Submission to the Privacy and Civil Liberties Oversight Board: Technical Issues Raised by the §215 and §702 Surveillance Programs

Steven M. Bellovin
https://www.cs.columbia.edu/~smb
Department of Computer Science, Columbia University

July 31, 2013

The recent disclosures of several surveillance programs under Section 215 of the USA PATRIOT Act and Section 702 of Foreign Intelligence Surveillance Act have raised a number of difficult legal and policy questions. In addition to the obvious legal questions, the various technical choices made by raises questions of their own. In this note, I will attempt to describe some of these technical issues and the policy choices they raise.

I should point out that I am not proposing, suggesting, or implying any particular answers to these questions. However, policy-makers should be aware of the implications of the underlying technologies.

Phone Call Databases and Machine Learning

One of the most controversial of the recent revelations is that the government is apparently collecting all telephone call records,[1] even for purely domestic calls between U.S. persons not implicated in terrorism. Legal issues aside,[2] there are several possible technical explanations for this. All raise policy issues.

The simplest answer, and one that has been cited in the press, is that phone companies do not retain data for long enough. An obvious answer is to mandate a longer retention period, much as the EU has done. This, however, has been controversial in privacy circles. Indeed, the EU's own data protection supervisor has expressed serious concerns about the EU's Data Retention Directive.[3] He also cited a letter from a large coalition of privacy and civil liberties organizations that oppose the directive.[4] While the existence of a controversy does not itself prove something wrong, it does suggest that this solution itself requires careful attention. If nothing else, in the U.S. there are very few laws regulating what the private sector can do with information they've collected; to many in the privacy community, requiring companies to record and retain more data is an invitation to more mischief.[5]

A second possible technical explanation is the lack of proper indexing of telephone company records. Their databases are designed to handle the sorts of queries they themselves need to make; these are not necessarily the same queries that an intelligence or law enforcement agency would make. Consider, for example, an ordinary phone

[1] Charlie Savage and Edward Wyatt, "U.S. Is Secretly Collecting Records of Verizon Calls," *New York Times* (June 5, 2013), https://www.nytimes.com/2013/06/06/us/us-secretly-collecting-logs-of-business-calls.html.

[2] There are, of course, many legal issues. For a discussion of the meaning of "relevant" and why it is legally important, see Jennifer Valentino-DeVries and Siobhan Gorman, "Secret Court's Redefinition of 'Relevant' Empowered Vast NSA Data-Gathering," *Wall Street Journal* (July 7, 2013), http://online.wsj.com/article/SB10001424127887323873904578571893758853344.html.

[3] Peter Hustinx, *Opinion of the European Data Protection Supervisor on the Evaluation report from the Commission to the Council and the European Parliament on the Data Retention Directive (Directive 2006/24/EC)*, 2011, http://www.edps.europa.eu/EDPSWEB/webdav/site/mySite/shared/Documents/Consultation/Opinions/2011/11-05-30_Evaluation_Report_DRD_EN.pdf.

[4] See http://www.vorratsdatenspeicherung.de/images/DRletter_Malmstroem.pdf.

[5] See, for example, the new data collection and advertising initiatives by cell phone companies described in Kashmir Hill, "How To Opt Out Of AT&T's Plan To Sell Everything It Knows About You And Your Smartphone Use," *Forbes* (July 3, 2013), http://www.forbes.com/sites/kashmirhill/2013/07/03/how-to-opt-out-of-atts-plan-to-sell-everything-it-knows-about-you-and-your-smartphone-use/.

bill for a land-line telephone. It shows the numbers called by the account-holder; it does not show incoming calls. Consequently, it is entirely possible that a query of the form "who has called so-and-so, a known terrorist?" cannot be answered efficiently by the phone company. If true, it is a strong argument for the data collection. The obvious alternative, requiring phone companies to be able to answer all such questions efficiently, raises issues of not just cost (who would pay for the extra programming and computer resources necessary?) but also of disclosure of sensitive information: that an intelligence agency regards some forms of queries as useful would be seen as revealing crucial information about our intelligence methods to our adversaries.

One approach to solving this last problem is a technology known as "encrypted search". This technology, long of interest to the intelligence community,[6] has been advocated as a way to permit the NSA to search phone company-resident databases without disclosing the search to the company or needing a local copy of the database.[7] Although encrypted search is generally considered a privacy-enhancing technology, just whose privacy is protected depends on how it is deployed. Using it as described here protects the queries, which is not a bad thing; however, that by itself does not protect the privacy of American citizens if the same queries are made. The advantage is that it obviates the need to transfer the entire database.

The third possible explanation raises the most difficult policy challenge of all: that the NSA et al. are using various forms of data mining technology to detect terrorists. Quite conceivably, their algorithms *require* data on innocent Americans in order to function properly. This requirement poses a direct challenge to the Fourth Amendment's requirement that "unreasonable" searches require a warrant "particularly describing the place to be searched, and the persons or things to be seized." Here, there is no particularity.

Data mining and machine learning are extremely important research areas today. They are behind the algorithms used by companies like Amazon and Netflix to make recommendations to their customers. These algorithms work by correlations, and do not rely on or presume causality.[8]

In one experiment, for example, researchers found that they could accurately predict a subject's ethnicity by monitoring cell phone location data.[9] It is not obvious why this should be possible; nevertheless, it works.

Other algorithms can be used to identify individuals based on their behavior patterns. In a database with hundreds of terabytes of Call Detail Records (CDRs), Cortes et al. were able to identify people who had changed their phone numbers; they used only the pattern of numbers that subjects called.[10] It is clear that this can work in principle; it was not obvious that it could be done effectively or efficiently in practice.

Another class of algorithms can detect unusual or significant patterns. Consider the diagram shown in Figure 1,[11] derived solely from membership lists of seven different organizations in pre-Revolutionary War Boston. There was no reliance on speeches, writings, etc. Using just this data, a very simple algorithm was able to identify Paul Revere as a very significant figure.

Many algorithms rely explicitly on the existence of a large amount of "normal" data, i.e., that belonging to people not of interest, in order to determine what is "abnormal". Consider the following thought experiment: is there a calling pattern peculiar to so-called "burner phones" (prepaid phones used briefly by people trying to avoid detection)? If so, might some sort of analysis of all calling records pick out such phones? I stress that I have no idea if this is possible or not; if it is, though, the analysis cannot be done without a very comprehensive database, one comprised primarily of records of no intrinsic national security interest whatsoever.

The fact these algorithms can learn useful things from improbable inputs makes it much harder to define "minimization" in this context. FISA requires minimization procedures that are "reasonably designed in light of the purpose

[6] See "Security and Privacy Assurance Research (SPAR) Program Broad Agency Announcement", December 29, 2010, `http://www.iarpa.gov/Programs/sso/SPAR/solicitation_spar.html`. Note: I am an active researcher in this field, and have received grants under this and other programs.

[7] See Siobhan Gorman, "Pressure Builds for Data-Sweep Alternative," *Wall Street Journal* (July 19, 2013), `http://online.wsj.com/article/SB10001424127887324448104578615881436052760.html`.

[8] Correlations can be extremely powerful and useful. Perhaps the best-known use is credit scoring: all of the major credit bureaus calculate scores based on correlations with things like job title, income, etc., without regard to *why* these predict a borrower's likelihood of repayment. That said, blindly following such models without paying attention to the larger context can lead to practices such as "red-lining".

[9] Yaniv Altshuler et al., "Incremental Learning with Accuracy Prediction of Social and Individual Properties from Mobile-Phone Data," in *IEEE Conference on Social Computing* (2012).

[10] C. Cortes, D. Pregibon, and C. Volinsky, "Communities of Interest," in *Proceedings of IDA 2001 — Intelligent Data Analysis* (2001), `http://citeseer.ist.psu.edu/cortes01communities.html`.

[11] This is from a blog posting by Kieran Healy, *Using Metadata to Find Paul Revere*, June 9, 2013, used with permission. `http://kieranhealy.org/blog/archives/2013/06/09/using-metadata-to-find-paul-revere/`.

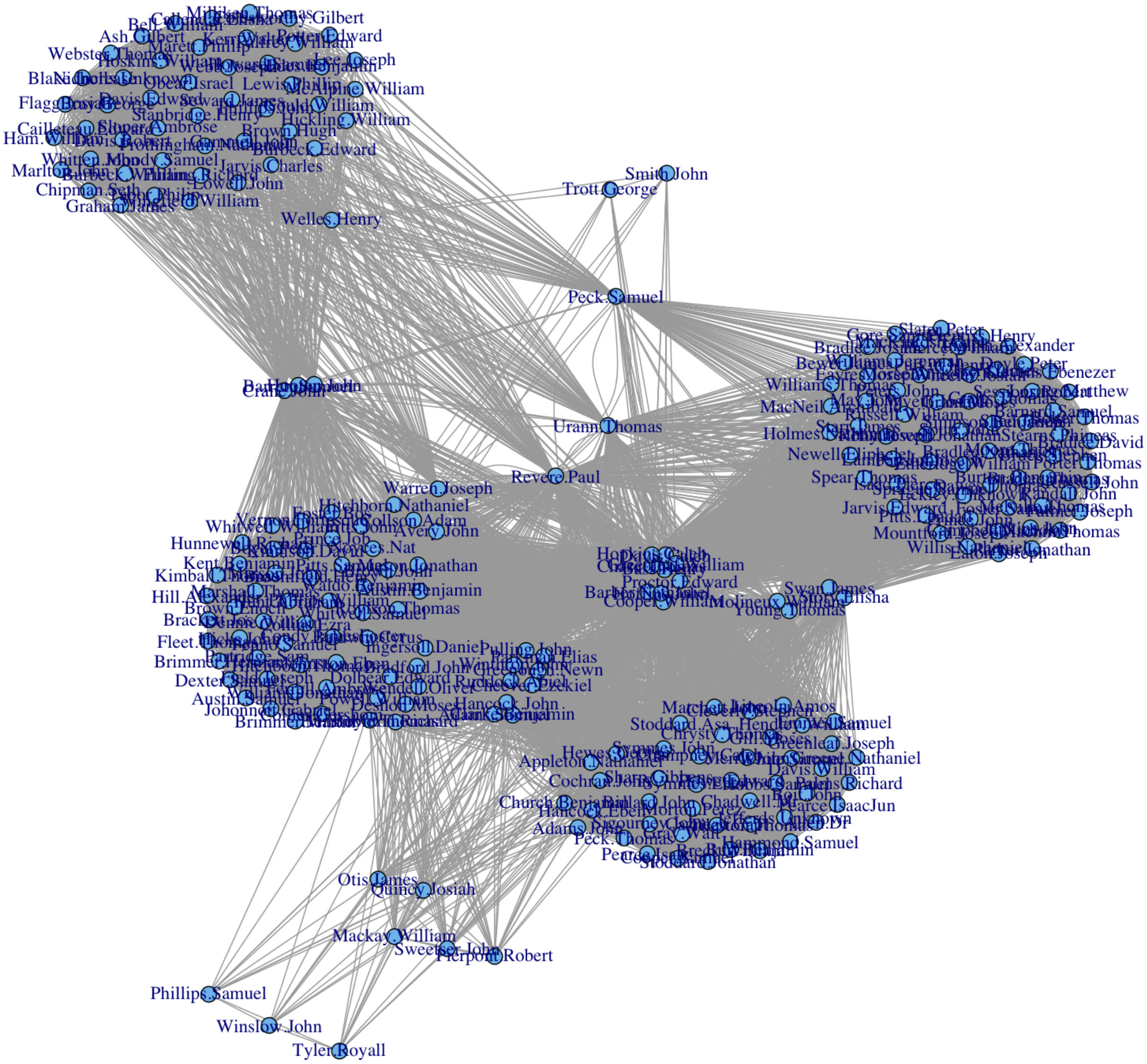

Figure 1: A diagram derived solely from membership lists of assorted pre-Revolutionary War organizations. Note how Paul Revere's name, in the center, jumps out the most connected person. (Diagram taken with permission from `http://kieranhealy.org/files/misc/revere-network-reduced.png`.)

and technique of the particular surveillance, to minimize the acquisition and retention, and prohibit the dissemination, of nonpublicly available information concerning unconsenting United States persons consistent with the need of the United States to obtain, produce, and disseminate foreign intelligence information".[12] However, given the way data mining works is *any* information not relevant? The law does permit retention of information if it is "consistent" with the need to produce foreign intelligence; do these algorithms mean that acquisition and retention of anything can be justified?[13]

It should be stressed that metadata is often far more revealing than content. Green has noted that "while encryption may hide what you say, it doesn't necessarily hide *who you're talking to*" (emphasis in the original).[14] In fact, hiding metadata can be extremely difficult. Metadata queries can be retrospective rather than prospective. You need to know in advance if you wish to tap a phone call; even after it has taken place, though, you can find out whom the parties were. It can also reveal surprising things. One experiment in network connectivity discovered that two researchers in a particular lab were meeting nightly; their relationship had not been known to their colleagues.[15] Location, it turns out, is an extremely powerful predictor. De Montjoye et al.[16] found that as few as four randomly-chosen time-space location points were enough to uniquely identify 95% of the subjects in their database.[17] They showed that even very coarse location data, what you would get just from tower location, was quite sufficient.

Other information that can be gleaned from cell phone usage patterns is even more surprising. Researchers at MIT have found that they can do psychological assessments based on how, when, and where phones are used:[18]

> Yves-Alexandre de Montjoye, an applied mathematician at the MIT Media Lab who worked on the study, hopes it will be possible someday to run psychological tests on entire cities or countries, simply by analyzing the citizens' phone records.
>
> "Our main goal is both to show that it is possible . . . and to see if this method can be applied on a larger scale," Montjoye said.
>
> For example, scientists might use phone records someday to estimate how many Americans are prone to depression, with no need to conduct millions of individual tests.

Current U.S. law regarding metadata (see the following section for a discussion of the difficulty of defining the term) was based on the assumption that individual data items in some sense stood alone. Here, though, we are dealing with data in the large. It may be necessary to write new statutes that specify how to treat these large datasets, and how to redefine concepts like "particularity", "relevance", and "minimization".

[12] 50 U.S.C. 1801(h)(1).

[13] For a longer discussion of mimization in this context, see Susan Landau, "Susan Landau on Minimization," *Lawfare (blog)* (June 18, 2013), http://www.lawfareblog.com/2013/06/susan-landau-on-minimization/.

[14] Matthew Green, "How to 'Backdoor' an Encryption App," *A Few Thoughts on Cryptographic Engineering (blog)* (June 17, 2013), http://blog.cryptographyengineering.com/2013/06/how-to-backdoor-encryption-app.html.

[15] Steven M. Bellovin et al., "Risking Communications Security: Potential Hazards of the "Protect America Act"," *IEEE Security & Privacy* 6, no. 1 (January 2008): 24–33, https://www.cs.columbia.edu/~smb/papers/j1lanFIN.pdf.

[16] Yves-Alexandre de Montjoye et al., "Unique in the Crowd: The privacy bounds of human mobility," *Scientific Reports* 3 (2013).

[17] On a number of occasions, the NSA has stated categorically that it is not collecting location data, e.g., "In addition, as we have repeatedly and publicly said, we are not collecting cell site location information under this program." (Letter from DNI James Clapper to Sen. Wyden, July 26, 2013, http://www.wyden.senate.gov/download/?id=285dc9e7-195a-4467-b0fe-caa857fc4e0d.) Numerous critics have pointed out that the phrase "as part of this program" occurs in all responses from the intelligence community. It is unclear from the public record if that is significant or not, i.e., if the NSA is indeed collecting location data but as part of a different program.

The leaked Verizon Business order required the company to provide CDRs; a number of sources state or imply that CDRs include cell site identifiers. For example, slides 23–24 of a slide deck from Verizon Wireless (Verizon Wireless Law Enforcement Resource Team (LERT), available at http://cryptome.org/isp-spy/verizon-spy.pdf) shows that their CDRs include cell site. Daniel and Daniel suggest use of cell tower data from CDRs as location evidence (Larry Daniel and Lars Daniel, *Digital Forensics for Legal Professionals: Understanding Digital Evidence From The Warrant To The Courtroom* (Waltham, MA: Syngress/Elsevier, 2012), p. 232). "Call detail information" is defined for regulatory purposes as "Any information that pertains to the transmission of specific telephone calls, including, for outbound calls, the number called, and the time, location, or duration of any call and, for inbound calls, the number from which the call was placed, and the time, location, or duration of any call" (47 CFR 64.2003). Note that it is easy to map from a cell site identifier to a physical location. Taken together, these sources suggest that if the NSA does not have a proxy for location data, it is because they are deliberately deleting it from the data they receive from wireless carriers.

[18] Hiawatha Bray, "Cellphone Data Mined to Create Personal Profiles," *Boston Globe* (July 8, 2013), http://www.bostonglobe.com/business/2013/07/07/your-cellphone-yourself/eSvTK1UCqNOE7D4qbAcWPL/story.html.

Defining “Metadata”

A lot of the current controversy has concerned collection of “metadata”. There is little information, though, on just what information falls under this label. Precisely how it is defined, though, has important legal and policy implications.

Kerr has noted that much of the collection may be justified under the third party doctrine.[19] How, though, to draw the line? One approach is to differentiate data given to a third party for its own use, versus information passed along for conveyance to some ultimate recipient. This fits nicely with the holding in *Smith v. Maryland*,[20] which said that customer-dialed digits had been “given” to the phone company and hence were unprotected by the Fourth Amendment, whereas access to the actual conversation would require a search warrant.

We can approach this question from a technical perspective. Computer and communications protocols are defined in terms of different functional entities; in a strong sense, information sent *to* a functional entity along with some data intended for a later entity is metadata to that later entity. We thus can try to distinguish between “end-to-end” information—data sent by one party to some ultimate recipient—and “hop-by-hop” information, data intended or expected to be processed along the way by intermediate nodes, for their own purposes and in furtherance of the communication.

The simplest analogy is the business letter. The destination and return addresses on the envelope are intended for the postal service; the recipient of the letter has little need for it. By contrast, the same information may appear within the body of the letter; that data is “payload” to the postal service, since it does not use that information.

The same distinction applies to email. The familiar From: and To: lines we see are, technically speaking, content, not metadata; the various mail systems along the path are passing it along, not using it. However, very similar lines, in the so-called SMTP dialog, are metadata, since they are generated and interpreted by intermediate mail handlers and are not given to the ultimate recipient. This distinction is very clear from the technical standards; the former is defined by RFC 5321[21] while the latter is defined by by RFC 5322;[22] these apply to very different functional elements.[23]

It is thus tempting to use these technical standards to define what is and isn’t metadata from a legal perspective. While this is certainly a good first approach, there are some caveats. First, the exact boundary is often not obvious from a quick glance; a detailed technical assessment is necessary. Second, invisible technological changes can change the boundary. Consider speech, which would seem to be one of the clearer examples of content, because it is intended solely for the end recipient and not for the phone system. Reality is rather more complex. A tone of a certain frequency, for example, turns off “echo cancellation” on phone calls; modems emit this tone when they answer calls.[24] The phone network must therefore listen to the call—the content—to detect the tone. Another example is “TASI” (Time Assignment Speech Interpolation), a 1960s scheme that detected silence and reused the channel for another conversation; it was used to stretch very expensive bandwidth on transoceanic calls.[25] Again, the phone network is using the content of the call, rather than just the dialed digits. It is, however, rather hard to argue that this optimization, entirely invisible to the people at either end, should be used to decide that calls are not private.

Technical boundaries are often blurred for varied reasons. TCP port numbers,[26] which specify which service—

[19] Orin Kerr, “Hints and Questions About the Secret Fourth Amendment Rulings of the FISA Court,” July 7, 2013, http://www.volokh.com/2013/07/07/hints-and-questions-about-the-secret-fourth-amendment-rulings-of-the-fisa-court/.

[20] Smith v. Maryland, 442 U.S. 735 (1979).

[21] J. Klensin, *Simple Mail Transfer Protocol*, RFC 5321 (RFC Editor, October 2008), 1–95, http://www.rfc-editor.org/rfc/rfc5321.txt.

[22] P. Resnick, ed., *Internet Message Format*, RFC 5322 (RFC Editor, October 2008), 1–57, http://www.rfc-editor.org/rfc/rfc5322.txt.

[23] Trying to draw the line for email can be much harder than that. Under certain unusual circumstances, even the SMTP dialog may be end-to-end in a legal sense. Unfortunately, it is rarely possible to tell if these circumstances will apply until after the interception has taken place.

There are odder situations possible. Most people use either the POP 3 protocol (J. Myers and M. Rose, *Post Office Protocol - Version 3*, RFC 1939 (RFC Editor, May 1996), 1–23, http://www.rfc-editor.org/rfc/rfc1939.txt) or IMAP 4 (M. Crispin, *INTERNET MESSAGE ACCESS PROTOCOL - VERSION 4rev1*, RFC 3501 (RFC Editor, March 2003), 1–108, http://www.rfc-editor.org/rfc/rfc3501.txt) to retrieve their email from their ISP. In the latter, the email headers can be treated as metadata by the server: they can be retrieved separately, searched, etc. POP 3 has limited, and optional, i.e., not universally implemented, support for retrieving headers. It is perfectly reasonable, under certain circumstances, for one person to use IMAP 4 from a smartphone but POP 3 from a laptop. Can a sender’s expectation of privacy depend on how the recipient will, in the future, read the email?

[24] A detailed explanation of why this is done is far beyond the scope of this note; see https://en.wikipedia.org/wiki/Echo_cancellation#Modems for a summary. Let it suffice to say that it is necessary to do this if modems are to work properly.

[25] See https://en.wikipedia.org/wiki/Time-assignment_speech_interpolation.

[26] J. Postel, *Transmission Control Protocol*, RFC 793 (RFC Editor, September 1981), 1–91, http://www.rfc-editor.org/rfc/

email, web browsing, etc.—an Internet connection is using are always end-to-end. However, for more than 25 years commercial routers—which operate at the Internet Protocol layer,[27] not the TCP layer—have provided the ability to filter on port numbers. This is widely used by ISPs for a variety of reasons, including spam control, preventing certain kinds of attacks, and enforcement of their terms of service. Does this widespread usage—and deliberate evasion of the functional divisions of the protocols standards—constitute a de facto redefinition of the port number as metadata?

The situation is even more complicated for wireless data. Because bandwidth is more limited, there is more reason for ISPs to tinker with the content of a connection to compress it. This can help them—they will need less spectrum to service their customer base—and it can often help customers, since they will perceive a higher effective bandwidth, at least if measured in pictures per second rather than megabits per second. Some wireless ISPs, at least, do this.[28]

Finally, there are types of information that, though metadata by most technical interpretations, are almost always end-to-end. The best example is the extra data fields in most digital pictures; these can include things like GPS location, camera and lens serial numbers, etc. This would clearly seem to be content from a legal perspective—but perhaps not in the context of a social network that let you ask "who else took a picture here?" The answer here is context-dependent.

It is clear from this analysis that metadata in the legal sense—that is, data not considered to be third party information under the Fourth Amendment—cannot be defined solely in technical terms. However, technical considerations must play a part in creating such definitions.

Enforcing Search Limits

A recent statement by DNI James R. Clapper said, in part

> By order of the FISC, the Government is prohibited from indiscriminately sifting through the telephony metadata acquired under the program. All information that is acquired under this program is subject to strict, court-imposed restrictions on review and handling. The court only allows the data to be queried when there is a reasonable suspicion, based on specific facts, that the particular basis for the query is associated with a foreign terrorist organization. Only specially cleared counterterrorism personnel specifically trained in the Court-approved procedures may even access the records.[29]

Apart from the issue of whether or not a court may, in fact, impose extra restrictions,[30] there is the purely technical issue of how such limits can be enforced. That is, assuming that such mandatory limits exist , either because a court has so ordered or because the NSA and other intelligence agencies believe that restrictions are necessary to stay within legal and constitutional strictures, are there suitable technical mechanisms used to ensure compliance?

There are two basic (and non-exclusive) approaches that may be taken. First, the database system itself may impose limits, refusing to return information in response to certain queries. Second, it can log all queries, under the assumptions that there will be a later audit of the query logs, to detect improper activity by analysts. Either of these approaches, limits or logging, can be implemented by either the database system or the analysts' own machines. All of these choices raise some difficult issues.

rfc793.txt.

[27] J. Postel, *Internet Protocol*, RFC 791 (RFC Editor, September 1981), 1–51, http://www.rfc-editor.org/rfc/rfc791.txt.

[28] See, e.g., Verizon's "Optimization Deployment—Terms and Conditions", http://support.verizonwireless.com//terms/network_optimization.html. Similar techniques have been used in the past by wired ISPs, especially during the dial-up era; see "Web images are saved as .art files sand not as .jpg or .gif files", http://help.aol.com/help/microsites/search.do?cmd=displayKC&docType=kc&externalId=220576&sliceId=1&docTypeID=DT_AOLTROUBLESHOOTING_1_1&dialogID=331712664&stateId=1%200%20885268094&radios=False.

[29] James Clapper, "DNI Statement on Recent Unauthorized Disclosures of Classified Information," June 6, 2013, http://www.dni.gov/index.php/newsroom/press-releases/191-press-releases-2013/868-dni-statement-on-recent-unauthorized-disclosures-of-classified-information.

[30] A number of legal scholars have questioned this ability. See, for example, Babak Siavoshy, "Does the Fourth Amendment regulate the NSA's analysis of call records? The FISC might have ruled it does," June 10, 2013, http://www.concurringopinions.com/archives/2013/06/does-the-fourth-amendment-regulate-the-downstream-analysis-of-call-records-by-the-nsa-the-fisc-might-have-ruled-it-does.html or Orin Kerr, "Why Does a *Terry* Standard Apply to Querying the NSA Call Records Database?", June 7, 2013, http://www.volokh.com/2013/06/07/why-does-a-terry-standard-apply-to-querying-the-nsa-call-records-database/.

The biggest problem with client-side enforcement, whether by actual limits or by logging, is that is easily bypassed, either intentionally or inadvertently. A clever but faithless person with authorized access can write custom software to do database queries, without bothering to log the queries. This is apparently what Bradley Manning did; indeed, he is alleged to have said “Weak servers, weak logging, weak physical security, weak counter-intelligence, inattentive signal analysis. . . a perfect storm.” Marcus Ranum summed up the Manning case nicely:[31]

> Then the other piece of the puzzle that I find is really interesting is the apparent inability of the people who lost the data, the original data holders, to tell what data was stolen and while it was being stolen. And this is an important message for anyone who is a CISO because it shows what can happen when your data leaks if you don’t have auditing and logging in place so that you can go back and say, “Well, OK if we do believe this guy leaked a bunch of information, what information did he actually access and when?” Of course, ideally you would get in front of that process and maybe detect the fact that somebody who really didn’t have a need to access this particular information was downloading [this information] in one fell swoop. That is kind of a red flag, I would think.

Even if all intelligence community employees are scrupulously honest and careful, there is still a potential problem: some technically trained analysts will wish to write their own query software, to do sophisticated yet legal inquiries to aid in their investigations. This is, of course, to be encouraged; however, logging can itself require difficult programming, especially if audit programs are reading the logs. Will the analysts’ custom software implement logging? Will it be done correctly?

For reasons like these, server-resident logging is often preferred. Indeed, it is mandatory precisely because of the faithless employee problem. Nevertheless, it has its own failure modes; in particular, it makes it difficult to log higher-level queries properly. An analogy here is helpful. Suppose that policies permit querying a personnel department database for the average salary of a group of employees, but bar queries for the salary of any single individual. What is “average”, though? Mean? Median? Perhaps the former is supported, i.e., is a primitive built into the database, but not the latter; both, however, are in accordance with the intent of the policy. Client-side query software can calculate the median by retrieving everyone’s salary and performing the appropriate calculations; it could then display only the final—and permissible—result. In the log files, though, it would appear as a sequence of impermissible queries; only the client side would know what was actually done with the data.[32]

A similar analysis applies to query enforcement. A clever employee can easily evade restrictions imposed by client-side software—indeed, Manning is alleged to have written his own programs, and it seems not improbable that Snowden did the same—while server-side enforcement can place unnecessary limits on legitimate activities.

Ultimately, the proper answer is likely a combination of all of these options. It is vital that auditing be done properly, though, for several reasons. One, of course, is its basic function: to catch (or deter) misbehavior. This requires “enough” auditing. A second consideration is to avoid unnecessarily burdening the vast majority of honest employees. Apart from hurting productivity and morale, overly-complex and formalistic procedures tend to elicit avoidance behavior: routine answers and click-throughs, without any real thought being devoted to the actual questions or process: “Click OK to approve this query as appropriate under 50 USC 1801 et seq. and to certify that you have read and are complying with the appropriate regulations.” Finally, due allowances must be made for the most sophisticated analysts, the ones who are doing unusual things while still complying with the law.[33]

[31] Tom Field, “Marcus Ranum on 2011 Security Outlook,” *Bank Info Security* (December 24, 2013), `http://www.bankinfosecurity.com/marcus-ranum-on-2011-security-outlook-a-3205/op-1`.

[32] Although useful as an analogy, this is in fact a poor design from a mathematical perspective. Unless there are very strict restrictions on which sets of employees’ salaries can be averaged, it is possible to manipulate the queries to learn any given person’s salary, though no single query violates the policy. This, of course, underscores the difficulty of some types of technical enforcement.

[33] Log analysis and audit is a very complex subject. Sophisticated algorithms employing *anomaly detection* can spot unusual access patterns; they can, however, also generate false alarms.

Insiders and System Administration

On many computer systems in many organizations, system administrators represent a very serious potential vulnerability. This has long been recognized in some quarters. In 1996, an NSA analyst wrote:[34]

> ~~(S-UO)~~ With system administrators, though, the situation is potentially much worse than it has ever been with communicators. In part, this is because the system administrators can so easily, so quickly, so *undetectably*, steal vast quantities of information. Communicators of the past usually sent only relatively short messages and "finished" documents, but today's system administrators can obtain full-length copies of entire reports, including draft versions, as well as informal e-mail messages, electronic calendar appointments, and a wide variety of other data. [emphasis in the original][35]

The article goes on to point out other nasties that rogue system administrators can do, including modifying assorted system files.

It is perhaps gratuitous to note that according to some published reports, Edward Snowden was a system administrator.[36] As such, he could not only bypass many security controls, he was undoubtedly far more aware than most of the system and network configuration, and what sorts of monitoring schemes were in place. Why were there not proper defensive measures?

In all fairness, this is a fiendishly difficult problem. I am unaware of *any* commercial packages that implement proper "two person control" mechanisms; indeed, the only research on the subject that I am aware of is my own.[37]

There are, of course, common work-arounds. The usual answer in the intelligence community is "compartmentalization", restricting individuals' access—and that includes system administrators—to just the types of information necessary to perform their jobs. That said, there are powerful drivers against too much isolation. One issue heavily discussed in the 9/11 Commission Report was problems with information sharing;[38] too much compartmentalization can hinder necessary sharing. Another driver is economic: more compartments may require more system administrators, which increases both cost and the number of people who must be trusted. (The NSA is considering reducing its system administration staff;[39] unless done very carefully, that will require giving *more* access to the remaining individuals.)

It is a truism in the computer security business that data that does not exist cannot be compromised. This includes both organizational misuse and misuse by individuals. Conversely, databases that do exist can be and are misused. In one recent case, a New York City police officer was charged with improperly looking up information on some of his fellow officers via the National Criminal Information Center system; officials suggest that he was doing this to monitor activities of his ex-girlfriend.[40] There were technical restrictions and audits, of course, but "both the instructor testifying at the Valle trial [a different case than this new allegation] and an Internal Affairs Bureau investigator who took the witness stand in an earlier case have conceded that officers can easily circumvent safeguards."

I am by no means suggesting that intelligence agencies should not collect or store information. That said, any form of collection does pose additional risks to personal privacy and security; an evaluation of the desirability of creating new databases of this type should take potential misuse into account as well. Put bluntly, it *will* happen; technical and personnel precautions will at best limit the extent.

[34] It is unclear to me why this paragraph was ever considered worthy of classification, since it is immediately and trivially obvious to anyone with any system administration background whatsoever.

[35] [Redacted], "Out of Control," Originally classified SECRET. There is another, and more heavily redacted, version at `http://www.nsa.gov/public_info/_files/cryptologic_quarterly/Out_of_Control.pdf`, *Cryptologic Quarterly* 15, Special Edition (1996), `http://www.gwu.edu/~nsarchiv/NSAEBB/NSAEBB424/docs/Cyber-009.pdf`.

[36] Scott Shane and David E. Sanger, "Job Title Key to Inner Access Held by Snowden," *New York Times* (June 30, 2013), `https://www.nytimes.com/2013/07/01/us/job-title-key-to-inner-access-held-by-snowden.html`.

[37] Shaya Potter, Steven M. Bellovin, and Jason Nieh, "Two Person Control Administration: Preventing Administration Faults through Duplication," in *LISA '09* (November 2009), `http://www.usenix.org/events/lisa09/tech/full_papers/potter.pdf`.

[38] National Commission on Terrorist Attacks Upon the United States, *Final Report of the National Commission on Terrorist Attacks Upon the United States* (2004), `http://www.9-11commission.gov/report/911Report.pdf`.

[39] Nicole Blake Johnson, "NSA Leak Raises Concerns About IT Hiring Practices," *Federal Times* (June 26, 2013), `http://www.federaltimes.com/article/20130626/IT03/306260006/NSA-leak-raises-concerns-about-hiring-practices`.

[40] Tom Hays, "NYC Cases Show Crooked Cops' Abuse of FBI Database," *Associated Press* (July 7, 2013), `http://bigstory.ap.org/article/nyc-cases-show-crooked-cops-abuse-fbi-database`.

Conclusions

The Section 215 and 702 programs have generated a lot of controversy. Beyond doubt, they pose thorny questions on the proper balance between personal privacy and national security. Some of these questions, though, turn on purely technical matters; understanding them is crucial to drawing the necessary lines.