the digital person

Technology and Privacy in the Information Age

daniel j. solove


NEW YORK UNIVERSITY PRESS *New York and London*

2 The Rise of the Digital Dossier

We currently live in a world where extensive dossiers exist about each one of us. These dossiers are in digital format, stored in massive computer databases by a host of government agencies and private-sector companies. The problems caused by these developments are profound. But to understand the problems, we must first understand how they arose.

A History of Public-Sector Databases

Although personal records have been kept for centuries,[1] only in contemporary times has the practice become a serious concern. Prior to the nineteenth century, few public records were collected, and most of them were kept at a very local level, often by institutions associated with churches? The federal government's early endeavors at collecting data consisted mainly in conducting the census. The first census in 1790 asked only four questions.[3] With each proceeding census, the government gathered more personal information. By 1860, 142 questions were asked.[4] When the 1890 census included questions about diseases, disabilities, and finances, it sparked a public outcry, ultimately leading to the passage in the early twentieth century of stricter laws protecting the confidentiality of census data.[5]

Government information collection flourished during the middle of the twentieth century. The creation and growth of government bureaucracy—spawning well over 100 federal agencies within the past century—led to an insatiable thirst for information about individuals. One such agency was the Social Security Administration, created in 1935, which assigned nine-digit numbers to each citizen and required extensive record-keeping of people's earnings.

Technology was a primary factor in the rise of information collection. The 1880 census required almost 1,500 clerks to tally information tediously by hand—and it took seven years to complete.[6] At the rapid rate of population growth, if a faster way could not be found to tabulate the information, the 1890 census wouldn't be completed before the 1900 census began. Fortunately, just in time for the 1890 census, a census official named Herman Hollerith developed an innovative tabulating device—a machine that read holes punched in cards.[7] Hollerith's new machine helped tabulate the 1890 census in under three years.[8] Hollerith left the Census Bureau and founded a small firm that produced punch card machines—a firm that through a series of mergers eventually formed the company that became IBM.[9]

IBM's subsequent rise to prosperity was due, in significant part, to the government's increasing need for data. The Social Security System and other New Deal programs required a vast increase in records that had to be kept about individuals. As a result, the government became one of the largest purchasers of IBM's punch card machines.[10] The Social Security Administration kept most of its records on punch cards, and by 1943 it had more than 100 million cards in storage.[11]

The advent of the mainframe computer in 1946 revolutionized information collection. The computer and magnetic tape enabled the systematic storage of data. As processing speeds accelerated and as memory ballooned, computers provided a vastly increased ability to collect, search, analyze, and transfer records.

Federal and state agencies began to computerize their records. The Census Bureau was one of the earliest purchasers of commercially available computers.[12] Social Security numbers (SSNs)—originally not to be used as identifiers beyond the Social Security System—became immensely useful for computer databases.[13] This is because SSNs en-

able data to be easily linked to particular individuals. In the 1970s, federal, state, and local governments—as well as the private sector—increasingly began to use them for identification.[14]

Beginning in the 1960s, the growing computerization of records generated a substantial buzz about privacy. Privacy captured the attention of the public, and a number of philosophers, legal scholars, and other commentators turned their attention to the threats to privacy caused by the rise of the computer.[15] Congress began to debate how to respond to these emerging developments.[16] In 1973, the U.S. Department of Health, Education, and Welfare (HEW) issued a report entitled *Records, Computers, and the Rights of Citizens,* which trenchantly articulated the growing concerns over computerized record systems:

> There was a time when information about an individual tended to be elicited in face-to-face contacts involving personal trust and a certain symmetry, or balance, between giver and receiver. Nowadays, an individual must increasingly give information about himself to large and relatively faceless institutions, for handling and use by strangers—unknown, unseen, and, all too frequently, unresponsive. Sometimes the individual does not even know that an organization maintains a record about him. Often he may not see it, much less contest its accuracy, control its dissemination, or challenge its use by others.[17]

These problems continued to escalate throughout the ensuing decades. Computers grew vastly more powerful, and computerized records became ubiquitous. The rise of the Internet in the 1990s added new dimensions to these problems, sparking a revolution in the collection, accessibility, and communication of personal data.

Today, federal agencies and departments maintain almost 2,000 databases,[18] including records pertaining to immigration, bankruptcy, licensing, welfare, and countless other matters. In a recent effort to track down parents who fail to pay child support, the federal government has created a vast database consisting of information about all people who obtain a new job anywhere in the nation. The database contains their SSNs, addresses, and wages.[19]

States maintain public records of arrests, births, criminal proceedings, marriages, divorces, property ownership, voter registration, workers' compensation, and scores of other types of records. State licensing regimes mandate that records be kept on numerous professionals such as doctors, lawyers, engineers, insurance agents, nurses, police, accountants, and teachers.

A History of Private-Sector Databases

Although the government played an integral role in the development of massive dossiers of personal information, especially early on, businesses soon began to play an even greater role. While the public-sector story concerns the quest for regulatory efficiency, the private-sector story involves money and marketing.

Long before the rise of nationwide advertising campaigns there was a personal relationship between merchant and customer. Local merchants lived next door to their customers and learned about their lives from their existence together in the community. To a large extent, marketing was done locally—by the peddler on the street or the shopkeeper on the corner. Mass marketing, which began in the nineteenth century and flourished in the twentieth century, transformed the nature of selling from personal one-to-one persuasion to large-scale advertising campaigns designed for the nameless, faceless American consumer.

Mass marketing consumed vast fortunes, and only a small fraction of the millions of people exposed to the ads would buy the products or services. Soon marketers discovered the power of a new form of marketing—targeted marketing. The idea was to figure out which people were most likely to consume a product and focus the advertising on them.

In the 1920s, the sales department of General Motors Corporation began an early experiment with targeted marketing. GM discovered that owners of Ford vehicles frequently didn't purchase a Ford as their next vehicle—so it targeted owners of two-year-old Fords and sent them a brochure on GM vehicles.[20] GM then began to send out questionnaires asking for consumer input into their products. GM be-

lieved that this would be a good marketing device, presenting the image of a big corporation that cared enough to listen to the opinions of everyday people. GM cast itself as a democratic institution, its surveys stating that it was "OF THE PEOPLE, FOR THE PEOPLE, BY THE PEOPLE." One GM print advertisement depicted a delighted child holding up the survey letter exclaiming: "Look dad, a letter from General Motors!" The campaign was quite successful—ironically not because of the data collected but because of GM's image of appearing to be interested in the consumer's ideas.[21]

Today, corporations are desperate for whatever consumer information they can glean, and their quest for information is hardly perceived as democratic. The data collected extends beyond information about consumers' views of the product to information about the consumers themselves, often including lifestyle details and even a full-scale psychological profile.

The turn to targeting was spurred by the proliferation and specialization of mass media throughout the century, enabling marketers to tap into groups of consumers with similar interests and tastes. The most basic form of targeting involved selecting particular television programs, radio shows, or magazines in which to place advertisements. This technique, however, merely amounted to mass marketing on a slightly smaller scale.

The most revolutionary developments in targeted marketing occurred in the direct marketing industry. The practice of sending mail order catalogs directly to consumers began in the late nineteenth century when railroads extended the reach of the mail system.[22] The industry also reached out to people by way of door-to-door salespersons. In the 1970s, marketers began calling people directly on the telephone, and "telemarketing" was born.

Direct marketing remained a fledgling practice for most of the twentieth century. Direct marketers had long accepted the "2 percent" rule—only 2 percent of those contacted would respond.[23] With such a staggering failure rate, direct marketing achieved its successes at great cost. To increase the low response rate, marketers sought to sharpen their targeting techniques, which required more consumer research and an effective way to collect, store, and analyze information

about consumers. The advent of the computer database gave marketers this long sought-after ability—and it launched a revolution in targeting technology.

Databases provided an efficient way to store and search for data. Organized into fields of information, the database enabled marketers to sort by various types of information and to rank or select various groups of individuals from its master list of customers—a practice called "modeling." Through this process, fewer mailings or calls needed to be made, resulting in a higher response rate and lower costs. In addition to isolating a company's most profitable customers, marketers studied them, profiled them, and then used that profile to find similar customers.[24] This, of course, required not only information about existing customers, but the collection of data about prospective customers as well.

Originally, marketers sought to locate the best customers by identifying those customers who purchased items most recently and frequently and who spent the most money.[25] In the 1970s, marketers turned to demographic information.[26] Demographics included basic information such as age, income level, race, ethnicity, gender, and geographical location. Marketers could target certain demographic segments of the nation, a practice called "cluster marketing." This approach worked because people with similar incomes and races generally lived together in clusters.

The private sector obtained this demographic information from the federal government. In the 1970s, the United States began selling its census data on magnetic tapes. To protect privacy, the Census Bureau sold the information in clusters of 1,500 households, supplying only addresses—not names. But clever marketing companies such as Donnelley, Metromail, and R. L. Polk reattached the names by matching the addresses with information in telephone books and voter registration lists. Within five years of purchasing the census data, these companies had constructed demographically segmented databases of over half of the households in the nation.[27]

In the 1980s, marketers looked to supplement their data about consumers by compiling "psychographic" information—data about psychological characteristics such as opinions, attitudes, beliefs, and lifestyles.[28] For example, one company established an elaborate tax-

onomy of people, with category names such as "Blue Blood Estates," "BohemianMix," "YoungLiterati,""Shotgunsand Pickups," and "Hispanic Mix."[29] Each cluster had a description of the type of person, their likes, incomes, race and ethnicity, attitudes, and hobbies.[30]

These innovations made targeted marketing—or "database marketing" as it is often referred to today—the hottest form of marketing, growing at twice the rate of America's gross national product.[31] In 2001, direct marketing resulted in almost $2 trillion in sales.[32] On average, over 500 pieces of unsolicited advertisements, catalogs, and marketing mailings arrive every year at each household.[33] Due to targeting, direct mail yields $10 in sales for every $1 in cost—a ratio double that for a television advertisement—and forecasters predict catalog sales will grow faster than retail sales.[34] Telemarketing is a $662 billion a year industry.[35] In a 1996 Gallup poll, 77 percent of U.S. companies used some form of direct mail, targeted email, or telemarketing.[36]

The effectiveness of targeted marketing depends upon data, and the challenge is to obtain as much of it as possible. Marketers discovered that they didn't have to research and collect all the information from scratch, for data is the perspiration of the Information Age. Billions of bytes are released each second as we click, charge, and call. A treasure trove of information already lay untapped within existing databases, retail records, mailing lists, and government records. All that marketers had to do was plunder it as efficiently as possible.

The increasing thirst for personal information spawned the creation of a new industry: the database industry, an Information Age bazaar where personal data collections are bartered and sold. Marketers "rent"lists of names and personal information from database companies, which charge a few cents to a dollar for each name.[37] Over 550 companies compose the personal information industry, with annual revenues in the billions of dollars.[38] The sale of mailing lists alone (not including the sales generated by the use of the lists) generates $3 billion a year.[39] The average consumer is on around 100 mailing lists and is included in at least 50 databases.[40]

An increasing number of companies with databases—magazines, credit card companies, stores, mail order catalog firms, and even telephone companies—are realizing that their databases are becoming one of their most valuable assets and are beginning to sell their data.

A new breed of company is emerging that devotes its primary business to the collection of personal information. Based in Florida, Catalina Marketing Corporation maintains supermarket buying history databases on 30 million households from more than 5,000 stores.[41] This data contains a complete inventory of one's groceries, over-the-counter medications, hygiene supplies, and contraceptive devices, among others. Aristotle, Inc. markets a database of 150 million registered voters. Aristotle's database records voters' names, addresses, phone numbers, party affiliation, and voting frequency. Aristotle combines this data with about 25 other categories of information, such as one's race, income, and employer—even the make and model of one's car. It markets a list of wealthy campaign donors called "Fat Cat." Aristotle boasts: "Hit your opponent in the Wallet! Using Fat Cats, you can ferret out your adversary's contributors and slam them with a mail piece explaining why they shouldn't donate money to the other side."[42] Another company manufactures software called GeoVoter, which combines about 5,000 categories of information about a voter to calculate how that individual will

The most powerful database builders construct information empires, sometimes with information on more than half of the American population. For example, Donnelley Marketing Information Services of New Jersey keeps track of 125 million people. Wiland Services has constructed a database containing over 1,000 elements, from demographic information to behavioral data, on over 215 million people. There are around five database compilers that have data on almost all households in the United States.[44]

Beyond marketers, hundreds of companies keep data about us in their record systems. The complete benefits of the Information Age do not simply come to us—we must "plug in" to join in. In other words, we must establish relationships with Internet Service Providers, cable companies, phone companies, insurance companies, and so on. All of these companies maintain records about us. The Medical Information Bureau, a nonprofit institution, maintains a database of medical information on 15 million individuals, which is available to over 700 insurance companies.[45] Credit card companies have also developed extensive personal information databases. Un-

like cash, which often does not involve the creation of personally identifiable records, credit cards result in detailed electronic documentation of our purchases.[46]

Increasingly, we rely on various records and documents to assess financial reputation.[47] According to sociologist Steven Nock, this enables reputations to become portable.[48] In earlier times, a person's financial condition was generally known throughout the community. In modern society, however, people are highly mobile and creditors often lack first-hand experience of the financial condition and trustworthiness of individuals. Therefore, creditors rely upon credit reporting agencies to obtain information about a person's credit history. Credit reports reveal a person's consistency in paying back debts as well as the person's loan defaulting risk. People are assigned a credit score, which impacts whether they will be extended credit, and, if so, what rate of interest will be charged. Credit reports contain a detailed financial history, financial account information, outstanding debts, bankruptcy filings, judgments, liens, and mortgage foreclosures. Today, there are three major credit reporting agencies—Equifax, Experian, and Trans Union. Each agency has compiled extensive dossiers about almost every adult U.S. citizen.[49] Credit reports have become essential to securing a loan, obtaining a job, purchasing a home or a car, applying for a license, or even renting an apartment. Credit reporting agencies also prepare investigative consumer reports, which supplement the credit report with information about an individual's character and lifestyle.[50]

Launched in 2002, Regulatory DataCorp (RDC) has created a massive database to investigate people opening new bank accounts. RDC was created by many of the world's largest financial companies. Its database, named the Global Regulatory Information Database (GRID), gathers information from over 20,000 different sources around the world.[51] RDC's purpose is to help financial companies conduct background checks of potential customers for fraud, money laundering, terrorism, and other criminal activity. Although some people's information in the database may be incorrect, they lack the ability to correct the errors. RDC's CEO and president responds: "There are no guarantees. Is the public information wrong? We don't have enough information to say it's wrong."[52]

Cyberspace and Personal Information

Cyberspace is the new frontier for gathering personal information, and its power has only begun to be exploited. The Internet is rapidly becoming the hub of the personal information market, for it has made the peddling and purchasing of data much easier. Focus USA's website boasts that it has detailed information on 203 million people.[53] Among its over 100 targeted mailing lists are lists of "Affluent Hispanics," "Big-Spending Parents," "First Time Credit Card Holders," "Grown But Still At Home," "Hi-Tech Seniors," "New Homeowners," "Status Spenders," "Big Spending Vitamin Shoppers," and "Waist Watchers."[54] For example, Focus USA states for its list of "New Movers":

> As much as 20% of the population moves every year. . . . New movers have a lot of needs in their first few months. . . . During this lifestyle change period, new movers tend to be more receptive to direct mail and telemarketing offers for a wide variety of products.

The database contains data about age, gender, income, children, Internet connections, and more. There is a list devoted exclusively to "New Movers With Children," which includes data on the ages of the children. A list called "Savvy Single Women" states that "[s]ingle women represent a prime market for travel/vacation, frequent flyer clubs, credit cards, investing, dining out, entertainment, insurance, catalog shopping, and much more."

There's also a list of "Mr. Twenty Somethings" that contains mostly college-educated men who Focus USA believes are eager to spend money on electronic equipment. And there are lists of pet lovers, fitness-conscious people, cat and dog owners, motorcycle enthusiasts, casino gamblers, opportunity seekers, and sub-prime prospects.[55] Dunhill International also markets a variety of lists, including "America's Wealthiest Families," which includes 9.6 million records "appended with demographic and psychographic data."[56] There are also databases of disabled people, consumers who recently applied for a credit card, cruise ship passengers, teachers, and couples who just had a baby. Hippo Direct markets lists of people suffering from "med-

ical maladies" such as constipation, cancer, diabetes, heart disease, impotence, migraines, enlarged prostate, and more.[57] Another company markets a list of 5 million elderly incontinent women.[58] In addition to serving as a marketplace for personal information, cyberspace has provided a revolution for the targeted marketing industry because web pages are not static—they are generated every time the user clicks. Each page contains spaces reserved for advertisements, and specific advertisements are downloaded into those spots. The dynamic nature of web pages makes it possible for a page to download different advertisements for different users.

Targeting is very important for web advertising because a web page is cluttered with information and images all vying for the users' attention. Similar to the response rates of earlier efforts at direct marketing, only a small percentage of viewers (about 2 percent) click the advertisements they view.[59] The Internet's greater targeting potential and the fierce competition for the consumer's attention have given companies an unquenchable thirst for information about web users. This information is useful in developing more targeted advertising as well as in enabling companies to better assess the performance and popularity of various parts of their websites.

Currently, there are two basic ways that websites collect personal information. First, many websites directly solicit data from their users. Numerous websites require users to register and log in, and registration often involves answering a questionnaire. Online merchants amass data from their business transactions with consumers. For example, I shop on Amazon.com, which keeps track of my purchases in books, videos, music, and other items. I can view its records of every item I've ever ordered, and this goes back well over six years. When I click on this option, I get an alphabetized list of everything I bought and the date I bought it. Amazon.com uses its extensive records to recommend new books and videos. With a click, I can see dozens of books that Amazon.com thinks I'll be interested in. It is eerily good, and it can pick out books for me better than my relatives can. It has me pegged.

Websites can also secretly track a customer's websurfing. When a person explores a website, the website can record data about her ISP, computer hardware and software, the website she linked from, and

exactly what parts of the website she explored and for how long. This information is referred to as "clickstream data" because it is a trail of how a user navigates throughout the web by clicking on various links. It enables the website to calculate how many times it has been visited and what parts are most popular. With a way to connect this information to particular web users, marketers can open a window into people's minds. This is a unique vision, for while marketers can measure the size of audiences for other media such as television, radio, books, and magazines, they have little ability to measure attention span. Due to the interactive nature of the Internet, marketers can learn how we respond to what we hear and see. A website collects information about the way a user interacts with the site and stores the information in its database. This information will enable the website to learn about the interests of a user so it can better target advertisements to the user. For example, Amazon.com can keep track of every book or item that a customer browses but does not purchase.

To connect this information with particular users, a company can either require a user to log in or it can secretly tag a user to recognize her when she returns. This latter form of identification occurs through what is called a "cookie." A cookie is a small text file of codes that is deployed into the user's computer when she downloads a web page.[60] Websites place a unique identification code into the cookie, and the cookie is saved on the user's hard drive. When the user visits the site again, the site looks for its cookie, recognizes the user, and locates the information it collected about the user's previous surfing activity in its database. Basically, a cookie works as a form of high-tech cattle-branding.

Cookies have certain limits. First, they often are not tagged to particular individuals—just to particular computers. However, if the website requires a user to log in or asks for a name, then the cookies will often contain data identifying the individual. Second, typically, websites can only decipher the cookies that they placed on a user's computer; they cannot use cookies stored by a different website.

To get around these limitations, companies have devised strategies of information sharing with other websites. One of the most popular information sharing techniques is performed by a firm called DoubleClick. When a person visits a website, it often takes a quick detour

to Doubleclick. Doubleclick accesses its cookie on the person's computer and looks up its profile about the person. Based on the profile, Doubleclick determines what advertisements that person will be most responsive to, and these ads are then downloaded with the website the person is accessing. All this occurs in milliseconds, without the user's knowledge. Numerous websites subscribe to Doubleclick. This means that if I click on the same website as you at the very same time, we will receive different advertisements calculated by DoubleClick to match our interests. People may not know it, but DoubleClick cookies probably reside on their computer. As of the end of 1999, Doubleclick had amassed 80 million customer profiles.[61]

Another information collection device, known as a "web bug," is embedded into a web page or even an email message. The web bug is a hidden snippet of code that can gather data about a person.[62] For example, a company can send a spam email with a web bug that will report back when the message is opened. The bug can also record when the message is forwarded to others. Web bugs also can collect information about people as they explore a website. Some of the nastier versions of web bugs can even access a person's computer files.[63]

Companies also use what has become known as "spyware," which is software that is often deceptively and secretly installed into people's computers. Spyware can gather information about every move one makes when surfing the Internet. This data is then used by spyware companies to target pop-up ads and other forms of advertising.[64]

Legal scholar Julie Cohen has noted another growing threat to privacy—technologies of digital rights management (DRM), which are used by copyright holders to prevent piracy. Some DRM technologies gather information about individuals as they listen to music, watch videos, or read e-books. DRM technologies thus "create records of intellectual exploration, one of the most personal and private of activities."[65]

Copyright holders are also using computer programs called "bots" (shorthand for "robots"). Also known as "crawlers" or "spiders," bots can automatically prowl around the Internet looking for information. Industry trade groups, such as the Recording Industry Association of America (RIAA) and the Motion Picture Association of America

(MPAA), have unleashed tens of thousands of bots to identify potential illegal users of copyrighted materials.[66] Spammers—the senders of junk email—also employ a legion of bots to copy down email addresses that appear on the web in order to add them to spam lists. Bots also patrol Internet chat rooms, hunting for data.[67]

As we stand at the threshold of an age structured around information, we are only beginning to realize the extent to which our lives can be encompassed within its architecture. "The time will come," predicts one marketer, "whenwe are well known for our inclinations, our predilections, our proclivities, and our wants. We will be classified, profiled, categorized, and our every click will be watched."[68] As we live more of our lives on the Internet, we are creating a permanent record of unparalleled pervasiveness and depth. Indeed, almost everything on the Internet is being archived. One company has even been systematically sweeping up all the data from the Internet and storing it in a vast electronic warehouse.[69] Our online personas—captured, for instance, in our web pages and online postings—are swept up as well. We are accustomed to information on the web quickly flickering in and out of existence, presenting the illusion that it is ephemeral. But little on the Internet disappears or is forgotten, even when we delete or change the information. The amount of personal information archived will only escalate as our lives are increasingly digitized into the electric world of cyberspace.

These developments certainly suggest a threat to privacy, but what specifically is the problem? The way this question is answered has profound implications for the way the law will grapple with the problem in the future.