DATA COLLECTION AND LEAKAGE

PHILIP HOWARD AND KRIS ERICKSON*

INTRODUCTION

Two questions have directed my research in the last few years. First, I have been exploring the role of consumer personally identifiable information in campaigns—where do political campaign managers get their data and how is it used? Second, in joint work with Kris Erikson, we have empirically assessed the extent of data breaches reported in the media during 1995–2005. Both of these questions involve issues of information security—how consumers can or cannot control access and use of their information.

I. THE MARKET FOR POLITICAL CONSUMER DATA

Discussion usually centers on the political implications of keeping private consumer data private. However, what has not been discussed equally is a much more explicitly political market in very political consumer data. I believe that there is a difference between what people say and what people do, and sometimes it is important to track both, and to consider both, of these patterns. We understand what people mean when they talk about privacy and what political campaign managers mean when they think they are talking about privacy. However, not all of these political campaign managers have exactly the same opinion. They have a diversity of opinion.

A number of professionals specialize in building various political organizations, such as websites for the NRA,[1] the Push Coalition,[2] and the NAACP.[3] Although one may assume that these groups would never have much to do with each other, they actually rely on the same small communi-

[*] Philip N. Howard is an associate professor in the Department of Communication at the University of Washington. He directs the National Science Foundation-funded World Information Access Project (www.wiaproject.org) and the Project on Information Technology and Political Islam (www.pitpi.org). His book, *The Digital Origins of Dictatorship and Democracy: Information Technology and Political Islam*, was recently published by Oxford University Press.

[1] National Rifle Association, http://www.nra.org (last visited Dec. 4, 2009).

[2] Rainbow Push Coalition, http://www.rainbowpush.org (last visited Dec. 4, 2009).

[3] NAACP, http://www.naacp.org (last visited Dec. 4, 2009).

ties of computer professionals for their data management. The quality of research is really interesting when you want to learn about the norms of a small community.

In traditional politics, individual policy preferences could only be meaningfully categorized by a small number of demographic variables. Those are the simple variables that political campaign managers use. But, now I can differentiate many gradations of variables and political culture.

So, there is a community of people who increasingly rely on political information technologies for attempting to reach voters. In 1960, it was a little over a dollar that presidential campaigns spent to get each person to actually show up on Election Day.[4] It is up to six dollars for the 2004 campaign.[5] Most of that goes toward television and other media. But, increasingly, campaigns' money allocations go to finding data on constituents to allow them to target voters more effectively.[6]

Politically minded individuals will probably have heard the terms "soft votes" and "hard votes." "Hard votes" describe a situation where a congressional leader promises his or her constituents something specific.[7] That is a hard vote to move. A "soft vote" is someone who has not expressed political opinions; that is, the voter is flexible in terms of committing to vote a certain way.[8] There are three consistent patterns to the decision making process for campaign managers who go after soft votes in particular. They decide to mine for data.[9] The goal of mining for data is not to figure out who is important in your district. It is actually about figuring who not to spend any time with. The major source of waste in a political campaign is to try to communicate with people you know are not going to vote for you. For example, African-Americans may get significantly less attention from political campaigns, because the assumption is that campaign managers do not need to spend much time trying to communicate with people who historically are statistically less likely to vote.[10]

A second set of technology issues that relate to political institutions are mistakes. Technology service providers make little red and blue maps

[4] PHILIP N. HOWARD, NEW MEDIA CAMPAIGNS AND THE MANAGED CITIZEN 146 (2006).

[5] *Id.* at 146.

[6] *Id.* at 147.

[7] For a discussion of hard votes, see, for example, Phuong Cat Le, *Powerful Voices Against Expanded Gambling*, SEATTLE POST-INTELLIGENCER, Mar. 1, 2003, at B1.

[8] HOWARD, *supra* note 4, at 90.

[9] For a discussion of data mining, see, for example, *id.* at 76–77.

[10] *See* Mark Hugo Lopez, Emily Kirby & Jared Sagoff, *The Youth Vote 2004*, CIRCLE (2005), http://www.civicyouth.org/PopUps/FactSheets/FS_Youth_Voting_72-04.pdf (discussing lower African American voter turnout).

of which states are going which way, and those maps are also used by the newsrooms. It is interesting to see that mistakes of these maps were eventually being loaded to the newsrooms and then loaded for distribution online. Counting and reporting mistakes can happen face-to-face in an election, but there are also transition mistakes by the polling stations trying to get data to election offices. Companies like Voter Data Services that are there to try to collect the sense of the public opinion on election day do it largely to generate news coverage, but they are also supposed to be a democratic double check on the elections process polling. Exit polling is very important and it is a process that is not free of mistakes.[11]

Finally, there are interpreted mistakes, and these are the kind of mistakes that we see watching the news, like television journalists giving data to viewers when they decide to call something. They wait before they call an election: What they are waiting for is for their producers to actually double check these other kinds of mistakes, and when they feel confident they accept the risk of being wrong. They go ahead and make their interpretation.

II. DATA BREACHES 1995–2005

Looking at a different category of Internet mistakes, in joint work with Kris Erikson, we conducted research on data breaches, performing a search of incidents of electronic data[12] loss reported in major U.S. news media from 1980–2006.[13] We also consulted lists of electronic data breaches compiled by third-party computer security advisories, such as the Identity Theft Resource Center (www.idtheftcenter.org) and Attrition.org. Our method yielded 589 incidents, 550 of which were successfully cross-checked with LexisNexis and Proquest to ensure accuracy and thirty-nine of which we discarded for involving citizens of other countries or for being unverifi-

[11] HOWARD, *supra* note 4, at 16.

[12] We defined electronic personal records as data containing privileged information about an individual that cannot be readily obtained through other public means. We define "personal data" to be information that should reasonably be known only to the individual concerned or be held by an organization under the terms of a confidentiality agreement. Electronic personal records therefore could include individuals' personal credit histories, banking information such as credit card numbers or account numbers, medical records, social security numbers, and grades earned at school.

[13] These included print publications with national circulation such as the *New York Times*, the *L.A. Times*, and *USA Today*, along with major broadcast news media. Because some news reports contained references to more than one incident, we employed a snowball methodology to expand our analysis by including additional security breaches mentioned in the same article. Duplicate entries were eliminated by comparing news stories on the basis of organizations involved, dates, and other incident details. In cases where papers reported different quantities of lost records, we chose the most conservative report.

able in major news media reports.[14] We are not venturing into the broader debate about the virtues and dangers of online anonymity;[15] we have chosen to focus only on data that are more sensitive than the information that we regularly volunteer in the course of surfing the web. We also focused only on incidents where compromised personal records were kept for a legitimate purpose by a company, institution, or government agency.[16] All of the incidents in our analysis deal with data that were maintained in electronic form, although in some cases compromised data were contained on a lost or stolen laptop computer.

III. THE SCALE OF LOST DATA

Between 1980 and 2006, some 1.9 billion records were reported compromised by government agencies, firms, hospitals, universities, and the military. This is the sum of compromised records from 529 cases in which some estimate of the volume of lost records was offered, though in 60 of these incidents the impact of the security breach was unknown. In a sense, this number of lost records is larger than we might expect because a few landmark incidents account for large portions of the total number of records compromised. On the other hand, the number of confirmed incidents—550 in all—may seem smaller than expected given the twenty-six year time frame of our search. Some articles report multiple incidents, and, of course,

[14] Our list of reported incidents is limited to cases where one or more electronic personal records were compromised through mishandling or theft. There are interesting advantages and disadvantages to using printed news sources to construct the history of computer hacking and breached private records. As stated above, the mainstream media often equate hackers with any crime involving a computer and use the misnomer "hacker" without a nuanced understanding of the history of more legitimate computer hacking. We use the term in this analysis because it is the most commonly used term in media reports where an intruder was deemed responsible for compromised data. While criminal records would certainly provide details about the prevalence of malicious intrusions, such records are extremely difficult to collect nationwide. Moreover, a survey of incidents composed through criminal records would significantly over-sample incidents where an individual hacker was at fault, and significantly undersample incidents where an organization was culpable but not deemed criminally negligent. Over the decade, journalists would not have discovered all incidents. However, journalists do their best to report the facts, and in the absence of a public agency that might maintain comprehensive incident records on privacy violations, news accounts provide a good accessible resource.

[15] For a discussion of anonymity online, see, for example, Konrad S. Lee, *Hiding from the Boss Online: The Anti-Employer Blogger's Legal Quest for Anonymity*, 23 SANTA CLARA COMPUTER & HIGH TECH. L.J. 135 (2006); Victoria Smith Ekstrand, *Unmasking Jane and John Doe: Online Anonymity and the First Amendment*, 8 COMM. L. & POL'Y 405 (2003).

[16] Consequently, "phishing," or spoofing scams where victims are deceived into volunteering their own personal information, are not included in our analysis. For a discussion of phishing, see, for exmaple, Rasha AlMahroos, *Phishing for the Answer: Recent Developments in Combating Phishing*, 3 ISJLP 595 (2008); Camille Calman, *Bigger Phish to Fry: California's Anti-Phishing Statute and its Potential Imposition of Secondary Liability on Internet Service Providers*, 13 RICH. J.L. & TECH. 1 (2006).

many incidents were covered by journalists on multiple occasions.

In 2004, the Census Bureau estimated that there were 217 million adults living in the United States.[17] We can conservatively estimate that for every U.S. adult, in the aggregate, at least nine private records have been compromised. Unfortunately, we cannot know how many of these compromised private records have actually been used for identity theft, or how many were sold to marketing companies. Identity theft can have a significant impact on an individual whose identity is stolen and can taint the reputation of the organization that was compromised. But in the incidents studied here, the security breach is often with commercial firms and, increasingly, educational institutions, rather than individuals.

A single incident, involving 1.6 billion compromised records at Acxiom, accounts for a large portion of the volume of records lost in the period from 2000–2006.[18] If this event is removed from this period, then thirty-two percent of the compromised volume and thirty percent of the incidents are related to hackers, forty-eight percent of the compromised volume and sixty-two percent of the incidents involve organizational behavior, and twenty percent of the compromised volume and eight percent of the incidents remain unattributed. If this event is removed from the volume of compromised records for the whole study period—between 1980 and 2006—then forty-five percent of the total volume of compromised records related to hackers, twenty-seven percent of the volume was attributed to the organization, and twenty-eight percent remained unattributed. If this event is removed from the total number of incidents for the whole study period, then thirty-one percent of the incidents involved hackers, sixty percent involved organizational management, and nine percent remain unattributed. Regardless of how the data is broken down, hackers never account for even half of the incidents or the volume of compromised records.

The majority of incidents involved commercial actors, less than a third of the incidents involved colleges and universities, and the remainder involved government, hospitals, and the military. When the exceptional loss

[17] U.S. CENSUS BUREAU, TABLE 1: ANNUAL ESTIMATES OF THE POPULATION FOR THE UNITED STATES AND STATES, AND FOR PUERTO RICO: APRIL 1, 2000 TO JULY 1, 2004 (2004), *available at* http://www.census.gov/popest/states/tables/NST-EST2004-01.pdf.

[18] For a discussion of the Acxiom data breach, see, for example, Caryn Rousseau, *Hacker Accesses Customer Information from Database Manager Acxiom*, SECURITYFOCUS, Aug. 7, 2003, http://www.securityfocus.com/news/6665; Jay Lyman, *Acxiom Database Hack Highlights Risk*, TECHNEWSWORLD, Aug. 11, 2003, http://www.technewsworld.com/story/31306.html; Laura Rohde, *Florida Hacker Indicted in Big Online Theft Case*, COMPUTERWORLD, July 22, 2004, http://computerworld.com/securitytopics/security/story/0,10801,94673,00.html.

of 1.6 billion personal records by Acxiom Corporation is removed,[19] the commercial sector still accounted for approximately 252 million individual compromised records, four times that of the next highest contributor, the government sector.[20] The education sector accounted for a small percentage of the overall quantity of lost records but accounted for thirty percent of all reported incidents, suggesting that educational institutions suffer from a higher rate of computer insecurity than might be anticipated. This could be explained by the fact that colleges and universities generally maintain large electronic databases on current and past students, staff, faculty, and alumni, and have an organizational culture geared towards information sharing. However, medical institutions—which presumably also maintain large quantities of electronic data—reported a significantly lower number of incidents of data loss.

In the early reports, most incidents were described as an unspecified breach or as the general result of hacker activity. However, for the period between 2000 and 2006, thirty-one percent of the incidents were about a breach caused by a hacker, eight percent of the incidents involve an unspecified breach, and sixty-one percent of the incidents involved different kinds of organizational culpability. For example, sometimes management accidentally exposed private records online, administrative error resulted in leaked data, or employees were caught using the data for activities not related to the work of the organization. On some occasions, staff simply misplaced backup tapes, while on others, computer equipment such as laptops were stolen.[21]

IV. CURRENT REGULATORY APPROACHES

Legislators at the federal and state levels have adopted two main strategies to address the problem of electronic record management. On one hand, they have directly targeted those individuals, computer hackers, whose actions potentially threaten the security of private electronic data through the Computer Fraud and Abuse Act (CFAA).[22] The second strate-

[19] *See* sources cited *supra* note 18.

[20] For a more complete exposition of the findings, see Kris Erickson & Philip N. Howard, *A Case of Mistaken Identity? News Accounts of Hacker, Consumer, and Organizational Responsibility for Compromised Digital Records*, 12 J. OF COMPUTER-MEDIATED COMM. 1229, 1237 (2007).

[21] *Id.* at 1239.

[22] 18 U.S.C. § 1030 (2006). For a discussion of the CFAA and criminal computer intrusion generally, see, for example, Susan W. Brenner, *Toward a Criminal Law for Cyberspace: Distributed Security*, 10 B.U. J. SCI. & TECH. L. 1 (2004); Orin S. Kerr, *Cybercrime's Scope: Interpreting "Access" and "Authorization" in Computer Misuse Statutes*, 78 N.Y.U. L. REV. 1596 (2003); *see also* Neal Kumar Katyal, *Digital Architecture as Crime Control*, 112 YALE L.J. 2261 (2003).

gy employed by regulators might be thought of as an indirect, or disciplinary, strategy through data breach notification legislation, which obliges institutions that manage electronic data to report any loss of that data to the individuals concerned.[23]

A. CFAA

The CFAA has been repeatedly strengthened in response to a perception that electronic data theft represents a material and growing concern.[24] The fact that punishments for computer intrusion now surpass those for many other more violent forms of crime suggests that federal legislators consider computer crime to constitute a serious threat to our personal and collective security. However, our data suggest that malicious intrusion by hackers makes up only a portion of all reported cases, while other factors, including poor management practices by organizations themselves, contribute more to the problem.

Surveying news reports of incidents of compromised personal records helps us to understand the diverse situations in which electronic personal records are stolen, lost, or mismanaged. More important, it allows us to separate incidents in which personal records have been compromised by outside hackers from incidents in which breaches were the result of an organizational lapse. Of course, we should expect organizations to perform due diligence and safeguard the digital records holding personal information from attack by malicious intruders. But often organizations are both the unwilling and unwitting victims of a malicious hacker. Through this study of reported incidents of compromised data, we found that two-fifths of the incidents over the last quarter century involve malicious hackers with criminal intent.

Surprisingly, however, the proportion of reported incidents involving hackers is smaller than the proportion of incidents involving organizational action or inaction. While thirty-one percent of the incidents reported clearly identify a hacker as the culprit, sixty percent of the incidents involve missing or stolen hardware, insider abuse or theft, administrative error, or accidentally exposing data online. The remainder of news stories record too little information about the breach to determine the cause—either organizations or individual hackers might be to blame for some of these incidents.

[23] For a discussion of state data breach notification statutes see, for example, Paul M. Schwartz & Edward J. Janger, *Notification of Data Security Breaches*, 105 MICH. L. REV. 913 (2007).

[24] For a discussion of CFAA evolution see, for example, Linda K. Stevens & Jesi J. Carlson, *The CFAA: New Remedies For Employee Computer Abuse*, 96 ILL. B.J. 144 (2008).

Organizations can probably be blamed for the management practices that result in administrative errors, lost backup tapes, or data exposed online. And even though an organization can be the victim of theft by its employees, we might still expect organizations to develop suitable safeguards to ensure the safety of client, customer, or member data. Even using the news media's expansive definition of hacker as a basis for coding stories, we find that a large portion of the security breaches in the United States are due to various forms of organizational malfeasance. One important outcome of the legislation is improved information about the types of security breaches. Many of the news stories between 1980 and 2004 report paltry details, with sources being off the record and vague estimates of the severity of the security breach. Since mandatory reporting legislation has been enacted in many states,[25] most news coverage provides more substantive details. In 2006, only ten of the 257 news stories were unable to make some attribution of responsibility for a security breach.

B. Data Breach Notification Legislation

While this directly addresses the problem of consumer protection by empowering individuals to protect themselves in case of lost or stolen data, it has probably been intended to produce secondary effects. Companies and institutions, wary of both the negative publicity and the financial costs generated by an incident of data loss, are encouraged to adopt more responsible network administration practices. Similarly, end-users are urged to weigh both the risk of doing business electronically and the costs associated with taking action once they are notified of a potential breach.

The differences we saw across data categories may be the result of strong privacy legislation in the arena of medical information,[26] compared to weak privacy legislation in the arena of educational and commercial information. The bulk of the reports occur in 2005 and 2006, after data breach legislation in California,[27] Washington,[28] and other states[29] took effect. There were three times as many incidents in the period between

[25] For a list of state data breach notification statutes, see National Conference of State Legislatures, State Security Breach Notification Laws, http://www.ncsl.org/programs/lis/cip/priv/breachlaws.htm (last visited Dec. 15, 2009).

[26] For a discussion of the Health Insurance Portability and Accountability Act, see, for example, Sharona Hoffman & Andy Podgurski, *Securing The HIPAA Security Rule*, J. INTERNET L., Feb. 2007, at 1.

[27] CAL. CIV. CODE §§ 56.06, 1785.11.2, 1798.29, 1798.82 (West 2008 and Supp. 2009).

[28] WASH. REV. CODE § 19.255.010 (West 2008).

[29] *See supra* note 25.

2005 and 2006 as there were in the previous twenty-five years.

Interestingly, the mandatory reporting legislation seems to have exposed educational institutions as a major source of leakage of private data. In total, thirty-eight percent of the incidents involved commercial firms, but specifically in 2005 and 2006, thirty-five percent of the incidents involved educational institutions. These kinds of organizations may have been the least equipped to protect the data of their students, staff, faculty, and alumni. For the majority of incidents, the news article reports some information about how the records were compromised. A closer reading of each of the incidents, however, reveals that most incidents involve different combinations of mismanagement, criminal intent, and, occasionally, bad luck. The hacker label is often used, even when the theft is perpetrated by an insider, such as a student or employee. Moreover, company public relations experts often posit that personal records were only *exposed*, not compromised, when employees post private records to a website or lose a laptop, and the company cannot be sure that anyone has taken specific advantage of the security breach.

First, it is noticeable that as more states require organizations to report compromised digital records, the overall volume of annual news stories on the topic has increased significantly. In fact, there were more reported incidents in 2005 and 2006 than in the previous twenty-five years combined. We found 126 incidents of compromised records between 1980 and 2004, and 424 incidents between 2005 and 2006. Just summing the incidents from 2005 and 2006, when mandatory reporting legislation was in place in many states, we find that sixty-eight percent of the stories concern data that were accidentally placed online or exposed through administrative errors, stolen equipment, or other security breaches such as employee loss of equipment or backup tapes.[30]

[30] Several factors might explain the pattern of increasing incidents and volume of compromised data over time. First, there is the possibility that the results are skewed due to the relative growth of new, fresh news stories devoted to this issue, and the loss of older stories that disappeared from news archives as time passed. Perhaps there have always been hundreds of incidents every year, but only in recent years has the severity of the problem been reported in the news. If this were the case, we would expect to see a gradually decaying pattern with greater number of reported cases in 2006 than in 2005, 2004, and so on. However, the dramatic difference in reported incidents between later years and early years suggests that this effect does not adequately explain our observations. A second possibility is that increased media attention or sensational reporting in 2005 and 2006 lead to a relative over-reporting of incidents, compared with previous years. While it is unlikely that media outlets have exaggerated the amount of electronic personal record loss, it is possible that in previous years a certain number of events went unreported in the media due to lack of awareness or interest in the issue of identity theft. A third possibility is that there were a greater number of reported incidents of data loss in 2005 and 2006, because institutions are maintaining and losing a larger quantity of electronic data, and because a changing legislative environment in many states is obliging institutions to report events publicly that

The Notification of Breach legislation that requires the prompt reporting of lost records in California came into effect in 2003; however, the legislation was not widely adopted and implemented by other states until 2005, which might help to explain the dramatic increase in reported cases. The Notification of Breach legislation in California, as many other states, requires notification when a state resident has been a victim of data loss, regardless of where the offending institution resides. Therefore, institutions located in states without Notification of Breach laws are still required to report cases to victims who live in states that have enacted this type of legislation, such as New York. The nature and complexity of many databases means that in many cases, compromised databases are likely to contain information about residents who are protected by notification of breach legislation, thus increasing the total number of reported cases.

V. FUTURE DIRECTIONS

Although computer hacking has been widely reframed as a criminal activity and has received increasingly harsh punishments, the legal response has obfuscated the responsibility of commercial, educational, government, medical, and military organizations for data security. The scale and scope of electronic record loss over the past decade would suggest that organizational self-regulation or self-monitoring is failing to keep our personal records secure and that the state has a more direct role to play in protecting personal information.

State-level initiatives have helped expose the problem by making it possible to collect better data on the types of security breaches that are occurring and to make some judgments about who is responsible for the breaches. If public policy can be used to create incentives for organizations to better manage personally identifiable information and punish organizations for mismanagement, such initiative would probably have to come at the state level. Electronically stored data might very well be weightless, but the organizations that retain personally identifiable information must shoulder more of the heavy burden for keeping such data secure.

This practice of using a risk/reward calculus to achieve policy objectives through legislation has been termed governing "in the shadow of the

may have gone unreported in previous years. The fourth possibility, and the most plausible one, is that mandatory reporting legislation has exposed both the severity of the problem and the common circumstances of organizational mismanagement. It is likely that a combination of factors explain our observations.

law" by some authors working in the critical legal studies[31] and governmentality literature.[32] One potential problem with this strategy is that the risks and rewards will be unequally distributed among various individual, state, and corporate actors. While a large corporation might possess the resources and technical skill necessary to encrypt data, secure networks, or hire external auditors, other institutions in the private or public sector may not find the risk of potential record loss worth the expenditure necessary to secure that data. Governing through this type of market discipline is likely to result in a wide spectrum of responses from differentially situated actors.

There are a number of alternatives open to lawmakers and policy advisors that could materially strengthen the security of electronic personal records in this country. Alternatives include setting stricter standards for information management, levying fines against institutions that violate information security standards, and mandating the encryption of all computerized personal data. However, the introduction of legislation to directly regulate institutions that handle electronic information would certainly be controversial. A wide variety of agencies, companies, and organizations manage personal records on a daily basis. This complexity would hinder the imposition of standardized practices such as encryption protocols. Corporations would probably balk at the prospect of having to pay fines or introduce expensive security measures, and accuse the government of heavy-handed interference. Others might argue that the imperatives of free-market capitalism demand that the government refrain from adopting punitive legislation, especially in order to maximize competitiveness.

[31] For a discussion of critical legal studies, see generally Adam Gearey, *Anxiety and Affirmation: Critical Legal Studies and the Critical "Tradition(s),"* 31 N.Y.U. REV. L. & SOC. CHANGE 585 (2007); E. Dana Neacsu, *CLS Stands for Critical Legal Studies, If Anyone Remembers*, 8 J.L. & POL'Y 415 (2000); John Henry Schlegel, *Alan and I: Of Community, Critical Legal Studies and All That*, 44 BUFF. L. REV. 636 (1996); Gregory G. Schultz, *Statutory Deconstruction: An Examination of Critical Legal Studies in Context*, 26 CUMB. L. REV. 459 (1995); Jason E. Whitehead, *From Criticism to Critique: Preserving the Radical Potential of Critical Legal Studies Through a Reexamination of Frankfurt School Critical Theory*, 26 FLA. ST. U. L. REV. 701 (1999).

[32] For a discussion of governmentality, *see* Jonathan Simon, *Driving Governmentality: Automobile Accidents, Insurance, and the Challenge to Social Order in the Inter-War Years, 1919–1941*, 4 CONN. INS. L.J. 521 (1997).