Written evidence submitted by David Mytton (IPB 10)

Server Density is a systems management and monitoring startup I founded in 2009. We provide software which monitors the availability and performance of our customer's server environments, providing metrics to allow them to maintain the reliability and performance of their systems. For example, our software monitors the 999 and 111 emergency response systems for the NHS North East Ambulance Service[1].

I have been a programmer for 15 years and built the original version of our software, now maintained by a team of 20. We collect billions of metrics and many hundreds of terabytes of data from our customers. As such, we have significant technical experience building large scale, distributed logging and data collection systems which we operate on behalf of our customers.

0. Summary

0.1. Internet Connection Records can easily be spoofed and provide no indication of intention.

0.2. It is trivial to mask all communication with your provider through usage of a VPN or Tor.

0.3. As such, Internet Connection Records are not useful.

0.4. Storing and making such retention data available for search is hugely expensive.

1. Internet Connection Records

1.1. The current draft bill provides a very broad definition of what has been referred to elsewhere as an Internet Connection Record. Section 78(1) allows for an order to "retain relevant communications data". This is defined in section 223(5) as "entity data or events data" and when ultimately referenced back to section 223(2) results in "everything" because of section 223(2)(a) "anything comprising speech, music, sounds, visual images or **data of any description**" (emphasis mine).

1.2. This definition is inconsistent with repeated statements such as Mrs May 15 March 2016 Column 820 Hansard: "To reiterate, internet connection records do not provide access to a person's full web browsing history. An internet connection record is a record of what internet services a device or person has connected to, not every web page they have visited."

1.3. The use of the term "communications data" is misleading because it implies that this is only related to services one might normally consider "communication" e.g. e-mail or text based messaging, when in fact the definition is so broad that "data of any description" is covered. This means any attempt to access any service whatsoever.

1.4. The use of the phrase "Internet Connection Records" implies an analogy to an itemised phone bill. When making a phone call, there is a single "connection" from a source phone number to a destination phone number. This analogy breaks down when applied to the internet because of how many different connections take place across many different supporting services.

[1] https://www.serverdensity.com/customer-nhs/

1.5. By way of example, if I make a call to a fictional Mr Smith then the record would show that my phone number connected to Mr Smith's number, with a timestamp. That would be the extent of the record. However, if I visit the BBC website at http://www.bbc.co.uk then just inspecting the browser networking history one can see there are 104 separate requests. Under the current draft bill, each of these would be an "Internet Connection Record":

Name	Method	Status	Domain
about:blank	GET	Finished	
l.php?id=INS-642345567&v=7212&x...	GET	200	edigitalsurvey.com
l.php?id=INS-642345567&v=7212&x...	GET	307	edigitalsurvey.com
fig.js	GET	200	fig.bbc.co.uk
cross_icon_small--soft-blue.svg	GET	200	homepage.files.bbci.co.uk
plus.svg	GET	200	homepage.files.bbci.co.uk
right-arrow--soft-blue.svg	GET	200	homepage.files.bbci.co.uk
cross_icon_small--pale-blue.svg	GET	200	homepage.files.bbci.co.uk
cross_icon_small--tower-grey.svg	GET	200	homepage.files.bbci.co.uk
arrow-left.svg	GET	200	homepage.files.bbci.co.uk
arrow-right.svg	GET	200	homepage.files.bbci.co.uk
app.js	GET	200	homepage.files.bbci.co.uk
GELIconsFull-Book.woff2	GET	200	homepage.files.bbci.co.uk
GillSansW01Light.woff2	GET	200	homepage.files.bbci.co.uk
live-withpadding.svg	GET	200	homepage.files.bbci.co.uk
right-arrow--grey.svg	GET	200	homepage.files.bbci.co.uk
video-cta-black.svg	GET	200	homepage.files.bbci.co.uk
right-arrow--tower-grey.svg	GET	200	homepage.files.bbci.co.uk
audio-cta-black.svg	GET	200	homepage.files.bbci.co.uk
iplayer-black.svg	GET	200	homepage.files.bbci.co.uk
right-arrow--pale-blue.svg	GET	200	homepage.files.bbci.co.uk
iplayer-radio-black.svg	GET	200	homepage.files.bbci.co.uk
cog.svg	GET	200	homepage.files.bbci.co.uk
main.css	GET	200	homepage.files.bbci.co.uk
cross_icon_small--grey.svg	GET	200	homepage.files.bbci.co.uk
p0306tt8.png	GET	200	ichef.bbci.co.uk
p0306trm.png	GET	200	ichef.bbci.co.uk
p03mt3s3.jpg	GET	200	ichef.bbci.co.uk
p03mthkj.jpg	GET	200	ichef.bbci.co.uk
p03mt3n7.jpg	GET	200	ichef.bbci.co.uk
p03k271h.jpg	GET	200	ichef.bbci.co.uk
p03gjtyz.png	GET	200	ichef.bbci.co.uk
p0306tp8.png	GET	200	ichef.bbci.co.uk
p0306tqc.png	GET	200	ichef.bbci.co.uk
_88852929_crabb_bbc.jpg	GET	200	ichef.bbci.co.uk
_88848259_88848258.jpg	GET	200	ichef.bbci.co.uk
_88847606_88847604.jpg	GET	200	ichef.bbci.co.uk
_88853959_pandya_afp.jpg	GET	200	ichef.bbci.co.uk
_88854219_iwobi_rex.jpg	GET	200	ichef.bbci.co.uk
_88835302_gettyimages-163804561....	GET	200	ichef.bbci.co.uk
p03nb7tm.jpg	GET	200	ichef.bbci.co.uk
p03n19ds.jpg	GET	200	ichef.bbci.co.uk

Diagram 1: A sample set of connections made when accessing http://www.bbc.co.uk

1.6. In the BBC example, my intention was to access the BBC website. However, you can see from diagram 1 that access attempts to separate 3rd party services were also logged e.g. edigitalsurvey.com. I had no intention to visit edigitalsurvey.com and without inspecting the browser requests, would be completely unaware that request had been made on my behalf. If this is being logged against me, there is the obvious potential for a website to create connection records for services that I knew nothing about.

1.7. Paragraph 230 of the Explanatory Notes states “They could be used, for example, to demonstrate a certain device had accessed an online communications service”. This is of course true but it fails to provide any context for that access - what it deliberate or accidental. If the purpose is to demonstrate that someone deliberately accessed a particular service then there would always be a defence that it was an unintended access attempt.

1.8. The Internet Connection Records Factsheet[2] explains “You may be able to see that a person has used, google.co.uk or facebook.com but you would not be able to see what searches have been made on google or whose profiles had been viewed on Facebook.”. This is because “It could never contain a full web address as under the law these would be defined as content.” Both Google and Facebook run services for developers whereby they can log in using their Google or Facebook accounts. Many websites also include Facebook “Like” buttons or Google “+1” buttons. This would mean that any website using those services would create a connection record that a person accessed “google.co.uk” or “facebook.com”. Without the full web address, such a connection record is useless because it is too easily polluted. With a full web address, the record becomes “content” and invades privacy. Neither option works.

1.9. The bill assumes that serving a notice upon the telecommunications operator means that operator would be able to access the connection records. It is trivial to install a client side VPN on my computer such that every request (including DNS lookups) is encrypted and impossible to inspect by the provider. For example, it is standard security best practice and part of my company policy that all employees accessing any network (e.g. wifi) they do not control (e.g. at a hotel) use a VPN. This would render any retention notice pointless because there would be no record, only encrypted packets of data between my client and the VPN provider.

1.10. For many VPN providers, I can pick the location of the VPN server I connect to. This may be within the UK or it may be in another country where the UK has no jurisdiction and/or no arrangement to share data or warrants.

1.11. I could also configure my own VPN on my own server. I can run a server on any small server provider anywhere in the world for less than £10 per month. I could easily configure it to retain no logs and with virtualised instances, could easily destroy it at will, then launch a new instance at another provider in any other country. There are perfectly legitimate business reasons for using a VPN.

1.12. Alternatively I could use Tor to encrypt and redirect all my traffic through an anonymous network. This would have the same effect as using a VPN and is even easier to use.

1.13. The result of this is that internet connection records have no useful purpose because they a) cannot demonstrate intention; b) are easy to circumvent.

1.14. The Factsheet provides 3 purposes:

[2] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/473745/Factsheet-Internet_Connection_Records.pdf

1.14.1. “To identify the sender of a communication” - as described in 1.6 - 1.8 above, this data could be polluted because of requests made on my behalf but without my knowledge or intention by any site or service I happen to access.

1.14.2. “To identify the communications services a person is using” - as described in 1.9. - 1.12. above, encrypting and/or anonymising my access through a VPN and/or Tor means I can trivially hide all my access attempts.

1.14.3. “To determine whether a person has been accessing or making available illegal material online”. See 1.14.1 above. It would be easy to cause access to illegal material without any knowledge or intention of any user accessing a particular service.

1.15. Since all the purposes of Internet Connection Records can be shown to be easy bypassed, it makes the provision pointless. Even a narrowing of the definition can be bypassed through the use of VPN or Tor.

2. Retention of records

2.1. The Communications Data DRAFT Code of Practice[3] makes various references to security, integrity and destruction of retained data. There are broad descriptions that various safeguards must apply to the retained data but there is no detail about the actual technical mechanism for this. Being responsible for retention of vast quantities of monitoring metrics, this is something my company has significant experience of.

2.2. Building the software infrastructure to retain such vast quantities of data is non-trivial. At Server Density we operate this in a multi-tenant architecture so we can build and manage a single system ourselves. All customers run off the same shared infrastructure with built in security and segregation to ensure there are no security issues involving cross-account leakage of data. However, this is expertise built up over the last 7 years involving multiple versions of our metrics storage software.

2.3. Various considerations and optimisations have to go into the design and iteration of future versions of such a system. These include:

2.3.1. Integrity of the data. Ensuring that what is stored is not modified either through human or system error.

2.3.2. Availability of the data. Geographic redundancy is necessary to avoid localised failures and geographic events (e.g. fire, flood, storm).

2.3.3. Performance of queries against the data. With such huge volumes of data, search queries have to return responses within a reasonable period of time for the user. This requires specialist database design and the right infrastructure to support querying that data.

---

3 https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/505411/Communications_Data_draft_Code_of_Practice.pdf

2.4. These requirements are not cheap to implement. The software to store the data has to be designed, built and maintained. The infrastructure also has to be designed, built and maintained. Google offers a cloud based product called BigQuery which is designed to perform complex queries across huge datasets. The recently published a blog post[4] which explains how such a sophisticated system operates to execute multi-terabyte queries. They estimated[5] that to run such a query in under 30 seconds would require x330 100 MB/sec hard drives, 3,300 Cores and a 330 Gigabit network:

2.5. “These are highly optimistic assumptions that ignore lots of complexity:

2.5.1. Even if you have the data on 300 disks, the data will need to be perfectly spaced among those disks, you need to be able to read at full speed from all of those disks at once, and each disk will need to be otherwise completely idle. And be sure to account for several factors of replication for redundancy.

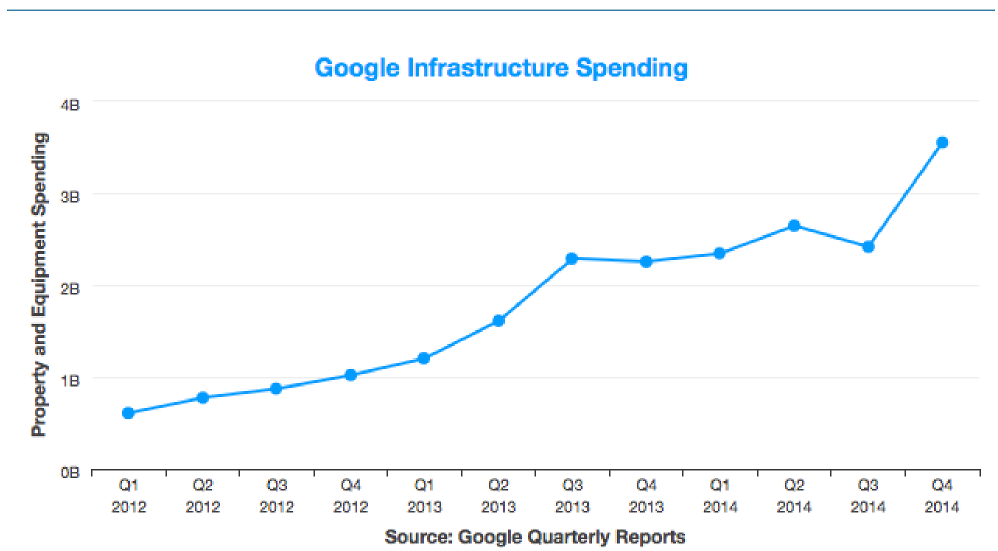
2.5.2. You would need to coordinate that many CPUs to start at the same time and get the same amount of work done. Assuming ~100 32-processor machines, one of the servers will fail every day on average, which will take all 3,300 CPUs offline, so you’ll need extra coordination to handle these failures without slowing down, including deploying additional computing redundancy, preferably across multiple zones.

2.5.3. To coordinate this much networking throughput so quickly, you’d need to make sure that networking throughput is evenly distributed across each instance, and you’ll likely worry about locality, as well as limits of total bisectional bandwidth.”

2.6. “BigQuery is powered by multiple data centers, each with hundreds of thousands of cores, dozens of petabytes in storage capacity, and terabytes in networking bandwidth. The numbers above — 300 disks, 3000 cores, and 300 Gigabits of switching capacity — are small. Google can give you that much horsepower for 30 seconds because it has orders of magnitude more.”

2.7. The current bill wording implies each communications provider must maintain its own infrastructure. Does that mean each provider must build their own system or purchase it from a 3rd party?

2.8. In 2014 Google spent $11 billion on infrastructure. Does the government expect to make similar investments?



oid-scaling-and-

Diagram 2 - Google Infrastructure Spending[6]

March 2016

[6] https://gigaom.com/2015/02/04/google-had-its-biggest-quarter-ever-for-data-center-spending-again/