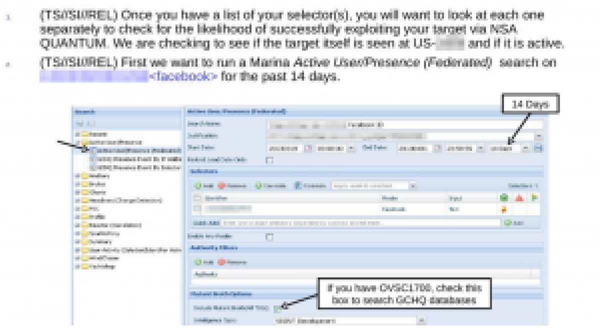
HOW NSA HUNTS METADATA “CONTENT” IN SEARCH OF YOUR DIGITAL TRACKS



3. (TS//SI//REL) Once you have a list of your selector(s), you will want to look at each one separately to check for the likelihood of successfully exploiting your target via NSA QUANTUM. We are checking to see if the target itself is seen at US-[redacted] and if it is active.
4. (TS//SI//REL) First we want to run a Marina *Active User/Presence (Federated)* search on [redacted]<facebook> for the past 14 days.

14 Days

If you have OVSC1700, check this box to search GCHQ databases

Der Speigel has posted a set of slides associated with their story on how NSA’s TAO hacks targets.

The slides explain how analysts can find identifiers (IPs, email addresses, or cookies) they can most easily use to run a Quantum attack.

Because NSA is most successful hacking Yahoo, Facebook, and static IPs, it walks analysts through how to use Marina (or “QFDs,” which may be Quantum specific databases) to find identifiers for their target on those platforms. If they can’t find one of them, it also notes, analysts can call on GCHQ to hack Gmail. Once they find other identifiers, they can see how often the identifier has been “heard,” and how recently, to assess whether it is a still-valid identifier.

The slides are fascinating for what they say about NSA’s hacking (and GCHQ’s apparent ability to bypass Google’s encryption, perhaps by accessing their own fiber). But they’re equally interesting for what they reveal about how the NSA is using Internet metadata.

The slides direct analysts to enter a known identifier, to find all the other known identifiers for that user, which are:

> determined by linking **content**

> (logins/email registrations/etc). It is worth verifying that these are indeed selectors associated to your target. [my emphasis]

This confirms something — about Internet metadata, if not yet phone metadata — that has long been hinted. In addition to using metadata to track relationships, they're also using it to identify multiple identities across programs.

This makes plenty of sense, since terrorists and other targets are known to use multiple accounts to hide their identities. Indeed, doing more robust such matching is one of the recommendations William Webster made after his investigation of Nidal Hasan's contacts with Anwar al-Awlaki, in part because Hasan contacted Awlaki via different email addresses.

But it does raise some issues. First, how accurate are such matches? The NSA slides implicitly acknowledge they might not be accurate, but it provides no clues how analysts are supposed to "verify[] that these are indeed selectors associated to your target." In phone metadata documents, there are hints that the FISC imposed additional minimization procedures for matches made with US person identifiers, but it's not clear what kind of protection that provides.

Also, remember NSA was experiencing increased violation numbers in early 2012 in significant part because of database errors, and Marina errors made up 21% of those. If this matching process is not accurate, that may be one source of error.

Also, note that NSA itself calls this "content," not metadata. It may be they've associated such content via other means, not just metadata collection, but given NSA's "overcollection" of metadata under the Internet dragnet, almost certainly collecting routing data that count as content, it does reflect the possibility they themselves admit this goes beyond metadata.

Moreover, this raises real challenges to NSA claims that they don’t know the “identity” of the people they track in metadata.

Now, none of this indicates US collection (though it does show that NSA continues to collect truly massive amounts of Internet traffic from some location). But the slide above does show NSA monitoring whether this particular user was “seen” at US-[redacted] in the last 14 days. US-[redacted] is presumably a US-associated SIGAD (collection point). (They’re looking for a SIGAD from which they can successfully launch Quantum attacks, so seeing if their target’s traffic uses that point commonly.) While that SIGAD may be offshore, and therefore outside US legal jurisdiction, it does suggest this monitoring takes place within the American ambit.

At least within the Internet context, Marina functions not just as a collection of known relationships, but also as a collection of known data intercepts, covering at least a subset of traffic. They likely do similar things with international phone dragnet collection and probably the results of US phone dragnet in the “corporate store” (which stores query results).

In other words, this begins to show how much more the NSA is doing with metadata than they let on in their public claims.

Update: 1/1/14, I’m just now watching Jacob Appelbaum’s keynote at CCC in Berlin. He addresses the Marina features at 22:00 and following. He hits on some of the same issues I do here.