DATA MINING RESEARCH PROBLEM BOOK, WORKING THREAD

Yesterday, Boing Boing liberated a fascinating 2011 GCHQ document from the Snowden collection on GCHQ's partnership with Heilbronn Institute for Mathematical Research on datamining. It's a fascinating overview of collection and usage. This will be a working thread with rolling updates.

In addition to BoingBoing's article, I'll update with links to other interesting analysis.

- A technical review from Conspicuous Chatter.

[1] The distribution list is interesting for the prioritization, with 4 NSA research divisions preceding GCHQ's Information and Communications Technology Research unit. Note, too, the presence of Livermore Labs on the distribution list, along with an entirely redacted entry that could either be Sandia (mentioned in the body), a US university, or some corporation. Also note that originally only 18 copies of this were circulated, which raises real questions about how Snowden got to it.

[9] At this point, GCHQ was collecting primarily from three locations: Cheltenham, Bude, and Leckwith.

[9-10] Because of intake restrictions (which I believe other Snowden documents show were greatly expanded in the years after 2011), GCHQ can only have 200 "bearers" (intake points) on "sustained cover" (being tapped) at one time. Each collected at 10G a second. GCHQ cyclically turns on all bearers for 15 minutes at a time to see what traffic is passing that point (which is how they hack someone, among other things). Footnote 2 notes that analysts aren't allowed to write up reports on this feed, which suggests

research, like the US side, is a place where more dangerous access to raw data happens.

[10] Here's the discussion of metadata and content; keep in mind that this was written within weeks of NSA shutting down its Internet dragnet, probably in part because it was getting some content.

> Roughly, metadata comes from the part of the signal needed to set up the communication, and content is everything else. For telephony, this is simple: the originating and destination phone numbers are the metadata, and the voice cut is the content. Internet communications are more complicated, and we lean on legal and policy interpretations that are not always intuitive. For example, in an HTTP request, the destination server name is metadata (because it, or rather its IP address, is needed to transmit the packet), whereas the path-name part of the destination URI is considered content, as it is included inside the packet payload (usually after the string GET or POST). For an email, the to, from, cc and bcc headers are metadata (all used to address the communication), but other headers (in particular, the subject line) are content; of course, the body of the email is also content.

[10] This makes it clear how closely coming up as a selector ties to content collection. Remember, NSA was already relying on SPCMA at this point to collect US person Internet comms, which means their incidental communications would come up easily.

> GCHQ's targeting database is called BROAD OAK, and it provides selectors that the front-end processing systems can look for to decide when to process content. Examples of selectors might be telephone numbers, email addresses or IP

> ranges.

[11] At the Query-Focused Dataset level (a reference we've talked about in the past), they're dealing with: "the 5-tuple (timestamp, source IP, source port, destination IP, destination port) plus some information on session length and size."

[11] It's clear when they say "federated" query they're talking global collection (note that by this point, NSA would have a second party (5 Eyes) screen for metadata analysis, which would include the data discussed here.

[11] Note the reference to increased analysis on serious crime. In the UK there's not the split between intel and crime that we have (which is anyway dissolving at FBI). But this was also a time when the Obama Admin's focus on Transnational Crime Orgs increased our own intel focus on "crime."

[12] This is why Marco Rubio and others were whining about losing bulk w/USAF: the claim that we are really finding that many unknown targets.

> The main driver in target discovery has been to look for known modus operandi (MOs): if we have seen a group of targets behave in a deliberate and unusual way, we might want to look for other people doing the same thing.

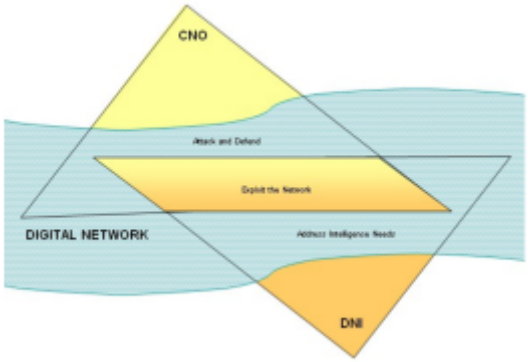
[12] This is reassuring but also interesting for the scope involved.

> It is important to point out that tolerance for false positives is very low: if an analyst is presented with three leads to look at, one of which is probably of interest, then they might have the time to follow that up. If they get a list of three hundred, five of which are probably of interest, then that is not much use to them

[13] GCHQ's first CNE was in the early 90s.

[13] Lists the protection of banks and other companies operating in the UK as part of CESG's defensive role. In the US we've adopted this model unthinkingly, even though US law enforcement doesn't have the same explicit role in providing for "economic well-being."

[15] This is NSA's conception of how hacking (CNO) and intelligence collection (DNI) intersect.



[16] In which GCHQ admits it doesn't do a lot of machine learning (which is what this research was supposed to address).

> There are a vast number of supervised machine learning algorithms which can often produce functions with high accuracies on real-world data sets. However, these techniques have had surprisingly little impact in GCHQ. There are various reasons why this has been the case but the principal reason has been the difficulty in creating training sets. In particular, the difficulty comes from knowing the desired output value for many training examples, either due to the required human effort and/or uncertainty in the desired output value. This difficulty is unlikely to be a one-off issue for an operational application. The nature of communications and our data changes with time and leads to "concept drift"; any algorithm must be periodically retrained.

[17] here are the areas where GCHQ has been successful:

> steganography detection (Random Forest) [I74], website classification (decision tree) [I36], protocol classification (Random Forest and neural network) [W1], spam detection (Random Forest) [I44], payphone detection (Random Forest) [I3] and drug smuggler detection (logistic regression) [I77].

Note steganography detection should be useful for the use of gaming consoles.

[17] GCHQ machine learning also affected because of holes (visibility problems) in the data.

[17] Note the redaction in footnote 6, which describes some entity the NSA's statistical advisory group worked with. Could it be the same entity as listed on the title page? Or a university or corporation? From a COMSEC standpoint, the US should have better expertise available via private industry.

[18] The asymmetry of metadata/content also affects the ability to do data mining bc you need content to truth the data.

[21] Lists contacts, timing, and geo behavior as bases for inferring a relationship between entities.

[21] In relationship scoring, GCHQ started with email comms. That's interesting, but already seems outdated by 2011. This passage also admits their facility for IDing location from an IP, something NSA pretends is not true for regulatory reasons (and FBI pretends even more aggressively).

[30] THis is an interesting admission, coming very late in their process of using billing records.

> In experiments carried out on billing records and SIGINT during the 2008 graph mining SWAMP at HIMR there was shown be

> a huge disparity between our view of the world and ground truth [I73]. CSEC have perform similar analyses with similar conclusions

[31] When NSA started aging off phone dragnet data it dealt with multiple time stamps. I think it was a different problem (arising from associating chains of texts that got recollected—and therefore permissively kept for 5 years from the new collection date—on multiple days), but that may not be the case.

> In particular the quality of the timing information is not as good as we might hope for. This presents at least two concrete problems. Firstly, our data tends to have second timestamps, which may be too coarse a measure for many applications. Does the granularity of the timestamps affect J our chances of finding causal flows? Secondly the clocks on our probes are not synchronised. This means that there is likely to be a constant offset between events happening on different bearers. Any technique to correct for this offset will both aid this problem area and be of general interest to the internal data mining and information processing community. Can we J correct for the clock offset between probes? Possible solutions may involve examining the same connection being intercepted on different bearers.

[32] Remember that NSA has invested a lot of work in mapping structure and devices. Note how this would interact with that process.

> We do have some truthing on flows that may exist in the data. Specifically, we have data on covert infrastructure (appendix F.3.3) used for exfiltrating data from CNE implants. These suspected flows can be used for both EDA and

> evaluation purposes. Further, we have lists of IPs that we suspect to be infected with the Conficker botnet (appendix F.3.4), either due to signatures collected or behavioural analysis

[36] This is the problem I keep talking about—but I find it a bit troubling that they don't consider the possibility that a pizza node is meaningful, particularly among targets (the Tsarnaev brothers) who have worked in that space.

> Removing pizza nodes (i.e. very high-degree nodes) is likely to be an essential prior component to get useful results. Intuitively, a pizza node is likely to be a large impersonal entity like a pizza parlour or an electricity supplier: the fact that two people both communicate with the pizza node gives us no reason to think that they are linked socially.

[38] When Obama limited the phone dragnet to 2 hops in 2014, that's what *analysts* were already doing. But there's some indication that tech people were doing more (which of course doesn't get audited). Plus, the NSA is not limited to two hops and 12333 data, and the old phone dragnet was in some senses fill for that. Note that Stanford has examined some of this in replication of the NSA dragnet, with a smaller dataset.

> Can we approximate the graph distance distribution, and see how it varies with the pizza threshold?
>
> This has a bearing on what hop distance we should choose for contact chaining. Conventionally, analysts focus on a 2-hop neighbourhood of their targets, but some work comparing billing records with SIGINT [I73] found that one needed to

> chain much, much further through SIGINT to reach a 2-hop neighbourhood from billing data. Can we use the SIGINT to billing J mapping (SOLID INK to FLUID INK—see appendix F.1.6) to help decide what the right thing to measure on a telephony graph is?

[40] Again, this reflects some uncertainty about the correlations GCHQ was making at a time when NSA was moving towards automating all of this. I wonder what FISC would say if it had seen this?