August 2014

Geoff Huston

What is Metadata, and Why Should I Care?

August 2014 is proving yet again to be an amusing month in the Australian political scene, and in this case the source of the amusement was watching a number of Australian politicians fumble around the topic of digital surveillance and proposed legislation relating to data retention measures. The politicians assured us that the proposed data retention measures were nothing untoward, and all that was being called for was the retention of "metadata" by Australian ISPs for a period of two years. We were being reassured that this was not an underhand introduction of a surveillance state, and at no stage was this a call for ISPS to retain a complete copy of the content of all traffic that they carry for their users. All that was being called for in these measures was to retain "metadata".

But then the wheels fell off. The Australian Attorney General explained this as retaining the domain names of sites visited by users, but not the contents of any session. But it was pointed out that if the ISP was not running a web proxy then the only way that an ISP would harvest domain names from the traffic flow would be to inspect the traffic flows on the wire and pick out the domain names from the content stream. Time to call in the Minister of Communications, who explained that "metadata" meant that the ISPs would not be collecting domain names, but instead would be retaining a record of IP addresses used by clients. The attempts to explain these measures were no clearer at the end of the week than at the start. ISPs were going to be required to collect some of this "metadata" stuff, but no politician could give a clear and coherent view of precisely what this data actually was.

Time to call in the next set of professionals, and David Irvine, the Director General of Australia's security organization, ASIO, and Andrew Colvin, the Australian Federal Police's Deputy Commissioner for National Security fronted the media after the government had bungled its attempt to explain the technical details behind its proposed legislation to force internet service providers to store non-content data for two years to aid law enforcement. Here's a press report of that press conference:

> "They confirmed they were after source IP addresses - the identifier of a user's connection to the internet - as opposed to destination IP addresses - the web pages and services they are connecting to - alongside non-content call data. "We have been accessing that data for many years legally, and all that we are changing actually, or seeking to change, is that the data - which is held by the companies for a commercial purpose for billing or other reasons - be held in such a way that we can continue to have access to it in an environment where that access has begun to diminish a little bit," Irvine said. Both clarified that law enforcement and security intelligence agencies were only legally allowed to access source IP addresses under non-warranted metadata requests. Any instance of an IP address pointing to a URL is classified as 'content' and would require a warrant to access." "Under metadata authorisation, law enforcement and security cannot access that data. If that was provided by ISPs it would not be permissible for us to use. That is content material," Colvin said.
>
> http://www.itnews.com.au/News/390823,afp-asio-asked-to-clarify-metadata-mess.aspx

So what's this all about? What's "metadata" in this context?

"Metadata" is, strictly speaking, data about data. It appears that the term came from the library folk, who used the term to describe what was in their catalogues, providing an abstract of the original material's content and context. In a data-dense world metadata is everywhere. It is attached to digital images to describe when, where and how the image was taken. It is often used in web content, in the belief that search engines use meta-terms in web content to seen their search engine's page ranking algorithms. It is not so generally used in the context of communications, and there is no agreed meaning of the term in that context. So what exactly were they referring to when they were asking the country's ISPs to retain this "metadata" for two years?

It's often that case that the Internet has borrowed many of its terms, and even underlying concepts, from telephony, and this appears to be no exception. Every attached device in the telephone network has a phone number. If you include the various international and area prefixes, this telephone number is, in theory, unique. And the theory goes that every telephone device is associated with a human user of one sort of another.

The simplest form of abstracted (or "meta") data in this network is the white pages telephone directory. There the human users are listed with their corresponding phone number. Typically what is circulated and used was the forward translation, or name to number, but of course reverse databases also exist, that can map a telephone number to a name.

Does the Internet has an analogous facility? Is there a directory where you can enter a name, or an IP address and it returns the IP address or end user details respectively? Well, yes, there is. There is a venerable service, called *whois*, which is an approximate analogy to this telephone directory. If you master the various forms of magical incantations of this application you can query an IP address, and it will return the details of the entity to whom the address was assigned, or you can entry a name and it will provide matches of this name against address assignment records. Like the white pages, these *whois* serves are also publicly available.

So are these *whois* records the metadata that the security agencies are after?

Probably not.

Generally these *whois* records only go as far in detail as the Internet service provider, and will not in general identify the end user. So if you use the *whois* tool to query an IP address it will probably provide the details of the communications service provider, but not the end user who was actually using the address. All these public *whois* records can tell you where to go to find out more, assuming that the ISP will tell you anything, but they are not, in and of themselves, useful to identify the end user.

What records does the ISP normally hold? In the very early days of the Internet each user had their own permanent address assignment, a lot like telephone numbers. But the dial-up Internet changed all that, and we moved on to the concept of temporary, or dynamic addresses. When you were connected to the Internet you were assigned a public IP address, but when you hung up the modem then the address was reclaimed, and would be re-assigned to another client. The next time you dialed in, chances were that you were assigned a different public IP address for this session. This concept of dynamically assigned addresses was taken through to the DSL internet, and to the cable modem Internet and it persists in all kinds of forms to this day. Given that most forms of wired connection are permanent connections then there is little technical need to perform dynamic address assignment. But once the practice of dynamic address assignment was established the industry monetized it, and charged extra for permanent addresses. So if you didn't pay a "business service" premium, then the ISP would give you a dynamically-assigned IP address.

How does the ISP keep track of which IP address has been assigned to which user, and across what time periods?

In this case the ISP uses what is commonly referred to as "Radius Logs" which provides a record of which IP address was assigned to which customer and when. These logs are usually private, but are accessible under conventional warrants and other forms of legal discovery as used by agencies. The query is typically of the form of an IP address and an approximate time, and query response from the Radius logs should reveal a username. This username can be used as a lookup key in to the ISP's customer database to generate the details of the end user who was using that IP address at that particular time.

This is not all that different from a reverse phone directory. You query it using a number and get back details of a user. It does not hold any details of the user's actual network activity per se. No details of web sites visited or their IP addresses. No details of any online activity at all, other than the general information that the user was "connected" at that time. And no details of any content that was passed between the user and anyone else on the network.

In terms of intrusive surveillance this sounds pretty innocuous, and about as innocuous as a reverse telephone directory. If this is all the metadata debate was all about then its time to move along as there really is nothing to see here other than some Australian politicians getting confused between names and numbers.

But it's not as simple as that.

While the model of each user being given the exclusive use of an IP address for an extended period of time was a widely used model for many years, that model assumed that there were enough IP addresses to go round. Unfortunately, that's no longer the case. We've run out of IP addresses (of the version 4 variety), and increasingly users are being forced to share IP addresses with other users. This means that the question "who was using this IP address at this time?" does not have a unique answer. Indeed, it may not even be a small answer of just one or two users, It may be that thousands of users were using the same public IP address at precisely the same time. And that means that when a regulation is enacted that forces ISPs to record logs to allow IP addresses to be traced to individual users for the past 2 years, for some ISPs that is a very different proposition than just keeping hold of the Radius Logs and the client data base for a couple of years.

When I look at my mobile device, I'm not using a public IP address. The address my device has is one drawn from a private address pool, network 10. When I use my mobile device to connect to a web site, the remote IP address it records for "me" is not this private network address, as that address is hidden from the network. The address is drawn from the provider's address block. If I wait for a few minutes, and connect to the same web site IP address that use for this next connection is different. And if I keep doing this, I'll use more IP addresses. There is now no single IP address that is uniquely associated with my mobile device. The public IP address may change rapidly. In some cases the public IP address may change each and every time my device accesses the network. And the address I use is not necessarily uniquely "mine" at any point in time. Others may be using that same IP address at the same time. Networks that behave in this way use a technology called "Carrier Grade NATS." And very commonly they are used to support mobile services networks.

How would a mobile network service provider with a CGN-equipped network, or any other CGN-equipped IP service network for that matter, be able to generate a log of the IP addresses used by each user? Radius logs are useless in this context, so to meet a regulatory requirement to uniquely identify an end user from public IP address records, the service provider needs to save the CGN log. The is a much larger log, as it details each connection made by every end user device, listing the time the connection was made and the time it was terminated and the public IP address used by the end user device. But that alone not going to be enough information to disambiguate one user from another. As we noted above, the objective of the CGN is to allow the same public IP address to be used by multiple users simultaneously. So how can you tell these users apart? If the intent of these regulatory measures is to allow end users to be unambiguously identified from ISP logs then the public IP address

is not good enough. One possible response might be to look at the NAT function, and record the TCP or UDP port addresses as well. But even this may not be enough as we gain more experience with CGNs. The most effective form of address-reuse and the one approach that maximizes the efficiency of use of each IP address is to use 5-tuple NAT bindings in the CGN. In this case it's necessary to use not only the source IP address, **but also the destination IP address**, as the only way to effectively disambiguate one customer from another. If the objective here is to put laws in place that requires ISPs to retain data that uniquely associates individual end users with IP addresses that they use on the public Internet, then in those cases where CGNs are being used, then the agencies who are after this data may not specifically require the collection of destination IP addresses, but that's what a CGN-operating ISP would need to collect and retain to comply with the overall provisions of such legislation.

This takes on a rather disturbing complexion, as we are no longer talking about a simple directory of IP addresses and end users. We're forcing those ISPs who operate CGNs to collect and retain a complete and comprehensive log of all connections, or a complete "call log" in telephone parlance in order to allow IP addresses to be mapped back to individual users. Now this particular metadata is all about what each and every user is doing on the Internet. This is every users' browsing history and more. Everything you do when you behind a CGN generates a CGN NAT binding record, and the complete binding record, including the destination IP address details need to be logged in order to disambiguate the use of the source IP address.

How prevalent are CGNs in today's Internet? At the moment we see them used most commonly in mobile networks. But the problem today is that while the Internet continues to grow, the supply of IP addresses stopped back in 2011. The only way IP networks can continue to grow at present is to share IP addresses across multiple customers, and the way we can do that is through various permutations of CGNs. We can expect more CGNs, not less, and we can expect this problem to get worse for as long as we persist in supporting the Internet using IPv4.

Maybe the politicians are right to be confused, as it certainly appears that the Director General of ASIO and the Deputy Commissioner for National Security at the Australian Federal Police appear to have an disturbingly incomplete picture of how Internet networks are constructed in these days of IP address exhaustion. Trying to make the claim that this form of "metadata" retention to permit public IP address records to identify end users will not morph from a simple log of assigned customer IP addresses into a comprehensive surveillance program that compels ISPs to capture and retain a comprehensive log of each user's online activity falls foul of technical credibility when we consider the situation of IP service networks that make use of Carrier Grade NATs. Perhaps there is something to worry about after all.

Disclaimer

The above views do not necessarily represent the views or positions of the Asia Pacific Network Information Centre.