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

310 Sherrerd Hall
Princeton University
Princeton ,NJ 08540-5233
+1 609 258 2203
feamster@cs.princeton.edu

RE: Docket No. 16-106, Protecting the Privacy of Customers of Broadband and Other Telecommunications Services.

Chairman Tom Wheeler
Commissioner Mignon Clyburn
Commissioner Jessica Rosenworcel
Commissioner Ajit Pai
Commissioner Michael O'Rielly
Federal Communications Commission
445 12th Street SW
Washington, DC 20554

May 27, 2016

Dear Chairman and Commissioners:

I am a Professor of Computer Science at Princeton University and the Acting Director of the Center for Information Technology Policy (CITP) at Princeton University. I have over fifteen years of expertise in networking research, in topics ranging from the engineering of large ISP backbone networks to broadband access performance measurement. In addition to my time at universities, where my work has won numerous awards, including the Presidential Early Career Award for Scientists and Engineers, I have also worked at AT&T, where algorithms I developed were used to help the backbone network operators diagnose faults and manage congestion within their networks. I also teach an online course on computer networking that has been taken by more than 100,000 students.

I write in reply to the FCC's proposed rulemaking in WC Docket No. 16-106, which aims to constrain the ways Internet service providers (ISPs) who fall under Section 222 of the Communications Act can collect and share Customer Proprietary Network Information (CPNI); it also prescribes how service providers must ask consumers to opt-in to the collection of CPNI and outlines baseline requirements for data security and breach notification. ISPs are also prevented from charging subscribers a premium for providing baseline privacy protections to consumers.

The proposed rulemaking raises significant concerns for (1) operators of ISP networks, who rely on network data to manage and secure their networks; (2) researchers, who rely on access to network data to develop new protocols and algorithms, and to study phenomena ranging from user behavior to pricing; (3) vendors and protocol developers, who rely on access to real network test traffic to debug software and protocol implementations. As written, the rule would harm all three of these stakeholders.

Section 222 does have specific exceptions that state that an ISP can disclose or permit access to CPNI "to protect users of those services and other carriers from fraudulent, abusive, and unlawful use".

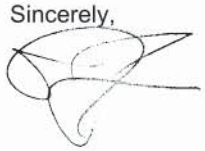
In this note, I illustrate how the proposed rulemaking might make it challenging for network operators to manage and secure their networks and may also set the research community back. In this comment, I will discuss:

- the network data that ISPs collect,
- how they use it,
- how and when it might be shared with other parties,

- if and when it might be retained
- to what extent each dataset contains private customer information

I will also point to specific studies where we have relied on network data collected from ISPs to perform a wide range of research studies. In all of the use cases I discuss in this note, the data collection does not require any opt-in from the user; doing so would significantly hamper any of the positive research efforts I outline in my note below. Of course, network data does carry privacy risks. To that end, I will discuss the privacy risks of each of the datasets I outline and make observations concerning possible approaches for protecting the user privacy in light of these risks. I have summarized all of the datasets that I discuss below in Appendix A.

Sincerely,



Nick Feamster

Section 1. How ISPs May Collect, Use, and Share Network Data

In this section, I provide a brief overview of the types of the data that ISPs may collect as part of the normal operation of running their networks. For each type of network data that ISPs collect, I discuss how network operators may use the data to help secure or operate the network.

1.1. Routing Data

Networks collect two types of routing data: data from the interior gateway protocol (IGP) which is responsible for establishing routes within the ISP's network; and data from the border gateway protocol (BGP), which is responsible for establishing routes between an ISP's network and other neighboring networks. Each type of routing data may be captured as snapshots (i.e., the state of the routing table at a given time) or as updates (i.e., a stream of data that represents changes to the state of the routing tables). Routing data generally represents the state of the network topology and thus does not contain any personal or private information concerning consumers. Therefore, the use or sharing of network routing data should not fall under the proposed rulemaking. For completeness, I briefly describe below how these two types of data are used to help network operators run their networks.

BGP Data. Information from BGP routes help network operators determine where traffic from their network is destined. In other words, the BGP routes can help network operators determine the paths that traffic takes as it traverses and leaves a network. Network operators can also affect BGP routes to help improve the performance of the network. For example, operators might change routing configuration to affect how traffic flows to neighboring networks to improve performance, relieve congestion, and so forth. This process is sometimes referred to as *traffic engineering*. In addition to using (and manipulating) BGP routes to improve network performance, operators also use the routes that they learn from other networks for security purposes; for example, BGP routes can help detect spoofing, by helping determine whether the purported source of network traffic is legitimate. In my own research, we have developed techniques to help network operators identify and takedown criminal "bulletproof" hosting domains, which can often be identified by unusual or anomalous routing patterns.

IGP Data. Intradomain routing information contains information about an ISP's internal network topology and how it changes over time, in response to failures and reconfiguration. The IGP carries only information about an ISP's internal network topology. A network operator can tune the IGP routing protocol configuration to improve the performance of the network (e.g., by better balancing traffic load across existing links). As with BGP routing information, IGP routing data does not contain customer information and should not be subject to the proposed rulemaking.

1.2 Traffic Data

Another class of data that ISPs collect is data about the *traffic* that traverses an ISP's network. In this section, I describe the data that ISPs may collect, how it helps them operate and secure their networks, and the potential privacy concerns associated with this data. I enumerate the types of network traffic data that ISPs collect to help operate and secure their networks. I also discuss how and when this data might be shared, as well as the potential privacy concerns associated with each class of data.

1.2.1 IPFIX Data

ISP routers are commonly configured to collect IPFIX data (sometimes referred to as "NetFlow", which refers to Cisco's implementation of the standard). IPFIX data captures statistics about each traffic flow that

traverses a that router. A *flow* is a set of packets that share the same source and destination IP address and port combination. IPFIX data records, for each flow, the start and end time for the flow, the number of bytes and packets in the flow, the protocol (i.e., TCP or UDP), and the source and destination of the flow. IPFIX statistics are often derived from samples of the traffic that traverses the router. For example, out of 1,000 packets that traverse a router, only one might be sampled for inclusion in IPFIX statistics. As a result, many short flows may not be recorded whatsoever.

Network operators use IPFIX data for a variety of performance and security tasks. The flow-level statistics can help network operators identify the existence of malware, networks of compromised machines, or other anomalous activity (including failures). Network operators may also use IPFIX data to facilitate planning, provisioning, and traffic engineering. In several projects at AT&T, we used a combination of routing and traffic data to help network operators predict how certain configuration changes might affect link utilization in different parts of the network. In an ongoing project with seven cable ISPs in the United States, I am using IPFIX data to study utilization patterns at interconnection points.

In some cases, network operators may also share IPFIX data with third parties. A common operational example is for denial of service (DoS) attack detection. Because IPFIX data can help operators detect anomalous traffic shifts that represent DoS attacks, ISPs sometimes share this data with third-party DoS detection and mitigation services (e.g., Arbor Networks). In this instance and others, sharing this data may be critical for protecting the users of the ISP and other carrier networks from abuse and attack.

Network operators may also share IPFIX data with researchers. I use IPFIX data collected at interconnection points to analyze utilization patterns. In another project related to DoS mitigation, we are using IPFIX data to better understand traffic attack patterns. In the past, we have also used IPFIX traffic traces from access ISPs to design and validate algorithms to detect botnets, large networks of compromised machines. Most recently, I have been using IPFIX data collected at the interconnection points from seven access ISPs in the United States—covering 50% of the US broadband subscriber population—to explore the characteristics and patterns of utilization between access ISPs and edge providers. Interestingly, this type of project that provides *exactly* the type of insight and analysis that the FCC is increasingly paying attention to. Preventing ISPs from sharing this type of data with researchers would impede progress on this research, ultimately inhibiting the public's visibility into ISP interconnection.

Preventing ISPs from collecting this data and sharing it with vendors of security services *or researchers* will harm the security and performance of the Internet and threatens to inhibit research innovation.

Privacy concerns. IPFIX data does carry privacy risks. Because it captures information about the destinations that an individual user visits, as well as the port numbers for that traffic, the statistics can carry information about user behavior, such as the websites a user is visiting, the applications they are visiting, and their general usage patterns. Traffic volume is also indicative of user behavior: traffic volumes can reveal whether a user is at home or away, whether the user is awake or asleep, and so forth. In my own research, we have demonstrated that even "metadata" about user traffic from a home network can be very revealing about user behavior; it can reveal everything from the devices that a user has and how they are being used, to fine-grained information about user presence and activity. In short, even though IPFIX records contain no information about the actual content of communication, information such as volumes, sources, and destinations can sometimes reveal private information about user behavior.

1.2.2 DNS Queries

The system that maps Internet names to IP addresses is called the Domain Name System (DNS). Access ISPs have access to traffic traces that represent the DNS names that users look up when visiting different Internet destinations. For example, if a user visited google.com, that user's browser would first need to generate a DNS query for google.com's IP address; typically, the user would send that DNS lookup to a DNS resolver in the ISP's network. As a result, the ISP might have considerable information about the DNS queries of individual users.

Because bots and malware that run on user machines typically rely on the DNS to "phone home" to control machines that coordinate attacks, DNS query information has proved to be incredibly useful in helping ISPs detect user compromise and infections. For example, DNS lookups to otherwise unpopular domains may indicate that a machine is compromised. In general, changes in DNS lookup patterns may indicate that a device is behaving incorrectly or may be compromised. In our own research, we have developed techniques to help network operators identify and shut down malicious DNS domain names that are linked to attack sites, such as phishing websites. We have also demonstrated that DNS lookups can serve as an early warning signal to identify spamming botnets, as sometimes the bots will issue DNS queries for DNS domains that they plan to use in future attacks.

ISPs often share this type of DNS query data with third parties who provide security services (e.g., botnet detection services). Damballa is one example of such a third-party service; the botnet detection that this service performs is largely based on analysis of DNS lookup information from users. Without access to such lookup information, the security of the network and the safety of users could be at risk.

Privacy concerns. As with IPFIX data, DNS lookup information carries privacy risks. A user's DNS lookups can reveal activity patterns, the website that a user is visiting, and (due to website fingerprinting attacks) possibly even the web *pages* that a user visits. DNS data can be incredibly revealing about user behavior and activities, as a result. This concern is likely to grow as consumers increasingly deploy IoT devices (e.g., thermostats, smart plugs) in their homes, as the DNS and IPFIX traffic from these devices may reveal an increasing amount of information about user behavior and activity.

1.2.3 Interface Byte Counters

Network operators often collect interface byte counters using a management protocol called the Simple Network Management Protocol (SNMP). SNMP counts the number of bytes that traverse a particular router or switch interface. SNMP typically "polls" each interface at a regular, fixed interval to record these byte counter values; a common polling interval is five minutes. SNMP byte count information can be extremely helpful for network operators in determining the level of utilization on a particular link or interface during a given period of time. This utilization information can be helpful for improving both the security and performance of the network. For example, byte counters can indicate increased utilization that might indicate the need to provision additional capacity; shifts in utilization may also indicate that a particular node or link in the network has failed, causing a significant shift in traffic. Byte counters are also useful for detecting denial of service (DoS) attacks.

Privacy concerns. Because SNMP byte counters are per-interface and do not contain information about IP addresses, the byte counters by themselves do not typically contain information that could be revealing about the patterns of a particular user. The only scenarios one might be cognizant of are if an SNMP byte counter were collected on a per-user granularity (e.g., at the CPE). Additionally, when combined with routing information, there may be certain cases where some information could be learned about usage patterns, but

in general it would be extremely difficult to extract any customer-specific information from SNMP interface byte counters.

1.2.4 Deep Packet Inspection

Deep packet inspection (DPI)—which may monitor and collect information from individual packets that traverse the network, including the contents of data traffic—has received disproportionate attention.

I believe that DPI is a red herring for two reasons. First, DPI is typically not widely deployed in many ISP networks. Several ISPs have stated in various forums that DPI capabilities are deployed on less than 10% of the link capacity in an ISP network; even if DPI were widely deployed, the cost of retaining the traffic that could be collected from DPI for any length of time would be prohibitive. Second, contrary to some conventional beliefs, ISPs often do not retain much of the data that they collect because the cost of doing so can be substantial; at some of the networking companies I have worked for, we have, in some cases, had to argue stridently that certain data be retained so that we could use it for a study or a research project.

Although the DPI discussion is a bit of a distraction, and although there are many uses of DPI that help operators manage and secure their networks, I also think it is worth pointing out that the discussions on encryption in this space are somewhat misguided, and are not addressing the real issues at hand. *Simply put, encryption and VPNs do not guarantee user privacy in the context of CPNI.* A recent report from Professor Peter Swire indicates that increasingly pervasive encryption makes it more difficult for ISPs to see certain user behavior, but Swire's report, while not incorrect, also does not tell the complete story:

- Many consumer IoT devices do not use end-to-end encryption.
- HTTPS/TLS connections still require a handshake where the SNI (i.e., the domain that a user is connecting to) is visible in cleartext.
- As mentioned in previous sections, the other data that ISPs can collect is still incredibly revealing.

The argument that we don't need to be concerned about DPI because of increasing deployment of end-to-end encryption is disingenuous and does not present the entire picture.

Swire's arguments for why Virtual Private Networks (VPNs) reduce the concern over DPI are also somewhat incomplete, for the following reasons:

- Many VPNs are difficult and unintuitive to use
- VPNs typically degrade performance and user experience, by taking user traffic on an indirect path that increases network latency.

Finally, the idea that users are increasingly mobile and thus will be more difficult for ISPs to track is not necessarily the case; take, for example, the community WiFi services that ISPs are increasingly deploying (e.g., CableWiFi, OptimumWiFi, XfinitiyWifi). In all of these cases, subscribers are connecting to their ISPs' networks from an increasingly diverse set of locations. If anything, users' tendencies to use the same ISP network as they move from one place to another makes it likely that ISPs will have *more* data about consumers, not less.

I have outlined some of these points in a previous letter to Chairman Wheeler. In summary, I think the public discussion thus far on DPI has been largely misguided. On the one hand, DPI is not as pervasively deployed as some might think, and DPI data is not retained as aggressively as many seem to think. Additionally, there are many valuable uses for data collected with DPI, ranging from protocol implementation testing to network research. On the other hand, encryption, VPNs, and mobility do not appreciably mitigate any privacy concerns. In summary, I do not think we need to be any more concerned about DPI than any of

the other network traffic types I have outlined earlier in this note, which are both incredibly useful and at times also can reveal information about consumers.

Section 2. Research Use of Network Data

In my fifteen years as a network researcher, I have used network data countless times to develop tools, systems, and protocols to help network operators manage and secure their networks. In some cases, the data has been directly provided by an Internet Service Provider (ISP). In other cases, we have built systems to collect the data ourselves. In many of these cases, the data that we have been privy to has contained sensitive information about individual users. Universities have processes in place that govern how any data that involves humans is collected, retained, and used; specifically, a university's Institutional Review Board (IRB) reviews the process that a researcher implements to collect such data and the processes that are in place to use it.

For the sake of illustration, I list just a few projects where I have received or used data of the nature described above. I have used:

- email traffic traces from campus, enterprise, and large ISP networks to develop new mechanisms for spam filtering.
- IPFIX data from a large backbone ISP to test new traffic engineering tools.
- DNS lookup data from access ISPs and enterprises to design algorithms to detect botnets and other network abuse.
- SNMP interface byte counters to study how traffic demands change in response to different access network provisioning strategies.
- IPFIX and DNS data from home networks to study user behavior and usage patterns, including how users respond to data caps imposed by ISPs
- IPFIX data, to study traffic utilization patterns at ISP interconnection points
- DPI data (i.e., packet traces) to study how different traffic sampling algorithms can affect the fidelity of traffic statistics for use in abuse and attack detection

In many of these cases, I either received the data from an ISP or I built and deployed a system to collect this data from the network. This list is not exhaustive; it is meant to illustrate the tremendous value that researchers derive from ISP data. **Preventing ISPs from sharing network data with researchers will prove to be a tremendous setback for innovation.**

Section 3. Explicit Recommendations in the Context of Proposed Rulemaking

I would like to make the following recommendations to the FCC in light of the points above.

Section 3.1 Exceptions

It should be apparent from the discussion above concerning how operators and researchers use network traffic data that the rulemaking should provide certain exceptions concerning collection and use of network data:

- *The rulemaking should provide an explicit exception for researchers.* As described above, network research fundamentally depends on cooperative data sharing agreements with ISPs.
- *The rulemaking should also provide an explicit exception for protocol developers and vendors.* Protocol developers and vendors often need real packet traces from ISPs to test for correct functioning and interoperability. The inability to receive traffic traces from ISPs will severely limit

vendors' and developers' ability to build and deploy network technology that functions correctly, safely, and securely. Some examples where limitations on such data sharing would have impacted development and deployment include protocols such as IPv6 and DNSSEC.

- *The rulemaking should provide an explicit exception for vendors who provide third-party security and network management services.* Many vendors provide third-party services to help network operators operate or secure their networks. ISPs should be permitted to route network traffic to and through these third-party services to the extent that doing so can make the network operate better or more securely.

Section 3.2 Choice, Transparency, and Security

The proposed rulemaking notes the need for consumers to have choice, transparency, and security when it comes to ISPs' collection and use of network traffic data. In light of the points above, I make the following recommendations:

- *When correctness, performance, or security is a consideration, ISPs should not be required to seek customer consent to collect the data, even if it contains potentially private information.* There are many situations where consumers cannot be expected to exercise "meaningful and informed control" over the circumstances in which their data is shared and used. Much of the network data above (i.e., IPFIX, DNS) **must** be collected, analyzed, and shared with third parties to ensure the correct and secure operation of the network. Imposing such a requirement would cripple network operations and security.
- *ISPs should be transparent about what they collect, to the extent that doing so is practical.* Consumers should know what information ISPs collect, and *why* that data is collected. ISPs could and should reasonably disclose the nature of the routing and traffic data that they collect—and for how long such data is retained—to the extent that making this information public does not place them at a competitive disadvantage.
- *ISPs should take precautions to protect data that may pose privacy risks to consumers.* While this recommendation almost goes without saying, it is also worth pointing out that some network data poses more privacy risks than others. For example, unencrypted packet traces are far more sensitive than SNMP interface byte counts. The level of protection—and the extent to which breach notification is required—should be commensurate with the level of risk that such a breach would pose to consumers.

Section 4. Summary and Conclusion

ISPs collect, use, and share a variety of network data to operate and secure their networks. ISPs also share these datasets with researchers in an effort to shed more light on network operations and security, user behavior, and economics. Requiring notification and opt-in for many of these datasets would hinder network operations and research; when data collection and sharing relate to performance, security, or research, user opt-in should not be required. Although each of these datasets provides opportunities unique and valuable insights, they do also carry variable privacy risks, which are not mitigated by end-to-end encryption, VPNs, or user mobility. In light of this, ISPs should certainly take measures to protect data that could pose risks to user privacy, but those measures should be commensurate with the risks that the data poses to consumers..

Appendix A. Table Of Data that ISPs May Collect, How it is Used, and Possible Privacy Risks

Data	How Used for Security and Performance	Possible Privacy Risks to Individuals
Routing and Addressing Data		
Routing Updates and Tables • BGP (interdomain routing data) • IGP (internal routing information)	· Security o Spoofing detection: Can be used to determine whether the purported source of network traffic is legitimate or spoofed o Identification of “bulletproof hosting domains” (useful for takedown) · Performance o Traffic engineering	· Not many (?): Routing data reveals properties of the network topology, but nothing in particular about user behavior
Traffic Data		
IPFIX/NetFlow (“Metadata”) • High-level statistics about individual connections (often highly sampled in large ISPs)	· Security o Detection of malware, botnets,□anomalous activity · Performance o Traffic engineering (ensuring that certain traffic flows do not congest links) o Provisioning and planning	· What destinations does an individual user visit? · What applications are they using, when,□and for how long? · Human behavior: o Is the user at home? Awake or asleep? o What devices are they using, and when?
Domain Name System (DNS) Lookups	· Security o Malware, bot detection o Other anomalous activities o Early warning for attacks (“DNS counter-intelligence”)	· Activity patterns · What sites are visited · Possibly which web *pages* are visited (fingerprinting)□
Deep Packet Inspection (DPI) *	· Security o Intelligence (fraud, abuse) o Traffic scrubbing · Performance o Application and protocol developers (scalability testing, user demand) o Implementation debugging	· [End-to-end encryption addresses some, but not all privacy risks.] o Many IoT devices do not encrypt o SNI header in TLS handshakes o TLS fingerprinting
Simple Network Management Protocol (SNMP) Counters	· Security o Denial of service (DoS) detection o Traffic anomalies · Performance o Provisioning	· Not many. Possibly some information about utilization/activity if byte counts are collected per-subscriber