Association for Computing Machinery (ACM)
US Public Policy Council of ACM (USACM)

usacm.acm.org
facebook.com/usacm
twitter.com/usacm

RESPONSE TO REQUEST FOR PUBLIC COMMENT
Big Data and Consumer Privacy in the Internet Economy
79 FR 32174
DOCUMENT NUMBER 2014-13195
NATIONAL TELECOMMUNICATIONS AND INFORMATION ADMINISTRATION
DEPARTMENT OF COMMERCE

RESPONSE FILED BY:
U.S. PUBLIC POLICY COUNCIL OF THE ASSOCIATION FOR COMPUTING MACHINERY

We submit the following comments on behalf of the U.S. Public Policy Council (USACM) of the Association for Computing Machinery (ACM).

Summary of recommendations:

The language of the Big Data reports from the Administration and the President's Council of Advisers on Science and Technology suggests that limiting data collection is increasingly infeasible and therefore should be deemphasized and that limiting use should receive increased attention. While big data poses challenges to collection limitation, the temptation to devalue it should be resisted and countervailing innovation, including privacy enhancing technologies (PETs), pursued. We must make it easy for someone without technical knowledge to select and apply a PET to that person's interactions with a potential data collector.

Responsible Frameworks: We recommend using a broadly construed risk-based approach to responsible use that accommodates multiple privacy risk models, and allows designers to accommodate variations in risk and exposure.

Notice and Consent: Consumer guidance should be attached to data supplied by the consumer only once. Organizations should be responsible for applying the guidance wherever and whenever the data is communicated or used.

Feasibility of Deletion: We agree that in some systems it is "practically impossible" to completely delete data for operational, technological or legal reasons. However, systems designers should build reasonably effective deletion capability into the system and document the capability and its limitations.

Extent of Regulation: Having sector-independent means of handling data of different levels of sensitivity could help address cost concerns and spur innovation in Big Data by simplifying the set of privacy rules.

Latent Information: In addition to proper access and physical security controls, contract language is a practical tool available to organizations that want to discourage attacks against latent information about individuals.

Privacy Preference Profiles: We believe such profiles, if properly managed and customizable, could mitigate privacy risks.

ABOUT ACM AND USACM

With over 100,000 members, the Association for Computing Machinery (ACM) is the world's oldest and largest educational and scientific computing society. The ACM U.S. Public Policy Council (USACM) serves as the focal point for ACM's interaction with U.S. government organizations, the computing community, and the U.S. public in all matters of U.S. public policy related to information technology. Our comments are informed by the research experience of our membership. Should you have any questions or need additional information, please contact our Public Policy Office at 212-626-0541 or at acmpo@hq.acm.org.

General Comments

We welcome the review of issues connected to the intersection of big data and consumer privacy. It is critical to examine now, before investments limit policy choices, how the ability to collect, analyze and use large amounts of information may affect society. The definition of big data used by the Office of Science and Technology Policy (OSTP) focuses on datasets so "large, diverse and/or complex, that conventional technologies cannot adequately capture, store and analyze them." But the questions in this RFC (and our responses) also apply to large datasets currently captured by conventional technologies. The ability to analyze collected data effectively typically follows the ability to capture and store such data. As capabilities change we must systematically revisit policies applied to datasets as our analytical abilities advance.

USACM submitted comments to the OSTP in connection with its Big Data Request for Information.[1] We understand that NTIA will be reviewing those submissions as it proceeds, and we appreciate the opportunity to present additional comments.

Before addressing specific questions in the RFC, we wish to address an apparent tension introduced by the Big Data reports issued by both the Administration and the President's Council of Advisers on Science and Technology (PCAST). Both reports articulate a concern that big data places additional challenges on the notice and consent model that helps address privacy concerns. The reports also note the importance of managing the use of collected information, and suggest that more policy attention is required on the context of data use. We agree that managing the use of collected data is an important part of preserving not only the privacy of individuals whose data has been collected, but also the

[1] http://usacm.acm.org/images/documents/BigDataOSTPfinal.pdf

security of collected information. We are concerned, however, that the language of both reports suggests that limiting the collection of personal data is inherently infeasible in a big data world and should be deemphasized while greater attention is focused on governing usage. The correction response to the challenges big data poses to collection limitation and other Fair Information Practice Principles (FIPPs), including those contained in the Consumer Privacy Bill of Rights, is promotion of countervailing innovation. All FIPPs are important to addressing privacy interests.

We also encourage the use of increased granularity in how big data is described. Collection and use of data matter, but there are many other actions by consumers, data collectors and other data users that deserve additional attention.

Privacy enhancing technologies (PETs) can offer mechanisms for addressing some of these challenges in ways that avoid a zero-sum approach that devalues worthy principles. Government and the private sector should support research on both potentially relevant PETS and on how to transfer the benefits of PETs to consumers. Methods and techniques that can help consumers understand, obtain, and employ PETs are needed. It should be made easier for someone without technical knowledge to select and apply a PET to their interactions with a potential data collector. At the same time, organizations collecting personal information should also be encouraged to deploy PETs suitable for enterprise environments and to directly embed PETs in consumer products where feasible.

Answers to specific questions in the RFC

3. Should a responsible use framework, as articulated in Chapter 5 of the Big Data Report, be used to address some of the challenges posed by big data? If so, how might that framework be embraced within the Consumer Privacy Bill of Rights? Should it be? In what contexts would such a framework be most effective? Are there limits to the efficacy or appropriateness of a responsible use framework in some contexts? What added protections do usage limitations or rules against misuse provide to users?

There are elements of the Consumer Privacy Bill of Rights that are consonant with a responsible use framework, (Individual Control and Respect for Context). Where a responsible use framework will have challenges is in addressing the wide variety of consumer interests and needs. For instance, some consumers (such as abused spouses) that have legitimate reasons for concealing certain information will have a different conception of reasonable use compared to others. Managing those differences will be important in mitigating the potential for misuse of information.

There will be instances where a technical system has to deal with uncertainty about the sensitivity of certain information (especially when the system contains free text, or if the information is mixed with other data sources). For responsible use to be meaningful, technical systems must be able to adapt to this uncertainty. A data system that has a five

percent chance of containing sensitive information should be treated differently than one that has a ninety-five percent chance.

We recommend using a broadly construed risk-based approach to responsible use, an approach that allows designers to accommodate variations in risk and exposure. Such an approach should accommodate multiple privacy risk models, including ones grounded in Fair Information Practice Principles.

4. What mechanisms should be used to address the practical limits to the "notice and consent" model noted in the Big Data Report? How can the Consumer Privacy Bill of Rights' "individual control" and "respect for context" principles be applied to big data? Should they be? How is the notice and consent model impacted by recent advances concerning "just in time" notices?

For consumers to be able to give meaningful notice and consent, mechanisms must be available for a consumer to provide broad guidance about their preferences on multiple categories of information. These preferences should be expressed once, not for each data collector, and be linked to the consumer as the data travels between data collection and use parties. Akin to a universal change of address notice, that would allow for consumers to change their preferences without identifying and signing in with each data collector. Identity solutions encouraged by the National Strategy for Trusted Identities in Cyberspace (NSTIC) may be able to provide such tools for consumers to manage their preferences and/or specify different preferences for different kinds of data.

Improved and heterogeneous approaches to notice deserve increased attention and development. These include the use of icons to convey standardized information and lexicons that can represent reference models of data flows.[2] The 'just in time' notice links the consent or choice decision to the context in which the user is more likely to understand their decision. However, just in time notices do not address the more complex challenges of ensuring that consumer preferences follow the data. Policies that allow for preferences to follow the data are sometimes called 'sticky.' To implement such policies, organizations must effectively collaborate in establishing and maintaining data provenance.

5. Is there existing research or other sources that quantify or otherwise substantiate the privacy risks, and/or frequency of such risks, associated with big data? Do existing resources quantify or substantiate the privacy risks, and/or frequency of such risks, that arise in non-big data ("small data") contexts? How might future research best quantify or substantiate these privacy risks?

---

[2] For more information on dataflow-based lexicons, see our 2011 comments to the Federal Trade Commission - http://usacm.acm.org/images/documents/FTCprivacyResponseFinal.pdf

Latanya Sweeney, currently FTC Chief Technologist, has led relevant and exemplary research on health data flows. The work is available at theDataMap.org.

7. The PCAST Report states that in some cases "it is practically impossible" with any high degree of assurance for data holders to identify and delete "all the data about an individual" particularly in light of the distributed and redundant nature of data storage. Do such challenges pose privacy risks? How significant are the privacy risks, and how might such challenges be addressed? Are there particular policy or technical solutions that would be useful to consider? Would concepts of "reasonableness" be useful in addressing data deletion?

We agree with the PCAST report on the practical impossibility of identifying and deleting all the data about an individual. It is also impractical to delete all information that is relevant to a particular category or event. There are three major difficulties:

- There may be legal or operational reasons for retention. It might be needed for auditing and litigation. It might be relevant to another person who wishes to retain it, e.g., 'John Jones was married to (or partied with) Jane Smith.'
- It is not feasible to track all copies of information within an enterprise. Legacy systems might need to be reengineered. End users' cutting and pasting into their own documents would need to be tracked. Disk backups contain information that is difficult to decode, since applications may not be in the same backup as the information. Deletion utilities may have trouble finding requested data within data dumps or other large datasets.
- If one allows deletion of selected content, it is often uncertain which objects (particularly free text or photos) *might* fit content description.

Reasonableness is useful in addressing data deletions, because there are certain circumstances where deletions would break the primary functionality of the system or there exist legal barriers to deletion. Sometimes sequestering data subject to a deletion request could address the relevant privacy concerns while still retaining needed data (which could be made available to users only under special circumstances). It may be useful to think of deletion less as a discrete action and more in terms of a dynamic spectrum of accessibility. Deletion then becomes an ongoing activity that steadily and progressively renders designated personal information inaccessible

One might support requests to delete data that the holder *can effectively determine* to fall within the scope of the deletion request. One might also require some minimum capability for making such determinations. The meaning of "reasonable" in the context of big data and consumer privacy needs to be defined and maintained by a specified organization (perhaps the National Institute of Standards and Technology or relevant sector-specific organizations).

System designers can and do provide users with functionality to delete specific and particularly sensitive data items. These techniques can be more effectively scaled to delete data by category using data tagging technology where data is tagged by the user or the system and the system links those tags to the data through the lifecycle of that data.

Requests to delete particular content reveal its existence. Therefore, in certain situations where humans can review deletion requests, the request should be used only for deletions.

8. The Big Data Report notes that the data services sector is regulated with respect to certain uses of data, such that consumers receive notice of some decisions based on brokered data, access to the data, and the opportunity to correct or delete inaccurate data. The Big Data Report also notes that other uses of data by data brokers "could have significant ramifications for targeted individuals." How significant are such risks? How could they be addressed in the context of the Consumer Privacy Bill of Rights? Should they be? Should potential privacy legislation impose similar obligations with respect to uses of data that are not currently regulated?

The differential treatment of information raises questions about the cost of having different controls in place for different sectors, and about the cost of determining what sector-specific rules are in effect in particular circumstances. Cross-cutting technologies, like applications that could link unregulated lifestyle data to regulated health data, or social network data to financial data, compound these questions. Small companies driving innovation in some sectors may not understand the complexity of rules specific to other sectors. Having sector-independent means of handling data of different levels of sensitivity could help address those cost concerns and spur innovation in Big Data by simplifying the set of privacy rules. This would also allow for more limited and focused sectoral rules when necessary

9. How significant are the privacy risks posed by unindexed data backups and other "latent information about individuals?" Do standard methods exist for determining whether data is sufficiently obfuscated and/or unavailable as to be irretrievable as a practical matter?

Standard methods do not exist for determining whether data is sufficiently obfuscated and/or unavailable as to be irretrievable. Fully addressing the issue requires addressing two types of threats: attempts to reidentify a specific individual versus reidentification in bulk. In both cases, the attacker can still download and analyze the whole backup dataset. This may be costly enough to dissuade attacks on an individual, but not to dissuade attacks in bulk (whose cost is spread over all reidentified individuals). In addition to proper access and physical security controls, contract language is a practical tool available to organizations that want to discourage this kind of attack.

10. The PCAST Report notes that "data fusion occurs when data from different sources are brought into contact and new, often unexpected, phenomena emerge;" this

process "frequently results in the identification of individual people," even when the underlying data sources were not linked to individuals' identities. How significant are the privacy risks associated with this? How should entities performing big data analysis implement individuals' requests to delete personal data when previously unassociated information becomes associated with an individual at a subsequent date? Do existing systems enable entities to log and act on deletion requests on an ongoing basis?

Several academic researchers have demonstrated the risks of re-identification, including in cases involving data sources from AOL Search,[3] Netflix,[4] and IMDB. The techniques for conducting these re-identifications are established and repeatable.

After a request for deletion, a system (e.g., at a data broker) may receive new information about the individual from a variety of original sources as well as from future linking. The issue of future deletions should be addressed in this more general setting. Both the privacy issues and the technological consequences seem similar.

The technical challenge is to 'cache an individual's request for deletion of a dataset, and to act on such requests on an ongoing basis. The task is not trivial, but the software effort does not seem huge. To our knowledge, systems do not currently exist that apply deletion requests continually. The run time cost can be reduced if deletion requests can be deferred and run in batches (e.g., weekly); that would leave a window of vulnerability for individuals, but is still better than current practice. Such an approach would be a risk management decision that must consider both the privacy risk and the mitigation cost

11. As the PCAST Report explains, "it is increasingly easy to defeat [de-identification of personal data] by the very techniques that are being developed for many legitimate applications of big data." However, de-identification may remain useful as an added safeguard in some contexts, particularly when employed in combination with policy safeguards. How significant are the privacy risks posed by re-identification of de-identified data? How can de-identification be used to mitigate privacy risks in light of the analytical capabilities of big data? Can particular policy safeguards bolster the effectiveness of de-identification? Does the relative efficacy of de-identification depend on whether it is applied to public or private data sets? Can differential privacy

---

[3] C. Christine Porter *De-Identified Data and Third Party Data Mining: The Risk of Re-Identification of Personal Information*, 5 Shidler J.L. Com. & Tech. 3 (Sep. 23, 2008), *at* http://www.lctjournal.washington.edu/Vol5/a03Porter.html.

[4] Narayanan, Arvid and Shmatikov, Vitaly, "Robust De-Anonymization of Large Datasets (How To Break Anonymity of the Netflix Prize Dataset," http://arxiv.org/pdf/cs/0610105.pdf

mitigate risks in some cases? What steps could the government or private sector take to expand the capabilities and practical application of these techniques?

It is harder to control re-identification in public datasets because access to those datasets is much easier and less effectively monitored. There may not be effective recourse against those who re-identify information from public datasets. De-identification alone would constitute a single (and potentially unacceptable) point of failure for the privacy of the individuals whose records make up a public dataset, highlighting the need to take a systems engineering approach to the management of privacy risk. Private datasets can be subject to other technical controls as well as organizational policies, codes of ethics, and laws, which may discourage those with access to the datasets from engaging in re-identification.

The utility of de-identification as a privacy risk control is hotly debated. The debate is complicated by the absence of any agreed framework for establishing confidence one way or the other under some set of conditions. The National Institute of Standards and Technology (NIST) or a similar standards body should conduct an evaluation of de-identification technologies against a range of datasets to better reflect the variety of data practices and circumstances. NIST performed a similar function when evaluating facial recognition technologies to assess the privacy risk in very practical terms

12. The Big Data Report concludes that "big data technologies can cause societal harms beyond damages to privacy, such as discrimination against individuals and groups" and warns "big data could enable new forms of discrimination and predatory practices." The Report states that "it is the responsibility of government to ensure that transformative technologies are used fairly" and urges agencies to determine "how to protect citizens from new forms of discrimination that may be enabled by big data technologies." Should the Consumer Privacy Bill of Rights address the risk of discriminatory effects resulting from automated decision processes using personal data, and if so, how? How could consumer privacy legislation (either alone or in combination with anti-discrimination laws) make a useful contribution to addressing this concern? Should big data analytics be accompanied by assessments of the potential discriminatory impacts on protected classes?

An impact assessment (distinct from a privacy impact assessment) focused on protected classes (and other relevant classes) would be a reasonable tool to address these concerns. Such assessments, whether focused on privacy, discrimination, or other concerns all serve as instruments for risk analysis and as such are consistent with a systems engineering approach. Such an assessment should characterize how the big data technology changes past practices in the relevant activity (housing, employment, benefits, etc.). In some situations, it may replace old, discriminatory practices, whether by human judgment or rule of thumb (e.g., based on just gender, age, and zipcode). In other cases, it may worsen undesirable discrimination, and reduce human judgment and review.

ACM US Public Policy Council (USACM)
1828 L Street NW, Suite 800
Washington, DC 20036
Tel: +1-212-626-0541
Fax: +1-202-667-1066
acmpo@acm.org
usacm.acm.org

13. Can accountability mechanisms play a useful role in promoting socially beneficial uses of big data while safeguarding privacy? Should ethics boards, privacy advisory committees, consumer advisory boards, or Institutional Review Boards (IRBs) be consulted when practical limits frustrate transparency and individuals' control over their personal information? How could such entities be structured? How might they be useful in the commercial context? Can privacy impact assessments and third-party audits complement the work of such entities? What kinds of parameters would be valuable for different kinds of big data analysts to consider, and what kinds of incentives might be most effective in promoting their consideration?

Organizations engaged in big data could benefit from the kind of advisory review that entities such as Institutional Review Boards provide. The structure and use of such entities could vary by sector, but providing visible accountability mechanisms can help achieve trustworthiness.

There must be effective detection of and recourse against offenders. A careful organization will detect and have recourse against employees who misuse personal information. Whistleblowers can sometimes reveal systemic bad practices. In situations where data access controls can be overridden under particular conditions (such as cyberattacks or other emergencies), there should be tools to encourage organizations to address excessive rates of overrides (such as reporting the frequency of overrides and rewarding low rates).

The Department of Health and Human Services requested comment in 2011 on proposed changes to its Common Rule concerning research involving human subjects. A good portion of those proposed changes addressed data collection and use. We encourage the NTIA to review what the Department has done in this area.

14. Would a system using "privacy preference profiles," as discussed in Section 4.5.1 of the PCAST Report, mitigate privacy risks regarding big data analysis?

Privacy risk is perceived differently by different individuals relative to their needs. In addition, individuals have demonstrated different skill and motivation levels in perceiving risks and setting controls to restrict use and sharing of their personal information. The privacy preference profile allows users to make judgments relative to others who have shared perceptions of risks, thus simplifying the challenge of individual participation. Explicit consent or preference expression is clearly advisable in particularly sensitive contexts. Profiles can also help organizations design their systems around common categories of risk based on user perceptions. We believe such profiles, if properly managed, could mitigate privacy risks. One way to do this is to make them customizable (allowing consumers to request tighter or looser control of particular categories of information in categories of circumstances).

One cannot (ever) expect that every piece of collected data will be properly tagged for every category that future legislators in some state, or some individuals consider sensitive. Medical record systems, particularly for specialized clinics, often manually tag a handful of categories,

but this does not scale to large number of categories, and does not apply new categories to legacy data. Some data categories may not be considered protected at the time of collection, or may not be understood by the data source. Automated tagging mechanisms will improve, but continue to be imperfect in their precision and recall. Preference profiles will have to address the inherent uncertainty about the sensitivity of some data.

15. Related to the concept of "privacy preference profiles," some have urged that privacy preferences could be attached to and travel with personal data (in the form of metadata), thereby enabling recipients of data to know how to handle the data. Could such an approach mitigate privacy risks regarding big data analysis?

The approach described in the question provides clear instructions and legal protection to a data holder, but serves individuals poorly. If a consumer's life situation changes, it is nearly impossible to adjust preferences with each data holder, on each copy held by that data holder. It would be preferable to have an individual's privacy preferences managed in one place, and provide data holders with a means to connect to that place. Allowing a few days delay in implementing requested changes might be permissible, if the time was used to enable the record holder to cache data and make data processing more efficient.

To pass preferences among organizations would require organizations to coordinate on standards to express, categorize and exchange these preferences. The kind of consumer data preference portability suggested by these questions is, right now, at best an emerging technology. In a similar vein, large-scale data provenance tracking has proved difficult. There does not presently exist the kind of cross-sector or cross-company exchanges of information that would facilitate the portability of personal data preferences.

One approach to consider is to provide stronger protection for data types that are widely considered sensitive, and emphasize tagging for them.

16. Would the development of a framework for privacy risk management be an effective mechanism for addressing challenges with big data?

A framework for privacy risk management would be beneficial for addressing challenges with big data and should leverage work on risk management across domains. These frameworks require analysts to examine specific threats, vulnerabilities and impacts based on contextual factors, including emerging privacy threats, laws, and hackers (among others), that require the introduction of mitigations intended to reduce risk. This form of analysis is typically done by experts; the framework provides consistency and encourages completeness with respect to the number and kinds of threats, vulnerabilities and impacts considered.

17. Can emerging privacy enhancing technologies mitigate privacy risks to individuals while preserving the benefits of robust aggregate data sets?

A significant challenge in this area (as addressed in our response to Question 11) is making such technologies easy to use by consumers. Another challenge (see our response to question 19) is making such technologies usable for data holders.

18. How can the approaches and issues addressed in Questions 14-17 be accommodated within the Consumer Privacy Bill of Rights?

Privacy Preference Profiles (Question 14) could be used to enhance the provisions of the Consumer Privacy Bill of Rights focused on Individual Control. If effectively designed and implemented, such profiles allow for consumers to customize how their data is collected and used, and could allow for those preferences to be changed. This would support the provisions in the Bill focused on Respect for Context. Privacy preference profiles would have to address the inherent uncertainty around data sensitivity to be effective.

Linking Privacy Preferences to Personal Data (Question 15). We consider linking preferences to data in poor service to consumers. It would not effectively recognize the provisions of the Bill that encourage Respect for Context. Unless there is a means to change privacy preferences post-linkage that is easy for the consumer, we do not think linkage would be well accommodated within the Consumer Privacy Bill of Rights.

Frameworks for Privacy Risk Management (Question 16) would be well accommodated within the Bill. Proper assessments of broadly construed privacy risk can identify and address risks related to potential violations of elements of the Bill, including Focused Collection. Other privacy risk models that articulate some combination of potential threats, vulnerabilities, and impacts should also be actively considered and employed where they might add value. Note that this is not an either/or proposition; multiple privacy risk models can and should be used where applicable.

Emerging Privacy Enhancing Technologies (Question 17) pose usability challenges to both consumers and data holders. Such technologies could not be effectively accommodated in the Bill if they were deployed prior to establishing such usability. It would also not be recommended that such technologies be widely deployed while they were still emerging. Mapping PETs, irrespective of maturity, to those elements of the Bill they support could be a useful exercise and help identify gaps in what should ideally be a broadly scoped research and development effort that includes analytical as well as design methods. Care should be taken not to overlook relatively simple techniques (which would still need to have their usability established), such as informative icons (which would map to Transparency).

19. What other approaches to big data could be considered to promote privacy?

Many PETs exist. To meet their potential, they need to be made usable by consumers, as discussed above (Questions 11 and 17). They also need to be feasible for and attractive to record holders. Few, if any record holders have expertise in every technique; small

businesses and organizations certainly cannot. For PETs to meet their potential, it would be helpful to have research in

- Descriptions of properties of each technique, a kind of nutrition label describing strengths and vulnerabilities.
- Automated tools that help a record holder determine how to use one or more privacy enhancing techniques to meet a privacy requirement. (Today, each researcher invents a tool, and then seeks practical problems that exactly fit its capabilities. This approach to technology transfer does not scale.)
- How to effectively integrate PETs (and privacy generally) into existing systems engineering practices.

One needs to give record holders incentives and reduce the cost of PETs. One might prioritize privacy capabilities that also are useful for the record holder's own needs. For example, privacy benefits from fine-grained controls over different types of content. Fortunately, some data brokers already want such controls to efficiently manage their sources' proprietary constraints and their customers' topic subscriptions.