De-Identification

Reduce Privacy Risks When Sharing Personally Identifiable Information

De-Identification

Unlock the value in your data

Privacy Analytics Inc. is commercializing the technology developed by the Electronic Health Information Laboratory (EHIL) at the Children's Hospital of Eastern Ontario Research Institute (CHEO RI). Lead by its principal researcher, Dr. Khaled El Emam, EHIL has become a leader within the research community in the area of de-identification. The lab has a number of peer-reviewed publications in the area of de-identification and hosts an annual conference on the subject. This whitepaper details the problem of de-identification and provides a high-level explanation of the technology developed at EHIL. This technology has been integrated into the Privacy Analytics software and is now available to organizations that manage personally identifiable data.

The Risks of Disclosing Personal Data

Today we live in a world where our personal information is being continuously captured in a multitude of electronic databases. Details about our health, financial status and buying habits are stored in databases managed by public and private sector organizations. These databases contain information about millions of people, and can provide valuable research, epidemiologic and business insight. For example, examining a drug store chain's prescriptions can indicate where a flu outbreak is occurring. To extract or maximize the value contained in these databases, data custodians must often provide outside organizations access to their data. In order to protect the privacy of the individuals whose data is being disclosed, a data custodian will "de-identify" information before releasing it to a third-party. De-identification ensures that data cannot be traced to the person about whom it pertains. What might seem like a simple matter of masking a person's identifiers (name, address), the problem of de-identification has proven more difficult and is an active area of scientific research.

The problem of de-identification affects a variety of industries including:

- **Registries.** Health care organizations (e.g., hospitals, clinics) currently submit patient data to registries. Data contained in these registries can be used for research and policy/administrative needs (such as a stroke or cancer registry). Often data is sent to a registry without patient consent under the assumption that it is de-identified.

- **Testing/Quality Assurance.** When developing or maintaining large information systems, there is the need to provide developers/QA teams with test data. Often, personal data is taken from a production system and must be de-identified before being provided to the testing team.

- **Prescription data.** Data brokers currently collect prescription data and sell the analysis derived from it to pharmaceutical companies. Personal information must be de-identified before being sent to a data broker.

- **Insurance claims.** Like pharmaceutical companies, insurance companies analyze claims data for actuarial and marketing reasons. De-identification is required to comply with privacy best practices, and in some jurisdictions, requirements.

- **National statistical agencies.** A census agency is the most commonly known provider of de-identified information. Census results are de-identified and provided/sold to third parties for further analysis.

Organizations, such as the ones listed above, are motivated to protect the privacy of personal information for several reasons, including:

- **Legislation.** Most governments have enacted legislation requiring organizations to adopt measures to protect personal data. For example, in the United States, health information is protected by the Health Insurance Portability and Accountability Act (HIPAA) and financial information by the Sarbanes-Oxley Act (SOX). Similar legislation exists in the European Union and Canada.

- **Litigation.** Should a person's private information be released by an organization without the person's consent, they have the right to file a complaint with a regulatory authority or take the organization to court. This can lead to a costly investigation or to litigation, even if no damages are awarded.

- **Cost.** If an organization inadvertently releases private information, privacy legislation often mandates that the people whose data was exposed must be notified. In addition to the cost of breach notification, an organization might face significant litigation and compensation costs, and increasingly, fines by regulators.

- **Reputation.** A privacy breach is a public relations disaster for an organization (public or private), and can directly affect the bottom line. Furthermore, breaches erode the public/client/patient trust in that organization.

To avoid a privacy breach, organizations currently use manual, ad-hoc methods to de-identify personal information. Given the lack of de-identification tools, there have been several high-profile incidents where improper de-identification has resulted in a privacy breach. Recent examples include:

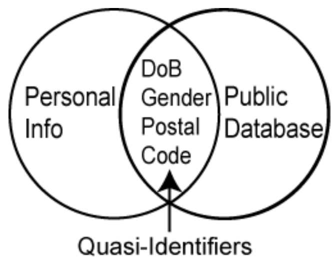
- Data from the Group Insurance Commission, which purchases health insurance for state employees in Massachusetts, was matched against the voter list for Cambridge, re-identifying the governor's record.
- Students were able to re-identify a significant percentage of individuals in the Chicago homicide database by linking with the social security death index.
- Individuals in an anonymized publicly available database of customer movie recommendations from Netflix were re-identified by linking their ratings with ratings in a publicly available Internet movie rating web site.
- A national broadcaster aired a report on the death of a 26 year-old female taking a particular drug who was re-identified from the adverse drug reaction database released by Health Canada.
- AOL put anonymized Internet search data (including health-related searches) on its web site. New York Times reporters were able to re-identify an individual from her search records within a few days.

These re-identifications were possible because the de-identification was not done properly and did not ensure that the risk were sufficiently low before the data was disclosed. Proper de-identification would have made those breaches highly unlikely.

EHIL's mandate is to develop technology to help organizations prevent privacy breaches such as these.

Quasi-Identifiers: The Devil is in the Details

When de-identifying records, many people assume that removing names and addresses (direct identifiers) is sufficient to protect the privacy of the persons whose data is being released. The problem of de-identification involves those personal details that are not obviously identifying. These personal details, known as *quasi-identifiers*, include the person's age, sex, postal code, profession, ethnic origin and income (to name a few).



EHIL has focused its research on the de-identification of quasi-identifiers. The research at EHIL has highlighted three unique types of re-identification attacks: prosecutor, journalist, and marketer. Algorithms to measure the risk of each type of attack have been developed and published in peer-reviewed journals (see the Publications section for details). Privacy Analytics Inc. has integrated these algorithms into an easy to use tool to allow organizations to measure re-identification risk.

Prosecutor risk

In this scenario, an intruder wants to re-identify a specific person in a de-identified database. Let's take the example of an employer who has obtained a de-identified database of drug test results. The employer is trying to find the test results of one of their employees (Dave, a 37 year-old doctor) and knows that Dave's record is in the de-identfied dataset.

The re-identification risk is measured by finding the unique combinations of quasi-identifiers in the anonymized dataset (these are called *equivalence classes*). To illustrate what is an equivalence class, let's take the following dataset containing the quasi-identifiers of sex, age and profession. The dataset also contains the person's latest drug test results (this is the sensitive data).

ID	Sex	Age	Profession	Drug test
1	Male	37	Doctor	Negative
2	Female	28	Doctor	Positive
3	Male	37	Doctor	Negative
4	Male	28	Doctor	Positive
5	Male	28	Doctor	Negative
6	Male	37	Doctor	Negative

In this dataset there are three equivalence classes: 28 year-old male doctors, 37-year-old male doctors and 28-year old female doctors. Since the employer knows that Dave is a 37 year-old doctor, there is a 1 in 3 chance (33%) of identifying Dave's record correctly. If however, the employer were attempting to identify a 28-year old female doctor, there would be a perfect match since only one record in that equivalence class exists. Since we cannot predict which equivalence class an intruder will attempt to match, we must assume the worst-case scenario, which is that the person they want to identify has the smallest equivalence class (denoted by k) in the database (i.e., 28-year-old female doctor). When de-identifying a dataset, a value of 5 for k (i.e., there are at least five records in any equivalence class) is often considered sufficient privacy protection.

Journalist risk

Journalist risk is also concerned with the re-identification of individuals. However, in this case the journalist does not care which individual is re-identified. The probabilistic risk profile here is quite different from that of prosecutor risk. In the journalist scenario, the anonymized data is a subset of a larger public database. The journalist doesn't know a particular individual in the anonymized dataset but does know that all the people in the dataset exist in a larger public database (which they have access to). A real-life example of a journalist attack occurred when a Canadian Broadcasting Corporation (CBC) reporter re-identified a patient in a de-identified adverse drug reaction database by matching her age, date of death, gender, and location with the public obituaries.

Previous research has shown that the smallest equivalence class found in the public database that maps to the anonymized dataset measures the risk of re-identification. To illustrate this, let's look at the following tables.

Original Database to Disclose

	IDENTIFYING VARIABLE	QUASI-IDENTIFIERS		
ID	Name	Gender	Year of Birth	Test Result
1	John Smith	Male	1959	+ve
2	Alan Smith	Male	1962	-ve
3	Alice Brown	Female	1955	-ve
4	Hercules Green	Male	1959	-ve
5	Alicia Freds	Female	1942	-ve
6	Gill Stringer	Female	1975	-ve
7	Marie Kirkpatrick	Female	1966	+ve
8	Leslie Hall	Female	1987	-ve
9	Bill Nash	Male	1975	-ve
10	Albert Blackwell	Male	1978	-ve
11	Beverly McCulsky	Female	1964	-ve
12	Douglas Henry	Male	1959	+ve
13	Freda Shields	Female	1975	-ve
14	Fred Thompson	Male	1967	-ve

2-Anonymization

	QUASI-IDENTIFIERS		
ID	Gender	Decade of Birth	Test Result
1	Male	1950-1959	+ve
2	Male	1960-1969	-ve
4	Male	1950-1959	-ve
6	Female	1970-1979	-ve
7	Female	1960-1969	+ve
9	Male	1970-1979	-ve
10	Male	1970-1979	-ve
11	Female	1960-1969	-ve
12	Male	1950-1959	+ve
13	Female	1970-1979	-ve
14	Male	1960-1969	-ve

Disclosed (k-Anonymized) Database (ς)

Matching

Identification Database (Z)

	IDENTIFYING VARIABLE	QUASI-IDENTIFIERS	
ID	Name	Gender	Year of Birth
1	John Smith	Male	1959
2	Alan Smith	Male	1962
3	Alice Brown	Female	1955
4	Hercules Green	Male	1959
5	Alicia Freds	Female	1942
6	Gill Stringer	Female	1975
7	Marie Kirkpatrick	Female	1966
8	Leslie Hall	Female	1987
9	Bill Nash	Male	1975
10	Albert Blackwell	Male	1978
11	Beverly McCulsky	Female	1964
12	Douglas Henry	Male	1959
13	Freda Shields	Female	1975
14	Fred Thompson	Male	1967
15	Joe Doe	Male	1961
16	Mark Fractus	Male	1974
17	Lillian Barley	Female	1978
18	Jane Doe	Female	1961
19	Nina Brown	Female	1968
20	William Cooper	Male	1973
21	Kathy Last	Female	1966
22	Deitmar Plank	Male	1967
23	Anderson Hoyt	Male	1971
24	Alexandra Knight	Female	1974
25	Helene Arnold	Female	1977
26	Anderson Heft	Male	1968
27	Almond Zipf	Male	1954
28	Alex Long	Female	1952
29	Britney Goldman	Female	1956
30	Lisa Marie	Female	1988
31	Natasha Markhov	Female	1941

The first table is the original dataset before anonymization. The records in the table are a subset of those found in the public database (Z). The dataset is anonymized (ζ) by removing names and aggregating the year of birth by decade (decade of birth). There are five equivalence classes in the anonymized table that map to the public database.

Equivalence class		Anonymized table		Public database	
Gender	Age	Count	Id	Count	ID
Male	1950-1959	3	1,4,12	4	1,4,12,27
Male	1960-1969	2	2,14	5	2,14,15,22,26
Male	1970-1979	2	9,10	5	9,10,16,20,23
Female	1960-1969	2	7,11	5	7,11,18,19,21
Female	1970-1979	2	6,13	5	6,13,17,24,25

This table shows that the smallest equivalence class in the public database (Z) that map to the anonymized dataset (ζ) is a male born in the 1950s (four records). Therefore, there is a one in four chance (25%) of re-identifying a record that falls in this equivalence class.

The problem with applying the existing journalist re-identification risk analysis is that the entire content of the public database (Z) is rarely known (e.g., due to cost, logistics). To overcome this limitation, the researchers at EHIL have developed a unique model to estimate the number of records found in each equivalence class in a public database. This allows the re-identification risk in the journalist scenario to be calculated and controlled without having access to the larger public database.

Marketer risk

In this scenario, an intruder wants to re-identify as many individuals as possible in a database. The marketer is less concerned if some of the records are misidentified. Therefore, rather than focus on individuals, here the risk pertains to everyone in the data set. Take for example a pharmaceutical company that obtained de-identified prescription data. They can attempt to match this data with their internal marketing database to create a mailing campaign (say, targeting doctors). They are not concerned if some of the mailers are sent to the wrong physicians (i.e., spam).

The marketer risk is measured by calculating the probability of matching a record in an equivalence class of the de-identified set with those in the matching equivalence class in the marketer's database. In the journalist example (see above), the first equivalence class (males ages 1950-1959) has three records that could be matched to one of four possible records in the public database. The expected number of records that a marketer can properly identify when randomly matching records in the de-identified dataset with those in the public database can be calculated for each equivalence class.

Equivalence class		Anonymized table		Public database		Expected # Correct Matches
Gender	**Age**	**Count**	**Record number**	**Count**	**Record number**	
Male	1950-1959	3	1,4,12	4	1,4,12,27	3/4
Male	1960-1969	2	2,14	5	2,14,15,22,26	2/5
Male	1970-1979	2	9,10	5	9,10,16,20,23	2/5
Female	1960-1969	2	7,11	5	7,11,18,19,21	2/5
Female	1970-1979	2	6,13	5	6,13,17,24,25	2/5
Expected number of identified records						2.35

A marketer would expect to correctly re-identify about 21% of the overall records in this scenario.

Data De-Identification Techniques

Once a dataset's risk of re-identification has been measured, it must be properly anonymized. De-identification techniques include:

- **Record suppression:** When a record's combination of quasi-identifiers present too high of a risk to be released, it must be dropped from the dataset.
- **Cell suppression:** A record can be further de-identified by suppressing/masking the value contained in a field (cell). For example, a field in a patient record containing a very rare disease would be suppressed.
- **Rounding:** For numerical data (dates, integers), the values can be rounded to further de-identify a record. For example, an age of 92 is rounded to 90.
- **Aggregation/Generalization:** Rare quasi-identifiers can be aggregated to provide better anonymization. For example, a low-population postal code can be aggregated to a larger geographic area (such as a city). A rare medical profession, such as perinatologist, can be aggregated to a more general obstetrician.

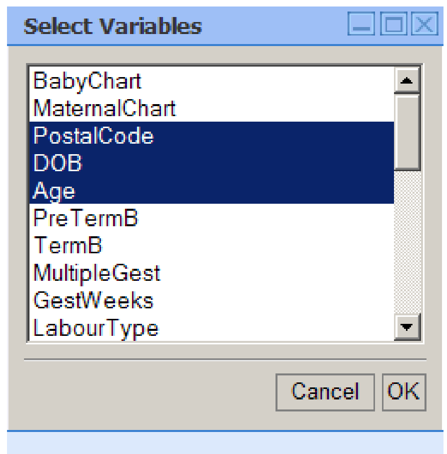
These de-identification techniques have been integrated into the Privacy Analytics application.

Privacy Analytics Risk Assessment Tool

The Privacy Analytics Risk Assessment Tool (PARAT) takes the guesswork out of de-identifying personal information. PARAT uses peer-reviewed techniques to measure and manage re-identification risk. Only PARAT can protect against all known types of re-identification attacks. It optimally de-identifies information to protect individual privacy while retaining the data's value. Using a simple four-step process, PARAT allows you to easily and safely release your valuable data.

Step 1: Variable Selection

To begin the process, the quasi-identifiers that are to be released must be selected from the dataset.



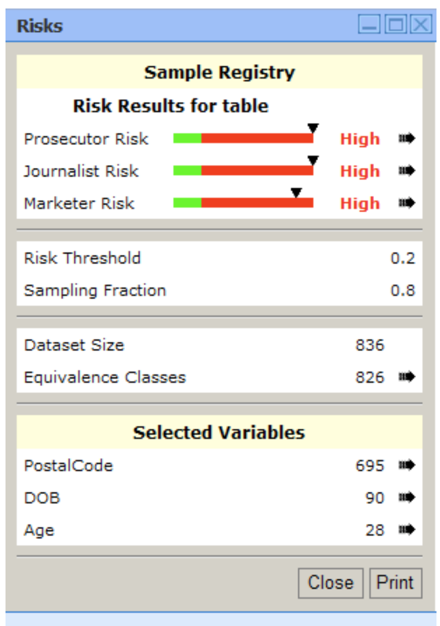
Once the quasi-identifiers are selected, you can rank them in order of importance (the variables' utility to the person using the de-identified data set). This ranking will be used during the de-identification process to determine the optimal anonymization that balances re-identification risk and data utility. For example, if age is ranked as the most important quasi-identifier and postal code as the least important, the de-identification process will attempt to keep age information intact while the postal code variable will be aggregated (i.e., grouped into larger geographic areas). Ranking allows you to maximize the utility of the de-identified dataset.

Step 2: Assign Acceptable Re-Identification Risk Threshold (Safety Index)

PARAT allows you to decide how much de-identification should be done before releasing a data set. The "amount" of de-identification is measured by the probability of accurately re-identifying a record (for prosecutor and journalist risk) or the expected number of records to be re-identified correctly (for marketer risk). For example, if the quasi-identifiers contained in a de-identified record can be associated with five individuals contained a public registry, the probability of re-identification is 0.2 (i.e., 1 in 5 chance of making the correct match). Achieving a lower probability of re-identification (lower risk) often means reducing the resolution of the released data (either suppressing records or aggregating variables). Ensuring a low re-identification risk might make the de-identified data less useful to the recipient because there is not enough data resolution for their needs. To balance the need for privacy with the need for data resolution, PARAT allows you to set the acceptable probability/risk of re-identification. Re-identification risk can be adjusted based on the profile of the person/organization requesting the information. For example, if data is to be released to the general public, a high degree of de-identification is required (e.g., a threshold of 0.05). However, if data is being shared within an organization (e.g., between government departments), a lesser amount of de-identification is needed (e.g., a threshold of 0.2). To help determine what is the right amount of de-identification, we provide a methodology to rate the risk of releasing data to a given person or organization. Risk based de-identification ensures that individual privacy is protected while maintaining the released data's value.

Step 3: Risk Analysis

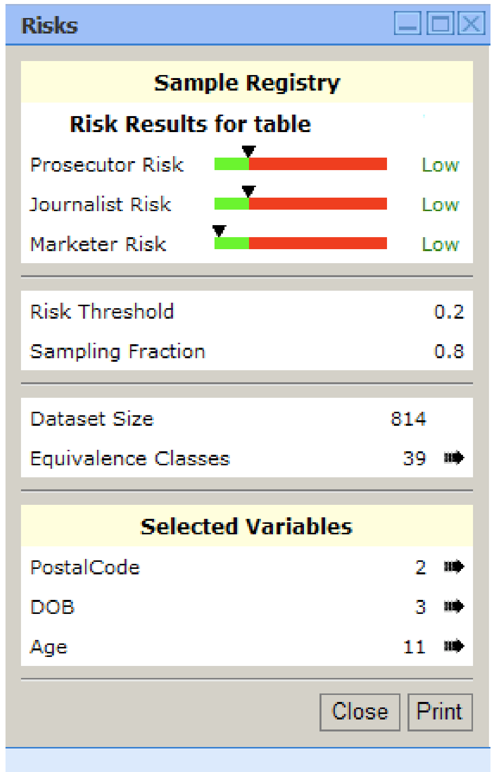
Once the acceptable threshold has been set, the risk analysis can be performed. PARAT calculates the dataset's risk for the three types of re-identification attacks: prosecutor, journalist and marketer.



In this example, a dataset containing the quasi-identifiers of postal code, date of birth and age has been analyzed with a re-identification risk threshold of 0.2. The results show the re-identification risk is high (above 0.2) for all three types of attacks: prosecutor, journalist, marketer. Of the 836 records in the dataset, 826 have a unique combination of quasi-identifiers (equivalence classes). The dataset contains 695 unique postal codes, 90 unique birth dates and 28 unique ages.

Step 4: De-Identification

To reduce the risk of re-identification below our acceptable threshold (0.2 in this example), PARAT will optimally de-identify the data. PARAT uses several techniques including suppression (removing high risk records) and aggregation (reducing the resolution of a given field).



After the de-identification process, the risk for all types of re-identification attacks has been reduced to acceptable levels. This was done by marking 22 records for suppression and aggregating quasi-identifier values. Postal code values are grouped into two areas, dates of birth are aggregated into three ranges and age into 11 ranges.

Age before de-identification

Count	Age
3	43
5	16
6	18
6	17
7	42
8	19
9	20
9	41
14	40
14	21
20	23
24	37
25	39
28	38
30	22
33	24
33	26
34	25
41	27
45	31
48	28

Age after de-identification

Count	Age
19	41-45
34	16-20
131	21-25
140	36-40
247	26-30
264	31-35

PARAT automatically produces the optimally anonymized dataset that meets the desired re-identification risk threshold.

Publications

F. K. Dankar and K. El Emam: “A method for evaluating marketer re-identification risk”. In *Proceedings of the 3rd International Workshop on Privacy and Anonymity in the Information Society*, 2010.

K. El Emam, F. Dankar, R. Issa, E. Jonker, D. Amyot, E. Cogo, JP. Corriveau, M. Walker, S. Chowdhury, R. Vaillancourt, T. Roffey, J. Bottomley: “A Globally Optimal k-Anonymity Method for the De-identification of Health Data .” In *Journal of the American Medical Informatics Association*, 16(5):670-682, 2009.

K. El Emam, A. Brown, and P. AbdelMalik: “Evaluating predictors of geographic area population size cutoffs to manage re-identification risk.” In *Journal of the American Medical Informatics Association*, March/April, 16(2):256-266, 2009.

K. El Emam, F. Dankar, R. Vaillancourt, T. Roffey, and M. Lysyk: “Evaluating Patient Re-identification Risk from Hospital Prescription Records.” In the *Canadian Journal of Hospital Pharmacy*, 62(4):307-319, 2009.

K. El Emam: “Heuristics for de-identifying health data.” In *IEEE Security and Privacy*, July/August, 6(4):58-61, 2008.

K. El Emam, and F. Dankar: “Protecting privacy using k-anonymity.” In the *Journal of the American Medical Informatics Association*, September/October, 15:627-637, 2008.

K. El Emam, E. Neri, and E. Jonker: “An evaluation of personal health information remnants in second hand personal computer disk drives.” In *Journal of Medical Internet Research*, 9(3):e24, 2007.

K. El Emam, S. Jabbouri, S. Sams, Y. Drouet, M. Power: “Evaluating common de-identification heuristics for personal health information.” In *Journal of Medical Internet Research*, 2006;8(4):e28, November 2006.

K. El Emam: “Overview of Factors Affecting the Risk of Re-Identification in Canada”, Access to Information and Privacy, Health Canada, May 2006.

K. El Emam: “Data Anonymization Practices in Clinical Research: A Descriptive Study”, Access to Information and Privacy, Health Canada, May 2006.

Contact Information

For more information contact us at

Privacy Analytics Inc.
151 Bay Street, Suite 1111
Ottawa, Ontario K1R 7T2
Canada
Tel: +1 613.797-5412
Fax: +1 613 369 4312
Email: info@privacyanalytics.ca
www.privacyanalytics.ca