TAKING THE PERSONAL OUT OF DATA: MAKING SENSE OF DE-IDENTIFICATION

YIANNI LAGOS*

INTRODUCTION

Data is powerful but scary. Many consumer services rely on data aggregation.[1] A navigation system, for example, uses geolocation data to help consumers circumvent rush hour traffic.[2] This is a useful service. The aggregation of geolocation data creates a privacy concern as companies have access to each and every place a person visits.[3] This is, at the very least, unsettling.

De-identification provides a solution. It is a process to prevent a personal identifier from being connected with information.[4] A car owner's name is an example of a personal identifier. The speed a car is going on a crowded highway is an example of information. De-identification involves deleting or masking direct identifiers, such as the car owner's name, and suppressing or generalizing indirect identifiers, such as the location of a person's home or work.[5] With de-identification, consumers get real-time traffic delay information while their privacy is protected.[6] This is a potential win-win-win for consumers, for privacy, and for businesses. It allows companies to analyze data to provide consumer services, it protects privacy by breaking the connection between the analytical information and personal identifiers, and it gives businesses increased flexibility to innovate by discovering novel uses of data.[7]

Despite the increased relevance of de-identification, there is not a universal

---

* Yianni Lagos is an attorney for Lagos & Lagos P.L.L. He is a former Legal and Policy Fellow at Future of Privacy Forum.

1. *See*, *e.g.*, Sean Madden, *How Companies Like Amazon Use Big Data to Make You Love Them*, FAST CO. (May 2, 2012, 8:30 AM), http://www.fastcodesign.com/1669551/how-companies-like-amazon-use-big-data-to-make-you-love-them.

2. Phillip Swarts, *Is Your Car Spying on You? GPS Tracks 'Consumers,' Identity Theft at Risk*, WASH. TIMES (Jan. 7, 2014), http://www.washingtontimes.com/news/2014/jan/7/no-privacy-behind-the-wheel-your-car-might-be-spyi/?page=all.

3. Many services seemingly without need to track location are doing so and selling that data to third-party advertisers. *See* Charles Arthur, *Android Torch App With Over 50m Downloads Silently Sent User Location and Device Data to Advertisers*, GUARDIAN (Dec. 6, 2013, 3:00 PM), http://www.theguardian.com/technology/2013/dec/06/android-app-50m-downloads-sent-data-advertisers.

4. *New Words & Slang*, MERRIAM-WEBSTER (June 15, 2007, 11:08 PM), http://nws.merriam-webster.com/opendictionary/newword_display_alpha.php?letter=De.

5. *Id.*

6. ANN CAVOUKIAN & KHALED EL EMAM, DISPELLING THE MYTHS SURROUNDING DE-IDENTIFICATION: ANONYMIZATION REMAINS A STRONG TOOL FOR PROTECTING PRIVACY 1 (2011), *available at* http://www.ipc.on.ca/images/Resources/anonymization.pdf.

7. *Id.* at 4-5.

definition. Policymakers are currently debating when data is sufficiently stripped of identifying information to be considered de-identified.[8] Some examples are obvious. Data is not de-identified if it contains a person's name.[9] Similarly, data is not de-identified if it contains a revealing email address, john.smith@gmail.com, or a phone number listed in the phonebook.[10] These are examples of direct identifiers.

The more difficult cases arise when only indirect identifiers are present. Data may not be de-identified even if it does not contain a direct identifier.[11] For example, if john.smith@gmail.com is replaced with a random number or pseudonym (such as 578294@gmail.com), data is not de-identified if indirect identifiers can re-associate the information to John Smith (re-identification).[12] Common indirect identifiers are date of birth, gender, and location.[13] Though it is not obvious that having location information would lead to identifying a living person, people tend to be located at two places most of the week—home and work. A public records search of a person's home and work could potentially lead to identifying the individual.[14] The starting and ending destinations should be generalized (taking a street address and turning it into a zip code) before the dataset becomes de-identified.[15]

---

8. *De-Identification*, FUTURE OF PRIVACY FORUM, http://www.futureofprivacy.org/de-identification/ (last visited Sept. 13, 2014) (discussing the debate over the definition of personal information); *see also* FED. TRADE COMM'N, PROTECTING CONSUMER PRIVACY IN AN ERA OF RAPID CHANGE: RECOMMENDATIONS FOR BUSINESSES AND POLICYMAKERS, at iv (2012), *available at* http://www.ftc.gov/sites/default/files/documents/reports/federal-trade-commission-report-protecting-consumer-privacy-era-rapid-change-recommendations/120326privacyreport.pdf (explaining the criteria for data to be successfully de-identified).

9. *See De-Identification*, *supra* note 8 (explaining when data is de-identified); *see also* FED. TRADE COMM'N, *supra* note 8.

10. *See* FED. TRADE COMM'N, *supra* note 8.

11. *Id.* at 33.

12. *Id.*

13. *See* Daniel C. Barth-Jones, *The "Re-identification" of Governor William Weld's Medical Information: A Critical Re-examination of Health Data Identification Risk and Privacy Protections, Then and Now* 5 (Privacy Ass'n, Working Paper, July 24, 2012) (*available at* https://www.privacyassociation.org/media/pdf/knowledge_center/Re-Identification_of_Welds_Medical_Information.pdf) (study found that twenty-nine percent of a population bore risk of *plausible* re-identification with three data points (full date of birth, gender, and five-digit ZIP code), though risk of actual re-identification was much lower given that the data set was incomplete.).

14. Browser tracking data associated with what is commonly referred to as a cookie similarly contains information that can be used to identify individuals easily, i.e., a username typed into a webpage. *See* JONATHAN MAYER, THIRD-PARTY WEB TRACKING: POLICY AND TECHNOLOGY, 2012 IEEE SYMPOSIUM ON SECURITY AND PRIVACY 415-16 (2012), *available at* http://www.academia.edu/2784919/Third-Party_Web_Tracking_Policy_and_Technology.

15. This Article uses the example of geolocation information, but many combinations of information pose privacy concerns. Credit card information combined with zip code, for example,

Indirect identifiers create a major problem with defining de-identification. Indirect identifiers, such as location, provide useful analytical information but also create a potential link back to an individual.[16] De-identified data must be specific enough for data to still be useful, but broad enough so it cannot be associated with an individual. This balancing is a difficult task that this Article explores.

Before defining data as sufficiently de-identified, this Article urges balance between protecting consumer privacy and ensuring companies can continue to innovate. Part I of this Article stresses the importance of ensuring any definition of de-identification includes adequate privacy benefits. Data still poses a risk to privacy unless it is sufficiently scrubbed of identifying information.[17] If the definition of "de-identified" is too lenient, it would create the false impression that data was now safe. This would be unfair to consumers. More importantly, this could undermine the trust necessary for a vibrant data-driven economy.

Part II of this Article looks at the privacy preserving aspects of data that do not rise to the level of "de-identified." There is a wide gap between de-identified information and information directly tied to a person's name. Information that can only be tied to a person through extensive research on where people live and work, for instance, does not pose the same privacy concerns as a credit card number tied directly to a person's name. That gap should be filled with an intermediate level of data that should be appropriately called "intermediate data." Intermediate data is information that is not easily linkable to a particular individual but is tied to a unique identifier. Intermediate data should not be confused with de-identified data, but there are still privacy protecting aspects to intermediate data.

Part III recognizes that any definition of de-identification should minimize the negative effects on innovation. Data is becoming a more integral part of our economy.[18] Many of the services we rely on in our daily lives, from GPS to social networking, cannot function without collecting data.[19] Just as importantly, the profit driver of the internet relies on information and advertisements to provide free services.[20] These companies are in a position of trust and have a constant pipeline of new information on consumers.[21] That trust is a prerequisite to innovation. Without trusting companies with data, the data driven economy

---

could lead to identification of an individual who purchases a very unique product.

16. *See* CAVOUKIAN & EMAM, *supra* note 6, at 11.

17. *See De-Identification*, *supra* note 8.

18. *Innovation & Data Use*, FUTURE OF PRIVACY FORUM, http://www.futureofprivacy.org/issues/innovation-data-use/ (last visited Sept. 13, 2014).

19. Adam Thierer, *Relax and Learn to Love Big Data*, U.S. NEWS & WORLD REPORT (Sept. 16, 2013, 12:10 PM), http://www.usnews.com/opinion/blogs/economic-intelligence/2013/09/16/big-data-collection-has-many-benefits-for-internet-users.

20. Quentin Hardy, *Troubles Ahead for Internet Advertising*, N.Y. TIMES (Aug. 29, 2013, 2:29 PM), http://bits.blogs.nytimes.com/2013/08/29/troubles-ahead-for-internet-advertising/?_php=true&_type=blogs&_r=0.

21. *Id.*

would suffer a significant setback.[22]

Part IV recommends balancing the privacy preserving aspects of de-identification with incentivizing companies to scrub data. Due to the power of technology companies, significant de-identification legislation is currently unlikely in the United States.[23] Even if passed, any statutory or regulatory definition of de-identification would almost assuredly be vague, as no specific definition of de-identification has been created. Today, it falls on companies to self-regulate. Companies will simply forgo de-identifying data if the definition of de-identification is too stringent, thus depriving consumers of a potentially powerful privacy protection.

I. BENEFITS OF DE-IDENTIFICATION

The privacy protecting benefits of de-identification depend on its definition. Yet there is no universal standard for when data has been scrubbed enough to be considered de-identified.[24] Any definition needs to live up to the name and provide true separation between a person's identity and his information.

Datasets are too varied for a simple definition. Those variations include the sensitivity of the data, the administrative safeguards used to protect the data, and the parties sharing the information.[25] Intimate medical details, for example, are more sensitive than preferences for shopping at Talbots or TJ Maxx.[26] Similarly, data released to the public at large creates more privacy concerns than data kept within a company.[27] All the variations of data need to be taken into account before defining the level of technical separation between peoples' identities and their information needed to call data de-identified.

One end of the spectrum, perfect de-identification, is not practical. If information has zero chance of being technically associated with a person or group of persons, there is no privacy risk in a dataset.[28] A useful dataset can never have zero chance of reconnecting a person to his information.[29] No statute

---

22. *Id.* (explaining how companies rely on data for advertising purposes).

23. Though the Federal Trade Commission has recently "called on Congress to protect consumers against the unchecked collection and sharing of their digital data," there is not an imminent chance of legislation getting through Congress. Steve Lohr, *New Curbs Sought on the Personal Data Industry*, N.Y. TIMES, May 28, 2014, at B1.

24. *See De-Identification*, *supra* note 8.

25. *See* CAVOUKIAN & EMAM, *supra* note 6, at 14.

26. *Id.* at 4 (explaining the sensitivity of health information).

27. INFO. COMM'R'S OFFICE ANONYMISATION: MANAGING DATA PROTECTION RISK CODE OF PRACTICE 6-9 (2012), *available at* http://ico.org.uk/for_organisations/guidance_index/~/media/documents/library/Data_Protection/Practical_application/anonymisation-codev2.pdf.

28. Joseph Jerome, *Making Perfect De-Identification the Enemy of Good De-Identification*, FUTURE OF PRIVACY FORUM, http://www.futureofprivacy.org/2014/06/19/making-perfect-de-identification-the-enemy-of-good-de-identification/ (last visited Sept. 13, 2014).

29. KHALED EL EMAM, GUIDE TO THE DE-IDENTIFICATION OF PERSONAL HEALTH INFORMATION 135 (2013).

or regulation should require the impossible standard of perfect unlinkability.

The other end of the spectrum, the ability to identify 100% of individuals with their information, is not de-identification, even if other administrative safeguards are in place to protect the data. Administrative safeguards protect data from being misused without technically altering the data itself, and include limiting access controls to trusted employees and providing cybersecurity measures to prevent data breaches from hackers.[30] These safeguards alone can never be enough to count as de-identification. A common industry practice is to hash a person's name to create a random unique identifier (taking John Smith and transforming it to 578294).[31] Many times a company retains the algorithm (commonly referred to as a key) to continue to transfer information associated with John Smith to the unique identifier 578294 but restricts access to the key to a limited number of employees.[32] The problem with retaining the key is that bad acting employees with access to the key technically can re-associate the information to John Smith.[33] Similarly, threats from government requests or outside bad actors are still significant when all of the individuals in a database could be identified.[34]

If administrative safeguards alone justified calling data de-identified that could potentially harm the data-driven economy.[35] Without trust, internet users may start withholding their personal information and refrain from online purchases, both essential ingredients to the expansion of the internet economy.[36] Companies would undermine consumer trust if they claimed data was de-identified that could in fact be easily re-associated with the individuals.[37] Misleading is not the answer.

Administrative controls, however, can provide important protections when used in addition to technical measures.[38] In the example above, if the key is

---

30. This Article combines the administrative and physical safeguards referred to in the Privacy Act of 1974 into one category: administrative safeguards. Privacy Act of 1974, 5 U.S.C. § 552a(e)(10) (2011).

31. Ed Felten, *Does Hashing Make Data "Anonymous"?*, TECH. AT FED. TRADE COMM'N (Apr. 22, 2012, 7:05 AM), http://techatftc.wordpress.com/2012/04/22/does-hashing-make-data-anonymous/ (explaining how hashing alone can lead to re-identification of an individual).

32. Edith Ramirez, Remarks at the Media Institute in Washington, D.C., at 7-8 (May 8, 2014) (transcript available at http://www.ftc.gov/system/files/documents/public_statements/308421/140508mediainstitute.pdf) (describing Target's use of algorithms).

33. *See* Felten, *supra* note 31.

34. *Id.* (explaining how hashing often fails).

35. Administrative safeguards provide a vital role in protecting consumer data and creating trust in the data driven economy. Those safeguards, however, should not justify calling data de-identified when data can be easily associated with an individual.

36. Ardion Beldad et al., *How Shall I Trust the Faceless and the Intangible? A Literature Review on the Antecedents of Online Trust*, 26 COMPUTERS IN HUM. BEHAV. 857, 859 (2010).

37. *See* FED. TRADE COMM'N, *supra* note 8, at 8-9.

38. "DeID-AT" is a short hand form of describing the combination of administrative safeguards and technical de-identification. *See* Yianni Lagos & Jules Polonesky, *Public vs. Non-*

destroyed, there is no direct way to re-identify John Smith.[39] Individuals may nonetheless be re-identified through the use of indirect identifiers.[40] The re-identification of Massachusetts Governor Weld's medical records using full date of birth, zip code, and gender is an example of using indirect identifiers to reconnect personal information with a person's identity.[41] That re-identification was from a publically released dataset.[42] If administrative safeguards were used to protect the data from the public, it is likely the data would have never been re-identified.

With non-public datasets protected by strong administrative measures, the ability to re-identify a small number of individuals poses much less of a privacy concern. With administrative controls, only a very limited number of individuals in the company or a skillful hacker who broke the controls could attempt to re-identify the dataset.[43] The reported examples of re-identification required the work of computer scientists who could only successfully identify a fraction of individuals in a public database.[44] The time and expertise needed to re-identify datasets is likely a barrier to bad actors. It is likely not worth a criminal's time. The easier it is to reconnect individuals with data, the more likely bad actors will hack a company database and attempt to re-identify individuals to their information.[45]

A major concern with administrative safeguards does not come from companies or bad actors but from the government. The National Security Administration (NSA) scandal raised the concern that company data will fall into the hands of the government with unknown consequences.[46] In theory the government could always request the information, but the threat of a government request is significantly reduced through the use of de-identification. It is likely not worth the government's time. Government requests are much more likely when 100% of a database is identifiable than when only one percent of a database could potentially be re-identified after significant effort.[47]

The benefits of administrative controls are dependent on the quality of those controls. Currently, companies have not been forthcoming with their different

---

Public Data: The Benefits of Administrative Controls, 66 STAN. L. REV. ONLINE 103, 104 (2014).

39. *See* Felten, *supra* note 31.

40. *See* Latanya Sweeney, *k-Anonymity: A Model for Protecting Privacy*, 10 (5) INT'L J. ON UNCERTAINTY, FUZZINESS & KNOWLEDGE-BASED SYSTEMS 557, 559 (2002).

41. *Id.* at 560.

42. Barth-Jones, *supra* note 13.

43. Restricting data to only trusted parties reduces privacy risk. *See* Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701, 1771 (2010).

44. Barth-Jones, *supra* note 13.

45. Felten, *supra* note 31.

46. Lisa Mascaro, *House Overwhelmingly Supports Bill to Curb NSA Domestic Spying*, LA TIMES (May 22, 2014, 7:57 PM), http://www.latimes.com/nation/politics/la-na-nsa-reforms-20140523-story.html#page=1 NSA legislation.

47. *Id*.

administrative techniques. Keeping the confidentiality of administrative safeguards does enhance their protections, but it is difficult to judge the efficacy of those programs without some disclosure. Additionally, the privacy preserving protections of administrative controls are lessened as more information is shared with outside parties (See Appendix A).[48] For datasets available to the public or released to a large number of individuals, administrative controls provide less of a benefit.[49] The benefits of technical de-identification, however, protect even data released to the public.[50]

Due to the protections of administrative controls, there should be a lesser requirement to remove indirect identifiers for internal databases than for public databases. A widely-used means of measuring the risk of indirect identifiers is k-anonymity.[51] K-anonymity measures re-identification risk by the number of individuals in a dataset with matching indirect identifiers.[52] If k equals three, then three individuals share all the same attributes in the dataset.[53] An example would be three people with the same birthday.[54] If k equals twenty, then twenty individuals share common attributes—a re-identification risk lower than when k equals three.[55] K-anonymity measures maximum risk by only considering the individuals with the smallest number of matching indirect identifiers or the biggest outlier (think the 100-year-old, 6'10", red head).[56] Public databases should have high k-values.[57] Current cell size precedents for public databases range from a k of three to a k of twenty.[58] With non-public databases, lower k-values should be acceptable.[59]

---

48. FED. TRADE COMM'N, *supra* note 8, at 33 (explaining companies should limit the amount of data shared with third parties to better protect privacy of consumers).

49. Even with publically available data, obscurity, or the difficulty in finding data, could still protect consumer privacy to a certain degree. *See* Woodrow Hartzhog & Evan Selinger, *Obscurity: A Better Way to Think about Your Data than Privacy*, ATLANTIC MAG. (Jan. 17, 2013), http://www.theatlantic.com/technology/archive/2013/01/obscurity-a-better-way-to-think-about-your-data-than-privacy/267283/.

50. *See* CAVOUKIAN & EMAM, *supra* note 6, at 4-5 (explaining the benefits of de-identification).

51. *See generally* Sweeney, *supra* note 40.

52. *Id.* at 5.

53. *See generally id.*

54. *See generally id.*

55. *See generally id.*

56. *See generally id.*

57. EMAM, *supra* note 29, at 279.

58. *Id.*

59. Deciding the exact level of k-anonymity needed involves looking at a number of factors that could include: administrative safeguards, sensitivity of the data, public or private data, the number of parties sharing the data, whether there is consumer choice, the purpose of using the data, and other factors. *See* Pierangela Samarati & Latanya Sweeney, *Protecting Privacy When Disclosing Information: k-anonymity and Its Enforcement Through Generalization and Suppression* 2-3, *available at* epic.org/privacy/reidentification/Samarti_Sweeney_Paper.pdf.

Average k-anonymity is another option for non-public databases. Instead of measuring only the individuals with the maximum risk (or lowest k), an average would take the mean risk of the entire database (or average k). For public databases, maximum risk is appropriate because many bad actors will likely focus exclusively on the easiest individuals to re-identify. That assumption may not hold true for non-public databases. Using average k-anonymity would give a more accurate measure of the risk to the entire database, while allowing companies to increase data utility. Using average k-anonymity does not necessarily mean that companies should allow for unique individuals in a dataset, or k values equal to one. Instead, companies should take into account both maximum k and average k when measuring risk.

II. INTERMEDIATE DATA

Data that does not rise to the level of de-identified still may have privacy preserving aspects. A data breach involving a person's name and credit card information, such as with the 2014 Target breach, creates significant danger of theft or other malfeasance.[60] A simple step of replacing a person's name with a random pseudonym could significantly reduce the harm from such a data breach.

Instead of generating a creative name for this intermediate level of data, the use of "intermediate identifiers" or "intermediate data" seems most descriptive. The most commonly used word to describe the intermediate category between fully identifiable and de-identified is "pseudonymized."[61] This word is fraught with confusion. An email address, for example, may be called a pseudonym, but "john.smith@gmail.com" does little to protect privacy.[62] Thus, a pseudonym alone has little to no privacy protection.[63] Intermediate data deserving of an intermediate category of privacy protection may be tied to a unique identifier

---

60. Though the Target breach has been reported as a point-of-sale breach, the high number of reported identify theft cases showcases the danger of combining personal identifiers with sensitive information, such as credit card data. *See Data Breach FAQ*, TARGET, https://corporate.target.com/about/shopping-experience/payment-card-issue-FAQ (last visited July 31, 2014).

61. MEP Jan Philipp Albrecht, European Union Committee on Civil Liberties, Justice, and Home Affairs, released a Draft Report on the General Data Protection Regulation that recognized such an intermediate category of data: "the rapporteur encourages the pseudonymous . . . use of services. For the use of pseudonymous data, there could be alleviations with regard to obligations for the data controller." JAN PHILIPP ALBRECHT, COMMISSION PROPOSAL FOR A REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL ON THE PROTECTION OF INDIVIDUALS WITH REGARD TO THE PROCESSING OF PERSONAL DATA AND ON THE FREE MOVEMENT OF SUCH DATA (GENERAL DATA PROTECTION REGULATION) 211 (2012), *available at* http://www.europarl.europa.eu/meetdocs/2009_2014/documents/libe/pr/922/922387/922387en.pdf.

62. Holding browsing tracking data in what is commonly referred to as a cookie identifier is another commonly cited example of using a pseudonym that does not tie to a particular individual. A cookie identifier, could potentially become used broadly enough to be indistinguishable from other common numerical identifiers such as a social security number.

63. *See* Felten, *supra* note 31 (explaining how hashing works).

(such as a random number) but must not be easily linkable to a particular individual (such as an email address).[64]

The previously discussed example of hashing an identifier with a key should be considered intermediate data.[65] In that example the name, John Smith, is transformed into a random unique identifier, 578294. The unique identifier 578294 is not easily identifiable to John Smith for parties without access to the key.[66] If the key is protected by sufficient administrative controls, the data may be considered intermediate data.[67] A hacker would need to gain access to both the dataset and the key.[68] That double layer of privacy protection provides a barrier in the case of a data breach, but the data is not yet de-identified because the random identifier could be converted back to the name John Smith by anyone with access to the key.[69]

A dataset must also go through the additional scrubbing to remove obvious indirect identifiers before becoming intermediate data. Indirect identifiers, such as date of birth and location, can lead to identifying a significant number of individuals in a dataset.[70] Though intermediate data does not need to be scrubbed to the same extent of de-identified data, obvious indirect identifiers, like a person's home address, should be removed or generalized.

It is also important that organizations cannot use an intermediate identifier to discriminate against an individual. If a pseudonym can still be used to reach an individual, it should not be considered intermediate data. Mobile phones often transmit information with a common number identifier (mobile ID).[71] A mobile ID is a pseudonym just as the random number 578294 could be characterized as a pseudonym.[72] The difference is the mobile ID can be used to discriminate against the phone owner.[73] Companies could discriminate against a mobile ID by charging a higher price for a mobile shopper based on where the person lives by tracking their mobile phone. In that scenario, it does not matter whether a company attaches a mobile ID to a person's name. The consumer is harmed regardless. Thus, when a pseudonym can be used to discriminate against an individual, that data should not be considered intermediate data and only minor

---

64. *Id.*

65. *Id.*

66. *Id.*

67. *Id.*

68. *Id.*

69. *Id.*

70. *See* CAVOUKIAN & EMAM, *supra* note 6, at 11 (describing quasi-identifiers).

71. Jennifer Valentino-DeVries, *Unique Phone ID Numbers Explained*, WALL ST. J. (Dec. 19, 2010, 9:40 PM), http://blogs.wsj.com/digits/2010/12/19/unique-phone-id-numbers-explained/.

72. *Id.*

73. *Value of Data*, FUTURE OF PRIVACY FORUM, http://www.futureofprivacy.org/issues/innovation-data-use/value-of-data/ (last visited Sept. 14, 2014) (explaining privacy advocates' concern that data will be used to discriminate against certain individuals); *see also* Omer Tene & Jules Polonetsky, *Privacy in the Age of Big Data: A Time for Big Decisions*, 64 STAN. L. REV. 63, 65 (2012).

alleviation on obligations are warranted.[74]

Only when a pseudonym is not directly tied to a person or their device should fewer restrictions apply to the uses of such information. The previously used example of hashing an identifier to create a random unique identifier 578294 is an example of a unique identifier that cannot be used to affect an individual. If companies restrict access to the key, the concern of discrimination is greatly reduced.[75] The consumer can no longer be reached by the identifier. Companies should have increased freedom to use intermediate data.[76]

The fact that data is not considered intermediate or de-identified data does not mean that companies should never use that information. Outright restrictions on data collection are rarely appropriate. Instead, increased consumer notice and control or use limitations are the appropriate responses.[77] There is currently a debate about whether to give consumers the ability to easily opt-out of broad scale collection of information, or whether companies should just be prohibited from using data for certain purposes.[78] Limitations on the uses of data, instead of collection, have the benefit of protecting consumers while allowing for non-harmful uses of the data.[79] Companies, however, have failed to provide comprehensive use restrictions that would give the public confidence that data would not be misused.

III. INNOVATION AND TRUST

Before defining the boundaries of de-identification and intermediate data, the effects on innovation should be considered in addition to the privacy implications.

---

74. The EU's General Data Protection Regulation currently has clauses allowing for a right to access and a right to data portability. ALBRECHT, *supra* note 61, at 53. The clauses allow consumers to see all data a company has about them and then transfer that information to their computer or to another company. *Id.* When companies hold data using only a number identifier, it is difficult for companies to verify the authenticity of such a request. *Id.* The data security concern of preventing identify theft outweigh the benefit to consumers of accessing information. Thus, policymakers should not allow a right to access or a right to data portability with data associated only with a mobile ID.

75. *See* Ramirez, *supra*, note 32.

76. The exact type of increased freedom should be decided on a case by case basis. Since consumers can no longer be reached directly by the data, companies should be able to use this information freely for research purposes if accompanied by a promise to not re-identify with the key. Companies should also be given increased freedom to share this data without also sharing the key.

77. *See generally* FED. TRADE COMM'N, *supra* note 8 (proposing changes for how consumers' data is handled).

78. *See* Wendy Davis, *Web Standards Group Moves Forward with Do-Not-Track Effort*, ONLINE MEDIA DAILY (Apr. 25, 2014, 5:11 PM), http://www.mediapost.com/publications/article/224423/web-standards-group-moves-forward-with-do-not-trac.html (discussing how some advertising groups preferred targeting use limitations to collection limitations).

79. *See generally* FED. TRADE COMM'N, *supra* note 8.

Innovation benefits consumers and businesses alike.[80] The large aggregation of data does lead to groundbreaking discoveries.[81] Studying the genetic code of large portions of the population could lead to breakthroughs in medicine.[82] Monitoring student performance could lead to the immediate recognition when a student is getting behind. On the other hand, organizations with access to a person's genetic code or a child's academic history create privacy concerns.[83] Finding the right balance recognizes both the potential for discovery and the privacy risk.

Less glamorous, but still vital to the economy, is corporate profit. Data leads to more profitable companies.[84] Facebook and Twitter exist because of data sharing.[85] The increasing use of data is leading to profitable companies that are providing services that consumers use every day.[86]

Many of these innovations are not possible without the aggregation of data.[87] The internet functions by assigning each user a unique IP address that is transmitted to every website visited.[88] Websites cannot function at the basic level of providing content to an individual computer without a minimum level of data collection.[89] Society needs to trust companies to protect and use data in appropriate ways to get these innovations.

With an endless stream of information from consumers, companies will always have the ability to exploit consumer information in inappropriate ways. The most stringent form of privacy protection would require companies to delete all information after this initial collection. This stringent requirement hurts consumers.[90] It deprives them of services without a corresponding benefit to privacy.[91]

The promise to protect data is similar to the promise to delete data. In both instances, trust is essential. Even if a company promised to delete all data

---

80. Johnathan Shaw, *Why "Big Data" Is a Big Deal*, HARV. MAG. (Mar. 2014), http://harvardmagazine.com/2014/03/why-big-data-is-a-big-deal.

81. *Id.*

82. *Id.* (explaining that data can be used for innovation in medicine).

83. *Id.* (describing the privacy concerns with uses of data).

84. Lin Jing & Chen Yingqun, *When Big Data Can Lead to Big Profit*, CHINADAILY (Apr. 21, 2014, 6:56 PM), http://www.chinadaily.com.cn/business/2014-04/21/content_17449249.htm.

85. *See* Shaw, *supra* note 80 (explaining that social media gathers significant amounts of data).

86. Jing & Yingqun, *supra* note 84.

87. Shaw, *supra* note 80.

88. Russ Smith, *IP Address: Your Internet Identity*, CONSUMER.NET (Mar. 29, 1997), http://www.ntia.doc.gov/legacy/ntiahome/privacy/files/smith.htm.

89. *Id.*

90. Society could change the driving laws to restrict cars from going above 5 mph. There would be a great reduction in car fatalities, but we as a society are moving faster than that.

91. Tim McGuire et al., *Why Big Data Is the New Competitive Advantage*, IVEY BUS. J. (July 2012), http://iveybusinessjournal.com/topics/strategy/why-big-data-is-the-new-competitive-advantage#.VBU31vldWAg (explaining the benefits of big data to businesses and consumers).

collected from consumers, it could simply change policies the next day and proceed to retain massive amounts of data. There is no privacy protection that does not involve some level of trust in companies.[92] Legal enforcement could similarly ensure companies follow through on promises to protect information just as easily as promises to delete information.[93]

Trust plays a crucial role in de-identification. Many debates on de-identification have centered on whether administrative safeguards should be used in conjunction with the technical transformation of the dataset.[94] Trust is essential for both controls.[95] Back to the key example, if the key is maintained, it is possible for the company to re-identify the information despite the administrative control.[96] Similarly, even if a company promises to destroy the key, the company could just retain the data, despite the supposed technical control. With both administrative and technical safeguards, companies can protect data or abuse it depending on their motivation.[97]

A natural reaction to the NSA's broad tracking practices is to restrict companies from collecting and retaining information, not just because we are afraid of what businesses are doing with the data, but because we are afraid government is going to get their hands on the information. Such a knee-jerk reaction could negatively impact the progress and innovation of society just as data aggregation is forming the center of many industries. There are valid reasons to restrict companies from collecting and using data, but companies should not be punished for the indiscretions of the government.

IV. BALANCING PRIVACY RISK WITH ADOPTION RATES

An often overlooked criterion in privacy protection is whether companies will actually protect data. The 2014 data breaches of eBay and Target show just how vulnerable consumer data is to potential hackers.[98] Those breaches involved both

---

92. EXECUTIVE OFFICE OF THE PRESIDENT, BIG DATA: SEIZING OPPORTUNITIES, PRESERVING VALUES 10-11 (2014) [hereinafter BIG DATA], *available at* http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_5.1.14_final_print.pdf (explaining how trust is essential to utilizing big data within government).

93. *See generally* FED. TRADE COMM'N, *supra* note 8.

94. "MITIGATING CONTROLS WORK IN CONJUNCTION WITH DE-ID TECHNIQUES TO MINIMIZE THE RE-ID RISK." HEALTH SYS. USE TECHNICAL ADVISORY COMM. DATA DE-IDENTIFICATION WORKING GRP., 'BEST PRACTICE' GUIDELINES FOR MANAGING THE DISCLOSURE OF DE-IDENTIFIED HEALTH INFORMATION 19 (2010), *available at* http://www.ehealthinformation.ca/documents/Data%20De-identification%20Best%20Practice%20Guidelines.pdf.

95. *See* BIG DATA, *supra* note 92.

96. *See* Felten, *supra* note 31 (explaining how hashing works).

97. Criminal sanctions could provide an incentive not to abuse data, but those criminal sanctions would presumably be just as effective in ensuring administrative safeguards as technical safeguards.

98. *See* Dave Johnson, *eBay Data Breach: What You Need to do Now*, CBS NEWS (May 25, 2014, 8:44 AM), http://www.cbsnews.com/news/massive-data-breach-at-ebay-change-your-

personal identifiers (name or username) and sensitive information (credit card numbers and passwords).[99] If that data was de-identified, the privacy invasions from the breaches would have been much less severe.[100]

Data categories fall on a spectrum from fully identified to intermediate data to strict de-identification, with many categories in between.[101] Strict de-identification provides the lowest risk of reconnecting individuals with their data, but not necessarily the greatest protection for consumers.[102] Companies may find that the privacy benefits of strict de-identification are outweighed by the large loss data utility, and thus decide to hold data in fully identified form.[103] Given the lack of general legislation on de-identification, consumers cannot benefit from the added privacy protection if companies refuse to de-identify data.[104]

Loose de-identification standards, conversely, may promote company adoption but provide little additional protections to consumers. Companies probably desire to call data de-identified in an attempt to tout their consumer-protective policies. If de-identification does not actually protect consumers, then such claims would create the false impression that companies are adequately protecting consumer privacy.

Any definition of de-identification must balance the added privacy protections of reducing re-identification risk with the lost privacy protection from companies refusing to scrub data altogether (See Appendix B).[105] The rate that companies scrub data is likely a function of the benefits and costs.

Companies benefit in two main ways from scrubbing non-public databases. First, regulations and self-imposed promises restrict how companies use and share personal data.[106] If companies scrub data to meet those standards, companies can

---

password-now/; *see also* Samantha Sharf, *Target Shares Tumble as Retailer Reveals Cost of Data Breach*, FORBES (Aug. 5, 2014, 9:16 AM), http://www.forbes.com/sites/samanthasharf/2014/08/05/target-shares-tumble-as-retailer-reveals-cost-of-data-breach/.

99. *See* Johnson, *supra* note 98; *see also* Sharf, *supra* note 98.

100. A wronged customer of Target has the civil remedy of negligence, but simple consumer protections such as de-identification are a preferred solution. It seems unjust to force a consumer to affirmatively sue a mega-company for damages as the only real remedy a consumer has available to him.

101. *See* Lagos & Polonesky, *supra* note 38.

102. *Id.*

103. EMAM, *supra* note 29, at 6.

104. The likelihood of general legislation in de-identification is low. Thus, companies are left to self-regulate with the help of advocacy groups and regulators. The fact that some fields, such as health care, currently require de-identification actually furthers the need for guidance on de-identification. The current definitions of de-identification are too broad to give companies any real guidance, and standards for healthcare data may not be appropriate in other areas. *See generally* FED. TRADE COMM'N, *supra* note 8.

105. *See De-Identification*, *supra* note 8.

106. The Federal Trade Commission currently enforces statements made in company privacy policies under Section 5(B) of the Federal Trade Commission Act. Federal Trade Commission Act, 15 U.S.C. § 45(a)(1) (2006).

use data for more purposes—in many cases without consent.[107] Companies can also more freely share datasets with business partners to gain further insights from the data.[108] Second, companies benefit from the lower risk of a data breach.[109] Data breaches have a large reputational cost, and reporting requirements may not apply to breaches of de-identified and potentially intermediate databases.[110]

De-identification, however, has costs. Companies weigh these benefits with the costs before scrubbing data. De-identification reduces data utility.[111] Data scrubbing techniques include data masking, suppression, and generalization—all of which reduce the statistical power of a dataset.[112] That is a loss for consumers and businesses alike.[113] Consumers lose out on novel new products and services specifically targeted to their interests, and companies lose out on profits.[114] Companies must also expend time and resources to scrub data that increase with stricter de-identification standards.[115] With data utility losses and implementation costs, companies may forgo any scrubbing under a strict standard because the costs outweigh the benefits.[116] Creating a reasonable definition of de-identification that companies will utilize should be the goal.

CONCLUSION

Defining de-identification is no easy task. Companies, regulators, and advocacy groups should adopt a pragmatic approach to de-identification. Any definition of de-identification should take into account the quality of the technical protections, the effects on innovation, and whether companies will actually use the tool. Intermediate data also has a vital role to play in protecting consumer privacy. A reasonable and clear definition of de-identification that companies choose to implement will go a long way in protecting consumer privacy.

---

107. In the European Union, for example, de-identification allows for public disclosure of data without violating individual privacy. Council Directive 95/46, 1995 O.J. (L 281) 26 (EC).

108. As the Federal Trade Commission recently advised, those business partners should be contractually bound not to re-identify. FED. TRADE COMM'N, *supra* note 8, at 21.

109. *See* CAVOUKIAN & EMAM, *supra* note 6, at 4-5.

110. *Data Breach FAQ*, *supra* note 60.

111. *See* CAVOUKIAN & EMAM, *supra* note 6, at 12.

112. EMAM, *supra* note 29, at 6.

113. Shaw, *supra* note 80.

114. *Id.*

115. *Id.*

116. *Id.*

APPENDIX A[117]

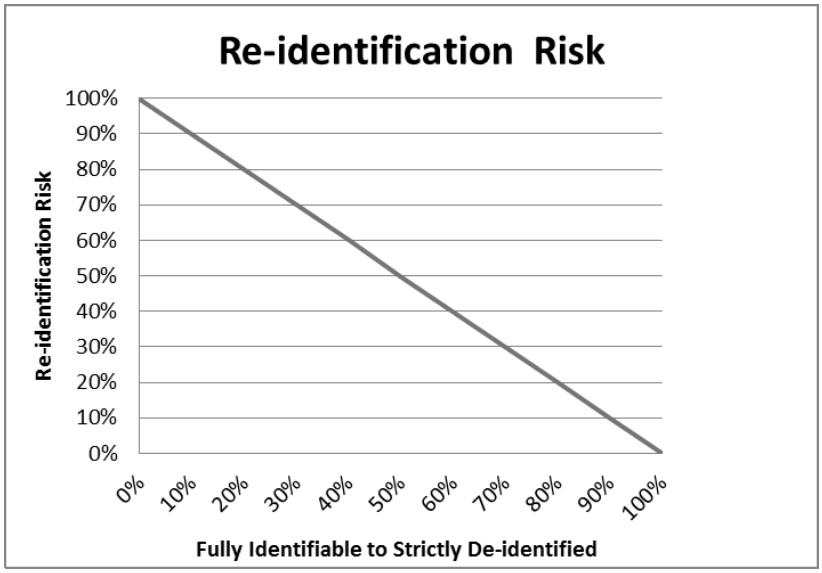
The benefits of using technical and administrative controls stem from the added protection of two independent events. Independence requires that the chance of a bad actor breaching administrative safeguards (administrative risk) is not correlated with the chance of a bad actor re-identifying data (technical risk). If administrative risk is independent from technical risk, then combining technical and administrative controls drastically reduces total privacy risk. As shown in the table below, if the probability of an administrative breach is one percent and the probability of a technical breach is one percent, the probability of a privacy breach drops to .01%. The benefits of this dual protection, however, decrease when companies share data with other business partners, as there is an independent chance of an administrative data breach within each company. With a relatively few number of companies exposed to the data, total privacy risk remains small, but when a company shares data with a hundred partners, the added benefits of combining administrative and technical controls can decrease or almost disappear. At a thousand companies, with the assumed one percent chance of breach, the data should be considered public. These statistics show large benefits of combining technical and administrative safeguards when a company confines data to a few trusted partners, but reduced benefits when the number of companies with access to the data increases.

# of Companies	**Technical Risk**	**Administrative Risk**	**Total Privacy Risk**
1	1%	1%	0.01%
2	1%	2%	0.02%
3	1%	3%	0.03%
4	1%	4%	0.04%
5	1%	5%	0.05%
6	1%	6%	0.06%
7	1%	7%	0.07%
8	1%	8%	0.08%
9	1%	9%	0.09%
10	1%	10%	0.10%
100	1%	64%	0.64%
1000	1%	100%	1.00%

117. Analysis and calculations were completed by the Author.

APPENDIX B[118]

Conventional thinking suggests that stricter de-identification standards best protect consumer privacy. The illustrative graph below shows that regulators can choose along a spectrum of the de-identification standard. The x-axis is the de-identification standard imposed, with 100% being the strictest standard. The y-axis is the re-identification risk. The line shows the re-identification risk for a given de-identification standard. When the de-identification standard equals zero (or fully identified data), the re-identification risk is 100%. As the de-identification standard increases, the re-identification risk decreases linearly. Re-identification risk, however, is only one component of overall consumer privacy protection.

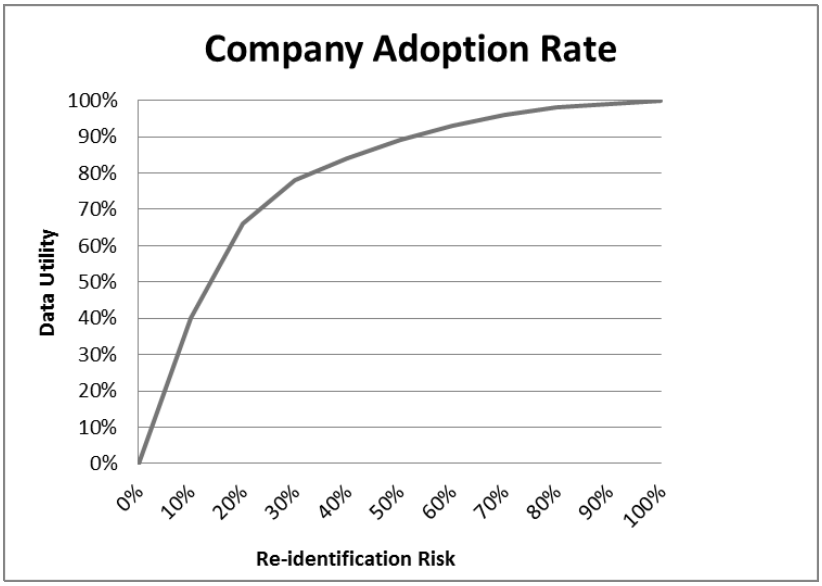


The privacy benefits provided by de-identification are meaningless unless companies de-identify public data. The graph below shows the relationship between re-identification risk and company adoption rates. The x-axis is re-identification risk. A strict de-identification standard is equivalent to a low re-identification risk. The y-axis is data utility. Stricter de-identification requirements reduce data utility. The line shows the percent of companies adopting De-ID AT for a given re-identification risk. Loss of data utility is a disincentive to companies, so when re-identification risk decreases, less companies adopt de-identification as a data protection tool.

Non-empirical thinking suggests that the number of companies de-identifying

118. Analysis and calculations were completed by the Author.

data does not follow a linear path. At a re-identification risk of zero, no companies will de-identify data because a re-identification rate of zero means the data has no utility. A small increase in re-identification risk from zero will have a large impact on data utility and company adoption because companies will add back the most useful data first (the steep part of the curve). At high risks of re-identification, a large increase in re-identification risk will have a small impact on data utility and company adoption because companies have already added back the most useful data (the flat part of the curve). In other words, when de-identifying data, companies remove the lowest value data first (i.e., low hanging fruit). Companies can therefore achieve reasonable de-identification with relatively little loss in data utility. If regulators require companies to reach strict de-identification standards, companies lose a relatively high amount of data utility.



The slope of the graph is dependent on the assumption that companies can efficiently de-identify data by using techniques that initially reduce the least useful data. For companies to efficiently de-identify, regulation needs to give room for companies to choose the appropriate techniques. The HIPAA statistical approach allows that flexibility. Under that approach, companies can de-identify data in any form, as long as a statistician certifies that the risk of re-identification is very small using accepted statistical and scientific principles and methods. That flexibility allows companies to de-identify data, while preserving data utility.