DISTANTFISH AND CORRELATIONS

For some time, I've been trying to track how the NSA does correlations, as a 2008 FISA Court opinion that almost certainly approves correlation has been withheld from release. By "correlation," NSA means that matching of known strong identifiers of a particular traffic. All such identifiers need to be tracked to track a target (indeed, France was not able to prevent the Bataclan attack because they had lost track of one of the key attackers).

One of the SIDToday newsletters the Intercept released today describes how a key tool to correlate identities, DISTANTFISH, works.

Here's how it describes DISTANTFISH's two functions:

> (S//SI) PSC works by processing application layer protocols to extract certain metadata fields that work as strong selectors for the client of the current application. These selectors are usually login names, client e-mail addresses, user numbers, and other unique metadata. If a selector is found to be that of a known terrorist, that session, as well as all others generated by the terrorist, is forwarded to NSA for analysis. The DISTANTFISH association algorithms are the primary way of determining which sessions the terrorist generated when the access is traditional passive collection. The collection of all user sessions is called the Aggregate Session and can be achieved by other methods, especially active efforts.
>
> (S//SI) However, PSC assumes that the strong selectors for a terrorist are known. The second objective for DISTANTFISH is to associate all strong

> selectors for SIGINT targets and store them in a database. Intelligence analysts use the database to discover new identities to add to the selectors for that terrorist. Work on this database has begun, but much work remains.

And here's how it worked to collect all the web activity of a particular target in Iraq in 2004.

> (S//SI) Project DISTANTFISH was created to target terrorist traffic on the Internet by providing two important services. First, it provides a database for discovering account identities for known terrorists to use as strong selectors (i.e. login names, e-mail addresses, or other elements that can be associated with a particular individual). Second, it provides information on which the same user generated computer sessions. Thus, if one session contains a strong selector for a terrorist, then all sessions can be collected. At the heart of this capability is an association service that can track an individual computer by the way it generates packets.
>
> (S//SI) From this association service, the DISTANTFISH team members were able to determine that the terrorist generated 107 computer sessions over eleven minutes, thus separating this traffic from that of the other 16 people in the web cafÃ©. As most of the supporting software is still under development, the data was manually examined resulting in the discovery of two additional MSN Messenger accounts and two Yahoo web mail accounts that the terrorist used, but that NSA had been unaware of. Since terrorists often abandon accounts for new ones, having a complete picture of the accounts used is critical for targeting the terrorists'

> traffic.

Remember, the USA Freedom Act requires “phone” companies, broadly defined, to turn over “session identifiers” under the guise of call records. Any such session identifier can be used to correlate identities in this fashion. I have long argued that is the point of USAF: to get tech companies to do correlations with a near perfect degree of accuracy rather than (in fact, in addition to) having the NSA correlate the IDs.