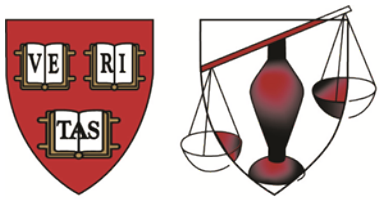

Berkman

The Berkman Center for Internet & Society
at Harvard University

Research Publication No. 2015-7
March 27, 2015

Integrating Approaches to Privacy across the Research Lifecycle: When is Information Purely Public?

David R. O'Brien
Jonathan Ullman
Micah Altman
Urs Gasser
Michael Bar-Sinai
Kobbi Nissim
Salil Vadhan
Michael Wojcik
Alexandra Wood

23 Everett Street • Second Floor • Cambridge, Massachusetts 02138
+1 617.495.7547 • +1 617.495.7641 (fax) • http://cyber.law.harvard.edu •
cyber@law.harvard.edu

Integrating Approaches to Privacy across the Research Lifecycle: When is Information Purely Public?

David R. O'Brien, Jonathan Ullman, Micah Altman, Urs Gasser, Michael Bar-Sinai, Kobbi Nissim, Salil Vadhan, Michael Wojcik, Alexandra Wood[1]

WORKING PAPER

March 27, 2015

[1] David O'Brien and Jonathan Ullman were the lead authors on the draft manuscript. David O'Brien and Micah Altman were primarily responsible for revisions. The workshop committee (Micah Altman, Michael Bar-Sinai, Kobbi Nissim, David O'Brien, Jonathan Ullman, Salil Vadhan, Michael Wojcik, Alexandra Wood) and Urs Gasser contributed to the conception of the report (including core ideas and statement of research questions); to the methodology (development of the workshop questions, use cases, conceptual model and taxonomies applied); to the project administration (coordination and management of the workshop and report writing process); and to the writing through critical review and commentary. Micah Altman, as workshop committee chair, and Salil Vadhan supervised the workshop process and report writing. Salil Vadhan, as managing PI of the Privacy Tools for Sharing Research Data project, led the funding acquisition supporting this workshop. The workshop and the writing of this report were supported by NSF grant CNS-1237235.

The co-authors would like to thank their fellow workshop participants for their contributions to the workshop and this report: John Abowd, Edoardo Airoldi, Aslan Askarov, Boaz Barak, Khaliah Barnes, Raef Bassily, Kristen Bolt, Scott Bradner, Mark Bun, Ran Canetti, Kenneth L. Carson, Eleni Castro, Stephen Chong, Mercè Crosas, Cynthia Dwork, Robert Gellman, Sharon Goldberg, Raquel Hill, Murat Kantarcioglu, Peter Katz, Henry Lam, David Lazer, Wendy Mariner, Dan O'Brien, Davide Proserpio, Sofya Raskhodnikova, Leonid Reyzin, Aleksandra Slavkovic, Adam Smith, Greta Lee Splansky, Peter Suber, Latanya Sweeney, Aurelia Tamò, Adam Tanner, Kit Walsh, John Wilbanks, Christopher Winship, Felix Wu, and Tanya Zlateva.

I. Background

The science of understanding human behavior, health, and interactions is being transformed by the ability of researchers to collect, analyze, and share data about individuals on a wide scale. However, a major challenge for realizing the full potential of such data science is ensuring the privacy of human subjects. And as new demonstrations and methods of reidentification continue to emerge, traditional approaches to protecting privacy are becoming decreasingly effective.

On September 24-25, 2013, the Privacy Tools for Sharing Research Data project at Harvard University, in collaboration with the Reliable Information Systems and Cyber Security Center at Boston University, held a workshop titled "Integrating Approaches to Privacy across the Research Lifecycle." Over forty leading experts in computer science, statistics, social science, law, and policy convened to discuss the state of the art in data privacy research. Participants considered how emerging tools and approaches from their various disciplines should be integrated in the context of real-world use cases involving the management of confidential research data.

This paper is part of a larger body of workshop materials that summarize the tools and use cases discussed during the event and map out a high-level research agenda to advance the integration of various methods of preserving confidentiality in research data. Additional materials produced in conjunction with the workshop are available on the workshop website.[2]

In the afternoon sessions of the workshop, participants separated into topical breakout sessions to discuss cross-cutting issues that emerged from discussions earlier in the workshop. This briefing paper identifies selected questions and issues raised during a breakout session that discussed the meaning of "public" in the context of using data about individuals for research purposes. This topic was motivated, in particular, by questions raised in earlier discussions at the workshop that revealed a lack of clarity about which data should be considered public and which data require a researcher to consider privacy implications. Although this paper does not discuss all of these questions, they are listed below for background.

- Do individuals expect a degree of privacy in the information they publicly share on the internet? How do the law and research ethics approach this question?
- Some individuals employ techniques to limit public exposure of data by, for example, using privacy controls on online services like to Facebook, deleting previously public information, selectively using services that purport to restrict certain data uses, and by following techno-social norms around information sharing that may not be recognized in non-internet contexts. How, if at all, should such conventions and behaviors shape legal expectations and ethical notions of privacy in research?
- Many web services, including social networks, have detailed terms of use that potentially restrict or prohibit secondary uses of data. When are researchers required to comply with these terms of use?
- Most websites are available in and cater to users located in different states, countries, and supranational regions, which may have different legal regulations and restrictions to use of data. When data is studied transnationally, are researchers obligated to honor data

[2] Privacy Tools for Sharing Research Data, http://privacytools.seas.harvard.edu/.

these laws and regulations that may apply in the country the user is located, where the services are hosted, and where the data in publicly accessed? To what extent are outcomes affected by stages in the data lifecycle, or to research that will be published in outlets with international reach?

- Sensors can be easily deployed in public locations to passively capture information about individuals. Is the data collected from such sensors public, and can researchers use it without restriction or IRB oversight?
- Government agencies often exercise discretion with respect to the scope of information they release publicly. What best practices should government actors follow in redacting information prior to release? And when researchers are aware that best practices have not been adopted or that stated de-identification policies have not been followed by others, do they have any obligations with respect to this data?
- Should IRBs be obligated to oversee human subject research that uses data from public sources, such as those mined from websites, from public sensors, or the government? Should researchers have an ethical duty to safeguard such information in their studies? How does the sensitivity of information factor into a determination whether information is public or private for research purposes? Should individuals on the internet have the ability to opt-out their information from studies?

II. Introduction: New Sources of Data

The technological developments of the last decade have provided social science researchers with new sources of rich information about individuals that enable them to analyze human behaviors with an unprecedented breadth and depth of scale.[3] The data can be mined directly from websites on the internet, collected by sensors placed in public locations, and released by the government. Much of this information is available for nominal cost and effort; researchers can quickly build large datasets from these sources. Most importantly, this information appears *public*, at least to first impressions, and it is capable of being used in research for a wide variety of purposes with seemingly minimal legal restrictions or ethical implications.

Social media and networking websites, like Facebook and Twitter, are among the many attractive new sources of social science research data.[4] As a whole, this category of website is increasingly being used around the world.[5] The web candidly captures and memorializes social interactions between individuals, and in many cases preserves them indefinitely. Although users can often restrict access to information using account settings and controls, not all do. For instance, a 2012 Pew Research report found that 42% of social media users' accounts are either

[3] *See* Victor Mayer-Schönberger and Keith Cukier, *Big Data: A revolution that will transform how we live, work, and think* (London: John Murray, 2013); David Lazer, et al., "Computation Social Science," *Science*, vol. 323, issue 5915 (February 6, 2009): pp.721-723.

[4] *See* Gary King, "The Changing Evidence Base of Social Science Research," in Gary King, Kay Scholzman, Norman Nie, eds., *The Future of Political Science: 100 Perspectives* (New York: Routledge Press, 2009).

[5] Maeve Duggan and Aaron Smith, "Social Media Update 2013," *Pew Research Internet Project* (December 30, 2013), http://www.pewinternet.org/files/2013/12/PIP_Social-Networking-2013.pdf. Increasingly use of the internet and social media websites appear to be a worldwide trend. *See*, *e.g.*, Simon Kemp, "Social, Digital & Mobile in 2014," *We Are Social*, January 8, 2014, http://wearesocial.sg/blog/2014/01/social-digital-mobile-2014/.

partially or completely available to the public without restriction.[6] Using automated tools, researchers can systematically collect information from the public portions of social media accounts without the need to interact with the individuals.[7] By creating fake profiles and friending users or by creating third-party applications on social network platforms, researchers can gain access to data that has been restricted to the public.[8] The data allows researchers to observe and analyze relationships, the sharing of information, conversations and interactions, creativity, media consumption, personality, reputation, and much more.

Researchers are also increasingly obtaining data from stores of government records, through open government data and e-government program or directly from government organizations.[9] Similarly, data can also be procured from businesses, such as telephone and utility providers, or collected using sensors, such as thermal imaging cameras, deployed in public places.[10] For example, the work of the *Boston Area Research Initiative* at the Radcliffe Institute for Advanced Study,[11] which seeks to promote original research by combining cutting edge social science model-based approaches, data mining and other big data methods that combine data from traditional sources with sensor data; and the work of the *Center for Urban Science and Progress* at New York University,[12] which uses big data methods and combinations of sensor data, such as thermal imaging and administrative data, to guide urban policy making and operations. Data generated by wireless phones, which are among the most ubiquitous devices in many countries, have also been used alone or in combination with data obtained from other sources in research studies.[13]

Despite the provocative insights that may result from these new veins of data, members of the research community are questioning these practices.[14] At the heart of the matter are some

---

[6] Mary Madden, "Privacy Management on Social Media Sites," Pew Research Internet Project (February 24, 2012), http://www.pewinternet.org/files/old-media//Files/Reports/2012/PIP_Privacy_management_on_social_media_sites_022412.pdf/.

[7] *See*, *e.g.*, Joesph Bonneau, Jonathan Anderson, and George Danezis, "Prying Data out of a Social Network," *Proceedings of the 2009 International Conference on Advances in Social Network Analysis and Mining* (2009), pp.249-254.

[8] *Id.*

[9] *See*, *e.g.*, US Government Open Data, https://www.data.gov/; New York City Open Data, https://nycopendata.socrata.com/.

[10] For an overview of many of these types of data sources, *see* Alex Pentland, *Social Physics: How Good Ideas Spread – The Lessons from a New Science* (New York: Penguin Press, 2014).

[11] Boston Area Research Initiative, http://www.bostonarearesearchinitiative.net/.

[12] Center for Urban Science and Progress, http://cusp.nyu.edu/.

[13] *See*, *e.g.*, "Using cell phone data to curb the spread of malaria," Harvard School of Public Health, Press Release, (October 11, 2012), http://www.hsph.harvard.edu/news/press-releases/cell-phone-data-malaria/; Juha K. Laurila, et al., "The Mobile Data Challenge: Big Data for Mobile Computing Research," 2012, https://research.nokia.com/files/public/MDC2012_Overview_LaurilaGaticaPerezEtAl.pdf; David Talbot, "African Bus Routes Redrawn Using Cell-Phone Data," *MIT Technology Review*, April 30, 2013, http://www.technologyreview.com/news/514211/african-bus-routes-redrawn-using-cell-phone-data/.

[14] *See*, *e.g.*, Ilka Gleibs, "Turning Virtual Public Spaces into Laboratories: Thoughts on Conducting Online Field Studies Using Social Network Sites," *Analyses of Social Issues and Public Policy*, vol. 00, issue 0 (2014): pp. 1-9; Jacquelyn Burkell, "Facebook: public space, or private space," *Information Communications & Society*, vol. 17, issue 8 (2014): pp. 974-985; R. Benjamin Shapiro and Pilar N. Ossorio, "Regulation of Online Social Network Studies," *Science*, vol. 339, issue 6116, (January 11, 2013): pp.114-145; Michael Zimmer, "But the data is already public: on the ethics of research in Facebook," *Journal of Ethics and Information Technology*, vol 12, issue 4 (December 2010).

difficult questions about the boundaries between public and private information, which implicate the law, ethical codes, and privacy theory. This paper identifies some of the questions raised by workshop participants and explores some of the contours of the debate around collecting and using data from these new public sources. The discussion has been supplemented with background information for context and, where possible, specific comments and questions from participants are noted.

III. Human Subjects Research, Private Information, and Public Information

Workshop participants questioned the extent to which the law and ethical codes adequately govern the use of information collected from sources, such as Facebook, Twitter, and sensors in public spaces, for research purposes. They contended that while the law may permit broad uses of this information, ethical issues may still lurk beneath the surface that are not currently addressed in practice.

In order to understand how the public-private distinction operates in practice, the law is a helpful starting point. Institutional research policies and ethical codes have an important relationship with law. The laws inform the policies and codes, setting baseline boundaries for the appropriate uses of information in practice, and to an extent ethical standards also inform the policies embodied in the laws crafted by legislators. Technological advances have also historically played a key role in shaping the concepts of privacy within the law. Perhaps now more than ever privacy law is in a state of evolutionary change as it continues to react to the technologies and practices that have emerged during last several decades – it is amorphous, complex, and unsettled.[15]

In the United States, information privacy law spans federal and state constitutional provisions, statutes and regulations, which are often described as being particular to organizations that deal with certain types of information (e.g., health care, educational records, and financial records) within specific industry sectors, the four common law "intrusion of privacy" torts, scenarios in which an agreement imposes contractual duties or restrictions, and relationships that give rise to special duties of confidentiality, such as those created between doctor and patient or lawyer and client. How these laws apply often depends critically upon the characteristics of the information, the actors, the uses, and other attendant circumstances.

Although information may appear to be open for the taking because it is featured on a publicly-viewable website or collected from a sensor capturing data in a public area, the information may be subject to a variety of privacy laws and regulations. Ethical codes may also prohibit certain uses of or methods of collection information that the law would otherwise permit. Most relevant to the public-private distinctions in research data are the regulations that cover human subject research, the intrusion of privacy torts, public records laws, and contracts. Other crosscutting legal issues – such as transnational jurisdiction and international laws – may also govern the collection and use of information. This section describes how these sources of law may be interpreted to apply to information obtained from seemingly public sources.

[15] *See* Daniel Solove, "A Brief History of Information Privacy Law," *Proskauer on Privacy, PLI* (2006), GWU Law School Public Law Research Paper No. 215, http://ssrn.com/abstract=914271.

A. US Human Subject Research Regulations

In a traditional social science study that uses information collected directly from interactions with human subjects, and not from the internet or other sources, academic researchers are required to structure their study in a manner that minimizes the risk of physical and psychological harms to the subjects. In many cases, depending on institutional policies, researchers will need to obtain approval from an IRB for research conducted in the United States, which is an institutional committee charged with ensuring the design of the study meets baseline criteria for safeguarding the human subjects from these harms. Other countries use similar review and oversight mechanisms for regulating research. Disclosing the nature of the study to potential participants, obtaining informed consent, and putting in place mechanisms to protect their privacy and confidentiality are the hallmarks of a properly designed study. This regime, which is derived from the US Department of Health and Human Services' (HHS) regulatory framework known as the "Common Rule"[16] and based on ethical guidelines created in the 1970s, applies to research in which the investigators directly interact with human subjects to collect and analyze potentially private information.

Not all research on humans is subject to the Common Rule. The Common Rule's reach is limited to studies that collect data through direct interactions with subjects or studies that involve subjects' private, identifiable information. Studies that only use public information have long been an exempt category.[17] The term public, in this sense, is synonymous with *benign* – if the information was collected from a public source, analyzing and disseminating it is not considered harmful to the person to whom it pertains. Equally important, if a research study falls into this category, institutions are not required to oversee it. The public-private distinction in the Common Rule owes its origins to privacy law, which has traditionally held that public information is not subject to privacy protections.[18] The distinctions between public and private information in US law are explored in later sections.

Once information has been made readily observable or sufficiently public it may no longer be considered private. According to the Common Rule's text, "private information" includes "information about behavior that occurs in a context in which an individual can *reasonably expect* that no observation or recording is taking place, and information which has been provided for a specific purpose by an individual and which the individual can *reasonably expect* will not be made public."[19] Here, the key factor for determining if this information may be subject to the Common Rule is whether an individual's expectations are "reasonable" – a term which often signals in the law a flexible, though somewhat unpredictable, benchmark based on an interpretation of appropriateness under the circumstances presented. The regulations do not

[16] 45 C.F.R. § 46, *et seq*.
[17] 45 CFR § 46.102(f) (defining "human subject" as "a living individual about whom an investigator obtains: (1) data through intervention or interaction with the individual, or (2) identifiable private information.") To the extent data mining is possible without interacting with a subject, which is often a trivial matter, and provided the information is not "private," the research is not subject to these regulations.
[18] *See*, *e.g.*, Ryan Calo, "The Boundaries of Privacy Harm," *Indiana Law Journal*, vol. 86 (Summer 2011): pp.1131-1162; Orin Kerr, "Applying the Fourth Amendment to the Internet: A general approach," *Stanford Law Review*, vol. 62 (2010), pp. 1027-1036.
[19] 45 CFR § 46(f)(2) (emphasis added).

provide any further guidance on the definition of reasonableness, leaving IRBs and researchers to exercise their judgment. As discussed in a later section below, the public's expectations of privacy have become more complex and intertwined with the contexts in which information is shared. Determining when these expectations exist or might be reasonable is not necessarily a simple calculation.

The information from new sources on the internet or from sensors is typically not subject to much oversight because it is often interpreted to be *public* based on the Common Rule's standard. As a consequence, researchers may not be required to obtain consent from the subjects or adhere to data security and confidentiality standards. Consider, for example, information about individuals on social networking websites. Most of this information is indexed and discoverable by using search engines, and it is capable of being copied and stored indefinitely. Once published to the web, a user effectively relinquishes her control, arguably making the information public by the Common Rule standard because anyone can observe it. Any expectation that this information would remain private or not be observed by others would probably not be considered reasonable.

While this interpretation reportedly prevails in practice, members of the research community have called into question whether it is ethical.[20] They are concerned that users of online social networks are unwittingly becoming subjects in research studies without consenting and may be exposed to harm. They also point to quantitative studies that show a surprising disconnect between user expectations and legal realities they face for their actions online as an argument that social network users are potentially vulnerable or may express their privacy preferences through normative expressions not recognized by current standards.[21] Debates on this topic have also spawned action. Several initiatives have emerged within the community to develop ethical guidelines that address user expectations of privacy in circumstances when the Common Rule may not treat the information as private.[22]

Until recently, the administrators of the Common Rule had not issued guidance on whether information obtained from these new sources should be treated as private or public. That

[20] *See*, *e.g.*, Ilka Gleibs, "Turning Virtual Public Spaces into Laboratories: Thoughts on Conducting Online Field Studies Using Social Network Sites," *Analyses of Social Issues and Public Policy*, vol. 00, issue 0 (2014): pp. 1-9; Jacquelyn Burkell, "Facebook: public space, or private space," *Information Communications & Society*, vol. 17, issue 8 (2014): pp. 974-985; R. Benjamin Shapiro and Pilar N. Ossorio, "Regulation of Online Social Network Studies," *Science*, vol. 339, issue 6116, (January 11, 2013): pp.114-145; Michael Zimmer, "But the data is already public: on the ethics of research in Facebook," *Journal of Ethics and Information Technology*, vol. 12, issue 4 (December 2010).

[21] *Id.* *See also* Mary Madden, "Privacy management on social media sites," *Pew Research Internet Project*, February 24, 2012, http://www.pewinternet.org/2012/02/24/privacy-management-on-social-media-sites/; Yabing Liu, Krishna P. Gummadi, Balachander Krisnamurthy, Alan Mislove, "Analzying facebook privacy settings: user expectations vs. reality," *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement* (2011): pp.61-70; Alessandro Acquisti and Jens Grossklags, "Privacy an Rationality in Individual Decision Making," *IEEE Security & Privacy* (January/February 2005), pp. 24-30.

[22] *See*, *e.g.*, Jason Gowans, "Draft: Code of Ethics & Standards for Social Data," Big Boulder Research Initiative, November 14, 2014, http://blog.bigboulderinitiative.org/2014/11/14/draft-code-of-ethics-standards-for-social-data/; Caitlin Rivers and Bryan Lewis, "Ethical research standards in a world of big data," F1000Research, http://f1000research.com/articles/3-38/v2; Annette Markham and Elizabeth Buchanan, Association of Internet Researchers, "Ethical Decision-Making and Internet Research: Recommendations from the AOIR Ethics Committee," August 2012, http://aoir.org/reports/ethics2.pdf.

changed in March 2013 when the Secretary's Advisory Committee on Human Research Protections (SACHRP)[23] at HHS issued a working document intended to serve as "the starting point for the development of FAQs and/or Points to Consider regarding the conduct and review of internet research."[24] This document was not mentioned by any of our workshop participants in the breakout session and as a result it was not discussed; however it marks an important development. In it, SACHRP explains that information intentionally posted or otherwise provided by individuals on the internet "should be considered public unless existing law and the privacy policies and/or terms of service of the entity/entities receiving or hosting the information indicate that the information should be considered 'private.'"[25] The guidance acknowledges that, whether reasonably or not, a subject's expectation that her information is private may support oversight of a study. This is significant since it deviates from how privacy law distinguishes between public and private information. In short, privacy law would afford little protection to information that is voluntarily made public.[26] The Common Rule has historically relied on this distinction to differentiate information that is private from information that is not.

The guidance states that four categories of websites can be considered "purely public" and are acceptable for widespread use:

> (1) sites containing information that, by law, is considered "public;"
> (2) news, entertainment, classified, and other information-based sites where information is posted for the purpose of sharing with the public;
> (3) open access data repositories, where information has been legally obtained (with IRB approval if necessary) and is made available with minimal or no restriction;
> (4) discussion fora that are freely accessible to any individual with Internet access, and do not involve terms of access or terms of service that would restrict research use of the information.[27]

Outside of the purely public categories, the guidance notes that determining whether websites can be used outside of the Common Rule can be difficult. Such decisions are riddled with nuance and may require researchers and IRBs to think rather deeply about the attendant circumstances, which could be subject to differing interpretations. For example, a website's architectural features, such as registration, authentication requirements, and content discoverability as well as techno-social norms within an online community, may suggest or indicate that information should be treated as private. The modification of website terms and privacy policies over time, shifting business models, and users who delete their content over time also pose considerations.[28]

While the SACHRP guidance is an important development, it is also limited in scope. For

---

[23] The Secretary's Advisory Committee on Human Research (SACHRP) "provides expert advice and recommendations to the Secretary . . . on issues and topics pertaining to or associated with the protection of human research subjects." HHS, "SACHRP Charter," (approved October 1, 2014), http://www.hhs.gov/ohrp/sachrp/charter/index.html. Note that its role within HHS is advisory, while the Secretary of Health is charged with regulatory oversight.

[24] SACHRP, US Dept. of Health and Human Services, "Considerations and Recommendations Concerning Internet Research and Human Subject Research Regulations, with Revisions," March 12, 2013, http://www.hhs.gov/ohrp/sachrp/mtgings/2013%20March%20Mtg/internet_research.pdf.

[25] *Id.* at p. 5.

[26] This is discussed in more detail in Section II. B.

[27] SACHRP, "Considerations and Recommendations," at pp. 8-9.

[28] *Id.* at pp. 5-6.

instance, it does not account for the global nature of the internet. Complex jurisdictional issues may arise when a study involves subjects from different countries, which could be difficult for researchers to ascertain based on a facial inspection and difficult to avoid altogether. By sweeping in users' information from other jurisdictions, researchers may be running afoul of international laws that prohibit the collection and use of data without consent despite the same manner of collection being lawful in the US. Not to mention normative privacy expectations may differ greatly between countries.

The SACHRP guidance is also silent about other emerging sources of data beyond the internet, including those from sensors located in public places. Images or recordings captured in publicly-observable areas, public records, cell phone transmissions, and the like can be attractive sources for passive data collection. Use of these data in studies also raises questions about whether it would be treated as public or private and subject to the Common Rule. The threshold determination is the same as for information collected from the web: the analysis turns on whether information is private and identifiable, and whether individual expectations of privacy are reasonable. In some cases, other statutes, laws, and regulations may concurrently govern the collection and use of this data. These will be explored in the next section.

B. US Information Privacy and Public Records Laws

The Common Rule regime only accounts for one potential source of privacy-related laws in the context of research data, and as noted above the regime does not necessarily apply to data gathered from public sources like the internet. Other laws, including the intrusion of privacy torts, the statutes and regulations that apply to specific types of data collection and use, and contractual limitations, may also constrain use of data from some sources. In this frame of reference, it may be useful to think of personal information in the US as being subject to a spectrum of possibilities, where one end represents information that is clearly subject to bright-line privacy rules and the other represents information that is considered public and unencumbered by law.

In the middle of this spectrum is a legal gray area where the application of law is less predictable. The wording of written laws, judicial interpretations, communal standards of appropriateness, and factual circumstances, such as the actors, the type of information, and how it is used, strongly influence where the information falls across this spectrum.

1. US Information Privacy Statutes and Regulations

Statutes and regulations govern information privacy at both the federal and state levels. Typically, they regulate how specific types of information can be collected, used, and disclosed in specific circumstances. Setting aside the Common Rule, which was explored in the previous section, these laws span financial records,[29] education records,[30] health information,[31] library

[29] *See*, *e.g.*, Fair Credit Reporting Act (FRCA) of 1970, Pub. L. 90-321, 84 Stat. 1127.
[30] *See*, *e.g.*, Family and Educational Rights Privacy Act (FERPA) of 1974, Pub. L. 93-380, 88 Stat. 57. (August 21,

records,[32] interception of wired and wireless communications,[33] and other categories of information.[34] These statutes and regulations comprise the bright-line rules of the information privacy laws – the categories of information to which they apply are almost universally regarded as sensitive, and the laws clearly state the circumstances in which they apply and the responsibilities of those who are regulated.

Many of these statutes and regulations are, in most cases, too narrowly focused to apply to researchers mining data from public sources. For instance, if health-related information is voluntarily posted to a social networking website in a publicly-viewable area by the person to whom it pertains, neither the information nor those who use it would be likely to be found subject to health information privacy laws.[35] Those laws apply to health care providers and a selection of other actors who handle the data in the course the business of health care. However, some statutes and regulations can present problems for researchers who collect information as intermediaries or from sensors located in public areas. For example, use of cameras or other sensors can implicate wiretapping laws in unexpected ways. Most state wiretapping laws require consent to record oral communications from at least one party to a conversation, and some states prohibit the collection of oral communications without the consent of all persons participating in the conversation.[36] If a researcher were to capture audio recordings of conversations without consent, even if the audio only captures what can be naturally overheard in public spaces, the researcher may be criminally liable under such laws.

2. US Public Records and Information

On the opposite end of the spectrum from information subject to bright-line privacy rules is "public information." Some information collected and used by the government, including personal information, may be considered a public record under the law. Freedom of information laws serve as the primary vehicle for public access to government records.[37] These laws, which are enacted at both the federal and state levels, enable members of the public to request access to inspect and copy records made or received by government entities.

The scope of records containing personal information subject to disclosure under these laws is

---

1974).

[31] *See*, *e.g.*, Health Information Portability and Accountability Act (HIPAA) of 1996, Pub. L. 104-191, 110 Stat. 1936 (August 21, 1996).

[32] *See*, *e.g.*, Mass.G.L. 78 § 7 (deeming the "part of the records of a public library which reveals the identity and intellectual pursuits of a person using such library shall not be a public record"); Ark. Code, §§ 13-2-703 (prohibiting disclosures of "library records which contain the names or other personally identifying details regarding the patrons of public, school, academic, and special libraries and library systems").

[33] *See*, *e.g.*, Electronic Communications Privacy Act (ECPA) of 1986, Pub. L. 99-508, 100 Stat. 1848 (October 21, 1986); Mass.G.L. 272 § 99 (prohibiting warrantless interception of wire and oral communications absent consent from all parties).

[34] Other types of regulated information include arrest and conviction records, cable television records, video rental records, retail transaction records, employment records, driver license records, use of social security numbers, tax records, telephone services, insurance records, and so on.

[35] *See*, *e.g.*, Standards for Privacy of Individually Identifiable Health Information, 45 CFR §§ 160, 164.

[36] Approximately 11 states are "two party consent states," and require the consent of all parties to a conversation.

[37] *See*, *e.g.*, Freedom of Information Act (FOIA), 5 USC § 552, http://www.law.cornell.edu/uscode/text/5/552; Massachusetts Public Records Law, Mass.G.L. Ch. 66, § 10, https://malegislature.gov/Laws/GeneralLaws/PartI/TitleX/Chapter66/Section10.

quite broad. It may include, for example, court filings, real estate deeds, arrest records and mugshots, records related to law enforcement investigations, certificates of death, marriage licenses, meeting minutes and transcripts, depending upon the jurisdiction and entity releasing the information. The freedom of information laws require government agencies to balance the confidential nature of the records against the public's interest in disclosure prior to permitting public access. However, the decision to withhold or release information is discretionary; government agencies are not compelled to withhold information unless another law would prohibit its release.[38] Practices can vary between jurisdictions and government entities depending upon the circumstances of a request for access and the nature of the record sought. Employment records, medical records, and other confidential information are often treated as sensitive and may be excluded from release. The exact information released or withheld is determined by the government entity on a case-by-case basis – some records are partially released with sensitive information that is redacted while others may not be released at all. Other statutes or regulations may also dictate that a record or specific information contained within it is exempt from public disclosure under freedom of information laws.

Once the information is made public through a public record, including any personal information released within those records, it is generally considered free to be used by the public for any purpose.[39] Indeed, public records are widely reused and redistributed on the internet in a variety of governmental, non-commercial, and commercial services. Examples include websites like http://mugshots.com, which purports to be a "search engine for . . . arrest records and booking photographs, mug shots," http://findthedata.org, which enables users to browse through a wide variety of public records and government data, and http://masslandrecords.com, which is the official Massachusetts government website for public land ownership records.

3. The Invasion of Privacy Torts

Among the broadest and perhaps least predictable of the US information privacy laws are the four "invasion of privacy" torts: (1) "intrusion on seclusion," (2) "public disclosure of private facts," (3) "false light," and (4) "appropriation of likeness." These torts were judicially crafted through the common law, and now are codified into written statutes or recognized as common law in nearly all US states.[40] Each tort is a distinct basis for a civil legal remedy – usually in the form of a lawsuit seeking damages or injunctive relief – when personal information expected to remain private is accessed or disseminated in a harmful manner. Of the four, intrusion on seclusion is perhaps the best suited to illustrating how these laws might apply to information voluntarily made public on the web or collected from other publicly-available sources.[41]

[38] *See* Ira Bloom, "Freedom of Information Laws in the Digital Age: The death knell of information privacy," *Richmond Journal of Law & Technology*, vol. 12 (Spring 2006): pp. 9-96.
[39] *See*, *e.g.*, "From Cradle to Grave: Government Records and Your Privacy," Privacy Rights Clearinghouse, https://www.privacyrights.org/cradle-grave-government-records-and-your-privacy.
[40] *See*, *e.g.*, Mass.G.L. 262 § 1B ("A person shall have a right against unreasonable, substantial or serious interference with his privacy. The superior court shall have jurisdiction to enforce such right an in connection therewith to award damages.").
[41] A "publication of private facts" necessitates that private information be published to the public at large, whereas here we only consider the data collection and use for research purposes. "False light" similarly concerns instances in which an individual "gives publicity to a matter concerning another," and the tort of "appropriation" provides a remedy for the appropriation of her likeness (e.g., name, image, etc) to the benefit of another. Restatement (Second) of Torts, §§ 652B, 652D.

However, the invasion of privacy torts have limited applications, especially in the context of the internet.[42]

Intrusion on seclusion is actionable when "one who intentionally intrudes, physically or otherwise, upon the solitude or seclusion of another or his private affairs or concerns . . . if the intrusion would be highly offensive to a reasonable person."[43] The act of intrusion must be unauthorized, highly offensive, and it must involve information that the individual reasonably expected to remain private. For example, trespassing on private property and peering into the windows of a home, searching through another's personal property, use of concealed cameras, and the unauthorized wiretapping of phone calls have been held unreasonable intrusions. The information does not have to be completely secret or unknown by others for it to be considered private, but the law does not protect facts that are already widely known about individuals unless an obligation to maintain confidentiality – a contractual agreement, specific statute, or legally recognized privilege – otherwise applies.

In a case involving an intrusion on a publicly-accessible website, the key question is whether the act of publishing information to a website causes an individual to circumstantially lose her expectation of privacy. After all, if it is observable in a public medium, how could the individual's expectation of privacy be reasonable? Most courts seem to agree that information posted publicly cannot reasonably be expected to remain private; however, the answer becomes murkier if the individual has taken steps to restrict access or otherwise limit dissemination of the information. Only a few courts have opined on this issue in civil litigation, and among those that have, the cases did not involve privacy tort claims but rather challenges to the scope of discovery requests in litigation, which raises questions about their precedential value. That said, the rulings suggest that individuals have little or no expectations in privacy for information they posted publicly on social networking sites or on publicly-viewable areas of the internet, even if the individual only expects a small number of individuals to read the information. Courts generally reason that the information was shared using a service that is intended to be a platform sharing information with others, and such information cannot reasonably expected to be private.[44] Some courts reason that the information is either viewable by the public at large or the individual cannot justify an expectation that his or her friends will keep information private.[45] Other courts have reasoned that by creating an account on a social networking site, like Facebook, the individual has consented to the possibility that others may eventually see the information regardless of his or her privacy settings.[46] The exceptions arise in cases where the individual uses privacy controls to limit the audience,[47] but there is some uncertainty as to whether privacy controls sufficiently seclude the individual and her affairs. Given the overall lack of cases and consensus on this issue, the law should be regarded as unsettled.

---

[42] *See*, *e.g.*, Ryan Calo, "The Boundaries of Privacy Harm," *Indiana Law Journal*, vol. 86 (Summer 2011): pp. 1131-1162.
[43] Restatement (Second) of Torts, § 652B.
[44] *See*, *e.g.*, *Guest v. Leis*, 255 F.3d 325 (6th Cir. 2001).
[45] *Reid v. Ingerman Smith*, LLP, 2010 WL 6720752 (EDNY 2012).
[46] *Loporcaro v. City of New York*, 35 Misc. 3d 1209(A) (April 9, 2012) (unreported).
[47] *See*, *e.g.*, *Ehling v. Monmouth-Ocean Hospital Service Corp.*, 2013 WL 4436539 (D NJ August 2013) (NO. 2:11-CV-03305 WJM) (holding Facebook wall posts are "configured to be private" for purposes of the Stored Communications Act); *Crispin v. Audiger*, 717 F.Supp.2d 965 (CD Cal. 2010).

4. Contracts, Terms of Service, and Privacy Policies

Contracts play something of a unique role on the internet. In its most basic form, a contract is an agreement between two or more parties that creates rights or obligations recognized by the law. Contracts are flexible instruments. Assuming a contract is validly formed and executed, parties to an agreement can contract for nearly anything within the boundaries of lawful activity, and should one party fail to meet those terms, the other party can seek redress in court.

Two types of standard-form contracts commonly encountered on the internet are relevant to this paper: "terms of service" agreement and the "privacy policy."[48] These instruments, which sometimes are called "terms of use," "terms of access," and other names, are used for a variety of purposes. They facilitate the sharing of data, create obligations of confidentiality, delineate access rights and acceptable uses of a service, describe ownership rights of user-generated and site-generated content, and much more.

The use of these agreements became popular on websites in the 1990s as a mechanism for allocating risks and responsibilities associated with emerging business models in retail sales and internet-based services, naturally evolving from the use of shrinkwrap licenses that accompanied software.[49] These agreements required the purchasers of the software to conditionally agree to terms prior to removing the shrinkwrap from a software retail box and installing it on a machine.

Today, terms of service are found on most websites. Their use has expanded beyond mere risk allocation to governing all aspects of the relationship between website operators and website end users.[50] They are often are presented during an account registration process for a web service, like use of a social network or webmail, and the user must click a button that states, in various phrasing, "I have read and agree to the terms" in order to complete the registration process. Many terms of service agreements are also hyperlinked on websites, usually near the page footer, and purport to apply to any browser of the site – that is, in exchange for visiting or using a website, the user implicitly agrees to the website's terms.[51] Whether the user is forced to acknowledge the terms of service or not, the terms are almost always a non-negotiable, take-it-or-leave-it offer, much like its ancestral cousin that captured a would-be user at shrinkwrap. Most websites will also reserve a right to modify their terms at will, and the terms often change

---

[48] It is worth noting that not every terms of service or privacy policy takes the form of or purports to be a contract between a user and service provider. This idea is explored briefly in the context of privacy policies in this section. For an in-depth discussion, *see* Allyson Haynes, "Online Privacy Policies: Contracting Away Control Over Personal Information?," *Pennsylvania State Law Review*, vol. 111 (Winter 2007): pp. 587-624; Ian Rambarran and Robert Hunt, "Are Browse-wrap Agreements All They Are Wrapped Up To Be?," *Tulane Journal of Technology and Intellectual Property*, vol. 9 (Spring 2007): pp. 173-202.

[49] *See* Mark Lemley, "Shrinkwraps in Cyberspace," *Jurimetrics*, vol. 35 (Spring 1995): pp. 311-323; Trotter Hardy, "The Proper Legal Regime for Cyberspace," *University of Pittsburgh Law Review*, vol. 55 (Summer 1994): pp.993-1055.

[50] *See*, *e.g.*, Nancy Kin, "Contract's Adaption and the Online Bargain," *University of Cincinnati Law Review*, vol. 17, (Summer 2011): pp. 1327-1370; Robert Dunne, "Deterring Unauthorized Access to Computers: Controlling Behavior in Cyberspace through a Contract Law Paradigm," *Jurimetrics Journal*, vol. 35 (Fall 1994): pp. 1-15.

[51] For example, this may be memorialized in an agreement as follows: "These Terms of Service constitute the agreement between [example.com] and you as a user who accesses, subscribes to access or otherwise establishes a connection ('user,' 'you,' or 'your') to the world wide web sites known as [example.com] (individually and collectively, the 'Site' and including any sub-domains, which are owned and controlled by [example.com])."

over time with and without notice to users.[52]

Although they are generally not required by federal law,[53] many websites employ privacy policies, which may be combined with or standalone from a website's terms of use agreement. The use of privacy policies emerged in the 1990s both organically and at the encouragement of the US Federal Trade Commission (FTC), as a means of disclosing information practices.[54] The earliest of these policies functioned as brief descriptions of a website's practices in a few sentences, or a certification seal, endorsed by third-party companies like TRUSTe, that suggested the website conformed with industry privacy norms.[55] The FTC encouraged internet businesses to adopt and expand privacy policies to provide consumers with notice of the information they collect and how they use it. According to scholars, "by 2001 virtually all of the most popular commercial websites had privacy notices."[56] Though no federal laws generally require websites to adopt these policies, some states laws do. California, for example, requires commercial websites that cater to users in California to "conspicuously post" an online privacy policy that describes the personally identifiable information collected and data practices.[57] Today, some privacy policies are structured as contractual agreements with website users, by expressly taking the form of a standalone contract or through incorporation by reference in the website's terms of service agreement, while other websites structure their privacies policy as a disclosure of practices rather than a contract. Regardless of their form, privacy policies have become an integral part of the FTC's consumer protection efforts. Should a website's information practices differ from those it voluntarily states in its policy, the FTC has jurisdiction to bring actions under its unfair and deceptive trade practices authority. State attorneys general may also be able to bring enforcement actions under similar state jurisdiction.

For researchers who seek to mine information from the internet, contracts that govern use of services can impose restrictions on collecting and using information obtained through websites. For example, according to its terms of service, Facebook requires individuals who collect information about other users to: obtain consent, make clear who is collecting the information,

---

[52] For example, Facebook's terms state that it will "provide [users] with seven (7) days notice . . . and an opportunity to comment on changes" to its Statement of Rights and Responsibilities. Facebook, "Statement of Rights and Responsibilities – Amendments," ¶ 14(1), (last revised November 15, 2013), https://www.facebook.com/legal/terms. Twitter's terms state that: "[Twitter] may revise these Terms from time to time, the most current version will always be at http://twitter.com/tos. If the revision, in our sole discretion, is material we will notify you via an @Twitter update or e-mail to the e-mail associated with your account. By continuing to access and use the Services after those revisions become effective, you agree to be bound by the revised Terms." Twitter, "Terms of Service – General Terms: Entire Agreement," ¶ 12(C), (last revised June 25, 2012), https://twitter.com/tos.

[53] Some specific types of websites are required to have privacy policy by federal law. For example, the Children's Online Privacy Protection Act (COPPA) requires websites that cater to children under the age of 13 to prominently post a privacy policy that details how the site uses personal information collected from children. 15 USC § 6502 *et seq*.

[54] *See* Federal Trade Commission, "Protecting Consumer Privacy in an Era of Rapid Change: A dynamic policy framework," Preliminary Staff Report, December 2010, http://www.ftc.gov/sites/default/files/documents/reports/federal-trade-commission-bureau-consumer-protection-preliminary-ftc-staff-report-protecting-consumer/101201privacyreport.pdf.

[55] Daniel Solove and Woodrow Hartzog, "The FTC and the New Common Law of Privacy," *Columbia Law Review*, vol. 114 (2014): pp. 583-676.

[56] *Id.* at 595.

[57] California Online Privacy Protection Act (CalOPPA), Cal. Bus & Prof. Code §§ 22575 *et seq*.

and post a privacy policy explaining what information is collected and how it is being used.[58] Pinterest's "acceptable use policy," which is incorporated into its terms of service, states that users are prohibited from "collect[ing] or stor[ing] personally identifiable information from Pinterest or its users without their permission," and "us[ing] any method to access, search, scrape, download, or change Pinterest or anything on it."[59] Twitter encourages and permits "broad re-use of its content" through its application programing interface within certain parameters.[60] However, its terms of service prohibit "scraping the services without the prior consent of Twitter."[61] While these terms may not have been written with academic researchers in mind, they may still have the effect of curtailing data mining from websites for research purposes without first obtaining permission from the website operators or complying with the terms and conditions.

Participants in this breakout session raised questions about the real-world implications of terms of service agreements for researchers. First, they questioned the enforceability of contracts on websites, given their mass-market non-negotiable form, onerous terms, and tendency to be ignored by users. Second, even if they were to be enforceable, participants questioned whether a website operator would invest the time and financial resources into pursuing a legal action in court against a researcher for violating the terms. Scholars and legal practitioners have also long scrutinized website contracting practices for similar reasons.[62] In practice, terms of service and similar contracts have scantily been tested in US courts; however, in the cases that have considered their validity, courts have generally found them to be enforceable.[63] For this reason, researchers should be aware that terms of service may still impact their ability to collect and use information obtained from websites, even if a website is unlikely to pursue a legal action. The SACHRP guidelines on internet research also suggest that researchers should not only be aware of these terms but also take them into account as they make decisions about using information posted to the internet for research purposes.[64]

C. International Information Privacy Laws

The subject of privacy in international law is too broad to provide a comprehensive review here, and much of it is well summarized by other sources.[65] Despite the risk of overgeneralization,

---

[58] "If you collect information from users, you will: obtain their consent, make it clear you (and not Facebook) are the one collecting their information, and post a privacy policy explaining what information you collect and how you will use it." Facebook, "Statement of Rights and Responsibilities – Protecting Other People's Rights," ¶ 5(7), https://www.facebook.com/legal/terms.

[59] Pinterest, "Acceptable use policy – Things you can't do," http://about.pinterest.com/en/acceptable-use-policy.

[60] Twitter, "Terms of Service – Restrictions on Content and Use of the Services," ¶ 8, (last revised June 25, 2012), https://twitter.com/tos.

[61] *Id.*

[62] *See*, *e.g.*, Leon E. Trakman, "The Boundaries of Contract Law in Cyberspace," *Public Contract Law Journal*, vol. 38 p.187-236 (2008); Victoria C. Plaut and Robert P. Bartlett III, "Blind Consent? A Social Psychological Investigation of Non-Readership of Click-Through Agreements," *Law and Behavior*, vol. 31(1) (2007); Mark Lemley, "Terms of Use," *Minnesota Law Review*, vol. 91 (December 2006): pp. 459-483.

[63] *See* Mark Lemley, "Terms of Use," *Minnesota Law Review*, vol. 91 (December 2006): pp. 459-483.

[64] *See* SACHRP guidelines *supra* note 24.

[65] *See* Daniel J. Solove and Paul M. Schwartz, *Information Privacy Law,* 4th ed. (New York: Aspen Publishers, 2011); Paul M. Schwartz and Daniel J. Solove, *Information Privacy: Statutes and Regulations* (New York: Aspen Publishers, 2011).

there are a number of aspects of international regulation of privacy that differ sharply from the US approach which are relevant both to the management of research data and to participants expectations of privacy.

First, in many parts of the world, protection of personal information is regulated in a fundamentally different way than in the United States, as described above. In the European Union, for example, "protection of personal data" is defined as a fundamental human right as it is part of the EU Charter of Fundamental Rights. This treatment is also reflected in the Organization for Economic Cooperation and Development's (OECD) privacy guidelines,[66] implemented in the EU data protection laws, and in the laws of individual member states, which provide omnibus protection for personal information in contrast to the sectoral approach used in the US. Many other countries beside those in the EU also adopt an omnibus approach to privacy. For instance, the Asia Pacific Economic Cooperation (APEC) privacy framework is based on the OECD privacy guidelines, as is Australia's privacy law.[67] Japan and a number of Middle Eastern countries also have uniform privacy protections, although were not derived from the OECD framework. Mexico and many South American states not only use an omnibus approach, but further adopt the concept of "habeas data," in which the subject of the data is treated as a data owner with inherent legal rights deriving from that status.[68]

Second, the omnibus privacy regulations in many other countries does not generally provide for unfettered use of information shared publicly. In most omnibus frameworks, data is shared for specific purposes, even when made publicly available, and the purpose for which the data was shared limits future use. Furthermore, individuals may have additional rights to examine, correct, or delete data that describes them – even when that data flows through third parties. This makes the public-private distinction more complex when applied to international data.

International privacy law is also rapidly evolving. A dramatic example of the implications of these rights is reflected in the recent European Court of Justice (ECJ) ruling against Google.[69] In this ruling the ECJ decided that an internet search engine operator is responsible for the "processing" of data involved in indexing and linking to third-party sites – and that search engines much honor requests by data subjects to remove those links where the data is irrelevant, no longer relevant or "excessive in relation to the purposes for which they were processed and in the light of the time that has elapsed."[70] In this case, the ECJ determined that search engines were both "processors" and "controllers" of data under the scope of the European Data Protection Directive, and that Google was subject to the directive because it "engages in the effective and real exercise of activity through stable arrangements" through its Spanish subsidiary.

---

[66] Organisation Economic Co-operation and Development (OECD), "Guidelines on the Protection of Privacy and Transborder Flows of Personal Data" (2013).
[67] Asia-Pacific Economic Cooperation (APEC), "APEC Privacy Framework" (December 2005), http://publications.apec.org/publication-detail.php?pub_id=390.
[68] *See* "Navigating the Gauntlet: A survey of data protection laws in three key Latin American countries," *Sedona Conference Journal*, vol. 14 (Fall 2013): p.137.
[69] Judgment of the Court (Grand Chamber), Case C-131/12, *Google Spain SL and Google Inc. v. Agencia Española de Protección de Datos (AEPD) and Mario Costeja González*, 2014.
[70] *Id.*

Another example of changes in global privacy law is the case *R. V. Spencer*, which was decided by the Supreme Court of Canada in June of 2014.[71] In this case the court's ruling included the finding that a request for information connecting an IP address to an individual constituted a search, and more generally, outlined an expansive approach to online privacy and a robust interpretation of the concept of reasonable expectation of privacy. Although this is a criminal case, which means the implications on research data sharing in other contexts is unclear, it illustrates that the notion of privacy rapidly evolving, and that there are some global trends towards strengthening global conceptions of privacy.

Even when research interacts solely with data collected by US based organizations, researchers are not fully insulated from global privacy law. The jurisdiction of international law is unsettled, may extend in unanticipated ways –further a number of countries such as the EU and members of the Association of Southeast Asian Nations (ASEAN) have explicitly attempted to either respectively assert jurisdiction of national data privacy law beyond national borders; or to hold local actors accountable for violations by third parties in other countries, involving privacy of data managed by the local actor.[72]

A particularly important example of how international law affects US organizations is the US-EU Safe Harbor framework, under which US companies self-certify that they will adhere to privacy protections that are deemed adequate by the European Commission – making these companies subject to enforcement by the Federal Trade Commission. Thousands of US companies have participated in the program and are thus subject to the privacy requirements of this framework.[73] Moreover, this framework is under increasing scrutiny from US and EU authorities. The FTC has increased enforcement; and as a result of challenges to data collection practices by Facebook, the Court of Justice for the European Union, the EU's highest court, is now actively reviewing the adequacy of the Safe Harbor regime.[74]

Researchers are not guaranteed to be insulated from international privacy regulation simply because their data collection efforts are conducted within the United States. This reflects a key area of concern that emerged in the breakout session and other discussions in the workshop. Management of data extends across multiple lifecycle stages. For example, data that is *collected* solely within the US may be *produced* in France or by its citizens. The data may have been originally *provided* with the expectation and under terms of use that appropriate local data protections would be followed. Many of these factors that should be taken into consideration may not be documented or readily accessible to a diligent research who inspects information prior to collection. Ethically, legally, and practically it is not safe to assume that the US definition of privacy is the sole relevant consideration.

---

[71] 2014 SCC 43 (Canada).
[72] Christopher Kuner, "Internet Jurisdiction and Data Protection Law: An International Legal Analysis (Part 1)," *International Journal of Law and Information Technology*, vol. 18 (2010): p.176.
[73] US-EU Safe Harbor Framework, http://export.gov/safeharbor/index.asp.
[74] *See* Loek Essers, "Europe's Top Court to Review Personal Data Exchange Between EU and US," *PC World*, June 2014, http://www.pcworld.com/article/2364920/europes-top-court-to-review-personal-data-exchange-between-eu-and-us.html.

IV. Unclear User Expectations of Privacy on the Internet

Although US privacy law may not recognize that individuals have expectations of privacy in internet postings and other public spaces, they may not expect that their personal information could be mined and used by anyone.[75] Scholarship and empirical studies in recent years suggests a disconnect between user expectations and corporate practices on the internet.[76] This might seem paradoxical given that many social networking websites are fundamentally one-to-many communication mediums and privacy is often thought of as a trade-off in an exchange for using free-of-charge web services. However, participants at our workshop hypothesized that, for a number of reasons, individuals may not view privacy as a binary choice in their online interactions; rather, individuals may subjectively take other contextual factors into account other than whether the information they post is viewable by others. Also, many individuals may not fully understand or appreciate the privacy consequences of publishing personal information available online.[77]

A. Contextual Expectations and the Role of Friction

According to leading scholars, the contexts in which individuals share information online play an important role in shaping users' expectations of privacy as well as cultural norms around the sharing and use of information.[78] Participants pointed out that decisions are also influenced by how alternatives, risks, and consequences are framed, and individuals may incorrectly perceive the extent to which privacy laws or technical controls work to preserve privacy and confidentiality. When information is shared in one context – for example, a post on a social network that a person expects to be only of interest to close friends – the individual may not anticipate that the information would be closely scrutinized by others, copied, or reused in other contexts. A number of lawsuits from recent years, to provide just one data point, demonstrate how gaps in contextual expectations for information use and sharing can upset users of social networking sites when new secondary uses of information are suddenly introduced.[79] Contextual

---

[75] *See*, *e.g.*, Eric Goldman, "The Privacy Hoax," *Forbes*, October 14, 2002, http://www.forbes.com/forbes/2002/1014/042.html.

[76] *See*, *e.g.*, Mary Madden, "Privacy management on social media sites," *Pew Research Internet Project*, (February 24, 2012), http://www.pewinternet.org/2012/02/24/privacy-management-on-social-media-sites/; Alessandro Acquisti and Jens Grossklags, "Privacy and Rationality in Individual Decision Making," *IEEE Security & Privacy*, pp.24-30 (January/February 2005).

[77] *See* Leslie K. John, Alessandro Acquisti, and George Lowenstein, "The Best of Strangers: Context Dependent Willingness to Divulge Personal Information Online," July 6, 2009, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1430482; Alessandro Acquisti and Jens Grossklags, "Privacy and Rationality in Decision Making," *IEEE, Security and Privacy* 24 (2005); Lior Strahilevitz, "A Social Networks Theory of Privacy," *University of Chicago Law Review*, vol. 72 (Summer 2005): pp. 919-988.

[78] Helen Nissenbaum, *Privacy in Context: Technology, Policy, and the Integrity of Social Life* (Stanford: Stanford Law Books, 2010).

[79] *See generally*, William McGeveran, "The Law of Friction," Legal Studies Research Paper Series, Research Paper No. 12-66, *University of Chicago Legal Forum* (2013), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2192191. Two prominent examples where companies have violated the integrity of their users' contextual expectations include Facebook's "Beacon" program and the Google "Buzz" service roll out, both of which prompted class action lawsuits over privacy concerns. *See* Louise Story, "The Evolution of Facebook's Beacon," *New York Times*, November 29, 2007, http://bits.blogs.nytimes.com/2007/11/29/the-evolution-of-facebooks-beacon/; Joshua Brustein, "The Negative Buzz Around Google's New Social Network," *New York Times*, February 11, 2010,

expectations and perceptions of how technology works may influence the privacy expectations of individuals as they publish information on the internet, and some empirical studies suggest that individuals spend very little time taking risks and consequences into account beyond the immediate context before them.[80]

Another recurring idea discussed during the workshop was the role of friction in accessing information online. Friction in this sense may be thought of as a function of the degree of effort required to gain access to information. Effort could include factors such as the financial cost of obtaining information, whether it is easy to locate or relatively obscure, the amount of time or energy necessary to access it, and so on. Friction, or the perception of it, likely plays a role in shaping individuals' privacy expectations, but at the same time technological developments continue to reduce the effort one needs to expend to gain access to information. On this point, workshop participants wondered to what degree should the law and ethics use friction to distinguish public and private information for use in research.

Personal information that can be located via a Google search for an individual's name may be valued differently than data in paper-based public records that can only be accessed by physically traveling to a courthouse near that individuals' home. As a result, users may have differing expectations about the risks associated with these different types of data based on how they perceive the efficacy of the friction and its effect on their privacy. They may also, perhaps unreasonably, expect that these costs and barriers will not be removed or easily overcome. Most internet users lack the expertise to adequately assess barriers to access, and some may possess inaccurate assumptions about the law, technology, and those with whom they share their information. Friction may also apply to information captured by sensors in public places. For example, a low voice in a public park may be barely audible except to a sufficiently-powerful microphone. Researchers may not be able to infer, absent a clear indication from a given user, whether a user expects her information to remain private.

Participants also pointed out that a lack of friction can play a role not only in shaping individual expectations but also in calculating harm. For instance, tweets are public, but an individual may have a reasonable expectation that an attacker looking at random Twitter accounts for stigmatizing information will not find her Twitter account. Thus, even though this information is public, her expectation of privacy may be colored by how obscure she perceives her account. However, if another individual mines Twitter accounts for a certain type of stigmatizing information and aggregates and links the information to the accounts of these users, then the search cost has been dramatically reduced and the realities of privacy may change drastically. In 2010, the website http://pleaserobme.com demonstrated how users can inadvertently share information that compromises the security of their home by aggregating public tweets from users that suggested, by inference or explicit reference, that the user was not at home.[81] In this case,

---

http://bits.blogs.nytimes.com/2010/02/11/the-negative-buzz-around-googles-new-social-network/.

[80] *See* Leslie K. John, Alessandro Acquisti, and George Lowenstein, "The Best of Strangers: Context Dependent Willingness to Divulge Personal Information Online," July 6, 2009, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1430482; Alessandro Acquisti and Jens Grossklags, "Privacy and Rationality in Decision Making," *IEEE, Security and Privacy*, vol. 24 (2005); Lior Strahilevitz, "A Social Networks Theory of Privacy," *University of Chicago Law Review*, vol. 72 (Summer 2005): pp. 919-988.

[81] *See* Dan Fletcher, "Please Rob Me: The risks of online oversharing," *Time*, February 18, 2010, http://content.time.com/time/business/article/0,8599,1964873,00.html.

the purpose of the website was to raise awareness of the potential for real-world harms, but it is easy to see how this concept could be exploited by malefactors.

Some of our workshop participants suggested that addressing the differences in the ease and cost of access to public data may be worth exploring as a solution, both in research and non-research contexts. For example, in the context of public records, certain data that is deemed less sensitive may be available to the public at large; whereas data that is more sensitive could be only be accessed for a fee or by expending greater effort in collecting it. Perhaps this requires, at a policy level, different legal and ethical standards that apply to distinct types of public data. An advantage of identifying different types of publicly-available information in this manner is that it would provide a basis for evaluating whether the use of data is appropriate. Another proposal from our discussion is to attach explicit costs to accessing data. The costs could be financial, in the form of a fee per record accessed, or architectural, possibly in the form of a tiered-access system in which access is made contingent through different levels of technical and legal restrictions based on a determination of the data's sensitivity. These explicit costs could be configured to ensure that even as the technology changes, the data will achieve a similar degree of openness to what was intended when it was created to account for a reduction in friction over time. In order to be effective, downstream uses would have to be bound not to remove these costs. Moreover, while the reduction of friction in accessing public records has brought with it privacy concerns, greater accessibility is also highly beneficial to the public. The public's interests would need to be carefully balanced in any solution that proposes to alter the costs of accessing public information, and that may render such a solution less attractive.

V. Concluding Thoughts

The social science and behavioral research fields are being revolutionized by these new sources of detailed information about humans. With them, researchers can quickly build large datasets for analysis using automated tools while simultaneously avoiding the managerial and oversight burdens associated with traditional methods of data collection. These trends are allowing researchers to test their hypotheses faster and conduct more research, which may yield new insights into human behaviors and provide enormous benefits to the public.

On the surface, information from these sources is often regarded as public. Indeed, the law in the US affords minimal privacy protections to information that has been voluntarily made public, and the Common Rule has traditionally allowed researchers to use public information with minimal oversight. However, the information widely shared on the internet and capable of being captured by sensors in public spaces can reveal surprisingly personal information about individuals, and many may not wish to have their information used in a research study without their consent. Moreover, individuals may not understand or appreciate the consequences of publishing information online, or expect unreasonably that their information will remain private.

Some members of the research community are questioning whether subjecting these new sources of data to the traditional standards of the Common Rule remain sensible. A number of community-driven initiatives have sought to fills these ethical gaps, and SACHRP at HHS has more recently issued an FAQ document intended to serve as a starting point for the development

of further guidance. Still, it is not clear to what extent researchers and institutions are following in practice the guidelines issues by HHS or the ethical codes being developed within the community. And, they remain subject to interpretation. Reasonable minds may disagree about the circumstances in which researchers should be ethically bound to disclose the existence of a study or be required to obtain consent from the subjects.

Researchers also face other practical issues mining information from the internet, including the challenges with disclosure and consent as well as potential collisions which may occur with the data ownership interests of website operators. Consider a study that using information mined from the internet is deemed subject to the Common Rule. Researchers may be required to disclose the existence of a study and obtain informed consent from participants in the study. It is not clear, however, at what point a researcher should inform an individual that their publicly-available information is being mined and included as part of a study, how they communicate it, or at what point explicit consent is required before any mining occurs. Participants were skeptical that an informed-consent regime could be created in this medium that provides a meaningful opportunity for individuals to opt-in or opt-out of research studies using current mechanisms like terms of use agreements or click-through agreements. Interacting with the users who generate data can also be difficult or even impossible. The regime may be difficult to scale. This is especially challenging given that most of the emerging examples of research involve data from a large number of users, users who are geographically diverse, and use of the data for purposes far from those the users could have anticipated when the data was created.

Collisions with other forms of law are also a danger. As individuals share information with services, like Facebook and Twitter, they may be agreeing to terms in which the services obtain licenses to information that entitle the website to certain exclusive copyright interests to use and display the data. These licenses can be enforced against others, including researchers, who attempt to use information posted online.[82] In addition, some website are known to use restrictive provisions in their terms of use agreements that inhibit the ability of others to mine data using automated techniques.[83] Finally, the legal and ethical responsibilities are not clear for the research institutions and the researchers and IRBs within them with respect to understanding, honoring and protecting expectations of privacy for participants who are the subject of US based internet research but reside in other countries with differing treatments of privacy.

A. Future Agenda for Defining "Public" for Research Purposes

In the short time of our breakout session, participants briefly mapped a few informal ideas they felt should be included in a future research agenda, such as a "public for research" standard that is incorporated into best practice principles.[84] Some participants felt that such a standard may not align perfectly with how US law currently defines these categories of information – for instance, "public for research purposes" may denote a narrower category of information than the

[82] *See*, *e.g.*, G. Ross Allen and Francine D. Ward, "Things Aren't Always As They Seem: Who really owns your user-generated content?," *Landslide*, vol. 3(2) (November/December 2010): pp. 49-54.
[83] *See*, *e.g.*, Peter A. Winn, "The Guilty Eye: Unauthorized Access, Trespass And Privacy," *Business Lawyer*, vol. 62 (August 2007): pp. 1395-1437; Michael Madison, "Rights of Access and the Shape of the Internet," *Boston College Law Review*, vol. 44 (2003): pp. 433-507.
[84] None of these ideas reflect consensus among the group, and further research is clearly needed to properly set a comprehensive agenda.

law. Others felt that researchers should be able to use at least as much as the law allows – "data is public if I don't need permission from the data holder or data subject to look at it, write it down, record it and republish it." Most participants seemed to agree that mechanisms for obtaining meaningful consent from subject whose information is collected from websites – easier to comprehend than a terms of service or privacy policy agreement – are needed to ensure subjects' participation is voluntary when it needs to be. Finally, many participants felt that a standard must also be nuanced enough to take into account the different degrees of publicity and be flexible over time to accommodate new types and uses of data.

The guidelines created by SACHRP as well as the efforts underway by groups within the research community to develop ethical codes, which were not discussed during the breakout session, do incorporate many of the agenda ideas that surfaced in our breakout session as well as ideas that were not discussed that seem prudent. However, it remains to be seen whether guidelines and codes will lead to predictable outcomes from both a legal and ethical point of view. A substantial gap exists between a researcher being able to obtain information without explicit permission and making a systematic determination that information is actually available to everyone on the internet without a reasonable expectation of privacy recognized by law or ethics. With this in mind, further research is needed to refine the approach and better understand the interests of stakeholder as well as the risks and potential harms to human subjects whose information is used in studies.