CS-523 Advanced topics on
Privacy Enhancing Technologies

Privacy-preserving Data Publishing I
Live exercises

Carmela Troncoso
SPRING Lab
carmela.troncoso@epfl.ch

Key	Gender	Zipcode	Age	Disease
Eric	M	1007	25	Cancer
Justine	F	1012	25	Heart Disease
Emma	F	1012	25	Flu
Helen	F	1012	*	Flu
Paul	M	1007	25	Cancer
Philip	M	1012	35	Herpes
Michel	M	1012	35	Cancer
Mory	M	1007	25	Cancer
Adrien	M	1007	25	Heart Disease
Mallory	M	1012	35	Flu
Camille	F	1012	25	Herpes
Samuel	M	1012	35	Cancer
Marco	M	1007	*	Cancer
Damien	M	1012	35	Flu

Consider only the *Gender, Zipcode, Age* attributes.

Which statement is TRUE?

(A) The database achieves k-anonymity with k = 4.

(B) The database does not achieve k-anonymity for any k.

(C) The database achieves k-anonymity with k = 1.

(D) The database achieves k-anonymity with k = 2.

Key	Gender	Zipcode	Age	Disease
Eric	M	1007	25	Cancer
Justine	F	1012	25	Heart Disease
Emma	F	1012	25	Flu
Helen	F	1012	*	Flu
Paul	M	1007	25	Cancer
Philip	M	1012	35	Herpes
Michel	M	1012	35	Cancer
Mory	M	1007	25	Cancer
Adrien	M	1007	25	Heart Disease
Mallory	M	1012	35	Flu
Camille	F	1012	25	Herpes
Samuel	M	1012	35	Cancer
Marco	M	1007	*	Cancer
Damien	M	1012	35	Flu

Consider *Gender, Zipcode, Age* *as quasi-identifying* attributes.

Which statement is TRUE?

(A) The database achieves k-anonymity with k = 4.

(B) The database does not achieve k-anonymity for any k.

(C) The database achieves k-anonymity with k = 1.

(D) The database achieves k-anonymity with k = 2.

The database achieves anonymity with k=4

Can Marco's or Helen age affect the result? k-anonymity is a property of a **published** dataset, thus to find the k parameter it does not matter what Marco's actual age was prior to sanitization.

Key	Gender	Zipcode	Age	Disease
Eric	M	1007	25	Cancer
Justine	F	1012	25	Heart Disease
Emma	F	1012	25	Flu
Helen	F	1012	*	Flu
Paul	M	1007	25	Cancer
Philip	M	1012	35	Herpes
Michel	M	1012	35	Cancer
Mory	M	1007	25	Cancer
Adrien	M	1007	25	Heart Disease
Mallory	M	1012	35	Flu
Camille	F	1012	25	Herpes
Samuel	M	1012	35	Cancer
Marco	M	1007	*	Cancer
Damien	M	1012	35	Flu

Consider the *Disease* attribute to be sensitive.
Which statement is TRUE?

(A) The database achieves 3-diversity.

(B) The database is differentially private.

(C) The database achieves 5-diversity.

(D) None of the above

Differential privacy is studied in the next lecture

Key	Gender	Zipcode	Age	Disease
Eric	M	1007	25	Cancer
Justine	F	1012	25	Heart Disease
Emma	F	1012	25	Flu
Helen	F	1012	*	Flu
Paul	M	1007	25	Cancer
Philip	M	1012	35	Herpes
Michel	M	1012	35	Cancer
Mory	M	1007	25	Cancer
Adrien	M	1007'	25	Heart Disease
Mallory	M	1012	35	Flu
Camille	F	1012	25	Herpes
Samuel	M	1012	35	Cancer
Marco	M	1007	*	Cancer
Damien	M	1012	35	Flu

Consider the *disease* attribute to be sensitive.

Which statement is TRUE?

(A) The database achieves 3-diversity.

(B) The database is differentially private.

(C) The database achieves 5-diversity.

(D) None of the above

If one takes the quasi-identifiers in the k-anonymity question, then the database is 2-diverse (the group Eric, Adrien, Mory and Marco only have 2 diseases: Cancer and Heart Disease), and the answer is *None of the above*.

Key	Gender	Zipcode	Age	Disease
Eric	M	1007	25	Cancer
Justine	F	1012	25	Heart Disease
Emma	F	1012	25	Flu
Helen	F	1012	*	Flu
Paul	M	1007	25	Cancer
Philip	M	1012	35	Herpes
Michel	M	1012	35	Cancer
Mory	M	1007	25	Cancer
Adrien	M	1007	25	Heart Disease
Mallory	M	1012	35	Flu
Camille	F	1012	25	Herpes
Samuel	M	1012	35	Cancer
Marco	M	1007	*	Cancer
Damien	M	1012	35	Flu

Consider *Age as quasi-identifying and* *Disease* as the sensitive attribute.

Which statement is TRUE?

(A) The database achieves 3-diversity.

(B) The database is differentially private.

(C) The database achieves 5-diversity.

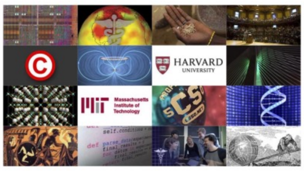
(D) None of the above

If one takes as quasi-identifier the age, then the dataset is 3-diverse as both groups (age 25 and age 35) have 3 diseases.

HarvardX and MITx: The First Year of Open Online Courses

Fall 2012-Summer 2013

Inside

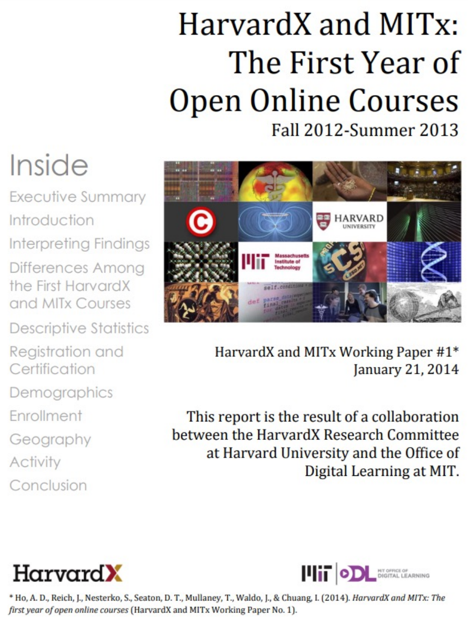


HarvardX and MITx Working Paper #1*
January 21, 2014

This report is the result of a collaboration between the HarvardX Research Committee at Harvard University and the Office of Digital Learning at MIT.

* Ho, A. D., Reich, J., Nesterko, S., Seaton, D. T., Mullaney, T., Waldo, J., & Chuang, I. (2014). *HarvardX and MITx: The first year of open online courses* (HarvardX and MITx Working Paper No. 1).

- 597,692 individuals registered for 17 online courses offered by Harvard and MIT through the EdX platform
- Data collected: students' demographics, engagement with course content, and final course grade
- "To meet these privacy specifications, the HarvardX and MITx research team (guided by the general counsel, for the two institutions) opted for a k-anonymization framework" [3]. A value of k = 5 "was chosen to allow legal sharing of the data" in accordance with FERPA. Ultimately, EdX published the 5-anonymized dataset with 476,532 students' records"



HarvardX and MITx:
The First Year of
Open Online Courses
Fall 2012-Summer 2013

Inside
Executive Summary
Introduction
Interpreting Findings
Differences Among the First HarvardX and MITx Courses
Descriptive Statistics
Registration and Certification
Demographics
Enrollment
Geography
Activity
Conclusion

HarvardX and MITx Working Paper #1*
January 21, 2014

This report is the result of a collaboration between the HarvardX Research Committee at Harvard University and the Office of Digital Learning at MIT.

HarvardX

* Ho, A. D., Reich, J., Nesterko, S., Seaton, D. T., Mullaney, T., Waldo, J., & Chuang, I. (2014). *HarvardX and MITx: The first year of open online courses* (HarvardX and MITx Working Paper No. 1).

- Published data **Xed**:
  - Basic demographics: self-reported level of education, gender, and year of birth, country (inferred from the student's IP address)
  - Activities and Results in 16 out of 17 courses
    - Results: enrolled, grade, certification status
    - Activities: e.g., number of posts in course
- K-anonymized (Generalization and supression) with respect to:
  - Q∗ = {enrolled in course 1, . . . , enrolled in course 16}
  - Qi = {gender, year of birth, country, enrolled in course i, number of forum posts in course i}

- If you were a student of one of these courses: what would be the privacy concerns? What adversaries would you worry about?

What courses were taken, and what was the outcome, when were courses taken

Adversaries:
- prospective employer (knows Q1 for all courses with certification)
- Classmate: knows activity on the shared courses, can de-anonymize with respect to those and then recover other courses' data
- Acquaintances with some information (discuss your experience with someone even without being classmates)

- If you were a student of one of these courses: what would be the privacy concerns? What adversaries would you worry about?

- If you were a student of one of these courses would you say it is safe?

No. k-anonymity with respect to one pseudo-identifier does not guarantee k-anonymity with respect to *the union of the quasiidentifiers*.

(7.1% of students (33,925 students) in Xed are unique with respect to the union of all Quasi-identifiers, and 15.3% have effective anonymity less than 5)

- If you were a student of one of these courses: what would be the privacy concerns? What adversaries would you worry about?
- If you were a student of one of these courses would you say it is safe?
- Does the order of k-anonymization matter?

Yes, k-anonymity is NOT resistant to post-processing.

In the case of Xed, they first k-anonymized with respect to Q*, and then with respect to Q1…Q16. The post-processing suppressed / generalized records that were needed for k-anonymity with respect to Q*

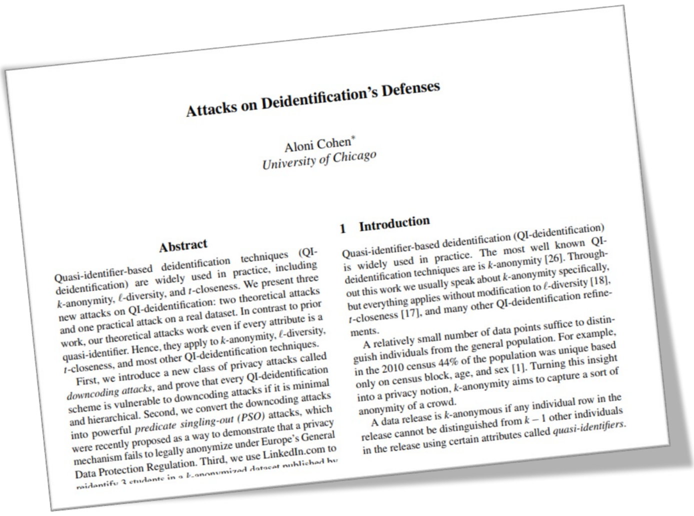
As a result, 245 students were unique and 753 had effective anonymity less than 5

- If you were a student of one of these courses: what would be the privacy concerns? What adversaries would you worry about?
- If you were a student of one of these courses would you say it is safe?
- Does the order of k-anonymization matter?
- If you found a unique record… how would you re-identify?

Use LinkedIn! Most of the information is there (with some noise) – especially if you have private paid access, such as recruiters. Once some information is found, Google can help complement

“ We reidentified 3 of the attempted 135 EdX students, each of whom registered for but failed to complete an EdX course.” (Cohen 2022, see next slide)

More on the Xed fiasco and other attacks



Attacks on Deidentification's Defenses

Aloni Cohen*
University of Chicago

Abstract

Quasi-identifier-based deidentification techniques (QI-deidentification) are widely used in practice, including k-anonymity, ℓ-diversity, and t-closeness. We present three new attacks on QI-deidentification: two theoretical attacks and one practical attack on a real dataset. In contrast to prior work, our theoretical attacks work even if every attribute is a quasi-identifier. Hence, they apply to k-anonymity, ℓ-diversity, t-closeness, and most other QI-deidentification techniques.

First, we introduce a new class of privacy attacks called *downcoding attacks*, and prove that every QI-deidentification scheme is vulnerable to downcoding attacks if it is minimal and hierarchical. Second, we convert the downcoding attacks into powerful *predicate singling-out (PSO)* attacks, which were recently proposed as a way to demonstrate that a privacy mechanism fails to legally anonymize under Europe's General Data Protection Regulation. Third, we use LinkedIn.com to

1 Introduction

Quasi-identifier-based deidentification (QI-deidentification) is widely used in practice. The most well known QI-deidentification techniques are is k-anonymity [26]. Throughout this work we usually speak about k-anonymity specifically, but everything applies without modification to ℓ-diversity [18], t-closeness [17], and many other QI-deidentification refinements.

A relatively small number of data points suffice to distinguish individuals from the general population. For example, in the 2010 census 44% of the population was unique based only on census block, age, and sex [1]. Turning this insight into a privacy notion, k-anonymity aims to capture a sort of anonymity of a crowd.

A data release is k-anonymous if any individual row in the release cannot be distinguished from $k-1$ other individuals in the release using certain attributes called *quasi-identifiers*.