PRINCETON UNIVERSITY

Nick Feamster
Professor, Department of Computer Science
Acting Director, Center for Information Technology Policy

310 Sherrerd Hall
Center for Information Technology Policy
Princeton University
Princeton, New Jersey 08540-5233
+1 609 258 2203
feamster@cs.princeton.edu

March 3, 2016

Dear Chairman Wheeler:

A report from Peter Swire entitled "Online Privacy and ISPs" recently came to my attention. I gather that this report pertains to a possible NPRM concerning privacy for CPNI. I am not taking a position in this policy debate, but certainly such important policy matters and decisions should be grounded in a solid understanding of the technical capabilities of ISPs. As I read Swire's white paper, however, I noticed many technical inaccuracies and omissions that warrant clarification and correction.

The premise of Swire's white paper is that ISPs have limited capability to monitor users' online activity, due to the fact that (1) users are increasingly using many devices and connections, so any single ISP is the conduit of only a fraction of a typical user's activity; (2) end-to-end encryption is becoming more pervasive, which limits ISPs' ability to glean information about user activity; and (3) users are increasingly shifting to use VPN traffic. On the surface, these observations reflect some basic misunderstandings of various Internet technologies, protocols, and trends.

I am an expert in Internet technologies and protocols, and I have spent nearly ten years studying user traffic in home networks. I have a deep understanding of the nature of traffic that is visible in home networks and what might be learned from it. Our research clearly demonstrates that we can learn a lot from the traffic that we can see from home networks. To claim that ISPs cannot learn about user activity from the traffic they can see is simply not true. It is worth noting that standard procedure acknowledges this strong capability: For example, when I conduct research using this traffic, I am not allowed to do so without first gaining clearance for human subjects research, due to the private and sensitive information that this traffic contains. We can learn so much from this traffic that we've written papers with conclusions about human behavior solely based on our analysis of the traffic that ISPs can see. I'll elaborate on some of the things that we can learn in my points below.

I'll discuss the three points in Swire's paper in turn, highlighting some technical inaccuracies and omissions and bringing out some of what we've learned in our past research, where appropriate.

First of all, the report cites increasing user mobility as the basis for the claim that a single ISP may not see a substantial fraction of user traffic. It is true that user mobility is increasing, but this by no means implies that a single ISP cannot track significant user activities in their homes, and even as they move. My technical expertise in studying home network traffic suggests, however, that there is much to be learned from the traffic that a user sends from their home network, even if they are also using other networks. The report doesn't discuss the extent to which most users rely on exclusively one Internet service provider (ISP) for home broadband Internet access, which places a substantial fraction of user activity within the purview of a single ISP; the shared WiFi access points that these ISPs now exacerbate this consolidation.

- Any single ISP may have visibility into many devices from a single subscriber, simply by virtue of having a gateway device in the home network. In a study that we performed in 2013 in about thirty home networks around the world, we saw that the median user in developed countries has about five devices connected to the home network at any given time.

- The traffic that an ISP can observe from such a gateway contains a significant amount of private information about user behavior. The same study from 2013 finds that network traffic can reveal significant information about user activity, including information about when a user is home; the number, type, and manufacturer of devices that they have connected to the network; and in some

cases even the waking and sleeping patterns of users in the home. It is worth noting that we can observe these features of user activity *even when traffic is encrypted.*

- The observation that "the average internet user has 6.1 connected devices, many of which connect from diverse locations…that are served by multiple ISPs" belies the fact that many mobile hotspot networks are run by common ISPs. As of mid-2015, for example, Comcast's Xfinity WiFi roaming service had deployed 10 million WiFi hotspots around the US, which users had accessed 3.6 billion times. The same providers are also offering voice-over-WiFi options, which will increase this convergence. While the report cites increasing mobility of users as evidence for diversification of last-mile access, this mobility alone does not imply diversification. In fact, the opposite may be true.

The report cites the fact that end-to-end encryption makes it difficult for an ISP to discern user activity due to the ISP's inability to inspect packet contents in a traffic flow. While it is true that end-to-end encryption is becoming more pervasive, this also does not by itself prevent the ISP from observing user activity from network traffic. End-to-end encryption does not conceal all information concerning user activity and traffic patterns. Although the data traffic from some devices may be encrypted end-to-end, various protocols and devices still leak a significant amount of information in cleartext.

- Nearly all Internet-connected devices use the Domain Name System (DNS) to look up domain names for specific Internet destinations. These DNS lookups are generally "in the clear" (i.e., unencrypted) and can be particularly revealing. For example, we conducted a recent study of traffic patterns from a collection of IoT devices; in that study, we observed, for example, that a Nest thermostat routinely performs a DNS lookup to `frontdoor.nest.com`, a popular digital photo frame routinely issued DNS queries to `api.pix-star.com`, and a popular IP camera routinely issued DNS queries to (somewhat ironically!) `sharxsecurity.com`. No sophisticated traffic analysis was required to identify the usage of these devices from plaintext DNS query traffic.

- Even when a site uses HTTPS to communicate with an Internet destination, the initial TLS handshake typically indicates the hostname that it is communicating with using the Server Name Indication (SNI), which allows the server to present the client with the appropriate certificate for the corresponding domain that the client is attempting to communicate with. The SNI is transmitted in cleartext and naturally reveals information about the domains that a user's devices are communicating with.

- The report cites the deployment of HTTPS on many major websites as evidence that traffic from consumers is increasingly encrypted end-to-end. Yet, consumer networks are increasingly being equipped with Internet of Things (IoT) devices, many of which we have observed send traffic entirely in cleartext. In fact, of the devices we have studied, cleartext communication was the norm, not the exception. While of course, we all hope that many of these devices ultimately shift to using encrypted communications in the future, the current state of affairs is much different. Even in the longer term, it is possible that certain IoT devices may be so resource-limited as to make cryptography impractical, particularly in the case of low-cost IoT devices. The deployment of HTTPS on major websites is certainly encouraging for the state of privacy on the Internet in general, but it is a poor indicator for how much traffic from a home network is encrypted.

The report cites the increasing use of VPNs as evidence that ISPs will have difficulty seeing user traffic. The report somewhat misunderstands how VPNs work, and probably also overstates their reach.

- First, the statement that VPNs will prevent ISPs from seeing DNS traffic depends on the configuration of the VPN tunnel. A VPN is simply an encrypted tunnel that takes the original IP packet and encapsulates the packet in a new packet whose destination IP address is the tunnel endpoint. But, the IP address for DNS resolution is typically set by the Dynamic Host Configuration Protocol (DHCP). If the consumer uses the ISP's DHCP server to configure the host in question (which most of us do), the client's DNS server will still be the ISP's DNS

server. Thus, the ISP will still continue to observe all of the user's DNS traffic, even if the user configures a VPN tunnel. It is possible for a user to configure their DNS server to not use their ISP's DNS server or to use a VPN that allows the client to use the VPN's DNS resolver, but this is by no means automatic. Even in cases where a VPN uses its own DNS resolver, the traffic for those queries by no means stay local: DNS cache misses can cause these queries to traverse many ISPs.

- Traffic from VPNs doesn't simply disappear: it merely resurfaces in another ISP that can subsequently monitor user activity. The opportunities for observing user traffic are substantial. For example, in a recent simple experiment we performed, web requests from Tor exit relays to the Alexa top 1,000 websites traversed more than 350 Internet service providers considering the DNS lookups from these exit relays, the traffic from these exit nodes traverses an additional 173 Internet service providers.

- Furthermore, VPN clients are typically for desktop machines and, in some cases, mobile devices such as phones and tablets. As previously discussed, IoT devices in homes will continue to generate more traffic. Most such devices do not support VPN software. While it is conceivable that a user could set up an encrypted VPN tunnel from the home router and route all home traffic through a VPN, typical home gateways don't easily support this functionality at this point, and configuring such a setup would be cumbersome for the typical user.

The report contains some other inaccuracies concerning the technical limitations of traffic inspection.. For example, the report claims that "limited processing power and storage placed technical and cost limits on [DPI] capability". Yet, this understates ISPs' capabilities in this context.

- Even software routers and switches can now process packets at rates of tens of gigabits per second. Hardware capabilities are even more impressive, and costs for these technologies continue to drop.

- In the last mile, data rates are substantially lower. Most home networks we have studied are sending traffic at only tens of megabits per second, even at peak rate. We have been able to perform packet capture on very resource-limited devices at these rates. While backhauling and storing all of this data surely poses a challenge, data analysis and reduction at the network edge is becoming increasingly feasible, particularly with the rise of software middleboxes and NFV, as well as the increasing capabilities of the CPE devices in homes.

- Targeted DPI (e.g., for specific applications or specific users) would also occur at much lower rates and would thus be well within the realm of most access ISPs' technical capabilities.

The technical inaccuracies concerning ISP capabilities in this document reflect some basic misunderstandings of Internet protocols, as well as the current Internet ecosystem. Such technically inaccurate descriptions can lead readers to incorrect conclusions. We must bring more technical experts to this discussion, lest we make decisions on technical inaccuracies and misunderstandings. Thank you for your consideration.

Sincerely,

Nick Feamster