10.

Protecting privacy while releasing data: Strategies to maximise benefits and mitigate risks

Joel Gurin, Matt Rumsey, Audrey Ariss & Katherine Garcia

Introduction

Privacy has become an urgent issue in data use. Traditionally, 'open government data' has been thought of as free, public data that anyone could use and republish. Now, the discussion is shifting to include data that may not be appropriate for wide, unfettered access, but can still be of use to non-government communities.[1] Data containing personally identifiable information (PII) cannot be released widely, but there are certain circumstances that could allow for its use in restricted or de-identified forms. By considering various levels of sensitivity in the datasets they manage, data stewards can provide several levels of openness and release datasets in different ways accordingly (Open Data Institute n.d.).

As more open government data has become available, data users in business, academia, and the non-profit community have come up against a conundrum. Many datasets in health, education, housing and other areas may have the most value when they are released with 'microdata' that can be analysed at the level of individual records. But releasing data at that level carries the risk of exposing PII that could threaten individuals' privacy if it were released openly. Government agencies must address the risks and sensitivities of making data available while at the same time maximising its accessibility and use.

1 Academic observers have been considering how best to balance open data and privacy concerns for several years. More recently, as the concept of open data is becoming accepted at all levels of government and the 'low hanging fruit' is released, government policy-makers and open data advocates have turned their attention to useful data that may be more difficult to release for a variety of reasons – including privacy concerns. Examples cited elsewhere in this chapter include the Open Data Institute's Data Spectrum; the Sunlight Foundation's work on 'microdata', privacy and criminal justice data; and the Center for Open Data Enterprise and White House Office of Science and Technology Policy's Open Data Roundtable Series.

Approaches to privacy are inevitably affected by political goals and considerations. In the US, for example, President Obama recognised the need for clear guidelines by establishing the Federal Privacy Council in February 2016 (Obama 2016), and the Federal Communications Commission under Obama instituted privacy protections for data collected by internet service providers. A few months into the Trump Administration, the Republican Congress eliminated those FCC protections. It remains to be seen how changing political dynamics in the US, and potentially other countries, will affect approaches to privacy policy. This chapter presents an analysis that should be helpful to any policy-maker who wants to study and address this issue.

Research context

'Microdata' is data released in its most granular, unaggregated form (Shaw 2014). The key question is: How can we maximise public access to and value from open granular information while protecting privacy? To answer this question, data and privacy experts have explored issues such as:

- What are the potential benefits of using unaggregated data (or microdata) for the public good?[2]
- What are the risks of using these datasets if they contain or could lead to the discovery of personally identifiable information, and how can those risks be minimised?[3]
- What are the best technical, policy and pragmatic approaches to ensure strong privacy protections while maximising the benefits of open data?[4]

Benefits of releasing microdata

Analyses of government-held microdata can advance public policy and social benefit through insight into public issues, better informed decision-making and improved delivery of public services. Microdata is already being used to improve the health and safety of citizens, the national transportation infrastructure, the criminal justice system, the quality of education, and the equity and stability of the country's housing market, among other uses. Here are examples of the benefits that highly detailed data can support.

2 See examples from transportation (Center for Open Data Enterprise 2015) and education (Park & Shelton 2012).
3 See, for example, Ortellado (2016).
4 See, for example, Altman et al. (2015); Borgesius et al. (2015); Dwork & Roth (2014); Ohm (2010).

Healthcare

A revolution in healthcare is underway, with data at its core. However, advances in this arena are also demonstrating the challenges and risks of greater health data utilisation. Health data has long been recognised as especially personal and sensitive information: it is already protected by the Health Insurance Portability and Accountability Act (HIPAA), and some experts believe that additional protections may be necessary (Podesta et al. 2014).

With proper privacy and security mechanisms in place, health and medical research institutions are able to share de-identified patient health information with doctors, allowing them to diagnose and treat disease more effectively. Large health datasets may be used to target services to underserved populations (Federal Trade Commission 2016). Research centers, drug companies, hospitals, and other institutions can analyse patient data to improve services and develop new treatments (Podesta et al. 2014).

The Precision Medicine Initiative (PMI) exemplifies the opportunities in analysing health microdata. Launched in 2015, the PMI is a US federal effort to 'enable a new era of medicine through research, technology, and policies that empower patients, researchers, and providers to work together toward development of individualised treatments' (The White House n.d.). If successful, it will allow for highly targeted treatments based on a range of inputs including personal medical histories and genetic analysis.

The PMI does not aim to make health data fully open to the public, but it relies heavily on data-sharing among clinicians and researchers with appropriate restrictions and safeguards. As the White House explains: 'to get there, we need to incorporate many different types of data [… including] data about the patient collected by health care providers and the patients themselves. Success will require that health data is portable, that it can be easily shared between providers, researchers, and most importantly, patients and research participants' (The White House n.d.).

Transportation

Around the world, untold numbers of commuters now check their mobile phones every day to see when the next bus will arrive. This information is at their fingertips thanks to open data (Press 2010). Ubiquitous travel apps have shown how open transportation data can improve public transit access, ease traffic congestion, and make citizens' lives easier.

Transportation microdata has potentially powerful applications when combined with other types of microdata. At a 2015 roundtable held with the US Department of Transportation and users of its data, participants flagged the need for crash data to be combined with hospital data 'to understand the long-term impacts of vehicle crashes and how different kinds of safety equipment

can mitigate injury' (Center for Open Data Enterprise 2015). Microdata from different sources can also be particularly useful for transit planners. For example, microdata on both travel patterns and commuters' income levels helps planners understand the obstacles faced by low income workers as they travel to their jobs, allowing for more efficient service delivery and equitable planning decisions (Tierney 2012).

Increasingly popular 'bike sharing' systems are another example of using transit microdata. These programmes generate mountains of data which are often released publicly, allowing advocates to push for expanded service, authorities to better target infrastructure investment, and researchers to ask tough questions about system equality. For example, a recent analysis of 22 million trips taken using New York City's Citi Bike system revealed that the bikes were heavily used for commuting purposes and rides were often concentrated in areas with robust bike lane infrastructure (Thomas 2016).

Criminal justice

Microdata can help improve the criminal justice process at several stages. It can be used to develop effective public policies, improve community relations, and correct unfair practices.

Recent high-profile efforts have focused on opening data about police practices and operations (Shaw 2015). The Sunlight Foundation has found that previous data releases 'have already paid off by improving outcomes that communities perceived as unfair. The case of released stop-and-frisk data provides an important example of this, where New York's public release of granular pedestrian stop data, and the analysis it permitted, led to the discovery that almost 9 out every 10 people stopped were entirely innocent, and that 9 out of every 10 people stopped were non-white' (The Sunlight Foundation 2014). Stop-and-frisk is a controversial practice during which police would stop and search pedestrians without a warrant. Allowing for better understanding of this data helped kick-start the repeal of what proved to be an ineffective and discriminatory policy.

Housing

Microdata on housing can help identify discriminatory lending patterns, surface structural vulnerabilities, and help policy makers prevent a future housing crisis. After the global financial crisis, the United States Congress took a number of steps to safeguard the country's financial system. Congress mandated the public release of data showing trends in the mortgage industry, in the interest of avoiding another housing bubble. As part of that effort, Congress strengthened requirements for publishing data under the Home Mortgage Disclosure Act (HMDA), a 1975 law designed to help prevent housing discrimination (Consumer Financial Protection Bureau 2015).

Data collected under HMDA, which is now implemented by the Consumer Financial Protection Bureau (CFPB), is released publicly every September. The data 'help show whether lenders are serving the housing needs of their communities; they give public officials information that helps them make decisions and policies; and they shed light on lending patterns that could be discriminatory' (Consumer Financial Protection Bureau n.d.). The CFPB is statutorily mandated to publicly disclose data under HMDA while developing appropriate protections for borrower privacy in light of HMDA's purposes.

Education

Microdata on student performance can help educational institutions provide students with the tools and support they need to build useful knowledge and skills. Data can be combined with mobile technologies and education software to personalise education (Podesta et al. 2014). To this end, the Obama administration took a number of steps to ensure that education data is properly leveraged, and pledged to 'work to develop a common trust mechanism for schools that want to exchange student data with each other and other qualified parties' (Park & Shelton 2012). So far, however, the difficulty of establishing that trust has been an obstacle to working with student data.

Risks of releasing microdata

The risks of releasing microdata from datasets containing PII are real and well documented. There is concern that releasing microdata from these sources could result in privacy violations, even if efforts have been made to 'anonymise' or 'de-identify' the data by stripping it of PII.

For many years it was thought that if a database was scrubbed of identifying information such as name, address, or social security number that privacy could be effectively protected. However, a growing body of research shows that this is often not sufficient to guarantee privacy. Furthermore, the increasing influence of big data has turned previously non-existent or inconceivable pieces of data into potentially identifying ones. There is also no standard definition of PII and wide variance in the way that various laws define the concept (Polonetsky et al. 2016).

The 'Mosaic Effect' is a common term for the idea that disparate datasets and information can be combined to expose sensitive information and negate attempts to protect privacy. Some high-profile examples have fueled these concerns. Latanya Sweeney's work showing that de-identified medical data can often be re-identified through linking or matching with other datasets is perhaps the most well-known instance (Sweeney 1997). In another well-known example, researchers were able to identify individuals from supposedly anonymised Netflix rating information a high percentage of the time with only the help of publicly available information from another source, the Internet Movie Database

(Narayanan & Shmatikov 2008). Another commonly cited example emerged when America Online (AOL) released 'anonymised' search results from 650 000 of its users. This turned out to be a case of very weak anonymisation, since AOL failed to consider the fact that individuals often perform web searches for their own names, rapidly allowing interested individuals to significantly narrow the list of potential names (Arrington 2006). While the Netflix and AOL examples took place several years ago, they exemplify a continuing concern.

Privacy concerns go beyond the technical difficulty of anonymising data. In a recent paper, Borgesius et al. (2016) highlight 'three kinds of concerns about releasing personal information as open data: (1) the chilling effects on people interacting with the public sector, (2) a lack of individual control over personal information, and (3) the use of open data for social sorting or discriminatory practices'. There is general consensus that there is no foolproof way to completely anonymise a dataset, because linking de-identified data to other sources of data can often give enough information to identify individuals (O'Hara 2011).

Loss of public trust

The chilling effects detailed by Borgesius et al. (2016) can be tied to a loss of public trust. As O'Hara put it, 'not only are privacy and transparency compatible, privacy is a necessary condition for a successful transparency programme' (O'Hara 2011). If individuals in a study don't trust that their privacy is being taken seriously, the programme in question will run into serious problems.[5]

Experience shows that it is critically important for the public to feel that privacy has been considered in the decision-making process around data release and sharing (O'Hara 2011). InBloom was a private data analytics company working with educational data from a number of states. The company's goal was to help teachers tailor assignments to better suit the needs of individual students. While 'there weren't any documented cases of InBloom misusing the information' that the company held, InBloom did not demonstrate to the community's satisfaction that the company was taking privacy seriously. There was serious pushback from parents and privacy advocates and the company was eventually forced to shut down (Kharif 2014). This lesson is applicable to government agencies and companies working with sensitive information.

Discriminatory practices

Scassa (2014) explains this risk in more detail as 'the potential for open government data – even if anonymised – to contribute to the big data environment in which citizens and their activities are increasingly monitored and profiled'. In January 2016, the US Federal Trade Commission (FTC) released a report looking at the

5 See, for example, Kharif (2014).

potential for big data to be used for discrimination (Federal Trade Commission 2016). That report followed a 2014 document released by the White House that assessed opportunities and risks associated with big data (Podesta et al. 2016).

Predictive policing has been cited as a data-driven area that has significant built-in risks of discrimination. For example, police reports may be used for predictive purposes, but neighborhoods with 'lots of police calls aren't necessarily the same places the most crime is happening. They are, rather, where the most police attention is – though where that attention focuses' is often directed by gender and racial biases (Isaac & Dixon 2017).

The 2014 White House report on big data and privacy, released right after InBloom announced that it was shutting down, used educational data as an example of this concern. 'As students begin to share information with educational institutions,' the report said, 'they expect that they are doing so in order to develop knowledge and skills, not to have their data used to build extensive profiles about their strengths and weaknesses that could be used to their disadvantage in later years' (Podesta et al. 2014).

Current legal and policy frameworks

A number of laws and guidelines provide a framework for ensuring privacy for individuals who share information with the government, and for communicating about privacy safeguards. Some of the broader legal and policy frameworks include the following:

Freedom of information laws

Freedom of information laws 'provide inspiration on how to strike a balance between privacy and transparency in the open data context [… they] typically aim to accommodate privacy interests, for example by reserving access to personal information to parties with particular interests, or by only making records available in secure reading rooms' (Borgesius et al. 2016). That said, these laws may have narrow privacy restrictions that do not protect against all the risks of misusing personal information.

Organization for Economic Cooperation and Development (OECD) Privacy Guidelines

First published in 1980, the OECD Privacy Guidelines were the first set of internationally agreed upon privacy principles (Kuschewsky 2013). They were updated and expanded in 2013. The Framework is widely utilised, but has been criticised for its 'risk-based approach [… as well as] for promoting business over privacy' (Borgesius et al. 2016).

Privacy impact assessments

US federal law requires government agencies to consider individual privacy broadly by requiring them to conduct Privacy Impact Assessments about their electronic information systems and data that may contain PII. These assessments can be useful when balancing the relative costs and benefits of releasing a dataset (Altman et al. 2015).

Fair Information Practice Principles

The Fair Information Practice Principles are 'a set of principles and practices that describe how an information-based society may approach information handling, storage, management, and flows with a view toward maintaining fairness, privacy, and security in a rapidly evolving global technology environment' (Dixon 2008). The principles are internationally recognised and were developed over decades by a number of international bodies including the US Departments of Health, Education, and Welfare, and the OECD (Dixon 2008). These principles have been lauded for their 'balance [between] privacy-related interests and other interests, such as those of business and the public sector' (Borgesius et al. 2016).

Methodology

The Center for Open Data Enterprise used a multimethod approach to identifying strategies to best manage data release and privacy protection. This included desk research; an Open Data Roundtable with legal, policy and technical experts on open data and policy; solicitation of expert feedback; and interviews. The sequence of work was as follows:

(1) Review of existing literature on data and privacy issues. From this, an initial framework for identifying the challneges, solutions, and experts was developed.

(2) Information collection through an online public survey. Questions assessed:
    - Respondents' evaluation of the key issues in open data and privacy
    - Effectiveness of current approaches used to address challenges in open data and privacy
    - Respondents' interest in participating in the roundtable

    The survey received 61 responses, which were used to inform the plan for the roundtable and preparation of background materials.

	Legal	Policy	Technical	Total
Academic	1	1	2	4
Company	1	1	6	8
Government	8	14	16	38
Non-profit	2	5	4	11
Total	**12**	**21**	**28**	**61**

(3) Preparation of a briefing paper for background to the Open Data Roundtable, based on literature review and survey responses.

(4) An all-day Open Data Roundtable, held on 24 March 2016, to address the issue: how to open granular information while protecting privacy. The roundtable brought together 75 participants from federal agencies, academia, the private sector, and non-profit organisations with technical, policy, and legal expertise. This facilitated discussion included presentations, small-group breakout sessions, reports back to the full group, and synthesis of findings by the Center for Open Data Enterprise.

	Legal	Policy	Technical	Total
Academic		1		1
Company	2	4	5	11
Government	13	21	13	47
Non-profit	2	11	3	16
Total	**17**	**36**	**21**	**75**

Roundtable participants were not asked to develop consensus recommendations but to provide individual observations and suggestions.

(5) Additional interviews with roundtable participants to provide additional details on existing projects and strategies.

Strategies for managing data release and privacy protection

While many government agencies are concerned about the privacy risk of opening data, policy-makers can create programmes and assessment tools that reduce these risks to release data for the public good. In developing their open data programmes, agencies should consider a range of strategies, and consider using them in combination to develop a holistic approach to data management. When truly sensitive data is at stake, agencies or cross-agency programmes will need to develop thorough, coordinated plans for privacy protection.

The responses to the survey, and the discussions at the roundtable itself, showed the need for a portfolio of strategies in addressing data privacy concerns. Some of the issues highlighted in the survey responses included the need to balance privacy risks against the public value of opening data; controlled access as

a strategy for handling sensitive data; the importance of community engagement; education about how data will be used; and building trust in the organisation that holds the data.

Participants at the roundtable also stressed the importance of including legal, policy, and data experts, as well as stakeholders including industry and civil society, to bring different perspectives to bear in devising privacy-protection strategies. The ultimate goal, they agreed, is to develop a portfolio of approaches for different situations. As one survey respondent put it, 'One size does *not* fit all use cases. The most appropriate method to protect data privacy and confidentiality depends on one's goals and objectives, risk tolerance, and audience.'

It is important to note that there is no one global view on privacy. Different areas of the world have different approaches, understandings, legal frameworks, and risk tolerances.[6] However, many of the strategies discussed in this paper should be useful for governments trying to strike a balance between privacy and openness, regardless of the local context.

Develop balancing tests

Agencies can balance the risks of releasing data against the potential for public good. They can thereby create customised privacy-protection programmes based on risk assessment for each type of data involved, recognising and assessing the actual risk for releasing a given dataset under different conditions. While the exact tradeoffs may be difficult to work out, the use of a 'balancing test' can be a useful framework for handling the risks and benefits of data release.

This is the approach the Consumer Financial Protection Bureau (CFPB) is planning to use to release data under the Home Mortgage Disclosure Act (HMDA). The CFPB is statutorily mandated to publicly disclose data under the HMDA while developing appropriate protections for borrower privacy in light of the HMDA's purposes. Following a recent rulemaking, the CFPB will use a balancing test with public input to determine the right balance of serving the public good and protecting individual privacy in this data release. The test, which has not yet been developed, will be used 'to determine whether and how HMDA data should be modified prior to its disclosure to the public in order to protect applicant and borrower privacy while also fulfilling the disclosure purposes of the statute' (Consumer Financial Protection Bureau 2014).

Balancing tests have also been explored in the academic literature around privacy and open data. Borgesius et al. (2016) propose a 'balancing framework to help public authorities address this question in different contexts. The framework takes into account different varying of privacy risks for different types of data. It also separates decisions about access and re-use, and highlights a range of

6 For a better understanding of the different views taken in Europe and the United States, see Van der Sloot (2011).

disclosure routes. A circumstance catalogue lists factors that might be considered when assessing whether, under which conditions, and how a dataset can be released.'

Customise privacy protection based on risk assessment for each agency or programme

Although there are risks to opening data, policy-makers can create programmes and assessment tools that reduce these risks. Data-sharing culture should recognise and assess the actual risk for releasing a given dataset under different conditions. The potential damage from someone breaking the code and learning where an individual went to college, for example, is much less than the potential harm from revealing that same person's medical history. For that reason, each agency should assess the true risk for every dataset that contains PII and choose strategies for managing those datasets accordingly.

When truly sensitive data is at stake, agencies or cross-agency programmes will need to develop thorough, coordinated plans for privacy protection. For example, the US Precision Medicine Initiative (PMI), which is intended to help patients personalise their health care, has developed a framework for protecting privacy without inhibiting this scientific work. The PMI is part of a new approach to disease treatment and prevention that 'takes into account individual variability in genes, environment, and lifestyle for each person'. The success of the PMI – and precision medicine more broadly – will require researchers, providers and patients to 'work together to develop individualised care' and will rely heavily on patient participation (National Institutes of Health n.d.). The PMI Privacy and Trust Principles 'articulate a set of core values and responsible strategies for sustaining public trust and maximising the benefits of precision medicine'. Developed by an inter-agency working group with expert consultation, they are broken down into six key areas: governance, transparency, respecting participant preferences, participant empowerment through access to information, data sharing, access, use, and data quality and integrity (The White House 2015).

Data governance in each agency should also consider a range of possible conditions and risks. Governance approaches make a distinction between 'good actors' and 'bad actors'. When data is released to good actors, such as qualified researchers, re-identification risk can be limited through agreements on conditions of data use. These kinds of agreements can provide a 'trust framework' to govern the use of data effectively. At the same time, trust frameworks are useless against 'bad actors' who want to breach privacy protections on purpose.

Agencies may want to use 'threat modeling' to identify worst-case scenarios and decide what measures they need to prevent them. Threat modeling is a concept applied to network security, where it involves identifying system objectives, vulnerabilities, and countermeasures to prevent or reduce the impact of potential threats to the system. The same concept can be applied to privacy

issues by developing scenarios where bad actors might try to break through security safeguards to identify individuals in a database, and planning effective preventive measures.

Apply differential access

It may be necessary to consider gradations of openness under different circumstances. For example, some kinds of data could be made 'open' only for sharing between federal agencies under certain conditions, or sharing only with qualified and vetted researchers, rather than opening it to the public at large. Approaches include:

- Inter-agency transfer of data that is controlled and kept securely between the two agencies involved.
- Federated model using a cloud repository and limiting access to trusted users. This model requires a secure way to upload data as well as secure ways to share it.
- Tiered access data-sharing programmes to allow levels of access to multiple types of users.
- Opt-in and permission-based mechanisms that enable individuals to make their data more widely available if they choose to. For example, individual patients have an incentive to share data about their condition in the hope that it will be used to find better treatments.

One of the first priorities of the Precision Medicine Initiative was a set of Privacy and Trust Principles that 'articulate a set of core values and responsible strategies for sustaining public trust and maximising the benefits of precision medicine'. They aim to ensure transparency, strong governance, and data quality while empowering patients and protecting privacy (The White House 2014). The principles for data sharing, access, and use, for example, include using methods to preserve the privacy of patients' records, prohibiting unauthorised re-identification of patients, and establishing multiple tiers of data access, from open to controlled, depending on the nature of the data. Overall, the Privacy and Trust Principles outline a strong framework for applying many current approaches to balancing data sharing with privacy.

Employ de-identification technologies

It seems to be impossible to create a method of de-identification that removes all the privacy risks of PII from public datasets while also retaining the full value of the data for analysis.[7] However, it may be possible to provide a secure level of

7 For a comprehensive look at the inability of anonymisation to function as a prescription for

de-identification if researchers can accept a loss of some detail and granularity in the resulting dataset. Approaches to de-identification include:

- Identifying individuals with unique ID numbers that make it possible to connect data about them in different datasets without revealing their identity.
- Dropping non-critical information to make re-identification more difficult. For example, one regular practice is to drop the last three digits of an individual's zip code.
- Using differential privacy and synthetic data. Differential privacy applies algorithmic research to the problem of data privacy. At its best, it 'can make confidential data widely available for accurate data analysis'. Over time, however, this method can also become vulnerable to re-identification. Therefore, 'the goal of algorithmic research on differential privacy is to postpone this inevitability as long as possible' (Dwork & Roth 2014). Synthetic data relies on 'a complex statistical model that generates a simulated population that has the same general features as the original data'. While it has several existing applications, there is no consensus on its broad usefulness (Callier 2015). These are both sophisticated tools that require resources and data science expertise to apply.

The technical challenge of de-identifying data is becoming increasingly complex. De-identification technology is difficult to apply to the range of data now available, including geospatial, medical and genomic, body-camera and other data. Finally, even if it is possible to de-identify data today, it could become possible to re-identify individuals as technology evolves in the future. If de-identification or related strategies are being used as part of a broader privacy protection strategy, 'The decision of how or if to de-identify data should thus be made in conjunction with decisions of how the de-identified data will be used, shared or released' (Garfinkel 2015).

Enhance data governance structures

New data governance structures can help manage privacy concerns. In the US, many agencies now handle privacy issues through a chief privacy officer, a disclosure review board, or other offices and organisational structures. To make privacy protection as effective as possible, governance structures and safeguards need to be integrated and aligned with goals for data release. Options include:

- Identifying a single agency leader (for example, a chief data officer) to centralise each agency's management of open government data and address privacy concerns.

privacy concerns, see Ohm (2010).

- Develop core sets of policies and procedures that can be customised for each agency.
- Create model infrastructure – a virtual central data hub where access to data and APIs is managed by a common set of metadata (security, definitional, sharing licences) and user agreements.

Build trust with the community

Individual privacy should be treated in the context of public good. Many datasets that include PII also include information that can have great public benefit. In these cases, it will be essential to craft approaches to privacy protection that respect individuals' rights while also making data available to the public, or to selected researchers, in a way that supports social and scientific goals.

It is also essential to communicate the goals of open data, and privacy safeguards for the data, to the community and individuals that have provided it. Individuals are understandably concerned that data about their health, education, employment, financial status, or other sensitive data should not be exposed or misused. Agencies and others that plan to use the data with appropriate privacy protections will need to be sure that the communities involved understand and are satisfied with their approach.

One successful example from the U.S. has been the Police Data Initiative (PDI), launched in May 2015 with an initial group of 21 police departments from across the country, along with a range of partners. Through the PDI, police departments are working with data and technology partners to overcome technical and other hurdles and improve data sharing and analysis. Working with police data poses challenges to security and privacy, including concerns about releasing data on potential perpetrators, victims, and individual officers' actions. Several police departments have taken this challenge as an opportunity to work with the community to find solutions together. For example, 'the New Orleans Police Department...previewed policing datasets with a group of young coders and their tech mentors [and] the Orlando Police Department worked with sexual assault and domestic violence victim advocates to figure out how to balance transparency with victim privacy'. By taking this kind of approach, a number of 'communities and police departments [are] using data as a way to engage in dialogue and build trust' (Wardell & Ross 2016).

Conclusion

There is no single, foolproof solution to the challenge of protecting privacy when open data is released. However, a combination of strategies can make it possible to tap the value of granular, detailed data while managing privacy risks. While some strategies involve technical approaches, others are based on policy, data governance, community outreach and communication. These strategies should

be applicable not only in the US, where this research was based, but in other countries and contexts around the world.

As technology and policy around privacy evolve, more research will help open data programmes optimise their strategies for privacy protection. Researchers may choose to focus on the potential and limits of different technical approaches; the conditions for success of different privacy-protection strategies; protocols for releasing data with different 'degrees of openness'; cultural and social expectations of privacy in different communities; or other topics that help to develop a multifaceted approach to privacy protection in the context of open data.

Acknowledgements

The Center for Open Data Enterprise thanks its open data partner Microsoft and open data supporter Booz Allen Hamilton for supporting the Center's work on the 2016 Open Data Roundtables and related research. We also thank participants in the Open Data Roundtable on Privacy, the White House Office of Science and Technology Policy for providing input and feedback throughout the process, and the Center team, fellows, and interns who have all contributed to this effort throughout 2016. Parts of this paper have also been published in the 'Open Data and Privacy' Briefing Paper (2016) and the Roundtable Report (2016) published by the Center for Open Data Enterprise.

About the authors

JOEL GURIN is President and Founder of the Center for Open Data Enterprise, a Washington-based non-profit that works to maximise the value of open data as a public resource. Before launching the Center in January 2015, he wrote the book *Open Data Now* and led the launch team for the GovLab's Open Data 500 study and Open Data Roundtables. He was Chair of the White House Task Force on Smart Disclosure and has served as Chief of the Consumer and Governmental Affairs Bureau of the US Federal Communications Commission and as Executive Vice-President of Consumer Reports. E-mail: joel@odenterprise.org

MATT RUMSEY is a researcher and consultant focused on government data and information policy. He previously worked at the Sunlight Foundation, where he helped develop and implement federal policy initiatives. He holds a Bachelors in History from the American University in Washington DC.

AUDREY ARISS is the Director of Research and Design at the Center for Open Data Enterprise in Washington DC, where she co-leads international initiatives. Audrey has spent the past five years focusing on the use of data and technology for economic development. Audrey holds a masters in International Affairs and Quantitative Methods from Columbia University and a bachelors degree in

Modern History and French from the University of Oxford. She was a 2013 Google Policy Fellow.

KATHERINE GARCIA is a communications professional and consultant focused on open data and sustainability. Previously, Katherine was a member of the founding team for the Center for Open Data Enterprise, where she managed communications and outreach. She earned her Master of Public Administration degree with an emphasis in policy analysis from Baruch College and her Bachelors in Communication at the University of California, Santa Barbara.

REFERENCES

Altman, M., Wood, A., O'Brien, D.R., Vadhan, S. & Gasser, U. (2015). Towards a modern approach to privacy aware government data releases. *Berkeley Technology Law Journal* 30(30). http://btlj.org/data/articles2015/vol30/30_3/1967-2072%20Altman.pdf

Arrington, M. (2006, 6 August). AOL Proudly Releases Massive Amounts of Private Data. *TechCrunch*. http://techcrunch.com/2006/08/06/aol-proudly-releases-massive-amounts-of-user-search-data/

Borgesius, F., Van Eechoud, M. & Gray, J. (2015). Open data, privacy, and fair information practice principles: Towards a balancing framework. *Berkeley Technology Law Journal* 30(3): 2074-2099. http://btlj.org/data/articles2015/vol30/30_3/2073-2132%20Borgesius.pdf

Callier, V. (2015, 30 July). How fake data could protect real people's privacy. *The Atlantic.* http://www.theatlantic.com/technology/archive/2015/07/fake-data-privacy-census/399974/

Center for Open Data Enterprise (2016). Briefing paper on open data and privacy'. http://www.opendataenterprise.org/convene.html

Center for Open Data Enterprise (2015). Improving safety data: A Roundtable with the US Department of Transportation. https://s3.amazonaws.com/odenterprise/DoT+Roundtable+Report.pdf

Consumer Financial Protection Bureau (2014). *Final Rule Home Mortgage Disclosure Regulation, Docket No. CFPB-20140-0019*. http://files.consumerfinance.gov/f/201510_cfpb_final-rule_home-mortgage-disclosure_regulation-c.pdf

Consumer Financial Protection Bureau (n.d.). *The Home Mortgage Disclosure Act: About HMDA.* http://www.consumerfinance.gov/hmda/learn-more

Consumer Financial Protection Bureau (2015) *CFPB Finalizes Rule to Improve Information About Access to Credit in the Mortgage Market* http://www.consumerfinance.gov/about-us/newsroom/cfpb-finalizes-rule-to-improve-information-about-access-to-credit-in-the-mortgage-market/

Data.gov (n.d.). Project Open Data Dashboard. http://labs.data.gov/dashboard/offices

Dixon, P. (2008). *A Brief Introduction to Fair Information Practices*. World Privacy Forum. https://www.worldprivacyforum.org/2008/01/report-a-brief-introduction-to-fair-information-practices/

Dwork, C. & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science* 9(3–4): 211-407. http://www.nowpublishers.com/article/DownloadSummary/TCS-042, 5

Federal Trade Commission (2016). Big data: A tool for exclusion or inclusion? https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf

Garfinkel, S. (2015). De-identification of personal information. National Institute of Standards and Technology, 8053. http://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf

Isaac, W. & Dixon, A. (2017, 10 May). Why big-data analysis of police activity is inherently biased, *The Conversation*, https://theconversation.com/why-big-data-analysis-of-police-activity-is-inherently-biased-72640

Kharif, O. (2014, 1 May). Privacy fears over student data tracking lead to InBloom's shutdown. *Bloomberg Business*. http://www.bloomberg.com/bw/articles/2014-05-01/inbloom-shuts-down-amid-privacy-fears-over-student-data-tracking

Kuschewsky, M. (2013, 23 September). Revised OECD privacy guidelines strengthen accountability principle. *Inside Privacy*. https://www.insideprivacy.com/international/revised-oecd-privacy-guidelines-strengthen-accountability-principle/

Narayanan, A. & Shmatikov, V. (2008). Robust de-anonymization of large datasets (How to break the anonymity of the Netflix prize dataset. University of Texas at Austin. http://arxiv.org/PS_cache/cs/pdf/0610/0610105v2.pdf

National Institutes of Health (n.d.). Precision Medicine Initiative Cohort Program https://www.nih.gov/precision-medicine-initiative-cohort-program

Obama, B. (2016, 9 February). Establishing the Federal Privacy Council Executive Order. The White House. https://www.whitehouse.gov/the-press-office/2016/02/09/executive-order-establishment-federal-privacy-council

O'Hara, K. (2011). Transparent government, not transparent citizens: A report on privacy and transparency for the Cabinet Office. London: Cabinet Office. http://eprints.soton.ac.uk/272769/

Ohm, P. (2010). Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review* 57: 1701-1777. www.uclalawreview.org/pdf/57-6-3.pdf

Open Government Guide (n.d.). Privacy and data protection. http://www.opengovguide.com/topics/privacy-and-data-protection/

Ortellado, D. (2016, 16 February). Reconciling criminal history open data and expungement. The Sunlight Foundation. http://sunlightfoundation.com/blog/2016/02/03/reconciling-criminal-history-open-data-and-expungement/

Park, T. & Shelton, J. (2012, 8 June). The power of open education data. The White House. https://www.whitehouse.gov/blog/2012/06/08/power-open-education-data-0

Press, E. (2010, 29 July). A case for Open Data in Transit. *Streetfilms.org*. http://www.streetfilms.org/a-case-for-open-data-in-transit/

Podesta, J., Pritzker, P., Moniz, E., Holdren, J., & Zients, J. (2014, 1 May). Big data: Seizing opportunities, preserving values. The White House Executive Office of the President. https://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf

Polonetsky, J., Tene, O., & Finch, K. (2016). Shades of gray: Seeing the full spectrum of practical data de-identification. *Santa Clara Law Review* 593. http://digitalcommons.law.scu.edu/lawreview/vol56/iss3/3

Scassa, T. (2014). Privacy and open government. *Future Internet* 6(2): 397–413. http://doi.org/10.3390/fi6020397

Shaw, E. (2015, 22 January). What do we want? Data about police practice! The Sunlight Foundation. http://sunlightfoundation.com/blog/2015/01/22/what-do-we-want-data-about-police-practice/

Shaw, E. (2014, 24 October). Exploring open data's microdata frontier. The Sunlight

Foundation. https://sunlightfoundation.com/blog/2014/10/24/exploring-open-datas-microdata-frontier/
Sweeney, L. (1997). Weaving technology and policy together to maintain confidentiality. *Journal of Law, Medicine & Ethics* 25(2&3): 98-110
The Open Data Institute (n.d.). Data spectrum: The data spectrum helps you understand the language of data. https://theodi.org/data-spectrum.
Sunlight Foundation (2014). The benefits of criminal justice data: Policing and beyond. The Sunlight Foundation. http://assets.sunlightfoundation.com/criminaljustice/sunlight-policy-brief-the-benefits-of-criminal-justice-data-policing-and-beyond.pdf
The White House (2015, 9 November). Precision Medicine Initiative: Privacy and trust principles. https://www.whitehouse.gov/sites/default/files/microsites/finalpmiprivacyandtrustprinciples.pdf
The White House (n.d.). The Precision Medicine Initiative. https://www.whitehouse.gov/precision-medicine
Thomas, K.E. (2016, 28 January) What 22 million rides tell us about NYC bike-share. *Next City.* https://nextcity.org/daily/entry/citi-bike-new-york-city-bike-share-data
Tierney, K.F. (2012). Use of the US Census Bureau's Public Use Microdata Sample (PUMS) by State Departments of Transportation and Metropolitan Planning Organizations. Transportation Research Board of the National Academies. http://onlinepubs.trb.org/onlinepubs/nchrp/nchrp_syn_434.pdf
Van der Sloot, B. (2011). On the fabrication of sausages, or of open government and private data. *eJournal of eDemocracy and Open Government* 3(2). http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2323771
Wardell, C. & Ross, D. (2016, 22 April). The Police Data Initiative year of progress. The White House. https://medium.com/the-white-house/the-police-data-initiative-year-of-progress-how-we-re-building-on-the-president-s-call-to-leverage-3ac86053e1a9#.58iuq5xo7