The Deidentification Dilemma: A Legislative and Contractual Proposal

Robert Gellman[1]

Version 2.4

July 12, 2010

Abstract

Deidentification is one method for protecting privacy while permitting other uses of personal information. However, deidentified data is often still capable of being reidentified. The main purpose of this article is to offer a legislative-based contractual solution for the sharing of deidentified personal information while providing protections for privacy. The legislative framework allows a data discloser and a data recipient to enter into a voluntary contract that defines responsibilities and offers remedies to aggrieved individuals.

Introduction

The goal of protecting the privacy interests of individuals often conflicts with ever-increasing demands for use of personal data to achieve potentially beneficial objectives in public health, law enforcement, national security, anti-terrorism, fraud prevention, and research in different fields of study. Conflicts over privacy can be reduced, moderated, and balanced in various ways.

Deidentification – the removal of identifiers from personal information used or disclosed for purposes unrelated to the purpose for which the information was originally obtained – is one method for protecting privacy while permitting other uses of personal information. However, deidentification does not always make it impossible to reidentify individuals. Reidentification is the linkage of deidentified personal information with an overt identifier belonging or assigned to any living or dead individual.

It is a premise of this article that statistical, encryption, or other mathematical approaches to deidentification that protect privacy will not provide solutions that will address every type of data and every data sharing activity. These approaches still have value, but they will sometimes or often not achieve maximum privacy protection. No matter how many identifiers have been removed or encrypted and no matter how much data has been coded or masked, it may still be the case that the remaining data can be reidentified. Further, the value of data for legitimate uses, such as research, may be significantly reduced when the data is processed extensively to protect privacy. In the absence of a technical solution to the possibility of reidentification, other approaches are needed.

The solution presented here focuses on controlling reidentification and providing accountability for promises not to reidentify information. This article offers a legislative-based

[1] The author is grateful to Steven Cope, Latanya Sweeney, Peter Winn, Mark Rothstein, and Susan Landau for their comments and assistance with this article.

contractual solution for the sharing of deidentified personal information while providing protections for privacy. The legislative framework allows a data discloser and a data recipient to enter into a voluntary contract that defines responsibilities and offers remedies to aggrieved individuals. This contractual solution can be useful whether personal information is deidentified in support of academic research or other objectives. The proposal is not a universal guarantee of privacy, nor will it work for all data exchanges. It will, however, provide another tool to support the sharing of personal data while addressing the privacy interests of data subjects.

In this article, deidentification means that personal information has been processed in some fashion to reduce the ability to identify individuals. It does not mean that information has been anonymized to the point where reidentification is never possible.

The Problem

A major challenge for deidentification is the vast amount of personal information available from public and private sources in the United States and, increasingly, elsewhere around the world. The more personal data that is available, the easier it can be to link deidentified data to a particular individual. The commercial collection, compilation, and exploitation of personal data in the United States are extensive. Sources include public records (e.g., voter registers, occupational licenses, property ownership and tax records, court records), commercial data (e.g., transaction information), and even nonidentifiable data (e.g., census data). Extensive profiles of individuals and households exist in commercial files that may include name, address, former addresses, telephone number, educational level, home ownership, mail buying propensity, credit card usage, income level, marital status, age, children, and lifestyle indicators that show personal interests in gardening, sports, and other activities. Private companies increasingly maintain health records outside the reach of health privacy laws applicable to health care providers and insurers.[2] Internet websites, including social networking sites, are recent institutions that are additional sources of personal information, including search requests, movies watched, and other activities and interests.[3] Cellular telephones now track the location of an individual at all times. So-called digital signage tracks individuals in public spaces, collecting detailed information about consumers, their behaviors, and their characteristics, like age, gender, and ethnicity.[4]

Personal information that no longer contains overt identifiers (name, identification number, email address, telephone number, etc.) can still be linked with known individuals. Identity can be ascertained from simple, basic, widely available non-unique identifiers (sometimes called *quasi-identifiers*). For example, Professor Latanya Sweeney, a leading academic authority on statistics, identification, and policy estimates that 87% of Americans are

---

[2]See Robert Gellman, *Personal Health Records: Why Many PHRs Threaten Privacy*, (World Privacy Forum 2008), http://www.worldprivacyforum.org/pdf/WPF_PHR_02_20_2008fs.pdf.

[3] See, e.g., JR Raphael, *People Search Engines: They Know Your Dark Secrets…And Tell Anyone*, PC World (March 2009), http://www.pcworld.com/article/161018/people_search_engines_they_know_your_dark_secretsand_tell_anyone.html; Latanya Sweeney, *Information Explosion*, in P. Doyle et al., Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies (2001).

[4] Pam Dixon, *The One-Way-Mirror Society: Privacy Implications of the New Digital Signage Networks* (2010), http://www.worldprivacyforum.org/pdf/onewaymirrorsocietyfs.pdf.

likely to be uniquely identified from date of birth, gender, and five-digit zip code.[5] Removing, generalizing, or coding these or other non-unique identifiers may make the task of reidentification harder, but the data may still be reidentified. At the same time, deidentified data sets may be less useful for research and other uses because of the difficulty of linking data sets or because the data will no longer support complete or precise conclusions.

The federal health privacy rule provides an example of the difficulty of achieving – and even defining -- deidentification.[6] The rule provides that individually identifiable health information is deidentified if seventeen specific fields of data are removed or generalized.[7] Data deidentified according to this standard falls outside the rule's scope, and the rule allows the data to be freely disclosed to anyone or published.

However, notwithstanding the rule's determination that the resulting data is deidentified, Professor Sweeney, testified that there is a 0.04% chance that data deidentified under the health rule's methodology could be reidentified when compared to voter registration records for a confined population.[8] If a database deidentified under HIPAA standards had one million names, then four hundred people could likely be reidentified. If other public, commercially available, Internet, or private records were also to be consulted, the chances of reidentification should increase.

The HIPAA deidentification process may be the most specific and detailed regulatory approach to deidentification. Yet, even HIPAA's extensive and carefully considered effort at deidentification does not achieve complete anonymity for all data. Indeed, there may not be a realistic and practical standard for absolute deidentification. Professor Sweeney put it this way: "I can never guarantee that any release of [deidentified] data is anonymous, even though for a particular user it may very well be anonymous."[9] As a general proposition, for most personal data, deidentification may be like absolute zero for temperature: a state that can be approached but never achieved. Even if data could be fully deidentified, the prize may not be worth the effort in many cases. The data may no longer have significant value for researchers and other users.

---

[5] Comments of Latanya Sweeney, Ph.D., Carnegie Mellon University, on the Department of Health and Human Service's Standards of Privacy of Individually Identifiable Health Information (2002), available at http://privacy.cs.cmu.edu/dataprivacy/HIPAA/HIPAAcomments.html.

[6] The federal rule for health privacy was issued under the authority of the Health Insurance Portability and Accountability Act, 42 U.S.C. § 1320d-2 note. HIPAA rules cover both privacy and security. 45 C.F.R. Part 160, 162, 164.

[7] 45 C.F.R. § 164.514(b)(2). The rule has an 18th, catchall, field covering "any other unique identifying number, characteristic, or code." Id. at § 164.514(b)(2)(i)(R). In addition to removing the specified identifiers, the entity making the disclosure cannot have actual knowledge that the information "could be used alone or in combination with other information to identify an individual." Id. at § 164.514(b)(2)(ii).

[8] National Committee on Vital and Health Statistics, *Report to the Secretary of Health and Human Services on Enhanced Protections for 'Secondary Uses' of Electronically Collected and Transmitted Health Data* (Dec. 21, 2007), at 36 n.16, available at www.ncvhs.hhs.gov/071221lt.pdf. The National Committee on Vital and Health Statistics is an advisory committee to the U.S. Department of Health and Human Services.

[9] National Committee on Vital and Health Statistics, Subcommittee on Privacy and Confidentiality, *Proceedings of Roundtable Discussion: Identifiability of Data* (Jan. 28, 1998), available at http://ncvhs.hhs.gov/980128tr.htm.

Others have also written about the shortcomings of deidentification. A June 2010 article by Arvind Narayanan and Vitaly Shmatikov offers a broad and general conclusion:

> The emergence of powerful re-identification algorithms demonstrates not just a flaw in a specific anonymization technique(s), but the fundamental inadequacy of the entire privacy protection paradigm based on "de-identifying" the data.[10]

Paul Ohm suggests that "[u]ntil a decade ago, the robust anonymization assumption worked well for everybody involved."[11] He provides examples of several well-publicized releases of supposedly deidentified data that were ultimately found to be identifiable, including the AOL research data release of search queries and the Netflix Prize Data Study that involved the release of 100 million movie ratings by Netflix customers.[12]

From a policy perspective, identifiability of personal information is best viewed as a continuum. At one end of the continuum, information is fully identifiable due to the presence of names, identification numbers, and the like. Shedding overt identifiers moves data down the continuum where it becomes harder to link the data with individuals, but data may still be identifiable even with all overt identifiers removed. While it may be possible at times to achieve provably absolute deidentification using encryption, coding, hashing, and other techniques, it seems unlikely that there is general solution that will work for all types of data, all types of users, and all types of activities. Thus, we continue to face the possibility that deidentified personal data shared for research and other purposes may be subject to reidentification.

Existing Legal Approaches

Statisticians have long been aware of deidentification issues and have developed many techniques to address the possibility of reidentification.[13] However, existing laws do little to untangle the deidentification dilemma. Indeed, they tend to make it worse. Laws often reflect an assumption that identifiability is a binary state; personal data is either identifiable or not. They tend to ignore the reidentification issue altogether. Laws establish vague or inconsistent standards for identifiability. Some examples:

> • The Privacy Act of 1974,[14] a U.S. law that applies mostly to federal agencies, defines *record* to mean a grouping of information about an individual that contains "his name, or the identifying number, symbol, or other identifying particular assigned to the individual, such as a finger or voice print or a

---

[10] Arvind Narayanan & Vitaly Shmatikov, *Myths and fallacies of "personally identifiable information"*, 53 Communications of the ACM 24, 26 (June 2010), available at http://userweb.cs.utexas.edu/users/shmat/shmat_cacm10.pdf.

[11] Paul Ohm, Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization, 57 UCLA L. Rev. ___ at 15 (forthcoming 2010), available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1450006.

[12] Id at 16-20.

[13] See, e.g., Federal Committee on Statistical Methodology (U.S.), *Statistical Policy Working Paper 22 - Report on Statistical Disclosure Limitation Methodology* (2005) http://www.fcsm.gov/working-papers/spwp22.html; K. El Emam, *Heuristics for de-identifying health data*, 6 IEEE Security and Privacy58-61 (2008).

[14] 5 U.S.C. § 552a.

photograph."[15] An identifier is an essential part of a record. The ability to infer identity or to reidentify a record is not sufficient or relevant, no matter how easy it may be to accomplish the reidentification. Further, the law treats a fingerprint as an identifier, when few people without access to a law enforcement fingerprint database could identify an individual from a fingerprint. The Act's concept of identifiability is muddled, at best.

• The Cable Communications Policy Act does not define *personally identifiable information*, but it excludes from the term "any record of aggregate data which does not identify particular persons."[16] However, even aggregate data can be used to reidentify individuals in some circumstances. The statute does not address that possibility.

• The Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA) defines *identifiable form* to mean "any representation of information that permits the identity of the respondent to whom the information applies to be reasonably inferred by either direct or indirect means."[17] CIPSEA's definition is one of the few that explicitly addresses the use of indirect inferences to permit identification, but it does not indicate the scope of effort that is necessary to render deidentified data identifiable. Further explication would presumably require parsing the meaning of *reasonably*.

• Canada's Personal Information Protection and Electronic Documents Act (PIPEDA) defines *personal information* as "information about an identifiable individual."[18] Thus, PIPEDA offers no standard for determining identifiability or anonymity, nor does it address the issue of reidentification. A treatise on PIPEDA suggests that truly anonymous information does not qualify. It also suggests that "caution should be exercised in determining what is truly 'anonymous' information since the availability of external information in automated format may facilitate the reidentification of information that has been made anonymous."[19] That advice may be helpful, but the statute itself is silent.

• The European Union (EU) Data Protection Directive defines *personal data* to mean "any information relating to an identified or identifiable natural person, and it defines an identifiable person as "an individual person . . . who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental,

---

[15] Id. at § 552a(a)(4).
[16] 47 U.S.C. § 551(a)(2)(A).
[17] E-Government Act of 2002, Pub. L. 107-347, Dec. 17, 2002, 116 Stat. 2899, 44 U.S.C. § 3501 note § 502(4)
[18] S.C. 2000, c. 5, § 2(1), available: http://www.privcom.gc.ca/legislation/02_06_01_01_e.asp.
[19] Perrin, S., H.H. Black, D.H. Flaherty, and T.M. Rankin, The Personal Information Protection and Electronic Documents Act: An Annotated Guide 54 (2001).

economic, cultural or social identity."[20] The task of parsing these words for a clear standard is helped somewhat by the Directive's introductory Recital 26, which states that privacy rules will not apply to "data rendered anonymous in such a way that the data subject is no longer identifiable." It also provides that "to determine whether a person is identifiable, account should be taken of all the means likely reasonably to be used either by the controller or by any other person to identify the said person."[21] Based on the recital, it seems apparent that the Directive uses a reasonableness standard to determine whether information is sufficiently deidentified to fall outside the Directive's ambit.

A further gloss on the Directive's meaning of *personal data* can be found in an opinion of the Article 29 Working Party, an organization established under the Directive. The opinion offers 26 pages of detailed and interesting explanation for establishing what is and is not personal data, and what is and is not identifiable. The length of the analysis is evidence of the complexity of identifiability under current conditions. The Working Party's conclusion that a determination of anonymity depends on the circumstances and calls for a case-by-case analysis is further evidence of the essential murkiness of the identifiability concept.[22]

• The Alberta Health Information Act defines *individually identifying* to mean when a data subject "can be readily ascertained from the information,"[23] and it defines *nonidentifying* to mean that the identity of the data subject "cannot be readily ascertained from the information."[24] This appears to limit the identifiability inquiry to the information itself. Alberta's data matching law,[25] regulates the creation of individually identifying health information by combining individually identifying or nonidentifying health information or other information from two or more electronic databases without the consent of the data subjects. The data matching requirements include submission of a privacy impact assessment to the commissioner for review and comment.[26] The Alberta law expressly addresses reidentification activities by anyone (at least, anyone using any electronic databases). The Act has an administrative process rather than a statutory standard for determining whether identifiable information is at stake.

---

[20] Directive on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data, Council Directive 95/46/EC, 1995 O.J. (L 281) 31, at Article 2(a), http://europa.eu.int/comm/internal_market/en/dataprot/law/index.htm.
[21] Id. at Recital 26.
[22] Article 29 Data Protection Working Party, *Opinion 4/2007 on the concept of personal data* 21 (2007) (WP 136), http://ec.europa.eu/justice_home/fsj/privacy/docs/wpdocs/2007/wp136_en.pdf.
[23] Alberta Health Information Act § 1(p) (1999), available at http://www.qp.alberta.ca/574.cfm?page=H05.cfm&leg_type=Acts&isbncln=9780779739493.
[24] Id. at § 1(r).
[25] Id. at § 32(2).
[26] Id. at § 68-72.

In general, statutes and rules that address identifiability and deidentification can be grouped roughly into three categories.[27] The first category includes standards that seek to determine whether data is sufficiently or potentially identifiable to warrant regulation. The standards can (a) be inward-looking (considering only the data); (b) be outward-looking (considering other data actually or potentially available elsewhere as well as the capabilities for reidentification generally available to individuals or experts); (c) require professional statistical judgment; or (d) consider the time, effort, or cost required for reidentification. More than one of these standards can apply at the same time. A standard can also reference a reasonableness test, either directly or indirectly. As is apparent, a multiplicity of standards is available in this category.

The second category uses an administrative process. The Alberta law calls for administrative privacy review in advance of some reidentification activities. This type of review may be possible in a small jurisdiction, but it would be impractical in a larger one.

The third category is a rule that requires the removal of specified data elements. The HIPAA health privacy rule is a leading example. Some of its shortcomings have already been discussed.

This limited review of statutes and rules suggests the wide variance in identifiability standards that can be found. In most cases, the statutes offer alternate word formulas that are probably more the result of casual drafting rather than of a carefully considered approach based on detailed study or analysis. Recognition by legislators and policy makers of the complexities presented by deidentification of personal data has been slow to develop. Laws badly trail the capabilities of modern computers and experts to use the vast pools of personal data available today. Current technology allows the reidentification of data that most casual observers would have thought was adequately deidentified. All existing regulatory approaches suffer from shortcomings. The HIPAA rule provides greater certainty, but that certainty is somewhat misplaced. CIPSEA expressly recognizes the possibility of reidentification, but it offers little practical guidance. It might well take years of litigation before any useful test emerged, and the result from litigation may not satisfy anyone. The Alberta administrative process will not scale to larger jurisdictions or will require a cumbersome bureaucracy.

A Contractual Solution

No legislation can establish meaningful standards for the creation of deidentified data that has full value for legitimate secondary users. That objective cannot be reached now and may be impossible to achieve generally. There will always be a tradeoff of some sort, involving the degree of identifiability of the data, the usability of the data, the privacy of data subjects, and the cost of the deidentification process. Technology can sometimes lessen these tradeoffs, but it cannot eliminate them all the time.

---

[27] See generally, Robert Gellman, *Privacy for Research Data*, Panel on Confidentiality Issues Arising from the Integration of Remotely Sensed and Self-Identifying Data, National Research Council, Putting People on the Map: Protecting Confidentiality with Linked Social-Spatial Data (2007) (Appendix A), http://books.nap.edu/catalog.php?record_id=11865.

What legislation can do, however, is to establish rules that will allow data disclosers and data recipients to agree voluntarily on externally enforceable terms that provide privacy protections for data subjects. That is the main purpose of the law proposed here, the Personal Data Deidentification Act (PDDA).

The key definition in the PDDA is *potentially identifiable personal information* (or PI^2). The definition of PI^2 builds on a definition of *personal information*, which is any information about an individual, whether it contains a personal identifier or not. *Potentially identifiable personal information* is any personal information without overt identifiers. PI^2 is a new concept in the PDDA, included to cover the wide range of personal information without overt identifiers that is likely to be reidentifiable. Because it cannot be known at any time whether information is reidentifiable, virtually all personal information that is not overtly identifiable is PI^2. Aggregate data (as opposed to microdata, which is data about an individual) is not addressed in the proposal.

The core proposal in the legislation is a voluntary *data agreement*, which is a contract between a data discloser and a data recipient. The PDDA will only apply to those who choose to accept its terms and penalties through a data agreement. The PDDA establishes standards for behavior and civil and criminal penalties for violations. In exchange, there are benefits to the discloser and recipient.

The proposal would not require all data disclosers and data users to comply with its requirements. Only those who voluntarily choose to reference the PDDA in a contract or equivalent document would be subject to the requirements. A mandatory solution may not be practical. There appears to be no way to write a definition that would encompass all data transfers, and there are too many data transfers to expect that one size will fit all.

A voluntary approach allows those who want the benefits to accept the obligations. A model for this approach to legislation is arbitration. Laws define, support, and provide for the enforcement of arbitration agreements, but it is typical for the parties to a contract themselves to decide whether they want to make use of arbitration at all. If they do not, then arbitration laws do not affect their activities.

The discloser who shares data under a data agreement proceeds with the knowledge that the recipient has accepted strict limits on data use and disclosure that are enforceable by the state and by data subjects. The discloser need not accept any obligation to police the agreement or to act on behalf of data subjects, other than to report on breaches of the agreement to a government agency. The main benefit to the discloser is that liability is capped or eliminated by the law. Further, the proposal includes a formal safe harbor, which exempts a discloser from liability for disclosure under a data agreement if (1) the recipient is a government agency, non-profit organization, or research organization that has not reported a breach of a data agreement in the five years prior to date of the agreement, and (2) the disclosure is for use in research ("systematic investigation designed to develop or contribute to generalizable knowledge, but does not include marketing research") or in a public health activity. The purpose of the safe harbor is to provide encouragement for data sharing for beneficial purposes.

The recipient who seeks data under a data agreement is in a better position to ask for data from a discloser because the data agreement and the law impose on the recipient defined and enforceable limits that protect the privacy of data subjects. A reluctant source might be encouraged by a would-be recipient to share data because of the existence of formal standards and limited liability. The recipient accepts the limits and the liability as a condition of receiving data.

The data subject benefits from disclosure under a data agreement because of strong rules prohibiting conduct that could reidentify data. The data subject also benefits because the law clarifies that the data subject is a third party beneficiary of the data agreement. That enables an aggrieved data subject to seek damages from a negligent party to the data agreement. Under current law, a data subject may be unable to sue relying upon an ordinary contract between a data discloser and a data recipient because the data subject is not a party to the contract. The data subject lacks privity – an adequate legal relationship – to the contract and cannot use the contract to enforce his or her interest. Today, only in some jurisdictions will a data subject be recognized as a third party beneficiary of a data use agreement and able to seek damages. In general, however, the requirement for privity can be a major obstacle to enforcement of privacy rights by data subjects.[28] The proposed law would clarify this issue in favor of data subjects.

Most of the obligations fall on the data recipient, which is appropriate because the recipient obtains new data vulnerable to reidentification. The recipient agrees to: 1) not reidentify or attempt to reidentify data received under the data use agreement; 2) take reasonable steps to prevent any related party from reidentifying or attempting to reidentify information received under a data agreement; 3) not further use or disclose any potentially identifiable personal information received under a data agreement except in accordance with the data agreement; 4) only disclose potentially identifiable personal information received under the data agreement to another person if the disclosure is allowed by the data agreement and if the disclosure is made pursuant to a data agreement subject to the Act; 5) maintain reasonable physical, administrative, and technical safeguards to protect against reidentification of potentially identifiable personal information received under a data agreement; and 6) inform a potential discloser in writing before entering into any data agreement of any actual or reasonably likely breaches of other data agreements that the recipient entered into during the past ten years. The last requirement provides a self-policing mechanism that obliges bad actors to tell others before they can obtain new data.

The fourth requirement – that data received under a data use agreement only be redisclosed under the original data agreement or under a new data agreement subject to the Act means that there must be a chain of trust if data is further disclosed. Data subject to a data use agreement will always be subject to the mandated protections. If allowed by the original data use agreement, the data recipient can become a data discloser with respect to the next recipient, and the protections continue in force because a new data use agreement is required.

---

[28] Joel Reidenberg, *The Privacy Obstacle Course: Hurdling Barriers to Transnational Financial Services*, 60 Fordham Law Review S137, S175 (1992).

Both the recipient and the discloser must: (1) report any breach of a data agreement that the recipient entered into to a national consumer protection/privacy agency and to each other; (2) publish a notice of the breach prominently on their respective public websites; and (3) maintain the website notice for two years. In addition, in the event that either party learns that potentially identifiable personal information under their data agreement has been reidentified, that party must comply with applicable security breach notification laws. The proposed law does not impose a security breach notification obligation of its own. It references existing obligations. It can be anticipated that parties to a data agreement will allocate responsibility for compliance with breach notification laws among themselves in some suitable manner.

The PDDA includes several carefully tiered criminal penalties for violations. The penalties range from civil penalties for failure to report or failure to post, to felonies for knowing and willful reidentification or attempted reidentification, to major felonies with the possibility of prison for knowing and willful reidentification or attempted reidentification with the intent to sell, transfer, or use personal information for commercial advantage, personal gain, or malicious harm. There is also a felony for disclosing PI^2 obtained under a data agreement subject to the Act in violation of the terms of the agreement.

Civil remedies are available to an individual whose PI^2 has been reidentified against a discloser or recipient who is negligently responsible for the reidentification. The PDDA specifically provides that a data subject is a third party beneficiary of a data agreement so that there will be no issue about a lack of standing to sue over the contract.

Other provisions require an appropriate government agency with oversight responsibilities for the Act to file a biennial report, review the law in five years, and prepare model data agreements. Another provision makes it clear that the PDDA does not change, override, or preempt any requirement or obligation established by other laws, and does not exempts anyone from complying with obligations under any law or rule for the protection of human research subjects. Finally, the proposed legislation has been drafted in a manner that is not directly tied to U.S. law. The same approach might have value in other jurisdictions.

Conclusion

The proposed PDDA seeks to strike a balance between the need to share deidentified personal information for research and other purposes and the inability to guarantee that the information is wholly deidentified. The solution is to allow data disclosers and data recipients to enter into a voluntary data agreement that defines the obligations of the parties, provides greater certainty about the potential liabilities, and allows individual data subjects to enforce their privacy interests when data has been reidentified. In order to support appropriate sharing, the legislation includes a safe harbor for a data discloser who shares data for a beneficial purpose.

Today's lack of clear definitions, deidentification procedures, and legal certainty can impede some useful data sharing. It can also affect privacy of users when the lack of clarity about deidentification results in sharing of identifiable data that could have been avoided

The proposed approach to the deidentification dilemma faced by data processors and policy makers will not solve every problem associated with personal data transfers and uses, but it will make available a new tool that fairly balances the needs and interests of data disclosers, data users, and data subjects. The solution could be invoked voluntarily by data disclosers and data recipients. Its use could also be mandated by regulation or legislation seeking to allow broader use of personal data for beneficial purposes.

*****

A BILL

To protect the privacy of potentially identifiable personal information by establishing accountability for the use and transfer of potentially identifiable personal information. [Version 4.4]

SECTION 1. SHORT TITLE.

This Act may be cited as the "Personal Data Deidentification Act".

SEC. 2. DEFINITIONS.

As used in this Act:

(1) DATA AGREEMENT.—The term "data agreement" means a contract, memorandum of understanding, data use agreement, or similar agreement between a discloser and a recipient relating to the use of personal information.

(2) DATA AGREEMENT SUBJECT TO THIS ACT.—The term "data agreement subject to this Act'' means a data agreement between a discloser and a recipient who have entered into an agreement described in section 3(a).

(3) DISCLOSER.—The term "discloser" means a person, who discloses potentially identifiable personal information to another person pursuant to a data agreement subject to this Act.

(4) OVERT IDENTIFIER.—The term "overt identifier" means any personal information that identifies or can readily be used to identify a particular individual, and includes a name, address, Social Security Number, account number, license number, serial number, telephone number, electronic mail address, Internet protocol address, webpage

address, or biometric, that alone or in combination with other information identifies or can readily be used to identify a particular individual.

(5) PERSON.—The term "person" means an individual, corporation, company, foundation, association, society, partnership, firm, non-profit organization, school, college, or university, or a department, agency, or other instrumentality of [Federal, State, or local] government.

(6) PERSONAL INFORMATION.—The term "personal information" means information about an individual that may or may not include an overt identifier.

(7) POTENTIALLY IDENTIFIABLE PERSONAL INFORMATION.—The term "potentially identifiable personal information" means any personal information without any overt identifiers.

(8) PUBLIC WEBSITE.—The term "public website" means a facility by which a person displays information to the general public on the Internet or any comparable successor technology.

(9) RECIPIENT.—The term "recipient" means a person who receives potentially identifiable personal information from another person pursuant to a data agreement subject to this Act.

(10) RESEARCH.—The term "research" means a systematic investigation designed to develop or contribute to generalizable knowledge, but does not include marketing research.

(10) REIDENTIFICATION.—The term "reidentification" means linking potentially identifiable personal information to an overt identifier belonging or assigned to any living or dead individual.

SEC. 3. DATA AGREEMENTS.

(a) AGREEMENTS SUBJECT TO ACT.—A person who enters into a data agreement that expressly references this Act by including the words "This data agreement is subject to the Personal Information Deidentification Procedures Act" or equivalent words, or who is required to be subject to this Act by statute or regulation for any disclosure or receipt of potentially identifiable personal information

(1) shall be bound by, and subject to, all of the terms of this Act with respect to potentially identifiable personal information disclosed or received under that data agreement, statute, or regulation; and

(2) may not terminate, revoke, suspend, or otherwise limit or restrict the application of this Act to potentially identifiable personal information disclosed or received under that data agreement, statute, or regulation.

(b) ADDITIONAL TERMS PERMITTED.—The parties to a data agreement subject to this Act may include additional terms to that data agreement that do not limit or undermine the terms required by this Act.

SEC. 4. DUTIES OF RECIPIENT.

A recipient under a data agreement subject to this Act shall—

(1) not reidentify or attempt to reidentify any potentially identifiable personal information received under that data agreement;

(2) take reasonable steps, including contracts, technical measures, or workplace rules, to prevent any employee, agent, consultant, contractor, affiliate, subcontractor, or other related party from reidentifying or making an attempt to reidentify any potentially identifiable personal information that the recipient received under that data agreement;

(3) not further use or disclose any potentially identifiable personal information received under the data agreement except in accordance with that data agreement;

(4) only disclose potentially identifiable personal information received under that data agreement to another person if the disclosure is allowed by that data agreement and if the disclosure is made pursuant to that data agreement or another data agreement subject to this Act;

(5) maintain reasonable physical, administrative, and technical safeguards to protect against reidentification of potentially identifiable personal information received under that data agreement;

(6) inform a potential discloser in writing before entering into any data agreement that will be a data agreement subject to this Act with the potential discloser of any actual or reasonably likely breaches of other data

agreements subject to this Act that the recipient entered into during the past 10 years;

(7) (A) promptly report any breach of a data agreement subject to this Act that the recipient entered into to—

(i) the [National Consumer Protection/Privacy Agency]; and

(ii) the discloser;

(B) promptly publish a notice of the breach prominently on the recipient's public website; and

(C) maintain the notice for two years; and

(8) in the event that the recipient learns that potentially identifiable personal information that the recipient obtained under that data agreement has been reidentified, comply with applicable [Federal or State] security breach notification laws.

SEC. 5. DUTIES OF DISCLOSER.

A discloser under a data agreement subject to this Act shall—

(1) (A) promptly report any breach of that data agreement to the [National Consumer Protection/Privacy Agency];

(B) promptly publish a notice of the breach prominently on the discloser's public website; and

(C) maintain the notice for two years;

(2) in the event that the discloser learns that any potentially identifiable personal information that the discloser disclosed under that data agreement has been reidentified, comply with applicable [Federal or State] security breach notification laws; and

(3) in the event that the discloser learns that any potentially identifiable personal information disclosed under that data agreement has been reidentified or may have been reidentified, immediately suspend further disclosures of potentially identifiable personal information to the recipient under the data agreement.

SEC. 5. SAFE HARBOR.

A discloser who lawfully discloses potentially identifiable personal information under a data agreement subject to this Act—

(1) to a recipient who is a government agency, non-profit organization, or research organization that has not reported a breach of a data agreement in the five years prior to date of the agreement,

(2) for use in research or in a public health activity,

shall not be liable under this Act or any other law to an individual who is the subject of potentially identifiable personal information disclosed pursuant to that data agreement for any damage resulting from that disclosure, except in the case of gross negligence on the part of the discloser.

SEC. 7. PENALTIES.

(a) CIVIL PENALTIES.—A person who fails to report a breach of a data agreement subject to this Act, or to post a notice in accordance with this Act, shall be subject to a civil penalty of not more than [$2,000] in an action brought by the [National Consumer Protection/Privacy Agency][Attorney General].

(b) FELONY OFFENSES.—

(1) A recipient, or any employee, agent, consultant, contractor, affiliate, subcontractor, or other related party of a recipient, who willfully discloses potentially identifiable personal information received under a data agreement subject to this Act in violation of this Act is guilty of a felony and shall be fined not more than [ ] or imprisoned not more than [ ] years, or both.

(2) A recipient, or any employee, agent, consultant, contractor, affiliate, subcontractor, or other related party of a recipient, who willfully reidentifies or attempts to reidentify potentially identifiable personal information received under a data agreement subject to this Act in violation of this Act is guilty of a felony and shall be fined not more than [ ] or imprisoned not more than [ ] years, or both.

(3) A recipient, or any employee, agent, consultant, contractor, affiliate, subcontractor, or other related party of a recipient, who willfully

reidentifies or attempts to reidentify potentially identifiable personal information received under a data agreement subject to this Act in violation of this Act with the intent to sell, transfer, or use personal information obtained under a data agreement subject to this Act for commercial advantage, personal gain, or malicious is guilty of a felony and shall be fined not more than [ ] or imprisoned not more than [ ] years, or both.

(c) FAILURE TO INFORM OFFENSES.—A person who fails to inform a potential discloser in writing before entering into any data agreement subject to this Act as required by section 4(8) is guilty of a misdemeanor and shall be fined not more than [ ].

SEC. 8. CIVIL REMEDIES.

(a) BENEFICIARIES OF THE AGREEMENT.—An individual who is the subject of potentially identifiable personal information that is disclosed pursuant to a data agreement subject to this Act shall be a third party beneficiary of that data agreement.

(b) LIABILITY FOR NEGLIGENCE OF DISCLOSERS AND RECIPIENTS.—If a discloser or recipient of potentially identifiable personal information pursuant to a data agreement subject to this Act fails to exercise reasonable care to prevent the reidentification of an individual who is the subject of the information, that individual may bring a civil action against the discloser or recipient if the individual suffers monetary harm, emotional harm, reputational harm, or public embarrassment as a result of such reidentification. Any individual entitled to damages under this subsection shall recover not less than $1000, and the court may award reasonable attorney fees and other reasonable litigation costs to an individual who substantially prevails.

SEC. 9. DUTIES OF [NATIONAL CONSUMER PROTECTION/PRIVACY AGENCY].

The [National Consumer Protection/Privacy Agency] shall—

(1) make a biennial report summarizing any activities under this Act [to the national legislature] and post the report on its public website;

(2) evaluate the operations of the Act and report to [the national legislature] within five years after the date of enactment of this Act; and

(3) within six months of the date of enactment, publish one or more model data agreements.

SEC. 10. OTHER LAWS.

Nothing in this Act changes, overrides, or preempts any requirement or obligation established by any other law. Nothing in this Act exempts any person from complying with obligations under any applicable law or rule for the protection of human research subjects.

#####