CHAPTER 2

Naked in the Sunlight

Privacy Lost, Privacy Abandoned

1984 Is Here, and We Like It

On July 7, 2005, London was shaken as suicide bombers detonated four explosions, three on subways and one on a double-decker bus. The attack on the transit system was carefully timed to occur at rush hour, maximizing its destructive impact. 52 people died and 700 more were injured.

Security in London had already been tight. The city was hosting the G8 Summit, and the trial of fundamentalist cleric Abu Hamza al-Masri had just begun. Hundreds of thousands of surveillance cameras hadn't deterred the terrorist act, but the perpetrators were caught on camera. Their pictures were sent around the world instantly. Working from 80,000 seized tapes, police were able to reconstruct a reconnaissance trip the bombers had made two weeks earlier.

George Orwell's *1984* was published in 1948. Over the subsequent years, the book became synonymous with a world of permanent surveillance, a society devoid of both privacy and freedom:

> ...there seemed to be no color in anything except the posters that were plastered everywhere. The black-mustachio'd face gazed down from every commanding corner. There was one on the house front immediately opposite. BIG BROTHER IS WATCHING YOU ...

The real 1984 came and went nearly a quarter century ago. Today, Big Brother's two-way telescreens would be amateurish toys. Orwell's imagined

London had cameras everywhere. His actual city now has at least half a million. Across the UK, there is one surveillance camera for every dozen people. The average Londoner is photographed hundreds of times a day by electronic eyes on the sides of buildings and on utility poles.

Yet there is much about the digital world that Orwell did not imagine. He did not anticipate that cameras are far from the most pervasive of today's tracking technologies. There are dozens of other kinds of data sources, and the data they produce is retained and analyzed. Cell phone companies know not only what numbers you call, but where you have carried your phone. Credit card companies know not only where you spent your money, but what you spent it on. Your friendly bank keeps electronic records of your transactions not only to keep your balance right, but because it has to tell the government if you make huge withdrawals. The digital explosion has scattered the bits of our lives everywhere: records of the clothes we wear, the soaps we wash with, the streets we walk, and the cars we drive and where we drive them. And although Orwell's Big Brother had his cameras, he didn't have search engines to piece the bits together, to find the needles in the haystacks. Wherever we go, we leave digital footprints, while computers of staggering capacity reconstruct our movements from the tracks. Computers re-assemble the clues to form a comprehensive image of who we are, what we do, where we are doing it, and whom we are discussing it with.

Perhaps none of this would have surprised Orwell. Had he known about electronic miniaturization, he might have guessed that we would develop an astonishing array of tracking technologies. Yet there is something more fundamental that distinguishes the world of *1984* from the actual world of today. We have fallen in love with this always-on world. We accept our loss of privacy in exchange for efficiency, convenience, and small price discounts. According to a 2007 Pew/Internet Project report, "60% of Internet users say they are not worried about how much information is available about them online." Many of us publish and broadcast the most intimate moments of our lives for all the world to see, even when no one requires or even asks us to do so. 55% of teenagers and 20% of adults have created profiles on social networking web sites. A third of the teens with profiles, and half the adults, place no restrictions on who can see them.

In Orwell's imagined London, only O'Brien and other members of the Inner Party could escape the gaze of the telescreen. For the rest, the constant gaze was a source of angst and anxiety. Today, we willingly accept the gaze. We either don't think about it, don't know about it, or feel helpless to avoid it except by becoming hermits. We may even judge its benefits to outweigh its risks. In Orwell's imagined London, like Stalin's actual Moscow, citizens spied on their fellow citizens. Today, we can all be Little Brothers, using our search

engines to check up on our children, our spouses, our neighbors, our colleagues, our enemies, and our friends. More than half of all adult Internet users have done exactly that.

The explosive growth in digital technologies has radically altered our expectations about what will be private and shifted our thinking about what *should* be private. Ironically, the notion of privacy has become fuzzier at the same time as the secrecy-enhancing technology of encryption has become widespread. Indeed, it is remarkable that we no longer blink at intrusions that a decade ago would have seemed shocking. Unlike the story of secrecy, there was no single technological event that caused the change, no privacy-shattering breakthrough—only a steady advance on several technological fronts that ultimately passed a tipping point.

PUBLIC ORGANIZATIONS INVOLVED IN DEFENDING PRIVACY

Existing organizations have focused on privacy issues in recent years, and new ones have sprung up. In the U.S., important forces are the American Civil Liberties Union (ACLU, `www.aclu.org`), the Electronic Privacy Information Center (`EPIC, epic.org`), the Center for Democracy and Technology (CDT, `www.cdt.org`), and the Electronic Frontier Foundation (`www.eff.org`).

Many devices got cheaper, better, and smaller. Once they became useful consumer goods, we stopped worrying about their uses as surveillance devices. For example, if the police were the only ones who had cameras in their cell phones, we would be alarmed. But as long as we have them too, so we can send our friends funny pictures from parties, we don't mind so much that others are taking pictures of us. The social evolution that was supported by consumer technologies in turn made us more accepting of new enabling technologies; the social and technological evolutions have proceeded hand in hand. Meanwhile, international terrorism has made the public in most democracies more sympathetic to intrusive measures intended to protect our security. With corporations trying to make money from us and the government trying to protect us, civil libertarians are a weak third voice when they warn that we may not want others to know so much about us.

So we tell the story of privacy in stages. First, we detail the enabling technologies, the devices and computational processes that have made it easy and convenient for us to lose our privacy—some of them familiar technologies, and some a bit more mysterious. We then turn to an analysis of how we have lost our privacy, or simply abandoned it. Many privacy-shattering things have happened to us, some with our cooperation and some not. As a result, the sense of personal privacy is very different today than it was two decades ago. Next, we discuss the social changes that have occurred—cultural shifts

that were facilitated by the technological diffusion, which in turn made new technologies easier to deploy. And finally we turn to the big question: What does privacy even mean in the digitally exploded world? Is there any hope of keeping anything private when everything is bits, and the bits are stored, copied, and moved around the world in an instant? And if we can't—or won't—keep our personal information to ourselves anymore, how can we make ourselves less vulnerable to the downsides of living in such an exposed world? Standing naked in the sunlight, is it still possible to protect ourselves against ills and evils from which our privacy used to protect us?

Footprints and Fingerprints

As we do our daily business and lead our private lives, we leave footprints and fingerprints. We can see our footprints in mud on the floor and in the sand and snow outdoors. We would not be surprised that anyone who went to the trouble to match our shoes to our footprints could determine, or guess, where we had been. Fingerprints are different. It doesn't even occur to us that we are leaving them as we open doors and drink out of tumblers. Those who have guilty consciences may think about fingerprints and worry about where they are leaving them, but the rest of us don't.

THE UNWANTED GAZE

The Unwanted Gaze by Jeffrey Rosen (Vintage, 2001) details many ways in which the legal system has contributed to our loss of privacy.

In the digital world, we all leave both electronic footprints and electronic fingerprints—data trails we leave intentionally, and data trails of which we are unaware or unconscious. The identifying data may be useful for forensic purposes. Because most of us don't consider ourselves criminals, however, we tend not to worry about that. What we don't think about is that the various small smudges we leave on the digital landscape may be useful to someone else—someone who wants to use the data we left behind to make money or to get something from us. It is therefore important to understand how and where we leave these digital footprints and fingerprints.

Smile While We Snap!

Big Brother had his legions of cameras, and the City of London has theirs today. But for sheer photographic pervasiveness, nothing beats the cameras in the cell phones in the hands of the world's teenagers. Consider the alleged misjudgment of Jeffrey Berman. In early December 2007, a man about

60 years old committed a series of assaults on the Boston public transit system, groping girls and exposing himself. After one of the assaults, a victim took out her cell phone. Click! Within hours, a good head shot was up on the Web and was shown on all the Boston area television stations. Within a day, Berman was under arrest and charged with several crimes. "Obviously we, from time to time, have plainclothes officers on the trolley, but that's a very difficult job to do," said the chief of the Transit Police. "The fact that this girl had the wherewithal to snap a picture to identify him was invaluable."

That is, it would seem, a story with a happy ending, for the victim at least. But the massive dissemination of cheap cameras coupled with universal access to the Web also enables a kind of vigilante justice—a ubiquitous Little-Brotherism, in which we can all be detectives, judges, and corrections officers. Mr. Berman claims he is innocent; perhaps the speed at which the teenager's snapshot was disseminated unfairly created a presumption of his guilt. Bloggers can bring global disgrace to ordinary citizens.

In June 2005, a woman allowed her dog to relieve himself on a Korean subway, and subsequently refused to clean up his mess, despite offers from others to help. The incident was captured by a fellow passenger and posted online. She soon became known as "gae-ttong-nyue" (Korean for "puppy poo girl"). She was identified along with her family, was shamed, and quit school. There is now a Wikipedia entry about the incident. Before the digital explosion—before bits made it possible to convey information instantaneously, everywhere—her actions would have been embarrassing and would have been known to those who were there at the time. It is unlikely that the story would have made it around the world, and that it would have achieved such notoriety and permanence.

There are many free webcam sites, at which you can watch what's happening right now at places all over the world. Here are a few:

```
www.camvista.com
www.earthcam.com
www.webcamworld.com
www.webworldcam.com
```

Still, in these cases, at least someone thought someone did something wrong. The camera just happened to be in the right hands at just the right moment. But looking at images on the Web is now a leisure activity that anyone can do at any time, anywhere in the world. Using Google Street View, you can sit in a café in Tajikistan and identify a car that was parked in my driveway when Google's camera came by (perhaps months ago). From Seoul, you can see what's happening right now, updated every few seconds, in Picadilly Circus or on the strip in Las Vegas. These views were always available to the public, but cameras plus the Web changed the meaning of "public."

And an electronic camera is not just a camera. *Harry Potter and the Deathly Hallows* is, as far as anyone knows, the last book in the Harry Potter series. Its arrival was eagerly awaited, with lines of anxious Harry fans stretching around the block at bookstores everywhere. One fan got a pre-release copy, painstakingly photographed every page, and posted the entire book online before the official release. A labor of love, no doubt, but a blatant copyright violation as well. He doubtless figured he was just posting the pixels, which could not be traced back to him. If that was his presumption, he was wrong. His digital fingerprints were all over the images.

Digital cameras encode metadata along with the image. This data, known as the Exchangeable Image File Format (EXIF), includes camera settings (shutter speed, aperture, compression, make, model, orientation), date and time, and, in the case of our Harry Potter fan, the make, model, and serial number of his camera (a Canon Rebel 350D, serial number 560151117). If he registered his camera, bought it with a credit card, or sent it in for service, his identity could be known as well.

Knowing Where You Are

Global Position Systems (GPSs) have improved the marital lives of countless males too stubborn to ask directions. Put a Garmin or a Tom Tom in a car, and it will listen to precisely timed signals from satellites reporting their positions in space. The GPS calculates its own location from the satellites' locations and the times their signals are received. The 24 satellites spinning 12,500 miles above the earth enable your car to locate itself within 25 feet, at a price that makes these systems popular birthday presents.

If you carry a GPS-enabled cell phone, your friends can find you, if that its what you want. If your GPS-enabled rental car has a radio transmitter, you can be found whether you want it or not. In 2004, Ron Lee rented a car from Payless in San Francisco. He headed east to Las Vegas, then back to Los Angeles, and finally home. He was expecting to pay $150 for his little vacation, but Payless made him pay more—$1,400, to be precise. Mr. Lee forgot to read the fine print in his rental contract. He had not gone too far; his contract was for unlimited mileage. He had missed the fine print that said, "Don't leave California." When he went out of state, the unlimited mileage clause was invalidated. The fine print said that Payless would charge him $1 per Nevada mile, and that is exactly what the company did. They knew where he was, every minute he was on the road.

A GPS will locate you anywhere on earth; that is why mountain climbers carry them. They will locate you not just on the map but in three dimensions, telling you how high up the mountain you are. But even an ordinary cell phone will serve as a rudimentary positioning system. If you are traveling in

settled territory—any place where you can get cell phone coverage—the signals from the cell phone towers can be used to locate you. That is how Tanya Rider was found (see Chapter 1 for details). The location is not as precise as that supplied by a GPS—only within ten city blocks or so—but the fact that it is possible at all means that photos can be stamped with identifying information about where they were shot, as well as when and with what camera.

Knowing Even Where Your Shoes Are

A Radio Frequency Identification tag—RFID, for short—can be read from a distance of a few feet. Radio Frequency Identification is like a more elaborate version of the familiar bar codes that identify products. Bar codes typically identify what kind of thing an item is—the make and model, as it were. Because RFID tags have the capacity for much larger numbers, they can provide a unique serial number for each item: not just "Coke, 12 oz. can" but "Coke can #12345123514002." And because RFID data is transferred by radio waves rather than visible light, the tags need not be visible to be read, and the sensor need not be visible to do the reading.

RFIDs are silicon chips, typically embedded in plastic. They can be used to tag almost anything (see Figure 2.1). "Prox cards," which you wave near a sensor to open a door, are RFID tags; a few bits of information identifying you are transmitted from the card to the sensor. Mobil's "Speedpass" is a little RFID on a keychain; wave it near a gas pump and the pump knows whom to charge for the gasoline. For a decade, cattle have had RFIDs implanted in their flesh, so individual animals can be tracked. Modern dairy farms log the milk production of individual cows, automatically relating the cow's identity to its daily milk output. Pets are commonly RFID-tagged so they can be reunited with their owners if the animals go missing for some reason. The possibility of tagging humans is obvious, and has been proposed for certain high-security applications, such as controlling access to nuclear plants.

But the interesting part of the RFID story is more mundane—putting tags in shoes, for example. RFID can be the basis for powerful inventory tracking systems.

RFID tags are simple devices. They store a few dozen bits of information, usually unique to a particular tag. Most are passive devices, with no batteries, and are quite small. The RFID includes a tiny electronic chip and a small coil, which acts as a two-way antenna. A weak

SPYCHIPS

This aptly named book by Katherine Albrecht and Liz McIntyre (Plume, 2006) includes many stories of actual and proposed RFID uses by consumer goods manufacturers and retailers.

current flows through the coil when the RFID passes through an electromagnetic field—for example, from a scanner in the frame of a store, under the carpet, or in someone's hand. This feeble current is just strong enough to power the chip and induce it to transmit the identifying information. Because RFIDs are tiny and require no connected power source, they are easily hidden. We see them often as labels affixed to products; the one in Figure 2.1 was between the pages of a book bought from a bookstore. They can be almost undetectable.

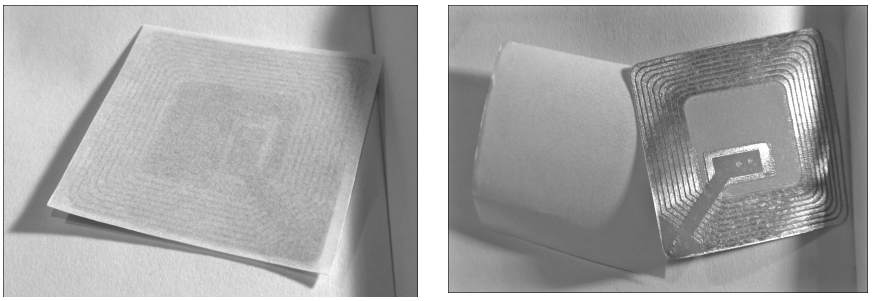


FIGURE 2.1 An RFID found between the pages of a book. A bookstore receiving a box of RFID-tagged books can check the incoming shipment against the order without opening the carton. If the books and shelves are scanned during stocking, the cash register can identify the section of the store from which each purchased copy was sold.

RFIDs are generally used to improve record-keeping, not for snooping. Manufacturers and merchants want to get more information, more reliably, so they naturally think of tagging merchandise. But only a little imagination is required to come up with some disturbing scenarios. Suppose, for example, that you buy a pair of red shoes at a chain store in New York City, and the shoes have an embedded RFID. If you pay with a credit card, the store knows your name, and a good deal more about you from your purchasing history. If you wear those shoes when you walk into a branch store in Los Angeles a month later, and that branch has an RFID reader under the rug at the entrance, the clerk could greet you by name. She might offer you a scarf to match the shoes—or to match anything else you bought recently from any other branch of the store. On the other hand, the store might know that you have a habit of returning almost everything you buy—in that case, you might find yourself having trouble finding anyone to wait on you!

The technology is there to do it. We know of no store that has gone quite this far, but in September 2007, the Galeria Kaufhof in Essen, Germany equipped the dressing rooms in the men's clothing department with RFID readers. When a customer tries on garments, a screen informs him of available sizes and colors. The system may be improved to offer suggestions about accessories. The store keeps track of what items are tried on together and what combinations turn into purchases. The store will remove the RFID tags from the clothes after they are purchased—if the customer asks; otherwise, they remain unobtrusively and could be scanned if the garment is returned to the store. Creative retailers everywhere dream of such ways to use devices to make money, to save money, and to give them small advantages over their competitors. Though Galeria Kaufhof is open about its high-tech men's department, the fear that customers won't like their clever ideas sometimes holds back retailers—and sometimes simply causes them to keep quiet about what they are doing.

Black Boxes Are Not Just for Airplanes Anymore

On April 12, 2007, John Corzine, Governor of New Jersey, was heading back to the governor's mansion in Princeton to mediate a discussion between Don Imus, the controversial radio personality, and the Rutgers University women's basketball team.

His driver, 34-year-old state trooper Robert Rasinski, headed north on the Garden State Parkway. He swerved to avoid another car and flipped the Governor's Chevy Suburban. Governor Corzine had not fastened his seatbelt, and broke 12 ribs, a femur, his collarbone, and his sternum. The details of exactly what happened were unclear. When questioned, Trooper Rasinski said he was not sure how fast they were going—but we *do* know. He was going 91 in a 65 mile per hour zone. There were no police with radar guns around; no human being tracked his speed. We know his exact speed at the moment of impact because his car, like 30 million cars in America, had a black box—an "event data recorder" (EDR) that captured every detail about what was going on just before the crash. An EDR is an automotive "black box" like the ones recovered from airplane crashes.

EDRs started appearing in cars around 1995. By federal law, they will be mandatory in the United States beginning in 2011. If you are driving a new GM, Ford, Isuzu, Mazda, Mitsubishi, or Subaru, your car has one—whether anyone told you that or not. So do about half of new Toyotas. Your insurance company is probably entitled to its data if you have an accident. Yet most people do not realize that they exist.

EDRs capture information about speed, braking time, turn signal status, seat belts: things needed for accident reconstruction, to establish responsibility, or to prove innocence. CSX Railroad was exonerated of all liability in the death of the occupants of a car when its EDR showed that the car was stopped on the train tracks when it was hit. Police generally obtain search warrants before downloading EDR data, but not always; in some cases, they do not have to. When Robert Christmann struck and killed a pedestrian on October 18, 2003, Trooper Robert Frost of the New York State Police downloaded data from the car at the accident scene. The EDR revealed that Christmann had been going 38 MPH in an area where the speed limit was 30. When the data was introduced at trial, Christmann claimed that the state had violated his Fourth Amendment rights against unreasonable searches and seizures, because it had not asked his permission or obtained a search warrant before retrieving the data. That was not necessary, ruled a New York court. Taking bits from the car was not like taking something out of a house, and no search warrant was necessary.

Bits mediate our daily lives. It is almost as hard to avoid leaving digital footprints as it is to avoid touching the ground when we walk. Yet even if we live our lives without walking, we would unsuspectingly be leaving fingerprints anyway.

It is almost as hard to avoid leaving digital footprints as it is to avoid touching the ground when we walk.

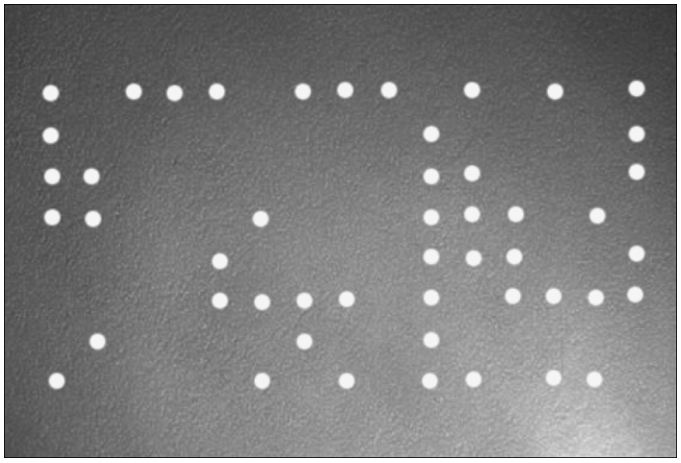
Some of the intrusions into our privacy come because of the unexpected, unseen side effects of things we do quite voluntarily. We painted the hypothetical picture of the shopper with the RFID-tagged shoes, who is either welcomed or shunned on her subsequent visits to the store, depending on her shopping history. Similar surprises can lurk almost anywhere that bits are exchanged. That is, for practical purposes, pretty much everywhere in daily life.

Tracing Paper

If I send an email or download a web page, it should come as no surprise that I've left some digital footprints. After all, the bits have to get to me, so some part of the system knows where I am. In the old days, if I wanted to be anonymous, I could write a note, but my handwriting might be recognizable, and I might leave fingerprints (the oily kind) on the paper. I might have typed, but Perry Mason regularly solved crimes by matching a typewritten note with the unique signature of the suspect's typewriter. More fingerprints.

So, today I would laserprint the letter and wear gloves. But even that may not suffice to disguise me. Researchers at Purdue have developed techniques for matching laser-printed output to a particular printer. They analyze printed sheets and detect unique characteristics of each manufacturer and each individual printer—fingerprints that can be used, like the smudges of old typewriter hammers, to match output with source. It may be unnecessary to put the microscope on individual letters to identify what printer produced a page.

The Electronic Frontier Foundation has demonstrated that many color printers secretly encode the printer serial number, date, and time on every page that they print (see Figure 2.2). Therefore, when you print a report, you should not assume that no one can tell who printed it.



Source: Laser fingerprint. Electronic Frontier Foundation. `http://w.2.eff.org/Privacy/printers/docucolor/`.

FIGURE 2.2 Fingerprint left by a Xerox DocuColor 12 color laser printer. The dots are very hard to see with the naked eye; the photograph was taken under blue light. The dot pattern encodes the date (2005-05-21), time (12:50), and the serial number of the printer (21052857).

There was a sensible rationale behind this technology. The government wanted to make sure that office printers could not be used to turn out sets of hundred dollar bills. The technology that was intended to frustrate counterfeiters makes it possible to trace every page printed on color laser printers back to the source. Useful technologies often have unintended consequences.

Many people, for perfectly legal and valid reasons, would like to protect their anonymity. They may be whistleblowers or dissidents. Perhaps they are merely railing against injustice in their workplace. Will technologies that undermine anonymity in political discourse also stifle free expression? A measure of anonymity is essential in a healthy democracy—and in the U.S., has been a weapon used to advance free speech since the time of the Revolution. We may regret a complete abandonment of anonymity in favor of communication technologies that leave fingerprints.

The problem is not just the existence of fingerprints, but that no one told us that we are creating them.

The problem is not just the existence of fingerprints, but that no one told us that we are creating them.

The Parking Garage Knows More Than You Think

One day in the spring of 2006, Anthony and his wife drove to Logan Airport to pick up some friends. They took two cars, which they parked in the garage. Later in the evening, they paid at the kiosk inside the terminal, and left—or tried to. One car got out of the garage without a problem, but Anthony's was held up for more than an hour, in the middle of the night, and was not allowed to leave. Why? Because his ticket did not match his license plate.

It turns out that every car entering the airport garage has its license plate photographed at the same time as the ticket is being taken. Anthony had held both tickets while he and his wife were waiting for their friends, and then he gave her back one—the "wrong" one, as it turned out. It was the one he had taken when he drove in. When he tried to leave, he had the ticket that matched his wife's license plate number. A no-no.

Who knew that if two cars arrive and try to leave at the same time, they may not be able to exit if the tickets are swapped? In fact, who knew that every license plate is photographed as it enters the garage?

There is a perfectly sensible explanation. People with big parking bills sometimes try to duck them by picking up a second ticket at the end of their trip. When they drive out, they try to turn in the one for which they would have to pay only a small fee. Auto thieves sometimes try the same trick. So the system makes sense, but it raises many questions. Who else gets access to the license plate numbers? If the police are looking for a particular car, can they search the scanned license plate numbers of the cars in the garage? How long is the data retained? Does it say anywhere, even in the fine print, that your visit to the garage is not at all anonymous?

All in Your Pocket

The number of new data sources—and the proliferation and interconnection of old data sources—is part of the story of how the digital explosion shattered privacy. But the other part of the technology story is about how all that data is put together.

On October 18, 2007, a junior staff member at the British national tax agency sent a small package to the government's auditing agency via TNT, a private delivery service. Three weeks later, it had not arrived at its destination and was reported missing. Because the sender had not used TNT's "registered mail" option, it couldn't be traced, and as of this writing has not been found. Perhaps it was discarded by mistake and never made it out of the mailroom; perhaps it is in the hands of criminals.

The mishap rocked the nation. As a result of the data loss, every bank and millions of individuals checked account activity for signs of fraud or identity theft. On November 20, the head of the tax agency resigned. Prime Minister Gordon Brown apologized to the nation, and the opposition party accused the Brown administration of having "failed in its first duty—to protect the public."

The package contained two computer disks. The data on the disks included names, addresses, birth dates, national insurance numbers (the British equivalent of U.S. Social Security Numbers), and bank account numbers of 25 million people—nearly 40% of the British population, and almost every child in the land. The tax office had all this data because every British child receives weekly government payments, and most families have the money deposited directly into bank accounts. Ten years ago, that much data would have required a truck to transport, not two small disks. Fifty years ago, it would have filled a building.

This was a preventable catastrophe. Many mistakes were made; quite ordinary mistakes. The package should have been registered. The disks should have been encrypted. It should not have taken three weeks for someone to speak up. But those are all age-old mistakes. Offices have been sending packages for centuries, and even Julius Caesar knew enough to encrypt information if he had to use intermediaries to deliver it. What happened in 2007 that could not have happened in 1984 was the assembly of such a massive database in a form that allowed it to be easily searched, processed, analyzed, connected to other databases, transported—and "lost."

Exponential growth—in storage size, processing speed, and communication speed—have changed the same old thing into something new. Blundering, stupidity, curiosity, malice, and thievery are not new. The fact that sensitive data

about everyone in a nation could fit on a laptop *is* new. The ability to search for a needle in the haystack of the Internet *is* new. Easily connecting "public" data sources that used to be stored in file drawers in Albuquerque and Atlanta, but are now both electronically accessible from Algeria—*that* is new too.

Training, laws, and software all can help. But the truth of the matter is that as a society, we don't really know how to deal with these consequences of the digital explosion. The technology revolution is outstripping society's capacity to adjust to the changes in what can be taken for granted. The Prime Minister had to apologize to the British nation because among the things that have been blown to bits is the presumption that no junior staffer could do that much damage by mailing a small parcel.

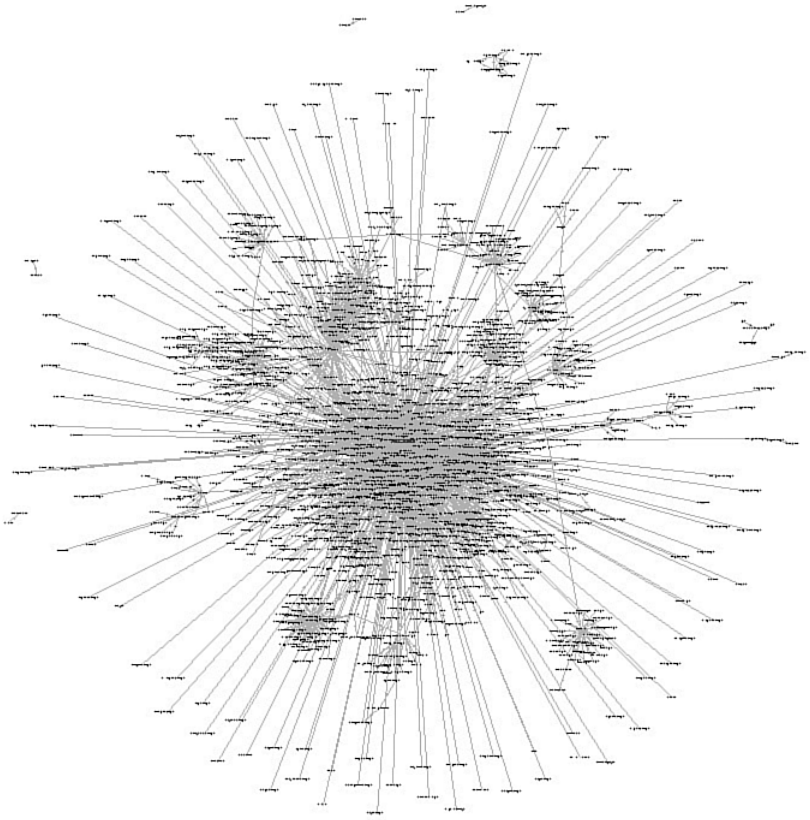
Connecting the Dots

The way we leave fingerprints and footprints is only part of what is new. We have always left a trail of information behind us, in our tax records, hotel reservations, and long distance telephone bills. True, the footprints are far clearer and more complete today than ever before. But something else has changed—the harnessing of computing power to correlate data, to connect the dots, to put pieces together, and to create cohesive, detailed pictures from what would otherwise have been meaningless fragments. The digital explosion does not just blow things apart. Like the explosion at the core of an atomic bomb, it blows things together as well. Gather up the details, connect the dots, assemble the parts of the puzzle, and a clear picture will emerge.

Computers can sort through databases too massive and too boring to be examined with human eyes. They can assemble colorful pointillist paintings out of millions of tiny dots, when any few dots would reveal nothing. When a federal court released half a million Enron emails obtained during the corruption trial, computer scientists quickly identified the subcommunities, and perhaps conspiracies, among Enron employees, using no data other than the pattern of who was emailing whom (see Figure 2.3). The same kinds of clustering algorithms work on patterns of telephone calls. You can learn a lot by knowing who is calling or emailing whom, even if you don't know what they are saying to each other—especially if you know the time of the communications and can correlate them with the time of other events.

Sometimes even public information is revealing. In Massachusetts, the Group Insurance Commission (GIC) is responsible for purchasing health insurance for state employees. When the premiums it was paying jumped one year, the GIC asked for detailed information on every patient encounter. And

for good reason: All kinds of health care costs had been growing at prodigious rates. In the public interest, the state had a responsibility to understand how it was spending taxpayer money. The GIC did not want to know patients' names; it did not want to track individuals, and it did not want people to *think* they were being tracked. Indeed, tracking the medical visits of individuals would have been illegal.



Source: Enron, Jeffrey Heer. Figure 3 from http://jheer.org/enron/v1/.

FIGURE 2.3 Diagram showing clusters of Enron emailers, indicating which employees carried on heavy correspondence with which others. The evident "blobs" may be the outlines of conspiratorial cliques.

So, the GIC data had no names, no addresses, no Social Security Numbers, no telephone numbers—nothing that would be a "unique identifier" enabling a mischievous junior staffer in the GIC office to see who exactly had a

particular ailment or complaint. To use the official lingo, the data was "de-identified"; that is, stripped of identifying information. The data did include the gender, birth date, zip code, and similar facts about individuals making medical claims, along with some information about why they had sought medical attention. That information was gathered not to challenge any particular person, but to learn about patterns—if the truckers in Worcester are having lots of back injuries, for example, maybe workers in that region need better training on how to lift heavy items. Most states do pretty much the same kind of analysis of de-identified data about state workers.

Now this was a valuable data set not just for the Insurance Commission, but for others studying public health and the medical industry in Massachusetts. Academic researchers, for example, could use such a large inventory of medical data for epidemiological studies. Because it was all de-identified, there was no harm in letting others see it, the GIC figured. In fact, it was such good data that private industry—for example, businesses in the health management sector—might pay money for it. And so the GIC sold the data to businesses. The taxpayers might even benefit doubly from this decision: The data sale would provide a new revenue source to the state, and in the long run, a more informed health care industry might run more efficiently.

But how de-identified really was the material?

Latanya Sweeney was at the time a researcher at MIT (she went on to become a computer science professor at Carnegie Mellon University). She wondered how hard it would be for those who had received the de-identified data to "re-identify" the records and learn the medical problems of a particular state employee—for example, the governor of the Commonwealth.

Governor Weld lived, at that time, in Cambridge, Massachusetts. Cambridge, like many municipalities, makes its voter lists publicly available, for a charge of $15, and free for candidates and political organizations. If you know the precinct, they are available for only $.75. Sweeney spent a few dollars and got the voter lists for Cambridge. Anyone could have done the same.

According to the Cambridge voter registration list, there were only six people in Cambridge with Governor Weld's birth date, only three of those were men, and only one of those lived in Governor Weld's five-digit zip code. Sweeney could use that combination of factors, birth date, gender, and zip code to recover the Governor's medical records—and also those for members of his family, since the data was organized by employee. This type of re-identification is straightforward. In Cambridge, in fact, birth date alone was sufficient to identify more than 10% of the population. Nationally, gender, zip code, and date of birth are all it takes to identify 87% of the U.S. population uniquely.

The data set contained far more than gender, zip code, and birth date. In fact, any of the 58 individuals who received the data in 1997 could have identified any of the 135,000 people in the database. "There is no patient confidentiality," said Dr. Joseph Heyman, president of the Massachusetts Medical Society. "It's gone."

It is easy to read a story like this and scream, "Heads should roll!." But it is actually quite hard to figure out *who, if anyone, made a mistake*. Certainly collecting the information was the right thing to do, given that health costs are a major expense for all businesses and institutions. The GIC made an honest effort to de-identify the data before releasing it. Arguably the GIC might not have released the data to other state agencies, but that would be like saying that every department of government should acquire its heating oil independently. Data is a valuable resource, and once someone has collected it, the government is entirely correct in wanting it used for the public good. Some might object to selling the data to an outside business, but only in retrospect; had the data really been better de-identified, whoever made the decision to sell the data might well have been rewarded for helping to hold down the cost of government.

It is easy to read a story like this and scream, "Heads should roll!." But it is actually quite hard to figure out who, if anyone, made a mistake.

Perhaps the mistake was the ease with which voter lists can be obtained. However, it is a tradition deeply engrained in our system of open elections that the public may know who is eligible to vote, and indeed who has voted. And voter lists are only one source of public data about the U.S. population. How many 21-year-old male Native Hawaiians live in Middlesex County, Massachusetts? In the year 2000, there were four. Anyone can browse the U.S. Census data, and sometimes it can help fill in pieces of a personal picture: Just go to `factfinder.census.gov`.

The mistake was thinking that the GIC data was truly de-identified, when it was not. But with so many data sources available, and so much computing power that could be put to work connecting the dots, it is very hard to know just how much information has to be discarded from a database to make it truly anonymous. Aggregating data into larger units certainly helps—releasing data by five-digit zip codes reveals less than releasing it by nine-digit zip codes. But the coarser the data, the less it reveals also of the valuable information for which it was made available.

How can we solve a problem that results from many developments, no one of which is really a problem in itself?