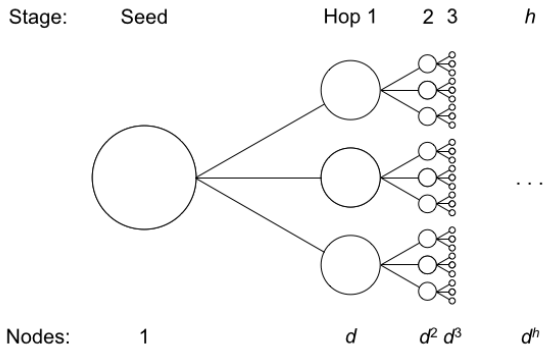
THREE-HOPPING THE CORPORATE STORE, IN THEORY

Stanford University has been running a project to better understand what phone metadata can show about users, MetaPhone, in which Android users can make their metadata available for analysis.

They just published a piece that suggests we could be underestimating the intrusiveness of the government's phone dragnet program. That's because most assumptions about degrees of separation consider only human contacts, and not certain hub phone numbers that quickly unite us.

> A common approach for calculating these figures has been to simply assume an average number of call relationships per phone line ("degree"), then multiply out the number of hops. If a single phone number has average degree *d*, and the NSA can make *h* hops, then a single query gives expected access to about d^h complete sets of phone records.[3, 4]
>
> 
>
>
> *We turned to our crowdsourced MetaPhone dataset for an empirical measurement. Given our small, scattershot, and*

> *time-limited sample of phone activity, we expected our graph to be largely disconnected. After all, just one pair from our hundreds of participants had held a call.*

> Surprisingly, our call graph *was* connected. Over 90% of participants were related in a single graph component. And within that component, participants were closely linked: on average, over 10% of participants were just 2 hops away, and over 65% of participants were 4 or fewer hops away!

In spite of the fact that just 2 of its participants had called each other, the fact that so many people had called TMobile's voicemail number connected 17% of participants at two hops.

> Already 17.5% of participants are linked. That makes intuitive sense—many Americans use T-Mobile for mobile phone service, and many call into voicemail. Now think through the magnitude of the privacy impact: T-Mobile has over 45 million subscribers in the United States. That's potentially tens of millions of Americans connected by just two phone hops, solely because of how their carrier happens to configure voicemail.

And from this, the piece concludes that NSA could get access to a huge number of numbers with just one seed.

> But our measurements are highly suggestive that many previous estimates of the NSA's three-hop authority were conservative. Under current FISA Court orders, the NSA may be able to analyze

> the phone records of a sizable proportion of the United States population with just one seed number.

This analysis doesn't account for one thing: NSA uses Data Integrity Analysts who take out high volume numbers — numbers like the TMobile voice mail number.

Here's how the 2009 End-to-End review of the phone dragnet described their role.

> 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 allowed 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].

And here's a (heavily-redacted) training module that describes what kind of massaging the tech people (which is a wider set of people than just the Data Integrity Analysts) do with dragnet data.

If the Data Integrity Analysts operate as multiple NSA documents say they do, this kind of quick inclusion of all Americans shouldn't happen — it's precisely the kind of noise NSA says it is trying to defeat.

There are just two problems with this then. First, as I have noted in the past, the inclusion or exclusion of high volume numbers will at times be a judgment call, and could lead to eliminating the most valuable pieces of intelligence in the dataset if targets knowingly or unknowingly exploit these high volume numbers. Similarly, it could easily be used — and may already have been — to make the dragnets totally unusable at critical times.

More importantly, this tech role receives far less oversight than the regular analysts do. And Dianne Feinstein's Fake FISA Fix might even eliminate some of the oversight on the position

now. So we have almost no way (and Congress seems to want to deprive itself of having a way) of ensuring these Data Integrity Analysts are doing what we think they're doing.

If NSA is doing what it says, then the Stanford analysis should be moot, because it doesn't account for that Data Integrity role. But ACLU's Patrick Toomey explained back in August, NSA has a very real incentive to get as much data picked up in queries and into the corporate store as it can.

> All of this information, the primary order says, is dumped into something called the "corporate store." Incredibly, the FISC imposes *no* restrictions on what analysts may subsequently do with the information. The FISC's primary order contains a crucially revealing footnote stating that "the Court understands that NSA may apply the full range of SIGINT analytic tradecraft to the result of intelligence analysis queries of the collected [telephone] metadata." In short, once a calling record is added to the corporate store, anything goes.
>
> More troubling, if the government is combining the results of *all* its queries in this "corporate store," as seems likely, then it has a massive pool of telephone data that it can analyze in any way it chooses, unmoored from the specific investigations that gave rise to the initial queries. To put it in individual terms: If, for some reason, your phone number happens to be within three hops of an NSA target, *all* of your calling records may be in the corporate store, and thus available for any NSA analyst to search at will.
>
> But it's even worse than that. The primary order prominently states that whenever the government accesses the wholesale telephone-metadata database,

> “an auditable record of the activity shall be generated.” It might feel fairly comforting to know that, if the government abuses its access to all Americans’ call data, it might eventually be called to account—until you read footnote 6 of the primary order, which *exempts entirely* the government’s use of the “corporate store” from the audit-trail requirement.

Not “defeating” numbers like the TMobile voice mail is a very easy way to populate the corporate store with very very broad swaths of US person data so as to be able to access it with much less stringent controls.

All of which demonstrates the urgency for more oversight into whether the Data Integrity Analysts are doing what they say they’re doing.