FEATURE

A DIGITAL TRAIL IS FOREVER

By Aaron Weiss

The data trail we generate through our everyday activity can be reassembled into a detailed account of our past, present, and possibly even future.

In 1976 the Eagles famously sang about the temptations of a certain hotel—you know, the one where "you can check out, but you can never leave." But if their cautionary tale were rewritten about life in today's digital environment maybe the updated lyric would be "you can log off, but you can never opt-out."

With apologies to Don Henley, we may not have a new hit song on our hands, but the metaphor rings true. Whether you participate actively, passively, or even not at all in life online, the new reality is that everyone is de facto opted-in. All that differs are matters of degree. Whether we post party pictures on Facebook, send email, make phone calls, swipe debit cards, live in a house, or walk down a street—we leave footprints and fingerprints in the form of a digital wake. But unlike traces through the air or water, our data trails are marked in permanent ink.

So what's new? Recordkeeping hardly began with the Internet. The answer is critical mass. Our old paper trails amounted to jigsaw puzzles with only a few pieces in place. Today, our digital trail leaves behind a forensic-quality virtual reality—one that can be pieced together from many sources of data to reconstruct our past, see into our present, and possibly even predict our future behavior.

It is impossible to talk about digital footprints without also talking about privacy. But the problem with talking about privacy is how to actually define it. Most would agree that your skin color is, for example, not private information since it is obvious and in plain sight. Is your birth date private? Weight? Home address? Income? Medical history? Last phone call made? Last Web search? Last one thousand Web searches?

The boundary between public and private might indeed lie somewhere in between, but pinning it down is not a simple exercise. In the U.S., the 1967 Supreme Court decision *Katz v. United States* established the boundary around an individual's "reasonable expectation of privacy." That is, you can reasonably expect that conversations held inside your home are private. You cannot reasonably expect that a conversation held while waiting on a line at Starbucks is private, since it is a public place. But, like Justice Potter Stewart's famous remark in the 1964 obscenity case Jacobellis v. Ohio, "I know it when I see it," the Court's definition of "reasonable" comes in shades of gray.

But times change, and in the years since 1967 the line between public and private—if it was ever clear at all—became less clear than ever. As daily life increasingly spills over from offline to online,

ILLUSTRATION BY RYAN ALEXANDER

boundaries are distorted or even erased. Is the whole Internet a public space, like standing in line at Starbucks? And if not, where and under what circumstances can we "reasonably" expect privacy online?

Sure, you can post vacation photos behind a password-restricted account. But what about the photos friends took that night you drank too much? In your own home? That they later posted to their own social networking accounts, open to the public, even tagged with your real name.

Not only does technology change over time, so do expectations. A 2004 MIT study entitled "Expectations on Privacy and Accountability" found that more than 50 percent of bloggers are under 30 years of age, and, of them, more than 55 percent publish under their real names. Plus, only a minority claim to limit publication of others' names or identities in their posts. Any parent whose kids are avid MySpace or Facebook users can vouch, probably with a cringe, how expectations of privacy are loosening.

Assembling Your Past

Not to take anything away from George Orwell, but it turns out that his oft-cited "Big Brother" falls far short of reality in an important way. Unlike Orwell's surveillance society that is centralized and state-controlled, today's surveillance is highly distributed. The records of our actions—movement, phone calls, email, and more—serve the interests of many different and sometimes unrelated parties.

It is the corner convenience store owner who captures our faces inside his shop for his own protection. It is the grocery store chain that records our purchases down to our preferred flavor of pop tart to manage inventory and sell marketing data. It is Google that indexes our email (at least, those sent through its Gmail system) to leverage its advertising business.

It comes as little surprise that virtually everything we do online leaves a distinct

digital trace. Every link you click on the Web is recorded by someone's Web server somewhere. It may know only a few nonspecific details about you—the link you clicked, the browser you're using, the city where your Internet connection is located—or if the site is one where you've volunteered your identity, perhaps with a login name, password, and profile.

Yet, even the technology savvy can find it jarring to see our past actions burned into history's digitized record. Subscribe to a VoIP service, as more people around the world are doing to replace traditional landlines, and you'll find detailed logs in your Web account of every single call placed and received—date, time, number, duration—a year ago or a minute ago. Mobile phones, too, provide instant logs of our actions, from calls to texts, in and out. We know this information has probably always been available, but while it used to be buried beneath layers of Ma Bell bureaucracy, it now sits in plain sight, only a click away, if one knows where to look.

For forensics investigators, our prolific digital wakes amount to a virtual time machine, enabling them to both reconstruct and dissect our pasts. It used to be that criminals might be tracked down through human fingerprints, if they were careless enough, but today we leave so many digital fingerprints it is virtually impossible to completely cover our tracks.

Traces

In addition to the literal logs of activity—from Web to email to IM chats to VoIP phone calls—many commonplace computer tasks leave behind "digital debris" that give clues to our behavior. When, for example, you delete a file, its contents aren't really erased from your hard drive. It may degrade, as new data is written in its place, but pieces could remain salvageable for a very long time. Software like Microsoft Word may keep temporary backup copies of documents, potentially recoverable in a legal investigation. Newer versions of both Windows and Mac OS X provide automatic file backup facilities—known as "shadowing," or in the Mac's case, literally "Time Machine"—which can also make forensics work easy pickings.

Then there's metadata. When you save a file—say, a document or photo from a digital camera—the computer saves more than just the data itself. Metadata includes information about the file (such as the date and time it was created). A Word document can contain metadata about the licensed user who created the document. A digital photo often contains a plethora of metadata, including not only date and time, but on some GPS-enabled models, location as well. Many cameras also record with each file detailed settings like shutter speed, f-stop setting, and white balance. But filesystems and cameras are just two sources of digital debris.

If you've tried to rollback the date on your PC to falsify entries in a patient database—like British physician Harold Shipman did between 1995 to 1998 to cover up his patients' suspicious deaths—investigators may discover that the Windows Event Log has recorded each time-warping action. Due in part to his digital trail, Shipman was convicted of killing 15 patients.

Or consider the more famous case of the serial killer known as BTK, or, in real life, security-alarm installer and church deacon Dennis Rader. Bragging about his crimes in electronic letters saved on floppy disks he mailed to local media, Rader apparently overlooked that deleted files can be recovered and, in his case, contain metadata that identified both him and the church that licensed the software—a church in which he happened to also be president.

If you are a criminal on the lam, being identified through digital forensics is not exactly your desired outcome. Your digital wake is not your friend. But what if you want to be found? To use an example growing so common it has led to its own websites, Flickr groups, and innumerable blog postings: Suppose you lose your digital camera. Suppose it is then recovered by someone with some tech smarts. Suppose that person can begin an Internet manhunt for you, using details from the camera's metadata, along with the subject matter of the pictures and the power of social networking groups like Flickr and Facebook. All this digging around in your digital wake has led to reunions between cameras and owners separated by distances as far apart as New York City and Australia, and Scotland and Italy, among (many) others. All of which, some civil libertarians say, begs the question: How should we define privacy in a world where everything we do becomes part of our globally accessible permanent record?

Peering into the Present

What are you doing right now? It's safe to say that you're reading this article, of course, but what if you were doing something else? It is hardly a stretch to observe that we are being observed. Directly or indirectly, our digital trails aren't confined to the past and indeed lead right up to this very moment.

If you're online, then, by definition, your data is passing through computers between your own and its destination. At the very least, this includes routers

operated by your ISP. So it is not difficult to see the value to business and government in deploying deep packet inspection, or DPI, technology.

Data packets are basically composed of two parts, a header and content data (sometimes known as a "payload"). This design is much like regular postal mail, which contains a payload enclosed inside a wrapper—like an envelope or box—containing information about the destination and possibly the contents (such as hazardous liquid or breakable materials).

It has long been possible for ISPs to also perform stateful packet inspection, or SPI, which essentially looks at just the header (wrapper) of packets moving in and out of the network. One reason to do this is to manage how different kinds of data are handled. For example, packets containing torrent transfers can consume a lot of network bandwidth, so an ISP might want to put them in the slow lane. (Or block them altogether, a practice that cable ISP Comcast dabbled with in 2007 and 2008, ultimately attracting the interest and ire of the U.S. Federal Communications Commission.)

But with advances in processing power, it is now increasingly feasible to implement DPI, which goes beyond the wrapper to look at the actual data content of packets. One argument for DPI is that SPI makes only rough determinations about traffic content and that by using the more granular DPI, ISPs are able to implement new business models, like tiered pricing for different kinds of services. (Some consumer advocates think this would not be a good business model for customers, potentially leading to higher prices for popular Internet services)

The truth about DPI that strikes fear into the blogs of privacy advocates is that it can be used to find, censor, or otherwise act on "forbidden" content. China reportedly uses DPI to enforce its "Great Firewall," filtering data packets related to political dissent and other taboo subjects. In the U.S., the National Security Agency is known to use DPI in conjunction with at least some telecom data centers (such as AT&T in San Francisco), presumably to monitor for terrorism-related "chatter," but few details are available about how precisely it uses the technology. DPI is also advocated by trade groups like the RIAA and MPAA to sniff out transfers of copyright-infringing content.

One concern about DPI, though, is that it could eventually become its own worst enemy. Savvy users foil DPI by encrypting their network activity through such means as SSH (Secure Shell network protocol) tunnels and VPN (virtual private network) connections. Widespread use of DPI could encourage more mainstream software to integrate encryption, possibly leading to a situation where most network traffic is encrypted—not unlike how the overprescription of antibiotics has led to drug-resistant strains of bacteria and potentially untreatable diseases.

A technology like DPI may cast an ominous shadow, but our digital trail can begin—and be intercepted—right at our fingertips. Imagine looking at a record of everything you typed. You wouldn't have to hack into email accounts or Web servers to dig up the most sensitive pieces of information.

Keyloggers, or malicious pieces of software that can be installed through spyware, are one way to sniff your trail in real time, though it still requires some degree of hacking into the target's machine. But why hack at all, when you can sniff keystrokes through their audio or RF fingerprints? It turns out that each key produces a slightly different acoustic profile and also a slightly different electrical signal that can be "seen" in the air with the right monitoring equipment.

Granted, there aren't likely to be keystroke sniffers lurking behind every office door. But the fact remains that a single keypress is the beginning of a digital trail that can be seized upon nearly instantaneously.

You don't necessarily need to be online to start a digital trail. The offline world, referred to by some in the technical community as "meatspace," is rarely completely offline. Take, for example, your cell phone. Unless it is powered off, it's probably logged onto the nearest tower. Which is why if you have a cell phone and you go missing, the authorities will start their hunt by identifying the tower most recently associated with your phone. Depending on the area and signal strength to your phone, this can narrow your location to anywhere from less than one to several dozen miles.

Why wait for the authorities to find you? Illustrating how expectations of privacy are changing, consider the Google service called Latitude. Used in conjunction with a supported device, like a Blackberry or Windows Mobile phone, you can broadcast your whereabouts to a circle of friends. Google will even show you a little map with markers where your friends are located right now, updated in real time.

Likewise, the major U.S. cell service providers now sell "family tracking" features that basically allow phones on

a shared plan to display one another's whereabouts in real time. These services are pitched as a way to keep track of kids who may be wandering, lost, or even abducted. But critics say these technologies are a slippery slope, asking how long until, say, the government proposes tracking vehicles to impose a tax on miles driven? Earlier this year just such a proposal was floated by several states, including Idaho, North Carolina, Oregon, and Rhode Island, as well as by President Obama's Transportation Secretary Ray LaHood (rejected by the White House). Both car rental agencies and auto insurance companies have also experimented with tracking vehicle location, though legal obstacles have limited these applications thus far.

A Google Maps camera car surveying the roads in Calabasas, CA.

Rake the Yard, Here Comes Google

A significant marker in our digital trails—our homes—doesn't involve movement at all. At a very personal level, our home feels like the ultimate form of privacy. Which might explain why residents of the U.K. village of Broughton recently formed a human chain to prevent entry to a Google camera car, the kind that drives around cities and photographs 360-degree views for use with the street view feature of Google Maps.

The villagers knew full well that once Google was through with them, the exteriors of their houses—like those in most major U.S. cities—would be readily viewable by anyone in the world via just a few mouse clicks. They're not the first to object; even within the U.S., a Pittsburgh-area couple attracted media attention with a lawsuit against Google, hoping to defend against its alleged invasion of their personal privacy.

In Google's favor, though, is the law. In the U.S—and to varying degrees in other countries—photographing homes and buildings is not a violation of privacy, so long as Google sticks to vantage points in public spaces. (Indeed, Google prevailed in the Pittsburgh case.) Many cities and towns have for years compiled their own photographic databases of houses and commercial buildings to augment property and tax records available to the public.

To the extent that personal data compiled locally has not raised the same concern, or in some cases outrage, as Google's globetrotting effort, is another shade to add to the ongoing question: Is our expectation of privacy influenced not only by the nature of our exposure but by its quantity as well? Did we cross an invisible boundary when our digital trails were made available to the eyes of billions of our fellow cybercitizens rather than mere thousands of local neighbors?

A Pebble Tossed in the Ocean

One way quantity absolutely matters is when trying to pluck a single trail out of millions of digital footsteps. Although it can be unnerving to think about how many traces we leave behind every action, both online and offline, consider that the same math applies to just about everyone. Tempting as it might be to think like a paranoid narcissist—"They're all watching me!"—our individual wakes are tiny, albeit permanent, ripples in a roiling sea.

With quantity, one might be tempted to think, comes anonymity. Yet anonymity, research is discovering, may provide false cover. In 2006 AOL inadvertently demonstrated the limitations of "anonymity" on a mass scale when it published more than 20 million search queries from more than 650,000 of its own users. The data was ostensibly intended to attract researchers, and

PHOTOGRAPH BY JON DELOREY

AOL "anonymized" the information by replacing user ID or IP with seemingly arbitrary identification numbers. But it quickly became clear to both researchers and privacy advocates that, given enough search terms, it is possible to identify individual users out of the millions. For example, a user whose searches include, say, weight loss, football, workplace, driving directions from home to various destinations, and violent fantasies could unwittingly provide enough information to accurately pinpoint the individual in question.

Similarly, combining quantities of personal digital information with correlations across multiple data sets can also effectively "out" individual identities from the mass of humanity. Consider that many active Internet users participate in multiple social networks. In 2006 a pair of researches from the University of Texas at Austin demonstrated that with sufficient data, one could identify Netflix users who have rated as few as five films by correlating public "anonymous" data from Netflix with public data at another popular movie site, the Internet Movie Database.

More recently, the same pair found that many users of the social networking site Twitter also use Flickr, and that one-third of users with accounts in both services could be identified by correlating their participation in each, combined with any personal information given out in only one or the other online community.

To some privacy advocates, these findings represent the very dangers that lurk behind our digital trails and they hope to combat through means beyond technology, including legal and political. But to others, notably among high school and college students—these findings may represent an accepted and even natural social reality that influences their expectations—or lack thereof—of privacy.

Foretelling the Future

It seems inevitable, really. Our digital trails already illuminate what we did and what we're doing, so why not take the next step and use them to try to predict what we'll do next? Reading the tea leaves of consumer behavior can provide a competitive business advantage, leading to increased profits. Which is why opinion research firms now monitor blog postings to take the pulse of both individual and public sentiment. Influential bloggers can be recruited (co-opted?) as unofficial corporate ambassadors, or even potential shills, privy to perks like inside information and material reward. Persistent complainers can be courted and soothed, so, for example, they are less motivated to post another aggrieved comment about a company or its products.

Advertisers are exploring how to see into the future with the help of models like "behavioral targeting" and "predictive advertising" that aim to anticipate consumers' future choices based on past events. Both send shivers down the spines of privacy advocates for their use of aggressive terms, calling customers "targets" and for the way these models depend on how they exploit our digital trails.

The behavioral marketing model looks at your most recent digital footprints as a clue to what is currently on your mind. Suppose, for example, you click on a series of travel sites when researching airfares, hotels, and car rentals in Florida. This trail of clicks suggests you probably intend to take a trip there. So, a behavioral ad network may show you ads that someone taking such a trip would likely be interested in, like discount tickets to Disney World. A more sophisticated behavioral network might look at your past surfing activity and determine that you are interested in fishing and baseball and display ads for sport fishing in Miami and tickets to spring training games in Tampa.

Predictive advertising takes a broader view of your digital trail, including information like your purchase history of goods and services, shared among retailers. Perhaps you like to buy shoes but tend to limit yourself to those in the $50-$100 range. The predictive ad network would not display ads for $25 shoes or $300 shoes, expecting that your past behavior is likely to signal your future choices.

Our digital trails may make it easier, and more profitable, for business to sell us stuff, but may also foretell our political choices and even our criminal inclinations. Digital forensics could theoretically help prevent crimes that seem likely to occur, rather than only solve those that have already claimed their victims.

Critics—privacy advocates and civil libertarians among them—fiercely oppose the breaches of privacy required to, say, predict future crimes. But in a world where virtually everything we touch leaves a digital record—a beacon declaring "I was here"—how much privacy is really left to expect? ◀

Aaron Weiss is a technology writer and Web developer living in upstate New York, as well as the human caretaker of livenudecats.com.

DOI: 10.1145/1540392.1540397