November 2014
Geoff Huston

Who's Watching?

Much has been said over the pasts year or so about various forms of cyber spying. The United States has accused the Chinese of cyber espionage and stealing industrial secrets. A former contractor to the United States' NSA, Edward Snowden, has accused various US intelligence agencies of systematic examination of activity on various popular social network services, through a program called "PRISM". These days cloud services may be all the vogue, but there is also an emerging understanding that once your data heads off into one of these clouds, then it's no longer necessarily entirely your data; it may have become somebody else's data too. And the rules and protocols relating to third party access to what used to be your data is no longer necessarily the rules and protocols as defined by your country's legislative and regulatory framework. Other rules and protocols that are used in other countries may apply for third party access to what used to be your data. And perhaps if you are not a citizen of this other country you may have few, if any, rights regarding the privacy of this data, or any rights regarding the secure handling of personally identifying information in this foreign regime.

Obviously, all of this has caused much public debate. For various intelligence agencies the Internet represents what they claim is an essential source of valuable information. This information, they say, is vital to their work of protecting the security and safety of the citizens of their country. For others this information gathering activity represents an abuse of privilege and power, as the more traditional process of judicial oversight and various checks and balances in executing warrants to eavesdrop on individual's activities appear to have been discarded in what looks to be an undisciplined rush to exploit this rich vein of online information.

Doubtless, this is a debate that will continue for many years to come, as finding the appropriate balance between these often conflicting interests is never an easy task. However, much of this public debate is carried out with a paucity of hard information. How is this online snooping carried out? Who is looking at whom? Can we see this digital snooping happen?

We saw an inadvertent instance of this form of online snooping when, in June 2012, a major Australian carrier, Telstra, appeared to breach the provisions of national telecommunications legislation when they apparently configured equipment in their mobile data network that intercepted customer's web fetches and sent a copy of these intercepted URLs to a third party located in the United States. Telstra gave every appearance of being unconcerned about this when they called such digital stalking "a normal network operation," while others appeared to be very concerned about the abuse of the carrier's role in performing such unauthorized eavesdropping on customers' traffic (see http://bit.ly/1u7kkzH for my perspective on this incident).

A year later, and with allegations of various forms of cyber spying flying about, it's probably useful to ask some more questions. What is a reasonable expectation about privacy and the Internet? Should we now consider various forms of digital stalking to be "normal"? To what extent can we see information relating to individuals' activities online being passed to others?

That last one is an interesting question, and in particular it's a question where we might be able to provide a small amount of data about such trafficking of information.

In our efforts to measure the extent of deployment of IPv6 and DNSSEC we present URLs to some 800,000 users each day, and we use the online ad delivery networks to try and ensure that these users are drawn in a relatively random fashion from across the entire Internet. All these URLs refer back to our server, and as each generated URL includes unique components within the DNS name part, we would expect to see at the server that each unique URL is used just once, and by one unique client. After all, it's a common expectation on the part of many Internet users that the web sites that your system contacts is essentially private information, so when you visit a web site using a unique URL, you would not conventionally expect a third party to eavesdrop on the session and capture this URL.

If this was truly the case, then each URL that we hand out to clients as part of our measurement program would be used once, and only once, and only by the client that received the URL. And most of the time that's exactly what we do see. But at times we see that the same unique URL is being used more than once. What can we understand from these cases? Are we seeing evidence of various forms of digital stalking?

Firstly, lets look at just one instance of potential stalking to illustrate how this data can be used to identify such activity.

```
10:21 120.194.53.0 GET /1x1.png?t10000.u3697062917.s1390349413.i333.v1794.rd.td
11:29 221.176.4.0  GET /1x1.png?t10000.u3697062917.s1390349413.i333.v1794.rd.td
```

It seems that this particular has been fetched twice, with a 68 second gap between the two.

```
10:21 120.194.53.0 – Origin AS = 24445 CMNET-V4HENAN-AS-AP Henan Mobile Communications Co.,Ltd
68 seconds later  -- SAME URL, different IP from a different network!
11:29 221.176.4.0 – Origin AS = 9808  CMNET-GD Guangdong Mobile Communication Co.Ltd.
```

That was a single instance of a form of stalking. What do we see across a far larger data set?

In the first 248 days of 2014 we presented some 123,110,633 unique URLs to clients. Most of these URLs were presented to the server from a single client IP address, as we would expect, but over this period some 317,309 URLs were presented to us more than once, from different client IP addresses. In some form or fashion the original fetch of the set of URLs from a client's IP address was subsequently duplicated using a different IP address. That's a stalking rate of around 1 on 400 of URLs, which, if this truly is an indicator of the level of digital stalking in todays Internet, then it's a disturbingly high figure.

Where is this happening? Are there locations where there is a higher rate of URL stalking then elsewhere. One way to answer this is to look at the rate of URL stalking per country. This is shown in Table 1, for the top 20 countries.

Rank	CC	Samples	Stalked	Rate (per 1M)	Country
1	IR	674	111	164,688	Iran
2	LA	28,506	2,875	100,855	Lao PDR
3	MO	38,761	2,954	76,210	Macao
4	SG	240,188	17,406	72,468	Singapore
5	HK	486,101	22,136	45,537	Hong Kong
6	CN	10,419,638	435,040	41,751	China
7	GB	872,124	28,845	33,074	United Kingdom
8	TW	1,769,367	36,823	20,811	Taiwan
9	JP	1,500,779	23,971	15,972	Japan
10	AU	293,193	4,620	15,757	Australia
11	US	4,491,711	53,370	11,881	United States of America
12	MY	1,035,434	10,214	9,864	Malaysia
13	AL	437,399	4,043	9,243	Albania
14	CA	947,922	6,244	6,587	Canada
15	KH	143,886	897	6,234	Cambodia

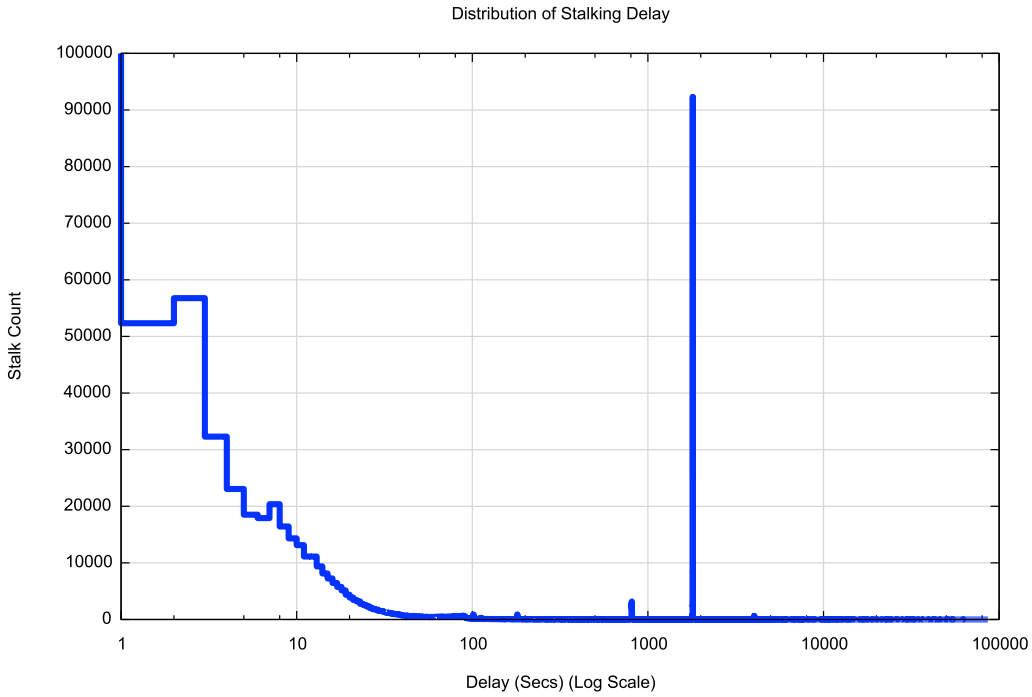
16	MM	16,411	97	5,910	Myanmar
17	MK	458,820	2,214	4,825	FYR Macedonia
18	BZ	8,139	35	4,300	Belize
19	MN	57,622	233	4,043	Mongolia
20	NZ	344,951	1,385	4,015	New Zealand

What addresses are performing this form of tracking of client activity? The second fetch was performed from 8,309 distinct source networks., and the distribution of these stalkers is far from even, as shown in the list of the top 20 stalker subnets.

Rank	IP Net	Count	AVGDly	AS
1	119.147.146.x	339,855	122.6	AS4134 CHINANET-BACKBONE,CN
2	101.226.33.x	53,181	1,502.2	AS4812 CHINANET-SH-AP China Telecom (Group),CN
3	180.153.206.x	51,592	1,528.0	AS4812 CHINANET-SH-AP China Telecom (Group),CN
4	112.64.235.x	33,067	1,470.8	AS17621 CNCGROUP-SH China Unicom Shanghai,CN
5	180.153.214.x	32,954	1,468.4	AS4812 CHINANET-SH-AP China Telecom (Group),CN
6	101.226.66.x	30,863	1,499.3	AS4812 CHINANET-SH-AP China Telecom (Group),CN
7	180.153.163.x	23,941	1,515.0	AS4812 CHINANET-SH-AP China Telecom (Group),CN
8	180.153.201.x	22,673	1,562.2	AS4812 CHINANET-SH-AP China Telecom (Group),CN
9	101.226.89.x	19,337	1,426.4	AS4812 CHINANET-SH-AP China Telecom (Group),CN
10	221.176.4.x	14,019	855.7	AS9808 CMNET-GD Guangdong Mobile.,CN
11	101.226.65.x	13,604	1,519.9	AS4812 CHINANET-SH-AP China Telecom (Group),CN
12	101.226.51.x	10,226	1,490.8	AS4812 CHINANET-SH-AP China Telecom (Group),CN
13	112.65.193.x	8,619	1,555.5	AS17621 CNCGROUP-SH China Unicom Shanghai,CN
14	66.249.93.x	8,306	31,355.1	AS15169 GOOGLE - Google Inc.,US
15	180.153.205.x	6,816	1,557.0	AS4812 CHINANET-SH-AP China Telecom (Group),CN
16	180.153.114.x	6,796	1,550.6	AS4812 CHINANET-SH-AP China Telecom (Group),CN
17	69.41.14.x	5,724	810.0	47018 CE-BGPAC - Covenant Eyes, Inc.,US
18	66.249.81.x	5,218	38,095.9	AS15169 GOOGLE - Google Inc.,US
19	66.249.88.x	4,817	31,119.7	AS15169 GOOGLE - Google Inc.,US
20	66.249.80.x	4,685	24,641.1	AS15169 GOOGLE - Google Inc.,US

Lets see if we can remove the factor of web proxy cache refresh from this data. While it's common to see web proxies behave in a mode that is not readily detectable, we also see web proxies that appear to operate in a mode that is more overt, where the proxy server appears to be given a feed of the URLs used by the community of users served by the proxy server and the proxy server separately queries the URL's server to fetch its own copy of the web object. Web proxies are very commonly deployed as a means of improving the cost efficiency of networks. What the proxy attempts to do is to reduce the extent of duplicate fetches of information to the client community that is served by the proxy. Not only does the network operator see some efficiencies in terms of reduction in total traffic loads presented to upstream transits, but also the users behind the proxy often see a much faster download time for proxy-served web objects.

So the prevalence of the use of web proxies in various developing economies in this table should not come as any particular surprise. One signal of web proxies is a cache refresh event, which is commonly set with a cache refresh timer of 15, 30 or 60 minutes. If we look at the time delay between the initial fetch and the second fetch, then a peak signal at these times would be a signal that there are web proxies at work. This is shown in Figure 1., and a strong 1800 second (30 minute) secondary fetch peak is evident in the data, with smaller signals at 900 and 3600 seconds.



Can we filter out what we assume to be the web proxies out of this data? One observation is that it is quite common to see the web proxy residing in the same Autonomous System as the client who is served by the web proxy. So what it we filter out all data where the original IP address and the shadow IP address are in the same originating AS? What does the table look like then?

Rank	IP Net	#	Avg Delay	AS
1	119.147.146.0	255,121	128.1	AS4134 CHINANET-BACKBONE No.31,Jin-rong Street,CN
2	101.226.33.0	50,257	1,543.3	AS4812 CHINANET-SH-AP China Telecom (Group),CN
3	180.153.206.0	48,808	1,574.6	AS4812 CHINANET-SH-AP China Telecom (Group),CN
4	112.64.235.0	32,800	1,507.1	AS17621 CNCGROUP-SH China Unicom Shanghai,CN
5	180.153.214.0	31,225	1,519.5	AS4812 CHINANET-SH-AP China Telecom (Group),CN
6	101.226.66.0	29,188	1,548.2	AS4812 CHINANET-SH-AP China Telecom (Group),CN
7	180.153.163.0	22,666	1,558.8	AS4812 CHINANET-SH-AP China Telecom (Group),CN
8	180.153.201.0	21,470	1,609.0	AS4812 CHINANET-SH-AP China Telecom (Group),CN
9	101.226.89.0	18,233	1,613.2	AS4812 CHINANET-SH-AP China Telecom (Group),CN
10	101.226.65.0	12,889	1,573.4	AS4812 CHINANET-SH-AP China Telecom (Group),CN
11	101.226.51.0	9,640	1,542.9	AS4812 CHINANET-SH-AP China Telecom (Group),CN
12	112.65.193.0	8,531	1,588.3	AS17621 CNCGROUP-SH China Unicom Shanghai,CN
13	221.176.4.0	8,324	749.6	AS9808 CMNET-GD Guangdong Mobile,CN
14	180.153.205.0	6,432	1,597.0	AS4812 CHINANET-SH-AP China Telecom (Group),CN
15	180.153.114.0	6,431	1,591.4	AS4812 CHINANET-SH-AP China Telecom (Group),CN
16	69.41.14.0	5,685	825.7	AS47018 CE-BGPAC - Covenant Eyes, Inc.,US
17	222.73.77.0	4,190	1,442.7	AS4812 CHINANET-SH-AP China Telecom (Group),CN
18	180.153.161.0	4,120	1,524.6	AS4812 CHINANET-SH-AP China Telecom (Group),CN
19	180.153.211.0	4,064	1,566.9	AS4812 CHINANET-SH-AP China Telecom (Group),CN
20	223.27.200.0	2,740	1.8	AS45796 BBCONNECT-TH-AS-AP BB Connect Co., Ltd.,TH

This has reduced the counts considerably, which supports the view that the predominant reason why we see duplicated URL fetches is a certain form of web proxy operation where the proxy server performs an independent fetch of the web object. When we filter out the instances of duplicated URL fetches where the original and the duplicate fetch IP addresses come from the same network (the same originating Autonomous System) the what is left appears to be systems located in China (18 of the top 20 are located in China), with the other two in the US and Thailand.

It is still feasible that these are proxy web servers, performing the proxy function for "downstream" networks. However, we also see a slightly different motivation for URL tracking in this list. On this list is a web filtering service located in the United States, Convenant Eyes (http://www.covenanteyes.com), where the intended functionality is that a feed of all URLs visited in a client system is sent "in an easy-to-read report to someone you trust," to quote their web site. It appears that the system also fetches these URLs as part of the reporting service.

The next filter I'll use on this list is to use the country of origin, and filter out all those instances where the client and the duplicate fetch system use IP addresses that are located in the same country. The resultant list is that of a set of servers who fetch a URL that was already fetched by a client, and where the client and this duplicate fetch server appear to be located in different countries.

Rank	IP Net	#	AVG Delay	AS
1	119.147.146.0	205,033	130.7	AS4134 CHINANET-BACKBONE,CN
2	101.226.33.0	6,198	1,576.1	AS4812 CHINANET-SH-AP China Telecom (Group),CN
3	180.153.206.0	6,120	1,608.3	AS4812 CHINANET-SH-AP China Telecom (Group),CN
4	180.153.214.0	3,827	1,561.0	AS4812 CHINANET-SH-AP China Telecom (Group),CN
5	112.64.235.0	3,819	1,544.9	AS17621 CNCGROUP-SH China Unicom Shanghai,CN
6	101.226.66.0	3,603	1,577.3	AS4812 CHINANET-SH-AP China Telecom (Group),CN
7	180.153.163.0	2,742	1,540.1	AS4812 CHINANET-SH-AP China Telecom (Group),CN
8	223.27.200.0	2,740	1.8	AS45796 BBCONNECT-TH-AS-AP BB Connect Co., Ltd.,TH
9	101.226.89.0	2,658	2,230.2	AS4812 CHINANET-SH-AP China Telecom (Group),CN
10	180.153.201.0	2,628	1,549.4	AS4812 CHINANET-SH-AP China Telecom (Group),CN
11	101.226.65.0	1,528	1,573.3	AS4812 CHINANET-SH-AP China Telecom (Group),CN
12	69.41.14.0	1,243	1,127.4	AS47018 CE-BGPAC - Covenant Eyes, Inc.,US
13	101.226.51.0	1,195	1,627.6	AS4812 CHINANET-SH-AP China Telecom (Group),CN
14	112.65.193.0	1,038	1,623.9	AS17621 CNCGROUP-SH China Unicom Shanghai,CN
15	64.124.98.0	906	1,288.9	AS6461 ABOVENET - Abovenet Communications, Inc,US
16	180.153.114.0	819	1,632.6	AS4812 CHINANET-SH-AP China Telecom (Group),CN
17	180.153.205.0	765	1,497.7	AS4812 CHINANET-SH-AP China Telecom (Group),CN
18	208.184.77.0	649	1,419.5	AS6461 ABOVENET - Abovenet Communications, Inc,US
19	222.73.77.0	535	1,373.8	AS4812 CHINANET-SH-AP China Telecom (Group),CN
20	180.153.211.0	517	1,450.6	AS4812 CHINANET-SH-AP China Telecom (Group),CN

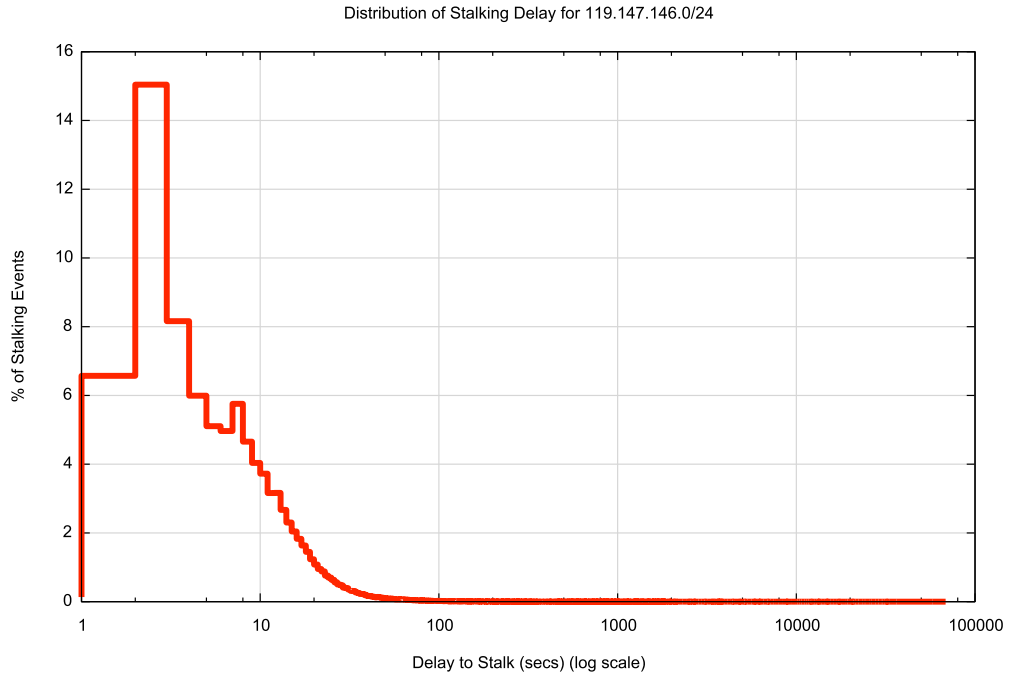
That first entry is quite exceptional. In the 248 day data collection window we saw some 205,000 instances of this duplicate URL fetch , while the second highest count was far lower, at 6,198 instances.

Lets take a closer look at the actions of the 119.147.146.x system. In what countries were the original clients located? The somewhat surprising answer is that almost every country is represented in this list. Whatever is happening here, there appears to have been a deliberate effort to sample web traffic from users located in almost every country. There are some countries, however, that see a higher rate of URL stalking by this particular stalker. Here's the top 25 countries where users that appear to be located in this countries are being stalked by the system located at 119.147.146.x.

Rank	**CC**	**Stalk Count**	**Country**
1	CN	136,402	China
2	TW	29,247	Taiwan
3	JP	23,174	Japan
4	HK	17,105	Hong Kong
5	SG	16,350	Singapore
6	GB	16,056	United Kingdom
7	VN	9,191	Vietnam
8	MY	9,077	Malaysia
9	US	8,524	United States of America
10	TH	4,529	Thailand
11	PL	4,114	Poland
12	AL	4,032	Albania
13	TR	3,465	Turkey

14	AU	3,463	Australia
15	PH	3,281	Philippines
16	CA	3,164	Canada
17	MA	3,111	Morocco
18	RO	2,990	Romania
19	RS	2,672	Serbia
20	BG	2,544	Bulgaria
21	MO	2,323	Macao
22	DZ	2,288	Algeria
23	MK	2,210	FYR Macedonia
24	ID	2,103	Indonesia
25	MX	1,928	Mexico

This particular stalker is echoing the original fetches within 3 seconds of the original fetch, which indicates that the stalking point is remarkably close to the user, perhaps right inside the user's browser.



What do we know about the stalker. One thing we have observed is that it uses a consistent User Agent string when it retrieves the URL. The string reveals that this system is reporting itself to be an instance of the Maxthon web browser.

What about the stalked victims? There is a fair deal of variety here, but we can list the top 10 most commonly used User Agent strings.

Rank	Count	User Agent String
1	6,068	Mozilla/5.0 (Windows NT 5.1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/28.0.1500.95 Safari/537.36 SE 2.X MetaSr 1.0
2	5,458	Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/28.0.1500.95 Safari/537.36 SE 2.X MetaSr 1.0
3	5,389	Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/33.0.1750.154 Safari/537.36
4	5,029	Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/32.0.1700.107 Safari/537.36
5	4,669	Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/28.0.1500.95 Safari/537.36 SE 2.X MetaSr 1.0
6	4,641	Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/31.0.1650.63 Safari/537.36
7	3,382	Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML, like Gecko)

		Chrome/31.0.1650.63 Safari/537.36
8	3,265	Mozilla/5.0 (Windows NT 6.1; WOW64; rv:26.0) Gecko/20100101 Firefox/26.0
9	3,084	Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/32.0.1700.107 Safari/537.36
10	2,915	Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/32.0.1700.76 Safari/537.36

They all look to be Windows devices, but this may well be an observation about the market share of Windows as distinct from an inference that this is the result of some form of malware that has been installed on victim's systems. One part of the user agent string is visible in some entries: the presence of the substring "MetaSr 1.0". This appears to be a signature of the "sogou" browser, which appears to be a browser that supports some form of Pinyin phonetic system input.

Wikipedia (http://en.wikipedia.org/wiki/Sogou) describes this browser as one that "adopts a "dual-core" (Google Chrome's WebKit and Internet Explorer's Trident layout engines) techniques and it connects to the cloud to recognize malicious websites and software."

That last sentence may be the critical one here. It is possible that this "connection to the cloud" may be a circumspect way of saying that the browser passes URLs off to a common server setup that performs a secondary retrieval.

The most benign explanation is that there is a browser that appears to be popular within the Mandarin speaking community that leaks URLs to some form of content grading system. Less benign explanations of this observed behaviour can speculate on browsers that deliberately include spyware to track the online actions of the users who use these browsers.

In relation to the scale of the entire Internet, our analysis of some 123 million web fetches across a 248 day period represents a microscopic proportion of the Internet's activity. However, the ability to detect anomalous behaviour within this microcosm of web activity is perhaps illustrative of what we should expect on the broader Internet. While this small data set does not show any clear and incontrovertible evidence of consistent digital stalking or cyber snooping of any form, it illustrates one extremely important maxim for the Internet – nothing on the Internet is completely private. Even when encryption can, to some extent, provide some privacy protection on the content of conversations and transactions on the Internet, you should always bear in mind that the sites you go to, and the time when you go to them, form part of a readily accessible pool of data that is not private. Its all forms part of our digital exhaust fumes that trail behind us as we travel through the Internet. And it should come as no surprise to learn that there are systematic efforts underway on the Internet to sniff these exhaust fumes, and collect this data about your online behaviour and interpret and use it in various ways.

So it's highly likely that from time to time, or even more often than that, on the Internet someone is indeed watching me and you.

This article is the writeup of a lightning talk presentation (http://bit.ly/1tZG2Ek) made at RIPE69 in November 2014 (http://bit.ly/1tEyHX7)

Disclaimer

The views expressed are the authors' and not those of APNIC.

About the Author

Geoff Huston B.Sc., M.Sc., is the Chief Scientist at APNIC, the Regional Internet Registry serving the Asia Pacific region.
www.potaroo.net