Identification: A Teaching Moment for Privacy and Databases

Matthew P. Dube
matthew.dube@maine.edu
University of Maine at Augusta
Augusta, ME, USA

Rocko Graziano
rocko.graziano@maine.edu
University of Maine at Augusta
Augusta, ME, USA

Course Databases/Data Mining
Programming Language None
Knowledge Unit Data Structures
CS Topics Tabular Data, Operations / Algorithms (Sets)
Resource Type Lecture Slides

SYNOPSIS

This learning experience helps students gain experience and proficiency with issues regarding the ethical collection and use of data. Students will gain an appreciation for the risks associated with record-level identification, where data attributes, however innocently collected, can and have been used to violate privacy and lead to discrimination against individuals and protected classes of individuals.

KEYWORDS

keys, databases, identification, data mining, privacy, proxy discrimination

ACM Reference Format:

Matthew P. Dube and Rocko Graziano. November 2023. Identification: A Teaching Moment for Privacy and Databases. In ACM *EngageCSEdu*. ACM, New York, NY, USA, 4 pages. https://doi.org/ 10.1145/3631985

1 ENGAGEMENT HIGHLIGHTS

This learning experience is designed to cross over several different types of computing courses to demonstrate how personal privacy can be compromised within a database. This experience works through numerous best practices in teaching computer science:

Interdisciplinarity: The examples discussed in the experience (either slides or addendum resources) reference two vastly different arenas: voter protection and spatial cognition. These issues provide relevance for students with interests in political science, geography, and language. This is critical because most examples for introductory database courses rely on business examples or educational data.

Discussions and groups: The exercise presented in the learning experience asks the students to play a game with a partner and use the insights from the game to analyze a dataset from a particular lens. They will reconvene and likely share that their approach worked in different ways, highlighting that the scope of the sample dictates identification potential.

Relevant content: The example in the learning experience is a relevant current event that fits in the context of political discussions of voter rights. The particular case intersects with voting laws in New York State. The embedding example could easily be applied by the students to something in their daily life where they were the only one that did something.

Additionally, the exercise focuses on a few seminal principles in critical computing:

Privacy: If a record can be identified by a non-personally identifiable piece of data, and that record furthermore includes personally identifiable information, that personally identifiable information can then be used to ascertain other information from various sources [1].

Marginalization: Sometimes an attribute in a dataset might not identify, but it does provide access to another table or information source that can provide discriminatory data. A prime example of that is the concept of red-lining, where communities can be used as a proxy for race or income [2].

2 RECOMMENDATIONS

Ultimately, there are a couple of important suggestions to provide to any instructor, independent of the context. These involve some of the issues of the tools used in the experience and also the context of the experience.

The Crazy Games implementation of Guess Who? is different from the normal game in one key area: the game presents an immutable set of questions that a player can ask so that the system can evaluate it appropriately. It is a helpful exercise to allow students to ask the appropriate questions. The instructor can print out an image of the people in the game so that students have access to it (or use it as a standalone slide). Students should be free to ask their partners whatever questions they want, not just those presented by the online game.

Do not show the students the table about the people in the game until they have played the game. The table can greatly simplify the task of identification by directing the types of questions you would ask about the individuals.

The activity presented relies upon student knowledge of places where personally identifiable information is required to be provided. Three obvious forms of that would be the social security number, the driver's license, or a passport [1, 3]. If a student has never had to give this information, be ready to comment about places, services, or applications that require these types of information. Some examples include:

ACM EngageCSEdu, November 2023.

ACM ISBN 978-8-4007-0470-3/23/11.
https://doi.org/10.1145/3631985

- Job application
- Financial aid application
- Insurance
- Educational institutions
- Healthcare institutions
- Airlines
- Financial institutions

Students will likely be quick to point out that the information in these tables is now connectable to each other and to governmental records due to the sharing of the identifier. Students may not identify that any part of the information now has the potential to lead back to the identifier (and thus to all of the information in all of the tables). Point out that if something were sufficiently identifiable even if it were innocuous, this would be enough to compromise all of the information stored about that person across all of these items. This is precisely how the voting records example in New York City was compromised: the voting registry had the name of the voter; the actual voting returns had the result of the vote; these were connected over a third item: the identity of the voting district [4].

A student may not be able to make the leap that the identification of an individual record in a database constitutes a violation of privacy. A loss of privacy is suffered when an action that should not have been discoverable and attributable to a person is subsequently discovered and thusly attributed. If a non-key attribute specifically ties a result to a name, that results in a loss of privacy. This is the precise reason that the US Census Bureau aggregates personal returns into census geographies: it protects the identity of individual responses [5].

Finally, the section on Proxy Discrimination [6] extends the lesson to consider the real-world impacts of decision support systems based on machine learning algorithms trained by data which can lead to discrimination against protected classes. An example from [16] related to Criminal Risk Assessments has been provided. This section may (and probably should) be updated with relevant, recent examples of such situations, leading to a discussion of the types of discrimination students may have unknowingly experienced.

3 POTENTIAL APPLICATION SPACES

Identification as a principle exists in many computer science courses, but in each case, the principle is found in a different way. Those differences suggest different recommendations for what context and circumstances to engage in this type of activity (or something similar).

3.1 Introductory data structures

In an introductory data structures course, students learn about the concept of a dictionary. A dictionary allows us to key off of any item that could be deemed an identifier. This approach to identification specifically motivates how we talk about primary keys in databases, and simultaneously neglects the risk of doing that very thing. Similarly, it neglects the concept of single record filtration beyond the key value. This lesson fits into that area where a student can distinguish between filtration of values versus filtration via a key.

3.2 Introductory programming

In an introductory programming course, an instructor might work with students creating functions or methods that can be used modularly. An example of identification through filtration can be used to frame this particular concept as a discussion of functional purpose. The same principle of the Guess Who? game also applies to developing functions that a) identify a particular phenomenon versus its logical opposite, and b) identify the exact category that a phenomenon belongs to. These two methods of identification can function in substantively different ways because the meaning of identification is different in both contexts. In a), the context is to determine necessary and sufficient conditions for that category versus in b), the context is to determine necessary and sufficient conditions to determine any category.

An example of that concept can be seen in spatial reasoning work meant to derive topological relations from measurements within a shapefile [7] or a raster image [8]. This slant of the activity exemplifies that the task of sufficient determination is less resource intensive, but is not a guarantee of global categorization. An example of this socially can be in a police lineup [9]. If the suspect is relatively distinct from the other participants in the lineup, very superficial traits may trip up the eye-witness. This particular practice is fundamentally related to differences in arrest rates in minority communities [10]. In the *Guess Who?* game, two characters have unique hair colors, embodying this principle.

3.3 Introductory data science

An introductory data science course should show applications of concepts to many disparate phenomena. This module can be used to demonstrate how identification can force or exacerbate social issues that were unintended, while also highlighting the prior linkage to geospatial or temporal language [7, 8].

3.4 Database design

In a database design course, primary keys are discussed at length. Primary keys represent a very narrow form of identification: the computer's perspective of a dictionary [3], motivated by functions from algebra. Often, a student will present a potential primary key off of knowledge of one object. That misconception is the origin of this activity because it demonstrates that global identification is much more complicated than that as a deductive process, but it is much more simplistic than the alternative. Nevertheless, this experience asks students to think through how a misconception can be a window into a different type of real problem that can get quite messy, and is frankly less controlled for, representing a design blind spot. This represents the ability to use a discrepant teaching event to our advantage [11].

3.5 Machine learning/data mining

A core learning outcome in a machine learning or data mining course is the premise of conditional probability as it relates to association rules. In that environment, we are not actually studying identification at all – we are studying combinations of logical information that seem to travel together at high rates, known as support (how frequent in the dataset) and confidence (how frequent given that the antecedent has occurred). If an observation has the lowest possible support (1 out of n), it will have the highest confidence (1 out of 1), and the combination of those two measures constitutes functional identity. This is the precise instance exhibited in the Dante DeBlasio case: the support for a rule concerning a vote from the particular voting district in New York City was exactly one vote, and thus the exact vote could be paired to another table that could extend the information [4]. Proxy discrimination [6] and the risks of perpetuating societal bias is itself an ethical issue which should be addressed in any introductory machine learning class.

3.6 SQL

An SQL course studies how to query data, and a core piece of that course is to discuss joins. Conceptually, joins naturally follow from foreign keys, but that is of course not necessary. Students will experience this concept as providing great power in a big data environment, but this experience will confront the challenge that is provided by having numerous data sources that share information as mutual fields in that type of environment, representing data versatility [12]. Students can then learn about views as a structural way to control more sensitive information, but that of course will not fully alleviate the problem [13].

3.7 Social aspects of computing

Identification and its corresponding privacy issues should be a core topic in any course that details the social ramifications of computing. As more and more personal data is collected, more and more people are put at risk [14]. This experience demonstrates how a real case that questions privacy was influenced by this exact principle and demonstrates to the students how something like that could happen.

3.8 Responsible conduct of research

This experience is designed to be used in a host of scenarios spanning different areas of computer science. It can also be used as an educational experience for students in a responsible conduct of research course who might have to consider the vulnerabilities that data can have. People sometimes cannot envision how data that is not at first personally identifiable can have a compromising nature. Great care needs to be considered in this environment to prevent haphazard identification, thus eliminating the expectation of privacy [15].

4 ACKNOWLEDGEMENTS

Our thanks to various students from our program whose anecdotal feedback either verbally or through course evaluations led to a rethinking of how to encapsulate this core knowledge of databases in a more understandable format. The authors also gratefully acknowledge Veronica Hotton, Ellie Harmon, Katherine Weatherford-Darling, Betina Tagle, and Elizabeth Reddy for prior collaboration and insights on integrating ethics in computing at a more granular level. Matthew P. Dube was supported by NSF Grant Award No. 2019470.

5 MATERIALS

Three elements are provided for this learning experience:

- Identity in Databases.pptx, a powerpoint slide deck which frames the lecture. This includes the warm-up exercise (the *Guess Who?* game), context for the main topics of the presentation. The notes for each slide serve as both the lesson plan and suggested voice-over for class discussion.
- Identity in Databases (instructor guide).pdf, a narrative with both detailed instructions and background material to prepare the instructor to lead the class discussion.
- CompanionData.xlsx, a collection of data used for the *Guess Who?* game and in the discussion on breast cancer research [17]

5 AUXILIARY MATERIALS

(1) https://www.crazygames.com/game/guess-who-multiplayer, the Guessing Game Activity used in the classroom exercise.

(2) Meaning of the Social Security Number*; Social Security Bulletin, November, 1982/Vol. 45, No.11; [online] https://www.ssa.gov/policy/docs/ssb/v45n11/v45n11p29.pdf , background on the format of SS numbers

(3) https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/?utm_source=pocket_saves, an article regarding predictive policing algorithms.

(4) https://towardsdatascience.com/how-discrimination-occurs-in-data-analytics-and-machine-learning-proxy-variables-7c22ff20792, background on proxy variables and how they may lead to discrimination.

REFERENCES

[1] Darrow, J. & Lichtenstein, S. (2008). *Do You Really Need My Social Security Number - Data Collection Practices in the Digital Age*, 10 N.C. J.L. & Tech. 1. Available at: https://scholarship.law.unc.edu/ncjolt/vol10/iss1/2

[2] Hunt, D. B,2005. Redlining, Encyclopedia of Chicago, http://www.encyclopedia.chicagohistory.org/pages/1050.html

[3] Bopp, C., Benjamin, L. M., & Voida, A. (2019). The coerciveness of the primary key: Infrastructure problems in human services work. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), 1-26.

[4] Clark, J., Cormack, L., & Wang, S. (2021). Privacy Concerns in New York City Elections. Technical Report, Princeton University.

[5] Young, C., Martin, D., & Skinner, C. (2009). Geographically intelligent disclosure control for flexible aggregation of census data. *International Journal of Geographical Information Science*, 23(4), 457-482.

[6] Tschantz, M. C., (2022). What is Proxy Discrimination? In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1993–2003. https://doi.org/10.1145/3531146.3533242

[7] Egenhofer, M., and Dube, M. 2009. Topological Relations from Metric Refinements. *Proceedings of the 17th ACM SIGSPATIAL 2009*, 158-167.

[8] Dube, M., Barrett, J., and Egenhofer, M. 2015. From Metric to Topology: Determining Relations in Discrete Space. *Proceedings of the 2015 International Conference on Spatial Information Theory*, 151-171.

[9] Colloff, M. F., Wade, K. A., & Strange, D. (2016). Unfair lineups make witnesses more likely to confuse innocent and guilty suspects. *Psychological Science*, 27(9), 1227-1239.

[10] Steblay, N. K., & Wells, G. L. (2020). Assessment of bias in police lineups. *Psychology, Public Policy, and Law*, *26*(4), 393–412. https://doi.org/10.1037/law0000287

[11] Longfield, J (2009). Discrepant Teaching Events: Using an Inquiry Stance to Address Students' Misconceptions. *International Journal of Teaching and Learning in Higher Education*, v21 n2 p266-271.

[12] Kapil, G., Agrawal, A., & Khan, R. A. (2016, October). A study of big data characteristics. In *2016 International Conference on Communication and Electronics Systems* (ICCES) (pp. 1-4). IEEE.

[13] Denning, D. E., Akl, S. G., Heckman, M., Lunt, T. F., Morgenstern, M., Neumann, P. G., & Schell, R. R. (1987). Views for multilevel database security. *IEEE Transactions on Software Engineering*, (2), 129-140.

[14] De Groot, J. (2022). The History of Data Breaches [Online]. Available: https://www.digitalguardian.com/blog/history-data-breaches.

[15] Throne, R. (2022). Adverse Trends in Data Ethics: The AI Bill of Rights and Human Subjects Protections. Available at SSRN: https://ssrn.com/abstract=4279922 or http://dx.doi.org/10.2139/ssrn.4279922.

[16] Angwin, Larson, Mattu, Kirchner (2016). Machine Bias, *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[17] Mangasarian, O., Street W., and Wolberg, W. 1995. Breast cancer diagnosis and prognosis via linear programming. Operations research 43.4, 570-577.