Penalties for Unauthorized Disclosure and Data Privacy

Len Burman,
Urban Institute/ Tax Policy Center

October 2022

The author welcomes feedback on this working paper. Please send all inquiries to lburman@urban.org.

Claire Bowen and Mayank Varia provided helpful comments on an earlier draft.

Abstract

The literature on data privacy has tended to ignore the value of penalties in strengthening data protection while allowing more and better data to safely be shared for research purposes. This note explains briefly why significant penalties and effective enforcement for those who make unauthorized disclosures of confidential data should be an essential complement to privacy-preserving technologies and data access rules designed to protect against disclosure. Evidence from tax evasion strongly suggests that penalties, along with high probability of detection, produce nearly perfect compliance with tax rules. In combination with other measures designed to protect data privacy, penalties for disclosure could significantly reduce the likelihood that anyone would even attempt an unauthorized disclosure. Moreover, penalties could allow more and better information to be safely released without increasing the overall risks to privacy.

The literature on data privacy has tended to ignore the value of penalties in strengthening data protection while allowing more and better data to safely be shared for research purposes. This note explains briefly why significant penalties and effective enforcement for those who make unauthorized disclosures of confidential data should be an essential complement to privacy-preserving technologies and data access rules designed to protect against disclosure. Evidence from tax evasion strongly suggests that penalties along with high probability of detection produce nearly perfect compliance with tax rules (Slemrod, 2019). In combination with other measures designed to protect data privacy, penalties for disclosure could significantly reduce the likelihood that anyone would even attempt an unauthorized disclosure. Moreover, penalties could allow more and better information to be safely released without increasing the overall risks to privacy.

How Adding Penalties and Enforcement Reduces Disclosure Risk

The model of disclosure risk used in the differential privacy literature does not account for the cost of inferring confidential information or the value of that information. It simply assesses the probability that any statistical analysis could change with the addition or subtraction of any record that could be in the dataset. This probability is measured as the sensitivity of an analysis to the most extreme value that could exist in a dataset—whether or not it actually is in the dataset. (Bowen, 2022) Model results or data must be altered so that this sensitivity is below a threshold determined by the data steward, denoted by ε. Larger ε requires less alteration to the information released but is less protective of privacy. A small ε protects against privacy loss—at the price of precision.

The differential privacy framework assumes that an intruder could possess all but one of the records in a dataset along with nearly infinite computing capacity and time to perform the computations. However, realistically, the acquisition of the needed data and computing power would be quite costly. The value of that last record would have to be enormous to justify such an attack. Alternatives to differential privacy make less extreme assumptions and are applicable to a broader range of statistical analyses (e.g., zero-concentrated differential privacy, the definition applied to the 2020 Census, or the differential-privacy-inspired definition by Chetty and Friedman, 2019), but they remain agnostic about the types of attack that might occur.

A simple economic model would posit that a potential intruder would only attack a dataset if the benefits were at least as great as the costs. Privacy protections reduce the net benefit by reducing the probability that useful information could be gleaned about any individuals. In the economic model, as in the differential privacy framework, lower ε translates into less risk.

But other policy choices could also reduce disclosure risk. Penalties on disclosure of confidential information would raise the cost of attempting to infer confidential information. Penalties are most effective when the probability of detection is high. For example, tax compliance studies have repeatedly found that compliance is very high (93 percent) for types of income where the IRS has substantial information about the source of income, but it is quite low where there is no information (37 percent). Slemrod (2019) concludes that there is "stark and compelling evidence for the primary importance of deterrence as an explanation of tax evasion." (p. 916)

There is no direct evidence about how legal sanctions affect the likelihood of unauthorized data disclosures, but the economics would be similar to those applying to tax evasion.[1] Thus, imposing legal sanctions including large financial penalties and possible jail time on anyone who uses confidential data about individuals or firms for purposes other than the specific purposes identified in a data-sharing agreement would substantially reduce the likelihood of disclosure. Internal Revenue Code section 6103 could be a good model for broader legislation. A government employee or contractor who discloses confidential taxpayer information for purposes other than permitted in the statute may be convicted of a felony and subject to fines, court costs, and up to five years in jail (Berggren 1999). Those penalties should be extended to *anyone* who discloses confidential information. This would include organizations that receive stolen confidential data and disclose it, and organizations that use computational or statistical techniques to reidentify individuals and firms in an anonymized dataset. That is, these sanctions would be a backstop to disclosure control methods.

With substantial penalties and effective enforcement and prosecution, the legal sanctions could reduce the probability of any disclosure as much as a very small ε, while preserving the integrity of the data for research purposes. The economic decision to disclose confidential data,

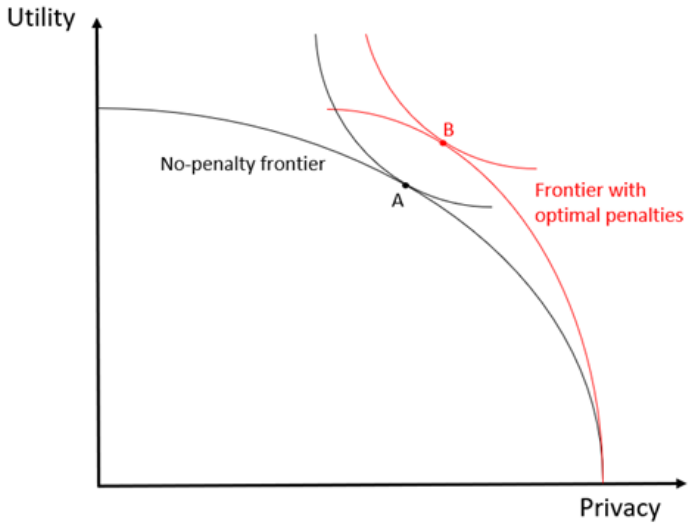
[1] This framework dates to Gary Becker's seminal analysis of the economics of crime (Becker 1968).

like the decision to evade taxes, amounts to a calculation that the benefits of disclosure exceed the costs. For data that are anonymized and subject to other statistical disclosure controls, the cost could include the cost of acquiring other data that could allow matching and computational resources applied to try to undo privacy protections. Applying more stringent privacy protections (e.g., lower ε) raises the cost of reverse engineering the confidential data, but raising the legal penalties and the probability of detection and prosecution would have the same effect.

The penalties might be designed to eliminate almost all disclosure risk by themselves, but the sanctions would have to be draconian and substantial resources would need to be devoted to enforcement. And there are cases where legal sanctions alone would have limited effect, as discussed below. An optimal mix of policy instruments would include a combination of privacy-protection measures and disclosure penalties and enforcement that minimize public and private costs.

This can be visualized by adding enforcement to the familiar tradeoff between privacy and utility (usefulness). The tradeoff is represented by a strictly concave downward sloping function, as depicted on Figure 1. Utility may be maximized by forgoing any privacy protection. In the context of differential privacy, this would correspond to $\epsilon \to \infty$. In the context of traditional statistical disclosure control methods, it would represent making the confidential data including personal identifiers public. Privacy protections, such as anonymizing the data, improve privacy with little cost to utility. But as privacy is enhanced (e.g., setting a small ε), less information may be released and the usefulness of the data for analysis declines. Eventually, the data may be protected from any possibility of disclosure, but only by preventing any data release (or randomizing the data that are released). This is the right-most point of the privacy-utility frontier. The optimal tradeoff may, in principle, be found by finding the point that maximizes social welfare, where equal levels of social welfare are represented by the convex curves on Figure 1. In the illustration, point A represents the optimal levels of utility and privacy in framework without penalties.

Figure 1. Privacy-Utility Trade-off with and without Disclosure Penalties



When penalties and enforcement are introduced, the utility-privacy frontier moves up and to the right. With penalties set at the optimal level, the optimal tradeoff between privacy and utility moves to point B on Figure 1.[2] Both privacy and utility increase. That is, more and better data may be released while the risk of reidentification of any individual record declines. This is clearly a gain for society.

Exemption for Good-Faith Security Research

There should be an exemption from legal penalties for privacy researchers who conduct research into re-identification techniques in a responsible manner (i.e., without publicly disclosing any personally identifiable information or doing anything else that would cause harm), so that data stewards can learn about vulnerabilities in their data. Such research could also help privacy researchers to get a better idea of realistic threats to privacy as opposed to the expansive theoretical risks inherent in the differential privacy framework.

The Library of Congress defined a "good-faith security research" exemption to the Digital Millenium Copyright Act, where "…'good-faith security research' means accessing a

[2] The optimal level of penalties is derived by adding penalties to the social welfare function. Penalties and enforcement may entail direct costs to society because enforcement is burdensome and indirect costs because the cost of administering an expansive enforcement regime requires higher taxes, bigger deficits, or cuts in other public services. If the parameter π represents the penalty level and ε the privacy-protection measure, then the optimal penalty level π^* occurs where social welfare is maximized given that ε is set at the optimal level.

computer program solely for purposes of good-faith testing, investigation, and/or correction of a security flaw or vulnerability, where such activity is carried out in an environment designed to avoid any harm to individuals or the public, and where the information derived from the activity is used primarily to promote the security or safety of the class of devices or machines on which the computer program operates." (Register of Copyrights, 2021, p. 258)

The Department of Justice applied the same exemption to the Computer Fraud and Abuse Act.[3]

Limitations of Disclosure Penalties

As noted, penalties alone are not sufficient to prevent all disclosure. There are at least three limitations. First, some potential data intruders might be outside the reach of any practical enforcement mechanism. Notably, entities in countries that do not have extradition treaties with the United States that cover crimes such as unauthorized release of confidential information might try to breach US databases with relative impunity. Second, it may be difficult to identify the source of a data leak, especially if the leaked information is not publicly disclosed. The leak could still be quite damaging to the individuals affected.[4] Third, if a press organization makes illegally acquired information public, they may be protected from sanctions by the First Amendment guarantee of press freedom.

These considerations make clear that privacy protections are essential for any released information. For example, one of the most famous data disclosures occurred when the investigative journalism organization, ProPublica claimed to have "obtained a vast cache of IRS information showing how billionaires like Jeff Bezos, Elon Musk and Warren Buffett pay little in income tax compared to their massive wealth." (Eisinger, Ernsthausen and Kiel, 2021) Even if ProPublica's reporting were protected by the First Amendment, their report would have been far less newsworthy if the data they received were subject to even minimal privacy protections, such as removing individual identifiers. Other standard privacy protections would have allowed at

[3] See https://www.justice.gov/opa/press-release/file/1507126/download, p. 4.

[4] To take an extreme example, a student told me that terrorists in Colombia had reportedly used illegally obtained income tax return information to select targets for kidnapping. The damage to those who are kidnapped is largely independent of whether the terrorists publicize the tax return information.

most the assertion of probabilistic statements, such as that some people probably exist who have very high incomes but pay little income tax, a fact that was well known before the ProPublica leak.

The question remains about how to prevent the release of confidential administrative data that has not been subject to any privacy protections, particularly if press publication is exempt from sanctions. One approach could be to invest in monitoring software that could detect any attempt to produce and download a database such as the one that ProPublica received. Sensitive government and private datasets have also been hacked, including the reported Russian breach of extremely sensitive data by infecting SolarWinds cybersecurity software. (Sanger, Perlroth and Schmitt, 2021) This suggests that investing in better cybersecurity measures also should be an important component of protecting privacy.

References

Becker, Gary S. 1968. “Crime and Punishment: An Economic Approach.” *The Journal of Political Economy* 76(2): 169-217.

Berggren, Mark. 1999. "IRC 6103: Let's Get to the Source of the Problem." *Chicago-Kent Law Review* 74(2): 825-853.

Bowen, Claire McKay. 2021. *Protecting your Privacy in a Data-Driven World*. Chapman and Hall/CRC.

Chetty, Raj, and John N. Friedman. 2019. "A Practical Method to Reduce Privacy Loss When Disclosing Statistics Based on Small Samples." *Journal of Privacy and Confidentiality* 9(2): 1-23.

Eisinger, Jesse, Jeff Ernsthausen and Paul Kiel. 2021. “The Secret IRS Files: Trove of Never-Before-Seen Records Reveal How the Wealthiest Avoid Income Tax.” ProPublica. https://www.propublica.org/article/the-secret-irs-files-trove-of-never-before-seen-records-reveal-how-the-wealthiest-avoid-income-tax.

Register of Copyrights. 2021. *Section 1201 Rulemaking: Eighth Triennial Proceeding to Determine Exemptions to the Prohibition on Circumvention: Recommendation of the Register of Copyrights*, 8 October. Washington, DC: The Register of Copyrights. https://cdn.loc.gov/copyright/1201/2021/2021_Section_1201_Registers_Recommendation.pdf.

Sanger, David E., Nicole Perlroth and Eric Schmitt. 2021. “Scope of Russian Hacking Becomes Clear: Multiple U.S. Agencies Were Hit.” *New York Times*. https://www.nytimes.com/2020/12/14/us/politics/russia-hack-nsa-homeland-security-pentagon.html.

Slemrod, Joel. "Tax Compliance and Enforcement." 2019. *Journal of Economic Literature* 57(4): 904-54.