Why a Systems-Science Perspective is Needed to Better Inform Data Privacy De-identification Public Policy, Regulation and Law

Daniel C. Barth-Jones, M.P.H., Ph.D.
Department of Epidemiology
Mailman School of Public Health
Columbia, University

Introduction

Systems sciences pursue an understanding of how parts (including individual actors and groups) within a larger system combine via their interactions to produce emergent phenomena which are greater than, or different from, a mere sum of the parts. Important examples of systems behaviors include the existence of threshold phenomena, which allows infectious diseases to spread as epidemics through a population, or the sudden crystallization of supersaturated solutions. I argue that data privacy policy for de-identification must take a systems perspective in order to better understand how combined multi-dimensional (i.e., involving both technical de-identification and administrative/regulatory response) interventions can effectively combine to create practical controls for countering wide-spread re-identification threats.

In his seminal "*Broken Promises*"[1] work on anonymization, Paul Ohm, puts forth a dystopic vision contending that the failure of perfect de-identification will lead all of us through inescapable cycles of accretive re-identifications toward our personal "databases of ruin". Narayanan and Shmatikov extend Ohm's argument by further contending that "any attribute can be identifying in combination with others"[2]. Such misconceptions ignore the fact that data uniqueness, replicability and accessibility all contribute importantly to the probability of realizing successful data re-identifications. By fallaciously conflating "what is possible" with "what is probable", Ohm's assertion of strictly accretive re-identification ignores important underlying mathematical realities regarding information entropy. However, when statistical disclosure limitation methods have been used to de-identify data, it is frequently possible to practically achieve orders of magnitude reductions in data re-identification probabilities, resulting in dramatic increases in "false-positive" re-identifications. Because false-positive data linkages will typically add incorrect data into electronic dossiers, they can help prevent strictly accretive re-identification, thus disrupting Ohm's supposed systemic "crystallization" of iteratively linked de-identified data into accurate dossiers for the vast majority of a population. While human rights based views of data privacy lead us to value the data privacy of all individuals, de-identification's lack of perfection cannot justify its wholesale abandonment, if on a systems level it is capable of preventing mass

[1] Paul Ohm, Broken promises of privacy: Responding to the surprising failure of anonymization a model for protecting privacy, UCLA LAW REVIEW 2010:(57):1701-1777.
[2] Arvind Narayanan and Vitaly Shmatikov, Myths and fallacies of "personally identifiable information", COMMUNICATIONS OF THE ACM, June 2010 53(6):24-26.

re-identification as part of multi-dimensional regulatory approaches. Together, systems modeling and quantitative policy analyses using uncertainty analyses provide us with some necessary scientific tools to critically evaluate the potential impacts of pseudo/anonymization in various regulatory schemas and should be pursued vigorously as a routine approach for conducting data privacy policy evaluations.

Systems Science and Public Policy

Systems sciences have been used for public policy evaluation for several decades to productively conceptualize, model and study interrelating actors and technologies as components within larger, complex and interacting systems. An important motivation for modeling any complex system is that there often exist feedback loops and nonlinear dynamics between the various components of a system that can result in emergent phenomena, or drive systems into persistent states or equilibriums, that differ importantly from what might be expected without a detailed understanding of the interactions between the actors or components within the systems.

Compelling examples exist where systems science yields valuable insights regarding how the emergent properties of interacting agents can cause critical transitions for the entire interacting system. This ranges from simpler phenomena like the sudden crystallization of supersaturated solutions, to much more complex interactions characterized in epidemic theory. In epidemic models, interacting susceptible, infectious and resistant individuals (e.g., in a classic "S-I-R" epidemic) experience an epidemic "threshold" above which epidemic disease transmission grows, and below which it dwindles and disappears. From such models, we have gained the critical insight that preventing epidemics depends on keeping the product of the disease's transmission below a critical threshold (called the Basic Reproduction Number). Additional systems science insights build off this Basic Reproduction Number theory to further guide public health policy by revealing the necessary vaccination levels that will successfully prevent epidemics of infectious diseases such as measles or polio.

A key tool in systems science is the use of mathematical and computer simulation models to formalize our understanding of the system and allow in-depth exploration of hypothetical intervention policy scenarios under varying conditions. Methods include a variety of simulation approaches such as compartmental/system dynamics models and agent-based modeling (ABM). Quantitative policy analyses utilizing uncertainty (e.g. Latin Hypergrid Sampling) and sensitivity analyses can be used to assess the implications of uncertainty about various model parameters for different scenarios in a probabilistically rigorous fashion. Use of such models provides crucial tools for the challenge of designing intervention policies which will balance various competing risks and benefits within the bounds of practical policy resource constraints. Most public policy evaluations face complex risk-benefit trade-offs for which stakeholders will often have conflicting goals and these risk-benefit trade-offs may change over time. In such cases, model-based evaluation can help with understanding of unintended consequences, time-dynamics and cost effectiveness of proposed interventions.

Re-identification Debate: Formalists versus Pragmatists

In their insightful recent review of the ongoing data de-identification debate[3], Rubinstein and Hartzog, examine the protracted exchanges between what they term the debate's "formalists" (valuing privacy guarantees and mathematical proof as the only meaningful basis for policy) and "pragmatists" (seeking workable and effective policy solutions and heuristics). Formalists have insisted that "*there is no evidence that de-identification works either in theory or in practice*" and, further, that "*attempts to quantify its efficacy are unscientific and promote a false sense of security by assuming unrealistic, artificially constrained models of what an adversary might do*".[4]

Pragmatists have countered that most of the re-identification attacks cited by the formalists: 1) have been conducted against data without any proper statistical disclosure limitation methods applied, 2) have been targeted to especially vulnerable subgroups and did not use statistical sampling to assure representative results, and 3) portray re-identification as broadly achievable, when there is not reliable evidence supporting this portrayal. Furthermore, in those limited cases when actual re-identifications have been demonstrated (as opposed to simply establishing that the data had a high proportion of unique observations), verifiable re-identifications have been quite rare. For example, in a recent review of 10 prominent re-identification attacks, only a total of a 268 individuals were re-identified from a total of more than 327 million data records.[5]

Dystopic Vision: Databases of Ruin

In his famous *Broken Promises* paper, Ohm argues that:

> "…*the power of reidentification* will create and amplify privacy harms. Reidentification combines datasets that were meant to be kept apart, and in doing so, *gains power through accretion: Every successful reidentification*, even one that reveals seemingly nonsensitive data like movie ratings, *abets future reidentification. Accretive reidentification makes all of our secrets fundamentally easier to discover and reveal*. Our enemies will find it easier to connect us to facts that they can use to blackmail, harass, defame, frame, or discriminate against us. *Powerful reidentification will draw every one of us closer to what I call our personal "databases of ruin."*[6] [emphasis added]

Ohm's argument is intended to be disturbing. Every piece of data making reference to each of us (even if de-identified) can, and will, be reassembled, to reveal our darkest secrets to our enemies, who will wield this information against us. Ohm's argument is cited by Narayanan and Shmatikov who state that "any information that distinguishes one person from another can be used for re-identifying anonymous data"

[3] Ira Rubinstein and Woodrow Hartzog, Anonymization and Risk, WASHINGTON LAW REVIEW, June 2016 91(2):703-760.
[4] Arvind Narayan and Edward Felton, No silver bullet: De-identification still doesn't work, July 9, 2014. http://randomwalker.info/publications/no-silver-bullet-de-identification.pdf.
[5] Daniel Barth-Jones, Testimony for the De-Identification and HIPAA Hearing for the National Committee on Vital and Health Statistics, May 24, 2016. http://www.ncvhs.hhs.gov/wp-content/uploads/2016/04/BARTH-JONES.pdf. Also see: Khaled El Emam et al. A Systematic Review of Re-Identification Attacks on Health Data. PLoS One 2011;6(12):e28071. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0028071.
[6] Ohm. *Supra* note 1. p. 1705.

and "the more information about a person is revealed as a consequence of re-identification, the easier it is to identify that person in the future".[7]

The essential disconnect here with respect to the development of judicious data privacy public policy is that the formalist's scenario of widespread "accretive re-identification" fails to distinguish "what is probable" from "what is possible". What is critically ignored in the formalist accretive *database of ruin* depiction is the fact that beyond simple data uniqueness, additional factors of data replicability and accessibility and actors motivations in response to administrative and legal data protections also contribute importantly to the probability of realizing successful data re-identifications.

Why De-identification Should Not Be Discarded

In spite of some advances in re-identification, such as the Narayanan/Shmatikov algorithm used in the Netflix re-identification attack,[8] rumors of de-identification's death have been greatly exaggerated. There are several key reasons why the popular "de-identification doesn't work" narrative fails to promote effective public policy by discarding an integral protection against widespread re-identification.

The first reason is that the vast majority of re-identification demonstrations have been conducted against data without any proper statistical disclosure limitation methods applied,[9] or have blatantly ignored the impact of disclosure controls where they have been applied.[10] The privacy protections provided by the data's distinguishability (i.e., uniqueness), dynamic replicability and/or accessibility from other sources can be non-trivial especially when viewed from a systems perspective. For example, it has become a nearly legendary data privacy axiom that 87 percent of the U.S. population can be re-identified by the combination of their 5-digit ZIP code, gender and full date of birth. Yet, a 2013 study using these quasi-identifiers was only able to re-identify 28 percent of the individuals using these extremely identifying data elements.[11] Furthermore, if the HIPAA de-identification standards had been applied to this data, requiring that only 3-digit ZIP codes and birth year could be included, it would be unlikely that anyone would have been re-identified with these data elements.[12]

The Formalist camp is likely to counter that such examples artificially constrict re-identification risks to only those created by specific "quasi-identifiers" (which may have widespread availability in the U.S.

[7] Narayanan and Shmatikov. *Supra* note 2.
[8] Arvind Narayanan and Vitaly Shmatikov, Robust De-anonymization of Large Sparse Datasets, PROCEEDINGS OF THE 2008 IEEE SYMPOSIUM ON SECURITY AND PRIVACY 2008:SP:111-125.
[9] Barth-Jones. Supra note 5.
[10] One example of Formalists completely ignoring systems-level impacts of statistical disclosure limitation practices can be found in the recounting of the Netflix re-identification attack. Data sampling has long been utilized as a disclosure control method and the Netflix data release included a sample with less than 10 percent of the original data present. This sampling constituted complete protection for the greater than 90 percent of the population who were protected by simply not being present within the sample and makes Narayanan and Shmatikov's conjectures about attacking an office colleague through a water-cooler conversation about their movie preferences much less threatening than portrayed. Such an attack would, of course, be doomed to fail in more than 90 percent of the cases, even if the attack knew their target to be a Netflix movie rater. From a systems perspective, protective impacts of sampling should not be ignored.
[11] Latanya Sweeney et al., Identifying Participants in the Personal Genome Project by Name, April 29, 2013: http://ssrn.com/abstract=2257732. Jane Yakowitz, Reporting Fail: The Reidentification of Personal Genome Project Participants, May 1, 2013: https://blogs.harvard.edu/infolaw/2013/05/01/reporting-fail-the-reidentification-of-personal-genome-project-participants/.
[12] Daniel C. Barth-Jones, The 'Re-Identification' of Governor William Weld's Medical Information: A Critical Re-Examination of Health Data Identification Risks and Privacy Protections, Then and Now, July 2012. http://ssrn.com/abstract=2076397.

through sales from "data brokers"). This point is not without some merit, but it is notable that an adversarial re-identification attempt by one of the leading formalists against 113,000 individuals within the Heritage Health Prize data set yielded zero re-identifications,[13] in spite of the attacker being free to use a broad array of patient medical characteristics, including, but not limited, to patient's age, gender, days in hospital, physician specialty, place of service, medical procedure and diagnosis codes, laboratory tests and prescription drug information.[14] And yet, the conclusion of this re-identification researcher, in spite of his own inability to re-identify anyone, is that "there is *no evidence* that de-identification works either in theory or in practice".[15][Emphasis added]

Another reason why the "de-identification doesn't work" mantra should be suspect is that it assumes as a default that the actors and forces of re-identification are omnipresent, omniscient, omnipotent and relentless. It is obviously highly distortive to assume every potential "data snooper" possesses the necessary motivation, time, resources, and requisite computer skills to implement even an exact matching re-identification attack, let alone a Fellegi-Sunter probabilistic record linkage, or a more sophisticated Narayanan-Shmatikov algorithm attack. Or further that, if such a data intruder is even existent, and so skilled, and unconstrained by monetary and financial resources, that intruder would not be finally constrained by the "limits of human bandwidth" [16] or by the legal prohibitions on re-identification if enacted.[17] In short, the prevailing motto of the formalist camp can be succinctly summarized as the essentially tautological proposition that "super-villains will always win".

Unfortunately though, the inconvenient truth is that we face some unavoidable trade-offs between the statistical accuracy of analyses conducted with de-identified data and the associated data privacy risks. Balancing between statistical accuracy and disclosure risks is essential because some popular de-identification methods (e.g. k-anonymity) can unnecessarily, and often undetectably, degrade the accuracy of de-identified data for multivariate statistical analyses or data mining (distorting variance-covariance matrixes, and masking heterogeneous sub-groups which have been collapsed as part of data generalization protections). This prevents us from being able to fully embrace a precautionary principle with respect to only data privacy risks. The precautionary principle might seem to be appropriate policy guidance if we faced only data privacy risks in isolation.[18] But we clearly face loss of potential benefits and important risks and harms on the other side of the equation with respect to making bad decisions and conducting bad science in any arena where de-identified data is used, if we over de-identify data in an attempt to achieve zero risks or completely abandon de-identification simply because it cannot provide perfect privacy protections.

Still, use of statistical disclosure limitation methods can reduce true positive re-identifications by orders of magnitude and increase false positive linkages. For example, with a k-anonymity of k=5, at least 80

[13] Arvind Narayanan. An Adversarial Analysis of the Reidentifiability of the Heritage Health Prize Dataset. May 27, 2011: http://randomwalker.info/publications/heritage-health-re-identifiability.pdf.
[14] Khaled El Emam et al. De-identification Methods for Open Health Data: The Case of the Heritage Health Prize Claims Dataset. J MED INTERNET RES 2012;14(1):e33 https://www.jmir.org/2012/1/e33/.
[15] Narayanan and Felton *supra* note 4.
[16] *infra* note 20.
[17] Robert Gellman, The Deidentification Dilemma: A Legislative and Contractual Proposal, July 12, 2010 https://fpf.org/wp-content/uploads/2010/07/The_Deidentification_Dilemma.pdf.
[18] Cass R. Sunstein, The Paralyzing Principle, REGULATION, Winter 2002-2003, p32-37.

percent[19] of the linkages made using specified quasi-identifiers would, in fact, be incorrect. Because "false-positive" linkages will typically add incorrect data in attempts to assemble electronic dossiers, used broadly and routinely, de-identification can help to prevent strictly accretive re-identification, thus disrupting systemic "crystallization" of iteratively linked de-identified data into accurate dossiers for the vast majority of a population. It should also be noted that by reducing the probability of successful re-identification, the use of effective de-identification also reduces the likelihood of re-identification even being attempted. This is precisely the sort of system-level feedback loop that adopting a systems science perspective can help us to recognize, and optimize, in order to successfully prevent widespread re-identification.

This also leads to perhaps the most important motivation for not abandoning de-identification protections. Which is that, in those cases where statistical disclosure limitation cannot produce data which has both very small re-identification risks and the necessary data accuracy/utility (and this will admittedly be the case sometimes), disclosure risk assessments are still extremely useful for helping us to set the boundaries beyond which data should not be released without requiring additional administrative and legal protections to minimize re-identification risks.

The picture painted by Ohm, Narayanan and Shmatikov is one in which the usual laws of "data thermodynamics" have somehow been suspended so our crystallizing data fragments can only grow, but never dissolve under the countervailing entropic forces of "data divergence"[20] (errors, discrepancies, differing coding schemas, missing data, etc.) and data dynamics over time. We need to actively design privacy protections at a systematic level, using combined technical de-identification practices and additional administrative and legal data privacy controls to assure that widespread accretive re-identification cannot occur. There is some notable irony involved in Ohm's database of ruin scenario because Ohm insightfully warned us about the misdirecting public policy implications of confusing statements about possibilities with inevitabilities in his paper *Myth of the Superuser*. [21]

No Silver Bullets: Why We Need Multidimensional Data Privacy Interventions

In his Superuser paper, Ohm presciently urged policymakers to stop using tropes of fear and called for better empirical work on the probability of online harms (among which data re-identification must presumably be included). Coming full-circle several years into the de/re-identification debate that Ohm initiated, Rubinstein and Hartzog accurately conclude that the first law of privacy policy is that "there are no silver-bullet solutions" and argue that the best way to move policy past the purported failures of anonymization is to instead focus on the process of minimizing risk of re-identification.[22]

---

[19] Note that this would often be a dramatic underestimate because k-anonymity examines quasi-identifer frequencies only within the sample data, but typically there will be many more people in the larger population sharing the same quasi-identifiers.

[20] Mark Elliot and Angela Dale, Scenarios of attack: the data intruder's perspective on statistical disclosure risk. Netherlands Official Statistics, Volume 14, Spring 1999, Special issue: Statistical disclosure control, p. 6-10.

[21] Paul Ohm, The Myth of the Superuser: Fear, Risk, and Harm Online, UC DAVIS LAW REVIEW, April 2008, 41(4):1327-1402. (p. 1360). Also see, Daniel Barth-Jones, Re-Identification Risks and Myths, Superusers and Super Stories (Part I) September 6, 2012 https://concurringopinions.com/archives/2012/09/re-identification-risks-and-myths-superusers-and-super-stories-part-i-risks-and-myths.html and (Part II) https://concurringopinions.com/archives/2012/09/re-identification-risks-and-myths-superusers-and-super-stories-part-ii-superusers-and-super-stories.html.

[22] Rubinstein and Hartzog, *supra* note 3.

Careful application of more rigorous statistical and scientific study design, along with systems science-based data intrusion scenario modeling for de/re-identification can importantly help resist having policymaking dominated by worst-case "superuser-esque" rhetoric presupposing universal re-identification. Better statistical and study design for re-identification experiments will help assure that we move beyond the collection of "anecdata" to provide representative estimates of data re-identification risks[23]. Systems science modeling will help us to better see the big picture of how we can best protect data privacy through combined technical and administrative/legal solutions. Modern probabilistic uncertainty analyses further allow us to examine impact of the considerable uncertainties that exist for re-identification scenarios and experiments. By enforcing evaluations using comprehensive and probabilistically consistent accounting at each step along the paths of events necessary to realize correct re-identifications for each and every member of a population, a systems science perspective can successfully redirect policy to deal with "what is probable" instead of only worst-case scenarios of "what is possible".

Doing this successfully will require a number of improvements in the current practice of "re-identification science" so it can better inform data privacy policy and practice. Re-identification researchers must focus on demonstrating re-identification risks on data where modern statistical disclosure control methods have actually been used. They must routinely use proper statistical random samples and scientific study designs in order to provide representative re-identification risk estimates for the populations they investigate. To better address the criticism that all data could potentially be identifying, they should also begin to implement and use ethically-designed re-identification experiments[24] to better characterize re-identification risks for quasi-identifiers and other data elements beyond simple demographics. It is also imperative that they design their re-identification experiments to purposefully demonstrate the boundaries where de-identification finally succeeds and to provide empirical evidence to justify any of their data intruder knowledge assumptions. They should verify any purported re-identifications and report their false-positive rates for supposed re-identifications so this information can be used in systems model evaluations of cumulative and multi-stage re-identification attacks. Finally, it is also important that they fully specify their re-identification threat models and report on a variety of realistic and relevant threats.

These improved re-identification research steps, combined with the use of systems modeling and quantitative policy analyses including uncertainty analyses can provide us with the necessary scientific tools to critically evaluate the potential impacts of pseudo/anonymization in various

[23] Daniel Barth-Jones, The Antidote for "Anecdata": A Little Science Can Separate Data Privacy Facts from Folklore, November 21, 2014: https://blogs.harvard.edu/infolaw/2014/11/21/the-antidote-for-anecdata-a-little-science-can-separate-data-privacy-facts-from-folklore/.

[24] Daniel Barth-Jones, Ethical Concerns, Conduct and Public Policy for Re-Identification and De-identification Practice, Harvard Law Petrie-Flom Center, Online Symposium on the Law, Ethics & Science of Re-identification Demonstrations: http://blogs.law.harvard.edu/billofhealth/2013/10/02/ethical-concerns-conduct-and-public-policy-for-re-identification-and-de-identification-practice-part-3-re-identification-symposium/.

regulatory schemas and should be pursued routinely when conducting data privacy policy evaluations.

While a human rights based view of data privacy leads us to equally value the data privacy of all individuals, de-identification's lack of perfection and failure to perfectly protect small percentages of individuals cannot justify its wholesale abandonment, if on a systems level it is capable of preventing mass re-identification as part of multi-dimensional regulatory approaches. This is all the more the case because a human rights perspective motivates us to care not only about potential data privacy harms, but to also broadly value and respect each person's welfare. Given this, we should not quickly jettison the considerable potential of newly emerging and evolving data science opportunities, or already existing research (such as has been occurring with de-identified medical data for decades) without conducting the necessary scientific data privacy policy evaluations needed in order to properly evaluate the efficacy of de-identification practices and the possible threats posed by re-identification risks.