THE PRIVACY PROBLEMS (?) OF OUTSOURCING THE DRAGNET

Both Ed Felten …

> I am reminded of the scene in Austin Powers where Dr. Evil, in exchange for not destroying the world, demands the staggering sum of "… one MILLION dollars." In the year 2014, billions of records is not a particularly large database, and searching through billions of records is not an onerous requirement. The metadata for a billion calls would fit on one of those souvenir thumb drives they give away at conferences; or if you want more secure, backed up storage, Amazon will rent you what you need for $3 a month. Searching through a billion records looking for a particular phone number seems to take a few minutes on my everyday laptop, but that is only because I didn't bother to build a simple index, which would have made the search much faster. This is not rocket science.

And Tim Edgar have started thinking about how to solve the dragnet problem.

> One helpful technique, private information retrieval, allows a client to query a server without the server learning what the query is. This would allow the NSA to query large databases without revealing their subjects of interest to the database holder, and without collecting the entire database. Recent advances should allow such private searches across multiple, very large databases, a key requirement for the program. The use of these cryptographic techniques would make the

> need for a separate consortium that holds the data unnecessary. I discussed this in more detail in my testimonybefore the Senate Select Committee on Intelligence last fall. Seny Kamara of Microsoft Researchpoints out these techniques were first outlined over fifteen years ago, while the state of the art is outlined in "Useable, Secure, Private Search" from IEEE Security and Privacy.

But I want to consider something both point to that President Obama said in his speech which both Felten and Edgar consider.

> Relying solely on the records of multiple providers, for example, could require companies to alter their procedures in ways that raise new privacy concerns.

I'm admittedly obsessed by this, but one processing step the NSA currently uses on dragnet data seems to pose particularly significant privacy concerns: the data integrity role, in which high volume numbers — pizza joints, voice mail access numbers, and telemarketers, for example — are "defeated" before anyone starts querying the database.

This training module from 2011 (and therefore before some apparent additions to the data integrity role, as I'll lay out in a future post) describes three general technical roles, the first of which would be partly eliminated if the telecoms kept the data.

- Ensuring production meets the terms of the order and destroying that which exceeds it (5)
- Ensuring the contact-chaining process works as promised to FISC (much of

this description is redacted) (7)

- Ensuring that all BR and PR/TT queries are tagged as such, as well as several other redacted tasks (this tagging feature was added after the 2009 problems) (9)

The first and third are described as "rarely coming into contact with human intelligible" metadata (the first function would likely see *more* intelligible data on intake of completed queries from the telecoms). But — assuming a parallel structure across these three descriptions — the redacted description on page 8 suggests that the middle function — what elsewhere is called the data integrity function — has "direct and continual access and interaction" with human intelligible metadata.

And indeed, the 2009 End-to-End Review and later primary orders describe the data integrity analysts querying the database with non-RAS approved identifiers to determine whether they're high volume identifiers that should be taken out of the dragnet.

Those analysts are not just accessing data in raw form. They're making analytic judgments about it, as this description from the E-2-E report explains.

> As part of their Court-authorized function of ensuring BR FISA metadata is properly formatted for analysis, Data Integrity Analysts seek to identify numbers in the BR FISA metadata that are not associated with specific users, e.g., "high volume identifiers." [Entire sentence redacted] NSA determined during the end-to-end review that the Data Integrity Analysts' practice of populating non-user specific numbers in NSA databases had not been described to the Court.

> (TS//SI//NT) For example, NSA maintains a database, [redacted] which is widely used by analysts and designed to hold identifiers, to include the types of non-user specific numbers referenced above, that, based on an analytic judgment, should not be tasked to the SIGINT system. In an effort to help minimize the risk of making incorrect associations between telephony identifiers and targets, the Data Integrity Analysts provided [redacted] included in the BR metadata to [redacted] A small number of [redacted] BR metadata numbers were stored in a file that was accessible by the BR FISA-enabled [redacted], a federated query tool that allowed3 approximately 200 analysts to obtain as much information as possible about a particular identifier of interest. Both [redacted] and the BR FISA-enabled [redacted] allowed analysts outside of those authorized by the Court to access the non-user specific number lists.
>
> In January 2004, [redacted] engineers developed a "defeat list" process to identify and remove non-user specific numbers that are deemed to be of little analytic value and that strain the system's capacity and decrease its performance. In building defeat lists, NSA identified non-user specific numbers in data acquired pursuant to the BR FISA Order as well as in data acquired pursuant to EO 12333. Since August 2008, [redacted] had also been sending all identifiers on the defeat list to the [several lines redacted]. [my emphasis]

That analytical judgment part is key: this does appear to be a judgment call about the distortion effect of the number balanced against its possible value. And as I've suggested, it is possible such judgment calls could strip the most important data from the database.

In addition, whether these tech people or others do the work, some analysts use raw data to test new chaining approaches and automatic queries, which has resulted in raw dragnet data ending up in places it didn't belong.

It wasn't until one of the three primary orders after September 3, 2009 (two of those have been withheld) that FISC required these techs to destroy the raw data when they were done with it. That didn't prevent the retention of over 3,000 files apparently used for this purpose on a server up until 2012.

> As of 16 February 2012, NSA determined that approximately 3,032 files containing call detail records potentially collected pursuant to prior BR Orders were retained on a server and been collected more than five years ago in violation of the 5-year retention period established for BR collection. Specifically, these files were retained on a server used by technical personnel working with the Business Records metadata to maintain documentation of provider feed data formats and performed background analysis to document why certain contact chaining rules were created. In addition to the BR work, this server also contains information related to the STELLARWIND program and files which do not appear to be related to either of these programs. NSA bases its determination that these files may be in violation of BR 11-191 because of the type of information contained in the files (i.e., call detail records), the access to the server by technical personnel who worked with the BR metadata, and the listed "creation date" for the files. It is possible that these files contain STELLARWIND data, despite the creation date. The STELLARWIND data could have been copied to this server,

> and that process could have changed the creation date to a timeframe that appears to indicate that they may contain BR metadata.

Which is to sum up: as of right now, it appears this role still requires both analytic judgment and access to human identifiable data in raw form. Verizon and AT&T presumably have their own automated function to do similar things for their own communities of interest, but that judgment call might be easier to automate than the one a tech analyst hoping to maximize the chances of finding a terrorist might make.

I'll let the tech folks debate ways to accomplish this without creating the dragnet in the first place. But it does seem to be one likely explanation for the addition privacy challenges the President referenced in his speech.