost. For those cases, 53% of the customers were offered credit monitoring to help salvage the relationship, while 47% of them were not offered even this small compensation for data that has been disclosed and the resulting fraud. Considering how many people an unhappy customer tells about their experience, organizations should evaluate the impact of leaving these data subjects to clean up the mess, and how that may cost them more than just the customer they know about.

The majority (57%) of the study's incidents do not indicate for certain in the reports whether credit monitoring is offered. The largest incident where credit monitoring was offered was 12.5 million. In that case, no indication was given that criminal use was present.

Geographic View

As mentioned previously, the United States accounts for the vast majority (2,557) of the incidents and records disclosed (630.5 million). Great Britain came in a distant second with 128 incidents and just over 30 million records. All told, a total of 28 countries reported incidents.

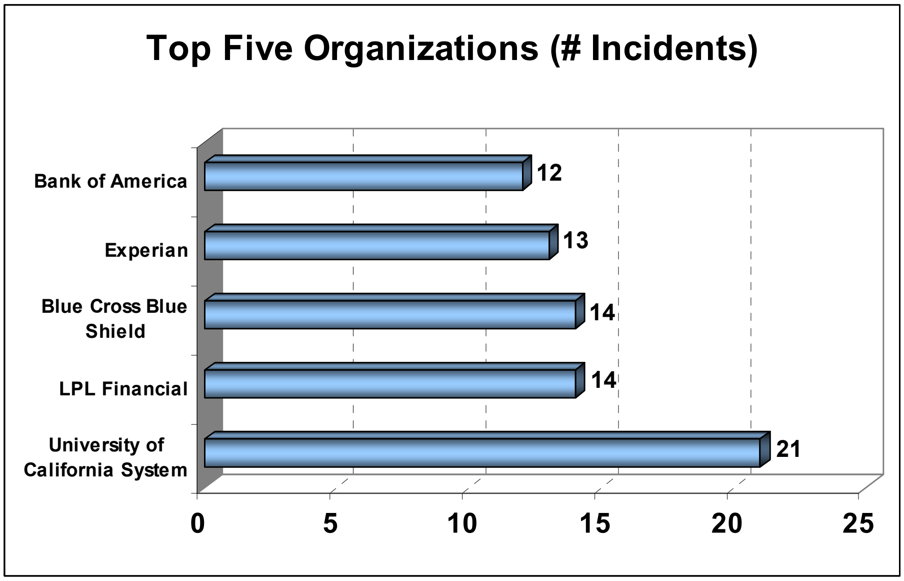
A combination of causes is likely making the United States the leader in data breach incidents. First, the sites that are tracking these cases are primarily in the United States and focused on the U.S. more than international incidents. Second, the majority of countries do not have laws requiring disclosure of data breaches, which makes learning about an incident far less likely.

There were data breaches recorded for each state in the U.S., despite laws in only 46 (the exceptions are New Mexico, Alabama, Kentucky and South Dakota). This is primarily because it is the residence of the data subject victims that is used in determining jurisdiction rather than the organization's location for many of the laws. Thus, while an organization may reside in a state without a breach disclosure law, if the people whose data they disclosed reside in a state with a reporting requirement, their incident must be reported [23].

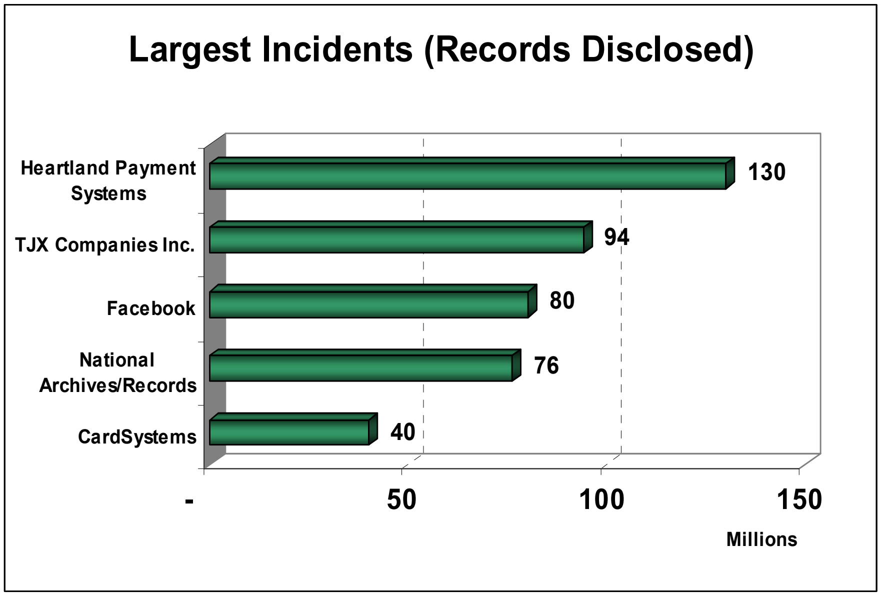
For incidents, New York (11%) and California (10%) are the leaders. New York's leading breach vector was laptop with 65 incidents. California's leading vector was also laptop with 54 incidents, but Hacking came in a very close second with 53 incidents. These two states have both had data breach laws on the books since 2005, which may account, in part, for their leading position in incidents reported.

Organizational Sectors

The first item of interest was to determine which organizations had the most incidents in the study. The Top Five Organizations graph shows the results. Three are from the Financial sector, which follows considering the value of the data they hold to the criminal element. The remaining two—Blue Cross Blue Shield and the University of California system are made up of multiple locations experiencing breaches.

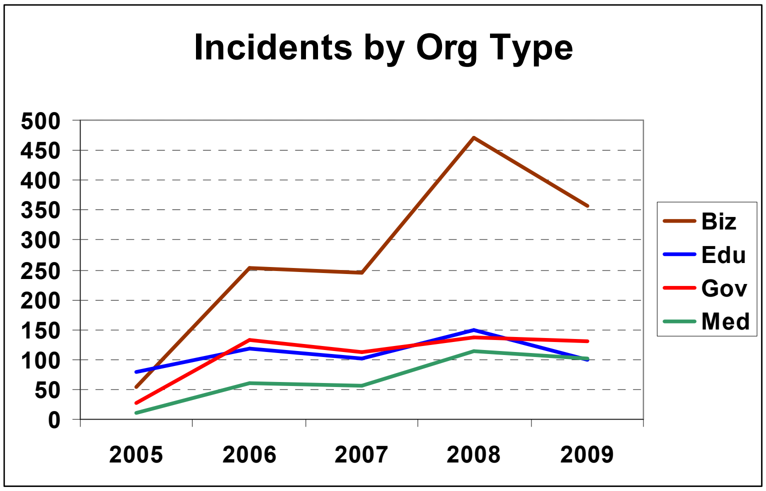


In the case of the University system, each university may function essentially as a separate entity, which has both good and bad points—it limits the breach scope to the local school's data, but unless there is sharing of knowledge in security controls and defense techniques that work, it compounds the vulnerabilities rather than the potential strengths. Blue Cross Blue Shield listed several geographic qualifiers to the breach incidents, implying that it too has a divided structure that may mean the responsibility for securing the data is decentralized.



The Largest Incidents (Records Disclosed) graph shows the biggest records lost breaches in the data set. All of these incidents caused a media sensation and were reported on extensively.

Organizations may hope never to be in this position—having a breach that gains national media attention—but since it remains a possibility, contingency planning to reduce the impact of the event would be a good step to take. This type of plan should be a part of the regular Incident Response planning, and even a part of the Business Continuity Plan. The time to determine how a major breach should be handled is not when the media are gathered at the front door.



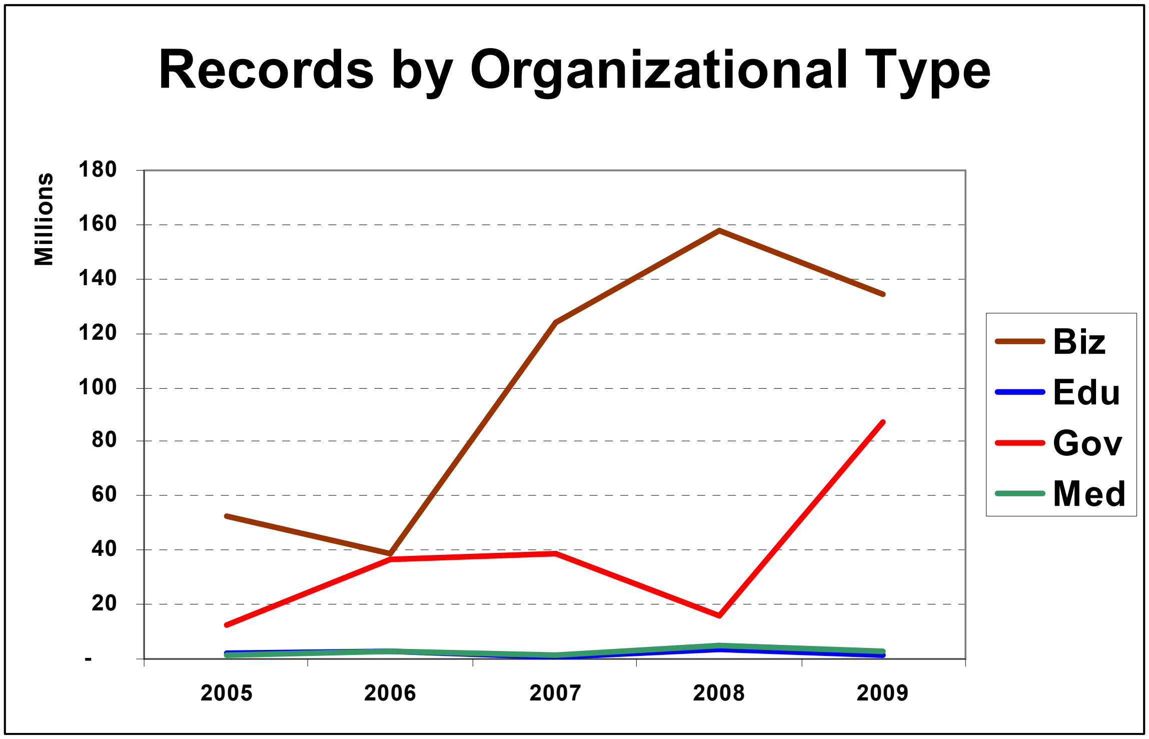
To study the organizational differences in the data set, each incident was classified into one of four major categories: Business (Biz), Education (Edu), Government (Gov) and Medical (Med). The Incidents by Org Type graph shows the trending for each sector over the course of the study. It illustrates the fact that the Business sector is the leader in incidents, both overall, and in four of the five years of the study. Only in 2005 did this sector take second place to the Educational sector.

Table T-10 shows the detail on the number of incidents by year by organizational type. With the Business sector responsible for almost half (49%) of all incidents, a closer look into the subsectors was warranted.

Table T-10: Number of Incidents per Year by Organizational Type

	Organizational Type				
Year	**Biz**	**Edu**	**Gov**	**Med**	**Total**
2005	54	79	26	11	170
2006	253	118	133	60	564
2007	245	102	112	55	514
2008	470	150	137	114	871
2009	356	100	131	101	688
Total	**1,378**	**549**	**539**	**341**	**2,807**

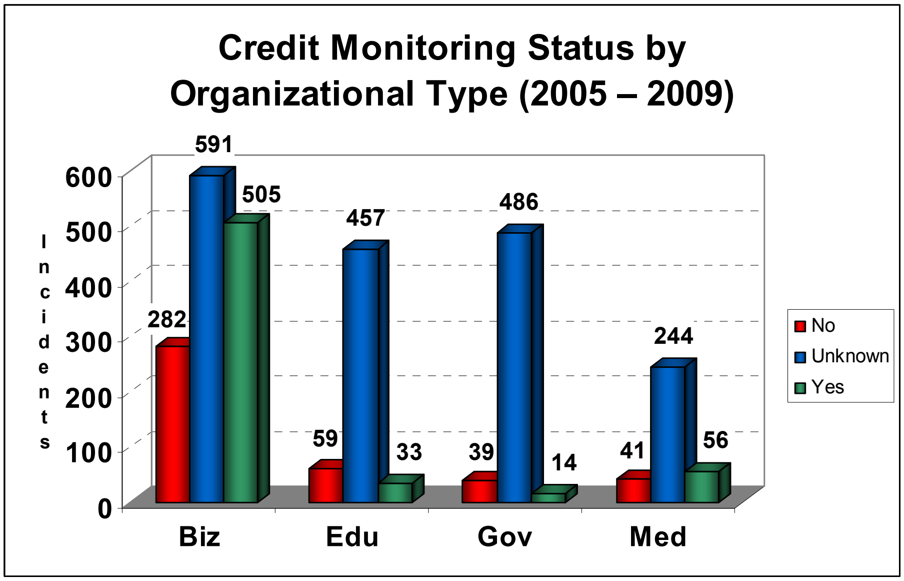
This leadership continues when looking at records disclosed in the Business sector as well. Government comes in a distant second in the records exposed section, although it ranks a very close third in incidents.



This indicates that per incident, the Government sector is losing significantly more records on average than the Educational sector.

Table T-11: Number of Records per Year by Organizational Type

	Organizational Type			
	Biz	Edu	Gov	Med
2005	52,292,228	1,882,896	12,743,605	1,636,834
2006	38,696,911	2,659,315	36,414,335	2,592,497
2007	123,852,691	938,610	38,677,012	1,281,100
2008	157,924,504	3,435,558	16,265,847	4,788,852
2009	134,496,065	1,488,398	87,307,862	2,555,039
Total	**507,262,399**	**10,404,777**	**191,408,661**	**12,854,322**

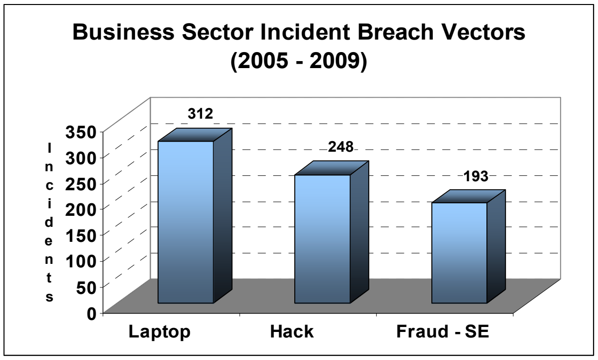


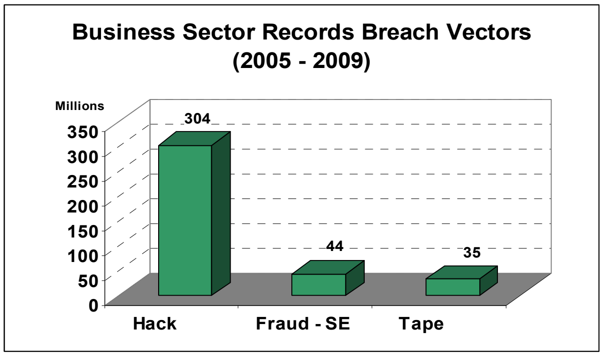
Another area of interest was to determine how the disclosing organizations are treating the data subject victims. Was any type of monitoring service offered to help them keep on top of any malicious use of the data disclosed? The Credit Monitoring Status by Organizational Type graph shows the results. In 37% of the cases, organizations chose to offer monitoring services. In 20%, they chose not to offer this service, and in the remaining incidents, the information was not provided. The Educational sector offered monitoring even less often, but the number of incidents where the outcome was known was very low. The same problem arises in the other two sectors—the number of known cases is so small that it is difficult to see this data as anything other than providing a general trend.

Each of these Organizational sectors was studied further to gain more insight into how they differ.

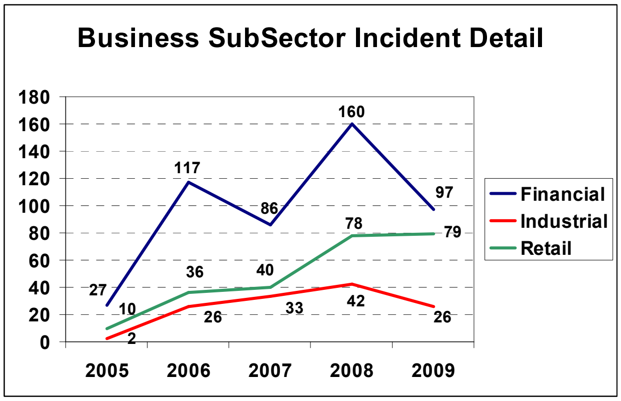
The Business Sector

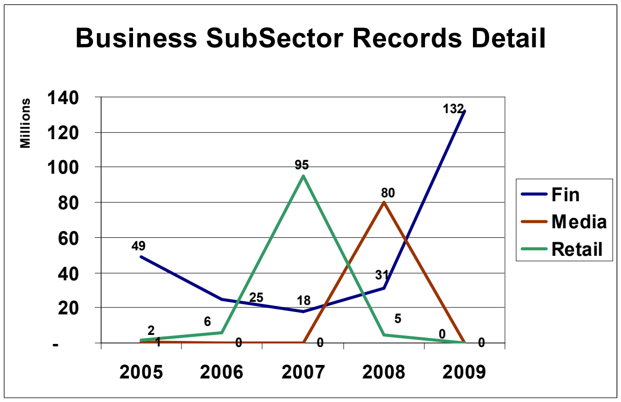
The Business sector was responsible for the majority of incidents and records with 1,378 and 507.2 million respectively. The following shows detail into the top three breach vectors for both incidents and records for this sector.





There are some large subsectors within the Business category. The largest of these by far is the Financial subsector. It is responsible for 487 incidents, or 50% of the incidents in the Business sector and 254,612,930 records. The second largest subsector within Business is Retail, which was responsible for 241 incidents and 21% of the records disclosed. The third highest Business subsector for incidents was the Industrial category with 126 incidents, but only 4% of the records disclosed. In contrast, the Media category was the third largest records disclosed accounting for 16% of the records, despite having only 2% of the incidents. The Media category has an average known incident records disclosure rate of 3.8 million, while the overall rate for the Business subsector is only 700,635. It should be noted that the Business subsector has a high rate of incidents where the records are listed as "unknown". Almost half of the incidents (47%) did not provide an estimated record loss figure. The average figure for all incidents in the study was 34%, so this subsector has a considerably higher incidence of uncertainty. The Financial category was closer to the average at 36% unknown, while the Media category should have more accurate figures with only 25% not reporting loss figures. In contrast, the Retail category did not quantify the number of records lost in 84% of incidents.


Business SubSector Incident Detail

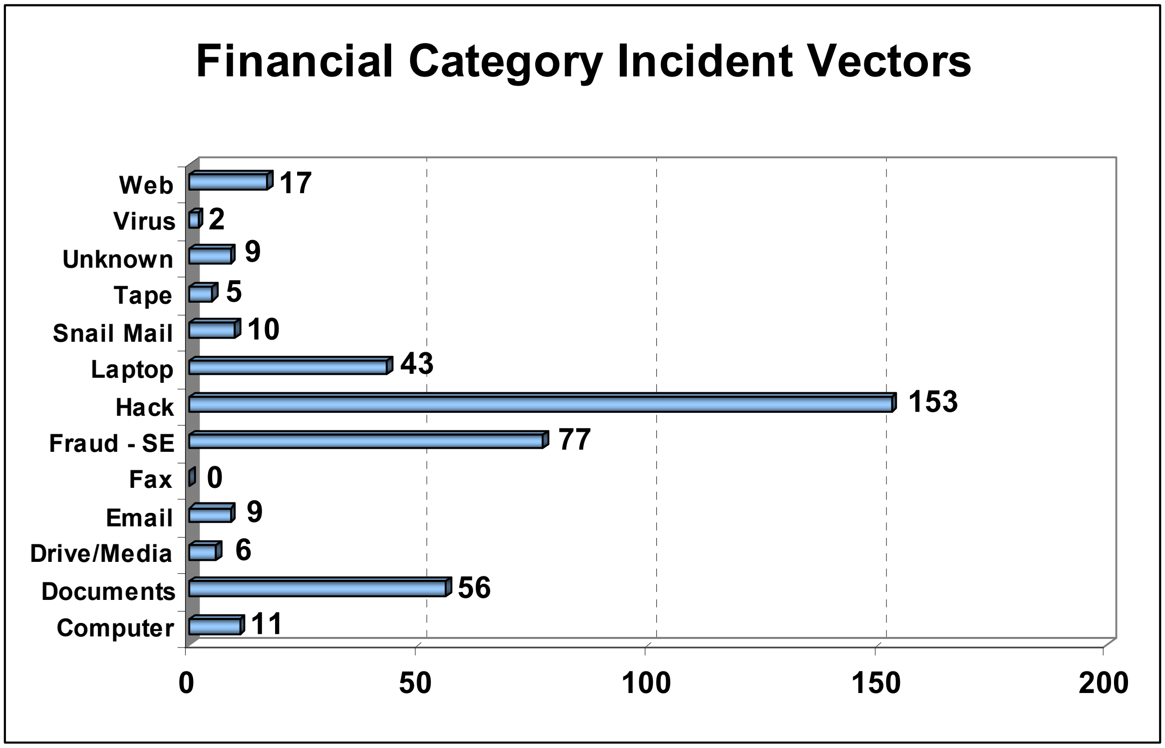

Business SubSector Records Detail

The Business Subsector Incident Detail graph focuses on the top three subsectors of the Business category for incidents. It shows a clear leadership in the Financial subsector, which is consistent with the high value of the data these companies maintain. Industry and Retail round out the leaders for number of incidents.

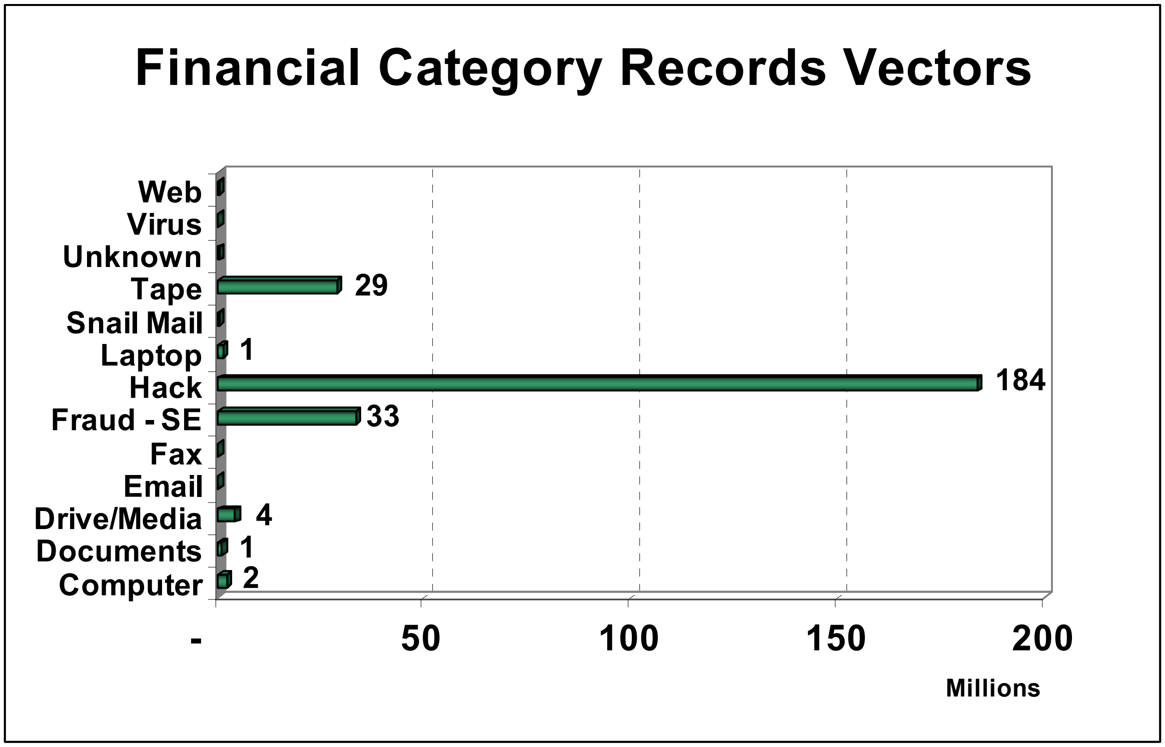
Breaking the Business subsectors down even further provides detail on the Financial category for both incidents and records disclosed over the course of the study. The Financial category actually had been on the decline for two years for records lost before a slight increase in 2008 and a sharp rise in 2009. The Business Subsector Records Detail graph shows the trend over time. The Media category actually had yearly totals of less than 1 million records disclosed for all but 2008, when Facebook's breach occurred. Without that breach, it would have had fewer than 120,000 records disclosed for the year. The spikes in Retail and Financial have similar causes—the TJX breach with 95 million records and the Heartland Payment Systems incident with 130 million records respectively.

Between these two graphs, it is interesting to see that both Financial and Retail are in the top three for records and incidents. The Industrial category, which was in third place for incidents, was not even close to the top for records with only 21 million disclosed. The Financial sector, in contrast, accounted for 254.6 million records.

Exploring the Financial category more closely, the Financial Category Incident Vectors graph illustrates how the breaches have occurred in this subset of the data. The Hack vector is almost twice as large as it's nearest sized vector (Fraud-SE), and the Documents vector surpasses the Laptop vector for the third highest by incident. This is in contrast to how strong a lead the laptop vector had for incidents in the overall data set.

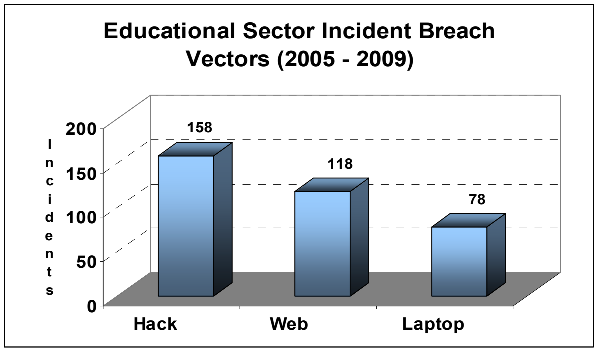


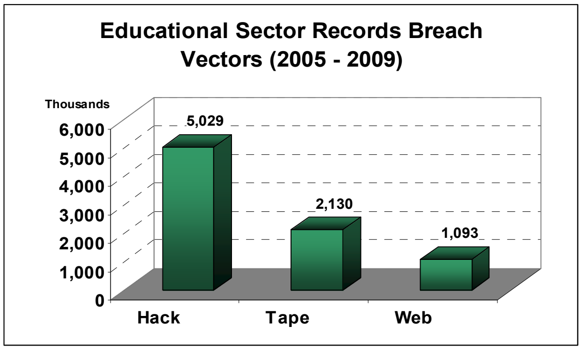
While the Hack vector is the clear leader in incidents, the difference between the Hack vector and all others is most glaring when looking at the records disclosed within the Financial category. As a vector, Hack accounts for 76% of the records in the Financial category.



The Educational Sector

Within the Educational category, Universities were the largest subsector, accounting for 463 of 549 incidents. This subsector was responsible for 9.8 million records, or 95% of the records disclosed for Education. As shown in the Educational Sector Incident Breach Vectors graph, the top three vectors were Hack, Web and Laptop. Hack was the leading vector in both records and incidents. As with many of the other sectors, Laptop is in the top three in incidents, but does not make the top vector for records disclosed.



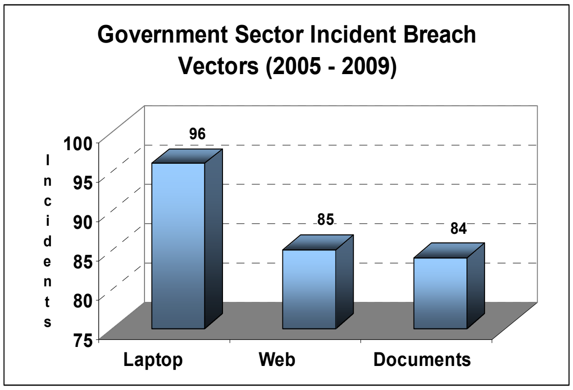


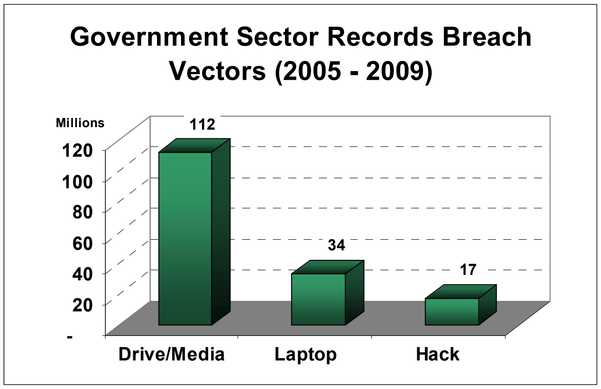
The top three vectors were Hack with 48% of the records, Tape with 20% of the records, and Web with 11%. The records disclosed in the Educational Sector (10.4 million) accounted for only 1% of the total records exposed in the study.

The Government Sector

Within the Government category, State government agencies were responsible for 230 of the incidents and 31.3 million records. The Federal government was responsible for 149 incidents and 146.9 million records. Clearly, the average loss per incident for the Federal government is significantly higher than for State. In fact, the mean records for the Federal breaches was 986,157 while the mean for State government was 136,091. The median figures are much lower, at 4,180 for Federal and 3,000 for State.

The following graphs illustrate the top three vectors for incidents and records in the Government sector. For incidents, there was a near-tie for the second place position between Web and Documents, with just one incident difference. In fourth place was the Drive/Media vector, which made the list of highest records disclosed vectors.

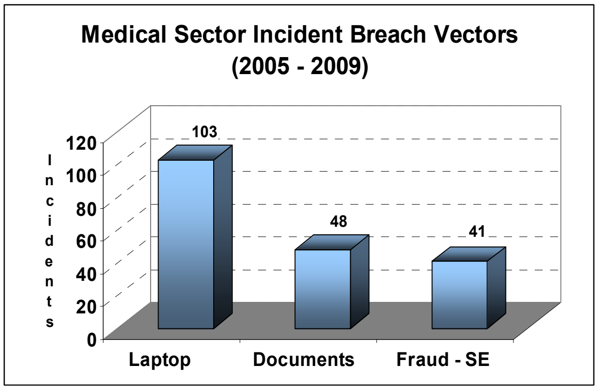


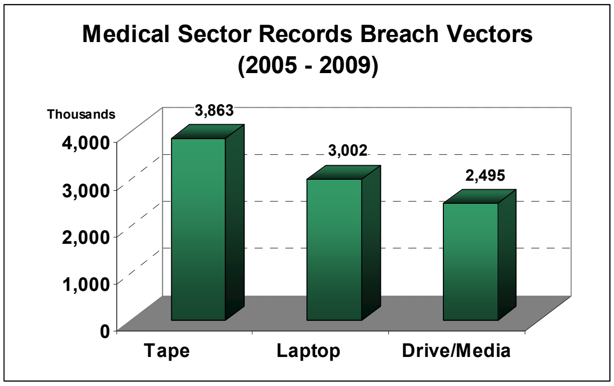


Additionally, as you can see above, Drive/Media was the leading vector for records disclosed by a wide margin—nearly triple its closest neighbor. With 191.4 million, the Government sector was responsible for 27% of the records disclosed in the study.

The Medical Sector

While this is the sector most commonly responsible for the disclosure of medical information, the most common data element exposed was SSNs (64% of cases). In 21% of cases in this sector, the combination of Name, Address, Social Security Number and medical information were simultaneously disclosed.





Finally, within the Medical category, 139 hospitals (the largest subsector) were responsible for losing 4.6 million records. Even with 12.8 million total records, the Medical sector was responsible for only 2% of the records disclosed in the study.

Data Types

The most common data types to be mentioned in the data breach laws are first and last name when combined with social security number. This is common across the majority of laws currently on the books requiring data breach disclosure. They vary in the other types that trigger a duty to report, and the most common ones are:

First (name or initial) and last name with:

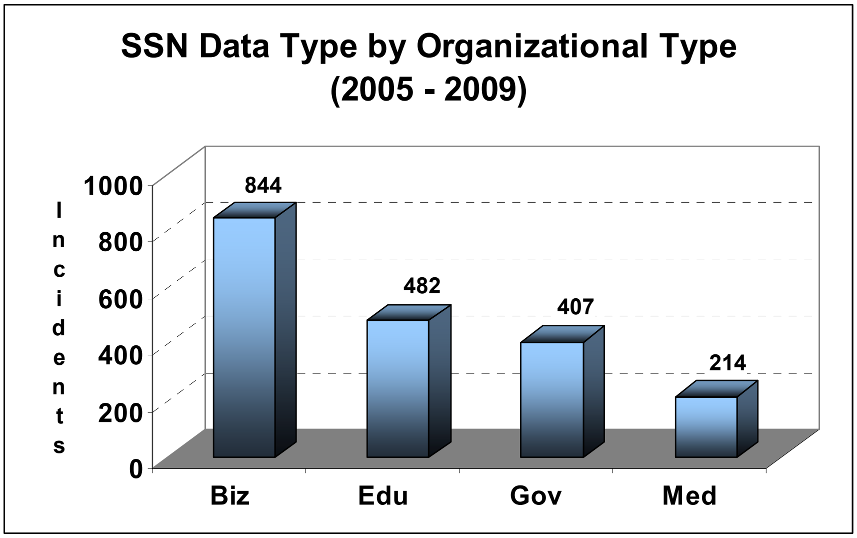
- Social Security Number (SSN)
- Driver's license or state identification number
- Credit card number or account information
- Other financial account information
- Date of Birth
- Medical information

Other items that can cause the media to take notice include email addresses (as in the case of Facebook), and miscellaneous data such as a private telephone number if a celebrity is involved (such as with Paris Hilton).

In many of these incidents, multiple data elements are lost at the same time. As mentioned earlier, since there are doubtless people who have had breaches expose their data numerous times, it is likely that different data elements are exposed with each breach. An area of future work would be to try and establish contact with people who have received these breach notifications and determine the rate of malicious use of the information—particularly to determine if the risk increases as the number of breaches involving one person's data increases. This data is particularly difficult to get to, however, since organizations would be reluctant to allow contact with individuals affected. It would also be useful to study the same set of people to determine if their attitudes to data privacy or the disclosing companies had changed as a result of the breaches.

Social Security Numbers

Social Security Numbers are highly valuable in conjunction with a person’s name. Without the name, generating SSNs is trivial, but knowing who they belong to is what makes them useful for identity theft and financial fraud. Some organizations continue to use them as a unique identifier for their records despite the risk. This is a data element that should be stored only when it must be present for the function of the data—such as for tax purposes. Other unique identifiers should be developed in place of using the SSN where it is not absolutely necessary. Data masking should be employed where possible as an extra precaution as well.

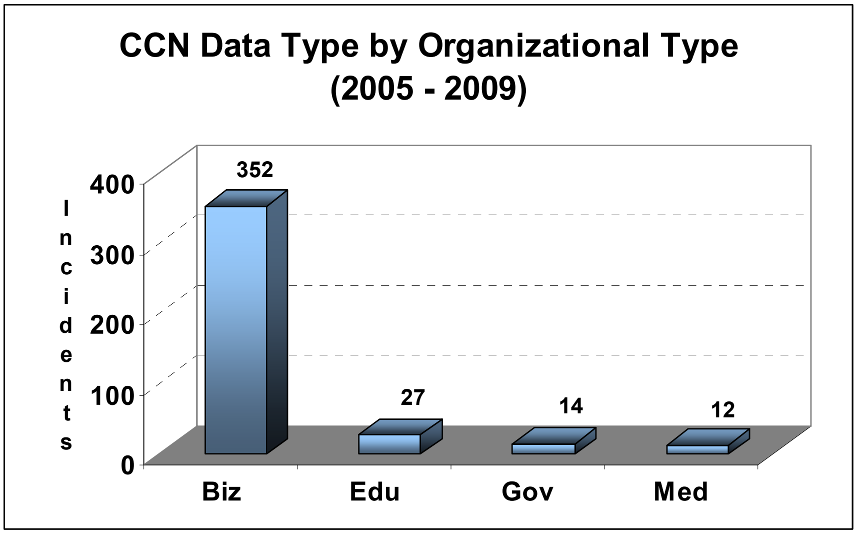


In 69% (1,947) of the cases, names and Social Security Number were disclosed. Customer data was disclosed in 40% of the cases—followed by Employees at 29%, Students at 18%, and Patients at 11%. All told, SSNs were the data element disclosed in 253.6 million records, or 35% of the records.

As shown here, the Business sector is the leader in losing Social Security Numbers by a significant margin. The Educational and Government sectors are not far apart, with the Medical sector trailing behind considerably.

Credit Cards

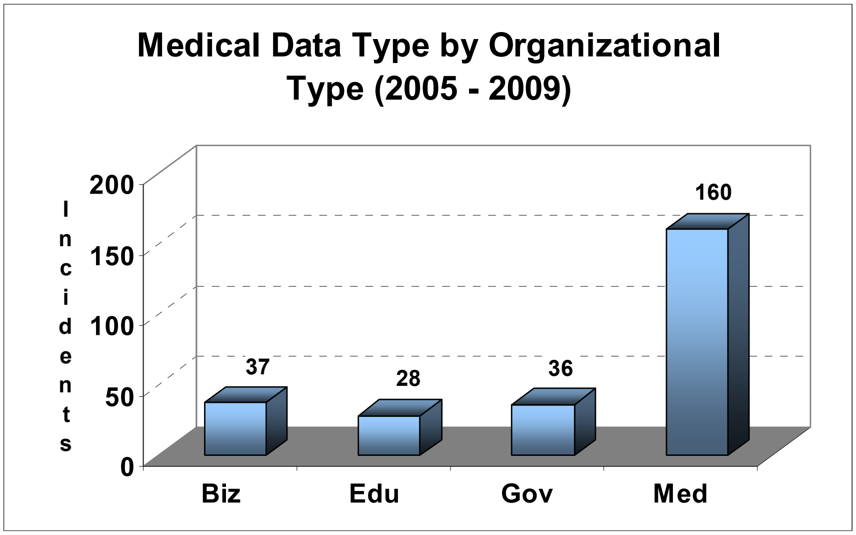
Credit card data has obvious financial fraud value, and continues to be targeted by malicious actors. The data is typically covered by the Payment Card Industry Data Security Standard (PCI DSS) issued by the major card brands. As shown below, the Business sector is the leader for incidents of disclosure.



Yet, even with the obvious value, it was the data item disclosed in only 14% of the incidents. That low number is deceiving, however, given the impact of the 328.1 million records disclosed. This one data element accounted for 45% of the records in the study. Clearly, this is a desirable target for data thieves. If we added the SSN and credit card records together, it would account for 80% of the records. However, in 72 of the cases, both SSN and credit card data was disclosed along with Name and Address. This accounted for 3.5 million records, which is included under both totals. So to get the combined total, we need to add the two together and subtract out 3.5 million, for a combined total of 578.2 million.

Medical Information

In 37 cases, the Business sector lost medical information about data subjects. These are data elements that would likely be covered under the Health Insurance Portability and Accountability Act of 1996 (HIPAA). In 11 of the cases, the data belonged to customers; in 6 of the incidents, the data belonged to employees; and for 17 instances, the data belonged to patients. The known number of records disclosed between these 37 incidents was 112,259, but 24 of the cases did not include a finite number for disclosed records.



In 28 cases, the Educational sector lost medical information about data subjects. The incidents affected employees in one case; patients in 17 cases; and students in 10 cases. The known number of records disclosed between these 28 incidents was 1,097,612, with only 5 of the cases reporting "unknown" disclosure figures.

In 36 cases, the Government sector disclosed medical information about data subjects. In 12 cases, customer data was disclosed; in just one case, employee data was disclosed (unfortunately it was accompanied by Social Security, Name and Address, Date of Birth and other data elements); in 22 cases, patient data was revealed and in one case, student medical data was compromised. The known number of disclosed records in the Government medical data cases was 711,564 records; with 13 incidents listing "unknown" disclosure totals.

Unsurprisingly, the Medical sector has the highest number of incidents disclosing medical data at 160. Since this sector is where the highest percentage of medical data resides, it makes sense that this is the type of data at risk. Of the incidents, in 19 cases, the data was about customers; in 7 the data was about employees; in 134, the data was about patients. The known records disclosed figure was 4,222,044 records; with 49 not reporting finite disclosure figures.

The ID Theft Critical Data Elements

In 168 cases where the Business sector lost customer data, the data lost included the identity theft critical data elements—those pieces of information that best facilitate identity theft when obtained together—namely, Name and Address when combined with other elements such as Social Security Number and Date of Birth. While Credit Card information can compromise just one account, these critical data elements can allow a criminal to compromise multiple accounts and even the victim's bank accounts and ability to obtain medical care. There have been cases where criminals have enough data to create new identification documents, and when arrested by Law Enforcement, have the charges count against the data subject victim [8].

> Surprisingly, in the case of criminal identity theft, in many cases, the victim’s identity can never truly be completely restored to pre-victim status. Many victims carry around proof that they were a victim of criminal identity theft for the rest of their lives [8].

Yet, in only 40 of those cases, credit monitoring was offered by the business to the data subjects. In fact, even in 5 of the 11 cases where confirmed instances of subsequent criminal use of the data were evident in the incident reports, the business that disclosed the data did not offer monitoring. In 58 cases, the Business sector lost the same three data elements of employee data, yet offered credit monitoring services in only 36 of those cases to their employees. Only three of those cases listed subsequent criminal use of the data, and none of those employees were confirmed to have been offered monitoring services.

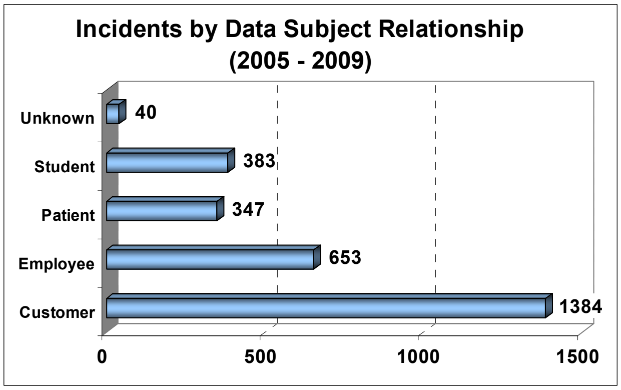
In the Educational sector, there were 39 cases where disclosed data included the three critical elements were listed. One case had confirmed criminal usage and none were confirmed to have been offered monitoring. There were 16 cases of student data with the same critical trio, with none of the cases having either confirmed usage or credit monitoring offered.

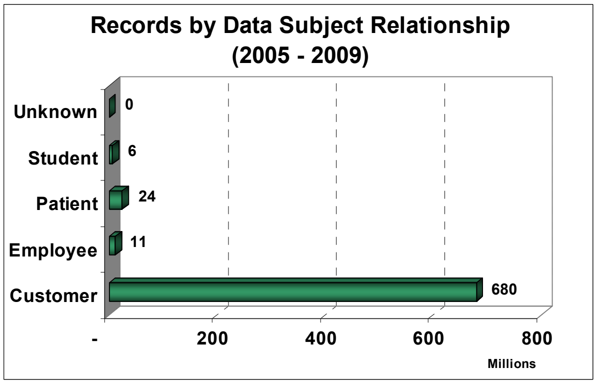
For the Government sector, 26 cases included the critical elements in customer data, with three cases having confirmed usage, none with monitoring. In fact, of the 520 incidents in the Government sector, only 13 cases showed credit monitoring being offered regardless of the data type lost. In eight of the cases, the critical trio of employee data was lost, and three of those cases indicated criminal usage of the data was present.

For the Medical sector, 29 cases included the critical trio, with 5 of them showing criminal usage of the data. Of the cases listed, only one confirmed credit monitoring was offered, and it was not one of the confirmed criminal use cases.

Relationships

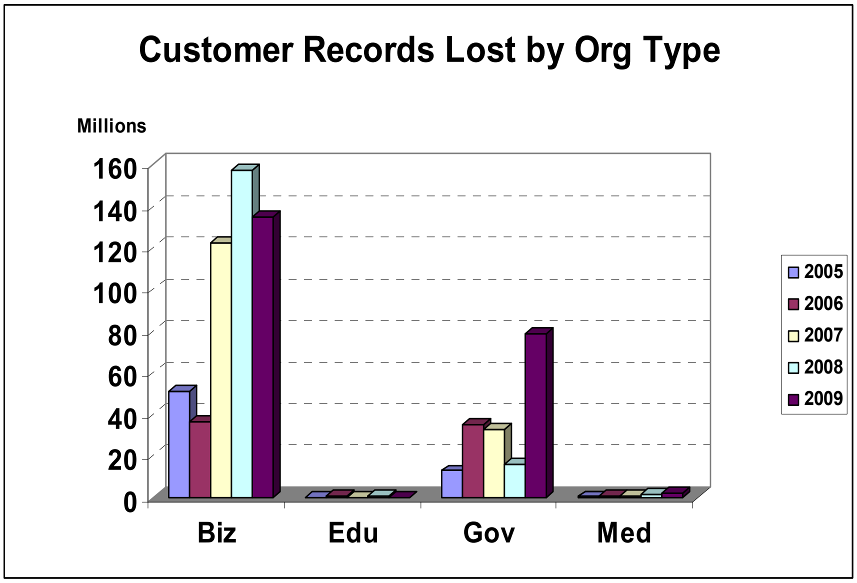
The next area of interest was the relationship between the organizations who have lost data and the people that data was about—the data subjects. Each incident was examined to determine whose data was affected in the breach. The subjects were identified as customers, employees, patients, students and unknown for those cases where the relationship was not clear. The Incidents by Data Subject Relationship graph breaks out whose data is being disclosed for the five years of the study.





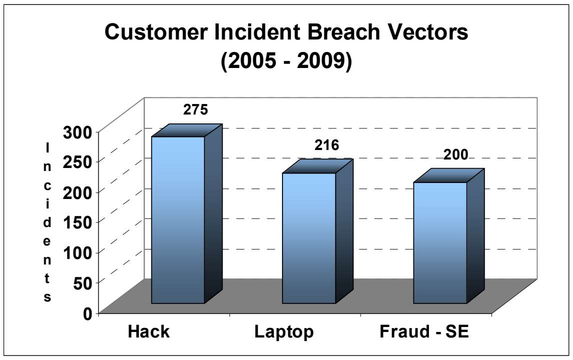
As evident above, the vast majority of the organizations are losing their customer's data. To look closer at the customer data, the information was broken up into records disclosed by sector.

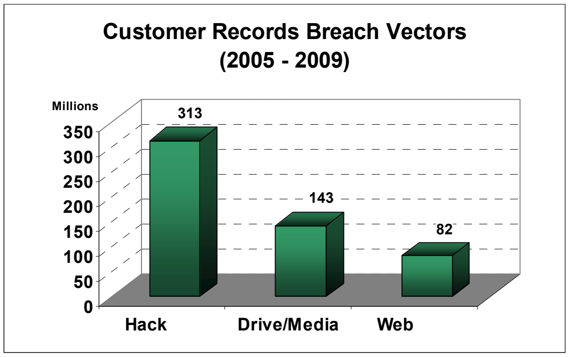
Customers



The Business sector incidents lost customer data 71% of the time. The Customer Records Lost by Org Type graph illustrates the trend in loss of Customer data over the course of the study. Business easily took the lead in records disclosed, while Government came in second. For incidents, when the relationship was customer, the data element lost was the SSN in 36% of the cases. The data element lost was a credit card in 25% of the incidents.

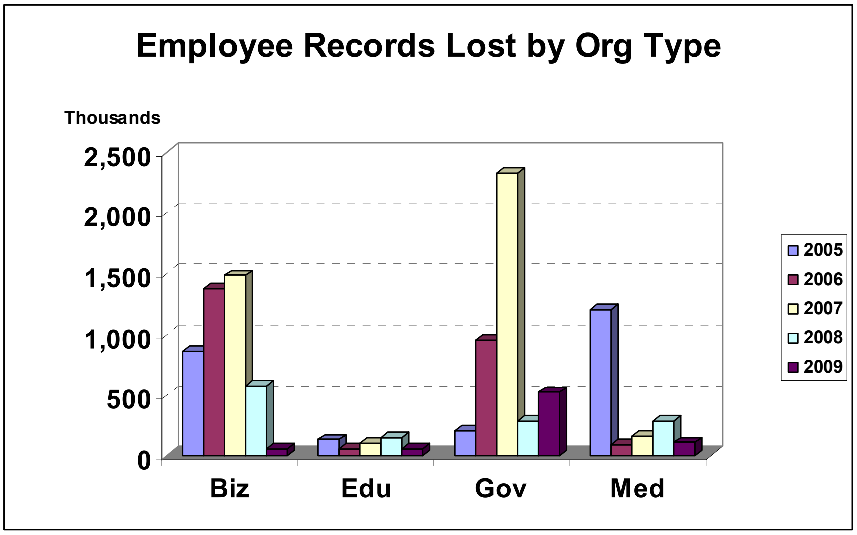
Looking at the vectors for this relationship, the incident and record leader was the Hack vector. The top three vectors in for this relationship are surprisingly close for incidents. For records disclosed, however, there is a wide difference between the three vectors, as shown below.





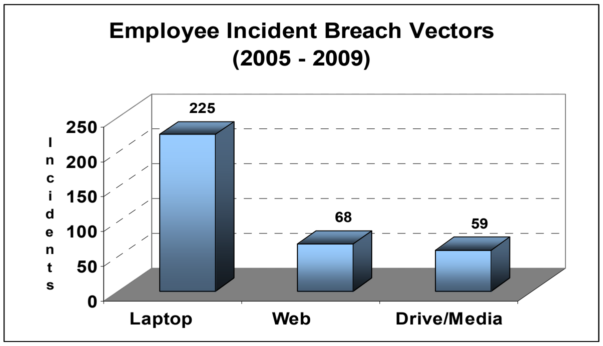
In 51 cases, the data elements disclosed included Name/Address, SSN and credit card data. Overwhelmingly, the data element is SSN with Name/Address (57% of incidents). In only 27% of the cases was the credit card the data element disclosed.

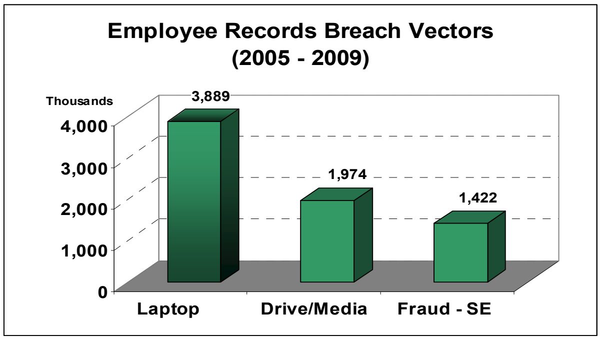
Employees



When the relationship was between an organization and its employees, the Business sector lost their employee's data 25% of the time. The Employee Records Lost by Org Type graph below shows the number of records where the relationship between the disclosing organization and the data subject was of employee/employer.

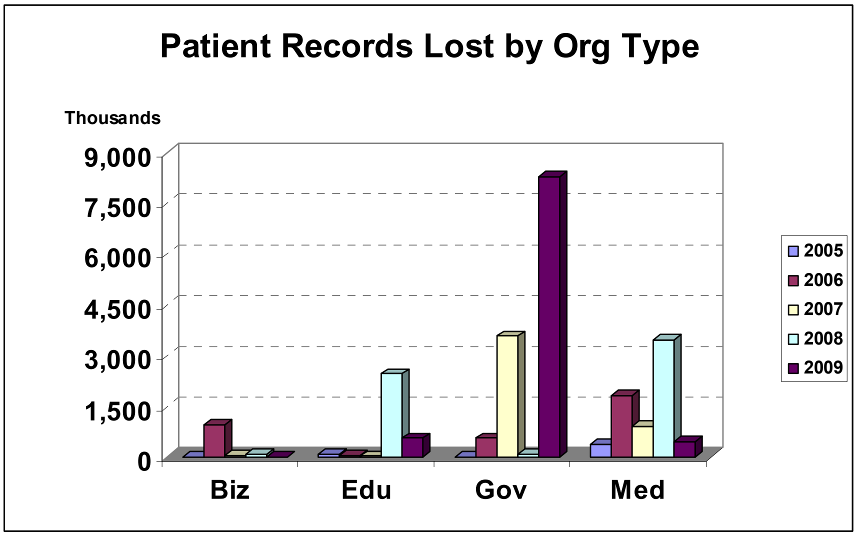
The top three incident vectors for the Employee relationship were Laptop, with a very strong lead; Web and Drive/Media. This is a rare case of the Laptop vector leading both incidents and records disclosed. The Laptop vector accounted for 34% of the employee data breach incidents, and 35% of the employee records disclosed over the course of the study (3.8 million).





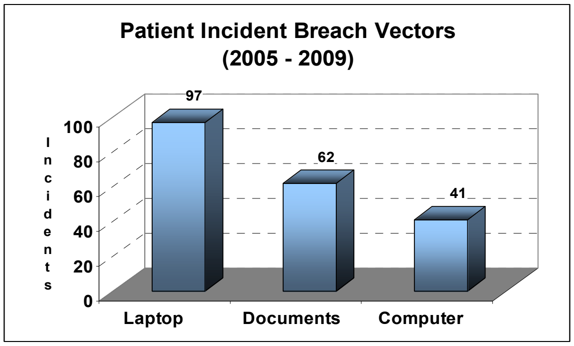
For 570 incidents (87%), the data element disclosed was the SSN along with Name/Address. Medical data was rarely disclosed (only 8 incidents), as were credit cards (11 incidents).

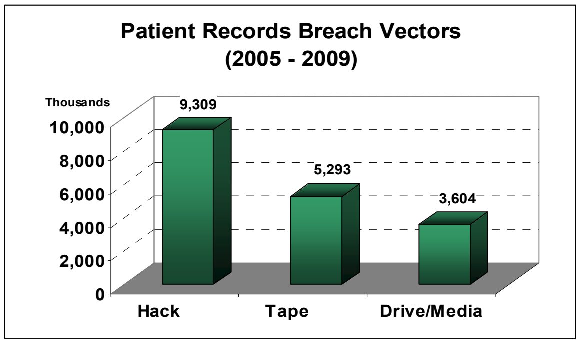
Patients



The Patient Records Lost by Org Type graph shows the records lost over the course of the study. The large total for Government in 2009 was due primarily to the Virginia Department of Health Professions breach responsible for 8.2 million records. Strangely, despite the current laws requiring notification, in 2008, 61% of those experiencing fraud related to medical information disclosure found out about it by having a billing company contact them about paying for the services rendered in their names [8].

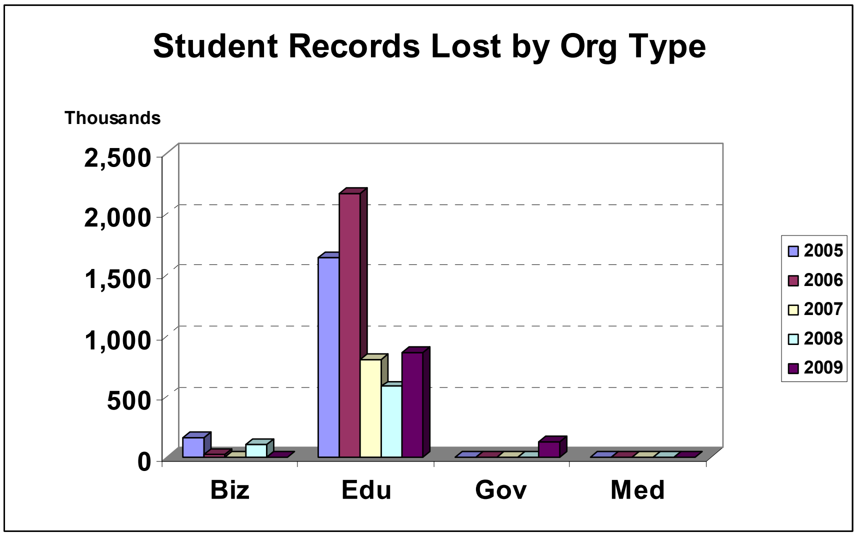
In 216 of the incidents, the data element disclosed in addition to name and address was SSN. In 41% of those 216 incidents, it was medical data being disclosed along with name, address and SSN.





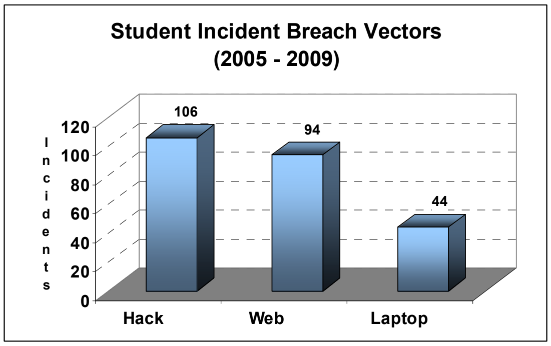
In looking at the incidents where Patient's data was disclosed, Laptop was the leading vector. In the vectors for number of records disclosed, however, Hack was the leader again with 9.3 million.

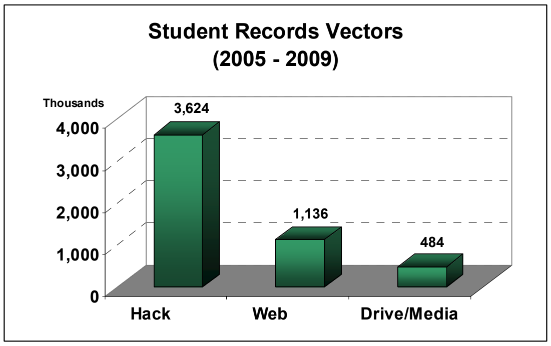
Students



The Educational sector, unsurprisingly, lost student data 66% and employee data 17% of the time. The Government sector lost customer data 56% and employee data 32% of the time. The Medical sector lost patient data 68% of the time.

When student's data is lost, the lead three vectors for incidents are Hack, Web and Laptop as shown below. Hack also is the leader for number of records disclosed for this relationship at 56% of the student records disclosed. In 90% of the incidents, it was a combination of SSN and Name/Address that was disclosed.

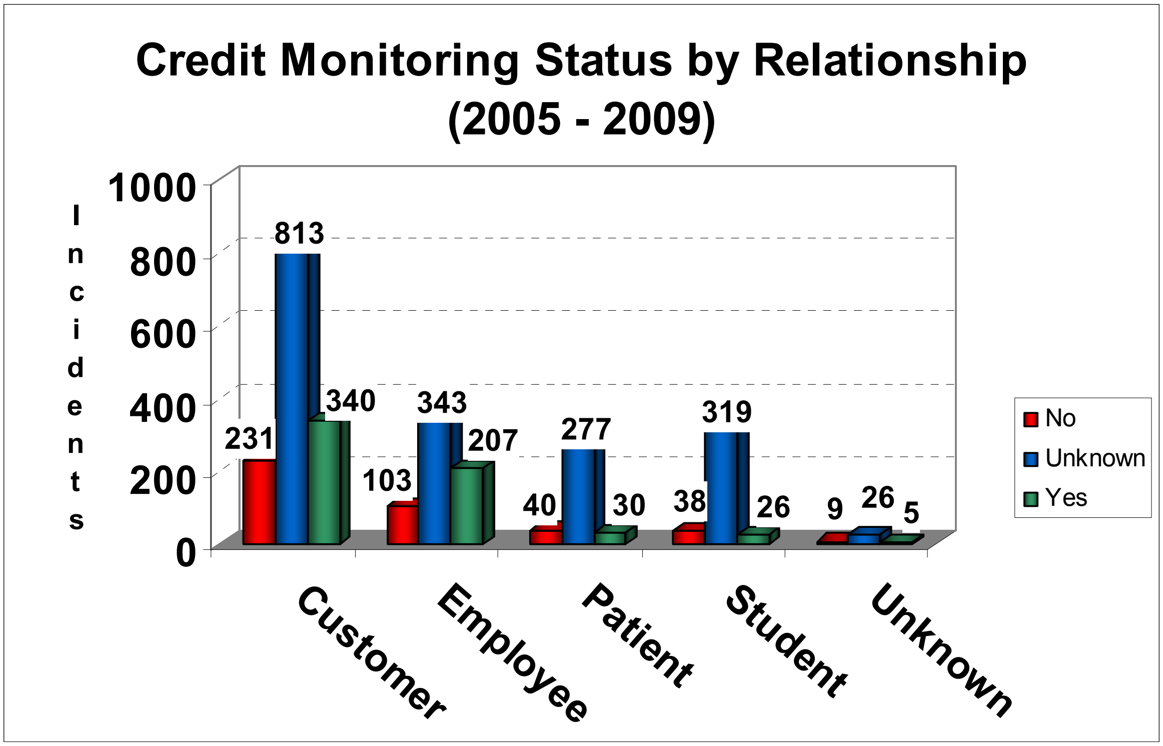




Credit Monitoring

Another aspect of interest, having defined whose data is being lost by relationship is to see how the victim data subjects are being treated—are they being offered credit monitoring to help minimize the impact of the disclosure? Earlier, in the Criminal Use section, we explored how these people were treated when there was subsequent fraudulent use of the data that was disclosed. The Credit Monitoring Status by Relationship graph shows which relationship was most likely to see some sort of monitoring offered.

As you can see from the graph below, the largest figure in each category is Unknown. These represent cases where the incident reports did not indicate whether this type of monitoring was offered. As you can see, however, for those incidents where the value is known, the customer relationship has the highest likelihood of being offered monitoring. These totals to not differentiate between what data type was disclosed.



Cost

The first lens to view costs of a data breach through is typically the impact on the disclosing organization. This is the cost that has had the most research, looking at the impact on stock prices [4, 5, 17, 18, 19]. According to the most recent Ponemon Institute study estimating the costs associated with data breaches, the cost of breaches has increased each year. Since there is no figure out yet for 2009, an estimate was made using the cost figure from 2008 for both the 2008 and 2009 cost per record figure. While holding the cost flat for 2009 over 2008 will likely produce a low estimate, it is the best figure available until the new data is released. Table T-12 shows an estimate of the cost based on the incidents tracked in this study and the costs computed in the Ponemon study [19].

Table T-12: Estimated Cost of Data Breaches/Year

Year	Records Disclosed	Cost Per Record	Total Breach Cost
2005	68,555,563	$138.00	$9,460,667,694
2006	80,363,058	$182.00	$14,626,076,556
2007	164,749,413	$197.00	$32,455,634,361
2008	182,414,761	$202.00	$36,847,781,722
2009	225,847,364	$202.00*	$45,621,167,528
Estimated Total	**721,930,159**		**$139,011,327,861**

*Cost figure from 2008.

The final estimate of the cost of the breaches reported in this study is over $139 billion over the past five years. If we add in the estimated records from our previous calculations, the cost rises to over $140 billion.

Table T-13: Estimated Cost of Data Breaches/Year with Median Estimated Records

Year	Records Disclosed	Cost Per Record	Total Breach Cost
2005	72,190,513	$138.00	9,962,290,794
2006	81,317,672	$182.00	14,799,816,304
2007	166,132,094	$197.00	32,728,022,518
2008	183,110,643	$202.00	36,988,349,886
2009	226,778,005	$202.00*	45,809,157,010
Estimated Total	**729,528,927**		**$140,287,636,512**

*Cost figure from 2008.

This does not include the costs suffered by the others who are affected by the breaches—the consumers whose data was disclosed, and the other companies who must incur costs related to breaches in their downstream or upstream channels.

This brings up another lens to view the costs of a data breach—that of down and upstream companies who are affected by a breach they did not cause. For instance, the case of the breach of the payment processor Heartland Payment Systems is an excellent example of these risks. The number of banks that were forced to reissue credit and debit cards is over 300, and some of those institutions have sued to recover some of their costs [13]. Most of the work that has been done in this area, deals with specific case studies on the larger incidents.

Finally, the impact to the data subject cannot be discounted in the cost of the breach. This cost is typically measured in less easily captured metrics—those of hours of productivity lost to trying to undo the damage, or monitor the situation for new activity.

> Victims reported spending an average of 68 hours repairing the damage done by identity theft to an *existing account* used or taken over by the thief. In cases where a *new account, criminal, governmental or a combination of several situations* were involved, respondents reported an average of 141 hours to clean up the fraud [8].

The costs associated with the data subject victims are difficult to correlate with the findings of this study, simply because it is unknown how many of these incidents result in direct use of the information. This type of data is captured in other studies, but without that direct tie, applying the cost figure would not be accurate.

CONCLUSION AND RECOMMENDATIONS

While it is not the intent of this study to provide recommendations to mitigate every type of disclosure vector, some recommendations were made throughout the document. They are summarized here:

- Organizations should ensure that their data lifecycle is managed end-to-end whether the data is on paper or in electronic form. They should also have controls to ensure that when the data is most vulnerable—on a laptop or other portable device—that it is further protected from the loss of the electronic storage mechanism.
- Organizations should have awareness programs that cover laptop physical security measures to discourage the practice of storing these items in vehicles unattended, of waiting to put them out of sight until after the person has reached their interim destination, and of leaving them unsecured in an office location. Any measures that minimize the laptop's time spent unattended and unprotected should help reduce the incidents of loss.
- Organizations should not rely on the login password to safeguard their data once physical custody has been lost. Additional controls on data at rest on computers or media should be in place so that loss of the item does not unavoidably mean disclosure of the data.
- When considering on-boarding a third party partner, companies should do a thorough examination of their security processes and procedures, to ensure they do not increase the risk to the data. These requirements should be included as contract items and include penalties for failure to protect the data entrusted to them. Finally, contract terms must define the process and expectations for data lifecycle management; including destruction or return of data should the partnership be terminated.
- Organizations should include data breach contingency planning in their incident response and/or disaster recovery plans. Any planned activity should also be tested at the same time the other aspects of the plan are tested to identify weaknesses and gaps.
- Organizations should refrain from using SSNs as unique identifiers. They should also take steps to make sure that sensitive data is only available to those who have a relevant business need to view it. Controls designed to prevent storing sensitive data on portable devices should also be implemented.
- Organizations that develop their own code should implement code reviews for such common exploits as SQL injection and Buffer Overflows.
- Organizations with internet-facing systems should have them regularly scanned for sensitive data that has been placed there against policy.

Until there is a Federal law that requires reporting all incidents to one agency, we will continue to have only a partial view into the data breach crisis. The creation of one law and one tracking organization would simplify matters for both the companies doing business in the United States, and the researchers who are trying to develop better metrics. Having original documents accessible to researchers would allow for the mining of additional metrics from the source material. Standard requirements for what is reported in each case would also help ensure the metrics gathered are complete from each record.

REFERENCES

[1] Attrition.org (2008). DLDOS (Data Loss Database - Open Source). http://attrition.org/dataloss/dldos.html

[2] Baker, W., Hylender, C. & Valentine, J. (2008) 2008 Data Breach Investigations Report. Verizon Business RISK Team. Verizon Business.

[3] Baker, W., Hutton, A., Hylender, C., Novak, C., Porter, C., Sartin, B., Tippet, P. & Valentine, J. (2009) 2009 Data Breach Investigations Report. Verizon Business RISK Team. Verizon Business.

[4] Campbell, K., Gordon, L., Loeb, M. and Zhou, L. (2003). The economic cost of publicly announced information security breaches: empirical evidence from the stock market. Journal of Computer Security. Vol. 11, Number 2003. pp. 431-448.

[5] Bania, P. (2008). Kon-Boot Ultimate Linux and Windows Hacking Utility. www.piotrbania.com.

[6] Cavusoglu, H., Mishra, B. and Raghunathan, S. (2004). The effect of Internet security breach announcements on market value: capital market reactions for breached firms and Internet security developers." International Journal of Electronic Commerce. Vol. 9, Number 1. 2004, pp. 69-104.

[7] Federal Trade Commission. (2007). FTC Facts For Consumers: Keeping Laptops from Getting Lost or Stolen. Federal Trade Commission, Bureau of Consumer Protection, Division of Consumer and Business Education.

[8] Foley, L. & Gordon, S. (2007). Identity Theft: The Aftermath 2007 Report. Identity Theft Resource Center. http://www.idtheftcenter.org/artman2/uploads/1/Aftermath_2007_20080529v2_1.pdf

[9] Foley, L, Barney, K. & Foley, J. (2009). Identity Theft: The Aftermath 2009. Identity Theft Resource Center. http://www.idtheftcenter.org/artman2/uploads/1/Aftermath_2009_20100520.pdf

[10] Hasan, R., & Yurcik, W. (2006). A Statistical Analysis of Disclosed Storage Security Breaches, Conference on Computer and Communications Security, Proceedings of the second ACM workshop on Storage security and survivability (pp. 1 - 8). Alexandria, Virginia, USA: ACM.

[11] Hoofnagle, C. (2008). Measuring Identity Theft at Top Banks (Version 1.0). Berkeley Center for Law and Technology. University of California, Berkeley.

[12] Identity Theft Resource Center. (2008). Identity Theft Resource Center 2008 Breach Report. http://idtheftmostwanted.org/ITRC%20Breach%20Report%202008.pdf

[13] Lee, S. (2006). Breach Notification Laws: Notification Requirements and Data Safeguarding Now Apply to Everyone, Including Entrepreneurs. Entrepreneurial Business Law Journal. Vol. 1:125.

[14] McGlasson, L. (2008). Heartland Data Breach Update: Now More Than 330 Institutions Impacted. BankInfoSecurity.com

[15] Moker, Kevin. (2008). Sound Assurance Incident Report Database. http://www.soundassurance.com/docs/incident_report_executive_summary_details.pdf

[16] Open Security Foundation (2008). DataLossDB. http://datalossdb.org/.

[17] Otto, P., Antón, A., Baumer, D. (2007). The ChoicePoint Dilemma. September/October 2007. Vol. 5, No. 5. pp. 15-23

[18] Ponemon Institute, LLC. (2007). 2007 Annual Study: U.S. Cost of a Data Breach. PGP Corporation and Vontu, Inc.

[19] Ponemon Institute, LLC. (2008). Airport Insecurity: The Case of Lost Laptops. Dell Corporation.

[20] Ponemon Institute, LLC. (2009). Fourth Annual U.S. Cost of a Data Breach Study. PGP Corporation.

[21] Privacy Rights Clearinghouse. (2008). A chronology of data breaches reported since the choicepoint incident (list). http://www.privacyrights.org/ar/ChronDataBreaches.htm.

[22] Romanosky, S., Telang, R., & Acquisti, A. (2008). Do Data Breach Disclosure Laws Reduce Identity Theft? Seventh Workshop on the Economics of Information Security, Center for Digital Strategies, Tuck School of Business, Dartmouth College, Hanover, NH.

[23] State of California Department of Consumer Affairs/Office of Privacy Protection. (2008). State Security Breach Notification Laws. National Conference of State Legislatures. http://www.ncsl.org/programs/lis/cip/priv/breachlaws.htm

[24] United States Census Population Clock. (2009). http://www.census.gov. Consulted 7/18/10.

APPENDIX A: DATA BREACH VECTOR DEFINITIONS

Computer: The Computer vector involves a non-laptop computer—frequently a desktop, but potentially a server or larger computer.

Documents: The Documents vector involves the loss, theft or inappropriate disposal of printed material.

Drive/Media: The Drive/Media vector involves portable hard drives, memory sticks, USB sticks, CD-ROMs and any other portable storage device.

Email: The Email vector involves data that is disclosed via email—whether to the wrong person, or other concerns.

Fax: The Fax vector involves the use of a fax machine to disclose information inappropriately.

Fraud - SE/Fraud-Social Engineering: The Fraud-SE vector involves malicious activities specifically designed to gain the attacker access to data via social engineering or other misrepresentation/pretexts.

Hack: The Hack vector involves attacking an organization's systems by exploiting vulnerabilities in the system's software, hardware or networking.

Laptop: The Laptop vector involves the loss, theft or disposal of portable computers.

Snail Mail: The Snail Mail vector involves the disclosure of information via the Postal Service or other courier.

Tape: The Tape vector involves the loss, theft or disposal of data stored on tape.

Unknown: The vector was not specified in the incident reports.

Virus: The Virus vector involves the disclosure of data as a result of a computer virus infection.

Web: The Web vector involves the disclosure of information by posting it on the web—either intentionally or accidentally.