

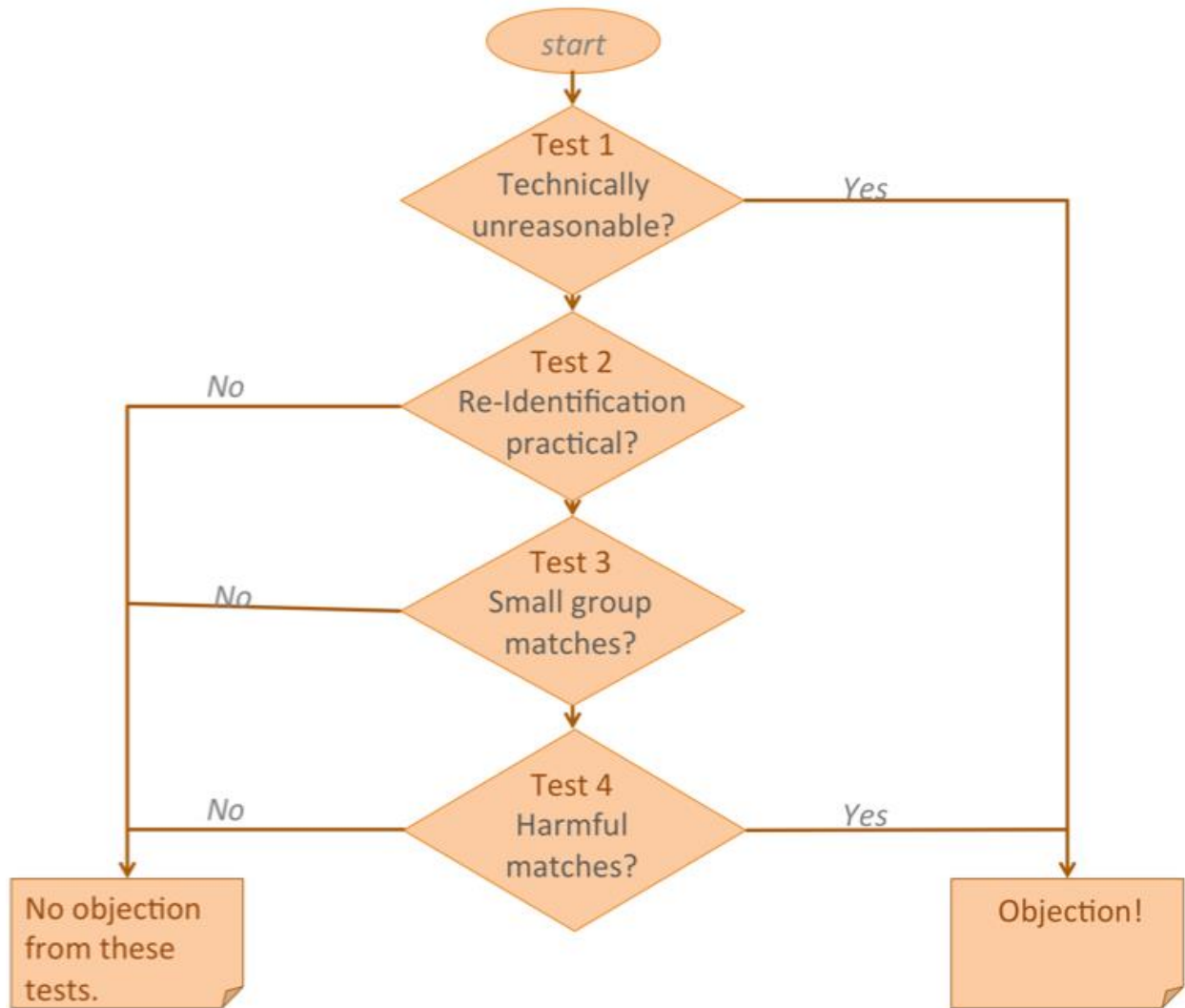


Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data

Latanya Sweeney, Michael von Loewenfeldt, and Melissa Perry

Highlights

- Leading data privacy experts produced four protocols that they said were popular ways to render personal information anonymous so it could be shared or sold publicly
- Experts relied on the HIPAA Safe Harbor, a flawed use of k-anonymity, an enclave, randomization, and standardized statistical values
- None of their protocols achieved the privacy protection they promised
- We were able to put names to the records in all of their protocols.



Sequence of 4 litmus tests to perform on a Sander Team protocol.

Abstract

Society trusts data privacy practitioners to make decisions about which fields of personal income, medical, or educational information can be shared publicly in accordance with laws and standards. How good are the decisions they make? They don't have to publish the protocols they use, and they often prohibit others from telling them about vulnerabilities found in the data. So, in the silence, these practitioners circularly assert that there are no problems. We had a unique opportunity in a legal setting to examine the real-world decision-making of a team of accomplished data privacy experts and to test the quality and accuracy of the decisions they make. The litigation, *Richard Sander et. al v. State Bar of California et. al.*, was over whether the release of requested data was required by California law [1]. During the lawsuit, an expert team of data privacy practitioners proposed four "best practice" protocols that they asserted were sufficient to protect the privacy of individuals whose

information was in the data. All four protocols claimed to leverage approaches widely used today in government, corporate, and research practice. This paper presents their protocols and shows, based on analysis that was made public during the trial, vulnerabilities that each protocol had to re-identifications – the ability to associate real names to “anonymized” data records.

Results summary: One protocol used a physical data enclave, two purported to produce a k -anonymous version of the data, and a fourth protocol developed a statistical model of the data. None of the protocols provided the privacy protection promised or commensurate with common expectations under public records laws. We demonstrate important lessons: (1) k -anonymity guarantees that an adversary cannot do better than guessing that a name matches to at least k records or, vice versa, that at least k people ambiguously match to a record. None of the “ k -anonymity” protocols were actually k -anonymous. (2) In today's data-rich, networked society, the k constraint must be enforced across all fields or scientific justification provided to exclude a field. The “ k -anonymity” protocols excluded some fields from k protection void of analytical rationale. We demonstrated ways to use those fields to help put names to records de-identified by these protocols. (3) We found small group re-identifications in all their protocols that were as harmful as unique re-identifications. (4) The physical data enclave limited access to the data, but still could not thwart hiding or memorizing sensitive information on targeted individuals. (5) All four protocols left the records of Black and Hispanic test-takers significantly more identifiable than the records of Whites. The Superior Court of California denied Sander's request for compelled disclosure of the data, and the California Court of Appeals upheld the decision. Our findings demonstrate how adversarial testing on de-identified data can point out vulnerabilities and improve real-world practice.

Introduction

A data privacy practitioner makes daily decisions about whether and how to share your personal data with others. Those decisions dictate what information about you will be available for everyone to see. What are the steps he takes to protect your privacy? The first step is obvious: he removes your name and address and any other explicit identifiers, such as your Social Security number, from the data. Then comes the hard part. He answers a lot of questions specific to the kind of data involved. Should your publicly available medical information contain your age or decade of birth? Is it okay for details about your income to be associated with the ages of your children and your hometown, race, and employer information? Should your courses, grades, test scores, and graduation dates be included in a release of educational data? The shared information, while appearing “anonymous,” can be used in combination with other information to triangulate to specific individuals or to small groups of named people, even if no directly identifying information appears in the data. So, the data privacy practitioner's goal is to provide useful information to researchers,

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018.
<http://techscience.org/a/2018111301>

businesses and others who receive the data, but to not provide so much information about individuals that others can associate the information with identifiable individuals.

Society relies on data privacy practitioners to protect individuals from the harm of being correctly, or even incorrectly, identified in shared data. This concern – the ability to relate data to specific individuals – is the primary, but not the only interest that is referred to as “data privacy.” Some, for example, consider the right to control the use of personal data an aspect of data privacy, whether or not the data is later capable of being associated with the individual at issue. This paper discusses the first problem only – the ability to associate shared records to named individuals – without discounting or rejecting the importance of data control or other aspects of privacy.

If no named individuals can be reliably associated with records in the data, then the particular privacy risk that is the subject of this paper is ameliorated in many data privacy settings, and so society can reap the benefits of sharing the data with others to improve services, reduce costs, assess policies, and advance science. But what if these trusted practitioners make bad or poor decisions? Then individual health, financial, or educational information that is believed to be anonymous can be vulnerable to re-identification. Others may associate the data with the individual, who, in turn, may suffer serious ramifications. Worse, the individual may never know about the original “anonymous” source or subsequent association. Being unaware of the data or her information’s vulnerability, she is also unable to correct or stop abuses from happening.

Privacy laws and regulations offer guidance on how to redact some kinds of shared data, such as medical, income, and educational data. However, this guidance is limited and deals poorly with new types of data and new techniques for re-identifying data. The practitioner still has to answer many open questions and should not comfortably rely on the assumption that formerly accepted techniques will continue to work.

There is tremendous variability in what data privacy practitioners may know. Practitioners include those who might know little or nothing about statistics, leading data privacy scientists who conduct experiments and provide scientific proofs of compliance, and professional statisticians who work in federal statistics offices worldwide. In many situations, there is no requirement that a practitioner have any specific knowledge, and, as a result, some data privacy practitioners may just be guessing. For example, the Privacy Rule of the Health Information Portability and Accountability Act (HIPAA) is the U.S. federal regulation that governs the sharing of patient health information by doctors, hospitals, and others involved in direct patient care or in the billing for that care [2]. HIPAA allows “someone skilled in the art” to make data sharing decisions about personal health data but does not actually define what the person’s skill requirements might be.

We do not expect today’s practitioners to be perfect. Instead, we expect them to make the best possible decisions, to be accountable for and transparent about the methodologies they

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018. <http://techscience.org/a/2018111301>

use, to learn from past experience, to improve as scientific knowledge about vulnerabilities and better privacy protections becomes available, and to make better decisions tomorrow.

A model for this kind of learning cycle comes from encryption. Society and governments have long needed an ability to share information secretly. At first, those seeking to encrypt data used ad hoc schemes. The resulting encrypted text looked so different from the original text that early encryption users wrongly believed that what they did was sufficient. National and business secrets relied on these methods. Eventually, others broke those naïve encryptions by showing how someone could learn the original text from the encrypted value. Smarter approaches emerged, and smart folks broke those too. The cycle continued until eventually we achieved the strong encryption we enjoy today. Without today's strong encryption, it would be impossible to make purchases, use email, or do online tasks that require a secure connection between computers.

This should happen in data privacy. Practitioners use contemporary methods, then data privacy scientists "break" those methods by exposing vulnerabilities, leading to the development of better methods. Eventually, if practitioners and scientists iterate through enough cycles, society will have strong privacy technologies capable of providing useful data in a variety of settings with guarantees of privacy.

How are today's data privacy practitioners doing? They publish datasets, but rarely do they publish analyses of why they believe a dataset is sufficiently protected. They tend to use Data Use Agreements that prohibit anyone from telling them about vulnerabilities found in a published dataset, making it hard for them to know what is not working. The lack of transparency and feedback makes learning and improving difficult and assessing performance infeasible.

Laws do not always help improve data privacy. For example, under HIPAA, improper handling of identifiable patient information can result in civil and criminal penalties. For example, an incidental data breach can cost \$50,000 or more. A knowing disclosure can result in a criminal penalty of \$250,000 and ten years' imprisonment [3]. However, if a data privacy practitioner redacts the data and determines that the risk of re-identification is very small, then the redacted version can be shared freely without concern for civil or criminal penalties (the "expert determination" clause of HIPAA) [4]. This allows data to flow free of penalties, but HIPAA never defines how small is small.

The practitioner should be "skilled in the art," according to HIPAA, but HIPAA does not describe what that skill, education, or experience should be, nor does HIPAA require the practitioner to publish the basis he used to determine that a medical dataset is sufficiently anonymous for sharing publicly. As a result, datasets appear in the public and are shared widely without knowing who made what decisions or why.

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018.
<http://techscience.org/a/2018111301>

When vulnerability is found in a publicly available dataset, and names are reliably put to records, the cycle of improvement and learning that should occur does not seem to happen. The primary reason is a clause that frequently appears in a data use agreement that prohibits anyone who receives the data from attempting to learn the identities of the individuals in the dataset. While that sounds like a good idea, these clauses also typically impose a gag order that prevents anyone from talking about vulnerabilities he finds or uses. These clauses do not necessarily stop anyone from identifying individuals in the data; they just stop the larger community from learning about those identifications (and therefore prevent addressing them). These prohibition clauses break the learning cycle by relying on an unproven assumption that the older methods work, and by preventing the identification and resolution of defects in those methods. The result is no learning and no improvement.

Worse, for those who may have the greatest incentive to exploit the data, a data use agreement with a prohibition clause is a questionable deterrent. For example, among the top multi-state acquirers of statewide hospital data are data analytic companies [5], many of which have data products that seem to rely on linking statewide hospital data with other data. Financial incentives may encourage the exploitation of data vulnerabilities and outweigh any concerns raised by the data use agreement. Conversely, a data use agreement that discourages telling anyone about known vulnerabilities effectively preserves opportunities for exploitation.

When no one can report data vulnerabilities, practitioners can wrongly interpret silence as adequate data protection even when serious vulnerabilities continue to exist.

There have been a few cases of scientific studies that demonstrate vulnerabilities in data, but even then, some practitioners have been reluctant to improve.

For example, Harvard researcher and co-author Latanya Sweeney purchased a copy of Washington State's publicly available hospital dataset for \$50 in 2012 [6]. It seemed to have all hospitalizations occurring in the state in the year, and included patient demographics, diagnoses, procedures, attending physician, name of the hospital, a summary of charges, and how the bill was paid. It did not contain patient names or addresses, only the U.S. residential postal codes known as ZIP codes.

Newspaper stories printed in Washington State for the same year that contained the word "hospitalized" often included a patient's name and residential information and explained the reason for the hospitalization, such as a vehicle accident or assault. Sweeney assembled a sample of 81 news stories and found that news information uniquely and exactly matched medical records in the Washington state database for 35 of the 81 sample cases (or 43 percent), thereby putting names to patient records. An independent news reporter verified matches by contacting patients and found them all correct (editors agreed not to publish any names without the explicit consent of the patient) [6][7]. Matches included high-profile cases such as politicians, professional athletes, and successful businesspeople. Some of the codes

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018. <http://techscience.org/a/2018111301>

included sensitive information beyond the purpose of the visit, such as drug and alcohol use and sexually transmitted diseases. Sweeney noted that this news information is the same kind of information that a banker, employer, family, friend, or neighbor might know about a patient and could therefore learn the same details from the public dataset. This is also the same kind of information that a holder of larger collections of related information, such as prescription data, medical marketing data, or health data gathered through mobile phone apps, has and could use to learn the health details on a large number of individuals.

After becoming aware of the experimental results, Washington State immediately addressed the problem by improving the protection of the publicly available version and making a more detailed version available through an application process [8][9]. A cycle of learning occurred, and privacy improved. But somehow the knowledge didn't generalize to the other 28 states having similar data, seemingly because the data privacy practitioners for the other states did not see or believe that the lesson of Washington State applied to their data. Rather than accept that historically accepted methods were inadequate, practitioners reflexively assumed that the vulnerabilities were anecdotal rather than systemic. Recently, experiments were replicated on the same kind of data, but from other states [10]. Proceeding state by state to demonstrate the same vulnerability seems inefficient, but practitioners seem resistant to change.

How can we get a glimpse into the decision-making of those who make data privacy determinations? Are practitioners making good decisions? Are they using state-of-the-art knowledge?

In this paper, we write about the unique opportunity we had through a lawsuit to see how some data privacy practitioners determine whether a dataset is sufficiently anonymous, and to test whether their decisions actually protect privacy. Co-author Sweeney provided independent analyses and experiments and presented her results as an expert witness. Co-authors Loewenfeldt and Perry were the attorneys who represented one of the parties.

The claimed data protection methods discussed in this paper provide a good example of the type of ad hoc assumptions about privacy that are common, and how those assumptions may be proven wrong (as they were here). This paper uses the legal proceeding as an illustration, but the same basic problems exist throughout the data privacy practice. This paper is intended as an illustration of these areas of concern and is not intended as a definitive statement of the specific facts of the particular proceeding or the legal standards and results therein. Where necessary to expand upon or make a point, this paper considers matters that were not part of the prior legal proceeding and presents examples of problems that did not necessarily play a role in that specific proceeding. The views expressed in this paper are the personal views of the authors expressed solely for purposes of academic exploration, and do not constitute a statement of the position of the author's clients, including The State Bar of California or its Board of Trustees.

Background

Richard Sander, a professor at the UCLA Law School, studies how race-based law school affirmative action policies relate to law school outcomes. Believing that data collected by the State Bar of California ("the Bar") from applicants for admission would be a good source of data for his research, in 2008 he requested that the Bar provide him with individual-level data on the race, law school, year of graduation, bar exam score, bar passage result, LSAT score, law school GPA, and undergraduate GPA for every individual who had attempted to pass the California bar examination between 1972 and 2007.

The Bar had never before publicly released data of the type that Sander requested. Neither had any other state bar in the country. The Bar stores confidential information about applicants, including an applicant's gender, ethnicity, and where and when the applicant attended law school. Further, in its entire history, the Bar never released scores to individuals who pass the bar examination. The Bar concluded that public release of this data would have been unprecedented and contrary to applicants' reasonable expectations of privacy given the many rules and statutes that govern the confidentiality of Bar admissions records.

The Bar rejected Sander's request. Months later, Sander sued the Bar to compel disclosure of the data he sought. An eight-year legal proceeding, *Richard Sander et. al v. State Bar of California et. al.*, ensued [1].

Professor Sander's original request called for "clustering" the data in a method that could loosely be described as providing 5-anonymity for some fields of data (but not all fields of data). In the legal proceedings, Professor Sander assembled a team of four experienced statisticians, academic researchers, and experts from a data privacy company. Each attested in the proceedings to having a statistical background but no computational expertise other than using statistical software. To be clear, none of them had expertise in computational data privacy or had ever worked for a federal statistics office. Still, these practitioners reportedly had been responsible for privacy preparations in dozens of major datasets and had published papers on data privacy.

The Sander Team proposed four new protocols. They asserted that the protocols were based on the team's understanding of "best practices" in the field and that each was sufficient to protect the privacy of individual bar exam takers while keeping the data useful for Professor Sander's study. Two protocols reflected various decisions Sander's practitioners made on which data to include, exclude, or aggregate. The third protocol relied on a physical enclave, which allowed visitors access to the data while in the enclave and limited the information that could leave the enclave. The final protocol involved constructing a statistical database of standardized (or relative) values.

After a trial, the Superior Court of California denied Sander's request for compelled disclosure of individual-level data from the Bar [1]. At the trial, the Court considered the following

questions: (1) Could the information be provided in a form that protected the privacy of applicants? and (2) Did any countervailing interest outweigh the public's interest in disclosure? The Court decided that, pursuant to Sander's proposed protocols, the information could not be provided in a form that protected the privacy of applicants and that numerous countervailing interests outweighed the public's interest in disclosure. The Court based its decision on five independently sufficient grounds.

1. Disclosure of the requested records would require the Bar to create new records, which no public agency is required to do under the California Public Records Act (CPRA). According to the Court, each of Sander's protocols requires substantial changes to the Bar's existing data and the creation of new records. For example, the protocols require the Bar to recode its original data into new values. Using the same reasoning, the Court found that the data enclave protocol is not a valid remedy under the CPRA, as it would require the Bar to create a data enclave.
2. Disclosure of the requested records is barred by California Business and Professions Code, which prohibits disclosure if data "may identify an individual applicant." The Court found that disclosure of the data pursuant to all of the protocols presented a risk that individual applicants may be re-identified from the data or rendered the data of minimal to no value such that disclosure would be unwarranted. Considering extensive testimony, the Court found that the percentage of unique records that exist after application of three of the four protocols is significantly higher than under acceptable norms. In particular, minority groups are more vulnerable to re-identification than their White counterparts. The Court also found considerable risk in "attribute disclosures," that is, inferences that can be drawn about applicants by virtue of their membership of a particular group. With respect to Sander's protocol that required the creation of a statistical database of the Bar's original data, the Court concluded that the database would offer the least value or utility of any of the protocols. "No purpose is achieved by requiring the State Bar to perform extensive computerized gymnastics to anonymize the data contained in the Admissions Database such that it might be subject to disclosure, when the information has minimal or no value."
3. California Government Code exempts from compelled disclosure "[r]ecords, the disclosure of which is exempted or prohibited pursuant to federal or state law." Because disclosure of the requested records is prohibited by Business and Professions Code, the Court found that the Bar met its burden under this California Public Records Act section [11].
4. Disclosure of the records is an unwarranted invasion of privacy and is thus exempt from disclosure pursuant to California Government Code [11]. In balancing the public and private interests served by disclosure or non-disclosure, the Court concluded that

individual applicants would suffer real-world consequences as a result of public disclosure of their private data.

5. The Bar showed that "the public interest served by not disclosing the record clearly outweighs the public interest served by disclosure of the record," and is thus exempt from disclosure pursuant to California Government Code [11]. The Court found that non-disclosure of the requested data would protect the general public from adverse consequences resulting from public disclosure of the data, protect the Bar's ability to collect and release data in the future, and protect the Bar from the burdens imposed by disclosure.

The California Public Records Act (CPRA)

At the time of trial, the governing privacy standard for the case was the California Public Records Act (CPRA) [11], which requires disclosure of governmental records to the public upon request, unless exempted by law. It is important to note that the general principles and standards discussed in this paper with respect to the CPRA are not unique to California. Indeed, many other state public records laws and the federal Freedom of Information Act (FOIA) [12] have similar standards. The CPRA was, in fact, modeled on its federal predecessor, FOIA.

Pursuant to the CPRA, anyone can request to inspect or request the disclosure of public documents [11]. The purpose does not have to be stated, and access to the released data cannot be conditioned on a data use agreement. In fact, once public records are provided to one requester in this setting, the same records must be produced to anyone in the public who seeks them. There are numerous sections in the CPRA that exempt certain records from compelled disclosure. For example, disclosure is not required when otherwise prohibited by federal or state law, if disclosure would constitute an unwarranted invasion of privacy, or if the public agency can show that the interests in non-disclosure clearly outweigh the interests in disclosure of the requested information.

Generally, public record requests cannot compel a government agency to create a new database; a requester can only ask that existing records be redacted (some items removed) or recoded (usually replacing values with less precise ones), and the effort involved to do so has to be reasonable [13][14][15]. Earlier, a court found the State Bar of California, an administrative arm of the California Supreme Court, to be subject to public record requests.

This is a fundamental principle of public records law, which provides access to *records*, not access to information. A local agency has no duty to create a record that does not exist at the time of the request [16][13]. "It is well settled that an agency is not required by FOIA to create a document that does not exist in order to satisfy a request . . . [A] requester is entitled only to records that an agency has in fact chosen to create and retain." [17][18]. Thus, while an

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018. <http://techscience.org/a/2018111301>

agency can be required to redact, extract, or rearrange existing data, an agency cannot be required to change its existing data or to create new data.

For example, a citizens' organization, Students Against Genocide, brought a FOIA action seeking release of reconnaissance photographs in a lower, non-classified resolution for the Department of State [19]. Despite the fact that producing the lower-resolution photographs was a technically trivial process, the court rejected Students Against Genocide's request because it would have required the Department of State to create new documents.

By way of another example, the Center for Public Integrity sought data from the Federal Communications Commission (FCC) and requested that the FCC replace individuals' responses with numerical ranges or an indication of whether the deleted responses were zero or greater than zero [20]. The court rejected such re-coding as creation of a new record.

If the circumstances under the various CPRA/FOIA exemptions are met, or disclosure requires an agency to create new documents, disclosure cannot be compelled, although under some circumstances an agency could choose to release the data voluntarily.

This paper does not attempt to address whether the trial court's specific factual and legal findings were correct. Instead, this paper uses the facts developed at trial as an illustration of data anonymity problems from a data privacy practitioner's perspective: the approach to protecting the data, and whether the approach is effective.

In order to appreciate both the protocols that the Sander Team presented and the approaches that we used to assess their protocols, you need to know how data having no names ("de-identified data") can have names associated with records in the data ("re-identification") and how to compute measures of risk ("binsizes") and compare those measures to a standard ("the HIPAA Safe Harbor") and to formal protection models (" k -anonymity"). We describe and unpack each of these terms and concepts below.

De-identification and the Safe Harbor Provision of the U.S. Health Insurance Portability and Accountability Act (HIPAA)

Sander and his team wanted to convince the court that its protocols were sufficient, so at times they made comparisons to the Safe Harbor Provision of the U.S. Health Insurance Portability and Accountability Act (HIPAA) [2].

HIPAA provides four ways of sharing health data beyond patient care. One is the expert determination provision described earlier. Another is the Safe Harbor provision. The HIPAA Safe Harbor provision is prescriptive. It requires eliminating 16 kinds of patient identifiers (including patient name, Social Security number, email address, and telephone, account, and all other record numbers) and generalizing date and geography information: dates must be reported as year, and the smallest reportable geographic subdivision is the first 3 digits of the

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018. <http://techscience.org/a/2018111301>

ZIP (postal) code (unless the three-digit zip code contains fewer than 20,000 individuals, in which case it is reported as 000) [21]. Personal health information redacted in this format can be shared widely, online or offline, with no restrictions and without a data use agreement.

The HIPAA Safe Harbor uses a traditional pillar of data privacy known as de-identification – the removal of explicit identifiers from data to make the result sufficiently anonymous. The rationale behind de-identification is simple. If an individual cannot be distinctly identified in data, then no one can be directly harmed, and so the data can be shared widely. The redactions should prevent others from learning the distinct identity of an individual (thereby protecting the individual from harm), while the dataset as a whole should retain useful information for worthy purposes.

The HIPAA Safe Harbor is convenient. A researcher can easily comply with the HIPAA Safe Harbor by merely making the appropriate data redactions. Visual inspection confirms compliance. No special computer programs, statistical modeling, or advanced analysis is necessary.

HIPAA does not require a zero risk of re-identification. In 2011 El Emam et al. conducted a review of 14 published re-identification attacks [22]. Of the 14 examples, the authors dismiss 11 as being conducted by researchers solely to demonstrate or evaluate the existence of a risk of re-identification, not to perform any actual re-identifications having results that are verified as being correct or not. They classify the work of Narayanan and Shmatikov [23] as in this category. Narayanan and Shmatikov demonstrated the possibility of re-identifying published Netflix rental histories from the (identified) movie reviews submitted by Netflix customers.

Of the remaining "three actual [or correct] re-identifications", El Emam and his co-authors dismiss two as having standards below those set by the HIPAA Safe Harbor. The authors promote the remaining study by Kwok and Lafky as being HIPAA compliant and as having a very low risk of re-identification [24]. Kwok and Lafky associated names to 2 of 15,000 (or 0.013 percent) HIPAA Safe Harbor-compliant hospital admission records of Hispanics by matching {ethnicity, year of birth, gender, first 3 digits of ZIP, and marital status} to marketing data that also included name and address. [In their very short 8-page paper, they generalize the rate to 0.22 percent based on undocumented assumptions, so we use the 0.013 percent that they actually reported as their experimental result.]

More generally, Sweeney used 1990 Census data to estimate that 0.04 percent of the United States population was uniquely identified by the basic demographic fields allowed by the HIPAA Safe Harbor – namely, {year of birth, gender, and first 3 digits of ZIP} [25]. Both the study by Kwok and Lafky and the study by Sweeney only examined demographic fields, and both found low likelihoods for unique re-identifications.

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018. <http://techscience.org/a/2018111301>

In a recent study, Sweeney et al. demonstrated a unique re-identification rate in HIPAA Safe Harbor data of 20 percent [26], which of course, is orders of magnitude greater than the 0.013 and 0.04 percent rates just discussed. This new study uses fields beyond demographics to make matches. With a 20 percent re-identification rate, the new finding makes the idea of comparing redacted datasets to a rate permissible under the HIPAA Safe Harbor extremely problematical. However, the work by the Sander Team predated this finding, so unless stated otherwise, we will use the 0.04 percent value for comparison.

Re-identification

When sharing personal data widely, the biggest privacy threat to de-identified data is re-identification – the ability for an interested party to use reasonable effort to match details in the de-identified dataset to distinct individuals. We use the term “named individual” to refer to having sufficient information to identify an individual by name. If specific records in a de-identified dataset can be associated with one or few named individuals, then we say the dataset is re-identified. Harm from a re-identification may result if sensitive information contained in the data becomes known about named individuals. For example, when Sweeney re-identified hospital discharge data released by Washington State, her re-identification exposed records that included sensitive information such as “references to venereal diseases, drug dependency, alcohol use, [and] tobacco use” [6].

A “unique re-identification” occurs when a record in the data matches to exactly one named individual. For example, Sweeney’s re-identification of de-identified health records from Washington State correctly matched one name to one record in 43 of the sample of 81 news stories [6].

A “group re-identification” occurs when one or a few records in the dataset match to a small number of named individuals. Both unique and group re-identifications raise privacy concerns. A one-to-few match or a few-to-few match can be just as damaging as a one-to-one match. For example, showing that a record in a de-identified dataset of lead poisoning cases belongs to one of few named individuals would allow all the individuals in the group to suffer the same adverse consequences, even though only one individual actually has the lead poisoning. As another example, a group re-identification of de-identified health records showing that six of seven named individuals have a genetic disposition toward cancer would result in each individual being equally likely (6 in 7) to have that condition, including the individual without the condition. It is well recognized that one-to-few and few-to-few re-identification poses privacy risks similar to unique re-identification [27].

Re-identification Strategy

A “re-identification strategy” is a means to assign identifying information to entities (e.g., individuals or addresses) whose information is believed to appear in de-identified records. Approaches typically include a stepwise process applied to various datasets, where one of

the datasets is the de-identified dataset itself. We use re-identification strategies to show whether a protocol presented by the Sander Team prevents re-identification.

The relevant outcome of a re-identification strategy is usually a set of sufficiently small group re-identifications. The total "number of re-identifications" is the number of records re-identified, regardless of whether the correct identification is included. If only unique re-identifications are of interest, then the number of re-identifications is the number of one-to-one associations found. When larger-sized groups are relevant, then the number of re-identifications is the number of groups. For example, consider a re-identification having 4 groups, with 2 named individuals in each group. One person in each of the two-person groups is believed to be the correct person, but the re-identification strategy does not distinguish which of the two named individuals that individual might be. Therefore, the number of re-identifications is 4, one individual from each group.

A re-identification is not necessarily correct. There may be strong reason to believe the association is correct, even if it is wrong. We use the term "correct re-identification" to identify whether a given re-identification actually identified the correct individual. If a reliable re-identification strategy strongly associates a record to an individual incorrectly, then the incorrect individual will likely suffer the same harm as if he was the correct individual. This is particularly true where the identification is unknown to the identified individual, who therefore has no ability to correct any mis-attribution. Therefore, incorrect re-identifications and correct re-identifications are both important.

In prior work, Sweeney introduced the notion of a "binsize" as the number of individuals that match one or more de-identified records indistinguishably [6][28][29]. Unique re-identifications have a binsize of 1, denoting a single one-to-one matchup, uniquely identifying the individual. A binsize of k lists k possible matches to a single record.

The number of unique re-identifications is the value at binsize 1 (we write $k=1$). Past government data sharing policies expected no re-identifications for binsizes of 5 or less ($k \leq 5$) (e.g., [30]). Recent government data sharing policies proscribe no re-identifications for binsizes of 10 or less ($k \leq 10$) (e.g., [31]). Guidelines for defamation cases suggest that a finding of defamation requires binsizes less than 20 ($k < 20$) (e.g., [32]) internationally or 25 (e.g., [33][34]) in the United States. Therefore, in discussions about the protocols from the Sander Team, we report the number of re-identifications for $k=1$, $k < 5$, $k < 11$, and $k < 20$, unless a different level is expressed by the Sander Team.

A re-identification strategy identifies a "risk pool" for groups 1 to k [35], comprising all distinct individuals named in the re-identified groups from size 1 to k . Risk pools are important because they identify others who may be harmed indiscriminately. In the prior example in which the results of a re-identification strategy were 4 groups of two named individuals (binsize = 2), then 8 named individuals are in the re-identification pool, and the total number of re-identifications is 4. Notice that the risk pool, as defined here, relates to a

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018. <http://techscience.org/a/2018111301>

re-identification strategy. Another re-identification strategy operating on the same de-identified dataset may generate a different risk pool.

***k*-anonymity**

How could data be released with limited or virtually no risk of re-identification? To eliminate risk of re-identification, data must adhere to a formal property that provides a privacy guarantee. Computer scientists have introduced such models. The first formal protection model was *k*-anonymity, which guarantees that each record released will ambiguously map to at least *k* other records [28][36]. Therefore, you cannot do better than guessing $1/k$ that any particular record belongs to a named individual. If data are *k*-anonymized, there would be *by definition* no small group re-identifications less than *k*, and each *k*-sized group would be indistinguishable. This guarantee would hold regardless of the amount or nature of redaction.

We introduce the notion of *k*-anonymity in this writing because the Sander Team often asserted that they made datasets adhering to *k*-anonymity (*k*-anonymous data) when, as we show, that was not the case.

Methods

We assess privacy (i.e., re-identification) risks in each of four protocols provided by the Sander Team, who decided on the number and nature of these protocols. The Team was provided a detailed version of the Bar data for the purpose of producing sufficiently anonymous versions of the data; these data had no names or explicit identifiers but were considered sensitive and identifiable. The Sander Team used these data to demonstrate its protocols. The assertion was that each protocol protects privacy while remaining useful for Professor Sander's study.

The proposed protocols have as their input the underlying raw data from the bar (Bar Data) and as their output a candidate dataset for public disclosure. We perform a privacy assessment on each protocol by identifying privacy risks and performing sample re-identifications to further demonstrate those risks. Our goal is to test the hypothesis proffered, but not demonstrated scientifically, by the Sander Team that their proposed protocols prevent re-identification of data of this sort.

Data and Tools

In preparing this paper, we used the public record from the litigation, including publicly filed expert reports concerning the underlying source data, descriptions and datasets from the Sander Team for four protocols, and information filed in the public Court record. We did not use any of the underlying private data except to the extent that analysis of it was presented in open court during the litigation. We also used other data and information available on the

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018. <http://techscience.org/a/2018111301>

Internet, the Stata program, a spreadsheet program and the Python programming language, all working on an off-the-shelf laptop. At times, we refer to and use data that Professor Sander received in response to other public record requests directly from law schools. Below are further descriptions of the datasets and protocols.

Bar Data

We term the relevant raw data held by the California Bar the "Bar Dataset." The Bar Dataset has fields for the race, law school, year of graduation, bar score, bar passage result, Law School Admission Test (LSAT) score, and law school grade point average (GPA) for every individual who attempted to pass the California bar between 1977 and 2008 ("Bar Data"). As set forth in the trial record, the Bar Dataset has a total of 139,338 rows, one row for each bar taker.

Figure 1 lists the fields of the Bar Dataset. The sample record set forth below describes a *hypothetical* (i.e., invented for purpose of illustration) White individual who graduated from Stanford Law School in August 2000 with a 75.1 GPA. The individual had an LSAT score of 136 and passed the bar after multiple attempts. His test scores, which even he does not know, appear in the data, and are listed as "1476..." as an example in Figure 1.

Field Name	Field Description	Sample Record
<i>recnum</i>	A made-up unique record number for this study	110240
<i>lawschool</i>	Name of law school	Stanford
<i>gradYr</i>	Month and Year of graduation (yyyymm)	200008
<i>LSAT</i>	LSAT score (10-48 scale or 120-180 scale)	136
<i>GPA</i>	Law School GPA (different scales possible)	75.1
<i>race</i>	Race/Ethnicity (8 distinct possible values)	White
<i>result</i>	Passed ("Pass") or not pass ("NoPass")	Pass
<i>tries</i>	Passed after multiple tries ("Multi" or blank)	Multi
<i>scores*</i>	List of scores by area of the exam	1476...

Figure 1. Fields of raw data about bar exam takers held by the California Bar. Values for *race* are: American Indian or Native Alaskan, Asian, Black, Filipino, Hispanic, Indian Subcontinent, Pacific Islander, or White. *The actual test scores were not part of the data used in the litigation but were part of what was requested to be released.

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018.
<http://techscience.org/a/2018111301>

<i>recnum</i>	<i>lawschool</i>	<i>gradYr</i>	<i>LSAT</i>	<i>GPA</i>	<i>race</i>	<i>result</i>	<i>tries</i>	<i>scores</i>
1001	Whittier	199806	141	91.78	White	Pass	Multi	
1002	Whittier	199807	128	85.09	Asian	Pass		
1003	Whittier	199807	132	70.36	Asian	Pass		
1004	Whittier	199808	134	70.36	Hispanic	Pass	Multi	
1005	Pepperdine	199810	143	70.59	White	Pass		
1006	Pepperdine	200006	132	92.65	White	Pass	Multi	
1007	Pepperdine	200006	144	84.2	White	Pass	Multi	
1008	Pepperdine	200006	148	67.45	White	Pass		
1009	Boston University	200608	141	98.65	White	Pass	Multi	
1010	Boston University	200608	148	67.51	White	Pass		
1011	Boston University	200608	151	70.94	Black	Pass	Multi	
1012	Boston University	200608	141	70.94	Black	Pass	Multi	
1013	Pace	200610	161	84.15	Hispanic	Pass		
1014	Regent	200610	163	70.36	Black	Pass	Multi	
1015	Southland	200611	151	70.59	Asian	Pass		
1016	New York	200611	136	81.75	White	Pass	Multi	
1017	South Bay	200612	137	70.94	White	Pass		
1018	Central	200612	138	86.9	White	Pass	Multi	
1019	Valley	200612	139	85	Asian	Pass		
1020	Drake	200601	139	80.38	Asian	Pass	Multi	
1021	Stanford	200106	136	80.2	Asian	Pass		
1022	Stanford	200106	157	82	Asian	Pass		
1023	Stanford	200106	148	82.21	Asian	Pass	Multi	
1024	Stanford	200107	158	80.37	Hispanic	Pass	Multi	
...

Figure 2. Hypothetical illustration of the first rows of data from the Bar Dataset and the Bar Pass Dataset. Values are invented for purpose of illustration. We cover the values for *scores* as a reminder that the actual data would include the bar scores. The *recnum* field only appears for the reader's benefit to track records across protocols.

A subset of the Bar Dataset containing the records of those individuals who passed the California Bar Exam, and by virtue of passing, satisfied that criterion to become members of the California Bar. Of the 139,338 individuals reported in the Bar Dataset, 116,535 of them eventually passed the Bar. We term this subset as the "Bar Pass Dataset," having 116,535 individuals (or rows) and the same fields as the Bar Dataset. The sample record in Figure 1 would be included in both the Bar Dataset and in the Bar Pass Dataset.

Protocol Data

The Sander Team provided 4 protocols. For each protocol, we received a textual description of the protocol, the Stata code to produce a version of data from the Bar Dataset that adheres to the protocol, and a dataset that the Sander Team asserted was the result of executing the protocol on the Bar Data. In summary, for each protocol, we had text and code descriptions of the protocol as well as a dataset that was the implemented instantiation of the protocol on the Bar Data. The underlying data and protocol datasets were under a protective order and not filed in the public record and were not used for purposes of this paper (although the paper reports analysis of those materials publicly revealed during the litigation).

In places, the textual description and the Stata code that we received from the Sander Team, which should have been consistent, were not. In cases where there was conflict between the written description and the code, we used the Stata code as the authoritative source unless it seemed in error.

In presenting this material here, we often simplified the data and protocol descriptions we received for presentation efficiency. In doing so, we took care not to change or alter the effect of the protocols on the data, or to make any changes that would otherwise impact our privacy assessments.

One of the protocols purported to use k -anonymity where k is 11; we term this the "11-Anonymity Protocol." Another protocol took the 11-Anonymity Protocol and made further changes to it, such as randomly removing records and GPAs; we term this the "Plus Protocol." The third protocol used a sequestered facility in which visitors access the data; we term this the "Enclave Protocol." Finally, the last protocol produced a database that reported relative test scores; we term this the "Standardized Protocol." Detailed descriptions of each of these protocols appear below.

11-Anonymity Protocol

As its title implies, the Sander Team asserts that the "11-Anonymity" Protocol adheres to k -anonymity where k is 11 (which is $k \leq 10$). It attempts to do so in the 9 steps enumerated in Figure 3. However, and this is a critical failure, it only enforces k -anonymity across certain fields in the data.

In the first step, the Sander Team 11-Anonymity Protocol drops records of unusual and older test-takers. In step 2, it reduces the numbers of races from 8 to 4, by generalizing designations of Asian, Indian Sub-continent, American Indian, Alaska Native, Filipino, and Pacific Islander into "Other" (see Figure 4a).

In step 3, the 11-Anonymity Protocol generalizes the law schools into categories by changing the data from the name of the school to Class One, Two, or Three, based on classifications

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018.
<http://techscience.org/a/2018111301>

provided by the Sander Team. Class One schools were supposed to have the most test-takers taking the Bar exam and Class Three schools to have the fewest test-takers. Additional processing occurs in subsequent steps on test-taker records from the Class Three schools.

In steps 4 and 5, the 11-Anonymity Protocol replaces the year of graduation with 3- or 6-year ranges for the popular Class One schools and with 9-year ranges for the less popular Class Two and Class Three schools.

The 11-Anonymity Protocol removes the names of the Class Three schools in step 9, but before that occurs, in steps 6 and 7, it appends information about the distribution of LSAT scores in each Class Three school graduation cohort. Specifically, it computes the average LSAT and the quintile or decile in which the test-taker's LSAT occurs among those having the same Class Three School and graduation period. It adds the information to test-taker records using some additional data fields. Quintiles are used for those test-takers in the 1982-1990 graduation cohort, and deciles are used for those in the 1991-1999 and 2000-2008 graduation cohorts. This only applies to test-takers in Class Three schools.

Finally, in Step 8, the 11-Anonymity Protocol recodes race in cases where the numbers of Blacks and Hispanics or the numbers of Whites and Others is less than 11, singly or jointly, in Class Three Schools. If a group of test-takers having the same Class Three school and graduation period has less than 11 Blacks or 11 Hispanics, then if the sum of the two is 11 or more, it changes the race of those Black and Hispanic test-takers to "Under Represented Minority" (see Figure 4b). If the numbers of Blacks and Hispanics combined still do not total to at least 11, then it blanks out the race of those Black and Hispanic test-takers.

Similarly, if a cohort of test-takers having the same Class Three school and graduation period includes less than 11 Whites or 11 "Others" (Asian, Indian Sub-continent, American Indian, Alaska Native, Filipino, and Pacific Islander), then if the sum of the two is 11 or more, the 11-Anonymity Protocol changes the race of those test-takers to "White and Other" (see Figure 4b). If the sum of the White and Other test-takers is still not at least 11, then it blanks out the race of those test-takers.

The 11-Anonymity Protocol's result is a dataset having the 14 fields listed in Figure 5. As an example of how the data appears, Figure 6 shows the results of applying the 11-Anonymity Protocol (Figure 3) on the hypothetical sample of the Bar Dataset in Figure 2. The changes are clear on visual inspection. Instead of reporting the graduation month and year (*gradYr* in Figure 2), graduation appears in multi-year ranges (*gradPeriod* in Figure 6). The race of Asians and all other test-takers that are not Black, White, or Hispanic is changed to "Other." Schools are additionally labeled as "Class One," "Class Two," or "Class Three" (*schoolCategory* in Figure 6).

Test-takers from Class Three schools report more LSAT information. The average LSAT score for the test-taker's school and graduation period (*avgLSAT* in Figure 6) and the decile within

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018.
<http://techscience.org/a/2018111301>

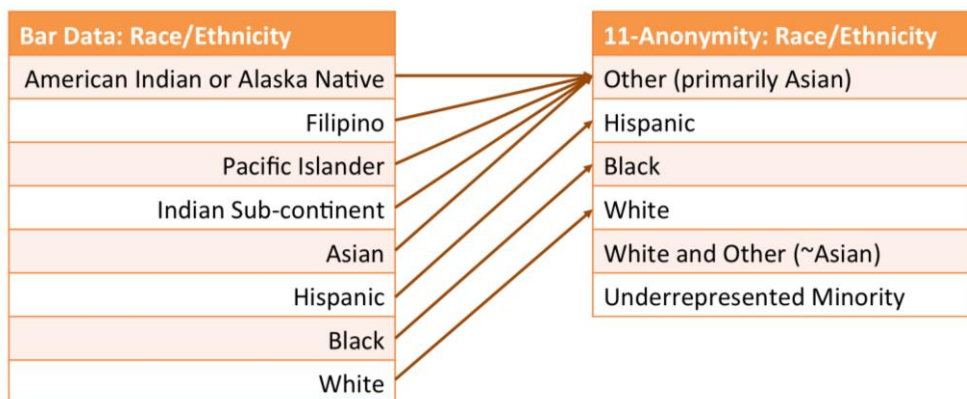
that distribution for the test-taker's LSAT score is appended (*decile00-08* in Figure 6). The fields *quintile82-90* and *decile91-99* are not shown in Figure 6 because in this example they have no values, but for some Class Three test-takers they would have values.

Step 1. Preliminary Steps	
1.1	DROP test-takers attending more than one law school
1.2	DROP test-takers who graduated prior to 1982
1.3	DROP test-takers from unaccredited and correspondence schools
1.4	DROP test-takers missing both LSAT and GPA scores
Step 2. Recode <i>race</i> from 8 values to 4 values as follows:	
2.1	Recode <i>race</i> (See Figure 2) IF <i>race</i> is one of: "White", "Black", or "Hispanic", THEN <i>race</i> stays the same ELSE <i>race</i> ="Other"
Step 3. ADD a field for named <i>schoolClass</i> and populate based on <i>lawschool</i> as follows:	
3.1	<i>schoolCategory</i> ="Class One" IF <i>lawschool</i> is one of: "California Western", "Loyola-Los Angeles", "Pepperdine", "McGeorge", "Santa Clara", "Southwestern", "Stanford", "UC Berkeley", "UC Davis", "UC Hastings", "UC Los Angeles", "UC San Diego", "Univ of San Francisco", "USC Law School", "Western State", "Whittier"
3.2	<i>schoolCategory</i> ="Class Two" IF <i>lawschool</i> is one of: "Chapman", "Golden Gate", "Thomas Jefferson", "Boston University", "Columbia", "Duke", "George Washington", "Georgetown", "Harvard", "New York University", "Northwestern", "Tulane", "University of Michigan", "University of Virginia"
	<i>schoolCategory</i> ="Class Three" OTHERWISE.
Step 4. ADD a field named <i>gradPeriod</i> to store the graduation year as a multi-year range	
Step 5. Aggregate graduation year (<i>gradYr</i>) based on school <i>schoolCategory</i> (from step 3) as follows:	
5.1	IF <i>schoolCategory</i> is "Class One", THEN:
	<i>gradPeriod</i> ="1982-1987" IF <i>gradYr</i> is one of: 1982,1983,1984,1985,1986,1987
	<i>gradPeriod</i> ="1988-1990" IF <i>gradYr</i> is one of: 1988,1989,1990
	<i>gradPeriod</i> ="1991-1993" IF <i>gradYr</i> is one of: 1991,1992,1993
	<i>gradPeriod</i> ="1994-1996" IF <i>gradYr</i> is one of: 1994,1995,1996
	<i>gradPeriod</i> ="1997-1999" IF <i>gradYr</i> is one of: 1997,1998,1999
	<i>gradPeriod</i> ="2000-2002" IF <i>gradYr</i> is one of: 2000,2001,2002
	<i>gradPeriod</i> ="2003-2005" IF <i>gradYr</i> is one of: 2003,2004,2005
	<i>gradPeriod</i> ="2006-2008" IF <i>gradYr</i> is one of: 2006,2007,2008
5.2	IF <i>schoolCategory</i> is "Class Two" or "Class Three", THEN:
	<i>gradPeriod</i> ="1982-1990" IF 1982 <= <i>gradYr</i> <= 1990
	<i>gradPeriod</i> ="1991-1999" IF 1991 <= <i>gradYr</i> <= 1999
	<i>gradPeriod</i> ="2000-2008" IF 2000 <= <i>gradYr</i> <= 2008
5.3	DROP <i>gradYr</i> field
Step 6. ADD fields: <i>avgLSAT</i> , <i>quintile82-90</i> , <i>decile91-99</i> , and <i>decile00-08</i>	
Step 7. Additional processing only for test-takers having <i>schoolCategory</i> ="Class Three" schools:	
7.1	Drop test-takers FROM Class Three schools having fewer than 20 test-takers with same <i>gradPeriod</i> (from step 4)
7.2	CREATE a new table "LSAT TABLE" having fields: <i>lawschool</i> , <i>gradPeriod</i> , and <i>meanLSAT</i> and populate with the average LSAT value for each <i>lawschool</i> , <i>gradPeriod</i> pair THEN sort LSAT TABLE by <i>gradPeriod</i> then <i>meanLSAT</i>

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018. <http://techscience.org/a/2018111301>

7.3	IF <i>gradPeriod</i> is "1982-1990", THEN LOOKUP <i>lawschool</i> for <i>gradPeriod</i> in LSAT TABLE to determine quintile of <i>lawschool</i> THEN SET <i>quintile82-90</i> to that quintile value (1, 2, 3, 4 or 5)
7.4	IF <i>gradPeriod</i> is "1991-1999", THEN LOOKUP <i>lawschool</i> for <i>gradPeriod</i> in LSAT TABLE to determine decile of <i>lawschool</i> THEN SET <i>decile91-99</i> to that decile value (1, 2, 3, ..., 8, 9, or 10)
7.5	IF <i>gradPeriod</i> is "2000-2008", THEN LOOKUP <i>lawschool</i> for <i>gradPeriod</i> in LSAT TABLE to determine decile of <i>lawschool</i> THEN SET <i>decile00-08</i> to that decile value (1, 2, 3, ..., 8, 9, or 10)
7.6	SET <i>avgLSAT</i> to the value of <i>meanLSAT</i> for <i>gradPeriod</i> , <i>lawschool</i> from LSAT TABLE
Step 8. Redact <i>race</i> based on cell size ($k < 11$) of those who passed the bar, as follows:	
FOR test-takers having the same values for <i>lawschool</i> , <i>gradPeriod</i> , and <i>result</i> = "Pass", DO:	
8.1	IF the number of test-takers having <i>race</i> = "Black" is less than 11, THEN: SET <i>race</i> = "Under Represented Minority" for Blacks and Hispanics
	IF the number of test-takers having <i>race</i> = "Under Represented Minority" is less than 11, THEN: ERASE <i>race</i> (blank the value out) for these black and Hispanic test-takers
8.2	ELSE IF the number of test-takers having <i>race</i> = "Hispanic" is less than 11, THEN: SET <i>race</i> = "Under Represented Minority" for these Blacks and Hispanics
	IF the number of test-takers having <i>race</i> = "Under Represented Minority" is less than 11, THEN: ERASE <i>race</i> (blank the value out) for these black and Hispanic test-takers
8.3	IF the number of test-takers having <i>race</i> = "Other" is less than 11, THEN: SET <i>race</i> = "White and Other" for these White, Asian, Indian, etc. test-takers
	IF the number of test-takers having <i>race</i> = "White and Other" is less than 11, THEN: ERASE <i>race</i> (blank the value out) for these White, Asian, etc. test-takers
8.4	ELSE IF the number of test-takers having <i>race</i> = "White" is less than 11, THEN: SET <i>race</i> = "White and Other" for Whites, Asians, Indian Sub-continent, etc.
	IF the number of test-takers having <i>race</i> = "White and Other" is less than 11, THEN: ERASE <i>race</i> (blank the value out) for these White, Asian, Indian, etc. test-takers
Step 9. ERASE <i>lawschool</i> (blank the value out) for all test-takers having <i>schoolCategory</i> = "Class Three"	

Figure 3. 11-Anonymity Protocol that is supposed to anonymize the Raw Dataset by producing the 11-Anonymity Dataset having fields described in Figure 4.



(a)

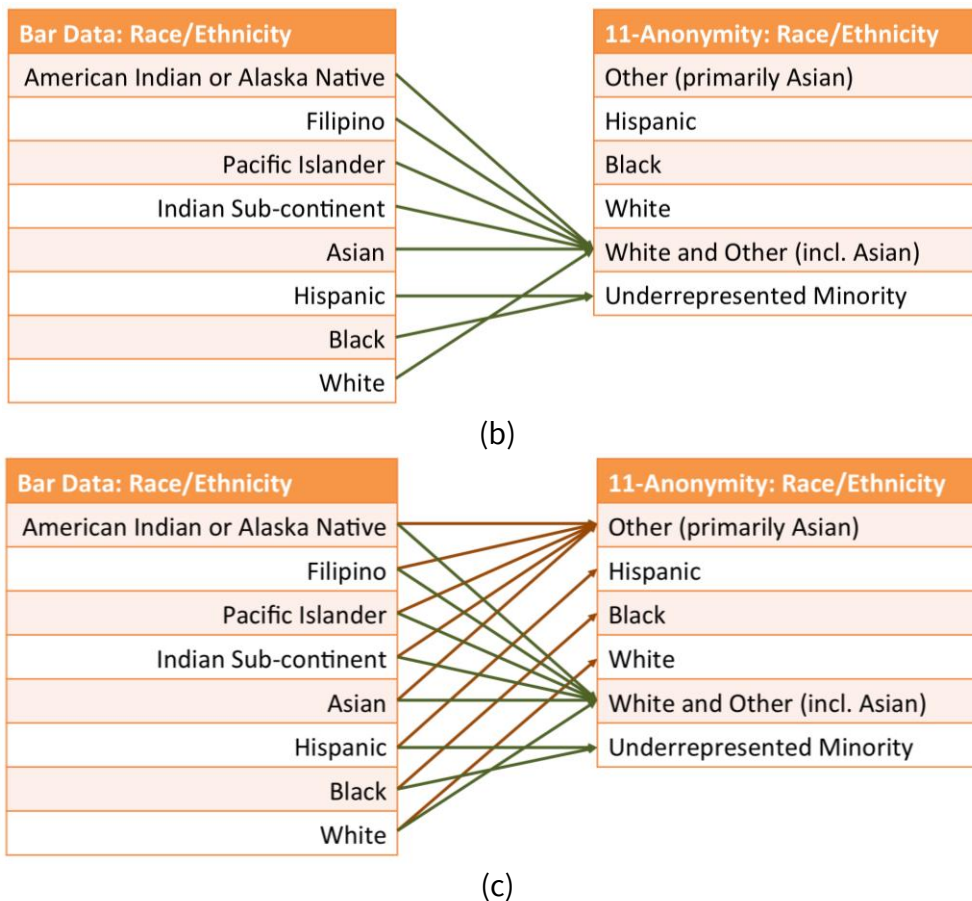


Figure 4. 11-Anonymity Protocol's recoding of race from Bar Data (left) to 11-Anonymity Protocol (right). (a) Reduces the number of race and ethnicity values from 8 (left) to 4 in the first steps of the protocol (see Step 2 in Figure 2). Then, in (b) recoding of race for less popular schools ("Class Three" schools), increases race values from 4 to 6. Race value "Underrepresented Minority" is Black or Hispanic, and "White and Other" is White, Asian, Indian Sub-continent, American Indian, Alaska Native, Filipino or Pacific Islander. (c) The original 4 values were distinct; however, the final 6 values overlap. For example, a Black test-taker could appear as "Black" or as an "Under Represented Minority." Similarly, an Asian test-taker could appear as "Other" or as "White and Other."

Field Name	Field Description
<i>recnum</i>	Unique record number for this study
<i>lawschool</i>	Name of law school (erased in some cases)
<i>gradPeriod</i>	Graduation in a 3, 6 or 9 year range
<i>LSAT</i>	LSAT score (10-48 scale or 120-180 scale)
<i>GPA</i>	Law School GPA (different scales possible)
<i>race</i>	Race/Ethnicity (6 overlapping values)
<i>result</i>	Passed ("Pass") or not pass ("NotPass")

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018.
<http://techscience.org/a/2018111301>

Field Name	Field Description
<i>tries</i>	Passed after multiple tries ("Multi" or blank)
<i>schoolCategory</i>	School popularity level (One, Two or Three)
<i>avgLSAT</i>	Average LSAT for original school and <i>gradPeriod</i>
<i>quintile82-90</i>	LSAT quintile among Class Three school's 1982-90 LSATs
<i>decile91-99</i>	LSAT decile among Class Three school's, 1991-99 LSATs
<i>decile00-08</i>	LSAT decile among Class Three school's 2000-08 LSATs
<i>scores*</i>	List of scores by area of the exam

Figure 5. Fields of 11-Anonymity Dataset as produced by the 11-Anonymity Protocol (Figure 2).

<i>rec num</i>	<i>lawschool</i>	<i>grad Period</i>	<i>LSAT</i>	<i>GPA</i>	<i>race</i>	<i>result</i>	<i>tries</i>	<i>school Category</i>	<i>avg LSAT</i>	<i>decile 00-08</i>	...
1001	Whittier	1997-1999	141	91.78	White	Pass	Multi	Class One			...
1002	Whittier	1997-1999	128	85.09	Other	Pass		Class One			...
1003	Whittier	1997-1999	132	70.36	Other	Pass		Class One			...
1004	Whittier	1997-1999	134	70.36	Hispanic	Pass	Multi	Class One			...
1005	Pepperdine	1997-1999	143	70.59	White	Pass		Class One			...
1006	Pepperdine	2000-2002	132	92.65	White	Pass	Multi	Class One			...
1007	Pepperdine	2000-2002	144	84.2	White	Pass	Multi	Class One			...
1008	Pepperdine	2000-2002	148	67.45	White	Pass		Class One			...
1009	Boston U	2000-2008	141	98.65	White	Pass	Multi	Class Two			...
1010	Boston U	2000-2008	148	67.51	White	Pass		Class Two			...
1011	Boston U	2000-2008	151	70.94	Black	Pass	Multi	Class Two			...
1012	Boston U	2000-2008	141	70.94	Black	Pass	Multi	Class Two			...
1013	Pace	2000-2008	161	84.15	Hispanic	Pass		Class Three	159	7	...
1014	Regent	2000-2008	163	70.36	Black	Pass	Multi	Class Three	153	6	...
1015	Southland	2000-2008	151	70.59	Other	Pass		Class Three	164	9	...
1016	New York	2000-2008	136	81.75	White	Pass	Multi	Class Three	142	1	...
1017	South Bay	2000-2008	137	70.94	White	Pass		Class Three	145	3	...

<i>rec num</i>	<i>lawschool</i>	<i>grad Period</i>	<i>LSAT</i>	<i>GPA</i>	<i>race</i>	<i>result</i>	<i>tries</i>	<i>school Category</i>	<i>avg LSAT</i>	<i>decile 00-08</i>	...
1018	Central	2000-2008	138	86.9	White	Pass	Multi	Class Three	142	1	...
1019	Valley	2000-2008	139	85	Other	Pass		Class Three	145	3	...
1020	Drake	2000-2008	139	80.38	Other	Pass	Multi	Class Three	149	5	...
1021	Stanford	2000-2002	136	80.2	Other	Pass		Class One			...
1022	Stanford	2000-2002	157	82	Other	Pass		Class One			...
1023	Stanford	2000-2002	148	82.21	Other	Pass	Multi	Class One			...
1024	Stanford	2000-2002	158	80.37	Hispanic	Pass	Multi	Class One			...
...

Figure 6. Excerpt of the first rows of interim data from the 11-Anonymity Dataset after executing the first 7 steps of the 11-Anonymity Protocol (Figure 3, Steps 1-7) on the hypothetical excerpt of the Bar Dataset (Figure 2). Changed content is outlined. Fields not shown are *quintile82-90* and *decile91-99* because they have no values, and *scores*, which is mentioned merely as a reminder that the final data would contain the actual bar scores. The *recnum* field only appears for the reader's benefit to track records across protocols.

In step 8, the 11-Anonymity Protocol recodes or redacts *race* values for Class Three schools, as the protocol deems appropriate. To understand the instructions in step 8, we have to consider other records beyond those that appear in Figure 6.

For example, Figure 7a displays hypothetical counts of 55 test-takers who graduated from Pace in 2000-2008. There are 25 White, 6 Black, 12 Hispanic, and 12 "Other" test-takers. One of the Hispanic test-takers is listed in Figure 6 (*regnum*=1013). The other 54 test-takers do not appear in the excerpt displayed in Figure 6. Because the number of Black test-takers is less than 11, and the total number of Black and Hispanic test-takers is 18, which is greater than 11, the protocol changes *race* for these test-takers to "Under Represented Minority" (Figure 7b). Figure 8 shows the change to *regnum*=1013 in the 11-Anonymity Dataset; *race* is now "Under Represented Minority" or "URM," which means Black or Hispanic in this protocol.

Figure 7a also displays hypothetical counts of 47 test-takers who graduated from Regent in 2000-2008. There are 27 White, 2 Black, 4 Hispanic, and 14 "Other" test-takers. One of the Black test-takers is listed in Figure 6 (*regnum*=1014). Because the number of Black test-takers is less than 11, and the total number of Black and Hispanic test-takers is also less than 11, the 11-Anonymity Protocol erases (or blanks out) the *race* value for these test-takers (Figure 7b). Figure 8 shows the change to *regnum*=1014 in the 11-Anonymity Dataset; *race* is now blank.

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018.
<http://techscience.org/a/2018111301>

Race recoding and redacting also works the same way for Whites and "Other." Figure 7a displays hypothetical counts of 63 test-takers who graduated from Southland in 2000-2008. There are 31 White, 12 Black, 14 Hispanic, and 6 "Other" test-takers. One of the "Other" test-takers is listed as Asian in Figure 2 and as "Other" in Figure 6 (*regnum*=1015). The other test-takers do not appear in the excerpt displayed in Figure 2 and Figure 6. Because the number of "Other" test-takers is less than 11, and the total number of "Other" and White test-takers is greater than 11, we change *race* for these test-takers to "White and Other" (Figure 7b). Figure 8 shows the change to *regnum*=1015 in the 11-Anonymity Dataset; *race* is now "White and Other" or "White&," which in this protocol means White, Asian, Indian Sub-continent, American Indian, Alaska Native, Filipino or Pacific Islander.

Here is the last *race* consideration. Figure 7a displays counts of 63 test-takers who graduated from New York in 2000-2008. There are only 7 White and 2 "Other" test-takers. One of the White test-takers is listed in Figure 2 and Figure 6 (*regnum*=1016). Because the total number of White and "Other" test-takers is less than 11, the protocol erases *race* for these test-takers (Figure 7b). Figure 8 shows the change to *regnum*=1016 in the 11-Anonymity Dataset; *race* is now blank.

In our hypothetical examples in Figure 6, the other Class Three schools – South Bay, Central, Valley, and Drake – have at least 11 occurrences of Black, Hispanic, White, and "Other" test-takers, so no changes to *race* occur for these.

Finally, in Step 9, the 11-Anonymity Protocol erases the names of all Class Three schools. Figure 8 shows the final excerpt of the 11-Anonymity Dataset based on the excerpt of the Bar Dataset in Figure 2.

<i>lawschool</i>	<i>gradPeriod</i>	<i>race="White"</i>	<i>race="Black"</i>	<i>race="Hispanic"</i>	<i>race="Other"</i>
Pace	2000-2008	25	6	12	12
Regent	2000-2008	27	2	4	14
Southland	2000-2008	31	12	14	6
New York	2000-2008	7	23	31	2

(a)

<i>lawschool</i>	<i>grad Period</i>	<i>race="White"</i>	<i>race="Black"</i>	<i>race="Hispanic"</i>	<i>race="Other"</i>
Pace	2000-2008	25	"Under Represented Minority" (URM)	"Under Represented Minority" (URM)	12
Regent	2000-2008	27			14
Southland	2000-2008	"White and Other" (White&)	12	14	"White and Other" (White&)
New York	2000-2008		23	31	

(b)

Figure 7. Examples of race redaction and recoding by the 11-Anonymity Protocol (Figure 3). (a) The counts by race of test-takers in the same school and graduation period and (b) redactions based on those counts. The value for *race* changed to "Under Represented Minority" for the 3 Black and 12 Hispanic test-takers who graduated from Pace in 2000-2008 and is blanked out for the 2 Black and to 4 Hispanic test-takers who graduated from Regent in 2000-2008. Similarly, the value for *race* changed to "White and Other" for the 31 White and 6 "Other" test-takers who graduated from Southland in 2000-2008 and is blanked out for the 7 White and to 2 "Other" test-takers who graduated from New York in 2000-2008.

<i>rec num</i>	<i>lawschool</i>	<i>grad Period</i>	<i>LSAT</i>	<i>GPA</i>	<i>race</i>	<i>result</i>	<i>tries</i>	<i>school Category</i>	<i>avg LSAT</i>	<i>decile 00-08</i>	...
1001	Whittier	1997-1999	141	91.78	White	Pass	Multi	Class One			...
1002	Whittier	1997-1999	128	85.09	Other	Pass		Class One			...
1003	Whittier	1997-1999	132	70.36	Other	Pass		Class One			...
1004	Whittier	1997-1999	134	70.36	Hispanic	Pass	Multi	Class One			...
1005	Pepperdine	1997-1999	143	70.59	White	Pass		Class One			...
1006	Pepperdine	2000-2002	132	92.65	White	Pass	Multi	Class One			...
1007	Pepperdine	2000-2002	144	84.2	White	Pass	Multi	Class One			...
1008	Pepperdine	2000-2002	148	67.45	White	Pass		Class One			...

<i>rec num</i>	<i>lawschool</i>	<i>grad Period</i>	<i>LSAT</i>	<i>GPA</i>	<i>race</i>	<i>result</i>	<i>tries</i>	<i>school Category</i>	<i>avg LSAT</i>	<i>decile 00-08</i>	...
1009	Boston U	2000-2008	141	98.65	White	Pass	Multi	Class Two			...
1010	Boston U	2000-2008	148	67.51	White	Pass		Class Two			...
1011	Boston U	2000-2008	151	70.94	Black	Pass	Multi	Class Two			...
1012	Boston U	2000-2008	141	70.94	Black	Pass	Multi	Class Two			...
1013		2000-2008	161	84.15	URM	Pass		Class Three	159	7	...
1014		2000-2008	163	70.36		Pass	Multi	Class Three	153	6	...
1015		2000-2008	151	70.59	White&	Pass		Class Three	164	9	...
1016		2000-2008	136	81.75		Pass	Multi	Class Three	142	1	...
1017		2000-2008	137	70.94	White	Pass		Class Three	145	3	...
1018		2000-2008	138	86.9	White	Pass	Multi	Class Three	142	1	...
1019		2000-2008	139	85	Other	Pass		Class Three	145	3	...
1020		2000-2008	139	80.38	Other	Pass	Multi	Class Three	149	5	...
1021	Stanford	2000-2002	136	80.2	Other	Pass		Class One			...
1022	Stanford	2000-2002	157	82	Other	Pass		Class One			...
1023	Stanford	2000-2002	148	82.21	Other	Pass	Multi	Class One			...
1024	Stanford	2000-2002	158	80.37	Hispanic	Pass	Multi	Class One			...
...

Figure 8. Excerpt of the first rows of the 11-Anonymity Dataset as produced by the 11-Anonymity Protocol (Figure 3) operating on the hypothetical excerpt of the Bar Dataset (Figure 2). Changed content is outlined in boxes. Fields not shown are *quintile82-90* and *decile91-99* because they have no values, and *scores*, which is mentioned merely as a reminder that the final data would contain the actual bar scores. The *recnum* field only appears for the reader's benefit to track records across protocols. Race value "URM" is Black or Hispanic, and "White&" is White, Asian, Indian sub-continent, American Indian, Alaska Native, Filipino, or Pacific Islander.

The actual 11-Anonymity Dataset provided by the Sander Team had 129,984 records, a 7 percent drop in the number of records from the original Bar Dataset (139,338).

Clearly, the Sander Team made a lot of decisions about what to keep and what to change in the 11-Anonymity Dataset. One important point to make here is that these protocols were not

developed on an *a priori* basis, but only after the Team was given restricted access to the underlying data for purposes of the litigation. In a non-litigation setting, and even in most public records litigation, an individual seeking data would not be given access to the private data in order to reverse-engineer an anonymization method. Yet, despite this unusual level of access, do their decisions actually protect privacy? They claimed that the data adhered to *k*-anonymity where *k* is 11. Does it? Before we test to find out, we introduce their other protocols.

Plus Protocol

The Plus Protocol begins where the 11-Anonymity Protocol ends. It makes further changes to law school names and GPA scores to further make values less specific. In step 1, the Plus Protocol performs the same instructions as the 11-Anonymity Protocol (Figure 3) performed, as described above. Then, the Plus Protocol randomly selects 25 percent of the test-takers and erases the name of the law school from those selected. In some cases, the law school name may already be blank; if so, it remains blank. Otherwise, it becomes blank.

In step 3, the Plus Protocol makes grade point averages less precise by rounding *gpa* to one decimal place for GPAs calculated on a 4- or 5-point scale and to whole numbers for GPAs on a 100-point scale.

In the final step, the Plus Protocol erases uniquely occurring *gpa* values in "Group One" and "Group Two" schools for each cohort having the same *lawschool*, *gradPeriod*, and *race*. The final result is a modification to the 11-Anonymity Dataset having the same fields but less information in the *lawschool* and *gpa* fields.

Step 1. Execute the 11-Anonymity Protocol (Figure 3)
Step 2. Redact 25 percent of the school names as follows:
2.1 ERASE <i>lawschool</i> in 25 percent of the records, randomly selected
Step 3. Reduce the scale of GPAs (reduce digits), as follows:
3.1 IF <i>gpa</i> is on 4 or 5 point scale, THEN: ROUND <i>gpa</i> to one decimal place (e.g., 3.18 becomes 3.2)
3.2 ELSE IF <i>gpa</i> is on 100 point scale, THEN: ROUND <i>gpa</i> to whole number (e.g., 87.6 becomes 88)
Step 4. Redact unique <i>gpa</i> values for the same <i>lawschool</i> , <i>gradPeriod</i> , and <i>race</i> as follows:
FOR test-takers having the same values for <i>lawschool</i> , <i>gradPeriod</i> , <i>race</i> , DO:
4.1 ERASE each unique <i>gpa</i>

Figure 9. Plus Protocol to purportedly anonymize the Bar Dataset by producing the Plus Dataset having fields described in Figure 4.

Figure 10 displays an example of applying the Plus Protocol to the excerpt of hypothetical data from the Bar Dataset (Figure 2). Because the first step of the Plus Protocol is the same as the 11-Anonymity Protocol, we examine the differences between the excerpt of hypothetical

values for the 11-Anonymity Dataset (Figure 8) and the excerpt of hypothetical values for the Plus Dataset (Figure 10). An additional 6 (or 25 percent) of the school names are blanked out. All the GPAs are now rounded whole numbers. The GPA for *recnum*=1010 is blanked out. To understand how it got erased, we have to examine the records of all the hypothetical White test-takers at Boston University who graduated in 2000-2008, of which *recnum*=1010 is one. Figure 11 shows the records for these hypothetical 20 test-takers. Three of the 20 test-takers have unique GPA values, so they are blanked out. Among these is *recnum*=1010, which is why in the excerpt in Figure 10, *recnum*=1010 has no *gpa* value.

The Plus Dataset produced by the application of the Plus Protocol by the Sander Team on the Bar Dataset had 98,932 records, a 29 percent drop in the number of records from the original Bar Dataset (139,338).

<i>rec num</i>	<i>lawschool</i>	<i>grad Period</i>	<i>LSAT</i>	<i>GPA</i>	<i>race</i>	<i>result</i>	<i>tries</i>	<i>school Category</i>	<i>avg LSAT</i>	<i>decile 00-08</i>	...
1001	Whittier	1997-1999	141	92	White	Pass	Multi	Class One			...
1002	Whittier	1997-1999	128	85	Other	Pass		Class One			...
1003	Whittier	1997-1999	132	70	Other	Pass		Class One			...
1004		1997-1999	134	70	Hispanic	Pass	Multi	Class One			...
1005	Pepperdine	1997-1999	143	71	White	Pass		Class One			...
1006	Pepperdine	2000-2002	132	93	White	Pass	Multi	Class One			...
1007	Pepperdine	2000-2002	144	84	White	Pass	Multi	Class One			...
1008	Pepperdine	2000-2002	148	67	White	Pass		Class One			...
1009		2000-2008	141	99	White	Pass	Multi	Class Two			...
1010	Boston U	2000-2008	148		White	Pass		Class Two			...
1011	Boston U	2000-2008	151	71	Black	Pass	Multi	Class Two			...
1012	Boston U	2000-2008	141	71	Black	Pass	Multi	Class Two			...
1013		2000-2008	161	84	URM	Pass		Class Three	159	7	...
1014		2000-2008	163	70		Pass	Multi	Class Three	153	6	...
1015		2000-2008	151	71	White&	Pass		Class Three	164	9	...
1016		2000-2008	136	82		Pass	Multi	Class Three	142	1	...
1017		2000-2008	137	71	White	Pass		Class Three	145	3	...
1018		2000-2008	138	87	White	Pass	Multi	Class Three	142	1	...

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018.
<http://techscience.org/a/2018111301>

<i>rec num</i>	<i>lawschool</i>	<i>grad Period</i>	<i>LSAT</i>	<i>GPA</i>	<i>race</i>	<i>result</i>	<i>tries</i>	<i>school Category</i>	<i>avg LSAT</i>	<i>decile 00-08</i>	...
1019		2000-2008	139	85	Other	Pass		Class Three	145	3	...
1020		2000-2008	139	80	Other	Pass	Multi	Class Three	149	5	...
1021		2000-2002	136	80	Other	Pass		Class One			...
1022	Stanford	2000-2002	157	82	Other	Pass		Class One			...
1023		2000-2002	148	82	Other	Pass	Multi	Class One			...
1024	Stanford	2000-2002	158	80	Hispanic	Pass	Multi	Class One			...
...

Figure 10. Excerpt of the first rows of the Plus Dataset as produced by the Plus Protocol (Figure 9) operating on the hypothetical excerpt of the Bar Dataset shown in Figure 2. Changed content from the 11-Anonymity Dataset (Figure 8) is outlined. Fields not shown are *quintile82-90* and *decile91-99* because they have no values, and *scores*, mentioned as a reminder that the final data would contain the actual bar scores. The *recnum* field only appears for the reader's benefit to track records across protocols. Race value "URM" is Black or Hispanic, and "White&" is White, Asian, Indian sub-continent, American Indian, Alaska Native, Filipino, or Pacific Islander.

<i>recnum</i>	<i>lawschool</i>	<i>grad Period</i>	<i>LSAT</i>	<i>GPA</i>	<i>race</i>	<i>result</i>	<i>tries</i>	<i>school Category</i>	...
1010	Boston U	2000-2008	148	68	White	Pass		Class Two	...
1179	Boston U	2000-2008	174	70	White	Pass		Class Two	...
9458	Boston U	2000-2008	156	70	White	Pass	Multi	Class Two	...
8721	Boston U	2000-2008	165	71	White	Pass	Multi	Class Two	...
6351	Boston U	2000-2008	173	93	White	Pass		Class Two	...
4021	Boston U	2000-2008	120	84	White	Pass		Class Two	...
2765	Boston U	2000-2008	165	67	White	Pass	Multi	Class Two	...
5799	Boston U	2000-2008	156	93	White	Pass		Class Two	...
1774	Boston U	2000-2008	149	67	White	Pass		Class Two	...
6202	Boston U	2000-2008	130	71	White	Pass		Class Two	...
5240	Boston U	2000-2008	123	71	White	Pass	Multi	Class Two	...
7259	Boston U	2000-2008	132	84	White	Pass	Multi	Class Two	...
1346	Boston U	2000-2008	150	70	White	Pass		Class Two	...
1553	Boston U	2000-2008	144	71	White	Pass		Class Two	...
5334	Boston U	2000-2008	143	82	White	Pass	Multi	Class Two	...
5300	Boston U	2000-2008	120	71	White	Pass		Class Two	...
8781	Boston U	2000-2008	141	87	White	Pass		Class Two	...
3379	Boston U	2000-2008	131	87	White	Pass		Class Two	...

3790	Boston U	2000-2008	128	80	White	Pass	Multi	Class Two	...
9265	Boston U	2000-2008	173	87	White	Pass		Class Two	...

Figure 11. Excerpt of hypothetical White test-takers from Boston University graduating in 2000-2008. These represent the totality of such test-takers in the dataset after completing the first 3 steps of the Plus Protocol (Figure 9). Three *gpa* values are unique (68 for *recnum*=1010, 82 for *recnum*=5334, and 80 for *recnum*=3790). In step 4 of the Plus Protocol, these 3 GPA values are erased or blanked out.

Enclave Protocol

The Enclave Protocol does not make as many sweeping changes to the Bar Dataset as did the prior two protocols. The idea is to transfer some of the privacy protection from being imposed on the dataset to being imposed on the environment in which the dataset is located. In the other protocols, once the dataset is constructed, it can be shared widely. The Enclave approach is different. Once the dataset is constructed, it is only available in a secure, sequestered physical room ("safe room"). This protocol assumes either that the public entity possessing the data agrees to create such a safe room or that a court compels that creation. We are unaware of any example of a data enclave created by court order in the United States, and the Sander Team could not to present an example at the trial.

The Enclave Protocol has both a protocol for producing a dataset and a protocol for the physical location of the data. We first describe the protocol for the physical location and then detail the steps necessary to produce the dataset that would be in the physical location.

The physical requirements for the Enclave Protocol are straightforward. Visitors to the safe room may access the Enclave Dataset only using the supplied computers and printers. The computers in the safe room contain the Stata program and word processing software, as well as the Enclave Dataset and freely available storage space. Recall, the lawsuit is not about what data only Professor Sander should receive or view for the purpose of his studies, but what data anyone in the public should be able to receive or view.

Visitors cannot bring electronic devices into the safe room. Visitors can bring in paper, pens, and manuals about the dataset. The safe room has a human operator who is responsible for maintaining the security of the room while visitors are present. The human operator also physically inspects all materials and printouts leaving the room. In particular, the Sander Team specifies that the only acceptable materials leaving the safe room are:

- copies of programs used, provided they do not contain any data values;
- regressions of the data, provided each regression is based on at least 400 observations and contains no more than 50 independent variables, and any dummy

variables that represent test-takers must contain at least 20 test-takers. This is not a k restriction but just a limit on what gross computations can leave the enclave; and,

- cross-tabulations of the data provided that no cell in a table has fewer than a count of 20 test-takers.

The operator would review all these types of outputs to make sure that any material that leaves complies with the requirements above.

To produce the Enclave Dataset for use in the safe room, the Enclave Protocol consists of a subset of the steps from the 11-Anonymity Protocol. Figure 12 enumerates the steps taken to construct the Enclave Dataset. In the first steps, the Enclave Protocol drops test-takers who graduated before 1982 and those attending more than one law school or an unaccredited or correspondence school. This is the same as was done in the first step of the 11-Anonymity Protocol; however, the protocol drops test-takers even if they are missing both LSAT and GPA scores.

In step 2, the Enclave Protocol recodes *race* from the original 8 values in the Bar Dataset to the same 4 values used in the 11-Anonymity Protocol – namely, White, Black, Hispanic, and Other. It also labels the schools with the same “Class One,” “Class Two,” and “Class Three” designations used in the 11-Anonymity Protocol to divide schools into sets purportedly based on the number of test-takers from the school taking the bar exam. Class One schools have the most test-takers, and Class Three schools should have the fewest. Also, steps 4 and 5 are exactly like those in the 11-Anonymity Protocol. The Enclave Protocol makes a new field (*gradPeriod*) that replaces the graduation year (*gradYear*) with the same 3-, 6-, or 9-year ranges.

Other than including test-takers for whom the data has no LSAT and GPA scores, the first steps of the Enclave Protocol are the same as those of the 11-Anonymity Protocol. Then some differences occur. At step 6, the Enclave Protocol drops all records for which there are not at least 10 test-takers having the same *lawSchool* and *gradPeriod*.

Lastly, in step 7, the Enclave Protocol merges the *race* values of Black and Hispanic test-takers and of White, Asian, Indian, and other test-takers if there are few of them. This is the same as in step 8 of the 11-Anonymity Protocol. When the sums of these test-takers were less than 11, the 11-Anonymity Protocol either replaced their *race* values with “Under Represented Minority” or “White and Other” or erased the *race* value altogether. The Enclave Protocol does the same in its step 7, except the threshold is 5 instead of 11.

Specifically, if test-takers having the same Class Three school and graduation period include fewer than 5 Blacks or 5 Hispanics, then if the sum of the two is 5 or more, the Enclave Protocol changes the race of those Black and Hispanic test-takers to “Under Represented Minority.” On the other hand, if even when the numbers of Blacks and Hispanics are

combined, the sum is still not at least 5, then the protocol blanks out the *race* of those Black and Hispanic test-takers. Similarly, if the set of test-takers having the same Class Three school and graduation period includes less than 5 Whites or 5 with *race*= "Other," then if the sum of the two is 5 or more, the protocol changes the *race* of those White and "Other" test-takers to "White and Other." On the other hand, if even when the numbers of Whites and "Other" are combined, the sum is still not at least 5, then the protocol blanks out the *race* of those White, Asian, Indian Sub-continent, American Indian, Alaska Native, Filipino, and Pacific Islander test-takers.

Executing the Enclave Protocol produces the Enclave Dataset having the fields listed in Figure 13. These are a subset of the fields for the 11-Anonymity Dataset (Figure 5); specifically, the LSAT distribution fields in the 11-Anonymity Dataset are not included in the Enclave Dataset.

Step 1. Preliminary Steps	
1.1	DROP test-takers attending more than one law school
1.2	DROP test-takers who graduated prior to 1982
1.3	DROP test-takers from unaccredited and correspondence schools
Step 2. Recode <i>race</i> from 8 values to 4 values as follows:	
2.1	Recode <i>race</i> (See Figure 2) IF <i>race</i> is one of: "White," "Black", or "Hispanic", THEN <i>race</i> stays the same ELSE <i>race</i> ="Other"
Step 3. ADD a field for named <i>schoolClass</i> and populate based on <i>lawschool</i> as follows:	
3.1	<i>schoolCategory</i> ="Class One" IF <i>lawschool</i> is one of: "California Western", "Loyola-Los Angeles", "Pepperdine", "McGeorge", "Santa Clara", "Southwestern", "Stanford", "UC Berkeley", "UC Davis", "UC Hastings", "UC Los Angeles", "UC San Diego", "University of Southern California", "Western State", "Whittier" <i>schoolCategory</i> ="Class Two" IF <i>lawschool</i> is one of: "Chapman", "Golden Gate", "Thomas Jefferson", "Boston University", "Columbia", "Duke", "George Washington", "Georgetown", "Harvard", "New York University", "Northwestern", "Tulane", "University of Michigan", "University of Virginia" <i>schoolCategory</i> ="Class Three" OTHERWISE.
Step 4. ADD a field named <i>gradPeriod</i> for storing the graduation year a larger time period	
Step 5. Aggregate graduation year (<i>gradYr</i>) based on school <i>schoolCategory</i> (from step 3) as follows:	
5.1	IF <i>schoolCategory</i> is "Class One", THEN: <i>gradPeriod</i> ="1982-1987" IF <i>gradYr</i> is one of: 1982,1983,1984,1985,1986,1987 <i>gradPeriod</i> ="1988-1990" IF <i>gradYr</i> is one of: 1988,1989,1990 <i>gradPeriod</i> ="1991-1993" IF <i>gradYr</i> is one of: 1991,1992,1993 <i>gradPeriod</i> ="1994-1996" IF <i>gradYr</i> is one of: 1994,1995,1996 <i>gradPeriod</i> ="1997-1999" IF <i>gradYr</i> is one of: 1997,1998,1999 <i>gradPeriod</i> ="2000-2002" IF <i>gradYr</i> is one of: 2000,2001,2002 <i>gradPeriod</i> ="2003-2005" IF <i>gradYr</i> is one of: 2003,2004,2005 <i>gradPeriod</i> ="2006-2008" IF <i>gradYr</i> is one of: 2006,2007,2008
5.2	IF <i>schoolCategory</i> is "Class Two" or "Class Three", THEN: <i>gradPeriod</i> ="1982-1990" IF 1982 <= <i>gradYr</i> <= 1990 <i>gradPeriod</i> ="1991-1999" IF 1991 <= <i>gradYr</i> <= 1999 <i>gradPeriod</i> ="2000-2008" IF 2000 <= <i>gradYr</i> <= 2008

5.3	DROP <i>gradYrfield</i>
Step 6. DROP all records for which there are not at least 10 test-takers having same <i>lawschool</i> , <i>gradPeriod</i>	
Step 7. Redact <i>race</i> based on cell size ($k < 5$) of those who passed the bar, as follows:	
FOR test-takers having the same values for <i>lawschool</i> , <i>gradPeriod</i> , and <i>result</i> = "Pass", DO:	
7.1	IF the number of test-takers having <i>race</i> ="Black" is less than 5, THEN: SET <i>race</i> = "Under Represented Minority" for Blacks and Hispanics IF the number of test-takers having <i>race</i> ="Under Represented Minority" is less than 5, THEN: ERASE <i>race</i> (blank the value out) for these black and Hispanic test-takers
7.2	ELSE IF the number of test-takers having <i>race</i> ="Hispanic" is less than 5, THEN: SET <i>race</i> = "Under Represented Minority" for these Blacks and Hispanics IF the number of test-takers having <i>race</i> ="Under Represented Minority" is less than 5, THEN: ERASE <i>race</i> (blank the value out) for these black and Hispanic test-takers
7.3	IF the number of test-takers having <i>race</i> ="Other" is less than 5, THEN: SET <i>race</i> = "White and Other" for these White, Asian, Indian, etc. test-takers IF the number of test-takers having <i>race</i> ="White and Other" is less than 5, THEN: ERASE <i>race</i> (blank the value out) for these White, Asian, etc. test-takers
7.4	ELSE IF the number of test-takers having <i>race</i> ="White" is less than 5, THEN: SET <i>race</i> = "White and Other" for Whites, Asians, Indian Sub-continent, etc. IF the number of test-takers having <i>race</i> ="White and Other" is less than 5, THEN: ERASE <i>race</i> (blank the value out) for these White, Asian, Indian, etc. test-takers

Figure 12. Enclave Protocol to purportedly anonymize the Raw Dataset for use in a physical safe room. The resulting Enclave Dataset has fields described in Figure 13. The Enclave Dataset is used within a private room on private computers with visual inspection of materials that leave the room.

Field Name	Field Description
<i>recnum</i>	Unique record number for this study
<i>lawschool</i>	Name of law school (erased in some cases)
<i>gradPeriod</i>	Graduation in a 3, 6 or 9 year range
<i>LSAT</i>	LSAT score (10-48 scale or 120-180 scale)
<i>GPA</i>	Law School GPA (different scales possible)
<i>race</i>	Race/Ethnicity (6 overlapping values)
<i>result</i>	Passed ("Pass") or not pass ("NotPass")
<i>tries</i>	Passed after multiple tries ("Multi" or blank)
<i>schoolCategory</i>	School popularity level (One, Two or Three)
<i>scores*</i>	List of scores by area of the exam

Figure 13. Fields of Enclave Dataset as produced by the Enclave Protocol (Figure 12).

Figure 14 shows the result of executing the Enclave Protocol on the hypothetical excerpt of records of the Bar Dataset (Figure 2). The outcome is similar to the result from the 11-Anonymity Protocol (Figure 8) with some notable exceptions. All law school names remain in

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018.
<http://techscience.org/a/2018111301>

the Enclave Dataset, and it has no additional LSAT fields. Changes to *race* for Class Three schools have less redaction in the Enclave Dataset because the threshold dropped from 11 to 5. Specifically, Figure 7 shows the counts of test-takers by race for the schools having the same school and graduation period as the Class Three schools listed in Figure 14 (Enclave Dataset) and Figure 8 (11-Anonymity Dataset). Pace had 6 hypothetical Black and 12 Hispanic test-takers who graduated in 2000-2008. When the threshold was 11, the *race* for these test-takers became "Under Represented Minority" or URM in the 11-Anonymity Data. A threshold of 5 allowed the *race* value for these test-takers to remain unchanged ("Hispanic" for *recnum*=1013) in the Enclave Dataset.

Regent had 2 hypothetical Black and 4 Hispanic test-takers who graduated in 2000-2008. A threshold of 11 led to *race* values for these test-takers being erased (11-Anonymity Data). A threshold of 5 led to *race* being "Under Represented Minority" or URM in the Enclave Data.

Similarly, for Southland and New York records, the lower threshold in the Enclave Protocol allowed *race* for *recnum*=1015 to be "Other" instead of "White and Other," as it was for the Asian test-taker in the 11-Anonymity Dataset. The New York record, *recnum*=1016, is "White and Other" in the Enclave Dataset instead of blank.

The actual Enclave Dataset provided by the Sander Team had 128,659 records, a 7 percent drop in the number of records from the original Bar Dataset (139,338).

<i>rec num</i>	<i>lawschool</i>	<i>gradPeriod</i>	<i>LSAT</i>	<i>GPA</i>	<i>race</i>	<i>result</i>	<i>tries</i>	<i>school Category</i>	<i>scores</i>
1001	Whittier	1997-1999	141	91.78	White	Pass	Multi	Class One	
1002	Whittier	1997-1999	128	85.09	Other	Pass		Class One	
1003	Whittier	1997-1999	132	70.36	Other	Pass		Class One	
1004	Whittier	1997-1999	134	70.36	Hispanic	Pass	Multi	Class One	
1005	Pepperdine	1997-1999	143	70.59	White	Pass		Class One	
1006	Pepperdine	2000-2002	132	92.65	White	Pass	Multi	Class One	
1007	Pepperdine	2000-2002	144	84.2	White	Pass	Multi	Class One	
1008	Pepperdine	2000-2002	148	67.45	White	Pass		Class One	
1009	Boston U	2000-2008	141	98.65	White	Pass	Multi	Class Two	
1010	Boston U	2000-2008	148	67.51	White	Pass		Class Two	
1011	Boston U	2000-2008	151	70.94	Black	Pass	Multi	Class Two	
1012	Boston U	2000-2008	141	70.94	Black	Pass	Multi	Class Two	
1013	Pace	2000-2008	161	84.15	Hispanic	Pass		Class Three	
1014	Regent	2000-2008	163	70.36	URM	Pass	Multi	Class Three	
1015	Southland	2000-2008	151	70.59	Other	Pass		Class Three	
1016	New York	2000-2008	136	81.75	White&	Pass	Multi	Class Three	
1017	South Bay	2000-2008	137	70.94	White	Pass		Class Three	
1018	Central	2000-2008	138	86.9	White	Pass	Multi	Class Three	

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018.

<http://techscience.org/a/2018111301>

1019	Valley	2000-2008	139	85	Other	Pass		Class Three
1020	Drake	2000-2008	139	80.38	Other	Pass	Multi	Class Three
1021	Stanford	2000-2002	136	80.2	Other	Pass		Class One
1022	Stanford	2000-2002	157	82	Other	Pass		Class One
1023	Stanford	2000-2002	148	82.21	Other	Pass	Multi	Class One
1024	Stanford	2000-2002	158	80.37	Hispanic	Pass	Multi	Class One
...

Figure 14. Excerpt of the first rows of the Enclave Dataset as produced by the Enclave Protocol (Figure 12) operating on the hypothetical excerpt of the Bar Dataset (Figure 2). Changed content is outlined. Race value "URM" is Black or Hispanic, and "White&" is White, Asian, Indian sub-continent, American Indian, Alaska Native, Filipino, or Pacific Islander. As a reminder, *scores* would contain the actual bar scores. The *recnum* field only appears for the reader's benefit to track records across protocols.

Standardized Protocol

The final protocol provided by the Sander Team is called the Standardized Protocol. It constructs a new statistical database from the Bar Dataset that reports standardized LSAT and GPA scores by test-taker. School names are removed. Race is one of four values: White, Black, Hispanic, or Other (Asian, Indian sub-continent, American Indian, Alaska Native, Filipino, or Pacific Islander). The original LSAT and GPA scores are also removed. The Sander Team stated that this was the least desired of the protocols because it reduced data utility and that the team preferred the 11-Anonymity Protocol.

Figure 15 lists the steps of the Standardized Protocol. In the first step, the protocol drops test-takers attending more than one law school, those who graduated prior to 1985, and those who attended unaccredited or correspondence schools. Then, in step 2, it performs the unusual step of dropping all test-takers who graduated between 1999 and 2005. It recodes the race of Asian, Indian sub-continent, American Indian, Alaska Native, Filipino, and Pacific Islander test-takers as "Other." The protocol aggregates graduation year into four 3- or 4-year bands: 1985-89, 1990-94, 1995-98, and 2006-08.

In steps 6, 7, and 8, the Standardized Protocol computes standardized LSAT scores. For each test-taker, it stores how many standard deviations the test-taker's *lsat* is from the mean LSAT of all the other test-takers that year. This is an annualized LSAT; it stores these values in the field *zLSATyr*.

In step 8, the Standardized Protocol divides test-takers into groups based on their law school and graduation year. For those groups having 20 or more test-takers, it stores the number of standard deviations the test-taker's *lsat* is from the group's mean and records this in the field *zLSAT*. It also records that this was based on a year of test-takers (*zLSATtype*="1Year"). If

there were not 20 test-takers for the law school and graduation year, then it tries using the 3- or 5-year ranges recorded (*gradPeriod*). If there are at least 20 test-takers having the same law school and graduation time range, the protocol computes and stores *zLSAT* and records that it used multiple years *zGPAType*="3-5Yr" to do so.

In step 9, the Standardized Protocol standardizes GPA scores as it did with the LSAT scores, described above. For those groups having 20 or more test-takers graduating from the same law school in the same year, it stores the number of standard deviations the test-taker's *gpa* is from the mean GPA of the group; it stores these values in the field *zGPA* and records that these values were based on a year of test-takers (*zGPAType*="1Year"). If there were not 20 test-takers, then it tries using the 3- or 5-year ranges recorded (*gradPeriod*). If there are at least 20 test-takers having the same law school and graduation time range, the protocol computes *zGPA* and records that it used multiple years *zGPAType*="3-5Yr" for the computation.

Figure 17 shows an excerpt of what the Standardized Protocol produces on the hypothetical Bar Dataset in Figure 2. Test-takers who graduated between 1999 and 2005, inclusive, were dropped. The names of the law schools and the actual LSAT and GPA scores were dropped. The race of the Asian test-takers was recoded to "Other." The graduation year was replaced with a 3- or 5-year range. Fields added contain standardized values and related information for the test-taker's LSAT and GPA.

The actual Standardized Dataset provided by the Sander Team had 85,364 records, a 39 percent drop in the number of records from the original Bar Dataset (139,338). However, several discrepancies existed between the textual description, the Stata program, and the Standardized Dataset. No two of them agree. We relied on the Stata program as the basis for the algorithm in Figure 15. For this writing, all further references to the Standardized Protocol and the Standardized Dataset will be to the algorithmic description in Figure 15 unless stated otherwise or obvious from context.

Step 1. Preliminary Steps	
1.1	DROP test-takers attending more than one law school
1.2	DROP test-takers who graduated prior to 1985
1.3	DROP test-takers from unaccredited and correspondence schools
1.4	DROP test-takers whose LSAT scores are not within the ranges: 10-48 or 120-180
Step 2. DROP test-takers who graduated between 1999 and 2005, inclusive	
Step 3. Recode <i>race</i> from 8 values to 4 values as follows:	
3.1	Recode <i>race</i> (See Figure 2) IF <i>race</i> is one of: "White," "Black", or "Hispanic", THEN <i>race</i> stays the same ELSE <i>race</i> ="Other"
Step 4. ADD a field named <i>gradPeriod</i> to store the graduation year as a multi-year range	
Step 5. Aggregate graduation year (<i>gradYr</i>) as follows:	
5.1	<i>gradPeriod</i> ="1985-89" IF <i>gradYr</i> is one of: 1985,1986,1987,1988,1989
5.2	<i>gradPeriod</i> ="1990-94" IF <i>gradYr</i> is one of: 1990,1991,1992,1993,1994
5.3	<i>gradPeriod</i> ="1995-98" IF <i>gradYr</i> is one of: 1995,1996,1997,1998

5.4	<i>gradPeriod</i> ="2006-08" IF <i>gradYr</i> is one of: 2006,2007,2008
Step 6. ADD fields to hold standardized LSAT values, namely: <i>zLSATyr</i> , <i>zLSAT</i> , <i>zLSATtype</i>	
Step 7. Produce annualized LSAT scores (<i>zLSATyr</i>), as follows:	
7.1	Remove all records whose <i>lsat</i> is not within 10-48 or 120-180.
7.2	FOR all test-takers having the same <i>gradYr</i> , DO:
7.2.1	COMPUTE the yearly average <i>lsat</i> (<i>meanYr</i>) and standard deviation
7.2.2	SET <i>zLSATyr</i> =number of standard deviations between <i>lsat</i> and <i>meanYr</i> for test-taker
Step 8. Produce standardized LSAT scores (<i>zlsat</i>) for <i>lawschool</i> groupings, as follows:	
8.1	FOR EACH group of at least 20 test-takers having the same <i>lawschool</i> and <i>gradYr</i> DO: COMPUTE the average <i>lsat</i> (<i>mean</i>) and standard deviation for the group SET <i>zLSAT</i> =the number of standard deviations <i>lsat</i> is from <i>mean</i> SET <i>zLSATtype</i> ="1Year"
8.2	FOR EACH group having fewer than 20 test-takers with the same <i>lawschool</i> and <i>gradYr</i> , DO: IF the group has at least 20 test-takers with the same <i>lawschool</i> and <i>gradPeriod</i> THEN: COMPUTE the average <i>lsat</i> (<i>mean</i>) and standard deviation for the group SET <i>zLSAT</i> =the number of standard deviations <i>lsat</i> is from <i>mean</i> SET <i>zLSATtype</i> ="3-5Yr"
8.3	ELSE: SET <i>zLSAT</i> to blank SET <i>zLSATtype</i> to blank
Step 9. ADD fields to hold standardized GPA values, namely: <i>zGPA</i> , <i>zGPAtype</i>	
Step 10. Produce standardized GPA scores (<i>zGPA</i>) for <i>lawschool</i> groupings, as follows:	
10.1	FOR EACH group of at least 20 test-takers having the same <i>lawschool</i> and <i>gradYr</i> , DO: COMPUTE the average <i>gpa</i> (<i>mean</i>) and standard deviation for the group SET <i>zGPA</i> =the number of standard deviations <i>gpa</i> is from <i>mean</i> SET <i>zGPAtype</i> ="1Year"
10.2	FOR EACH group having fewer than 20 test-takers with the same <i>lawschool</i> and <i>gradYr</i> DO: IF the group has at least 20 test-takers with the same <i>lawschool</i> and <i>gradPeriod</i> THEN: COMPUTE the average <i>gpa</i> (<i>mean</i>) and standard deviation for the group SET <i>zGPA</i> =the number of standard deviations <i>gpa</i> is from <i>mean</i> SET <i>zGPAtype</i> ="3-5Yr"
10.3	ELSE: SET <i>zGPA</i> to blank SET <i>zGPAtype</i> to blank
Step 11. ADD field <i>instate</i> to identify California from non-California schools	
Step 12. Label California and non-California schools, as follows:	
12.1	FOR EACH test-taker from a California school, DO: SET <i>instate</i> to "CA"
12.2	FOR EACH test-taker NOT from a California school, DO: SET <i>instate</i> to blank
Step 13. DROP fields to <i>lawschool</i> , <i>gradYr</i> , <i>lsat</i> , <i>gpa</i>	

Figure 15. Standardized Protocol that is supposed to anonymize the Bar Dataset. The resulting Standardized Dataset has the fields described in Figure 16. Many differences exist between the Stata code for the Standardized Protocol, as provided by the Sander Team, and the textual description of the Standardized Protocol, also provided by the Sander Team. The algorithmic description above defers to the Stata code in cases of ambiguity.

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018. <http://techscience.org/a/2018111301>

Field Name	Field Description
<i>recnum</i>	Unique record number for this study
<i>gradPeriod</i>	Graduation in a 1-, 3-, or 5-year range
<i>race</i>	Race/Ethnicity: White, Black, Hispanic, Other
<i>result</i>	Passed ("Pass") or not pass ("NotPass")
<i>tries</i>	Passed after multiple tries ("Multi" or blank)
<i>zlsat</i>	Standardized LSAT score based on school and <i>zlsatType</i>
<i>zlsatType</i>	Time period for <i>zlsat</i> cohort ("1Year" or "3-5Yr")
<i>zlsatPop</i>	Annual standardized LSAT score for graduation year
<i>zgpa</i>	Standardized LSAT score based on school and <i>zgpaType</i>
<i>zgpaType</i>	Time period for <i>zgpa</i> cohort ("1Year" or "3-5Yr")
<i>instate</i>	"CA" if California school; blank otherwise
<i>scores*</i>	List of scores by area of the exam

Figure 16. Fields of the Standardized Dataset as produced by the Standardized Protocol (Figure 15).

<i>rec num</i>	<i>grad Period</i>	<i>zLSAT</i>	<i>zLSAT yr</i>	<i>zLSAT type</i>	<i>zGPA</i>	<i>zGPA type</i>	<i>race</i>	<i>result</i>	<i>tries</i>	<i>scores</i>
1001	1995-98	0.015	0.076	1Year	-1.041	1Year	White	Pass	Multi	
1002	1995-98	-1.846	-1.594	1Year	0.360	1Year	Other	Pass		
1003	1995-98	-0.985	-1.312	1Year	-0.425	1Year	Other	Pass		
1004	1995-98	-0.229	-1.579	1Year	-1.307	1Year	Hispanic	Pass	Multi	
1005	1995-98	-0.009	-0.076	1Year	0.070	1Year	White	Pass		
1006										
1007										
1008										
1009	2006-08	-1.291	-0.484	1Year	-0.375	1Year	White	Pass	Multi	
1010	2006-08	-1.224	-0.335	1Year	0.900	1Year	White	Pass		
1011	2006-08	1.401	0.058	1Year	-1.657	1Year	Black	Pass	Multi	
1012	2006-08	-0.875	0.395	1Year	-1.138	1Year	Black	Pass	Multi	
1013	2006-08	3.009	1.383	3-5Yr	0.724	3-5Yr	Hispanic	Pass		
1014	2006-08	1.053	0.786	3-5Yr	0.264	3-5Yr	Black	Pass	Multi	
1015	2006-08	0.338	0.582	3-5Yr	-0.830	3-5Yr	Other	Pass		
1016	2006-08	-1.255	-0.799	3-5Yr	-1.664	3-5Yr	White	Pass	Multi	
1017	2006-08	-0.154	-1.182	3-5Yr	-0.192	3-5Yr	White	Pass		
1018	2006-08	-0.054	-0.653	3-5Yr	-0.285	3-5Yr	White	Pass	Multi	
1019	2006-08	1.563	0.320	3-5Yr	1.123	3-5Yr	Other	Pass		
1020	2006-08	-2.449	-1.645	3-5Yr	-0.014	3-5Yr	Other	Pass	Multi	

Sander Team scientifically meet their stated analytical objectives of k -anonymity protection, anonymity, and HIPAA compliance.

The protocols from the Sander Team may seem complex and daunting and capable of adjusting the data enough that no one can be re-identified, but the Sander Team did not provide testing, warranty, or privacy guarantees. The datasets they produced may look anonymous, but just because data looks anonymous does not make data anonymous. We need scientific proof. Proving a dataset offers a guaranteed level of privacy means showing that the dataset maintains its guarantee independent of an attacker. In this paper we do not attempt to demonstrate all vulnerabilities in the proposed protocols. We merely discuss examples of tests we conducted of the Sander Team's assertion that the protocols anonymize the data.

The hypothesis we seek to test is: *the protocols work as promised*. With perfectly working protocols, some things should not be possible, and we test for such things. For example, both analyzing and implementing a protocol should be within the technical know-how and available time of a government office employee. Few government entities have statisticians or data privacy experts available to respond to data requests. So, the protocols should describe actions that can be accomplished by existing staff and that are clearly adequate to achieve the expected privacy-protecting results [13][14][15].

In order to test this hypothesis, we examine whether the protocols withstand some reasonable litmus tests.

For each protocol, this paper discusses the following sequence of litmus tests.

- Litmus Test 1.** Is the construction of the dataset technically reasonable to accomplish by government staff?
- Litmus Test 2.** Is there privacy vulnerability in the resulting dataset? If so, can we develop one or more practical re-identification strategies to demonstrate the vulnerability?
- Litmus Test 3.** If we actually devise practical re-identification strategies in Test 2, then can we demonstrate that at least one of them reliably associates names uniquely or to a small group of records in the protocol's dataset?
- Litmus Test 4.** If we get to Test 4 and actually have some unique or small group re-identifications, then can we describe harms that might result to those who were matched?

If we evaluate a protocol and the answer to each of the four tests is "no," then the protocols withstood our litmus tests (Figure 18). An outcome of all "no's" means our litmus tests did not disprove the hypothesis. That does not, of course, mean there may not exist other re-identification strategies that would disprove the hypothesis (and says nothing about whether compelled production would be legally required). It just means we did not find any scientific evidence to disprove the hypothesis in this study using our litmus tests.

If we evaluate a protocol and the answer to each test is "yes," then we show that small group re-identifications are possible, and we demonstrate personal harm to specific test-takers that could result if the dataset were shared [13][14][15].

In the next subsections, we describe how we operationalize each of our litmus tests.

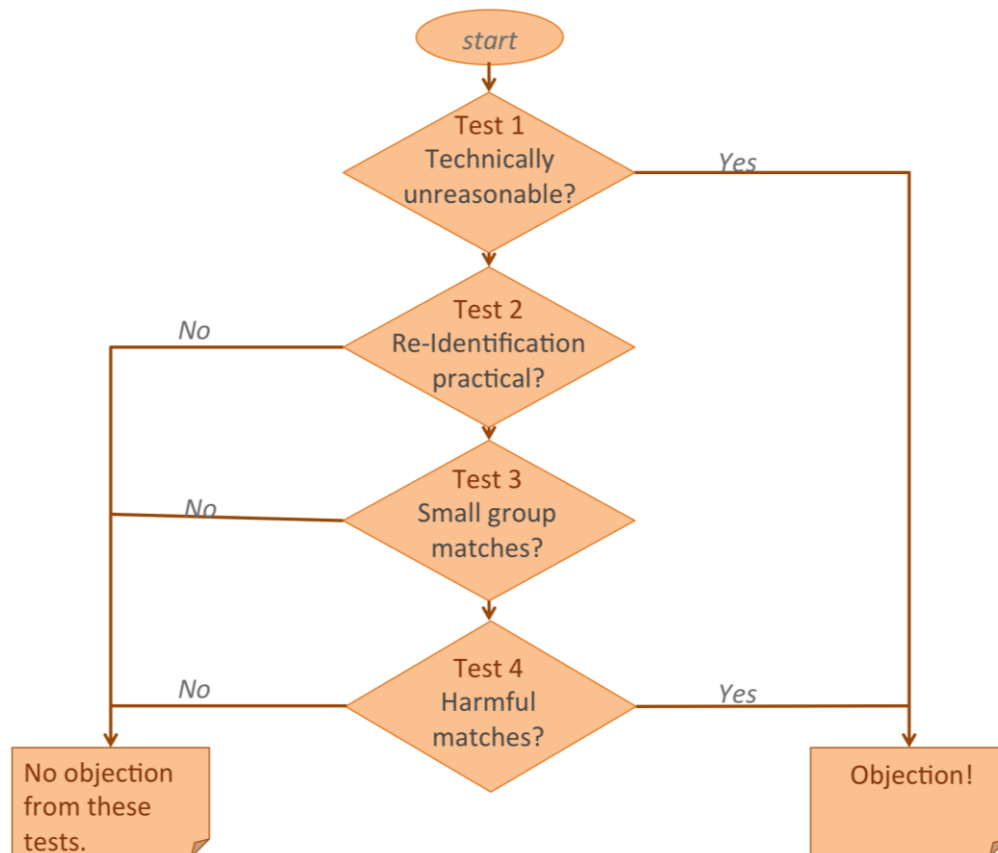


Figure 18. Sequence of 4 litmus tests to perform on a Sander Team protocol. There are two outcomes. A "no" outcome on tests 2, 3, or 4 means the protocol passes our litmus tests ("No objection from these tests."), but this outcome offers no scientifically generalizable proof that the protocol protects privacy. However, if the construction of the dataset is technically unreasonable (Test 1), or if re-identifications exist and are harmful (Test 4), then the outcome is "Objection!", which is an objection based on scientific evidence.

Approach: Litmus Test 1 – Unreasonable

We can end our assessment after Test 1 if we find that the construction of the dataset is technically unreasonable (Figure 18). We do not mean this as a legal question, but rather a technical determination based on the usability of the protocol by government staff. As discussed above, there is a generally accepted legal framework for FOIA and similar public records laws stating that a public record request cannot compel a government agency to create a new database, but only to produce existing records (with redaction where appropriate). As discussed above, the trial court found that none of the Sander Team protocols satisfied that requirement. For this paper, however, we will disregard that concern because the same examples would be relevant in connection with a voluntary disclosure of data. We also assume that the protocols are presented to the government in an acceptable form; our conclusion that implementation of a method is technically reasonable does not imply that creation of the method or testing the efficacy of the method are technically reasonable. To the contrary, development of an anonymization method requires a substantial degree of skill, and ad hoc methods are likely to leave data unprotected.

In this study, we use the popular spreadsheet program Excel to measure the expertise and effort involved in executing a protocol and determine whether a protocol is too burdensome or technically unreasonable for a government agency.

For example, once data are loaded into an Excel spreadsheet, we can use two or three mouse clicks to erase a value from a cell or delete an entire column or row of values. We can also recode information in Excel using nested IF statements. Below is an example of nested IF statements in Excel that would recode a *race* value in cell D3 to be "White," "Black," "Hispanic," or "Other."

```
=IF(D3="White", "White",  
    IF(D3="Black", "Black",  
        IF(D3="Hispanic", "Hispanic", "Other")))
```

We entered "excel tutorial" into the search bar on google.com and found, among the top results, several websites offering online tutorials to learn Excel (e.g., [37][38][39]). These websites listed basic knowledge of Excel as including the following capabilities: working with cells, ranges, formulas, and functions. Advanced knowledge includes: sorting, filtering, making a pivot table, and using lookup and reference functions (e.g., VLOOKUP and INDIRECT). A view of the HELP glossary in Excel includes even more topics beyond the advanced tutorial topics, such as macros. We use the designations of basic and advanced knowledge to determine the level of expertise required to implement a protocol and we use the number of keystrokes involved to measure effort. We do not consider topics beyond advanced tutorial knowledge in this writing.

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018. <http://techscience.org/a/2018111301>

There is no expectation that protocols must be implemented in Excel. A protocol could be implemented using some other program. We use Excel in this paper because Excel files are shared regularly, and the Excel program is available widely. So, we use it to determine the implementation burden of a protocol. In fact, the original protocols provided by the Sander Team were written in Stata, even though there was no evidence that any relevant personnel at the Bar knew how to program in Stata.

Any step of a protocol that only involves basic Excel topics is allowable for purposes of this litmus test. For example, recoding *race* into four values can be done by replicating the IF statement (shown above) down a column; this involves basic Excel knowledge, so it is practicable (whether or not it can be legally compelled).

We also consider for these purposes that a step of a protocol is allowable if it involves a single advanced topic used in a straightforward manner or requires nesting or inter-connecting basic topics with an advanced topic. For example, constructing a simple pivot table of counts (Figure 7a) is a single advanced topic; therefore, the step of the protocol that requires it is allowable. However, a step is "not allowable" if the step involves nesting or inter-connecting advanced topics, because common Excel tutorials consider such expertise to be beyond the average Excel user. For example, constructing a pivot table from pivot table results is considered too complicated for the average Excel user, so it is not allowable here.

Expertise is part of what determines the burden a protocol imposes. The other part is the effort involved. Decades ago, the individual doing the redacting had to read the printed pages. The average single-spaced page contains about 3,000 characters [40] and takes the average individual about 4 minutes to read aloud [41]. We use the "read aloud" time to account for the time needed to decide what needs redacting and to do the redacting, even though current efforts would be aided by a computer program. Our idea here is to ascertain what can be accomplished in the same amount of time a government employee might have used previously. We estimate that in one hour, an individual could redact about 15 pages or 45,000 characters.

Today, a protocol is more likely to require typing than reading. The speed of the "average typist" is 12,000 characters (or keystrokes) per hour [42]. Under the Freedom of Information Act, there is usually no charge for the first 2 hours [43]. We assume our government agency wants to provide the data and is willing to spend twice that time to do so. Therefore, we say that it is "acceptable" for an individual to spend up to 4 hours implementing a protocol, and the implementation may require up to 48,000 keystrokes of typing commands, entering values or clicking mouse buttons.

In summary, we say a protocol is "technically reasonable" if each step in the protocol requires allowable expertise in Excel and the effort involved consumes an acceptable number of keystrokes, including loops and advanced functions. Otherwise, we consider the protocol to be "technically unreasonable" for purposes of this paper

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018. <http://techscience.org/a/2018111301>

One final point bears mentioning. As discussed above, the protocols here were developed by experts in litigation after being provided with restricted access to the data. This allowed an extent of reverse engineering of solutions that would not exist in an ordinary records request (although it may exist with a voluntary disclosure of data). We ignore this limitation for purposes of the following analysis.

Approach: Litmus Test 2 – Practical Re-identification

In Litmus Test 2, we assess whether a protocol has privacy vulnerability; if so, we develop a stepwise re-identification strategy to demonstrate the vulnerability.

The basic idea of a re-identification is to put a name to one or few records in the protocol's dataset. The names have to come from somewhere. With this example, the most useful data to help re-identify records is information about named individuals that includes some of the same fields of information in the Bar Dataset. In the following subsections we provide examples of this kind of information found publicly available online, including commencement programs, attorney license profiles, resumes, bios, alumni lists, law school club memberships, and photographs.

We also describe a set of tools we made using the Python programming language to help us harvest and use these kinds of online information. Afterwards, we describe the steps we took to show privacy vulnerability and craft a re-identification strategy as Litmus Test 2. Notice that the effort requirement shifts in the remaining litmus tests. We are no longer monitoring the time and knowledge of government staff, as we did in Litmus Test 1. Instead, we are assessing the resources available to a recipient of the released data to re-identify the data.

Approach: Litmus Test 2 – Online Information

Here is our walk through relevant online information to introduce the nature and extent of publicly available information. In this analysis, we only consider the vulnerability of the data to re-identification by a stranger. There are other obvious potential attackers who would have much more comprehensive data. For example, a law school administrator may have comprehensive records for graduates of that school. In reporting the availability of information on the Internet, we do not mean to imply that a data privacy professional should only be concerned with re-identification by random strangers using publicly available data. As you will see, however, here even that most remote risk is realized.

Online Commencement Data. Most students graduate from law school in the spring, and several schools make graduation lists and commencement programs available online (e.g., [44][45][46][47]).

Commencement programs list the name of the school, date of graduation, and names of the graduates. They often include graduation honors received and may include photographs of

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018.
<http://techscience.org/a/2018111301>

graduates and their hometowns or undergraduate schools, or the names of student club officers.

Law school graduates earn a Juris Doctor ("J.D."). Figure 19 shows the first page of J.D. graduates in the 2002 commencement program for the Pepperdine University School of Law (Pepperdine) [44]. The first three students listed are Jay Spagnola, Daniel Droog, and Jesse Cripps. All 3 of them graduated summa cum laude on May 17, 2002 (the day of the commencement).

Graduation honors are often designated as "summa cum laude," "magna cum laude," or "cum laude." Often a web page at the school's website describe how honors are determined at the school. For example, at some schools, honor distinctions depend on explicit ranges of GPAs, with the uppermost range designated as summa cum laude, followed by magna cum laude, and so on. At other schools, honor distinctions are based on GPA ranking in the graduating class. For example, at Pepperdine in 2002, summa cum laude is given to students whose GPAs rank in the top 2 percent of the graduating class, magna cum laude is given to the next 5 percent, and cum laude is given to the next 18 percent [48]. At Pepperdine, the student having the highest GPA is the Valedictorian, and the student having the second highest GPA is the Salutatorian. So, among the students in the 2002 graduating class from Pepperdine, Jay Spagnola had the highest GPA, and Daniel Droog had the second highest GPA (Figure 19).

Bar Exam Inference. The California Bar offers its exam in July and February, so a law school student who graduates in the spring has his first opportunity to take the exam about 2 months after graduating law school. If he passes on that first attempt, he can be admitted to the California Bar (assuming all other requirements are met) in approximately November or December, which is 6-7 months after graduating and in the same calendar year as his May graduation. Once admitted, he has a license to practice law as an attorney in the state of California. An individual's admissions date is a matter of public record, while the date or number of times each individual took the bar exam is not.

If an individual does not pass and takes the exam again at the next opportunity, the earliest he could do so is in about 3 months after notification. Therefore, a repeat bar taker may take the exam as often as twice a year.

As described, there is a relationship between number of tries at the bar exam an attorney might have before passing, the date of his graduation from law school, and the date of his admission. The graduation date places an exact earliest date on bar admission (December for a May graduate). Thus, an individual's date of bar admission gives the maximum number of attempts at the bar exam.

For example, the earliest that students who graduated from Pepperdine in May 2002 could have taken the exam is July 2002. Those who passed would have been admitted in November

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018. <http://techscience.org/a/2018111301>

or December 2002. Those who did not pass on the first attempt and decided to take the exam again at the next possible time would have done so in February 2003. Those who passed then would have been admitted in April or May 2003. So, we know that a May 2002 graduate of Pepperdine having a bar admission date of December 2002 passed the bar on her first attempt. A graduate who passed in April 2003 may have made 1 or 2 attempts.

Attorney License Data Online. The State Bar of California maintains a website that allows the public to search for information about attorneys who are members of the California Bar [49]. Members of the public can learn whether a particular individual is admitted to and in good standing with the Bar.

There are numerous ways to acquire data from the State Bar of California's website. One can enter a name using the website's quick search option [49]. The result reports the individual's bar number, current standing, current city, and the date admitted to the Bar. It also provides a link to learn more information about the attorney, and information on that web page includes the law school attended. Figure 20 shows the search result from the State Bar of California's search website for Daniel Droog, one of the summa cum laude graduates of Pepperdine in 2002 mentioned earlier. Attorney Droog was admitted to the California Bar in April 2003, so he may have passed the bar exam after 1 or 2 attempts. In comparison, searches for Jay Spagnola and Jesse Cripps reported them both as being admitted to the California Bar in December 2002, which means they passed the bar on their first attempts.

Online Resumes and Bios. Biographical information about law school graduates and members of the Bar often appear on law office websites, in news articles, and in online resumes at repositories such as LinkedIn.com. Figure 21 shows the online resume at LinkedIn for Daniel Droog [50], who graduated summa cum laude from Pepperdine in 2002 (Figure 19) and passed the bar in April 2003 (Figure 20). His LinkedIn profile includes his photograph, a history of his employment since graduating, and details about his graduation. From his picture, we also learn that he is White. Besides photos, some online resumes include GPA (e.g., Figure 22 and Figure 23) and LSAT scores (e.g., Figure 24 and Figure 25) and sometimes both (e.g., Figure 26).

Online Alumni Lists. School alumni often have publicly available websites. Figure 27 shows a highlight of 6 of the 151 alumni profiles of 2005 graduates of Loyola Law School that are publicly available on the Loyola alumni website [51]. Some profiles include photographs. All profiles include the graduate's name.

Similarly, LinkedIn provides an index of resumes having the same graduation year from a given school. For example, Figure 28 shows an excerpt of 20 of more than 100 LinkedIn profiles for Pepperdine 2002 graduates [52].

Online Club Memberships. Club memberships may also allow racial or ethnic inferences. For example, Figure 29 shows excerpts from the publicly available 2006 Commencement

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018. <http://techscience.org/a/2018111301>

program from Stanford Law School listing officers of the Asian & Pacific Islander Law Students Association, the Black Law Students Association, the Chinese Law Association, the Native American Law Students Association, the South Asian Law Students Association, and the Stanford Latino Law Students Association [53]. Similarly, Figure 30 shows the names of officers of the McGeorge Law School Black Law Students Association [54].

Online Photos. Race can often be inferred from online photographs. Figure 31 shows an assortment of photos of Black graduates from McGeorge Law School, including one of Marcia Randle, who is also listed as Vice-President of the McGeorge Black Law Student Association in 2002-2003 (Figure 30).

Juris Doctor	
JAY PAUL SPAGNOLA <i>Valedictorian</i> <i>summa cum laude</i> <i>Law Review</i> B.A., University of Maryland, College Park	CYNTHIA BURNS McCAUGHEY <i>magna cum laude</i> B.S., University of Redlands
DANIEL D. DROOG <i>Salutatorian</i> <i>summa cum laude</i> <i>Law Review, Honor Board</i> B.A., Dordt College	TYLER CHRISTOPHER NEAL <i>magna cum laude</i> <i>Law Review</i> B.A., Stanford University
JESSE A. CRIPPS, JR. <i>summa cum laude</i> <i>Law Review, Dean's Award</i> B.A., Pepperdine University	JONATHAN BELL RUBENSTEIN <i>magna cum laude</i> <i>Law Review</i> B.A., University of California, Los Angeles
BETH ALICIA NUNNINK <i>summa cum laude</i> <i>Law Review, Dean's Award</i> B.A., Luther College	NICOLLE TAYLOR <i>magna cum laude</i> <i>CETL Certificate, Law Review</i> B.S., Pepperdine University

Figure 19. Excerpt from the program for the 2002 commencement at Pepperdine School of Law, publicly available online. [44]



THE STATE BAR OF CALIFORNIA
Protecting the Public and Enhancing the Administration of Justice

[Home](#) > [Public](#) > [Attorney Search](#) > [Attorney Profile](#)



ATTORNEY SEARCH

Daniel Dale Droog - #224596

Current Status: Active

This member is active and may practice law in California.
See below for more details.

Profile Information

The following information is from the official records of The State Bar of California.

Bar Number:	224596	Phone Number:	(713) 372-9071
Address:	Chevron Upstream and Gas Litigation Management Group 1400 Smith FI 5 Houston, TX 77002	Fax Number:	Not Available
County:	Non-California	e-mail:	Not Available
District:	Outside California	Undergraduate School:	Dordt Coll; Sioux Center IA
Sections:	None	Law School:	Pepperdine Univ SOL; Malibu CA

Status History

Effective Date	Status Change
Present	Active
4/9/2003	Admitted to The State Bar of California

Figure 20. Excerpt from California Bar search results for Daniel Droog, as publicly available online [55]. Profile above shows he graduated from Pepperdine and was admitted to the Bar in 2003.

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018. <http://techscience.org/a/2018111301>

in Search

Daniel D. Droog
Senior Counsel for Chevron Global Upstream & Gas

Experience

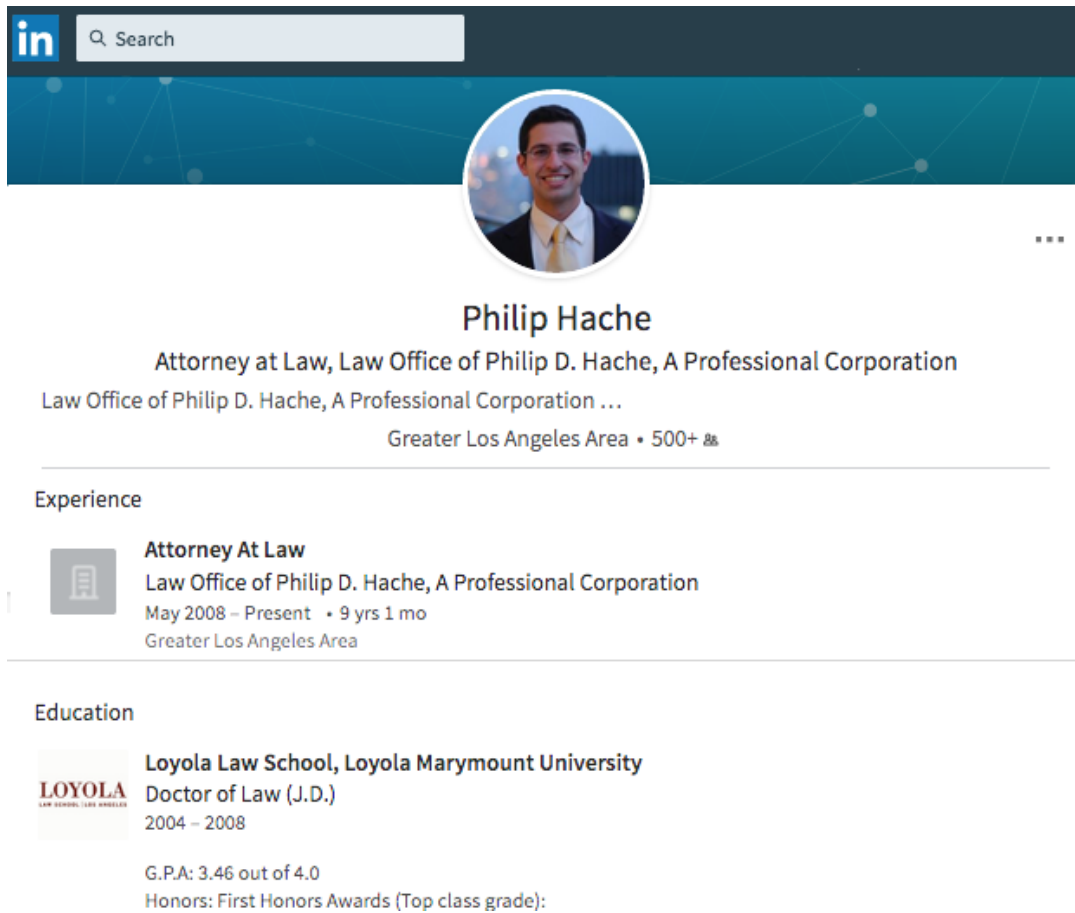
- Chevron** Senior Counsel Chevron Upstream & Gas
Chevron
Jan 2013 – Present • 4 yrs 5 mos
Houston, Texas Area
- Partner**
Shipley Snell Montgomery Droog LLP
Mar 2006 – Dec 2012 • 6 yrs 10 mos
Houston, Texas Area
- Associate**
Baker Botts LLP
Aug 2004 – Mar 2006 • 1 yr 8 mos
Houston, Texas Area
- Judicial Clerk**
Hon. Harold R. DeMoss Jr., United States Court of Appeals for the Fifth Circuit
Aug 2002 – Aug 2004 • 2 yrs 1 mo
Houston, Texas Area

Education

- Pepperdine School of Law**
Juris Doctorate, Law, summa cum laude, 2 out of 205
1999 – 2002
Activities and Societies: Editor-in-Chief of the Pepperdine Law Review: 2001-2002, Staff Member:

Figure 21. Collage excerpt of Daniel Droog's LinkedIn profile, as publicly available online [50]. Profile above shows he graduated from Pepperdine in 2002 summa cum laude with the second highest ranked GPA (Figure 19).

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018. <http://techscience.org/a/2018111301>



Philip Hache
Attorney at Law, Law Office of Philip D. Hache, A Professional Corporation
Law Office of Philip D. Hache, A Professional Corporation ...
Greater Los Angeles Area • 500+ 🌐

Experience

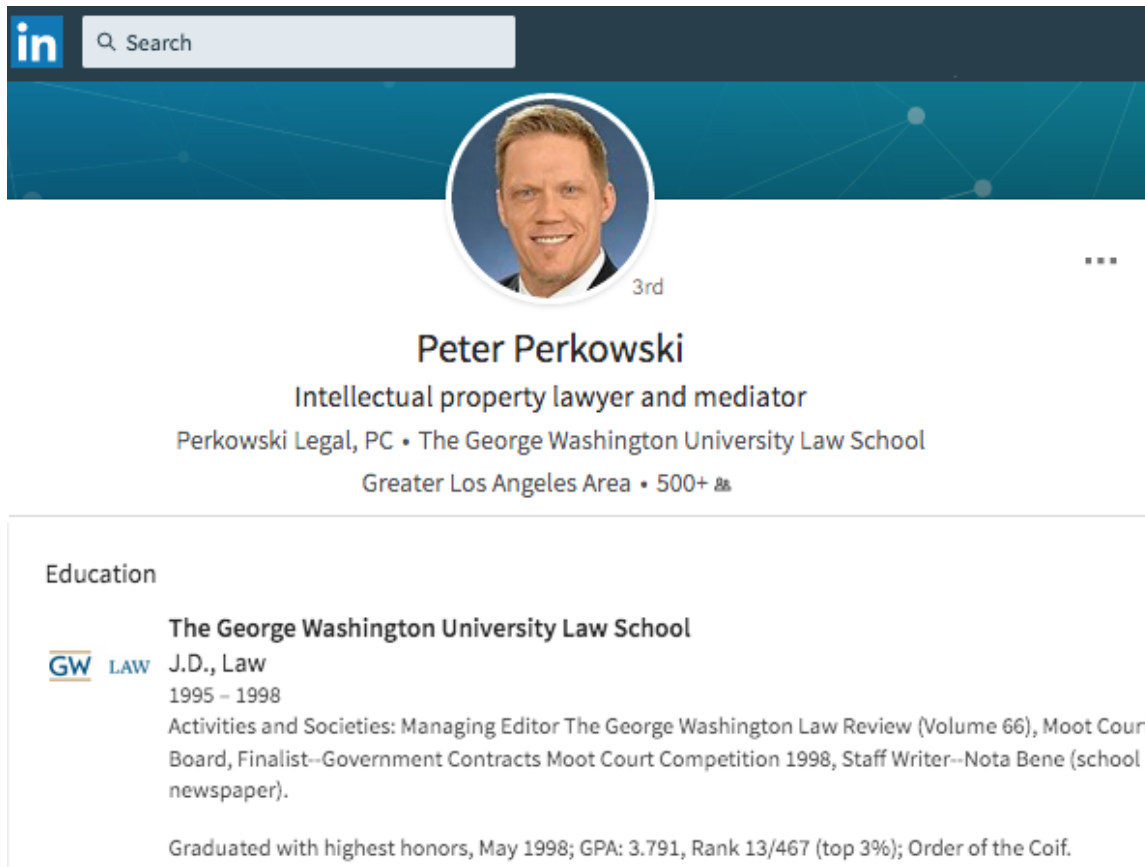
Attorney At Law
Law Office of Philip D. Hache, A Professional Corporation
May 2008 – Present • 9 yrs 1 mo
Greater Los Angeles Area

Education


Loyola Law School, Loyola Marymount University
Doctor of Law (J.D.)
2004 – 2008
G.P.A: 3.46 out of 4.0
Honors: First Honors Awards (Top class grade):

Figure 22. Collage excerpt of Philip Hache's LinkedIn profile, as publicly available online [56]. Profile above shows he graduated from Loyola Law School in 2008 with a reported GPA of 3.46.

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018.
<http://techscience.org/a/2018111301>



in Search

 3rd

Peter Perkowski
Intellectual property lawyer and mediator
Perkowski Legal, PC • The George Washington University Law School
Greater Los Angeles Area • 500+ &

Education


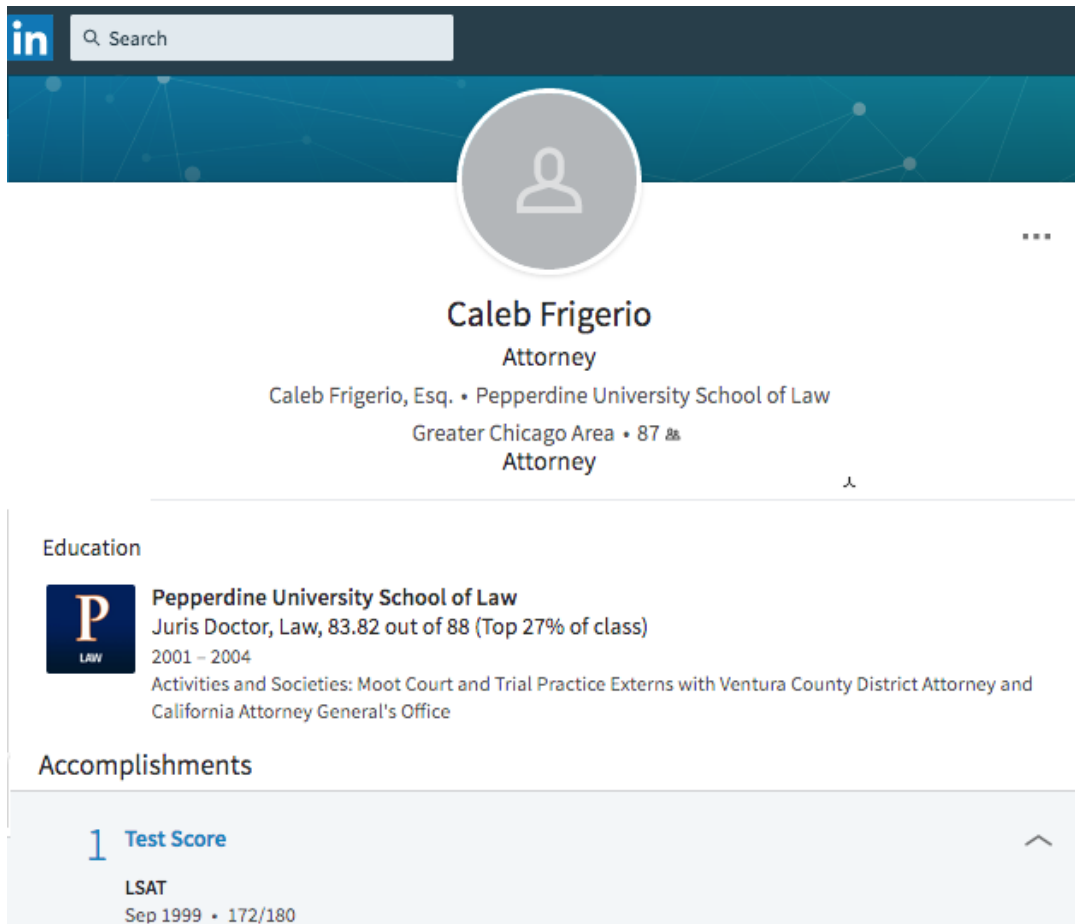
 **The George Washington University Law School**
J.D., Law
1995 - 1998
Activities and Societies: Managing Editor The George Washington Law Review (Volume 66), Moot Court Board, Finalist--Government Contracts Moot Court Competition 1998, Staff Writer--Nota Bene (school newspaper).
Graduated with highest honors, May 1998; GPA: 3.791, Rank 13/467 (top 3%); Order of the Coif.

Figure 23. Collage excerpt of Peter Perkowski's LinkedIn profile, as publicly available online [57]. Profile above shows he graduated from George Washington University Law School in 1998 with a reported GPA of 3.791.

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018. <http://techscience.org/a/2018111301>



The image is a screenshot of a LinkedIn profile for Caleb Frigerio. At the top, there is a dark blue header with the LinkedIn logo and a search bar. Below the header is a circular profile picture placeholder. The name "Caleb Frigerio" is displayed in a large font, followed by "Attorney" in a smaller font. Below that, it says "Caleb Frigerio, Esq. • Pepperdine University School of Law" and "Greater Chicago Area • 87 & Attorney".


The "Education" section is highlighted. It features the Pepperdine University School of Law logo, which is a blue square with a white "P" and "LAW" below it. To the right of the logo, the text reads: "Pepperdine University School of Law", "Juris Doctor, Law, 83.82 out of 88 (Top 27% of class)", "2001 - 2004", and "Activities and Societies: Moot Court and Trial Practice Externs with Ventura County District Attorney and California Attorney General's Office".

The "Accomplishments" section is also highlighted. It shows a blue bar with the number "1" and the text "Test Score". Below this, it says "LSAT" and "Sep 1999 • 172/180".

Figure 24. Collage excerpt of Caleb Frigerio's LinkedIn profile, as publicly available online [58]. Profile above shows that he graduated from Pepperdine in 2008 and entered with a reported LSAT score of 172.


Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018. <http://techscience.org/a/2018111301>


in Search



Marla Chabner
Law Offices Of Marla Chabner, Specializing In Estate Planning
Law Office of Marla Chabner • University of Southern California Law School

Education

 **University of Southern California Law School**
Doctor of Law (JD), Law
1990 – 1993
Activities and Societies: Clerkship for the Honorable Robert Boochever, 9th Circuit District Court, 1992.

 **University of Kansas**
Bachelor of Architecture (B.Arch.), Architecture
1981 – 1985

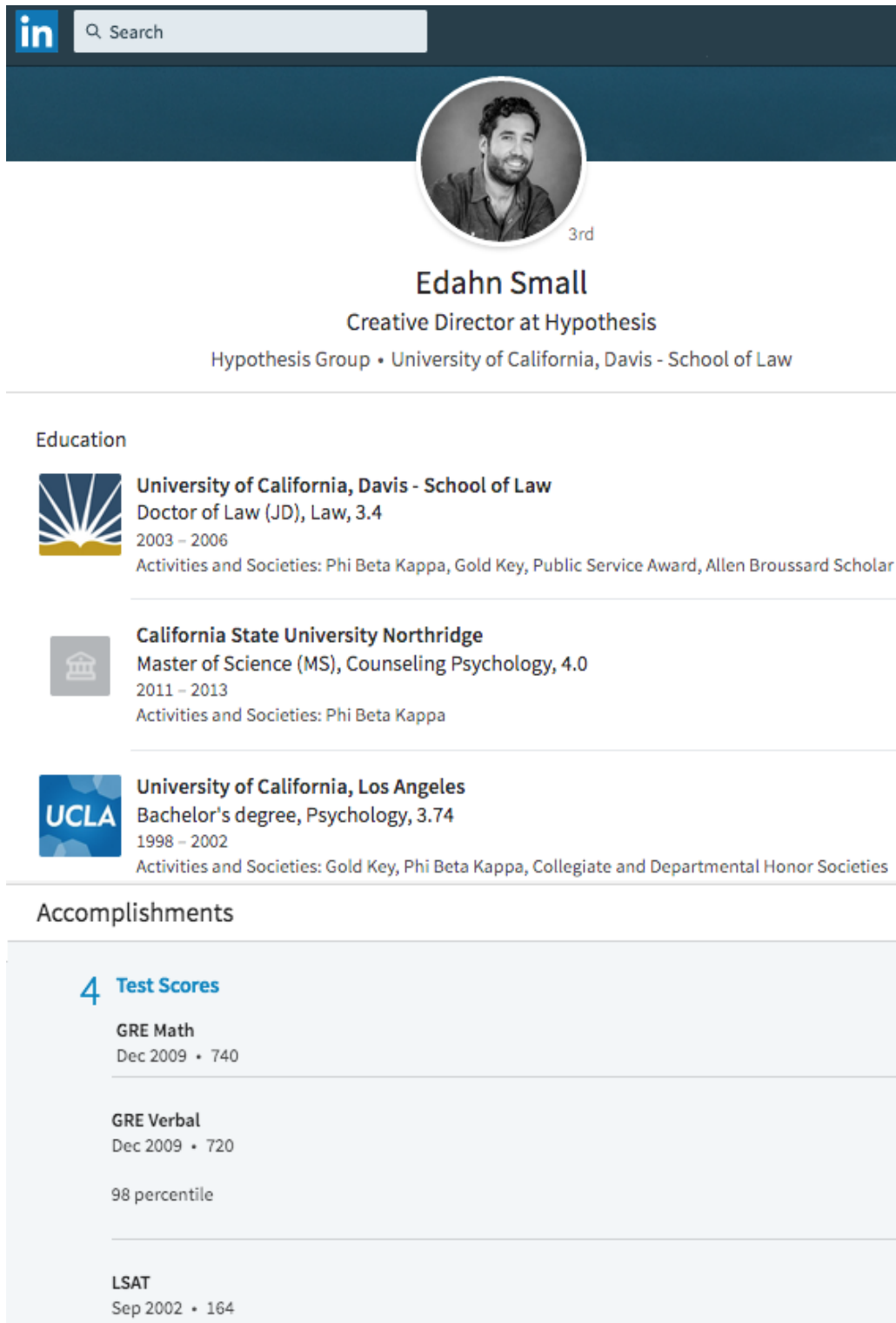
Accomplishments

1 Test Score


LSAT
44
Score is 99%.


Figure 25. Collage excerpt of Marla Chabner's LinkedIn profile, as publicly available online [59]. Profile above shows that she graduated from the University of Southern California Law School in. She received her undergraduate degree at the University of Kansas 1993 and had a reported LSAT score of 44 (on the earlier scale with 48 as the maximum score).


Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018. <http://techscience.org/a/2018111301>



Education

 **University of California, Davis - School of Law**
Doctor of Law (JD), Law, 3.4
2003 – 2006
Activities and Societies: Phi Beta Kappa, Gold Key, Public Service Award, Allen Broussard Scholar

 **California State University Northridge**
Master of Science (MS), Counseling Psychology, 4.0
2011 – 2013
Activities and Societies: Phi Beta Kappa

 **University of California, Los Angeles**
Bachelor's degree, Psychology, 3.74
1998 – 2002
Activities and Societies: Gold Key, Phi Beta Kappa, Collegiate and Departmental Honor Societies

Accomplishments

4 Test Scores

GRE Math
Dec 2009 • 740

GRE Verbal
Dec 2009 • 720

98 percentile

LSAT
Sep 2002 • 164

Figure 26. Collage excerpt of Edahn Small's LinkedIn profile, as publicly available online [60]. Profile above shows that he graduated from UC Davis in 2006 with a reported GPA of 3.4 and entered with an LSAT score of 164.

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018. <http://techscience.org/a/2018111301>







 <p>Natalie Griffiths Greater Los Angeles Area Attorney at InterCon Security Systems, Inc. Legal Services</p> <p>Show more</p>	 <p>Paul Efstratis San Francisco Bay Area Law Practice</p> <p>Show more</p>	 <p>Payam Khodadadi Los Angeles, California Associate at McGuireWoods LLP Legal Services</p> <p>Show more</p>
 <p>Peter C. Leonard Greater Los Angeles Area CPA and Attorney Accounting</p> <p>Show more</p>	 <p>Rachel Stilwell Greater Los Angeles Area Owner at Law Offices of Rachel Stilwell Law Practice</p> <p>Show more</p>	 <p>Raven Sarnoff San Francisco Bay Area Partner at Sarnoff + Sarnoff, APLC Law Practice</p> <p>Show more</p>

Figure 27. Excerpt from publicly available online alumni page of 2005 graduates of Loyola Law School [51].

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018. <http://techscience.org/a/2018111301>

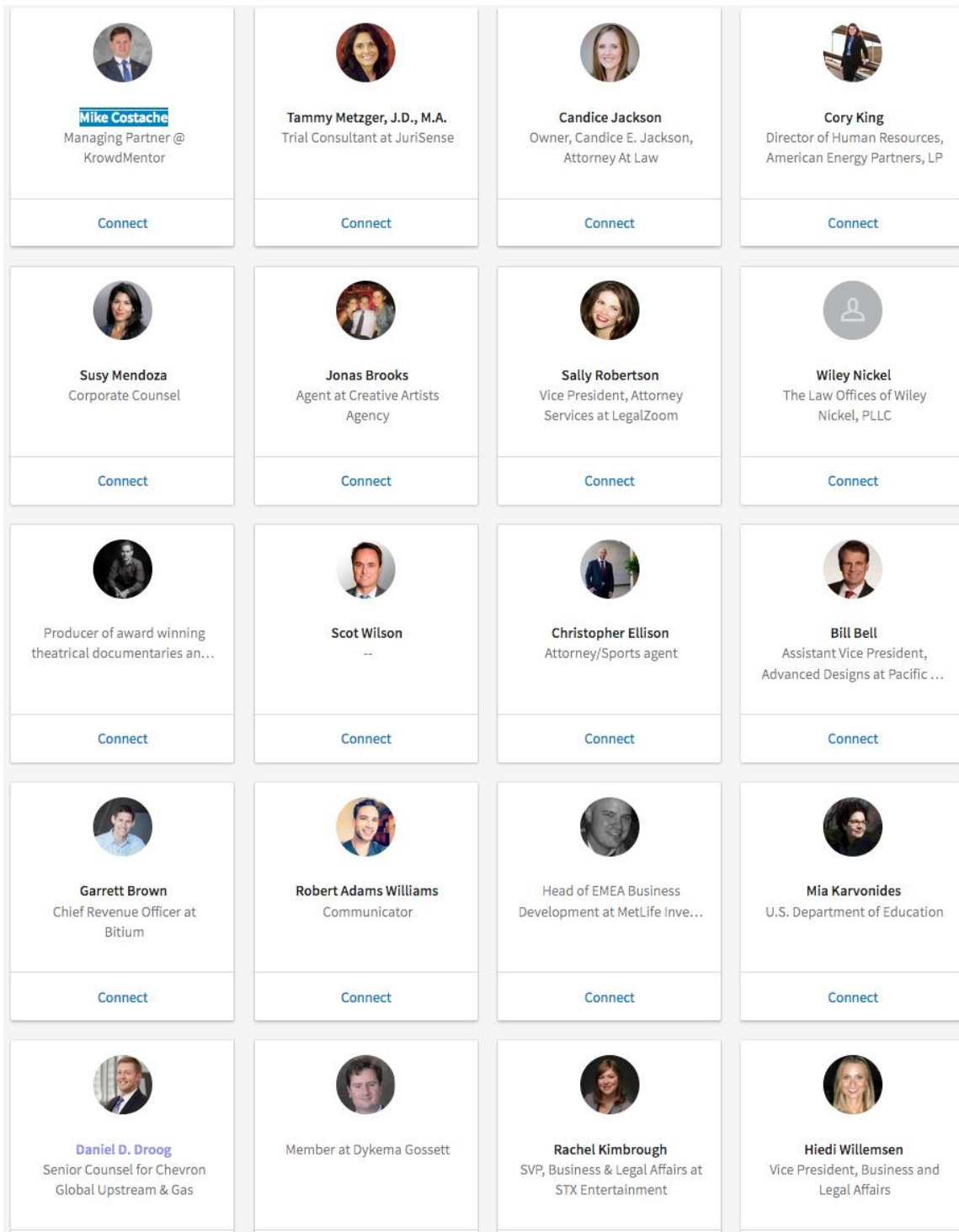


Figure 28. Excerpt from publicly available online page of LinkedIn profiles for 2002 graduates of Pepperdine [52]. Daniel Droog, mentioned earlier in Figure 19, Figure 20, and Figure 21, appears on the bottom row.

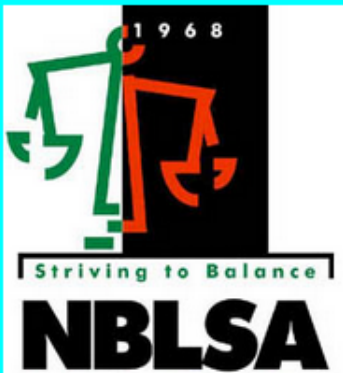
Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018.
<http://techscience.org/a/2018111301>

<p><i>Asian & Pacific Islander Law Students Association</i> Sebyul Chun (2L), <i>President</i> Eric Chan (2L), <i>Secretary/Alumni Relations</i> Sebastian Chan (2L), <i>CFO/Immigration</i></p> <p><i>Black Law Students Association (BLSA)</i> Fred O. Smith (2L), <i>President</i> Lauren Kestenbaum (2L), <i>Vice President</i> Andrea Manka (2L), <i>Vice President</i> Afam Onyema (2L), <i>Vice President</i></p> <p><i>Chinese Law Association</i> Christopher Kramer (1L), <i>Co-President</i> Yi Zhang (LLM), <i>Co-President</i> Adam Rachlis (1L), <i>VP and Treasurer</i></p>	<p><i>Native American Law Students Association</i> Timothy Sanders (2L), <i>President</i> Colin C. Sampson (3L), <i>Secretary/Treasurer</i></p> <p><i>South Asian Law Students Association</i> Saswat Bohidar (2L), <i>Co-President</i> Jenifer Rajkumar (2L), <i>Co-President</i></p> <p><i>Stanford Latino Law Students Association</i> Ann Marie Rosas (2L), <i>Co-Chair</i> Michael Angelo (2L), <i>Co-Chair</i> Emilia Petersen (2L), <i>Co-Chair</i> Juliana Chereji (2L), <i>Co-Chair</i> Joel Hernandez (2L), <i>Co-Chair</i></p>
--	---

Figure 29. Club officers of some race- and ethnicity-specific clubs at Stanford Law School. Shown are excerpts from a publicly available 2006 commencement program [53].

McGeorge School of Law Black Law Student Association

A member of



**BLSA
PAST/PRESENT
BOARD
MEMBERS**

2003-2004 Board:
Denise Williams
(Vice-President);
Samantha Munroe
(Treasurer);
Christy LaPierre
(Minority Affairs
Representative).

2002-2003 Board:
Shakira Pleasant
(President); Marcia
Randle (Vice-
President); Mishael
Pine (Secretary);
Daune Kirrene
(Treasurer).

**BLSA
PICTURES**

Figure 30. Club officers of the McGeorge School of Law Black Law Student Association [54].



Figure 31. Photos of black graduates from McGeorge School of Law. (a) Venus Johnson [61], (b) Dustin Johnson [62], (c) Marcia Randle [63], (d) Anthony C. Williams [64], and (e) William Bishop [65].

Approach: Litmus Test 2 – Our Tools

We used the Python programming language to write programs to harvest online content, so that we could build our own datasets and infer race from names. Below is more information about these tools.

Harvesting Online Content. We have just listed examples from a smorgasbord of online content sharing some fields of information with the Bar Dataset. Because these materials are online, the information can be harvested by automated means in order to produce standalone datasets for convenient processing.

For example, we created our own copy of the Attorney License Data (Figure 20) as follows. First, we wrote a simple Python script to compile a local dataset having fields $\{Barnum, name, datepassed, lawschool, undergrad\}$ to store each member's bar number, name, date passing the bar, law school and undergraduate school attended, respectively. We term this the Attorney Dataset. Here is how we did it.

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018. <http://techscience.org/a/2018111301>

The URL for a specific Bar number appears at the end of the URL. For example, the URL for Daniel Droog having Bar number 224596 (Figure 20) is members.calbar.ca.gov/fal/Member/Detail/224596.

By strategically searching URL ranges, we found that the earliest Bar number assigned to an individual who was admitted in 1977 is 73307, and the last Bar number assigned to an individual who was admitted in 2008 is 262000. Therefore, sequential URLs, from members.calbar.ca.gov/fal/Member/Detail/73307 to members.calbar.ca.gov/fal/Member/Detail/262000, provided all the URLs of Bar members that are the subject of this report. We additionally mined profiles to Bar number 280000 to capture all members who were admitted through December 2011.

In 5 minutes, we wrote a Python script (Figure 32) to capture all the web pages for the Bar members that are the subjects of this study. This simple script is termed a "scraper." Sample code for scrapers written in Python and instruction to write scrapers in Python are readily available online. Python is a free programming language that is also readily available online and is automatically shipped with many computers, such as Mac laptops.

Our Python script ran for 3 days on a laptop connected to the Internet. It captured 198,850 pages. The pages for judges and deceased members of the Bar were excluded. In 45 minutes, we wrote another Python script to compile the Attorney Dataset from the mined web pages.

We use this as both an example of the ease at which we can make standalone datasets from online content and an explanation of how we constructed the Attorney Dataset.

```
import mechanize
from mechanize import Browser
import time

for num in range(73307, 262000+1):
    br = mechanize.Browser()
    response = br.open("http://members.calbar.ca.gov/fal/Member/Detail/" + str(num))
    content= response.read()      # the text of the page
    text_file = open("results" + "/" + str(num) + ".html", "w")
    text_file.write(content)
    text_file.close()
    time.sleep(0.2) #do not overtax web server
```

Figure 32. Python script to capture all California Bar listings for Bar numbers 73307 to 26200, which corresponds to those admitted from 1977 through 2008. Each listing (or web page) is stored in a pre-existing folder (named "results") on the local computer's hard drive. The filename for each captured profile is the bar number followed by ".html."

Inferring Race from Names. We wrote a Python program that used a list of the 151,671 most popular last names and their frequency by race and ethnicity in the 2000 U.S. Census [66] to

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018. <http://techscience.org/a/2018111301>

infer Hispanic and Asian associations from last names. Similarly, we wrote a Python program that used lists of first names given predominantly to Black and White babies [67][68] to infer race from first name.

Approach: Litmus Test 2 – Matching Strategies

The examples above identify numerous online sources that have overlapping information with the Bar Dataset. Commencement programs, for example, include the name of the law school, graduates, and club officers and identify students having honors GPAs. Resumes and online bios may additionally include photos, honors designations and explicit GPA and LSAT scores. The Attorney Dataset, which contains Attorney license information, includes the attorney's name, law school, and Bar admission date. Figure 33 summarizes these data sources and their fields.

Some fields allow inferences to other fields. For example, we may infer *gpa* from honors designation. We may infer *race* from clubs, names and photos. The Attorney Dataset can combine with other sources that include graduation year to provide inferences on the number of times the attorney may have taken the bar (*tries* in the Bar Dataset). Figure 33 shows which fields can have inferred values. The only information not appearing in any other dataset or able to be inferred from any other datasets is *scores* in the Bar Dataset. Conversely, the Bar Dataset is the only source that has no names.

	Bar Dataset	Grad Programs	Resumes Bios	Alumni lists	Club Members	Photos	Attorney Dataset
name							
lawschool							
gradYr							
lsat							
gpa		honors	honors	honors			
race		club name	club name photo	club name photo	club name photo	name photo	name
result							
tries		gradYr& passdate	gradYr& passdate	gradYr& passdate	gradYr& passdate	gradYr& passdate	gradYr& passdate
scores							

	Bar Dataset	Grad Programs	Resumes Bios	Alumni lists	Club Members	Photos	Attorney Dataset
clubs							
honors							
photos							
passdate							
# of individuals in dataset	complete	complete	partial	partial	partial	partial	complete
Additional fields	recnum						barnum

Figure 33. Summary of data sources (top row) and fields (left column). The type of data is identified at the top of the column, from Bar Dataset to Attorney Dataset. A blank shaded cell means the data tends to have information for the field. A shaded cell with the name of a field identifies sources of inferences. For example, GPA is not usually provided in Grad Programs (Commencement programs), but honors do appear (shaded honors for Grad Programs), and GPA can sometimes be inferred from honors (shaded GPA with honors inside). Values for *tries* (number of times bar exam taken) can be inferred from *gradYr* and the bar *passdate* in the Atty Dataset; these inferences require two data sources. The value of scores is only provided in the Bar Dataset, which has no name. Bottom row states whether the source tends to have all individuals (complete) or some (partial). Bar Dataset additionally has the field *recnum* and Attorney Dataset the field *barnum*. Not shown are other data sources released under other public records requests.

Private and semi-public data sources exist too. As part of another public records request, for example, Professor Sander received data from various California law schools, including from the University of California, Davis (UC Davis). We obtained a copy of the UC Davis data (UC Davis Dataset) to determine whether that data could be linked to requested protocol datasets.

The UC Davis Dataset has 10 fields: *gradYear*, *ethnicity*, *admission_outcome*, *enrolled*, *grad_outcome*, *lsat*, *undergraduate_GPA*, *LSAT_index*, *gpa*, and *bar_result*. Several of these fields overlap with information in the Bar Dataset. See Figure 34 for an excerpt of the UC Davis Dataset. There are 59,072 records in total, including students who applied but were not accepted or did not attend, and there are 2001 records for UC Davis graduates. The data reports the year of graduation for most students. (Note: for some students, the graduation year was grouped into periods such as "1994-1996." However, the dataset was sorted by graduation year prior to the changes so that a time period change for one record is preceded

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018. <http://techscience.org/a/2018111301>

and followed by an actual graduation year, thereby making the graduation year understood regardless of the time period. So, we reverted the graduation periods back to graduation years for all records.)

gradYear	ethnicity	lsat	LSAT_index	undergraduate_GPA	bar_result
1994	Asian/Pacific Islander	173	7.0	3.001	P
1994	White/Other/No race/ethnicity disclosure	168	6.9	3.625	P
1994	White/Other/No race/ethnicity disclosure	164	6.8	3.234	P
1994	American Indian/Hispanic	160	6.7	3.143	P
1994	White/Other/No race/ethnicity disclosure	168	7.0	3.006	P
1994	American Indian/Hispanic	154	6.5	2.363	P
1994	White/Other/No race/ethnicity disclosure	163	7.0	3.376	P
1994	White/Other/No race/ethnicity disclosure	171	7.1	3.183	P
1994	White/Other/No race/ethnicity disclosure	156	6.8	3.104	P
1994	Asian/Pacific Islander	156	6.7	2.701	P
1994	Asian/Pacific Islander	152	6.6	2.606	P
1994	White/Other/No race/ethnicity disclosure	174	7.1	3.308	P
1995	White/Other/No race/ethnicity disclosure	162	6.8	3.326	P
1995	White/Other/No race/ethnicity disclosure	170	7.3	3.451	P
1995	White/Other/No race/ethnicity disclosure	161	6.9	3.334	P

Figure 34. Sample of the UC Davis Dataset received by Professor Sander that matches to records in Bar Dataset. Shown are LSAT and GPA scores of some graduates who passed the bar.

The overlap of fields in external data sources makes it easy to see how different data sources may combine to associate names to records in the Bar Dataset. Below we describe 42 different ways to combine fields from data sources in order to possibly put names to one or a few records in the Bar Dataset.

Many test-takers share the same *gradYr* and *lawschool*, a combination readily available in graduation programs and alumni lists. Matching on those fields alone would match hundreds of names to hundreds of records. That is not useful. We want strategies that yield small group matches, where one or a few named individuals match to one or a few records. Small group matches require combining *gradYr* and *lawschool* with other fields in the Bar Dataset that have many distinct values or have rarely occurring values. LSAT and GPA scores are examples of fields having many distinct values, and Black, Hispanic, and Asian test-takers are examples of rarely occurring values. The number of times it takes to pass the bar (*pass, tries*) may help divide the numbers further. We introduce re-identification strategies based on combining *gradYr* and *lawschool* with *lsat, gpa, race, and (pass, tries)*. That gives 14 combinations or the 14 matching plans that appear in Figure 35. Another 14 matching plans are possible by not including *lawschool* and another 14 by not including *gradYr*, but these are not expected to be as productive if *lawschool* and *gradYr* are available. We demonstrated earlier that all these fields may appear publicly or be inferred from publicly available information.

Matching Plan	Fields Involved
1	<i>gradYr, lawschool, lsat</i>
2	<i>gradYr, lawschool, gpa</i>
3	<i>gradYr, lawschool, race</i>
4	<i>gradYr, lawschool, (pass, tries)</i>
5	<i>gradYr, lawschool, lsat, gpa</i>
6	<i>gradYr, lawschool, lsat, race</i>
7	<i>gradYr, lawschool, lsat, (pass, tries)</i>
8	<i>gradYr, lawschool, gpa, race</i>
9	<i>gradYr, lawschool, gpa, (pass, tries)</i>
10	<i>gradYr, lawschool, race, (pass, tries)</i>
11	<i>gradYr, lawschool, lsat, gpa, race</i>
12	<i>gradYr, lawschool, lsat, gpa, (pass, tries)</i>
13	<i>gradYr, lawschool, gpa, race, (pass, tries)</i>
14	<i>gradYr, lawschool, lsat, gpa, race, (pass, tries)</i>
15-28	Same as 1-14 without <i>lawschool</i>
29-42	Same as 1-14 without <i>gradYr</i>

Figure 35. Matching plans to associate readily available public information online to records in the Bar Dataset. Matching plans 1-14 are the primary ones using the most fields. These plans translate to protocol datasets by replacing related fields with their derived

replacements in the protocol dataset (e.g., *gradPeriod* for *gradYr* in the 11-Anonymity Protocol).

Litmus Test 2 consists of testing whether any of the matching plans in Figure 35 are feasible given the dataset produced from a protocol and, if so, whether they describe an actual re-identification strategy that uses that matching plan.

We evaluate the feasibility of a matching plan (re-identification risk) by estimating or computing the number of unique or small group matches in the protocol dataset. Based on those counts, we report whether the risk is sufficiently small in terms of k -anonymity (Litmus Test 2a) and HIPAA (Litmus Test 2b).

We started testing the hypothesis that the protocols perform perfectly, i.e., that absolutely no individual could be matched reliably to any record. The Sander Team did not claim perfect protection but rather postulated that the protocols adhered to k -anonymity, where k is 11 for the 11-Anonymity and Plus Protocols and k is 5 for the Enclave Protocol. If so, that means there should be no unique or small group re-identifications less than k possible in those protocol datasets (the definition of k -anonymity).

More than 15 years ago, when k -anonymity was first introduced, there were fewer datasets available on which to link. Thus, k -anonymity allowed the k requirement to be enforced only on known shared fields. In today's data-rich environment, the opposite posture exists. One cannot know whether a field is or later will be shared. All fields thus have to be subject to the k -requirement unless strong evidence exists for a field to be exempted. Earlier in this writing, we showed evidence that each field in the Bar Dataset can be found publicly (Figure 33) with the exception of the bar scores. Therefore, all the fields of the Bar Dataset would be subject to the k -requirement. Any new fields generated by a protocol from information in the original fields would be subject to the k -requirement too. This was a primary flaw in the Sander Team's protocols – the assumption that many of the fields of data did not need to be anonymized.

As mentioned earlier, we report the risk of small re-identifications for $k=1$, $k<5$, $k<11$ and $k<20$ applied across all the fields for each re-identification strategy, and we report the number of distinct individuals included in the re-identifications (re-identification pool).

If the number of re-identifications is less than the protocol's proclaimed value of k , then the result for Litmus Test 2 is that a sufficient number of small group re-identifications exist ("yes" on Figure 18) and testing then proceeds to Litmus Test 3. On the other hand, if we find the number of small group re-identifications is less than k , then we proceed to test the number of unique re-identifications (Litmus Test 2b) as follows.

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018. <http://techscience.org/a/2018111301>

The Sander Team claimed that the number of individuals that might be uniquely re-identified in a protocol dataset would not exceed the prior experimental results for the HIPAA Safe Harbor (0.04 percent). If the number of unique re-identifications is less than 0.04 percent of the records as claimed, then Litmus Test 2 would find an insufficient number of small group re-identifications ("no" on Figure 18). That was not our finding, and so testing proceeds to Litmus Test 3.

In summary, to support the hypothesis posited by the Sander Team, we would expect the number of unique re-identifications to be less than the noted HIPAA standard and the number of small group re-identifications to be 0 for all groups of size smaller than k . Data contrary to that hypothesis causes a failure of Litmus Test 2.

Approach: Litmus Test 3 – Small Group Matches Possible

When testing advances to Litmus Test 3, we know a protocol is technically reasonable (Litmus Test 1) and has some specific re-identification risks (Litmus Test 2). Litmus Test 3 checks for real-world examples to substantiate the risk. We test at least one re-identification strategy, and if we find too many small-group matches, then we do not need to check any other re-identification strategies to have a result. On the other hand, if we do not find a sufficient number of small-group matches with the first re-identification strategy we test, then we test additional strategies. If we exhaust all the re-identification strategies without finding a sufficient number of small-group re-identifications, then we have not disproved the hypothesis.

Approach: Litmus Test 4 – Harmful Matches

In Litmus Test 4, we assess whether the small-group matches described generally in Litmus Test 2 and specifically in Litmus Test 3 could actually cause harm to individuals in the re-identification pool. We do this by demonstrating the nature and specificity of possible harms.

Results

In the prior section, we operationalized our litmus tests, so we are now ready to begin testing. We test each protocol in turn.

The 11-Anonymity Protocol

We find it technically reasonable to execute the 11-Anonymity Protocol, but it does have grave privacy vulnerabilities. One critical problem is that the 11-Anonymity Protocol does not provide k -anonymity for $k=11$, as asserted by the Sander Team. Another is that the number of unique records remaining in the data (46 percent) dwarfs the old HIPAA standard (0.04 percent). As a result, we were able to demonstrate small-group re-identifications and the

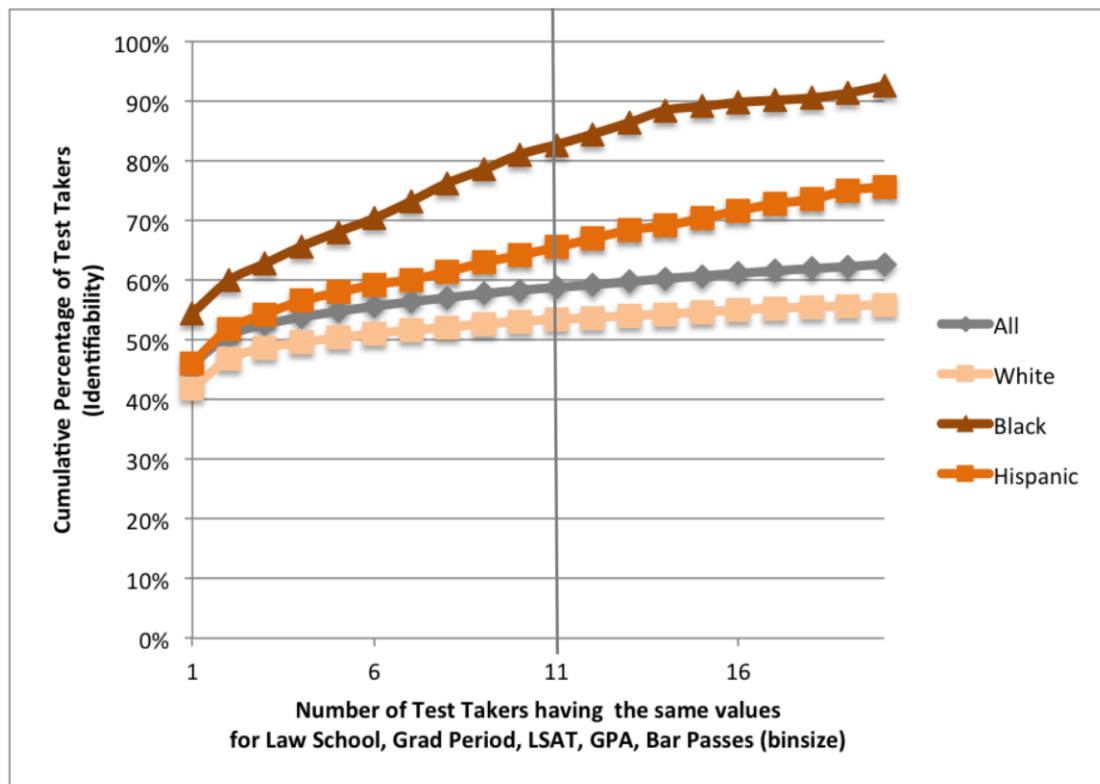
potential for personal harms. Therefore, results from our litmus tests scientifically disproved the Sander Team's hypothesis that their 11-Anonymity Protocol protected privacy.

Steps 1 through 6 of the 11-Anonymity Protocol (Figure 3) use basic Excel commands. Steps 7 could use the PERCENT() or PERCENTRANK() functions with an additional spreadsheet, and Step 8 could use a pivot table with nested IF statements afterwards. Overall, the 11-Anonymity Protocol is technically reasonable to execute as we define that term above.

The 11-Anonymity Dataset had 131,910 records. We counted the number of test-takers having the same values for *lawschool*, *gradPeriod*, *lsat*, *gpa*, *race*, and bar passes and found that 60,572 (46 percent) were unique, 77,568 (58 percent) were in binsizes less than 11, and 82,722 (62 percent) were in binsizes less than 20. Figure 36 itemizes these values by binsize and race/ethnicity.

Results are even worse for records identified as Black or Hispanic. For Blacks, we found that 2,568 (54 percent) were unique, 3,908 (83 percent of all Black test-takers) were in binsizes less than 11, and 4,316 (91 percent) were in binsizes less than 20.

And for Hispanics, we found that 3,414 (46 percent) were unique, 4,853 (66 percent of all Hispanic test-takers) were in binsizes less than 11, and 5,560 (75 percent) were in binsizes less than 20.



	All	%White	%Black	%Hispanic	%Asian	%URM	%White/Asian
binsize	(n=131910)	(n=81857)	(n=4724)	(n=7404)	(n=16270)	(n=1466)	(n=1023)
1	46%	42%	54%	46%	35%	51%	81%
2	51%	47%	60%	52%	39%	54%	93%
3	53%	49%	63%	54%	41%	55%	97%
4	54%	50%	66%	57%	43%	56%	99%
5	55%	50%	68%	58%	45%	57%	100%
6	56%	51%	70%	59%	46%	59%	
7	56%	52%	73%	60%	47%	60%	
8	57%	52%	76%	61%	48%	61%	
9	58%	53%	79%	63%	49%	62%	
10	58%	53%	81%	64%	50%	65%	
11	59%	53%	83%	66%	50%	65%	
12	59%	54%	85%	67%	50%	65%	
13	60%	54%	86%	68%	51%	67%	
14	60%	54%	89%	69%	52%	68%	
15	61%	55%	89%	70%	53%	70%	
16	61%	55%	90%	72%	53%	71%	
17	62%	55%	90%	73%	53%	74%	
18	62%	56%	91%	74%	54%	76%	
19	62%	56%	91%	75%	54%	76%	

Figure 36. Re-identification risk in 11-Anonymity Dataset for test-takers having the same law school, graduation period, LSAT, GPA, and bar passes (first or multiple attempts) for k from 1 to 19. Reported by races and all races. URM = Black and Hispanic.

The 11-Anonymity Protocol purported to adhere to k -anonymity where $k=11$. If that were true, the number of test-takers in bins of sizes less than 11 would be 0 and not 60,572 (46 percent). Therefore, we found vulnerability with the 11-Anonymity Dataset. Almost half the records are unique, implying almost half the records are uniquely identifiable!

An additional 12 percent of the records (77,568 or 58 percent, less the 46 percent unique) are in groups of sizes less than 11. With so many small groups in the 11-Anonymity dataset, there are many ways to demonstrate real-world vulnerabilities.

For example, Class One schools maintain the most detail under the 11-Anonymity Protocol. In comparison to the fields for the Bar Dataset, fields for Class One schools have *gradPeriod* (3- or 6-year ranges) instead of *gradYr*, and 6 overlapping values for *race* instead of 8 distinct values; all other fields remain the same. Of the 14 primary matching plans (Figure 35), changes made to Class One Schools by the 11-Anonymity Protocol only preclude Matching

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018.
<http://techscience.org/a/2018111301>

Plan 3. The remaining 13 describe other vulnerabilities not addressed by the 11-Anonymity Protocol.

For convenience, we look closer at Matching Plans 2, 5, 7, 10, and 14 in the following five examples.

Example 1. Matching Plan 10 focuses our attention on the combination of fields *gradPeriod*, *lawschool*, *race*, and bar passes.

Here is a re-identification strategy for the 11-Anonymity Protocol based on Matching Plan 10 and a small group of Hispanic or Asian bar passers from the same school in the same graduation period. We focus on Hispanic or Asian attorneys because we can often infer race/ethnicity from last names for Hispanics and Asians. First, we acquire lists of graduates from a Class One school for the requisite graduation years. We harvest the names and infer race from their names. We then look up Hispanic or Asian graduates by name in the Attorney Dataset to get the date each was admitted to the bar (passed the exam). We then learn which attorneys passed on the first or perhaps multiple tries, and the number of named attorneys in each category could be less than 11.

In fact, in the 11-Anonymity Dataset is a small-group bin consisting of 11 Hispanic attorneys who graduated from Pepperdine (a Class One school) during the same 3-year graduation period. According to the 11-Anonymity Dataset, 4 of them passed the exam on their first attempt, and 7 passed after more than one attempt. Who are these graduates?

We downloaded graduation programs from Pepperdine for the noted graduation years. We then harvested all names from the commencement programs. Our computer program inferred that 12 had last names most often associated with Hispanics in the U.S. Census. We found 9 of these Pepperdine graduates in the Attorney Dataset. Of them, 4 passed the bar the same year as graduation, and 5 passed later. Therefore, these named attorneys matched the groups of 4 and 7 records in the 11-Anonymity Dataset, respectively. The size of each group, 4 names to 4 records and 5 names to 7 records, is less than 11. In other words, the 11-Anonymity Protocol did not work. This is a real-world example of small-group re-identifications in the 11-Anonymity Dataset.

Example 2. Matching Plan 7 focuses our attention on the combination of fields: *gradPeriod*, *lawschool*, *lsat*, and bar passes.

Here is a re-identification strategy for the 11-Anonymity Protocol based on Matching Plan 7. Find an online profile for a practicing attorney who reports an LSAT score, law school, and graduation year. Again, for convenience, we focus on Class One schools. Look up the attorney in the Attorney Dataset to see whether bar passage was on the first attempt or on multiple attempts. Then, see how many records in the 11-Anonymity Dataset have the same *gradPeriod*, *lawschool*, *lsat*, and bar passes.

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018. <http://techscience.org/a/2018111301>

Marla Chabner graduated from the University of Southern California Law School (USC) in 1993, according to her LinkedIn profile (Figure 25). She now reportedly practices law as an estate planner. Her LinkedIn profile also states that her LSAT score was 44.

When we searched for her name in the Attorney Dataset, we did not find anyone with her name. Realizing that women often change their last name upon marriage, we searched for an attorney having her first name and graduating from the USC in 1993. We found Marla Ann Smith, whose undergraduate degree was from the University of Kansas. The LinkedIn profile of Marla Chabner lists her undergraduate degree as also being issued from the University of Kansas. The State Bar's website lists Ms. Smith's email address as MChabner@ChabnerLegalFinancial.com. So, we were confident that Marla Ann Smith is now Marla Chabner. She passed the bar on her first attempt.

How many individuals in the 11-Anonymity Dataset graduated from the USC between 1991 and 1993 with an LSAT score of 44 and passed the bar on the first attempt? One. We uniquely identified Marla Chabner's record in the 11-Anonymity Dataset. Two individuals graduated in the same 3-year period from USC with a 44 LSAT score, but she was the only one who passed the bar on her first attempt. Again, the 11-Anonymity Protocol failed to provide the promised *k*-anonymity protection. If Ms. Smith's bar score records were publicly released with this data, they would be easy identifiable through this method.

Example 3. Matching Plan 2 focuses our attention on GPA instead of LSAT; the fields of interest are *gradPeriod*, *lawschool*, and *gpa*.

Here is a re-identification strategy for the 11-Anonymity Protocol based on Matching Plan 2. We find an online profile with a GPA, law school, and graduation year. This time we focus on Class Two schools; the graduation period in Class Two schools is 9 years instead of 3. How many records in the 11-Anonymity Dataset have the same *gradPeriod*, *lawschool*, and *gpa* for a Class Two school?

Peter Perkowski graduated from the George Washington University Law School, a Class Two school, in 1998 with a 3.791 GPA, according to his LinkedIn profile (Figure 23). How many individuals in the 11-Anonymity Dataset graduated from the George Washington University Law School between 1991 and 1999 with a GPA of 3.791? No one in the 11-Anonymity Dataset had that exact GPA, but one individual who graduated the George Washington University Law School between 1991 and 1999 had a GPA of 3.79. We uniquely associated a name to only one record in the 11-Anonymity Dataset. Again, the 11-Anonymity Protocol failed to provide the promised *k*-anonymity protection.

Example 4. Matching Plan 2 revisited to use honors designation. The fields of interest are *gradPeriod*, *lawschool*, and *gpa*.

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018.
<http://techscience.org/a/2018111301>

We could continue on and on with other matching plans and re-identification strategies, or we could use the same re-identification strategies with different data sources. For example, here is another re-identification strategy for Matching Plan 2, which involves the fields *gradPeriod*, *lawschool*, and *gpa*. Select a Class One or Class Two school that assigns honors to the top ranked students. Find the names of those students in graduation lists, graduation programs, or profiles. See which of them passed the Bar to confirm they should be in the 11-Anonymity Dataset. Then, search for the top GPAs in the 11-Anonymity Dataset for the school in that time period. The result is a match of those names to those records. Race may allow us to further subdivide the matches.

Here's how that would work. Recall the Pepperdine commencement programs for 2000, 2001, and 2002. The Valedictorian has the top GPA for the class, and the Salutatorian has the second highest. Julie Trotter, Ashlea (Wright) Montgomery, and Jay Spagnola were the Pepperdine Valedictorians for 2000, 2001, and 2002, respectively, and Lindsey Duro, Michelle (Murray) Garland, and Daniel Droog were the Salutatorians. All 6 of them are members of the Bar, so their scores are in the 11-Anonymity Dataset. If we look at the top 6 top GPA scores for Pepperdine among 2000-2002 graduates in the 11-Anonymity Dataset, then what do we know? If the top 6 scores are from the top 2 individuals each year, then we have 6 names matched to 6 records correctly. Of course, it could be that all 6 scores were for individuals who all graduated the same year. In that case, only 2 of the 6 names would be correct. Altogether, by matching the 6 names to the 6 records, we know 2, 3, 4, 5, or 6 of them are correct. Regardless, at least 2 are correct, so the 11-Anonymity Protocol failed again to provide the promised k -anonymity protection.

Example 5. Matching Plan 10 revisited to scale the number of re-identifications using a two-step approach. First we use Matching Plan 5 (or 14), and then Matching Plan 10. The fields of interest are *gradPeriod*, *lawschool*, *lsat*, and *gpa* and then *gradPeriod*, *lawschool*, *race*, and *bar passes*.

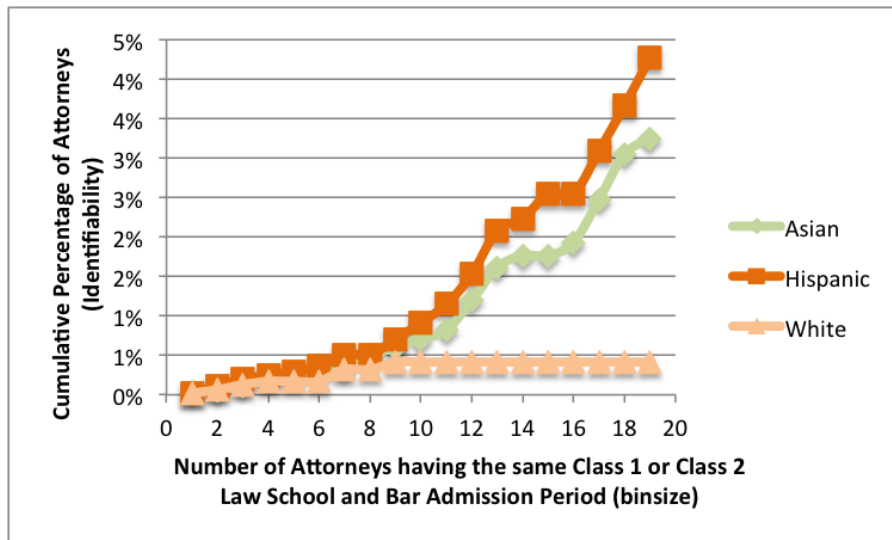
There are many more matching plans and re-identification strategies to consider. The vast number of unique records and small-sized groups in the 11-Anonymity Dataset allow a plethora of possibilities, but so far, the examples we provided are small-scale. They were sufficient to show the failure of the 11-Anonymity Dataset, but how do they scale? We could iterate and automate, but even then, how many attorneys might be re-identified using an approach like that in Example 1, for example?

We took the Attorney Dataset and inferred race by last name for each attorney admitted from 1982 to 2008 who graduated from a Class One or Class Two law school; we found a total of 109,083 attorneys. We then counted the number of small groups of Hispanics, Asians, and Whites admitted to the Bar from the same law school in the same graduation period that the 11-Anonymity Protocol uses for the law school. We excluded Blacks because racial inference by last name for Blacks is not as predictive, often overshadowed by Whites having the same

last name – i.e., some of the “White” attorneys may be Black. We inferred race for 9,412 Asian attorneys, 5,691 Hispanic attorneys and 75,035 White attorneys. The percentage of attorneys labeled as Asian (10 percent) or Hispanic (6 percent) was much less than the percentage of those labeled as White (83 percent).

The approach we took in the first example relied on small groups of graduates from the same school during the same graduation period and of the same race. Figure 37 shows the number of small groups we found of attorneys from the same school admitted to the Bar in the same time periods used by the 11-Anonymity Protocol. There are 3 unique Asians and 2 unique Hispanics, 24 Asians and 28 Hispanics in groups of size 5 or less, and 66 Asians and 86 Hispanics in groups of size 11 or less.

Of course, to actually do the re-identifications, we would also need graduation years, which are not in the Attorney Dataset. We would have to use online alumni lists, graduation programs, or profiles to find the corresponding graduation years. But that step is not necessary for us to appreciate these estimates. The results show that the approach we took in the first example scales a bit, but how can we get it to scale much more?



Binsize	Asian		Hispanic		White	
	Number Bins	Cumulative Percentage (n=9412)	Number Bins	Cumulative Percentage (n=5691)	Number Bins	Cumulative Percentage (n=75035)
1	3	0%	2	0%	2	0%
2	1	0%	4	0%	2	0%
3	2	0%	3	0%	1	0%
4	2	0%	1	0%	0	0%
5	1	0%	1	0%	0	0%
6	0	0%	1	1%	2	0%

Binsize	Asian		Hispanic		White	
	Number Bins	Cumulative Percentage (n=9412)	Number Bins	Cumulative Percentage (n=5691)	Number Bins	Cumulative Percentage (n=75035)
7	1	0%	2	1%	0	0%
8	2	0%	0	1%	1	0%
9	1	0%	2	1%	0	0%
10	1	1%	2	2%	0	0%
11	1	1%	2	2%	0	0%
12	3	1%	3	3%	0	0%
13	3	2%	4	3%	0	0%
14	1	2%	1	4%	0	0%
15	0	2%	2	4%	0	0%
16	1	2%	0	4%	0	0%
17	3	2%	3	5%	0	0%
18	3	3%	3	6%	0	0%
19	1	3%	3	7%	0	0%

Figure 37. Estimated re-identification risk in 11-Anonymity Dataset for attorneys having the same law school and graduation period from Class 1 and Class 2 Schools for k from 1 to 19 based on their occurrence in the Attorney Dataset. See the 11-Anonymity Protocol for graduation periods and school class determinations. Race inferred by last name.

One way to get large numbers of re-identifications is to use another source of data, preferably one that has fields with lots of distinguishing values (e.g., LSAT and GPA scores). The UC Davis Dataset (Figure 34) is one such dataset. It provides the graduation year, LSAT, GPA, and bar result for all the graduates from UC Davis from 1994 to 2008. Clearly, law schools have this kind of information on their students, as evidenced by this dataset originating from UC Davis in response to another public records request. Once it is in the public domain, others have it too. We understand that Professor Sander received a copy of the UC Davis Dataset. In fact, Professor Sander received the same kind of data from other California schools and from schools around the country [69]. How might having this kind of information impact re-identification rates? The answer depends on how often LSAT and GPA scores are unique.

We changed the graduation year in the UC Davis Dataset to have the same graduation periods described in the 11-Anonymity Protocol; we term this "School Data." We then counted the number of unique bins for different combinations of fields. Figure 38 shows the counts and percentages. Selecting a row and a column and the cell at which they intersect reports the number and percentage of unique records found. Not surprisingly, having all 4 fields, LSAT,

GPA, Race, and Bar for a graduation period provided the greatest number of unique records (1,979 of 2,001 or 99 percent). Any combination that included both LSAT and GPA resulted in at least 97 percent of the records being unique. These GPAs have as many as 3 digits to the right of the decimal. That means that LSAT and GPA (with 3 digits to the right of the decimal) combined are close to being a unique identifier for each student, similar to a Social Security number, though not 100 percent unique. This by itself is persuasive evidence that individual-level LSAT and GPA scores should not be publicly revealed.

GPA drives the number of unique records. There are 527 (26 percent) unique GPAs, regardless of graduation period. Knowing the student's GPA and the fact that they graduated from UC Davis is enough to uniquely identify the records of 527 students. In comparison, only 5 students have a unique LSAT when graduation period is not included.

When we include graduation period, the number of unique records having the graduation period and GPA is 1,445 (72 percent), compared to 18 (1 percent) for LSAT instead of GPA. See Figure 38.

	LSAT	GPA	Race	Bar	LSAT, GPA	Race, Bar
LSAT	18 (1%)	1959 (98%)	118 (6%)	50 (2%)	1959 (98%)	227 (11%)
GPA	1959 (98%)	1445 (72%)	1675 (84%)	1564 (78%)	1959 (98%)	1726 (86%)
Race	118 (6%)	1675 (84%)	0 (0%)	0 (0%)	1971 (99%)	0 (0%)
Bar	50 (2%)	1564 (78%)	0 (0%)	0 (0%)	1971 (99%)	0 (0%)
LSAT, GPA	1959 (98%)	1959 (98%)	1971 (99%)	1971 (99%)	1959 (97%)	1979 (99%)
Race, Bar	227 (11%)	1726 (86%)	0 (0%)	0 (0%)	1979 (99%)	0 (0%)

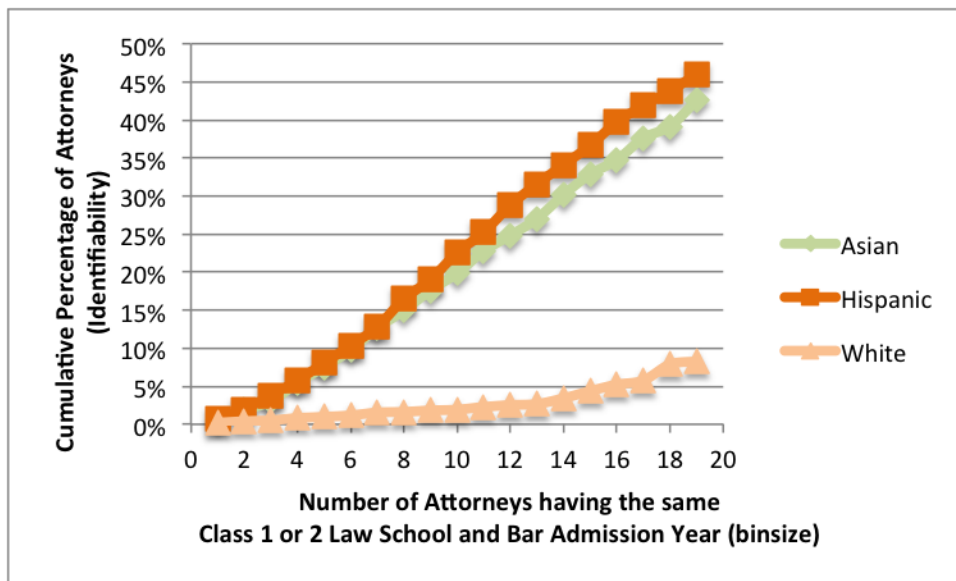
Figure 38. Number and percentage of unique occurrences of combinations of fields in the UC Davis Dataset for graduation periods based on the 11-Anonymity Protocol. The value for Bar is pass/fail. There are a total of 2,001 records. GPA values have a precision of 3 decimal places (e.g., 3.123 and 2.784). The intersection of a row and column report the number of unique records found with that combination of fields. Each combination includes the graduation period – i.e., the combination of graduation period and GPA is unique for 1,445 records.

A school, researcher or anyone else holding such School Data can use it to match names to the records in the 11-Anonymity Dataset at scale. Here is the re-identification strategy. First, you match on $\{lawschool, gradPeriod, lsat, gpa\}$. These matches will be about 97 percent unique (Figure 38). Of course, matching on $\{lawschool, gradPeriod, lsat, gpa, race, bar\}$ would yield about 99 percent unique matches. If you are the school from which the data originates, then you already have the names of all the students, so you have now learned the Bar scores of each student.

If you are not the originating school, then the matched records now include the graduation year from the school data and the bar scores from the 11-Anonymity Dataset. This combined data holds the sensitive information in a more identifiable form than the original 11-

Anonymity data because we can now use the graduation years to help match to names. For example, suppose we want to put names to the records of Hispanic and Asian attorneys. We can use a graduation program or alumni list, as described before, but now instead of being constrained to the graduation periods, we have the graduation year. Instead of getting some re-identifications of about a hundred individuals (Figure 37), the vulnerability dramatically increases to almost half of all Hispanic and Asian attorneys. Here is how we made this estimate.

We revisited our earlier counts of the number of small groups of Hispanics, Asians, and Whites admitted to the Bar in the same graduation period from the same law school (Figure 37). This time we looked at what happens when we replace graduation period with graduation year. That is, we look at the vulnerability based on the year of admission to the Bar and not on a period of years. Figure 39 shows the number of small groups found. There are 48 (1 percent of all Asian labeled attorneys) unique Asians and 65 (1 percent of all Hispanic labeled attorneys) unique Hispanics, 513 (5 percent) Asians and 543 (6 percent) Hispanics in groups having fewer than 5 named individuals, and 1,869 (20 percent) Asians and 2,122 (23 percent) Hispanics in groups having fewer than 11 named individuals. Almost half (43 percent Asians and 46 percent Hispanics) of all the Asian and Hispanic attorneys are in binsizes less than 20. That means Hispanic and Asian attorneys are highly susceptible to this re-identification strategy.



Binsize	Asian		Hispanic		White	
	Number Bins	Cumulative Percentage (n=9412)	Number Bins	Cumulative Percentage (n=5691)	Number Bins	Cumulative Percentage (n=75035)
1	48	1%	65	1%	20	0%
2	63	2%	60	2%	9	0%
3	37	3%	58	4%	4	1%

Binsize	Asian		Hispanic		White	
	Number Bins	Cumulative Percentage	Number Bins	Cumulative Percentage	Number Bins	Cumulative Percentage
		(n=9412)		(n=5691)		(n=75035)
4	57	5%	46	6%	7	1%
5	39	8%	43	8%	3	1%
6	35	10%	36	10%	3	1%
7	39	13%	33	13%	5	2%
8	27	15%	43	16%	1	2%
9	28	18%	27	19%	2	2%
10	21	20%	33	23%	1	2%
11	25	23%	23	25%	3	2%
12	16	25%	28	29%	2	3%
13	16	27%	20	32%	1	3%
14	21	30%	17	34%	5	3%
15	17	33%	16	37%	6	4%
16	11	35%	18	40%	5	5%
17	16	38%	12	42%	3	6%
18	8	39%	10	44%	11	8%
19	17	43%	11	46%	2	8%

Figure 39. Estimated re-identification risk in 11-Anonymity Dataset for attorneys having the same law school and graduation year from Class 1 and Class 2 Schools for k from 1 to 19 based on their occurrence in the Attorney Dataset. Race inferred by last name.

Of course, there are many other re-identification strategies and matching patterns not discussed here. Nevertheless, we demonstrated in numerous ways that the 11-Anonymity Protocol leaks re-identifications and does not keep its k -anonymity promise (that no individual or record would be associated with a group of less than 11).

The last litmus test we impose describes harm that can result from unique and small group re-identifications in the 11-Anonymity Dataset.

Data provided under a public records request is data put in the public domain. Uses and users cannot be constrained except by the privacy protections imposed on the data. Any reasonable individual familiar with the Internet can do many of the re-identification examples we show. Programming skills and the availability of other data can dramatically scale efforts. Even the ability to re-identify small numbers of individuals can expose individuals to harm.

We realize that the showing of harm may not be relevant to the important point that records are re-identifiable. Professor Sander did not promise no harm, just that the records were not

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018.
<http://techscience.org/a/2018111301>

identifiable. However, the Sander Team question that even if some re-identifications were possible, would any harm result? Here are some examples.

Assume an individual passes the bar and begins working as an attorney. Having bar scores and LSAT and GPA scores made public may undermine professional standing by encouraging comparison among practitioners based on the scores (which may have no legitimate impact on the quality of legal representation). During the legal proceedings, we showed examples of how unflattering GPA scores could be associated with highly successful politicians, judges, and other high-profile individuals. If such data were made public, insinuations about academic performance could fuel speculation about other candidates. Solely to address raised concerns, judges, candidates for office, and even attorneys being hired or promoted may be forced to reveal personal academic records that would otherwise remain private.

Current or prospective employers and data brokers of personal information can use academic records to locate the bar examination scores of applicants and make hiring and promotion decisions based on the information found, even though the State Bar's policy is not to share the information because of the risk of unwarranted inferences being drawn from the data (a risk the trial court specifically found in the litigation). This puts test-takers at a disadvantage because at test time, many test-takers did not necessarily attempt to get the highest score possible as much as to achieve a passing score. The Bar exam is a pass/fail test.

A law school can use the information to match bar scores to students and then assess professors based on student performance. For example, student scores on the Constitutional Law part of the exam, if revealed, could impact the student's Constitutional Law professor. Each of these uses creates a motive to try and re-identify data, even though the consequence in this case is not to the student who is the subject of the re-identified data.

Even more disturbingly, the 11-Anonymity Protocol leaves Black, Hispanic, and Asian students particularly vulnerable to re-identifications, imposing an unfair and disproportionately adverse impact on underrepresented minorities.

Even incorrect matches may still cause harm to individuals. For example, suppose a unique match reveals a low score or numerous failed attempts to pass the Bar. The implication or potential inference that the scores may belong to that individual remains, even if false. Alternatively, if the data reveals that a job applicant is one of three individuals, two of whom barely passed the bar on the second attempt while the third took six months off and then scored highly on the first attempt, an employer could easily decide not to hire any of the three for fear that the individual hired would be one of the two lower performers. Of course, an employer could ask how many times an individual took the exam, but without the ability to perform a re-identification, there is no way for the employer to validate the answer; the individual has plausible deniability.

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018. <http://techscience.org/a/2018111301>

Because of these potential harms and the kinds of unique and small group re-identifications discussed, the 11-Anonymity Protocol fails our litmus tests.

The Plus Protocol

We find it technically reasonable to execute the Plus Protocol, but it does have grave privacy vulnerabilities. One critical problem is that the Plus Protocol does not provide k -anonymity for $k=11$, as asserted by the Sander Team. Another is that the number of unique records remaining in the data (31 percent) dwarfs the old HIPAA standard (0.04 percent). As a result, we could demonstrate small group re-identifications and the potential for personal harms. Therefore, results from our litmus tests provide scientific objections to sharing the Plus Dataset. Below are the litmus test results.

The Plus Protocol starts with the 11-Anonymity Dataset and then rounds GPAs and randomly drops 25 percent of the records. Our assessment of the Plus Protocol involves our revisiting the elements of our assessment of the 11-Anonymity Protocol to see what impact these changes may have on outcomes.

Step 1 of the Plus Protocol is to implement the 11-Anonymity Protocol. One way to execute Step 2 is to make a column of random numbers using the function RAND() or RANDBETWEEN(). Sort the data based on the random numbers and delete the first 25 percent of rows. The ROUND() function within an IF() statement can round the GPAs to one decimal place if the number is on a 4- or 5-point scale and to no decimal places if it is on a 100-point scale.

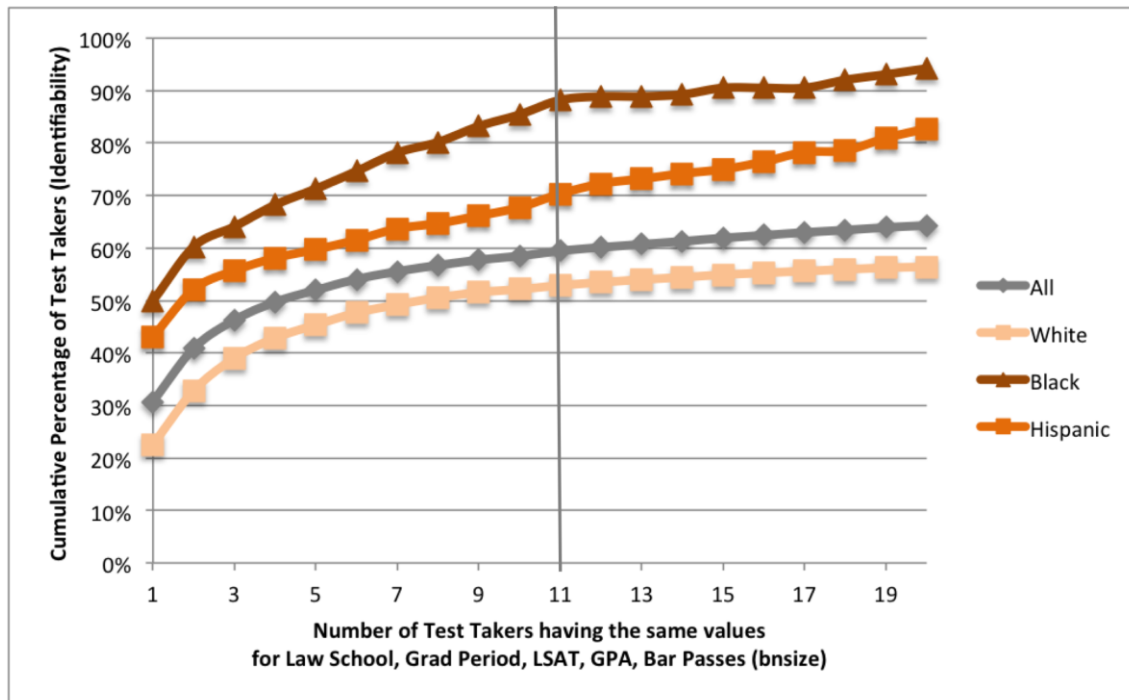
The last step erases each uniquely occurring GPA. This requires a series of advanced steps. First, make a pivot table to identify the GPA values that are unique. Put those values in a column ("uniques"). Then, make a column beside the GPAs; this will be used to identify which GPAs to erase. The new column will contain a nested IF() statement that uses ISERROR() and VLOOKUP, such as: IF(ISERROR(VLOOKUP(gpa, uniques, 1, FALSE))), "OKAY", "ERASE"). Then sort the new column and erase the GPA values noted as ERASE. The pivot table and VLOOKUP are advanced commands used sequentially. The command ISERROR() is uncommon, but not advanced. So, the Plus Protocol is technically reasonable to execute.

The Plus Dataset had 98,932 records. We counted the number of test-takers having the same values for *lawschool*, *gradPeriod*, *lsat*, *gpa*, *race*, and bar passes and found that 30,385 (31 percent) were unique, 57,879 (59 percent) were in binsizes less than 11, and 63,254 (64 percent) were in binsizes less than 20. Figure 40 itemizes these values by binsize and race/ethnicity.

Overall, the percentages of unique records dropped from 46 percent in the 11-Anonymity Dataset to 31 percent in the Plus Dataset, but not for Blacks or Hispanics. In the 11-Anonymity Dataset, the percentage of unique records for Whites was 42 percent; that dropped to 22

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018. <http://techscience.org/a/2018111301>

percent in the Plus Dataset. But the percentage of unique Blacks and Hispanics stayed about the same: 54 percent in the 11-Anonymity Dataset and 50 percent in the Plus Dataset for Blacks, and 46 percent in the 11-Anonymity Dataset and 43 percent in the Plus Dataset for Hispanics.



	%All	%White	%Black	%Hispanic	%Asian	%URM	%White/Asian
binsize	(n=98932)	(n=61438)	(n=3580)	(n=5579)	(n=12154)	(n=1103)	(n=726)
1	31%	22%	50%	43%	31%	47%	48%
2	41%	33%	60%	52%	39%	55%	72%
3	46%	39%	64%	56%	42%	56%	86%
4	50%	43%	68%	58%	44%	59%	92%
5	52%	45%	71%	60%	46%	62%	96%
6	54%	48%	75%	62%	48%	63%	99%
7	56%	49%	78%	64%	49%	65%	100%
8	57%	51%	80%	65%	49%	65%	
9	58%	52%	83%	66%	50%	66%	
10	59%	52%	85%	68%	51%	66%	
11	59%	53%	88%	70%	52%	71%	
12	60%	53%	89%	72%	53%	72%	
13	61%	54%	89%	73%	53%	75%	
14	61%	54%	89%	74%	54%	77%	

	%All	%White	%Black	%Hispanic	%Asian	%URM	%White/Asian
binsize	(n=98932)	(n=61438)	(n=3580)	(n=5579)	(n=12154)	(n=1103)	(n=726)
15	62%	55%	91%	75%	55%	81%	
16	62%	55%	91%	76%	55%	83%	
17	63%	56%	91%	78%	57%	86%	
18	63%	56%	92%	79%	58%	86%	
19	64%	56%	93%	81%	58%	89%	

Figure 40. Re-identification risk in Plus Dataset for test-takers having the same law school, graduation period, LSAT, GPA, and bar passes (first or multiple attempts) for k from 1 to 19. Reported by races and all races. URM = Black and Hispanic.

The Plus Protocol purported to adhere to k -anonymity where $k=11$; if so, the number of test-takers in bins of sizes less than 11 would be 0 and not 30,385 (31 percent). Therefore, we found vulnerability with the 11-Anonymity Dataset. Almost one-third of the records are unique.

Example 6. We provide an example similar to Example 1 that also uses Matching Plan 10's combination of fields: *gradPeriod*, *lawschool*, *race*, and bar passes.

Five records in the 11-Anonymity Dataset ambiguously describe attorneys who were 2000-2002 Pepperdine graduates and who passed the Bar on more than one attempt. The race "Other," assigned to these records, indicates that they are not Black, White, or Hispanic and are likely to be Asian. Similar to Example 1, we use a computer program to associate race with last names we found in the commencement programs, looked up those names designated as Asian in the Attorney Dataset, and then compared Bar admission dates to graduation dates. We matched 8 names to the 5 records. This is yet another example of a small group re-identification of size less than 11 in the 11-Anonymity Dataset. What happens to these records in the Plus Dataset?

We found all the records present. We then repeated the step of randomly erasing 25 percent of the records, and no changes to any of these records resulted. This is a persistent example of a small group re-identification of size less than 11 in the Plus Dataset.

Recall Example 1. In the 11-Anonymity Dataset were 11 Hispanic attorneys who graduated from Pepperdine (a Class One school) during the same 3-year graduation period, of which 4 were found to have passed the bar the same year as graduation and 5 to have passed later. Using the Attorney Dataset, we put 4 names to the group of 4 records and 5 names to the group of 7 records.

As shown above, the Plus Protocol might not reduce the number of records appearing in the Dataset, but what if it did? Suppose 1 of the 4 (25 percent) records for those who passed the

bar the same year as graduation had been erased and that 2 records from the group of 7 also had been erased. The names would remain unchanged, so we end up with 4 names matched to 3 records in one group and 5 names matched to 5 records in the second group. So, even with our deliberately erasing 25 percent of the records in this re-identification, small-group matches remained. The Plus Protocol did not work at thwarting re-identifications of groups having fewer than 11 individuals.

Re-identifications involving LSATs may similarly rely on whether the record is erased. Recall that Example 2 used the fields *gradPeriod*, *lawschool*, *lsat*, and *bar* passes. The Plus Protocol changes GPAs but leaves LSATs unchanged. So, the Plus Protocol might drop some LSAT matches when randomly erasing rows, but not necessarily. The odds of a record for a given LSAT being dropped are 1 in 4. This could reduce the number of re-identifications or increase the number of false re-identifications, but it does not stop re-identifications from occurring.

Other examples that we discussed previously involved GPAs. What happens with GPAs? We recoded GPA values in the UC Davis Dataset to see how uniqueness changes as the number of significant digits change. Figure 41 shows that, as the precision of GPA values changes, the number of unique records drops. The drop in unique matches seen when the GPA is rounded from 2 digits to the right of the decimal point to one digit, as in the Plus Protocol, is dramatic. Clearly this step reduces the number of unique re-identifications based on GPA. Just in case, the Plus Protocol, after redacting GPAs, then erases any remaining unique GPAs. In Figure 41, that would mean that 2 unique GPAs would be erased.

	GPA 3 digits (e.g., 3.123)	GPA 2 digits (e.g., 3.12)	GPA 1 digit (e.g., 3.1)
GPA	1445 (72%)	216 (11%)	2 (0%)
LSAT	1959 (98%)	1602 (80%)	500 (25%)
Race	1675 (84%)	631 (32%)	60 (3%)
Bar	1564 (78%)	376 (19%)	11 (1%)
Race, Bar	1726 (86%)	801 (40%)	110 (5%)

Figure 41. Number and percentage of unique occurrences of combinations of fields in the UC Davis Dataset for graduation periods with varying GPA scales. The value for Bar is pass/fail. There is a total of 2,001 records. The intersection of a row and column report the number of unique records found with that combination of fields. Each combination includes the graduation period.

Suppose we removed all GPA values. Then, 6 of the 14 primary matching patterns (Figure 35) would remain. Many re-identification strategies are possible using those matching patterns; one is demonstrated in Example 6. So, the Plus Protocol, like the 11-Anonymity Protocol, leaks re-identifications and does not keep its *k*-anonymity promise (that no individual or record would be associated with a group of less than 11).

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018. <http://techscience.org/a/2018111301>

All the harms discussed in the 11-Anonymity Protocol remain. None of the changes made by the Plus Protocol reduces harms. Quite the opposite, the Plus Protocol increases potential harms to minorities because their records are even more likely to be unique than those of their White counterparts. In the Plus Dataset, 50 percent of all Black records are unique, 43 percent of all Hispanic records are unique, and 31 percent of all Asian records are unique. But only 22 percent of all White records are unique. That means a Black or Hispanic attorney is about twice as likely to be uniquely associated with a record than a White attorney.

The Plus Protocol fails our litmus tests because of the potentially harmful unique and small group re-identifications that remain in the data and its disparate impact on minorities.

The Enclave Protocol

We find it technically reasonable to execute the Enclave Protocol, but it too has grave privacy vulnerabilities. One critical problem is that the Enclave Protocol does not provide k -anonymity for $k=5$, as asserted by the Sander Team. Another is that part of its protection is a privacy filter on information leaving the safe room. However, sensitive and harmful information can still leave regardless of the filter. Therefore, results from our litmus tests provide scientific objections to sharing the Enclave Dataset in a safe room, as described by the Enclave Protocol. Below are details for each finding by litmus test.

The Enclave Protocol is a version of the 11-Anonymity Protocol that claims to achieve k -anonymity for $k=5$ instead of $k=11$. All other technical data specifications remain the same. So, for the same reasons that we found the 11-Anonymity Protocol technically reasonable to execute, we find the Enclave Protocol technically reasonable to execute.

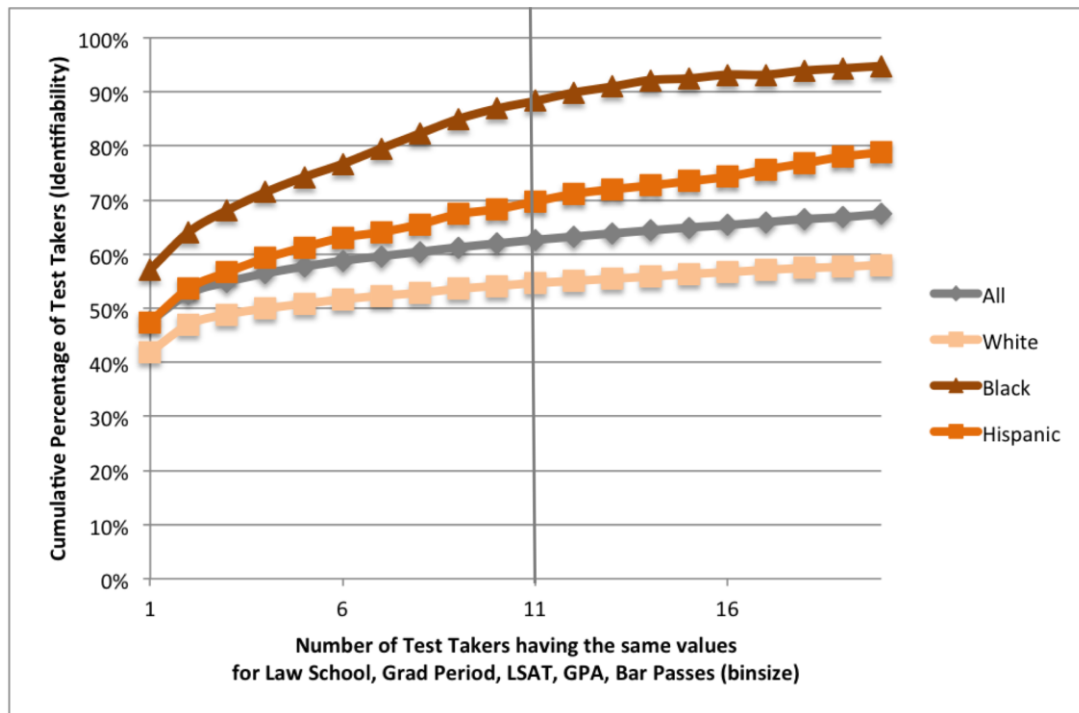
The Enclave Dataset had 128,659 records. We counted the number of test-takers having the same values for *lawschool*, *gradPeriod*, *lsat*, *gpa*, *race*, and *bar passes* and found that 61,084 (47 percent) were unique, 79,727 (62 percent) were in binsizes less than 11, and 86,012 (67 percent) were in binsizes less than 20. Figure 42 itemizes these values by binsize and race/ethnicity.

As was the case with the 11-Anonymity Dataset and the Plus Dataset, results are worse for Blacks and Hispanics in the Enclave Dataset. For Blacks, we found that 2,665 (57 percent) were unique, 4,059 (87 percent of all Black test-takers) were in binsizes less than 11, and 4,407 (94 percent) were in binsizes less than 20.

And, for Hispanics, we found that 3,653 (47 percent) were unique, 5,260 (68 percent of all Hispanic test-takers) were in binsizes less than 11, and 6,010 (78 percent) were in binsizes less than 20.

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018.

<http://techscience.org/a/2018111301>



	%All (n=128659)	%White (n=76294)	%Black (n=4670)	%Hispanic (n=7700)	%Asian (n=11263)	%URM (n=1058)	%White/Asian (n=4135)
1	47%	42%	57%	47%	37%	53%	58%
2	53%	47%	64%	54%	42%	60%	65%
3	55%	49%	68%	57%	46%	66%	69%
4	56%	50%	71%	59%	48%	73%	71%
5	58%	51%	74%	61%	51%	76%	74%
6	59%	52%	77%	63%	52%	83%	76%
7	60%	52%	80%	64%	53%	83%	77%
8	60%	53%	82%	66%	54%	85%	78%
9	61%	54%	85%	67%	55%	85%	79%
10	62%	54%	87%	68%	56%	87%	80%
11	63%	55%	88%	70%	57%	90%	80%
12	63%	55%	90%	71%	57%	92%	81%
13	64%	55%	91%	72%	58%	92%	81%
14	64%	56%	92%	73%	59%	92%	83%
15	65%	56%	93%	73%	61%	93%	84%
16	65%	57%	93%	74%	61%	93%	84%
17	66%	57%	93%	76%	61%	93%	85%
18	66%	57%	94%	77%	62%	95%	85%
19	67%	58%	94%	78%	63%	95%	85%

Figure 42. Re-identification risk in Enclave Dataset for test-takers having the same law school, graduation period, LSAT, GPA, and bar passes (first or multiple attempts) for k from 1 to 19. Reported by races and all races. URM = Black and Hispanic.

The Enclave Protocol purported to adhere to k -anonymity where $k=5$; if so, the number of test-takers in bins of sizes less than 5 would be 0 and not 72,665 (47 percent). Therefore, we found vulnerability with the Enclave Dataset. Almost half the records are unique.

Of course, the Enclave Protocol has an additional guard. A human inspects all information leaving the safe room so that visitors are limited to entering the room with paper, pens, and manuals about the dataset and leaving with limited materials that do not contain data values, such as copies of programs used, and cross-tabulations with no table having fewer than 20 test-takers.

Example 7. We provide another example, similar to Examples 1 and 6, that also uses Matching Plan 10's combination of fields: *gradPeriod*, *lawschool*, *race*, and bar passes.

In each of the protocols, a larger percentage of Blacks have unique records than any other group: 54 percent in the 11-Anonymity Protocol, 50 percent in the Plus Protocol, and 57 in the Enclave Protocol. And almost 90 percent of all the records for Blacks were in groups having no more than 20 individuals, regardless of the protocol. These numbers suggest that anyone with access to any of these protocol datasets can learn small-group information about Black attorneys, merely by knowing the attorney is Black.

The Enclave Protocol is particularly vulnerable to re-identifications of targeted attorneys. There is less risk of large-scale re-identification, but risks to targeted individuals do exist.

Because the information in the enclave is public, there are no limits on who can enter the safe room. A reporter, private investigator, or representative from a legal data mining company can enter the enclave with basic knowledge of an attorney's name, law school, bar admission date, and graduation date, and make inferences about the attorney's LSAT, GPA, and bar scores. There are many ways to get the requisite information. An attorney's name, law school, and Bar admission date are public information on all practicing attorneys at the State Bar's website and many other places. Graduation date seems readily available on most attorneys' online biographies and resumes. Here are re-identification examples that could be done in the safe room, notwithstanding the attempted protections of the Enclave Protocol.

By visual inspection, locate a group of Black attorneys online who all graduated from the same Class One or Class Two law school and in the same graduation period. Look up the year each was admitted to the Bar in the Attorney Dataset to infer the number of times each may have taken the bar exam. (Visit the safe room.) Look up the records for Black test-takers from the school in the time period. The names now match those records.

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018. <http://techscience.org/a/2018111301>

We found the group of photographs of 5 seemingly Black attorneys that appear in Figure 31. One of them, Marcia Randle, was also Vice President of the McGeorge Black Law Student Association (Figure 30). All of them graduated from the McGeorge School of Law in the same graduation period. Each passed the Bar in the same year as graduation. Exactly 5 such records were found in the dataset, thereby putting these 5 names to those 5 records. This re-identification result would be realized on the 11-Anonymity Dataset. It might possibly occur in the Plus Dataset, in whole or part, depending on whether any of the records were randomly erased. What happens in the Enclave Dataset? If the re-identification is done in the Enclave, the visitor could walk out having memorized critical information about the 5 GPAs, LSATs, or bar scores.

Memory is not the only way for scores to leave the enclave. For example, scores could be encoded in a program where letters represent digits, or where the text of a seemingly relevant comment merely stores values. Steganography is the study of ways to hide information in plain sight – like spies writing in invisible ink, or notes hidden in images [70]. Here is an example of a program comment that could be written inside of the enclave to embed the GPAs 3.1, 3.21, 2.7, 3.45, and 3.31. (A program comment is ignored when the program runs, so comments provide a simple way to hide messages.) When the program leaves the enclave, it passes visual inspection because it just looks like a program, but what also leaves are the private values.

```
# Check for GPAs in the range of 3.1 to 3.21 then 2.7 to 3.45, and finally 3.1 to 3.31.
```

The person could further write the program to actually use those values. As long as the results created a pivot table with more than 20 cells, the results can leave too, even though the comment and not the results contains the critical information.

Example 8. We provide yet another example similar to Examples 1, 6, and 7 that also uses Matching Plan 10's combination of fields: *gradPeriod*, *lawschool*, *race*, and *bar passes*. As in Example 7, we focus on targeted individuals.

Minority attorneys are particularly vulnerable to targeted re-identifications, but frankly, so are most high-profile individuals, including politicians, attorneys involved in high-profile cases, law enforcement, or business, or those seeking promotion in a competitive setting. Even if the individual's information is in a bin of fewer than 20 individuals, sensitive information could still be learned.

We identified a high-profile individual (Individual X) and searched for the individual's law school, graduation period, and race for those passing the bar the same year as graduation, all of which we learned from public information about Individual X. We found 6 records matching the criteria; see Figure 43. Notice that all the records had GPAs, and the GPA for all of these individuals was less than 3.0. Even without revealing which record is Individual X's, the data reveals an attribute about all of them: that they earned less than a 3.0 GPA in law

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018. <http://techscience.org/a/2018111301>

school, thereby strongly suggesting that the targeted individual had less than a 3.0 average in law school. The assumption would be that Individual X had a GPA less than 3.0, and Individual X would be left to confirm or refute the assumption.

Law School	LSAT	GPA	Grad Period
Law School Y	36	2.93	Same
Law School Y	25	2.67	Same
Law School Y	31	2.3	Same
Law School Y	24	2.68	Same
Law School Y	24	2.46	Same
Law School Y	38	2.14	Same

Figure 43. Re-identification in Enclave Dataset of test-takers having the same law school, graduation period, race and bar passes. Learned are LSAT and GPA scores.

This inference is possible using the 11-Anonymity Dataset. It might remain in the Plus Dataset, in whole or part, depending on whether all the records were randomly erased. What happens in the Enclave Dataset? The data remains, of course. And the visitor to the enclave need only walk out with the knowledge that the high-profile individual had a GPA less than 3.0.

The Enclave Protocol fails our litmus tests because of the potentially harmful unique and small group re-identifications that remain in the data even with the safe-room filtering.

The Standardized Protocol

The Excel knowledge required to execute the Standardized Protocol is beyond our standards of what is technically reasonable because numerous steps require nested advance topics in Excel and doing them manually is too time consuming. We also found that the Standardized Protocol failed our privacy litmus tests. Here are the details.

The first two steps of the Standardized Protocol involve dropping rows from unaccredited schools and those who graduated prior to 1985 and between 1999 and 2005. These activities involve sorting data, which is an advanced Excel topic, and then deleting highlighted rows, which are basic Excel operations. These are allowable effort. For example, sorting the entire dataset by increased graduation year (*gradYr*) requires a few mouse clicks. The unwanted records appear at the beginning of the dataset, so highlighting those records and then deleting the selected rows can be done with a few clicks. Similarly, the records of unaccredited schools can be located after sorting the dataset by *lawschool* and then deleting the rows belonging to those 37 schools.

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018.
<http://techscience.org/a/2018111301>

Step 3 of the Standardized Protocol recodes *race* into "Black," "White," "Hispanic," and "Other." As described earlier, this recoding can be done in Excel using nested IF statements in a new column, which is allowable effort. The computed values can then be copied and pasted into the same column so that only the values remain and not the IF statement.

Steps 4 and 5 are similar to Step 3. Add a new column *gradPeriod* and then either use nested IF statements as in Step 3, this time to aggregate graduation year (*gradYr*) into 3-, 4- or 5-year periods, or, alternatively, sort by *gradYr* and then manually fill down the ranges for *gradPeriod*. Either way, this is allowable effort.

Step 6 just adds some columns, which is easily done by writing the name of the fields at the top of adjacent unused columns. The new columns add the fields *zLSATyr*, *zLSAT*, and *zLSATtype*. Later, we will discuss how Step 8 populates these fields using values computed in a separate table constructed in Step 7.

So far, the first 6 steps use allowable effort, but beginning with Step 7, the Standardized Protocol starts getting more complicated.

Step 7.1 makes sure that LSAT scores are within one of two valid ranges: 10-48 or 120-180. Sort the data on *lsat* and remove all records that do not have values within these ranges. This requires manually traversing approximately 100,000 records and making appropriate deletions. A faster way is to insert a blank row after the heading row, search for the smallest legal LSAT score, 10, insert a new row just before that row, and then delete the rows between and including the blank rows to remove all records having LSAT scores less than 10. Repeat this process to delete records between 48 and 120 and then just delete those records having LSAT scores greater than 180. All these steps are allowable.

Step 7.2.1 involves the construction of a lookup table that holds the average LSAT value for each year; we term this the Yearly Average Lookup Table. After Step 3, the data contains 17 years of records, from 1985 to 1998 and 2006 to 2008, so the Yearly Average Lookup Table will have 17 rows, one for each graduation year, and contain the average LSAT score of test-takers having that graduation year. The paragraphs below describe how to construct the Yearly Average Lookup Table.

First, sort the dataset by graduation year (*GradYr*). Then, make a separate table having the columns *GradYr*, *zLSATyr* and *stdev* (e.g., Figure 45d). These are the headings for the Yearly Average Lookup Table – namely, graduation year and the average LSAT score and standard deviation of test-takers having that graduation year. Next, enter the 17 distinct years under *GradYr* using visual inspection or a pivot table to generate the list of years. Then, use the AVERAGE function on *lsat* values in the ranges of cells for each year to compute the average *lsat* of test-takers for that year. A visual traversal or search by year through the *GradYr* column in the dataset will reveal the row numbers for the first and last test-taker of each year. For example, assume *lsat* is column D in the data and that there are 1,868 records for

1985, from row 2 to 1869, and 4,770 records for 1986, from row 1870 to 6639. The value in the Yearly Average Lookup Table for 1985 is AVERAGE(D2:D1869) and for 1986 is AVERAGE(D1870:D6639). Repeat this for each of the remaining years, using the row indices appropriate for records having that graduation year. Then, replicate these formulae using STDEV instead of AVERAGE to compute the accompanying standard deviation values.

Here is an alternative way to construct the Yearly Average Lookup Table, but it uses nested advanced topics. Instead of looking up the indices by hand, we use a pivot table to provide the indices. First, we create a pivot table on *GradYr* to show the graduation years and number of records for each graduation year. Assume we copy the pivot table results with *GradYr* in column T and the total number of test-takers for the year in column U, and results start at row 2 (see columns T and U in Figure 45a). From our examples, the first row of the pivot table will report 1,868 records for 1985, 4,770 records for 1986, and so on. We add 2 columns with labels *IndexFrom* and *IndexTo* to get indices for the INDIRECT function to use (see columns V and W in Figure 45a). We use a formula and values from the pivot table to produce the indices for the *Isat* column (D) that correspond to each graduation year. For 1985, the range is from rows 2 to U2+1, so we write “=D2” in the *IndexFrom* column and the concatenation of D to the value in U2+1 (or “(“D”&(U2+1))” in the *IndexTo* column. For 1986, the range begins at 1870, which is 1868 in U2+2. The range ends at the row that is the sum of 1868 and 4770+1. So, we write “(“D”&(U2+2))” in V3 and “(“D”&(SUM(U\$2:U3)+1))” in V4. Finally, we keep the pattern going where the number of the beginning row in the range is the sum from U2 to the current year plus 2, and the ending row number is the sum from U2 to the next row plus 1. Figure 45a shows the full list of formulae to produce the indices, and Figure 45b shows the computed results.

Now that the pivot table is extended to have two columns to list the indices of rows to use for each year (Figures 45a and 45b), we can compute the values for the Yearly Average Lookup Table using the INDIRECT function. Figure 45c shows the replicated formula “=AVERAGE(INDIRECT(V2):INDIRECT(W2))” for computing the AVERAGE for each graduation year and “=STDEV(INDIRECT(V2):INDIRECT(W2))” for computing the standard deviation (STDEV).

	Pivot Table with Indices (Formulae View)			
	[T]	[U]	[V]	[W]
[1]	GradYR	Count	IndexFrom	IndexTo
[2]	1985	1868	=“D2”	=("D"&(U2+1))
[3]	1986	4770	=("D"&(U2+2))	=("D"&(SUM(U\$2:U3)+1))
[4]	1987	4891	=("D"&(SUM(U\$2:U3)+2))	=("D"&(SUM(U\$2:U4)+1))
[5]	1988	4682	=("D"&(SUM(U\$2:U4)+2))	=("D"&(SUM(U\$2:U5)+1))
[6]	1989	4639	=("D"&(SUM(U\$2:U5)+2))	=("D"&(SUM(U\$2:U6)+1))
[7]	1990	4337	=("D"&(SUM(U\$2:U6)+2))	=("D"&(SUM(U\$2:U7)+1))

[8]	1991	2620	=("D"&(SUM(U\$2:U7)+2))	=("D"&(SUM(U\$2:U8)+1))
[9]	1992	614	=("D"&(SUM(U\$2:U8)+2))	=("D"&(SUM(U\$2:U9)+1))
[10]	1993	1649	=("D"&(SUM(U\$2:U9)+2))	=("D"&(SUM(U\$2:U10)+1))
[11]	1994	4854	=("D"&(SUM(U\$2:U10)+2))	=("D"&(SUM(U\$2:U11)+1))
[12]	1995	5113	=("D"&(SUM(U\$2:U11)+2))	=("D"&(SUM(U\$2:U12)+1))
[13]	1996	4932	=("D"&(SUM(U\$2:U12)+2))	=("D"&(SUM(U\$2:U13)+1))
[14]	1997	4943	=("D"&(SUM(U\$2:U13)+2))	=("D"&(SUM(U\$2:U14)+1))
[15]	1998	4064	=("D"&(SUM(U\$2:U14)+2))	=("D"&(SUM(U\$2:U15)+1))
[16]	2006	2088	=("D"&(SUM(U\$2:U15)+2))	=("D"&(SUM(U\$2:U16)+1))
[17]	2007	5344	=("D"&(SUM(U\$2:U16)+2))	=("D"&(SUM(U\$2:U17)+1))
[18]	2008	83	=("D"&(SUM(U\$2:U17)+2))	=("D"&(SUM(U\$2:U18)+1))

(a)

Pivot Table with Indices (Value View)				
	[T]	[U]	[V]	[W]
[1]	GradYR	Count	IndexFrom	IndexTo
[2]	1985	1868	D2	D1869
[3]	1986	4770	D1870	D6639
[4]	1987	4891	D6640	D11530
[5]	1988	4682	D11531	D16212
[6]	1989	4639	D16213	D20851
[7]	1990	4337	D20852	D25188
[8]	1991	2620	D25189	D27808
[9]	1992	614	D27809	D28422
[10]	1993	1649	D28423	D30071
[11]	1994	4854	D30072	D34925
[12]	1995	5113	D34926	D40038
[13]	1996	4932	D40039	D44970
[14]	1997	4943	D44971	D49913
[15]	1998	4064	D49914	D53977
[16]	2006	2088	D53978	D56065
[17]	2007	5344	D56066	D61409
[18]	2008	83	D61410	D61492

(b)

Yearly Average Lookup Table (Formulae View)			
	[P]	[Q]	[R]
[1]	GradYR	zLSATyr	stdev
[2]	1985	=AVERAGE(INDIRECT(V2):INDIRECT(W2))	=STDEV(INDIRECT(V2):INDIRECT(W2))

Yearly Average Lookup Table (Formulae View)			
	[P]	[Q]	[R]
[3]	1986	=AVERAGE(INDIRECT(V3):INDIRECT(W3))	= STDEV (INDIRECT(V3):INDIRECT(W3))
[4]	1987	=AVERAGE(INDIRECT(V4):INDIRECT(W4))	= STDEV (INDIRECT(V4):INDIRECT(W4))
[5]	1988	=AVERAGE(INDIRECT(V5):INDIRECT(W5))	= STDEV (INDIRECT(V5):INDIRECT(W5))
[6]	1989	=AVERAGE(INDIRECT(V6):INDIRECT(W6))	= STDEV (INDIRECT(V6):INDIRECT(W6))
[7]	1990	=AVERAGE(INDIRECT(V7):INDIRECT(W7))	= STDEV (INDIRECT(V7):INDIRECT(W7))
[8]	1991	=AVERAGE(INDIRECT(V8):INDIRECT(W8))	= STDEV (INDIRECT(V8):INDIRECT(W8))
[9]	1992	=AVERAGE(INDIRECT(V9):INDIRECT(W9))	= STDEV (INDIRECT(V9):INDIRECT(W9))
[10]	1993	=AVERAGE(INDIRECT(V10):INDIRECT(W10))	= STDEV (INDIRECT(V10):INDIRECT(W10))
[11]	1994	=AVERAGE(INDIRECT(V11):INDIRECT(W11))	= STDEV (INDIRECT(V11):INDIRECT(W11))
[12]	1995	=AVERAGE(INDIRECT(V12):INDIRECT(W12))	= STDEV (INDIRECT(V12):INDIRECT(W12))
[13]	1996	=AVERAGE(INDIRECT(V13):INDIRECT(W13))	= STDEV (INDIRECT(V13):INDIRECT(W13))
[14]	1997	=AVERAGE(INDIRECT(V14):INDIRECT(W14))	= STDEV (INDIRECT(V14):INDIRECT(W14))
[15]	1998	=AVERAGE(INDIRECT(V15):INDIRECT(W15))	= STDEV (INDIRECT(V15):INDIRECT(W15))
[16]	2006	=AVERAGE(INDIRECT(V16):INDIRECT(W16))	= STDEV (INDIRECT(V16):INDIRECT(W16))
[17]	2007	=AVERAGE(INDIRECT(V17):INDIRECT(W17))	= STDEV (INDIRECT(V17):INDIRECT(W17))
[18]	2008	=AVERAGE(INDIRECT(V18):INDIRECT(W18))	= STDEV (INDIRECT(V18):INDIRECT(W18))

(c)

Yearly Average Lookup Table (Value View)			
	[P]	[Q]	[R]
[1]	GradYR	zLSATyr	stdev
[2]	1985	31.343	8.477
[3]	1986	35.688	6.363
[4]	1987	34.448	6.863
[5]	1988	36.144	6.682
[6]	1989	31.871	6.875
[7]	1990	34.362	6.277
[8]	1991	31.221	7.262
[9]	1992	30.382	11.472
[10]	1993	35.822	10.474
[11]	1994	36.414	6.862
[12]	1995	141.036	17.734
[13]	1996	144.733	39.376
[14]	1997	152.386	25.256
[15]	1998	151.733	17.473
[16]	2006	155.952	11.373

Yearly Average Lookup Table (Value View)			
	[P]	[Q]	[R]
[17]	2007	157.227	9.373
[18]	2008	151.692	23.362

(d)

Figure 44. Pivot Table formulae (a) and values (b) used to list the indices for computing the annual average LSAT scores and standard deviation (d). The formulae that use the pivot table values in (b) to produce the annual values in (d) appear in (c).

The problem with this approach is that we are now nesting the advanced topic of a lookup table with INDIRECT, and the values for INDIRECT come from nesting a pivot table with SUM and concatenation (&). Knowing how to do this is beyond the basic topics covered in basic Excel tutorials, so this alternative approach is not allowable. Of course, the more laborious activity we described earlier, of manually looking up the indices and entering them manually, achieves the same goal and is allowable. It just consumes more keystrokes and mouse activity.

In Step 7.2.2 we compute the standardized LSAT for each individual using the Yearly Average Lookup Table. This is simply the test-taker's *lsat* minus the average LSAT divided by the standard deviation. Values for the average and standard deviation come from the Yearly Average Lookup Table based on the test-taker's graduation year. Specifically, if *gradYr* is column C and *lsat* is column D in the dataset and the Yearly Average Lookup Table has column P for graduation year, column Q for the average LSAT score ($zLSAT_{yr}$), and column R for the standard deviation, then the formula for the test-taker's normalized LSAT on row 2 is $(D2 - VLOOKUP(C2, \$P\$2:\$R\$18,2))/ VLOOKUP(C2, \$P\$2:\$R\$18,3)$. We replicate this formula for all test-takers.

Altogether, the first 7 steps can be achieved with allowable operations.

Step 8 seeks a similar calculation to Step 7, but instead of standardized LSAT scores based on the annualized mean, Step 8 computes standardized LSAT scores per school and graduation year for schools having at least 20 graduates and per school and graduation period for schools having fewer than 20 graduates. Our approach involves making two lookup tables, one based on graduation year and the other based on graduation period. We use a pivot table to decide which table to use. Here are the details.

There are about 3,000 combinations of school names and graduation years to consider. We make a new table called the Yearly Cohort by GradYr Lookup Table, which models the Yearly Average Lookup Table. The new table will have 3 columns: SchoolGroup, zLSAT, and stdev, which for each combination of school and graduation group will include the average (zLSAT) and standard deviation (stdev) for the group. Instead of 17 rows, the new table will have

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018.
<http://techscience.org/a/2018111301>

about 3,000 rows. Then, we make a second table, called the Yearly Cohort by GradPeriod Lookup Table having the same 3 columns and 3,000 rows as the Yearly Cohort by Year Lookup Table.

We begin with the Yearly Cohort by Year Lookup Table by sorting the data by school name and, within school name, by graduation year. Here is how we fill the column SchoolGroup with its 3,000 values. We add a new column to the dataset that is the concatenation of school name and graduation year. Replicate the formula down the column and then paste a copy of the concatenated values of *lawschool* and *gradYr* in the Yearly Cohort Lookup Table as values under the SchoolGroup column. The SchoolGroup column in the Yearly Cohort Lookup Table is now complete. Concatenation (&) is an advanced topic, but it is not nested, so it is allowed.

Make a pivot table of counts of SchoolGroup to identify which combinations of school name and graduation year have 20 or more test-takers and which do not. Then, copy the values of the pivot table into its own lookup table, the LSAT Type Table, having columns SchoolGroup and Total Individuals. Making a pivot table and copying its values into a lookup table involves an advanced topic but is allowable.

Now we have to compute the average and standard deviation for each school group based on the row indices of the groups in the dataset. In Step 7, to find the row indices, we found we could either manually inspect the data to learn the row indices, or nest a lookup table with INDIRECT, where the values for INDIRECT come from nesting a pivot table with SUM and concatenation (&). Visual inspection of the row indices used allowable activities – i.e., only those activities requiring a reasonable level of Excel knowledge, but using the nested operations was not allowable because the level of Excel knowledge required exceeded the criterion. So, we must visually inspect the data to learn the row indices needed to complete the average and standard deviation values in the Yearly Cohort Lookup Table.

As described earlier, the Standardized Dataset from the Sander Team had 85,364 records. These have to be visually inspected to learn the beginning and ending row indices for each {*lawschool*, *GradYr*} group for those groups in the LSAT Type Table having 20 or more test-takers. Moving down the rows of the dataset, highlighting those rows belonging to the same group, provides a good quality control practice but requires more than 85,000 keystrokes (one click per row) just to move through the dataset, and that does not include actually writing the row indices themselves in the formulae. This is almost twice the number of keystrokes allowed (45,000 maximum), so it cannot be done.

Suppose instead of clicking through each row, we use page down to visually detect the row indices. That reduces the navigational keystrokes to about 1,400 for viewing 60 rows on a screen at a time. Once the first and last indices for a group are visible, we enter them into the lookup tables: two indices for each AVERAGE and STDEV formula, which is 4 indices per lookup table. The indices go from 2 to 85365 in the Standardized Dataset, which means the average number of digits in a row index is 4.8. There are 3,000 rows in the lookup table that

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018.
<http://techscience.org/a/2018111301>

would require entering 4 indices into the formulae of each. That is, 4.8 digits x 4 indices x 3,000 groups = 57,600 keystrokes. But the maximum number of keystrokes allowed for the entire protocol is 45,000, and this is just for one table. So, manual processing is not allowable. We can stop without replicating the work needed to do the second table for those groups having less than 20 test-takers, which might be another 50,000 strokes, and computing the standardized LSAT values for test-takers.

Achieving Step 8 by manual means requires too many keystrokes. Step 9 requires a repeat of Step 8 but with GPA instead of LSAT. That means the number of keystrokes for the protocol is about 4 times the maximum allowed, and the alternatives to reduce the manual number of keystrokes requires too much advanced knowledge.

Privacy test. Our litmus test approach allows us to stop if the production of the dataset is technically unreasonable. We just showed that the Standardized Protocol requires skill or time beyond reasonable. But instead of stopping, we will continue anyway and execute privacy litmus tests on the Standardized Protocol. The claim of the Standardized Protocol is that because all explicit school names and GPA and LSAT values no longer appear in the data, then no one should be re-identified. Let's test whether we can associate names with groups of 20 or fewer records in a dataset produced by the Standardized Protocol.

Step 10 of the Standardized Protocol (Figure 15) computes standardized GPA values for groups having at least 20 test-takers sharing the same school name and graduation year. For those groups having less than 20 test-takers, in Step 10.2, the protocol expands the graduation year to a period of 3 to 5 years and computes standardized GPA values for groups having at least 20 test-takers with the same school name and graduation period. But if the number of test-takers having the same school name and graduation period is also less than 20, then the protocol blanks out the value for the standardized GPA (Step 10.3). The same approach applies to LSAT values in Step 8. This means a test-taker whose record has blank standardized GPA and LSAT values in the Standardized Dataset is one of fewer than 20 test-takers having the same school and 3- to 5-year graduation period. While these records have blank standardized GPA and LSAT values, they still have the test-taker's race (Step 3), whether the law school was in California (Step 12), and the bar passage result and bar scores. How might we find these individuals and then match them to records in the dataset to learn personal bar scores?

Example 9. We provide an example that uses Matching Plan 14's combination of fields: graduation year, law school name, GPA, LSAT, race, and bar passes as available in the Standardized Dataset, having blanked out values for standardized GPA and LSAT scores.

Here is a re-identification strategy. We identify attorneys in the Attorney Dataset who passed the bar in the same 3- to 5-year graduation periods as those listed in the Standardized Protocol (Figure 15) but where: (1) fewer than 20 attorneys from the same law school passed the bar within the time period; (2) each has a last name most often given to Asians; (3) each

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018.
<http://techscience.org/a/2018111301>

graduated within the same graduation period listed in the Standardized Protocol; and (4) each graduated from an accredited U.S. law school not in California. We then compute how these named records match to records in the Standardized Dataset.

The oldest graduation period in the Standardized Protocol is 1985-1989. During that time, 26,818 attorneys in the Attorney Dataset passed the Bar. Of these, 883 graduated from 114 accredited non-California USA schools [71], where each school had fewer than 20 bar passers within 1985-1989. Among these, 17 attorneys had last names most often associated with Asians in the U.S. Census data [66]. Below are the names of the 17 attorneys.

Attorney Name	Graduation Year
Kyoung Joo Sue Oh	1987 See [72]
Rebecca Lyn Chiao	
Rita Mankovich Irani	1982 See [73]
*David Harris Solo	1987 See [74]
Christopher Paul Wee	
Sangeeta Jain	1988 See [75]
Lila Leianne Choy	
Joseph Michael Gabriel	1985 See [76]
John William Lau	
Pomphyilia G. Baker Bow	
*Judith Michiko Sasaki	1985 See [77]
Liem Hieu Do	1983 See [78]
Victoria Iusam Chau	1986 See [79]
Katherine Landey Pang	1982 See [80]
Steven Susumu Kondo	1978 See [81]
Christopher John Van Son	
Diep Ngoc Nguyen	1985 See [82]

Figure 45. List of the 17 named attorneys who passed the bar between 1985 and 1989 from accredited, non-California U.S. law schools that had fewer than 20 bar passers in 1985-1989 and whose last name is most often associated with Asians in U.S. Census data [66]. We questioned “Joseph Gabriel” and found his image appeared to be an Asian male [83]. The 7 highlighted rows are attorneys who had graduation years between 1985 and 1989. These 7 named attorneys match to a small group of rows in the Standardized Dataset. *Two attorneys had graduation and bar passage dates in the same year, which means they must also match to the even smaller group of records in the Standardized Dataset of those who passed the bar on the first attempt.

In the Standardized Dataset, there will be a comparable number of Hispanic and Asian (“Other”) bar passers who graduated between 1985 and 1989 from an accredited non-California school having fewer than 20 test-takers between 1985 and 1989. However, there is

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018.
<http://techscience.org/a/2018111301>

not a one-to-one correspondence between the list of attorneys from the Attorney Dataset and the list of test-takers from the Standardized Dataset. Some of the attorneys who passed the bar between 1985 and 1989 did not necessarily graduate during that same time period, and similarly, some of the test-takers who graduated between 1985 and 1989 and passed the bar did not necessarily pass between 1985 and 1989. Attorneys who overlap these two lists both graduated and passed the bar between 1985 and 1989.

We searched online for the graduation date of each of the 17 attorneys having last names associated with Asians and learned that 7 had graduation dates between 1985 and 1989 and 4 did not. We were unable to learn the graduation dates of 6 of the attorneys. We know these 7 named attorneys must match at least 7 of the records in the Standardized Dataset for test-takers who passed the bar, have "Other" race, graduated between 1985 and 1989, and have blank values for standardized GPA and LSAT scores. The number of records in the Standardized Dataset will be larger than 7 and will match ambiguously to these 7 named attorneys, allowing us to learn a range of applicable bar scores. Further, 2 of the 7 named attorneys passed the bar the same year as graduation (Figure 45), so these two named attorneys may match to an even smaller number of records of those who passed on the first try.

This re-identification strategy also applies to attorneys having last names given most often given to Hispanics and also to the other time periods, 1990-1994, 1995-1998, and 2006-2008, listed in the Standardized Protocol.

Example 10. We provide another example that uses Matching Plan 14's combination of fields: graduation year, law school name, GPA, LSAT, race, and bar pass result as available in the Standardized Dataset with school names removed and GPAs and LSATs replaced with standardized scores.

In this re-identification strategy, we show how someone who holds external, but similar, data from a Class One California law school can use the Standardized Dataset to learn the bar scores of test-takers. The external dataset may already contain the name of the test-takers, or it may contain other information sufficient to put names to the records using one of the re-identification strategies we described earlier (Example 6). Obviously, any Class One California law school already holds this data, and so does anyone who obtained the data from the school. As described earlier, Professor Sander acquired this kind of information using other public record requests. In fact, Professor Sander obtained the UC Davis Dataset through a public records request.

In each of the graduation periods listed in the Standardized Protocol, namely, 1985-1989, 1990-1994, 1995-1998, and 2006-2008, UC Davis had hundreds of bar passers. So, records of test-takers from UC Davis in the Standardized Dataset appear with LSAT and GPA scores standardized annually (Steps 8.1 and 10.1 in Figure 15).

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018.
<http://techscience.org/a/2018111301>

We can independently compute the standardized GPA and LSAT scores for each UC Davis student in the years 1995-1998 and 2006-2008 using the UC Davis Dataset. Beyond GPA and LSAT, this dataset also includes race and bar passage result.

Stata usually reports standardized scores with 7 digits to the right of the decimal place (e.g., 0.4367097). After we compute the standardized scores, we have two 8-digit values, race, and bar passage result for each UC Davis test-taker per year. This combination of values may be unique for each student, so that the combination can serve as a unique identifier in the Standardized Dataset, thereby allowing us to associate a single UC Davis student record in the UC Dataset to a specific record in the Standardized Dataset. The critical question is the uniqueness of these identifiers. If they are unique, will they remain unique even when merged with data from other years from UC Davis and with standardized data from other schools?

We computed standardized GPA and LSAT scores for UC Davis for each of the years 1995-1998 (4 years) and 2006-2008 (3 years). We then compared the results from all 7 years to see how unique the standardized GPA and LSAT scores, combined with race and bar result, may be. UC Davis had a total of 907 test-takers in 7 years. Of these, 804 (89 percent) had unique standardized GPA scores and only 42 (5 percent) had unique standardized LSAT scores. But 897 (99 percent) had unique GPA and LSAT scores combined. When we added race and bar pass result to the standardized GPA and LSAT scores, 903 values, or all but four (statistically 100 percent), were unique.

Of course, as the number of digits to the right of the decimal point decreases, the number of unique combinations can decrease. The UC Davis Dataset reports GPA with 3 digits to the right of the decimal, and the Bar Dataset has 2 digits, as discussed earlier. Repeating our calculations using 2 digits to the right of the decimal on the 907 test-takers, we found that 305 (34 percent) had unique standardized GPA scores, and 842 (93 percent) had unique GPA and LSAT scores combined. When we added race and bar pass result to the standardized GPA and LSAT scores, 877 (97 percent) remained unique. During the legal proceeding, we demonstrated how unique these combinations were by locating UC Davis records in the Standardized Dataset. We found those combinations unique in the Standardized Dataset for "in-California" schools for all the records tested, allowing us to uniquely associate bar scores with UC Davis student records.

Of course, other re-identification strategies seem possible too. Regardless, these exemplars are sufficient to show that the Standardized Protocol fails our privacy litmus tests because of the potentially harmful unique and small group re-identifications that remain in the data even when school names are removed and LSAT and GPA scores replaced with standardized scores.

Discussion

The Sander Team, an experienced group of data privacy practitioners, gave us four protocols that were supposed to protect privacy and be technically reasonable to implement. The two surprising outcomes were that none of the protocols provided the k -anonymity protection promised. Worse, about half of all the records were unique when none should have been. A final protocol constructed a statistical database by replacing GPA and LSAT scores with standardized values. Even then, sufficient information remained in the data to allow names to be associated with the records.

Three of the protocols claimed to provide k -anonymity protection. None did. For a protocol to provide k -anonymity, we should visually see at least k copies of each record across all the fields, and at least k records should be suppressed altogether. The protocols attempted to k -anonymize the law school, graduation period, and race fields only for those who passed the bar. This led to two problems.

First, you cannot elect to k -anonymize records having some value for a field and no other values for the same field. Either the field is subject to k -anonymization or it is not. In the case of the protocols, they choose only records of individuals who passed the bar. There are three values for passing the bar. An individual either passed the bar (1) on his first attempt, (2) on multiple attempts, or (3) not at all. The protocols enforce k -anonymity on the first two jointly, regardless of the value being the first or multiple attempts. This fails k -anonymity and leaves information vulnerable, as we demonstrated in Example 1 in which 11 Hispanic attorneys who graduated from Pepperdine during the same 3-year graduation period split into 4 who passed on their first attempt, and 7 who passed after more than one attempt. We then put names to the group. Notice how each group is less than size 11. See also Examples 6, 7, and 8.

The second problem with the attempted k -anonymization is the lack of inclusion of all fields, or any proof that the selected fields are the only fields necessary to provide the k -anonymity guarantee. For example, we provided examples that re-identified individuals based on adding one more field, namely if they passed the bar on the first or on multiple attempts (Examples 1, 6, 7, and 8). We showed that all the fields are knowable (e.g., Figures 20, 26, and 29), and therefore all of the fields should have been part of the k -anonymization. All fields should be subject to the k requirement unless proof exists that excluding a field would not impact any link attempt. k -anonymity was first introduced in 1997. Since then, there has been an explosion in available information, so enforcement across all the fields is particularly important today.

The Sander Team asserted that some of the protocols were compliant with the HIPAA Safe Harbor. It is not clear how to make such a comparison. The fact that more than half the records (50 percent) were unique dwarfs the 0.04 percent of allowable uniques adopted by some as the old HIPAA Safe Harbor standard [24]. In fact, having more than 50 percent of the

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018. <http://techscience.org/a/2018111301>

records being unique is more than double a recent finding that 20 percent of records in a HIPAA Safe Harbor-compliant database were unique[26].

HIPAA also requires protection against re-identification using information the recipient may know or hold. The idea is that if the recipient of the data could have other information on which they could link to further re-identify individuals in the data, then the recipient may not receive that data [21]. The UC Davis Dataset, which Professor Sander received under a prior public record request, is indicative of the kind of datasets he received under prior public record requests to law schools from around the country [69]. As we showed, having these datasets makes it even easier to re-identify test-takers at scale. This concern is not limited to Professor Sander. Once a dataset is made available to Professor Sander under a public records request, others could request the same information from those law schools (or someone could download from his project's website a copy of some of the data that Professor Sander received [69]). So, a data analytics company or a data broker, as examples, could acquire the same data, do re-identifications, and then sell the resulting data product or a service based on the re-identified data.

As stated earlier, the HIPAA Safe Harbor has a geography requirement, namely that the smallest reportable geographic subdivision is the first 3 digits of the ZIP (postal) code (unless the three-digit zip code contains fewer than 20,000 individuals, in which case it is reported as 000). In comparing their protocols to HIPAA Safe Harbor, the Sander Team does not address geography. The name of the law school seems to be the field related to geography because it geographically situates test-takers. If so, the implication is that the name of any law school having less than 20,000 students should be suppressed. One protocol did suppress all law school names, but the other 3 protocols only suppressed the names of Class Three law schools, leaving the names of the more popular Class One and Class Two schools.

One of the protocols made no promises of k-anonymity or of a HIPAA level of protection because of a false belief that transforming the data into a statistical database of standardized scores means no one can be re-identified. Producing a statistical database requires time and expertise beyond what is reasonable to require of a government agency and may be beyond what a willing agency can do. The remedy is not to just produce a protocol for the agency to blindly execute, because nuances of the original data may matter in a way that the expert producing the protocol may be unaware.

The statistical database constructed still had privacy problems. We showed that we could make inferences about missing values. For example, a record with no standardized value for GPA and LSAT meant, in compliance with the algorithm used to construct the statistical database, that the test-taker was from a school having fewer than 20 test-takers in a specific 3- or 5-year period. We showed we could use this information to identify some of these test-takers by name.

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018. <http://techscience.org/a/2018111301>

As another example, the standardized scores themselves can be the basis of a unique identifier because the scores can be computed by a school using its own student records and then compared to the publicly shared statistical database to learn the privately held bar scores. This can be done by the school, as well as by anyone receiving a copy of school's data through a public records request, as Professor Sander had done.

No one should believe that the examples we provided are the only possible re-identifications. They are just examples that demonstrate some of the possible ways names could reliably match to records in protocol datasets. During litigation, we provided an early round of re-identification examples involving GPAs in response to earlier versions of the Sander Team protocols. The Sander Team modified the protocols in response (e.g., removing some of the digits to the right of the decimal place). The protocols we reported in this writing are the improved versions. However, the Sander Team seemed to overfit to the specific example re-identifications we first provided about GPAs and did not address the bigger issues (e.g., LSATs and failed k -anonymity enforcement) that those re-identifications represented. There is something fundamentally wrong with improvements that so narrowly match specific re-identification examples. It assumes that the burden of proof that a dataset is sufficiently anonymous as promised lies on the recipient of the anonymized data and not on the party who claimed to anonymize it.

The idea of an exemplar is that the promised protection is not as claimed. It is not the case that any given exemplar should be accepted as demonstrating the extent of the problem. Instead, evidence of such exemplars should be seen as a tip of the iceberg, being indicative of many other possible re-identifications using the demonstrated strategy and many other different strategies and data sources. Small numbers of Hispanic and Asian attorneys graduate from many schools, not just 11 in Pepperdine in a single 3-year period. Lots of online resumes and bios post GPA and LSAT scores, not just the samples in this writing. Vulnerability in these protocols is not limited to a few cases, a few matching plans, or even the few demonstrated strategies. It is beyond litmus testing to compute how many can be re-identified. With 60,572 unique records, there are a lot of possibilities.

In the Sander Team's report, they assert that there are no recorded incidents of a breach of a data enclave in which records were re-identified. Perhaps this is true, but how would the administrators know? It is certainly not in the interest of a data aggregator or other person who deliberately circumvents the security of a data enclave to publicly announce that fact.

If the State Bar of California or other State Bars across the United States made this type of data publicly available, even using the proposed protocols, it might fuel speculation on the academic performance of judges and other high-profile individuals merely because these data are made public. Judges, candidates for office, and even attorneys being hired or promoted, may be forced to reveal personal academic records that would otherwise remain private, solely to address raised concerns.

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018. <http://techscience.org/a/2018111301>

Finally, the protocols imposed a disparate impact on minorities, leaving their data more vulnerable to re-identification. We are not asserting that this type of data cannot be properly anonymized (whether the steps necessary to do so can be legally compelled is a separate question). In fact, formal protection models like k -anonymity can provide guarantees of protection, but they have to be properly implemented. As the example of this litigation illustrates, a critical aspect of proper implementation includes all fields in the k -anonymity protocol rather than simply assuming that certain fields do not need to be k -anonymized.

Acknowledgements

This paper is further explanation of problems analyzed during a legal proceeding. The views expressed in this paper are the personal views of the authors on the academic question raised here. They are not intended as the authors' views on the merits of the litigation or the result achieved therein, and do not represent the position of any of the authors' clients including the State Bar of California or its Board of Trustees.

References

1. Sander v. State Bar, No. CPF-08-50880, 2016 WL 6594874 (Cal. Super. Nov. 11, 2016)
2. U.S. Health Insurance Portability and Accountability Act of 1996. 45 CFR Parts 160 and 164. February 2003. U.S. Health Insurance Portability and Accountability Act (HIPAA; Pub.L. 104–191, 110 Stat. 1936, enacted August 21, 1996)
3. U.S. Department of Health and Human Services. Summary of the HIPAA Privacy Rule. Accessed March 10, 2017. <https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/>
4. U.S. Health Insurance Portability and Accountability Act of 1996. Safe Harbor. 45 CFR 164(b)(1).
5. Robertson J. Public Records Requests for State Discharge Data (updated with Maine). Bloomberg News. ForeverData.org. Collection 1007. November 2012. <https://foreverdata.org/1007/>
6. Sweeney L. Only You, Your Doctor, and Many Others May Know. *Technology Science*. 2015092903. September 29, 2015. <http://techscience.org/a/2015092903>
7. Robertson J. Who's Buying Your Medical Records. Bloomberg News. June 5, 2013. <http://www.businessweek.com/news/2013-06-05/states-hospital-data-for-sale-leaves-veteran-s-privacy-at-risk>

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018.
<http://techscience.org/a/2018111301>

8. Engrossed Substitute Senate Bill 6265. State of Washington. 63rd Legislature. 2014 Regular Session. <http://apps.leg.wa.gov/documents/billdocs/2013-14/Pdf/Bills/Senate%20Passed%20Legislature/6265-S.PL.pdf>
9. Comprehensive Hospital Abstract Reporting System (CHARS). Washington State Health Department. <http://www.doh.wa.gov/DataandStatisticalReports/HealthcareinWashington/HospitalandPatientData/HospitalDischargeDataCHARS>
10. Combine to one citation: Yoo J, Thaler A, Sweeney L, and Zang J. Risks to Patient Privacy: A re-identification of patients in Maine and Vermont statewide hospital data. *Technology Science*. 2018100901. October 9, 2018.
<http://techscience.org/a/2018100901>
11. California Public Records Act. CPRA: California Government Code sections §§ 6250 through 6276.48
12. U.S. Freedom of Information Act, 5 U.S.C. § 552
13. League of California Cities. The People's Business: A guide to the California Public Records Act. April 2017. <https://www.cacities.org/Resources/Open-Government/THE-PEOPLE%E2%80%99S-BUSINESS-A-Guide-to-the-California-Pu.aspx>
14. California Attorney General's Office. Summary of the California Public Records Act. August 2004. http://ag.ca.gov/publications/summary_public_records_act.pdf
15. California Special Districts Association. California Public Records Act Compliance Manual for Special Districts: a guide to understanding the California Public Records Act. 2015. <http://www.bwslaw.com/tasks/sites/bwslaw/assets/Image/2015-Public-Records-Act-Guide.pdf>
16. Haynie v. Superior Court, 26 Cal.4th 1061, 1075 (2001).
17. Laroche v. U.S. S.E.C., No. C-05-4760 (CW), 2006 WL 2868972, at *3 (N.D. Cal. Oct. 6, 2006)
18. Yeager v. Drug Enforcement Admin., 678 F.2d 315, 321 (D.C. Cir. 1982)
19. Students Against Genocide v. Department of State, 257 F.3d 828 (D.C. Cir. 2001)
20. Center for Public Integrity v. Federal Communications Commission, 505 F.Supp.2d 106, 114 (D.D.C. 2007)
21. U.S. Department of Health and Human Services. Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018.
<http://techscience.org/a/2018111301>

Insurance Portability and Accountability Act (HIPAA) Privacy Rule. November 26, 2012. Accessed March 10, 2017. <https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/#safeharborguidance>

22. El Emam K, Jonker E, Arbuckle L, and Malin B. "A Systematic Review of Re-Identification Attacks on Health Data." *PLoS ONE*, vol. 6, no. 12, Dec 2011, pp. 1-12.
23. Narayanan A and Shmatikov V. "Robust De-anonymization of Large Sparse Datasets. Proceedings of the IEEE Symposium on Security and Privacy, 2008, pp. 111-125
24. Kwok P and Lafky D. Harder Than You Think: A case study of re-identification risk of HIPAA-compliant records. https://www.researchgate.net/publication/265077763_Harder_Than_You_Think_A_Case_Study_of_Re-identification_Risk_of_HIPAA-Compliant_Records
25. Sweeney L. Simple Demographics Often Identify People Uniquely. Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh 2000.
<http://dataprivacylab.org/projects/identifiability/>
26. Sweeney L, Yoo J, Perovich L, Boronow K, Brown P, Brody J. Re-identification Risks in HIPAA Safe Harbor Data: A study of data from one environmental health study. *Technology Science*. 2017082801. August 28, 2017.
<https://techscience.org/a/2017082801>
27. Federal Committee on Statistical Methodology. Report on Statistical Disclosure Limitation Methodology. Statistical Working Paper 22. December 2005.
<http://www.hhs.gov/sites/default/files/spwp22.pdf>
28. Sweeney L. k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10 (5), 2002; 557-570.
<http://dataprivacylab.org/dataprivacy/projects/kanonymity/kanonymity.html>
29. Sweeney L. Patient Identifiability in Pharmaceutical Marketing Data. Data Privacy Lab Working Paper 1015. Cambridge 2011.
<https://dataprivacylab.org/projects/identifiability/pharma1.html>
30. Alexander, L. and Jabine, T. Access to social security microdata files for research and statistical purposes. *Social Security Bulletin*. 1978 (41) No. 8.
31. California Department of Health Services. Public Aggregate Reporting – Guidelines Development Project. Version 1.6 August 25, 2014.
<http://www.dhcs.ca.gov/dataandstats/data/DocumentsOLD/IMD/PublicReportingGuidelines.pdf>

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018. <http://techscience.org/a/2018111301>

32. Romanos A. Defamation Update (New Zealand Defamation Law). 2014. <http://www.defamationupdate.co.nz/guide-to-defamation-law>
33. Who can Sue for Defamation. Digital Media Law Project. Berkman Center for Internet and Society. Harvard University. <http://www.dmlp.org/legal-guide/who-can-sue-defamation>
34. Ross S. Deciding Communication Law: Key Cases in Context. Lawrence Erlbaum Associates. Mahwah, New Jersey. 2004. p507. <https://www.amazon.com/Deciding-Communication-Law-Context-Routledge/dp/0415647150>
35. Sweeney L. Privacert Risk Assessment Server. 2004. <http://privacert.com/>
36. Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10 (5), 2002; 571-588. <http://dataprivacylab.org/dataprivacy/projects/kanonymity/kanonymity2.pdf>
37. Excel Easy. 2016, 2017. <http://www.excel-easy.com/>
38. Microsoft Excel Training. 2016, 2017. <https://support.office.com/en-us/article/Excel-training-9bc05390-e94c-46af-a5b3-d7c22f6990bb>
39. Excel Central. The Smart Method. 2016, 2017. <http://excelcentral.com/>
40. OlgaS. How Many Words in One Page? 2016, 2017. <http://anycount.com/WordCountBlog/how-many-words-in-one-page/>
41. Graduate Student Issues Committee. The Classical Association of the Middle West and South. 2016, 2017. <https://camws.org/gsic/conference.php>
42. Winslow N. Average Typing Speed Per Minute of All Levels –Identify Yours Here! 2016, 2017. <http://typefastnow.com/average-typing-speed>
43. U.S. Department of Justice. Frequently Asked Questions. FOIA.gov. 2016, 2017. <https://www.foia.gov/faq.html>
44. Pepperdine Digital Collections. Commencement Programs, 1973, ..., 2016. <http://pepperdine.contentdm.oclc.org/cdm/search/collection/p15730coll17/searchterm/law%20school%20commencement/mode/all/order/nosort/page/2>
45. Stanford Commencement Program 2002. <http://law.stanford.edu/wp-content/uploads/sites/default/files/child-page/183686/doc/slspublic/slsgrad2002-program.pdf>

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018.
<http://techscience.org/a/2018111301>

46. University of Virginia. Commencement Final Exercises. 2003.
https://majorevents.virginia.edu/sites/majorevents.virginia.edu/files/2003_Finals_Program.pdf
47. Golden Gate. Commencement 1909,...,2016.
http://digitalcommons.law.ggu.edu/commencement/?utm_source=digitalcommons.law.ggu.edu%2Fcommencement%2F9&utm_medium=PDF&utm_campaign=PDFCoverPages
48. Pepperdine Transcript Legend. Pepperdine University. Accessed 2016.
<https://www.pepperdine.edu/registrar/content/transcriptlegend.pdf>
49. State Bar of California. Attorney Search. 2016, 2017.
<http://members.calbar.ca.gov/fal/membersearch/quicksearch>
50. Daniel Droog LinkedIn Profile. 2016, 2017. <https://www.linkedin.com/in/dddroog/>
51. Loyola Law School. Alumni US. 2005 Graduates. 2016, 2017.
http://alumnus.net/loyola_law_school_lo-7875-year-2005#alumni
52. Pepperdine University School of Law. Alumni Graduated 2002. 2016, 2017.
<https://www.linkedin.com/school/8382772/alumni/?educationEndYear=2002&filterByOption=graduated>
53. Stanford Law School. Commencement 2006 Graduation Program. 2016, 2017.
<https://law.stanford.edu/index.php?webauth-document=child-page/183682/doc/slspublic/slsgrad2006-program.pdf>
54. McGeorge School of Law Black Law Student Association. 2016. Archived at
<http://www.oocities.org/mcgeorgeblsa/>
55. State Bar of California. Attorney Profile. Daniel Droog.
<http://members.calbar.ca.gov/fal/Member/Detail/224596>
56. Philip Hache LinkedIn Profile. 2016, 2017. <https://www.linkedin.com/in/philhache/>
57. Peter Perkowski LinkedIn Profile. 2016, 2017.
<https://www.linkedin.com/in/pperkowski/>
58. Caleb Frigerio LinkedIn Profile. 2016, 2017. <https://www.linkedin.com/in/caleb-frigerio-588266aa/>
59. Marla Chabner LinkedIn Profile. 2016, 2017. <https://www.linkedin.com/in/marla-chabner-5075a947/>

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018.
<http://techscience.org/a/2018111301>

60. Edahm Small LinkedIn Profile. 2016, 2017. <https://www.linkedin.com/in/edahn-small-6503a23/>
61. Venus Johnson. University of the Pacific McGeorge School of Law. 2016.
http://www.mcgeorge.edu/Venus_Johnson.htm
62. Dustin Johnson. University of the Pacific McGeorge School of Law. 2016.
<https://www.mcgeorge.edu/profiles/alumni/dustin-johnson>
63. Marcia Randle. LinkedIn Profile. 2016. <https://www.linkedin.com/in/anthony-c-williams-98b2392>
64. Anthony C. Williams. LinkedIn Profile. 2016. <https://www.linkedin.com/in/anthony-c-williams-98b2392>
65. William Bishop. 2016. http://www.mcgeorge.edu/documents/pacific_law_sp04.pdf
66. U.S. Census Bureau. Frequently Occurring Surnames from the Census 2000. September 15, 2014 https://www.census.gov/topics/population/genealogy/data/2000_surnames.html
67. Bertrand M and Mullainathan S. Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. NBER Working Paper No. 9873. July 2003. <http://www.nber.org/papers/w9873> (As of January 9, 2013).
68. Fryer R and Levitt S. The Causes and Consequences of Distinctively Black Names. *The Quarterly Journal of Economics*. Vol 59 (3) August 2004.
<http://pricetheory.uchicago.edu/levitt/Papers/FryerLevitt2004.pdf> (As of January 9, 2013).
69. Sander R. Project SEAPHE. 2016. <http://www.seaphe.org/databases.php>
70. Newman L. Hacker Lexicon: What is Steganography. *Wired*. June 26, 2017.
<https://www.wired.com/story/steganography-hacker-lexicon/>
71. Every ABA Accredited Law School in the United States.
<http://www.lawyeredu.org/aba-accredited-schools.html#louisiana> (As of June 21017)
72. Kyoung Joo Sue Oh. Martindale Attorney Profile.
<https://www.martindale.com/encino/california/kyoung-joo-sue-oh-270000-a>
73. Rita Mankovich Irani. Alumni US Profile. http://alumnus.net/duquesne_university_-9250-3

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018.
<http://techscience.org/a/2018111301>

74. David Harris Solo. Alumni US Profile. http://alumnius.net/florida_state_univer-8118-year-1987-1991#alumni
75. Sangeeta Jain. Martindale Attorney Profile. <https://www.martindale.com/north-oaks/minnesota/sangeeta-jain-209551-a/>
76. Joseph Michael Gabriel. Attorney Profile. <http://www.ecombase.info/los-angeles/california/joseph-michael-gabriel-78980-a/>
77. Judith Michiko Sasaki. Professional Profile. <http://radaris.com/p/Judith/Sasaki/>
78. Liem Hieu Do. Martindale Attorney Profile. <https://www.martindale.com/westminster/california/liem-hieu-do-284802-a/>
79. Victoria Iusam Chau. Alumni US Profile. http://alumnius.net/university_of_hawai3-8207
80. Katherine Landey Pang. Martindale Attorney Profile. <https://www.martindale.com/san-francisco/california/katherine-l-delsack-223223-a/>
81. Steven Susumu Kondo. Martindale Attorney Profile. <https://www.martindale.com/vista/california/steven-susumu-kondo-281328-a/>
82. Diep Ngoc Nguyen. Lawyer Profile. <http://www.lawyers.com/san-jose/california/diep-ngoc-nguyen-252515-a/>
83. Joseph Gabriel. Digital Domain Profile. <http://www.digitaldomain.com/leadership/joseph-gabriel/>

Authors

Latanya Sweeney is Professor of Government and Technology in Residence at Harvard University, X.D. and Nancy Yang Faculty Dean of Currier House at Harvard, Director of the Data Privacy Lab at Harvard, Editor-in-Chief of Technology Science, and was formerly the Chief Technology Officer of the U.S. Federal Trade Commission. She earned her PhD in computer science from the Massachusetts Institute of Technology and her undergraduate degree from Harvard. More information about Dr. Sweeney is available at her website at latanyasweeney.org. As Editor-In-Chief of Technology Science, Professor Sweeney was recused from the review of this paper.

Michael von Loewenfeldt is a partner at Kerr & Wagstaffe and is certified as a specialist in Appellate Law by the State Bar of California Board of Legal Specialization. He heads the

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018. <http://techscience.org/a/2018111301>

firm's appellate practice and is also an accomplished trial court lawyer. Michael has substantial experience in privacy rights, local government, employment, and insurance law. After receiving his J.D. from the University of California at Berkeley (Boalt Hall) in 1995, Michael served as a law clerk to the Honorable Sandra B. Armstrong, United States District Court for the Northern District of California. He was also a member of the California National Guard (Mechanized Infantry) from 1987 until 1994.

Melissa Perry is a civil litigator with a wealth of experience at Kerr & Wagstaffe LLP in San Francisco. Prior to joining the firm, Melissa was a Law Fellow at Legal Aid at Work, handling gender and disability discrimination cases. She graduated from Georgetown University Law Center, where she was a student attorney for the Domestic Violence Clinic. Melissa served as the Senior Notes Editor for the Georgetown Journal on Poverty Law and Policy. While at law school, she externed at the Equal Employment Opportunity Commission's Office of Federal Operations, Lambda Legal, Lawyers' Committee for Civil Rights, and the Disability Rights Education and Defense Fund. Prior to attending Georgetown, Melissa was a Teach For America corps member in Arkansas, teaching high school English Language Arts.

Referring Editor: [TBD]

Citation

Sweeney L, Von Loewenfeldt M, Perry M. Saying it's Anonymous Doesn't Make It So: Re-identifications of "anonymized" law school data. *Technology Science*. 2018111301. November 13, 2018. <http://techscience.org/a/2018111301>

Data

[TBD]