Washington Journal of Law, Technology & Arts

Volume 5 | Issue 1 Article 3

9-23-2008

De-identified Data and Third Party Data Mining: The Risk of Re-identification of Personal Information

C. Christine Porter

Cite as: C. Christine Porter, *De-Identified Data and Third Party Data Mining: The Risk of Re-Identification of Personal Information*, 5 Shidler J.L. Com. & Tech. 3 (Sep. 23, 2008), *at* http://www.lctjournal.washington.edu/Vol5/a03Porter.html.

DE-IDENTIFIED DATA AND THIRD PARTY DATA MINING: THE RISK OF RE-IDENTIFICATION OF PERSONAL INFORMATION

C. Christine Porter[1]

Abstract

Recent computer science research demonstrates that anonymized data can sometimes be easily re-identified with particular individuals, despite companies' attempts to isolate personal information. Netflix and AOL are two examples of companies that released personal data intended to be anonymous but which was re-identified with individual users with the use of very small amounts of auxiliary data. Re-identification of anonymized data may expose companies to increased liability, as the information may no longer be treated as anonymous. In addition, companies may violate their own privacy policies by releasing anonymous information to third parties that can be easily re-identified with individual users. The potential for third parties to re-identify anonymous information with its individual source indicates the need for both increased privacy protection of anonymized information and increased security for databases containing anonymized information.

Table of Contents

Introduction
The Law on Anonymized Data
The Risk of Re-Identification of "Anonymous Data"
Consequences of Violating Privacy Policies
Conclusion

INTRODUCTION

<1>In 2006, Netflix published customer movie-rankings data that it anonymized by replacing names with random numbers and removing personal details.[2] This data came from rankings customers assigned to movies while logged into their personal accounts. Two researchers at the University of Texas were able to de-anonymize some of the Netflix data by comparing it with non-anonymous users' movie ratings posted by those users in the Internet Movie Database ("IMDb").[3] The researchers discovered that very little information about a Netflix subscriber was needed in order to identify that subscriber in the anonymous database.[4] Given a user's public IMDb movie ratings, the researchers were able to uncover *all* of the users' private movie ratings entered into the Netflix system.[5] That researchers successfully re-identified a portion of the anonymized data with individual Netflix consumers shows the potential security problems with anonymous data.

<2>It has long been assumed that anonymous consumer data does not need the same protections as data that can be identified with a particular customer. Companies sometimes sell this ostensibly de-identified information to third-party data miners, even when the information is particularly sensitive, such as financial and health care information. Companies with explicit privacy policies (such as Netflix), health care providers such as pharmacies, and financial institutions like credit card companies, may release data after it has been de-identified.

<3>This article explores the potential legal problems arising from the increasingly strong possibility that anonymized information may be re-identified with a particular individual using only a small set of auxiliary information that is publicly available. Companies which fail to exercise reasonable precautions to protect sensitive information may violate financial confidentiality laws or medical privacy laws. Even with less sensitive consumer information, the loss of consumer privacy could lead to greater company liability, particularly when companies violate their own privacy policies by releasing information that is easily re-identified. In addition, security breaches such as the one experienced by AOL in 2006, in which it accidentally released anonymous information that was easily re-identified with individuals, lead to negative publicity that can be costly.[6]

THE LAW ON ANONYMIZED DATA

<4>U.S. law has no general right of information privacy parallel to the 1995 Data Protection Privilege that exists under European Union ("EU") law.[7] While the U.S. has no over-arching privacy law, there are some privacy protections for particular categories of information under existing statutes. For example, the Fair Credit Reporting Act ("FCRA") protects the privacy of a person's financial information under some circumstances.[8] The Gramm-Leach-Bliley Act ("GLBA") protects some kinds of financial data.[9] Medical data is protected by the Health Insurance Portability and Accountability Act ("HIPAA"),[10] and children's data by the Children's Online Privacy Protection Act ("COPPA").[11]

<5>These existing privacy regulations do not typically protect information that has been modified so that the data subject cannot be identified.[12] The use of de-identified financial information, for example, is permitted by the Federal Trade Commission ("FTC") as long as it is aggregated. However, if the information can be re-identified by joining it with auxiliary data, it would seem to be still in the realm of sensitive information.[13] To some extent, privacy concerns in the U.S. still reflect the assumption that if confidentiality is breached, it will be primarily through *deliberate* releases of personally identifiable information.[14]

<6>But even information usually given greater privacy protection may not be statutorily protected from use by third parties if it is anonymized. For example, the GLBA requires financial institutions to provide consumers with the opportunity to opt out of having their nonpublic personal information shared with third parties,[15] and HIPAA expressly prohibits such sales or any sharing of patient information outside of a covered entity without express authorization from the patient.[16] But de-identified data is treated differently. The GLBA's corresponding regulations state that if information is "aggregate information or blind data that does not contain personal identifiers such as account numbers, names, or addresses," it should not be considered personally identifiable information, and is not regulated by the statute.[17] If this information can be easily re-identified with consumers, should the data be treated as de-identified data? What happens when the data does become re-identified?

<7>Privacy concerns are arguably stronger with medical records and consumer data from pharmacies than with other kinds of consumer data. HIPAA protects the privacy of all personally identifiable health information.[18] However, the corresponding regulations state that covered entities can release such information to third parties if it is properly de-identified.[19] Pharmacies commonly sell this de-identified information to data mining companies, who in turn sell it to pharmaceutical companies. But recent concerns about the possible re-identification of this data have prompted the enactment of state legislation to ban this data mining of medical information; however, federal district courts in Maine and New Hampshire have struck down recently-enacted privacy laws on First Amendment grounds.[20] A federal court in one recent case called any concern about patient privacy "illusive," refusing to recognize a significant risk of re-identification.[21]

<8>Recent research underscores the risk that third parties could join "anonymized" data with a small amount of auxiliary data from another database and de-anonymize the data. In 2007, two researchers in the computer science department at the University of Texas published a paper entitled "How to Break Anonymity of the Netflix Prize Database."[22] Netflix, the largest online DVD rental service, publicly released a database with movie rankings in connection with a contest. The names and other personal details were removed from the rankings; yet the Texas researchers were able to re-identify this information with very little auxiliary information. While the theory behind the ability to break into the database is difficult for a lay person to follow, once the steps required to break the anonymity of the database are disclosed, a high level of technical knowledge is not needed to attain access to the potentially sensitive information contained in the database.[23] In addition, current tests used by companies to determine if their anonymous databases can withstand such adversarial attacks may not be sufficient.[24]

Perhaps the ability of third parties to discover information about an individual's movie rankings is not too disturbing, as movie rankings are not generally considered to be sensitive information. But because these same techniques can lead to the re-identification of data, far greater privacy concerns are implicated. Even as far back as 1997, a researcher was able to de-anonymize medical records by joining them with a publicly-available voter database.[25] Anecdotal evidence suggests algorithms already exist that can re-identify patient information with prescription drug information after third party data mining companies ostensibly de-identify the information.

<9>Sometimes technical expertise is not even needed for a third party to de-anonymize data. As researchers have recently pointed out, re-identification is easier when dealing with a population that has a unique combination of identifiers.[26] After AOL accidentally published users' searches in 2006, reporters for the *New York Times* were able to take groups of searches made by anonymized individual users on AOL and re-identify an individual simply from the searches she made.[27] This individual, Thelma Arnold, confirmed to the newspaper that she had made these searches. The *New York Times* article also stated that bloggers were able to identify other individuals from the searches.[28]

<10>In August 2006, the Electronic Frontier Foundation ("EFF")[29] filed a complaint with the Federal Trade Commission against AOL. [30] The complaint accused AOL of violating the Federal Trade Commission Act[31] by intentionally or recklessly disclosing Internet search histories of more than half a million AOL users in March to May 2006.[32] Section 5(a) of the Federal Trade Commission Act prohibits deceptive acts or practices affecting commerce.[33] EFF's complaint alleged that by falsely leading consumers to believe AOL would protect consumer privacy, AOL violated Section 5(a) of the Federal Trade Commission Act.

<11>In its complaint, the EFF made detailed allegations about the sensitive information from Internet searches AOL published. The data disclosed by AOL was publicly available as a downloadable file for ten days before AOL removed it.[34] The disclosure made public such sensitive search queries as "how to tell your family you're a victim of incest," "how to kill your wife," "will I be extradited from NY to FL on a dui charge," and "my baby's father physically abuses me."[35] AOL included a warning and disclaimer with this information, illustrating its awareness of the sensitive nature of the information.[36] The EFF reviewed this information and found many examples of search histories that could personally identify a particular AOL subscriber or household. Some of these search histories contain personally identifiable information such as addresses, birth dates, and Social Security numbers.[37]

<12>A particularly worrisome problem with these types of security breaches is that once an individual's privacy is breached by re-identification, future privacy breaches become easier. "In general, once any piece of data has been linked to a person's *real* identity,

any association between this data and an anonymous *virtual* identity breaks anonymity of the latter."[38] If a Netflix subscriber's rankings were re-identified, for example, then that person can never again disclose any information about her movie viewing, because it can then be traced back to her real identity using the Netflix Prize dataset.[39]

CONSEQUENCES OF VIOLATING PRIVACY POLICIES

<13>Re-identification of anonymized data with individual consumers may expose companies to increased liability. If data is re-identified, this may be due to the failure of companies to take reasonable precautions to protect consumer data. In addition, companies may violate their own privacy policies by releasing anonymous information to third parties that can be easily re-identified with individual users. As discussed below, the FTC has made examples out of several companies for not properly protecting personal data.

<14>In 2005, the FTC filed a complaint against ChoicePoint, after the third-party data broker's failure to take reasonable precautions to protect financial data resulted in numerous instances of identify theft. ChoicePoint admitted that a problem with its screening procedures allowed a group of criminals to access the personal financial information of thousands of people, in violation of federal consumer protection law.[40] The FTC filed a complaint in January 2006, alleging, among other violations, that ChoicePoint "has not employed reasonable and appropriate measures to secure the personal information it collects for sale to its subscribers."[41] The FTC announced a large civil penalty for ChoicePoint: $10 million to the Commission, and $5 million to redress consumer harms.[42] These penalties were assessed because ChoicePoint violated federal consumer protection laws by failing to maintain reasonable procedures to protect personal financial data and by making false and misleading statements about its privacy policies.[43]

<15>Most companies are now aware of the need to adhere to their online privacy policies and the potential consequences for non-compliance. In 1998, the FTC made a public example out of GeoCities, which was at that time a "virtual community" that hosted members' Web pages and provided other services to 1.8 million members.[44] GeoCities' Web site included statements assuring members that their personal information would not be shared without their permission.[45] However, GeoCities sold the information to third parties, who used it for targeted advertising.[46] The case settled with a consent order prohibiting GeoCities from misleading consumers about its data-collection practices.[47] This action was a message to companies that they could not deceive consumers by posting a privacy policy and then ignoring it.

<16>Companies that violate their own privacy policies may face liability beyond the possibility of an FTC action.[48] State courts may adopt HIPAA Privacy Rules as minimum standards of care for breach of confidentiality under state common law.[49] If the information released would be highly offensive or humiliating to a reasonable person and is widely disclosed, the company may be liable for the tort of invasion of privacy by public disclosure of private facts.[50] Additionally, if the privacy policy is viewed as a contract, the consumer may have the ability to bring an action for contract damages.[51] The policy might be viewed as an offer, for which the user's use of the site or submission of information is an acceptance, with either being sufficient consideration to support the finding of a contractual obligation.[52]

<17>At this point, online consumers may expect to see a privacy policy and companies may not want to violate these policies for fear of losing consumer trust. Companies have learned to avoid problems by releasing data to third parties that has been detached from personal identifiers. However, if that information can be easily re-identified with those persons, and companies release the information to third parties even thought they are aware that such a re-identification is a significant possibility, companies may be liable in contract or tort. The lack of case law at this point makes it difficult to predict

what claims a court would be willing to uphold.[53] Although cases are occasionally brought outside of the health care context, parties have so far been able to settle their claims.

<18>The danger of re-identification of personal information often goes un-acknowledged by companies. On the Netflix Prize webpage, Netflix referred to its own privacy policy and stated that since a person probably could not even identify his or her *own* data, publication of the rankings created no privacy concern.[54] This failure to recognize the ability of third parties to re-identify data is not unique to Netflix; in fact, it is almost standard in privacy policies to assert that anonymous information collected by the company cannot be linked to personal data by the third parties receiving the anonymous information.[55]

<19>Even companies who intend to abide by the promises conveyed in their privacy policies sometimes have trouble keeping those promises when they run into financial difficulties. For example, Toysmart, an Internet-only retailer, had a very strict privacy policy.[56] However, when Toysmart had to file for bankruptcy, its previously confidential customer information became another asset that needed to be put up for sale.[57] Retailers like Toysmart that are experiencing financial difficulties may need to sell their anonymous databases of consumer information along with the personally identifiable information.[58] Even if a company selling an anonymous database has a strict privacy policy, it is possible that the next company will not adhere as strictly to the privacy policies of the preceding company. This is even more of a concern with health care and financial information.

CONCLUSION

<20>The increasing ability of third parties to re-identify anonymous information with its individual source indicates the need for both increased privacy protection of anonymized information and increased security for databases containing anonymized information. Anonymous systems should be subjected to adversarial attacks to test their ability to withstand such attacks; however, if the researchers who broke the anonymity of the Netflix database are correct, even those tests may not be enough to ensure that anonymous information cannot become re-identified with individual consumers.[59]

[<< Top]

Footnotes

1. C. Christine Porter, University of Washington School of Law, J.D. program Class of 2009; University of Washington, M.A., 2003. Thank you to Professor Jane K. Winn of the University of Washington School of Law, Peter Winn, lecturer at UW Law School, and Jeffrey Bashaw and N. Elizabeth Mills, student editors, for their invaluable advice and comments on drafts of this Article.

2. Bruce Schneier, *Why Anonymous Data Sometimes Isn't*, WIRED, Dec. 13, 2007, *available at* http://www.wired.com/politics/security/commentary/securitymatters/2007/12/securitymatters_1 .

3. Arvind Narayanan & Vitaly Shmatikov, *Robust De-anonymization of Large Sparse Datasets*, PROC. OF THE 29TH IEEE SYMPOSIUM ON SECURITY AND PRIVACY, May 2008, *available at* http://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf (published previously as *How to Break the Anonymity of the Netflix Prize Database*, March 2, 2007, *available at* http://arxiv.org/PS_cache/cs/pdf/0610/0610105v1.pdf).

4. *Id.*

5. Narayanan & Shmatikov, *supra* note 3.

6. Michael Barbaro & Tom Zeller, Jr., *A Face is Exposed for AOL Searcher No. 4417749*, N.Y. Times, Aug. 9, 2006, at A1, *available at* http://www.nytimes.com/2006/08/09/technology/09aol.html?ex=1312776000. *See* discussion *infra* ¶¶ 11-13.

7. Council Directive 95/46, On the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data, 1995 O.J. (L 281) (EC), *available at* http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:EN:NOT. *See, e.g.*, Julia M. Fromholz, *The European Union Data Privacy Directive*, 15 Berkeley Tech. L.J. 461, 472 (2000).

8. 15 U.S.C. § 1681-1681x (2008).

9. Gramm-Leach-Bliley Act of 1999, Pub. L. No. 106-102, 113 Stat. 1338 (codified as amended in scattered sections of 12 and 15 U.S.C. (2008)).

10. Health Insurance Portability and Accountability Act of 1996, Pub. L. No. 104-191, 110 Stat. 1936 (codified as amended in scattered sections of 18, 26, 29, and 42 U.S.C. (2008)).

11. 15 U.S.C. §§ 6501-6506 (2008).

12. Benjamin Charkow, *The Control over the De-Identification of Data*, 21 Cardozo Arts & Ent. L.J. 195, 196 (2003).

13. See Jane K. Winn, *Can a Duty of Information Security Become Special Protection for Sensitive Data Under US Laws?* (Social Sci. Research Network, Working Paper), *available at* http://ssrn.com/abstract=1265775.

14. Douglas J. Sylvester & Sharon Lohr, *The Security of Our Secrets: A History of Privacy and Confidentiality in Law and Statistical Practice*, 83 Denv. U.L. Rev. 147, 187 (2005).

15. 15 U.S.C. § 6802(b) (2008).

16. 45 C.F.R. § 164.508 (2008).

17. 16 C.F.R. § 313.3(o)(2)(ii)(B) (2008).

18. 29 U.S.C. § 1181.

19. 45 C.F.R. § 164.502(d) (2008).

20. In IMS Health Inc. v. Ayotte, 490 F. Supp. 2d 163 (D.N.H. 2007), the district court in New Hampshire struck down the state's law banning the use of de-identified health data, ruling that the Act unlawfully limited commercial free speech. The case is currently before First Circuit Court of Appeals. In IMS Health Corp. v. Rowe, 532 F. Supp. 2d 153 (D. Me. 2007), a district court enjoined Maine's new law on the same basis.

21. *IMS Health Corp.*, 532 F. Supp. 2d at 162. For a discussion of why the concern about re-identification of patient information is well-founded, see Brief for Electronic Privacy Information Center as Amicus Curiae Supporting Defendant-Appellant, IMS Health v. Ayotte, 490 F. Supp. 2d 163 (2007) (No. 06-cv-280-PB).

22. Narayanan & Shmatikov, *supra* note 3.

23. *Id.*

24. Schneier, *supra* note 2.

25. Latanya Sweeney, *Weaving Technology and Public Policy Together to Maintain Confidentiality*, 25 J.L. Med. & Ethics 98, 110 (1997).

26. See Brief for Electronic Privacy Information Center as Amicus Curiae Supporting Defendant-Appellant, *IMS Health v. Ayotte*, 490 F. Supp. 2d 163

(Aug. 14, 2007) (No. 06-cv-280-PB), for a discussion of the relative ease of re-identification for a unique population, such as the sole female in a male-dominated working environment.

27. Barbaro & Zeller, *supra* note 6.

28. *Id*.

29. According to its Web site, the Electronic Frontier Foundation is a 501(c)(3) organization founded in 1990 to protect civil liberties in the digital age. *See* Electronic Frontier Foundation, http://www.eff.org (last visited Aug. 31, 2008).

30. Complaint filed with the FTC by the Electronic Frontier Foundation, In the Matter of AOL LCC, Aug. 14, 2006, *available at* http://w2.eff.org/Privacy/AOL/aol_ftc_complaint_final.pdf.

31. 15 U.S.C §§ 41-58 (2008).

32. Complaint at 1, filed with the FTC by the Electronic Frontier Foundation, *supra* note 32.

33. 15 U.S.C. § 45(a)(1).

34. Complaint at 4, filed with the FTC by the Electronic Frontier Foundation, *supra* note 32.

35. *Id.* at 4-5.

36. *Id.* at 5.

37. *Id.* at 4-5.

38. Narayanan & Shmatikov, *supra* note 3.

39. *Id.*

40. For a summary of the ChoicePoint breach, see G. Martin Bingisser, *Data Privacy and Breach Reporting: Compliance with Various State Laws*, 4 SHIDLER J.L. COM. & TECH. 9 (2008), at http://www.lctjournal.washington.edu/Vol4/a09Bingisser.html.

41. Complaint, United States v. ChoicePoint, No. 06-CV-0198 (N.D. Ga. Jan. 30, 2006), *available at* http://www.ftc.gov/os/caselist/choicepoint/0523069complaint.pdf.

42. Press Release, FTC, ChoicePoint Settles Data Security Breach Charges; to Pay $10 Million in Civil Penalties, $5 Million for Consumer Redress, FTC File No. 052-3069 (Jan. 16, 2006), *available at* http://www.ftc.gov/opa/2006/01/choicepoint.shtm.

43. *Id. See also* Anita Ramasastry, *Whose Credit Report Is It, Anyway? It's Time for States to Pass Credit Freeze Laws That Give Consumers Control over Their Credit Profiles*, FINDLAW, Feb. 6, 2006, http://writ.news.findlaw.com/ramasastry/20060206.html.

44. Scott Killingsworth, *Minding Your Own Business: Privacy Policies in Principle and in Practice*, 7 J. INTELL. PROP. L. 57, 60 (1999).

45. *Id.*

46. *Id.*

47. Press Release, FTC, Internet Site Agrees to Settle FTC Charges of Deceptively Collecting Personal Information in Agency's First Internet Privacy Case (Aug. 13, 1998), *available at* http://www.ftc.gov/opa/1998/08/geocitie.shtm. The FTC does not have authority to require Web sites to post privacy policies or to prescribe their

content, but under Section 5 of the FTC Act, it has broad enforcement power over "deceptive acts or practices." See Killingsworth, *supra* note 44, at 61.

48. There is no private right of action under the FTC Act. *See* Peter A. Winn, *Confidentiality in Cyberspace: The HIPAA Privacy Rules and the Common Law*, 33 RUTGERS L.J. 617, 618 (2002).

49. *Id.* at 619.

50. *See generally* Killingsworth, *supra* note 44.

51. *See id.* at 91.

52. *See id.*

53. Patrick F. Gallagher, *The Internet Website Privacy Policy: A Complete Misnomer?*, 35 SUFFOLK U. L. REV. 373, 392 (2001).

54. Narayanan & Shmatikov, *supra* note 3, quoting the Netflix Prize Page ("Even if, for example, you knew all your own ratings and their dates you probably couldn't identify them reliably in the data because only a small sample was included"). *See also* Netflix Prize Page, http://www.netflixprize.com/faq (last visited Aug. 31, 2008).

55. For an example of a privacy policy with promises about the anonymization of collected information, see Kaplan Test Prep and User Admissions Agreement, http://www.kaptest.com/privacy_statement.jhtml (last visited on Aug. 22, 2008).

56. Gallagher, *supra* note 53, at 375.

57. *Id.*

58. The FTC also brought an action against Toysmart.com for attempting to sell its consumer information to others while its privacy policy statement said it would not do so, even though Toysmart.com was facing bankruptcy at the time. Fed. Trade Comm'n v. Toysmart.com, Inc., No. 00-CV11341RGS, 2000 WL 1523287 (D. Mass Aug. 1, 2000).

59. Schneier, *supra* note 2.