Before the

**National Telecommunications and Information Administration**

Washington DC

In the Matter of )
)
Developing the Administration's Approach ) Docket No. 180821780–8780–01
)
to Consumer Privacy )

**Comments by Ben Kaiser, Graduate Researcher, Princeton University**

Submitted November 9, 2018

My name is Ben Kaiser and I am a graduate researcher in the Center for Information Technology Policy at Princeton University. I am responding to NTIA's request for comments on developing the Administration's approach to consumer privacy. The proposal uses the terms "personal data" and "personal information" interchangeably to describe consumer data that requires privacy protections. The definitions of these key terms will greatly impact the ability of the policy to achieve the desired outcomes; in particular, an excessively narrow definition will limit users' control over their personal information (Outcome #2) and inhibit accurate risk modeling (Goal #4) by leaving out information that directly affects users' privacy risks.

I urge NTIA to clarify two important facets of this definition:

1. The information protected by NTIA's privacy rules must include both the *content* of users' data and communications as well as *metadata* that describes the content, because even without access to content, metadata can be intimately revealing.
2. NTIA's definition of Personally Identifiable Information (PII) must expand upon existing U.S. legal definitions of the term to include the pseudonymous identifiers used by third-party trackers because those identifiers do not sufficiently disambiguate users' real identities.

**Metadata must be covered by NTIA's policy**

The distinction between content and metadata is clear when considering communications data: the text of a message or recording of a call is the content while the identities of the participants, time of transmission, and other descriptive information are the metadata. The distinction also exists for non-communications data: a document stored in the cloud has content (the text it contains) as well as metadata like its size, owner, and times of creation, access, and modification. For Internet traffic, metadata exists at various layers: packet-level metadata includes source and destination IP addresses while application-level metadata includes website titles and keywords indexed for search.

Content clearly reveals a person's activities, behaviors, and beliefs and therefore require privacy protection. However, research has demonstrated that through simple analysis, metadata can also reveal sensitive information. Even if metadata has been anonymized by stripping personal identifiers, the user whom the metadata describes can be identified through pattern analysis and correlation with public information sources.  Personal information about that user can then be inferred.

As an example, consider telephone call metadata. Using a de-identified set of call records and public data sources like Google and Facebook, researchers at Stanford were able to connect phone numbers to the names of their owners and their city of residence.[1] For personal calls, the researchers were able to determine the nature of the relationship between participants (e.g., siblings, parent/child, or boss/employee), and they were also able to identify calls to sensitive organizations such as mental health facilities, sexual and reproductive health facilities, financial services, and religious organizations. Patterns in this data can be further analyzed to reason about protected personal traits: calls to mental health providers followed by a call to an insurance company suggest a patient searching for a new therapist; long calls to family members followed by a short call to Planned Parenthood suggests pregnancy or other family planning concerns; a dearth of phone calls after sundown on Friday may indicate a Shabbat-observant Jew.

[1] Mayer, J., Mutchler, P., & Mitchell, J. C. (2016). Evaluating the privacy properties of telephone metadata. *Proceedings of the National Academy of Sciences*, *113*(20), 5536–5541. https://doi.org/10.1073/pnas.1508081113

All of this information is clearly personal, sensitive, and worthy of privacy protections, and this research shows that if the metadata is not protected, neither is the personal information. Therefore privacy policies that fail to protect metadata fail to truly grant users control over their personal information.

The same principle holds for Internet traffic: an Internet Service Provider (ISP) has access to traffic metadata that can reveal customers' Web browsing habits, search queries, and the times and locations at which they are active. Internet traffic metadata represents a compounded privacy risk because it is collected not only by intermediary service providers but also by third-party web trackers.[2] A 2016 study found 81,000 unique third-party web trackers in existence, including advertising companies, analytics services, social media companies, content providers like news services, and content hosting platforms.[3] While users may be notionally aware that call records are accessible to their phone provider and browsing records are collected by their ISP, this type of third-party collection is more surreptitious, less understood by users, and thus even more deserving of regulatory protection.

**User tracking data must be covered by NTIA's policy**

In some cases these third-party web trackers collect information that includes personal identifiers, which allows for activity to be directly related to a particular person. Tracking companies sometimes argue that the information they collect is "pseudonymized" or "de-identified" and therefore is not personally revealing about any specific user.[4] These

[2] Mayer, J. R., & Mitchell, J. C. (2012). Third-Party Web Tracking: Policy and Technology. In *2012 IEEE Symposium on Security and Privacy* (pp. 413–427). San Francisco, CA, USA: IEEE. https://doi.org/10.1109/SP.2012.47

[3] Englehardt, S., & Narayanan, A. (2016). Online Tracking: A 1-million-site Measurement and Analysis. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security - CCS'16* (pp. 1388–1401). Vienna, Austria: ACM Press. https://doi.org/10.1145/2976749.2978313

[4] A. Narayanan. (2011) There is no such thing as anonymous online tracking. [Online]. Available: http://cyberlaw.stanford.edu/blog/2011/07/there-no-such-thing-anonymous-online-tracking

terms describe processes by which a purportedly unattributable identifier is attached to a person's data, obscuring the person's identity. Because that identifier does not personally identify any user, the argument goes, there is no meaningful privacy impact.

This notion that collected information is only worthy of receiving privacy protections if it is associated with PII underlies much of existing U.S. data privacy law, such as state data breach regulations. The definition of PII used in these laws is fairly narrow; it typically includes users' real names or account usernames, Social Security numbers or other unique identifiers, credit card numbers, medical information, and a few other specific categories.[5]

Tracking companies rely on these narrow definitions to sustain their practices. However, the data they collect can be easily de-anonymized to reveal PII, allowing the data (content or metadata) to be linked to a users' real identity and rendering the regulations ineffective. This can be done through correlation with other data sets that the tracking company can access; recently, Stanford and Princeton researchers correlated anonymized browsing histories with public social media data to de-anonymize users.[6] Other sources of data that can be used for de-anonymization include incidentally leaked information or purchased data sets from websites whose business model is to collect and sell PII (so-called "lead-generation sites").[7]

It is therefore essential that Federal privacy regulations define PII broadly enough that these trackers cannot claim that the supposedly de-identified information they collect is

[5] National Conference of State Legislatures (2018). Security Breach Notification Laws. [Online]. Available: http://www.ncsl.org/research/telecommunications-and-information-technology/security-breach-notification-laws.aspx

[6] Narayanan, A., & Shmatikov, V. (2008). Robust De-anonymization of Large Sparse Datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)* (pp. 111–125). Oakland, CA, USA: IEEE. https://doi.org/10.1109/SP.2008.33

[7] A. Narayanan. (2011) There is no such thing as anonymous online tracking. [Online]. Available: http://cyberlaw.stanford.edu/blog/2011/07/there-no-such-thing-anonymous-online-tracking

exempt while maintaining a store of data that does in fact reveal personal user information. Any data that is relatable to an identifiable person by means of parsing or analysis must be included within that definition.

**NTIA's policy must respect consumers' preference for control over tracking information**

Consumers have a well-established desire to not be tracked across the Web.[8] They are unable to protect themselves because they do not understand the threats they face and as research from Carnegie Mellon has shown, they cannot readily use the privacy tools provided to them.[9] Therefore, to achieve NTIA's stated goals of developing an outcome-based policy that allows users to exercise reasonable control over their personal information, metadata in general and web tracking data specifically must be covered by NTIA's policy.

[8] Turow, J., & King, J., & Hoofnagle, C. J., & Bleakley, A., & Hennessy, M. (2009). Americans Reject Tailored Advertising and Three Activities that Enable It. Technical report. https://repository.upenn.edu/asc_papers/524/
Purcell, K., & Brenner, J., & Rainie, L. (2012). Search Engine Use. Technical report. http://www.pewinternet.org/2012/03/09/search-engine-use-2012/

[9] Leon, P., Ur, B., Shay, R., Wang, Y., Balebako, R., & Cranor, L. (2012). Why Johnny can't opt out: a usability evaluation of tools to limit online behavioral advertising. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12* (p. 589). Austin, Texas, USA: ACM Press. https://doi.org/10.1145/2207676.2207759