# I Spy?

by William Menke, December 20, 2013

I read Michael Chertoff's opinion piece[1] in *USA Today*, which concerns NSA's telephone monitoring program, while flying home from San Francisco. He gives the following description of the program:

> " … US telephone metadata collected by the NSA are anonymized records that memorialize calls from one number to another, and their duration. The content and the subscriber information are not disclosed until a suspicious number is identified."

and then puts forward the flowing scenario:

> "An analyst may then check the database to determine other numbers that have been in contact with the suspicious number. If officials identify the number of an al-Queda safe house, say, in Syria, the database gives them the ability to determine if a US phone has been in contact with the safe house."

As we fly over the western desert, I am imagining a serious and well-muscled NSA officer, who looks a lot like Liev Schreiber in black suit and tie, hunched over a keyboard and using his index finger to peck out a phone number. However, what Mr. Chertoff's scenario and my caricature misses is the awesome power of computers.

Stuck in Seat 40C, and having already polished off *USA Today*, the flight magazine and the shopping catalog, I began to think of what I would do if I had access to several years of anonymized phone logs. That's a huge pile of *very rich* data!

I'd start out by de-anonymized the phone numbers. The great majority of phone numbers are not confidential, and even those that are confidential are not truly secret. Many name/number combinations are published in *telephone directories*, web sites and other public forums. Even those that aren't may be available in documents, such as real estate records, that can be obtained legally, just for the asking. In fact, I would be surprised if the NSA *didn't* have a division that routinely collects contact information for everyone, everywhere and that it predates their telephone program by many years. It's information that can come in handy in a myriad of settings.

I'd use the telephone metadata to reconstruct *networks*, meaning groups of telephone numbers that are in routine communication with one another. Superficially, you might think that the task would be overwhelming. Almost everyone uses a phone and almost everyone belongs to several distinct networks (relatives, business associates, club members, etc.) who regularly converse with one another. There may well be billions of networks. But computers are very good at finding patterns in data[2]. Uncovering a few billion networks from a few trillion phone calls is by no means an overwhelming task. Furthermore, the results need not be 100% correct. So what if,

say, the telephone number for Olympic Pizza is accidentally included in a network that is otherwise composed of members of a Sloatsburg NY athletic club, on account of club members ordering pies from that restaurant. The issue of whether the restaurant should really be considered part of the network can be resolved later – or not at all, if the available information is good enough to permit a decision to exclude the network from further scrutiny.

I would then do my best to characterize all the networks and to flag those that might possibly represent threats. Geography will be important; a network that crosses international boundaries will obviously be of more interest than one limited to, say, Lewiston, Maine. The types of commercial enterprises included in the network will hint at its function. It would be pretty east to distinguish, say, a book club from a gun club on that basis. The time of day that a network was most active might be diagnostic, especially if had a pattern of activity that matched a different time zone than its nominal location. And, of course, information about any individuals linked to the phone numbers, their ages, ethnicities, citizenship, occupations, convictions and socio-economic status could be extremely useful in assessing whether a given network might pose a threat. Billions could be winnowed down to millions and maybe even thousands on this basis, especially since, as before, the results need not be perfect. The list of suspicious networks only needs to be accurate enough to include a substantial percentage of threats and be small enough to enable further screening, either by more specialized software or by human analysts.

Correlation of network activity with world events might provide a further insight a network's character. It would be relatively easy to assemble minute-by-minute timelines of when news of various sorts breaks on all the world's major media outlets. And again, I would be surprised if the NSA were not already generating such timelines on almost every imaginable subject, because they have so many obvious uses. A network whose activity tracks, say, a *Mid East Assassinations* timeline is likely to be of much more interest to the NSA than one that tracks *US Football*. And, of course, a network whose activity picked up *just before* breaking news would raise red flags.

Michael Chertoff's unimaginative scenario notwithstanding, the telephone monitoring debate is not really about the information that might be available to a single NSA officer or even the complete ensemble of them. It is about very large scale compute-based information mining, where the most important human decisions are made up-front and include choices about what level of accuracy of results to accept, and where human scrutiny of results occurs very late in the information vetting process.

[1]Michael Chertoff, Opposing View: Don't hamstring data collection, USA Today, December 18, 2013.

[2]Consider that credit card companies routinely monitor hundreds of millions of transactions every day for signs of fraud.