3 The EU Data Retention Directive in an Era of Internet Surveillance

Hal Roberts and John Palfrey

Introduction

The European Union (EU) enacted a directive on data retention in 2007 that requires all member countries to mandate the retention by telecom companies of the sender, recipient, and time of every Internet or other telecom communication. The directive requires the collection of the Internet Protocol (IP) address, user ID, phone number, name, and address of every sender and recipient, but explicitly excludes (but does not forbid) the monitoring of content itself. All the monitored data must be retained for a period ranging between six months and two years, contingent upon the local law of each member state. Telecom companies must promptly give these data to law enforcement authorities upon request to assist with serious crimes, overseen by a national public authority monitoring the data retention practices.[1] As of its effective date of April 2009, all Internet service providers (ISPs) in EU countries must comply with the relevant national implementations of the directive.

The monitoring required by the EU data retention directive amounts to a form of surveillance. The directive does not take the form of surveillance that most quickly leaps to mind: the two men in a van with headphones listening to phone conversations of an unwitting crook in a seedy apartment. But it has functional similarities. The directive requires that the participating states collect personal data about their citizens without the citizens' consent. It enables states to use these data to control some of the subjects of monitoring (including by arresting them). What matters most about the directive and its relationship to surveillance is its impact on citizen activity and its place in the growing constellation of surveillance activities online.

The Internet is a ''surveillance-ready'' technology. There is a wide range of choices for any state that wishes to know more about its citizens. This digital information comes in the form of bits of data that flow through rivers and into oceans of data. These rivers are full of information that passes by a given point, or series of points, in a network and can be intercepted; these oceans are stocked with information that can be searched after the fact; and the rivers arise from springs that can be watched at the source. The data involved are held in private hands as well as public.

This chapter paints a simplified picture of the technical landscape of Internet surveillance, as well as the place of the EU data retention directive within that landscape, by taking up a series of short cases about surveillance. We examine how these cases inform (and are informed by) the technical questions of what *data* are being *actually* and *potentially* monitored on the Internet and whom we are *trusting* to access that data.

We break Internet surveillance into three broad categories: network, server, and client. The Internet is composed of clients and servers, in essence a series of devices that talk to one another through the network. Every bit of data on the Internet is traveling or residing at one or more of these locations at any given time. As such, any given Internet activity must happen at one (or more) of these locations. We treat any surveillance happening on the end user device as client-side surveillance, including both software tools like workplace keylogging systems and hardware tools like keyboard tapping devices. We treat any surveillance happening on a machine that predominantly accepts requests, processes them, and returns responses as server-side surveillance. And for simplicity, we treat everything between the client and the server as the network, including the wires over which the data travel and the routers that direct the traffic.

We argue in this chapter that the EU data retention directive introduces new risks related to the networks of trust created by each category of surveillance. In the section on network surveillance we argue that trust can only be rerouted around the network, rather than removed from it, and that the EU data retention directive may cause users to reroute trust in the network in ways that both reduce the amount of useful data available to law enforcement and encourage users to expose more of their network data to non-EU states. In the section on server surveillance we argue that users struggle to evaluate how the data they submit to servers are used and combined and that the EU data retention directive is likely to increase this problem by requiring ISPs to store more server data (which will be used and combined in ways opaque to most users). In the section on client surveillance we argue that the client has become an intensely complicated battleground of trust played out by sophisticated actors with their own agendas, resulting in widespread leaks of data from the client. This section argues that the EU data retention directive will place a large new set of private data onto this battleground—to the detriment of people from around the world.

Network Surveillance

The most obvious kind of Internet surveillance takes place on the network between the clients and servers. Government agencies collect data from within network ISPs, including not only wiretap-like data about specific subjects with warrants but also entire streams of data for mining without judicial oversight. But the network is a diverse place. There are a wide variety of different actors with different access to data. As a re-

sult, there are a wide variety of other cases of network surveillance: about users who try to get around surveillance but end up exposing themselves to different sorts of surveillance in unexpected ways, about the collection of extensive user data without meaningful consent for targeted advertising, and about how criminal organizations exploit the trust model of the network to facilitate illegal surveillance.

The U.S. Communications Assistance for Law Enforcement Act (CALEA) requires that telecommunications companies have the ability to respond quickly and fully to wiretap requests even when using the newly digital telephone switches. The current interpretation of this law includes not only traditional telephone service but also Internet telephone services like Skype and, pending a ruling by the Federal Communications Commission, maybe even sender and recipient data for all Internet traffic. Such a ruling would render CALEA in a sense analogous to the EU data retention directive. The FCC has ruled that ISPs can forward their entire data stream to independent ''trusted third parties'' to handle the wiretapping implementation, exposing data streams of entire ISPs to these third parties.[2]

More intrusively, the National Security Agency (NSA) is apparently mining the full stream of data passing through major ISP backbones in the United States.[3] There are limits to our understanding, as laypersons, of this process, for obvious national security reasons. We know that the equipment used for this surveillance is capable of executing highly sophisticated queries on the data passing through the backbone. We know that the NSA is engaged in some level of warrantless surveillance of the international communications of U.S. citizens, but we do not know precisely what is being done with the data.[4] We are left with indeterminate, circumstantial evidence about the existence and function of the surveillance that leaves unanswerable questions about what data the NSA is making available to whom.

Relakks is one of many proxy tools available on the Internet that encrypts and reroutes traffic to avoid monitoring or filtering by anyone, such as the NSA, monitoring the user's local ISP. When someone in China uses Relakks to request a page from the BBC, the connection goes from the user's ISP in China to Relakks in Sweden (where Relakks is hosted) and then from Relakks in Sweden to the BBC in Britain (and back along the same route). The Chinese ISP can only see a connection to Relakks in Sweden, hiding the ultimate destination of the request (and bypassing any filtering as well). But in 2008, Swedish lawmakers authorized the the Försvarets Radioanstalt (FRA), or National Defense Radio Establishment, a government agency responsible for signals intelligence, to monitor the content of all international Internet and phone traffic (including that of Relakks) without a warrant, requiring all Swedish ISPs to install FRA monitoring equipment.[5] Now, Relakks users (at least, those who do not follow Swedish politics) are unwittingly handing their complete Internet data streams over to the Swedish government as they seek to avoid monitoring on their own local networks.

Users have reason to be concerned about forms of surveillance on their local networks beyond government monitoring. In June and July of 2007, several British Telecom (BT) Internet customers noticed strange problems with their Internet connections that they tracked down to a spyware company, 121media.[6] BT insisted that it had nothing to do with the suspicious behavior, and 121media refused to comment on the grounds of customer (BT's) privacy.[7] But in 2008, 121media, renamed Phorm, publicly announced a deal with BT to target advertising at the ISP's customers.[8] Phorm soon afterward admitted, in response to media reports of the 2007 activity related to 121media and BT, that it had already tried its targeted advertising on tens of thousands of users on the BT network with BT's help but without the knowledge of the users.[9] Phorm claims that it does not store any personally identifying information or browsing histories in the process of targeting ads; Phorm says it only stores information about the kinds of sites each user visits (expensive cars, rugby sites, and so on) connected to the user only by a randomly generated unique ID.[10] Privacy advocates have reacted strongly against Phorm's announcement and justifications, but BT continues to push for a full rollout of the system.[11]

As with the Relakks case, user efforts to circumvent network monitoring can have unpredictable results on the network of trusted relationships that tie together Internet activity. In February 2008, a Pakistani ISP responded to a government request to ban a video on YouTube.com by (probably accidentally) blocking a majority of the entire Internet from accessing the whole site for a few hours.[12] During the few days Pakistan was blocking YouTube.com locally, many Pakistanis bypassed the block using a tool called Hotspot Shield.[13] AnchorFree describes Hotspot Shield as a privacy tool: ''You remain anonymous and protect your privacy.''[14] But AnchorFree makes money by injecting ads into Web pages, and users of the tool give AnchorFree complete access to all data exchanged while Web browsing with the tool. AnchorFree implies (but never explicitly says) that it does not monitor its users' traffic, but it nonetheless has both the ability to snoop on the data at any time and a business model based on processing user data for advertisers. Thus, Pakistan is monitoring its citizens' Internet traffic to block content it does not like, and citizens are accessing the blocked content by using a tool that circumvents Pakistan's filters. But the circumvention tool is at least potentially just a monitoring tool for the different purpose of advertising.

To make sense of these cases, we need to understand what data are available on the network and to whom they are accessible. Three sorts of data are vulnerable to surveillance on the network: routing information, the actual content of the data stream, and contextual signatures. All Internet data packets must include the IP address of the ultimate recipient, and most data packets (including all Web and e-mail traffic) also include the IP address of the sender. Users can hide routing information on the network by using proxies, like Relakks or HotSpot Shield, which forward communication between a client and a server. In addition to the routing data, the packets contain both protocol-specific data (data about the URL requested, the referring URL, the user agent,

and so on for Web requests; data about the originating e-mail server, the ''from:'' e-mail address, the date, and so on for e-mails) and the content proper of the communication. This content includes, but is not limited to, any data submitted to Web sites, any Web pages retrieved, and any e-mails sent or received.

Users can encrypt the content of their network communications to hide their activities from network monitoring. Encryption is most effective when it is applied end-to-end, as by using Hypertext Transfer Protocol Secure (HTTPS). In this case, the entire stream of data from the client to the server is encrypted, allowing no one on the network between the sender and the receiver to read the content. But end-to-end encryption has to be supported by both the client and the server, and many servers do not support encrypted communication at all or for all pages. In these cases, a user can connect through a proxy like Relakks or HotSpot Shield to encrypt the data from the client to the proxy. But such encrypting proxies still use unencrypted channels to talk to servers that do not support encryption. As a result, the proxied content remains readable to any intervening routers on the network. For instance, even though content between Relakks and the user is encrypted, the requests and responses between Relakks and Google.com are not encrypted.

These apparently secure connections between Internet users leak other, contextual forms of information as well. The information about those requests and responses—think of it as ''metadata'' to the ''data'' of the communication itself—can be observed by anyone on the network in between. Communications that are both encrypted and proxied can hide both the routing information for, and the content of, a communication from the network between the client and the proxy. But even proxied and encrypted data leaks some of its metadata: information about the timing, number, and size of the packets as well as the fact that the communication is proxied and encrypted, may be observed. Different sorts of traffic generate different signatures of packet size and timing that can allow easy identification of the nature of the communication. These signature-based monitoring methods have reportedly been used, for example, to block proxied file-sharing traffic.

The Internet consists of billions of links between clients, servers, and routers, making comprehensive surveillance of the entire network very difficult. But in practice, all Internet traffic flows through a much smaller number of links between routers, and those routers are controlled by a much smaller yet number of autonomous systems (ASs). These ASs are the independent entities (mostly ISPs) that have the ability to route traffic on the Internet. There are fewer than 100,000 of these ASs in the world. In practice, the vast majority of Internet traffic flows through an even much smaller number of those ASs. For instance, virtually all 300 million Internet users in China connect to the Internet through only five big ISP ASs.[15] For a combination of technical, business, and policy reasons, a disproportionate amount of global Internet traffic flows through a few very large ASs in the United States, including most traffic between Europe and Asia.[16] Traffic between ASs within a given country (or sometimes even a

given city) often flows through an Internet exchange point (IXP), a physical network node that connects geographically close ASs. There are fewer than 200 major IXPs in the world, which together carry much of the Internet's local traffic.[17] IXPs keep local traffic local, so unlike the ASs responsible for routing intracountry and intracontinental traffic, most IXPs are located in (and therefore potentially under the jurisdictional control of) the country whose traffic they carry.

This topology of the Internet has several implications for the actors trusted with access to Internet data. The first is that a large majority of users need to access the Internet through an AS (usually an ISP). The situation should be familiar by now: data from these users is therefore vulnerable to surveillance by someone controlling that AS. The second is that there are a relatively small number of these ASs within any given country and an even smaller number of IXPs, so monitoring all the network traffic in a given country is a manageable task of making the small number of ASs and IXPs monitor their networks (though some of these ASs can be big complex organizations in themselves). This rule applies doubly for international Internet traffic, which is controlled by an even smaller number of ASs disproportionately located in the United States.[18] And the United States is capable of monitoring a large portion of international Internet traffic through a few ASs based in the United States, including even traffic flowing between non-U.S. countries.

A user can move her trust around from provider to provider. In the end, though, she must ultimately trust someone on the network (barring the unlikely event of widespread adoption of end-to-end encryption of networked digital communications). In practice, almost every user of digital networked technologies ends up trusting more actors over time. A user who tries to get around surveillance of her local ISP connection, for instance by Phorm, has only a few choices of ISPs in the United Kingdom, most of whom have been reported to be considering adding Phorm monitoring to their networks. A user may choose to stay on the possibly monitored local network but use a service like Relakks to proxy and encrypt her data as it travels through the local, Phorm monitoring, ISP. The user will avoid Phorm monitoring in the process, but at the cost of trusting not only Relakks but also the Swedish ISP through which Relakks talks to the Internet *and* the Swedish government that monitors the data flowing through all Swedish ISPs. And her local ISP will still be able to tell that she is proxying and encrypting all her data through a third party, which fact itself might prove suspicious to a law enforcement agency. Finally, the user's data are vulnerable to network monitoring at any point along which they travel, from her local ISP to the server's ISP to any ISPs between. A portion of her data is likely to travel outside of Europe through one of a few ISPs in the United States that process a disproportionate share of international Internet traffic.

Against the backdrop of these forms of network surveillance, we can see that the EU data retention directive has the potential to distort the network of trust through which Internet data flows. The precise effects of these distortions are difficult to predict. Some

number of EU users may react to the increased monitoring by using encrypted proxies in non-EU countries to avoid local ISP surveillance, maybe to hide illegal activity but maybe to hide legal but sensitive personal activity. The directive may encourage these proxy users to route their connections through non-EU countries to avoid the data retention mandate altogether.

The net effect of this process may very well be that law enforcement has less access to useful data. Likewise, smart users may proxy their data through autocratic countries like China that are least likely to share data with EU countries to avoid monitoring by another EU country compliant with the data retention directive. If enough users resist monitoring through the use of proxies, ISPs will be pressured to try to block such proxied traffic. If ISPs were to take this step, they would likely start an arms race between proxy tools and proxy blocking tools analogous to the current arms race between malware and antivirus software (as has happened in China with its attempts to block filtering circumvention tools). This arms race could further strain the network through the same sort of knock-on effects of the malware arms race: users wanting to avoid monitoring would become increasingly dependent on increasingly sophisticated anonymizing tools and therefore become increasingly vulnerable to the developers of such tools (some with good, some with bad intentions as with the current developers of antivirus tools). We have already seen this process with music downloading tools, which have become a vector for malware infections as the music industry has driven them underground.[19]

The popularity of tools that circumvent various sorts of filtering of music downloading tools points to the possibility that large numbers of users will attempt to route around the monitoring, particularly if authorities brand illegal music downloading a ''serious crime.'' But users also have a tendency to accept gradual erosions of privacy without resorting to resistance.[20] Even relatively small-scale usage of proxy tools could have the strong effect of making those few users very susceptible to false suspicion by law enforcement authorities. Given the widespread availability of proxying tools and widespread publication of the EU data retention directive, it seems safe to assume that many of the most serious criminals will use such tools, reducing the effectiveness of the monitoring of such users. Internet service providers can use contextual information about connections to detect the use of proxies and will be tempted to track which users are resisting monitoring. This possibility raises the risk that merely resisting monitoring will label a user as suspicious.

Server Surveillance

As with the network, the collection of data on Internet servers is highly concentrated among a few big actors. Even though there are hundreds of millions of servers on the Internet, a few large entities like Google, Yahoo, Facebook, and Wikipedia capture a large proportion of Internet traffic. Virtually all these big sites collect data about their

users. For example, Google and Yahoo collect search queries (among many other sorts of data), Facebook collects the social maps of its users, Wikipedia collects the editing histories of its users, and so forth. The collection of these sorts of data does not look like the typical Internet surveillance performed by the NSA and other government actors, but it represents a second type of surveillance on the network, equivalent in scope to the collection of data flowing across the network.

Google in particular (but not alone) has collected a tremendously large and intrusive amount of personal data about its users through the operation of its various services. Google argues that users should not be overly concerned about Google's data collection: ''we remember some basic information about searches. Without this information, our search engine wouldn't work as well as it does or be as secure. [What this information] doesn't tell Google is personal stuff about you like where you live and what your phone number is. . . . Logs don't contain any truly personal information about you.''[21]

User search data surely help in the way Google says they do (''improve our search results,'' ''maintain security,'' and ''prevent fraud''). And their explanations of the privacy implications of the data usage of Google's engineers are accurate in a narrowly technical sense. But the importance of data is determined by the larger world in which it exists—by the other data that it connects to. Google's statement that a cookie does not tell them ''personal stuff about you like where you live'' is only true in the narrow sense that your driver's license number does not tell the police where you live. Even though the cookie itself is just a random string of gibberish letters, it can indeed be used to look up personal information ''like where you live.'' For example, the cookie connects to all searches performed by a single person. Many people search for their own names at some point and for their own addresses at some point (if for no other reason than to see their houses in Google Maps). The cookie connects those two searches to the same (otherwise anonymous) person, thus potentially identifying the name and address of the person behind the random gibberish of a particular cookie. Researchers have consistently shown the ability to crack the identity of individual users in these kinds of data collections with anonymous but individually unique identifiers.[22]

It is likely that Google's collection of search terms, IP addresses, and cookies represents one of the largest and most intrusive single collections of personal data online or off-line. Google may or may not choose to do the relatively easy work necessary to translate its collection of search data into a database of personally identifiable data, but it does have the data and the ability to query personal data out of the collection at will. Even assuming perfect security to prevent data leaks, Google is still subject to many sorts of government requests for its data. Witness, for instance, the success of Viacom in subpoenaing the complete log of every video ever viewed on YouTube.com (which is owned by Google) in the context of copyright litigation.[23]

In addition to its search engine, Google's AdWords displays ads on about 35 percent of all advertising Web pages.[24] Google logs instances of consumers clicking on ads

through its AdWords system and watches the advertisers through AdWords auctions that determine the value of advertising topics. Content providers track the value of those advertising topics to determine which sorts of content to publish. The advertisers use the AdWords system as a stateless form of market research to target consumers without knowing anything about them. Content providers watch consumers to determine which sorts of content generate the most interest. All this monitoring happens in real time. Google adjusts the placement of ads in real time according to the current results of an ad auction; content providers watch the profitability of their content in real time and make adjustments to attract more ad-clicking customers; advertisers adjust their bids and update their ads in real time to attract more users. The effect of the system is continuous but stateless market research that is constantly adjusting to the current interests of users rather than the historical interests over time tracked by user profiling organizations like comScore, Phorm, and NebuAd.

Google and other service providers only have access to data that is sent directly to them over the network: the client address, the request itself, and any content explicitly submitted by the user. All these server-collected data, other than the client address, may be encrypted while traveling over the network, giving the end server access to some data that are not available on the network. For Web servers, the protocol data may include cookies, which are often used to assign a pseudonymous identity that persists between separate requests. Users voluntarily (and knowingly, at least in theory) submit vast amounts of such data to Internet servers. Users are aware that they are submitting names, addresses, and credit card numbers to Amazon when buying merchandise, personal e-mail to Microsoft through Hotmail, and movie preferences to Netflix when renting videos.

What determines the risk of privacy intrusions—and what ties this case to the narrative of online surveillance—is not just the collection of data but rather what the actors controlling the servers do with the data, with whom they share the data, and how the data are combined with other data. The act of collecting a credit card number to execute a purchase for a user is presumably acceptable and necessary in the modern global economy. But using the credit card number to request data about a user's purchase history from the credit card company in order to target advertising at him may not be so acceptable. Likewise, it is fine for Microsoft to collect personal e-mails through Hotmail, but it would become a concern if Microsoft were to sell its users' e-mail content to a consumer research company. And Netflix seems innocuous when using video preference information for its own recommendation engine, but many users would be uncomfortable if they found out Netflix was combining its users' video rental history with (even public) information from users' social networking pages to make video recommendations.

The issue of combining data is particularly relevant (but not unique) to server-based surveillance because, in comparison to network and client surveillance, the domain of the data collected is generally much more limited. However, these domains of data can

almost always be combined either with themselves or with other domains to create a much more personal, intrusive set of combined data—for instance, by combining the Google search data with the Google search cookies to identify users by cookie, or by combining the Google search IP addresses with ISP logs (as required by the EU data retention directive) to identify users by IP address.

All these different possible uses and combinations of data represent networks of trust. A rational user ought to evaluate these decisions about trust carefully, though in practice few have the time to do so with any level of sophistication. The example of Google demonstrates the complexity of these issues of trust—about how data are collected, who has access to it, and what is done with it. Google is collecting vast amounts of information from users in ways that are not clear to most users, yet most users eagerly accept the arrangement that Google offers them. Most users presumably understand that they are giving Google access to their search terms, but some may not understand that Google is storing these data. Yet other users are likely not to understand that Google generates its revenue through its advertising brokerage business. It is not at all clear that clicking on any Google AdWords ad takes you to a Google server first and only then redirects you to the clicked ad. Nor is it clear that by clicking on a Google AdWords ad, you are sharing with the advertiser the fact that you searched for or browsed content about a given subject. The potential of Google's vast store of user data creates a serious risk of disclosure throughout this network of trust regardless of whether Google's intentions are in fact good for its customers.

The exchange of data between user and server establishes a relationship of trust. A survey by the Internet security company WebSense found that 60 of the 100 most popular sites on the Internet had hosted malicious code at some point in the past year.[25] The examples that WebSense cites are attacks on the Web pages displayed to users rather than the back-end servers, so they do not give direct access to user data stored on the servers. But they do hijack the identity of the server on behalf of the attacker, allowing the attacker to present a portion of the Web page as if it is coming from the trusted server. The result is a variety of attacks that collect data on behalf of an attacker posing as the trusted server. These widely prevalent Web site attacks allow attackers to insert themselves into users' networks of trusted actors by way of the infected sites.

Google tried to use the EU data retention directive as part of the rationale behind its eighteen-month data retention period for certain user data, though some observers contend that the directive does not apply to Google as a "content provider," as opposed to a "communication service."[26] EU commissioners beat back that particular argument and have aggressively lobbied for Google to reduce the amount of data it keeps and how long it keeps the data under its privacy directive.[27] This exchange demonstrates precisely the problem at the intersection of the EU data retention directive and the way that surveillance works in the networked public sphere today. The same EU authority that is responsible for guarding against the retention of personal

data by service providers like Google is at the same time requiring ISPs to increase the amount of data they are storing and the length of time over which they are storing it. ISPs are companies just like Google, and requiring that they store the data required by the data retention directive imposes the same sorts of risks of unwarranted surveillance.

Client Surveillance

In addition to surveillance on the network and on servers, there is a range of different actors directly watching users' machines in various ways. The number of different actors trying to control access to the client has turned the end user's computers into a constant battleground of surveillance. Malware creators try to infect machines to steal valuable personal information. Anti-virus developers watch computers to find these malware and other sorts of snooping software, including sophisticated family and workplace surveillance systems. Even personal computing equipment—say, a laptop—can itself be turned into a surveillance system, through the use of keyboard logging devices that are installed inside the box of the computer, making them virtually impossible for the casual user to detect.

SpectorSoft was selected as *PC Magazine*'s editor's choice for ''monitoring'' software and is the largest provider of such software for home, small business, and corporate use. It captures virtually every kind of activity on the client computer and can send all of the monitored data to a remote computer for viewing.[28] Since becoming the market leader in 2004, SpectorSoft has stopped advertising to spouses, citing legal ambiguity and spousal abuse, but it continues to market itself for use monitoring minors and employees, neither of whom are legally protected from such monitoring in the U.S.[29] SpectorSoft reported in 2007 that its software was installed on over 400,000 desktops, 60 percent of which were home users.[30] All major anti-virus products classify SpectorSoft as spyware and attempt to remove it, but SpectorSoft actively avoids detection by either the monitored subject or anti-virus or other anti-spyware software.[31]

In addition to these monitoring products like SpectorSoft, which take the form of software, a variety of companies make hardware keyboard logging devices. These devices are generally small plugs that sit between the USB plug of the logged keyboard and the USB plug of the logged computer but can also be cards that sit inside the computer case. The devices record every key pressed on the keyboard and can be used to capture passwords, emails, typed documents, and any such information entered on the keyboard. Relative to software keyloggers, these devices are much easier to install given physical access and are impossible to detect via anti-virus software. The only way to detect some of these devices is physically to open the computer case.[32]

The most intrusive and prevalent example of client-side surveillance software is the collection of various bots, viruses, worms, trojans, and other malware that infect a

significant chunk all Internet connected computers. As a whole, the set of malware-infected computers have the ability to collect all of the client data of hundreds of millions of computers, though in practice data collected is usually limited to obviously profitable data like credit card numbers.[33] Most of these infected computers are organized into large botnets—networks of infected computers remotely controlled by a single entity. The Conficker botnet, one of the biggest currently, now controls several million infected computers.[34] Botnets like Conficker perform a variety of illicit activities including sending spam, committing click fraud, subjecting servers to denial of service attacks, and stealing financial information. One recent study of spam distribution determined that the Storm botnet was responsible for twenty percent of all spam sent in the first quarter of 2008, and Storm was just the biggest of many botnets at the time.[35]

The direct impact of most of these activities on any given infected user is usually relatively small: outgoing spam only costs the user bandwidth, credit card theft is generally insured by the credit card company, and click fraud costs a user nothing directly. But the potential for greater abuse of personal data, both individually and collectively, is difficult to overstate given the vast number of malware-infected computers, the complete access of the malware to the infected computers' data, and the increasing sophistication of the criminal organizations that run them.[36] A recent report by two of this volume's co-editors, Ron Deibert and Rafal Rohozinski, and their respective teams, on the use of a small botnet to surveil a wide variety of embassies and other highly sensitive sites in southeast Asia, demonstrates the potential for harm represented by the botnet surveillance. The Information Warfare Monitor found clear evidence that the botnet, which they call GhostNet, had wide ranging abilities on the client: from copying locally stored files to watching the physical spaces through the webcams.[37] The study presented only circumstantial evidence pointing to Chinese involvement—the servers commanding the botnet were mostly located in China, and the infected sites were all of high, regional value to China. But the documentation of this particular botnet demonstrates that it would be straightforward for China or another state (or private) actor to perform wide-scale client-side surveillance through a botnet.

To protect oneself from viruses, bots, worms, and other such malware, most experts recommend installing anti-virus systems on all client computers. But this anti-virus software is itself highly intrusive, operating at the most fundamental levels of the operating system, incurring significant performance penalties, and attempting to avoid the notice of malware (which is itself trying to detect the anti-virus systems to disable them). Anti-virus tools have the capability to do the same sorts of harm that a piece of malware can do, including both stealing data from and disabling the host computer. Trust in Symantec and the other anti-virus vendors not to snoop or harm the computer is mostly well founded, but fake anti-virus systems have now become one of the most common types of malware precisely because of the need to trust anti-virus systems and the difficulty of determining which anti-virus systems to trust.

Various other sorts of actors surveil users through their clients as well. ComScore is one of the biggest of several companies that collect data about Internet users for market research. It collects the entire Web browsing stream, including encrypted requests, from the 2 million members of its worldwide ''consumer panel.'' ComScore connects its online data with a variety of sources of off-line data, including supermarket purchases and automobile registrations. ComScore has admitted to using its collected data to log in to its members' online banking accounts to verify reported incomes.[38] ComScore recruits these panel members from a wide variety of countries, including many from Europe and Asia, through a combination of sweepstakes, network performance improvement tools, claims of antimalware protection, and (according to ComScore) a sincere desire by panel members to improve the efficiency of the Internet. ComScore discloses to the panel members that the software is monitoring their Web browsing activities, but it also keeps a strong separation between the company itself and the operations that collect the data—currently OpinionSquare and PermissionResearch—by not directly naming the tools or the organizations that operate them anywhere on ComScore.com or even in its SEC annual report filing. And it has had to recreate those operations at least once to evade detection by antispyware tools.[39] ComScore sells access to these data, estimated by ComScore at 28 terabytes collected per month in 2007, as market research to many of the largest companies in the world. The U.S. Privacy Act of 1974 prohibits the U.S. federal law enforcement and other government agencies from importing data from ComScore (or LexisNexis or other private database) en masse, but the law does allow the agencies to perform queries through ComScore or other private data sources about specific people.[40]

Governments have various levels of access to data collected through anti-virus software market research, and malware. Some governments allegedly also use their own client software to collect data directly. Direct evidence of government client surveillance is rare, but examples occasionally pop up. For instance, the U.S. Drug Enforcement Agency (DEA) has been documented as installing a keylogger on a suspect's machine to capture the encryption keys necessary to read the suspect's PGP-encrypted email.[41] And we know major anti-virus companies have complied with court orders to ignore such U.S. government spyware.[42] There is strong evidence that the German police are aggressively pursuing the use of client-side software to tap calls on Skype.[43] In Denmark, parliament approved a law that explicitly gives law enforcement agencies the authority to install keylogging software on a suspect's computer.[44] And, thanks to researchers at the Citizen Lab, the world knows that a Chinese version of Skype was logging sensitive messages to servers as mandated by the Chinese government.[45] The software used by government agencies for surveillance in all these examples is functionally indistinguishable from client malware—the whole point of the software is to collect data from the subject without knowledge or consent.

Client-side surveillance provides the most complete access to user data in comparison to network or server surveillance. Every bit of data sent, received, viewed, played,

or typed on a computer is vulnerable to client-side surveillance, though most client surveillance tools collect all possible data. Malware mostly targets various sorts of directly profitable data, including e-mail addresses and bank account information. The most sophisticated anti-virus tools monitor all data stored on and transmitted to the computer, checking the data for malware signatures, but not keyboard or screen activity. Market research tools like ComScore typically monitor all network traffic, whether encrypted or not, but not stored data or keyboard or screen activity. Workplace and family monitoring tools usually monitor keystrokes and periodic screenshots of the activity on the computer screen but not stored or network data. The GhostNet report demonstrated that active malware is even capable of activating and recording the webcams and microphones of infected computers.

Indeed, the biggest problem for client surveillance tools is often dealing with the sheer amount of data. For example, even one screenshot a minute on a single computer can generate a daunting amount of data. This problem is magnified when applied over a large set of monitored clients. Botnets (networks of malware-infected computers controlled from a single point) only search for a limited set of data, like credit card numbers, which they can easily sell, presumably because of the difficulty (and therefore unprofitability) of sifting through the vast trove of other sorts of data on infected computers. Likewise, a primary challenge of corporate anti-virus systems that must manage entire networks of clients is to manage the resulting flood of data about infections and vulnerabilities in a network of clients.

Nonetheless, any client-side program has at least the *potential* to access every sort of data that resides on or passes through the computer. So, a keylogger may only monitor keystrokes, but that restriction is mostly the choice of the tool (and its developers) once it has been installed. Even non-surveillance-oriented programs (screen savers, games, chat programs, and so on) potentially have complete access to data once they have been installed. Most computers try to make it difficult for an arbitrary program to take over a computer, but a constant stream of vulnerabilities gives client programs access to the entire computer. And this same level of access applies to most hardware devices installed on the computer as well. Even devices that do not directly have the ability to access a shared bus or run a driver may have the ability to infect clients with malware, as shown by cases like virus-carrying digital photo frames.[46]

Unlike server and network surveillance, client surveillance is always theoretically detectable. Any change in the client behavior (whether processing data, storing it, or sending it over a network) requires some detectable change to the client. In practice, there is a long history of surveillance tools using increasingly sophisticated methods to hide themselves, including through rootkits that embed themselves into the deepest layers of the client operating system. But even with these sophisticated methods, there are always small changes in behavior that at least theoretically make the tools detectable. But detecting these small changes in the large number of malware, spyware, and other surveillance tools is beyond the capabilities of even the most sophisticated user.

Detection of the most advanced client surveillance tools requires the use of some other monitoring tool, such as an anti-virus system, designed specifically for this purpose. The latest versions of malware and other surveillance tools change themselves constantly to avoid detection even by sophisticated anti-virus systems. This cycle of increasingly sophisticated detection and evasion requires constant monitoring of every networked client, by anti-virus tools, by mal- or spyware, or often by both.

In a strict sense, the EU data retention directive applies directly only to network and server monitoring. But the routers that will be used in the implementation of the directive to collect network data are client devices themselves. As such, they are vulnerable to a stack of hardware, operating system, and applications just like any other client. Any actor within that network may potentially have access to the data, so adding the monitoring box to the ISP network potentially adds all those actors to the network trusted with the client data. Most of those actors (including hardware manufacturers, operating system developers, application developers) have access to all the data potentially collected through network surveillance and not just the legally monitored data, so we have to consider the flow of both the actual and potential data mandated by the directive through this network of trust around the monitoring tools. The sophistication of client surveillance tools at both collecting data and hiding themselves from detection demonstrates the possibility of an attacker installing such code undetected on a data retention tool. The number of well-publicized, active exploits against a range of routers (not to mention counterfeit routers) means that the risk of this occurring is high.

Conclusion

This chapter provides a typology of the different sorts of Internet surveillance through cases about Internet surveillance tools. For each set of cases, it is important to focus on the *actual* and *potential data* monitored and *networks of trust* through which the data necessarily flow. This typology is intended to serve as a starting point for analysis of the steady stream of cases about Internet surveillance; for the analysis of those cases not only from a technical frame but also from social, political, legal, and other frames; and for the application of the resulting road map to specific questions about surveillance that arise over time. As new stories about surveillance emerge, this road map can provide a context for determining whether and in what ways those stories tell us anything new about Internet surveillance. For example, one might ask whether recently reported iPhone viruses represent a new sort of surveillance or how the monitoring required for Comcast's BitTorrent throttling (described in the U.S.-Canada Overview presented later in this book) compares to existing examples of surveillance.

This typology also provides a frame for considering the impact of the EU data retention directive, which is likely to have important effects beyond its explicit scope. These

effects apply both to the data potentially accessible to monitoring and to the way users will relate to the networks of trusted actors who have access to these data. As the cases about client tools show, surveillance is both widespread and very difficult to detect on a range of client devices, including the routers and other computers necessary to implement the EU directive. As a result, an unintended outcome of the directive will be that both the mandated data (senders, recipients, and time stamps) and the potentially collected data (the content of the whole data stream) will be exposed to potential access by a whole network of new actors. The cases about server tools and surveillance show that the directive's requirements, when applied to server-side companies like Google, will have the effect of increasing the data stored by those companies and thereby further pushing the already strained trust relationship between users and servers. Similar mandates related to data retention should be viewed in a similar context of growing Internet surveillance practices in the OSCE member states and elsewhere around the world.

Finally, the cases about network surveillance show that many users, including many of the serious criminals that the directive is meant to track, have an incentive to choose to use available rerouting methods to avoid monitoring. The increased use of rerouting proxies will both pose a privacy risk to those users and potentially route a significant portion of EU Internet traffic through countries unfriendly (at least in terms of data sharing) to the EU. As a result, the intended purpose of the EU data retention directive may be thwarted in dangerous ways. The unintended consequences of the EU data retention directive are likely to prove costly.

Notes

1. European Union, *Directive 2006/24/EC of the European Parliament and of the Council of 15 March 2006 on the retention of data generated or processed in connection with the provision of publicly available electronic communications services or of public communications networks and amending Directive 2002/58/EC*, 2006.

2. Federal Communications Commission, *Order FCC 06–56* (Federal Communications Commission, May 12, 2006).

3. John Markoff and Scott Shane, ''Documents Show Link between AT&T and Agency in Eavesdropping Case,'' *New York Times*, April 13, 2006, http://www.nytimes.com/2006/04/13/us/nationalspecial3/13nsa.html.

4. Ibid.

5. Sara Sundelius, ''Sweden Adopts Controversial Law to Allow Secret Tapping of E-mails, Phone Calls,'' *International Herald Tribune*, June 18, 2008; DW-World Staff, ''Swedish Government Clears Hurdles to Pass Surveillance Bill,'' *DW-World.de*, June 19, 2008, http://www.dw-world.de/dw/article/0,2144,3421627,00.html.

6. Frank Rizzo, ''thinkbroadband: sysip.net (BT and 121Media),'' discussion forum, thinkbroadband.com, July 2, 2007, http://forums.thinkbroadband.com/bt/3047764-sysipnet-bt-and-121media.html; Filippo Spike Morelli, ''To Own, To Be Owned, or What Else? BT and Its Proxies,'' *SpikeLab.org*, July 9, 2007, http://www.spikelab.org/blog/btProxyHorror.html; Ryan Naraine, ''Spyware, Rootkit Maker Stops Distribution,'' *eWeek.com*, May 10, 2006, http://www.eweek.com/c/a/Security/Spyware-Rootkit-Maker-Stops-Distribution/.

7. Chris Williams, ''ISP Data Deal with Former 'Spyware' Boss Triggers Privacy Fears,'' *The Register*, February 25, 2008, http://www.theregister.co.uk/2008/02/25/phorm_isp_advertising/.

8. Phorm, Inc., ''BT PLC, TalkTalk and Virgin Media Inc Confirm Exclusive Agreements with Phorm,'' phorm.com, February 14, 2008, http://cyber.law.harvard.edu/pubrelease/accesscontrolled/phorm-launch-agrement.html.

9. Chris Williams, ''ISP Data Deal with Former 'Spyware' Boss Triggers Privacy Fears,'' *The Register*, February 25, 2008, http://www.theregister.co.uk/2008/02/25/phorm_isp_advertising/; Chris Williams, ''BT Admits Misleading Customers over Phorm Experiments,'' *The Register*, March 17, 2008, http://www.theregister.co.uk/2008/03/17/bt_phorm_lies/print.html.

10. Phorm, Inc., ''Phorm: No Personal Information,'' phorm.com, http://privacy.phorm.com/no_personal_info.php.

11. Philip Stafford, ''BT to Begin Further Trials of Ad Technology,'' *FT.com*, July 17, 2008, http://www.ft.com/cms/s/0/34c59420-5356-11dd-8dd2-000077b07658.html.

12. Ethan Zuckerman, ''How a Pakistani ISP Briefly Shut Down YouTube,'' My Heart's in Accra, February 25, 2008, http://www.ethanzuckerman.com/blog/2008/02/25/how-a-pakistani-isp-briefly-shut-down-youtube/.

13. Sylvie Barak, ''Pakistan Becomes VPN Routing Hot-spot,'' *The Inquirer*, February 26, 2008, http://www.theinquirer.net/gb/inquirer/news/2008/02/26/pakistan-becomes-vpn-routing.

14. AnchorFree, Inc., ''AnchorFree History,'' anchorfree.com, 2008, http://anchorfree.com/about/history/.

15. Berkman Center for Internet and Society, ''Mapping Local Internet Control,'' http://cyber.law.harvard.edu/netmaps.

16. Cooperative Association for Internet Data Analysis, *IPv4 Internet Topology Report as Internet Graph* (Cooperative Association for Internet Data Analysis, August 1, 2007), http://www.caida.org/research/topology/as_core_network/.

17. Jacco Tunnissen, ''Global Internet Exchange Points / BGP Peering Points / IXP,'' Bgp4.as, 2008, http://www.bgp4.as/internet-exchanges.

18. Josh Karlin, Stephanie Forrest, and Jennifer Rexford, ''Nation-State Routing: Censorship, Wiretapping, and BGP,'' arXiv.org (March 18, 2009), http://arxiv.org/abs/0903.3218.

19. Ryan Naraine, ''Spyware Floods in through BitTorrent,'' *eWeek.com*, June 15, 205, http://www.eweek.com/index2.php?option=content&task=view&id=9664; Dan Ilett, ''CA Slaps Spyware

Label on Kazaa,'' *CNET News.com*, http://news.cnet.com/CA-slaps-spyware-label-on-Kazaa/2100-1025_3-5467539.html.

20. Gary T. Marx, ''Soft Surveillance: The Growth of Mandatory Volunteerism in Collecting Personal Information—'Hey Buddy Can You Spare a DNA?,''' in *Surveillance and Security: Technological Power and Politics in Everyday Life*, ed. Torin Monahan (New York: Routledge, 2006), 37–56; Daniel J. Solove, *The Digital Person* (New York: New York University Press, 2004).

21. Google Inc., ''Google Search Privacy: Plain and Simple,'' August 8, 2007, http://www.youtube.com/watch?v=kLgJYBRzUXY.

22. Michael Barbaro and Tom Zeller, Jr., ''A Face Is Exposed for AOL Searcher No. 4417749,'' *New York Times*, August 9, 2006, http://www.nytimes.com/2006/08/09/technology/09aol.html; Arvind Narayanan and Vitaly Shmatikov, ''How to Break Anonymity of the Netflix Prize Dataset,'' arXiv.org, October 18, 2006, http://arxiv.org/abs/cs/0610105.

23. Miguel Helft, ''Google Told to Turn Over User Data of YouTube,'' *The New York Times*, July 4, 2008, http://www.nytimes.com/2008/07/04/technology/04youtube.html.

24. Attributor Corporation, ''Get Your Fair Share of the Ad Network Pie,'' March 30, 2008, http://www.attributor.com/blog/get-your-fair-share-of-the-ad-network-pie/.

25. Websense Security Labs, *Websense Security Labs: State of Internet Security Q1–Q2, 2008* (Websense Security Labs, 2008), http://www.websense.com/securitylabs/docs/WSL_Report_1H08_FINAL.pdf.

26. Google Inc., ''Google Log Retention Policy FAQ,'' March 14, 2007, http://www.seroundtable.com/google_log_retention_policy_faq.pdf; Kevin J. O'Brien and Thomas Crampton, ''E.U. Probes Google Over Data Retention Policy,'' *New York Times*, May 26, 2007, http://www.nytimes.com/2007/05/26/business/26google.html.

27. European Commission Article 29 Data Protection Working Party, ''Opinion 1/2008 on Data Protection Issues Related to Search Engines,'' April 4, 2008, ec.europa.eu/justice_home/fsj/privacy/docs/wpdocs/2008/wp148_en.pdf.

28. SpectorSoft, ''Spector Pro 2008 for Windows,'' http://www.spectorsoft.com/products/SpectorPro_Windows/.

29. Camille Calman, ''Spy v. Spouse: Regulating Surveillance Software on Shared Marital Computers,'' *Columbia Law Review 105*, no. 2097 (2005), http://www.columbialawreview.org/pdf/Calman-Web.pdf.

30. Ellen Messmer, ''Spouse-vs.-Spouse Cyberspying Dangerous, Possibly Illegal,'' *Network World*, August 16, 2007, http://www.networkworld.com/news/2007/081607-spouse.html?page=2.

31. Computer Associates, ''eTrust PestPatrol, EBlaster,'' August 16, 2004, http://www.ca.com/securityadvisor/pest/pest.aspx?id=55479.

32. KeyCarbon, ''Record Keystrokes on Laptop Keyboard with Tiny 50 × 60mm Device. Software Free—Plug Device into Base of Laptop, It Begins to Record Immediately,'' keycarbon.com, http://www.keycarbon.com/products/keycarbon_laptop/overview/.

33. Tim Weber, ''Criminals 'May Overwhelm the Web','' *BBC News*, January 25, 2007, http://news.bbc.co.uk/1/hi/business/6298641.stm; PandaLabs, *Quarterly Reports PandaLabs* (April—June 2008), (PandaLabs, July 1, 2008).

34. John Markoff, ''Worm Infects Millions of Computers Worldwide,'' *New York Times*, January 23, 2009, http://www.nytimes.com/2009/01/23/technology/internet/23worm.html.

35. MessageLabs, *MessageLabs Intelligence: Q1/March 2008:* ''One Fifth of All Spam Springs from Storm Botnet,'' (MessageLabs, April 1, 2008).

36. Nicholas Ianelli and Aaron Hackworth, *Botnets as a Vehicle for Online Crime*, (CERT Coordination Center, December 1, 2005), http://www.cert.org/archive/pdf/Botnets.pdf.

37. Information Warfare Monitor, ''Tracking GhostNet: Investigating a Cyber Espionage Network'' (Citizen Lab/the SecDev Group, March 29, 2009), http://www.tracking-ghost.net.

38. ComScore, Inc., ''ComScore Methodology,'' ComScore.com, http://cyber.law.harvard.edu/pubrelease/accesscontrolled/comscore-tech.html; Evan Hansen, ''Net Privacy and the Myth of Self-regulation,'' *CNET News.com*, October 16, 2001, http://www.news.com/Net-privacy-and-the-myth-of-self-regulation/2010-1071_3-281580.html.

39. ComScore, Inc., *SEC Filing, 10-K Annual Report*, 2008; Stefanie Olsen, ''ComScore: Spyware or 'Researchware'?,'' *CNET News.com*, December 20, 2004, http://news.cnet.com/ComScore-Spyware-or-researchware/2100-1032_3-5494004.html.

40. Electronic Privacy Information Center, ''EPIC Privacy Act of 1974 Page 5 U.S.C. § 552a,'' epic.org, August 26, 2003, http://epic.org/privacy/1974act/.

41. Declan McCullagh, ''Feds Use Keylogger to Thwart PGP, Hushmail,'' *CNET News.com*, July 10, 2007, http://news.cnet.com/8301-10784_3-9741357-7.html.

42. Declan McCullagh and Anne Broache, ''Will Security Firms Detect Police Spyware?'' *CNET News.com*, July 17, 2007, http://news.cnet.com/Will-security-firms-detect-police-spyware/2100-7348_3-6197020.html.

43. Louis Charboneau, ''Skype Encryption Stumps German Police,'' *Global and Mail*, November 22, 2007; Kim Zetter, ''Leaked Documents Show German Police Attempting to Hack Skype,'' *Wired.com*, January 29, 2008, http://blog.wired.com/27bstroke6/2008/01/leaked-document.html.

44. Privacy International, *PHR2006—Kingdom of Denmark* (Privacy International, December 18, 2007), http://www.privacyinternational.org/article.shtml?cmd%5B347%5D=x-347-559545.

45. Nart Villeneuve, ''Breaching Trust: An Analysis of Surveillance and Security Practices on China's TOM-Skype Platform'' (Information Warfare Monitor/ONI Asia, October 1, 2008), http://www.nartv.org/mirror/breachingtrust.pdf; Josh Silverman, ''Answers to Some Commonly Asked Questions about the Chinese Privacy Breach,'' Skype Blogs, October 4, 2008, http://share.skype.com/sites/en/2008/10/answers_to_some_commonly_asked.html.

46. David Goldsmith, ''Digital Hitchhikers,'' SANS Internet Storm Center, December 25, 2007, http://isc.sans.org/diary.html?storyid=3787.