Final Project Write-up: Hawks, Rats, and De-Anonymization

CS 105: Privacy and Technology
Esther An, Samantha Marks, Michael Moorman, Andrew Palacci

December 10, 2023

1. Introduction

1.1 Background

Online data privacy is a topic that is frequently at the forefront of the news cycle and our minds, as we are constantly reminded of the dangers of data leaks and privacy violations. Because of this, many individuals take significant steps to protect their data, such as using fake names, email addresses, and other personal information, which would seem like a reasonable approach to protecting one's personal details online.

However, even if an individual is supposedly anonymous online, it is still the case that quasi-identifiers (attributes that do not inherently identify someone uniquely, but can be used to narrow the search for them) can be used to re-identify them. We present a case study of this phenomenon, in which we demonstrate how data from a seemingly harmless survey can be used to re-identify individuals in the Harvard student body.

1.2 Project Motivation

At Harvard, it is extremely common to receive an email from a fellow undergraduate asking for your participation in a survey for a psychology class numbered 1-1000 (take your pick). These are nearly always marketed as "anonymous" surveys, and are sent to multiple house mailings lists, reaching hundreds of students. However, due to the fact that these surveys often ask for various quasi-identifiers of respondents, such as gender, ethnicity, place of residence, etc., we believed it may be possible to re-identify individuals who respond to these surveys and wanted to investigate this hypothesis.

So, we set out to create a survey that would be sent to various house mailing lists, and then attempt to re-identify respondents by linking responses with data from the Harvard College Facebook, a HarvardKey protected database of Harvard College students' concentrations, houses, and other personally identifying information. We also 5-anonymized the Harvard College Facebook, using house, concentration, and year as quasi-identifiers, and created a website in which a Harvard College student could enter their concentration, house, and social class and see how many other individuals at Harvard match that criteria.

2. Methods

2.1 Survey

The survey[1] was marketed as an "anonymous psych survey" upon release to different Harvard house mailing lists. The survey sought responses from Harvard students with a house affiliation. Freshmen were not included among survey respondents.

The following questions were included in the survey:

- Concentration
- Additional Concentration (for those with joint or double concentrations)
- Social Class (based on year of matriculation)
- Academic Year
- House
- Two of these defend you, the others are hunting you. Which two will you choose to defend you?
  - 50 Hawks
  - 10 Crocodiles
  - 3 Brown Bears
  - 15 Wolves
  - 1 Hunter with a rifle
  - 7 Cape Buffalo
  - 10,000 Rats
  - 5 Gorillas
  - 4 Lions
  - 20 Geese
- Why? (as a follow-up to the "psychology" question above)

2.2 Re-identification

Our re-identification of respondents to our survey was done with an algorithm, rather than manually. The Harvard College Facebook allows for

[1] https://forms.gle/NeYTMwkHBq8zjZKcA

downloads of its data in CSV format, with columns such as `Name`, `Email`, `Concentration` (including double and joint concentrations), `Social Class`, `House`, `Hometown`, and more. This proved especially useful, as we were able to isolate the relevant columns used for re-identification, namely `Concentration`, `Social Class`, and `House`.

Given this data, and the data collected from our survey in CSV format, we were able to write a Python script using only the standard library that identifies respondents by:

1. For each response in the survey dataset,
   (a) Iterating through the Harvard Facebook data, filtering out rows that do not match at least one of the `Concentration`, `Social Class`, and `House` attribute, and
   (b) Counting the remaining matches, returning the value in the `Name` column in the Harvard Facebook data in the case of a unique match.

These simple steps alone allowed for efficient and effective re-identification of respondents to our survey. Special care did have to be taken to ensure double concentrations were correctly accounted for, as the Facebook imposes an arbitrary ordering on concentrations in the case of a double concentration. Besides edge cases such as this, the algorithm is relatively straightforward.

The time complexity of this algorithm is not fantastic, as it takes precisely $\Theta(pq)$ time, where p is the number of rows in the Facebook, and q is the number of survey responses. However, because these are relatively small values, on the order of thousands and hundreds, respectively, this algorithm still runs in a fraction of a second in practice.

2.3 k-anonymization

k-anonymization is a model for protecting the privacy of people whose information is included in a dataset. k-anonymity protection is afforded if "the information for each person contained in [a] release cannot be distinguished from at least $k-1$ individuals whose information also appears in the release" [1]. Note that all datasets are minimally 1-anonymous. To assess the level of k-anonymity in a dataset, quasi-identifiers can be selected to distinguish between individuals. Quasi-identifiers are a set of attributes that, in combination, can be linked with external information to re-identify, or reduce uncertainty, about individuals [2].

We 5-anonymized the Harvard Facebook dataset (see Data Availability section for access). This dataset is a listing of all Harvard College students who agreed to have their information shared on the Harvard Facebook and whatever information they consented to being included in the database. When downloaded, it contains each student's name, email, house (dormitory), year (academic), and concentration (major). The quasi-identifiers we selected were house, year, and concentration. The motivation for this is explained in the Discussion (Section 4.3). We discarded the name and email categories, as these are generally not asked for or collected in "anonymous" surveys to the student body.

To successfully 5-anonymize the Harvard Facebook dataset, we utilized pandas, a Python package for working with datasets, to implement generalization. This technique involves creating bins for values belonging to a column and replacing the values in the column with the larger bin that they belong to, effectively reorganizing all entries by reorganizing them into wider groups. We wrote the code implementing this in `k-anon-final.ipynb` (see the Code Availability section for access). We generalized the dataset to achieve 5-anonymity using three steps:

1. Binning Concentrations
2. Binning Houses
3. Removing Special Concentrations and the Second Concentration from Double and Joint Concentrations

We now describe how we carried out these three binning steps.

2.3.1 Binning Concentrations

The following concentration bins, which are organized by Harvard University on the official Concentrations web page, were defined.[2] The bins organized by Harvard University are designed such that some concentrations are assigned to more than one bin. The striked-out concentrations are those that were represented redundantly and were thus removed from a given bin. The bins used in the "Binning Concentrations" step of achieving 5-anonymization are those defined by the following categories and all concentrations below them that are not striked out. We had to make the decisions regarding how to organize the concentrations represented redundantly such that each concentration is represented exactly once, in one of the ten bins.

Concentration Bins:

- Arts
  - Art, Film, and Visual Studies

[2] Concentrations. (n.d.). Harvard College. https://college.harvard.edu/academics/liberal-arts-sciences/concentrations.

  - ~~*English*~~
  - Music
  - Theater, Dance, & Media
- Engineering
  - Biomedical Engineering
  - Electrical Engineering
  - Engineering Sciences
  - Environmental Science and Engineering
  - Mechanical Engineering
- History
  - Anthropology
  - Classics
  - ~~*Comparative Study of Religion*~~
  - East Asian Studies
  - History and Literature
  - History and Science
  - History of Art and Architecture
  - ~~*Near Eastern Languages and Civilizations*~~
  - ~~*Philosophy*~~
  - South Asian Studies
- Languages, Literatures, and Religion
  - ~~*Classics*~~
  - Comparative Literature
  - Comparative Study of Religion
  - ~~*East Asian Studies*~~
  - English
  - Folklore and Mythology
  - Germanic Languages and Literature
  - ~~*History and Literature*~~
  - Linguistics
  - Near Eastern Languages and Civilizations
  - ~~*Philosophy*~~
  - Romance Languages and Literature
  - Slavic Literatures and Cultures
  - ~~*South Asian Studies*~~
- Life Sciences
  - ~~*Biomedical Engineering*~~
  - Chemical and Physical Biology
  - Human Developmental and Regenerative Biology
  - Human Evolutionary Biology
  - Integrative Biology
  - Molecular and Cellular Biology
  - Neuroscience
  - Psychology
- Math and Computation
  - Applied Math
  - Computer Science
  - Mathematics
  - Statistics
- Physical Sciences
  - Astrophysics
  - Chemistry
  - Chemistry and Physics
  - Earth and Planetary Sciences
  - ~~*Environmental Science and Public Policy*~~
  - ~~*Mathematics*~~
  - Physics
- Qualitative Social Sciences
  - African and African American Studies
  - ~~*Anthropology*~~
  - ~~*Comparative Study of Religion*~~
  - Government
  - ~~*History and Literature*~~
  - ~~*Linguistics*~~
  - Philosophy
  - Social Studies
  - ~~*Sociology*~~
  - Studies of Women, Gender, and Sexuality
- Quantitative Social Sciences
  - ~~*Applied Math*~~
  - Economics
  - Environmental Science and Public Policy
  - ~~*Government*~~
  - ~~*Psychology*~~
  - Sociology
  - ~~*Statistics*~~
- Special Concentration
  - Special Concentration

To implement this binning, the concentrations associated with each Harvard University-organized category were initially found by filtering the concentrations by each bin name on the Harvard University Concentrations web page. The Harvard Facebook dataset was then queried using "Find All" for every instance of each concentration, and the name of each concentration was replaced with their associated bin as defined above.

2.3.2 Binning Houses

The following house bins were created to reorganize all 12 houses and the Dudley Co-op into larger categories:

- **River East**: Dunster, Leverett, and Mather
- **River Central**: Adams, Lowell, and Quincy
- **River West**: Eliot, Kirkland, and Winthrop
- **Quad**: Cabot, Currier, Pforzheimer, and Dudley

To implement this binning, each house value was replaced by its bin name using the pandas Python library.

2.3.3 Removing Special Concentrations and the Second Concentration from Double and Joint Concentrations (Truncating Concentrations)

Entries (rows) with the value "Special Concentration" in the "Concentration" column were dropped; hence this procedure required some (minimal) row suppression.

Additionally, values of second concentrations, which would be included for students as the second area of study of a double concentration or the allied field in a joint concentration, were dropped. The row of data was not completely removed: note that the rest of the information about the individual with a double or concentration was retained. In order to implement this step, we cleaned and reformatted the Harvard Facebook dataset, which puts all concentrations in the same column, to parse out second concentrations into a separate column. We then simply dropped this column from use (hence requiring column suppression in the reformatted version of the dataset).

Since multiple concentrations in the "Concentration" column were separated by commas, which are also used in the names of some concentrations themselves, the implementation of this parsing required several intermediate steps. First, all commas in the database were temporarily replaced with the '>' symbol using "Find All and Replace" in Excel. We then searched for all concentration names containing a comma, of which there are only a few (e.g., "Art, Film, and Visual Studies"), and restored them to the original concentration names with commas. We then used "Find All and Replace" again to replace all remaining '>' symbols with a ';'. Using pandas, we split the "Concentration" column on the ';' symbol to split it into two columns: one for the first concentration listed under "Concentration" and one for the second concentration of the double or joint.

3. Results

3.1 Survey

258 Harvard students responded to the "anonymous psych survey".

The survey respondents were skewed towards Dunster House residents, with 27.1% of respondents from Dunster (70 respondents). The next most-represented houses were Pforzheimer (10.5% of respondents) and Kirkland (10.5% of respondents).

There was a fairly even breakdown among students of different academic years, with 32.9% being sophomores (85 respondents), 39.9% being juniors (103 respondents), and 27.1% being seniors (70 respondents).

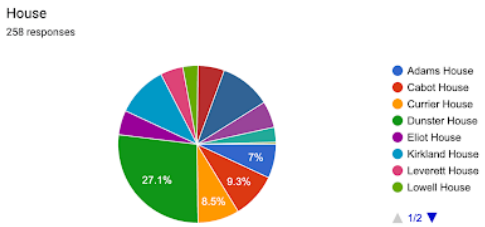

Figure 1: *Respondents by Harvard house.*

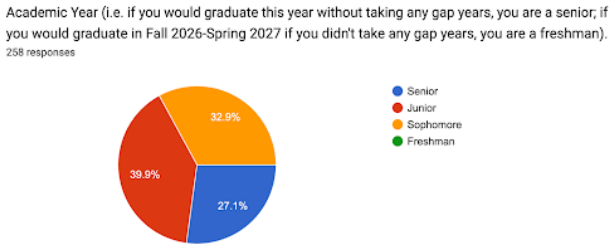

Figure 2: *Respondents by academic year.*

There was a slight skew towards Computer Science among respondents in terms of their main (first) concentration, second concentration, and secondary, with 17.8% of all respondents reporting computer science as their primary field of study (46 respondents), 12.1% of respondents with a double or joint concentration reporting computer science as their second field of study (89 respondents), and 16% of respondents with a secondary reporting computer science as their secondary (20 respondents). This may reflect the popularity of the CS concentration among the student population. For reference, among the 5,775 students included in the Harvard Facebook, 12.3% of students have a Computer Science affiliation in their concentration (713 students concentrate in only Computer Science or Computer Science and another field of study).

The second most prevalent primary concentration among respondents was Government (8.5%, 22 respondents). The second most prevalent additional concentration was Statistics (10.6%, 7 respondents). The second and third most prevalent secondaries were Economics (12%, 15 respondents) and Global Health and Health Policy (11.2%, 14 respondents), respectively.

3.2 Re-identification

Among the 258 students we surveyed, we were able to re-identify 62 of them algorithmically (this does not include one respondent that proudly shared his full name and email in his survey response, in

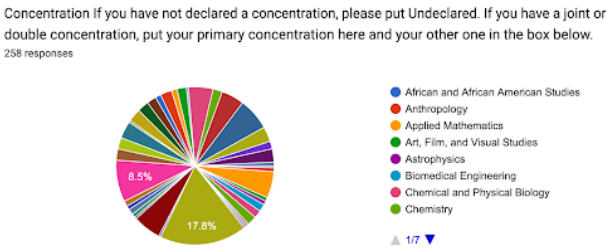

Figure 3: *Respondents by concentration.*

hopes of being contacted about the survey's findings). This represents 24% of our responses.

We were also able to classify an additional 25 respondents, or 9.7%, as 2-anonymous. This means that of all of the students at Harvard, we could attribute the response to one of two people (without even considering the content of the survey response itself). Furthermore, this could pose a risk when *both* of the people who have a shared 2-anonymous identity submit survey responses—if any or all of their responses overlap, we (the malicious re-identifiers) could gain knowledge about both respondents.

One example of this (in the case of our less-than-malicious survey) would be if two people that share the same 2-anonymous quasi-identifiers responded with "Hunter, Rats" and "Rats, Hawks"—even without full re-identification, we would still be able to attribute the choice of rats to both respondents. This clearly extends to 3-anonymity, 4-anonymity, etc., and could also pose a far greater risk for respondents in a survey with more sensitive questions.

3.3 k-anonymization

We successfully 5-anonymized the Harvard Facebook dataset. We did so while upholding the goal of minimizing the amount of information lost through row or column suppression.

To do so, we used the generalization approach defined in the Methods section. The motivation for this is explained in Discussion (Section 4.3).

We will show why the steps we took to make the Harvard Facebook 5-anonymous through generalization minimized the amount of information lost by showing that we maximized the amount of information retained. The initial Harvard Facebook dataset before we implemented our adjustments was 1-anonymous.

3.3.1 Binning Concentrations

To identify the reason the Harvard Facebook dataset in its original form was 1-anonymous, we found the minimum number of occurrences of any unique value within each column, doing so for each column independently. The most unique value in the "House" and "Year" columns, or the one with the fewest number of duplicates, had 123 duplicates and 49 duplicates, respectively. The most unique value in the "Concentration" column with the fewest number of duplicates had 1 value—that is, just the "Concentration" column alone led to 1-anonymity. Hence we had to generalize this column by binning concentrations, as described in Methods. We aimed to put concentrations in bins that were as descriptive and logical as possible, which was why we based them off of the categories of concentrations that Harvard lists as options under the "Fields of Study" descriptions on the official Harvard Concentrations web page.

However, after binning concentrations as described in Methods (Section 2.3.1), the Harvard Facebook dataset was still 1-anonymous. Hence we needed to implement additional generalization to achieve 5-anonymity.

3.3.2 Binning Houses

Looking through the number of entries within each combination of quasi-identifiers, we noticed two major patterns:

1. Many people with joint, double, or special concentrations fell into groupings with fewer than 5 people, and most frequently belonged to groupings of size 1 (meaning they were unique and could readily be re-identified).
2. Even some respondents without a joint or double concentration fell into a group with fewer than 5 people in it, as demonstrated in these two cases:
    - ("Adams", "Senior", "Arts"): 1 person
    - ("Dunster", "Sophomore", "Arts"): 1 person

These observations pointed us in two potential directions: (1) remove the second concentration from joint and double concentrations so that students are only associated with one concentration, or (2) try another form of generalization. We decided that less information would be lost if we first binned the Houses into logical categories than if we removed second concentrations, tried generalizing further by placing concentrations into bigger bins (which would then necessarily be less descriptive of the concentrations that fall within them or provide less intuitive groupings), or generalized across year (as there are only 4 years and students perceive the experience and their maturity level in each as distinct, versus the 13 Houses that students tend to place into groupings). We thus binned house as listed in Methods (Section 2.3.2), using groupings conventionally understood by Harvard students, with the exception that

we included Dudley in the Quad since it is physically closer to the Quad than any of the River Houses.

However, even after binning the houses in addition to the concentrations, the Harvard Facebook dataset was still 1-anonymous. Hence the dataset required additional generalization.

3.3.3 Truncating Concentrations

Looking through the number of entries that fall into each combination of quasi-identifiers, we noticed that pattern 1, enumerated in the Results section above (Section 3.3.2), persisted.

Because nearly everyone with a joint, double, or special concentration is unique (meaning that there exists only one entry with their grouping of quasi-identifiers), we decided that we had to remove their second concentration, and eliminate entries with special concentrations. After doing so, the Harvard Facebook dataset was 5-anonymous.

We then went back and verified that the binning of Harvard houses was actually necessary to achieve 5-anonymity by implementing only the binning of concentrations and removal of second and special concentrations, skipping the binning of Houses. After doing so, the Harvard Facebook dataset returned to being 1-anonymous. Thus, we confirmed that we still needed to bin by house to achieve 5-anonymity.

Therefore, we showed that we can 5-anonymize the Harvard Facebook dataset through the generalization mechanisms described above, and do so in a way that minimizes the amount of potentially valuable information lost.

4. Discussion

4.1 Survey

The survey was marketed as an "anonymous psych survey" upon release to different Harvard house mailing lists to fully simulate the Harvard student experience of completing anonymous surveys for other students.

In designing the survey, we wanted to request the minimum required information to re-identify individuals by linking the survey response dataset with the Harvard Facebook dataset. Hence the only questions with "identifying" information were those asking for the surveyed student's house, concentration (including their main concentration and second concentration if the student is completing a joint or double), secondary, and social/academic year.

4.2 Re-identification

A few interesting conclusions arose from our re-identification process. In general, we learned that, while it may be hard to achieve k-anonymity for a given k for a dataset, it is still hard to perfectly re-identify people, especially given quasi-identifiers that leave little room for uniqueness, given the fact that there are only 13 houses (including Dudley), and 5-6 valid social classes.

4.2.1 Reflections on Re-identification

We think our success rate of 24% may strike the reader as rather low. However, we believe that our success rate may be lower than the actual success rate of the average survey. This is because we had many more students who responded with a Computer Science concentration than any other concentration, and our percentage of respondents concentrating in Computer Science was disproportionately larger than the percentage of students concentration Computer Science. This was mentioned earlier in Results 3.1, where we also showed that Computer Science is a large concentration. This means that individuals with a Computer Science concentration are less unique, and hence less re-identifiable. So having a disproportionately large number of Computer Science concentrators responding likely has made our success rate smaller.

On top of this, while our experimental success rate still gives most individuals a solid chance of not being re-identified, our standard for our privacy and data protection should not lie in anywhere in the realm of "solid chance". With only three rather broad data points, we were able to link a quarter of respondents with only one other data set. Imagine the catastrophic results if some malicious actor (or purportedly benevolent one) were to attempt something like this with more datasets to link or more quasi-identifiers to match. We find it very likely that given these extra resources, a large proportion of our data can be uniquely linked to each of us, especially within an organization like Harvard.

Additionally, consider that this re-identification was done entirely algorithmically. We believe that many of us hold the belief that we are safe from re-identification and profiling online due to the fact that we don't have anyone who would want to spend the time focused on us alone to re-identify us. Firstly, this assumption may be wrong entirely, and secondly, it isn't as though these actors must do any manual labor or put in the effort to single you out: they can re-identify (and thus exploit!) you algorithmically, with little effort or time.

The key take-away is that we are not as anonymous as we seem, and that while we knew that re-identification could be done, it turns out to be much easier than one might think.

4.2.2 Feedback from Re-identified Individuals

Re-identified individuals were informed via an email explaining that they had been re-identified and, as proof of re-identification, with their response to the "psychology" question described in the survey section above. They were asked to respond to a follow-up survey which asked the following questions:

1. How concerned were you about being re-identified when responding to anonymous surveys prior to receiving this email? (1 = Not concerned; 5 = Extremely concerned)
2. How concerned do you feel about being re-identified from "anonymous" surveys now? (1 = Not concerned; 5 = Extremely concerned)
3. How likely are you to respond to "anonymous" surveys in the future? (1 = Very unlikely; 5 = Very likely)
4. Any additional comments?

Approximately 44% of re-identified students (22 students) responded to the follow-up survey. In their responses, some students expressed feeling betrayal at having been re-identified, while others wondered whether the re-identification was implemented using AI of some sort. Others shared that they were less concerned given that the survey had not asked for any sensitive information, while some shared that they typically responded to anonymous surveys with the awareness that the survey may not be fully and truly anonymous.

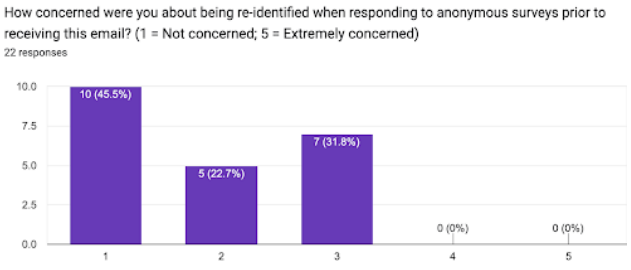

Figure 4: *Concern about re-identification prior to our project notification.*

0% of respondents expressed feeling great or extreme concern about being re-identified when responding to anonymous surveys prior to receiving our "You were identified." email, while 31.8% of respondents expressed feeling great or extreme concern

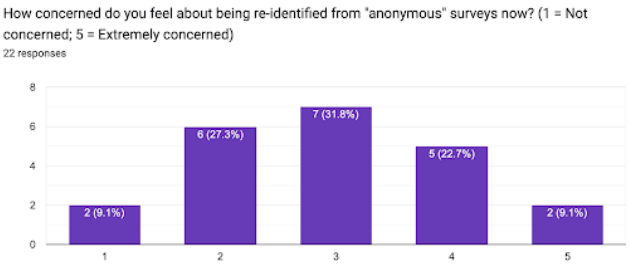

Figure 5: *Concern about re-identification after our project notification.*

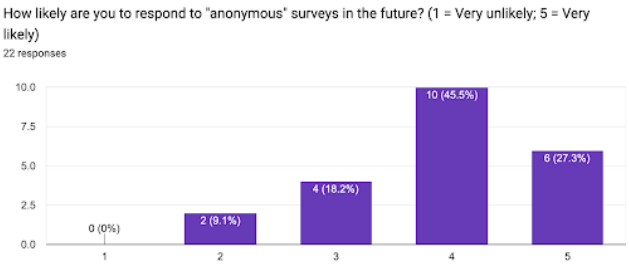

Figure 6: *Likelihood of responding to future "anonymous surveys.*

about being re-identified after receiving our follow-up email. Hence our follow-up and proof of concept seems to have sparked some concern among the students who were re-identified.

However, when asked whether they would still respond to "anonymous" surveys in the future, 72.8% of respondents answered that they would be likely or very likely to respond.

4.3 k-anonymization

We 5-anonymized the Harvard Facebook dataset, using house, concentration, and year as quasi-identifiers. We did so with the goal of 5-anonymizing the responses to the typical "anonymous" Harvard survey, which often asks for a student's house, concentration, and year, and can be linked to the Harvard Facebook dataset.

By 5-anonymizing the Harvard Facebook dataset with our chosen quasi-identifiers, we ensure that all surveyors which ask for these three things cannot reasonably re-identify a respondent based on those attributes alone. This is because 5-anonymizing the Harvard Facebook dataset with our chosen identifiers is equivalent to 5-anonymizing a survey in which all Harvard College students respond, ensuring that any survey with any number of respondents which have any set of these quasi-identifiers are indistinguishable from at least 4 other people when linked to the

Harvard Facebook. We note that the Harvard Facebook was the only source we found which readily allows one to make this linkage without impractically time-consuming manual searching, so we view this restricted access to data as a generally strong protection of student anonymity.

Additionally, based on our motivation, we had to choose a technique of 5-anonymization which translated to practical steps a surveyor could take in making a 5-anonymous survey. We chose not to use entry (row) suppression, as this would correspond to individuals not being able to take a survey in order to remain anonymous, which is not practical. We also chose not to use general column suppression, as having responders' House, Concentration, and Year may be valuable information for the surveyor in order to analyze their population sampled. Thus the generalization approach we described in Methods (Section 2.3) was the most practical method for translating how to anonymize the Harvard Facebook dataset to how to make an anonymous survey and preserved the most amount of information.

To make this translation: instead of asking for respondents' exact concentrations, surveyors can provide the bins we listed in Methods (Section 2.3.1) as the options. They would detail to the respondent which concentrations belong to which bin, and instruct those with joint, double, or special concentrations to select the bin which suits them best. This ensures that nobody is left out from being able to take the survey, while still capturing the essence of the information that could be learned from their concentration. Additionally, instead of exact concentrations, surveyors would provide the bins we listed in Methods (Section 2.3.2) as the options. They would detail to the respondent which Houses belong to which bin. As students tend to group Houses into the bins we made, this still captures the essence of the information that could be learned from a respondent's House. Therefore, the survey made through this method still contains valuable information for the surveyor to use, while providing strong protection of the respondent's anonymity.

Moreover, we suggest that surveyors request students pursuing a Special Concentration to select the concentration "bin" that describes their concentration best.

5. Addendum: Website

While our survey is certainly a good demonstration of the lack of anonymity the Harvard College population truly possesses, only the 24% that were reidentified would really feel the impact of this. So, we decided to create a website in which a Harvard College student could enter their concentration, house, and social class, in order to see how many other individuals at Harvard match that criteria. One may come to find that they are one in thirty, or even that they are unique!

In addition to the form in which a student can find out how unique they are at Harvard, we have a page explaining the goal of our project, which is that we wish to demonstrate the sensitivity of our seemingly harmless personal data, like the quasi-identifiers used in our survey. Additionally, we have a page with numerous tips for individuals who are now (rightly) concerned about the privacy of their data, explaining how best to protect their data.

5.1 Project Limitations

Naturally, for a website demonstrating the dangers of one's data being public, we were hesitant to make this website public. Consider that given access to this tool, a malicious actor could determine a student's house if they knew the student's year and concentration, if those alone were enough to identify the student. This may be a long-shot scenario, but again, we'd prefer not to take chances when it comes to the safety and privacy of the students at Harvard.

We considered how to avoid this problem. Without removing this tool entirely, there would be no way to avoid this without protecting our website behind HarvardKey access. Because Harvard students already have access to the Harvard College Facebook, this really poses no increased risk to the community. However, Harvard is not extremely forthcoming with the steps to protect a website behind HarvardKey, and in fact seems a bit reluctant to. We have been unable to find proper instructions on how to even request the usage of the HarvardKey protection system, and for this reason, our website is not currently hosted online, despite being entirely ready for deployment.

We intend to push HUIT further in order to host our website behind HarvardKey, and if not possible, we intend to use Google Authentication, ensuring that those who access our website via Google are required to use a `college.harvard.edu` email address.

6. Conclusion

6.1 Summary of Insights

Our project resulted in several fascinating insights. Overall, we learned that the difficulty of success-

fully 5-anonymizing a dataset and the difficulty of successfully re-identifying a significant portion of survey respondents were not at odds and could in fact simultaneously be true: edge cases of individuals with unique combination of house and concentration (especially concentrations in the arts) made it such that, while it was difficult to successfully re-identify a large portion of respondents, it could be exceedingly easy to re-identify (and thus challenging to protect the anonymity of) select members of the Harvard community.

We also observed that while our success in re-identifying individuals came as a surprise for most survey respondents, most, after further reflection, felt that they should have expected the lack of assurance of complete anonymity afforded to them given the quasi-identifiers they exchange when communicating with other Harvard affiliates on a daily basis.

6.2 Future Directions

In the future, it seems important to conduct a similar analysis on communities beyond the organization of Harvard. This may be important for other college institutions like Harvard University, or even for corporate entities that are inclined to regularly anonymously survey their employees regarding the company. Different quasi-identifiers would have to be identified for different types of organizations.

We would also be interested in conducting further analysis on whether redundancy in responses can lead to unique identification of respondent's information/opinions, even if their information is shared via a dataset that is k-anonymized with $k > 1$.

7. Data Availability

The dataset used was the Harvard Facebook. This is not a publicly available dataset. Access is restricted to Harvard affiliates. For those with access, it can be downloaded as a table in Microsoft Word from this link: https://facebook.college.harvard.edu/

To convert this to a .csv file, as used by the code that can be accessed in the section below, the contents of this table can be copy-pasted into Microsoft Excel, or a similar spreadsheet editor.

8. Code Availability

The code used in this project is publicly available, although the dataset it relies upon is not publicly accessible (see the Data Availability section above). The code can be accessed from here: https://github.com/mjmoorman03/CS105FinalProject

9. Answers to Questions from In-Class Presentation

Question 1:

What if a survey was set up so that respondents used Google Authentication to log in, enabling the surveyor to see the respondents' emails, which may contain their names?

Answer: This would not be compatible with creating an anonymous survey. Surveyors should not set up surveys in a way such that personal, uniquely identifying information about recipients—such as their email addresses—are collected. For instance, in the Harvard Facebook, people can be queried by their email, which enables surveyors to easily identify respondents even if their email address does not comprise of their full name. Additionally, as people often put their emails on public websites and many documents, one can search the Internet or other databases for their email, and find a lot of information about them beyond their experience at Harvard, including personal details regarding that person and their identity. Our survey was created using Google Forms, and we did not select the option to record respondents' emails, so this is one way to prevent this.

Question 2:

In terms of responses, in general, what actual content about themselves can the respondent tell you, and how that can be used to identify them?

Answer: People referencing other Harvard-related affiliations, including club/organization involvements, in survey responses can further help identify those individuals, especially if the survey issuer personally analyzing the response data also has "implicit" knowledge of the student body that may not clearly present itself in the dataset but can be used to link to a survey response dataset. This implicit knowledge may further be used in unexpected "linkage attacks". Essentially, any information that is descriptive of an individual, however simple, can be used to identify them if it offers details about a characteristic that is relatively unique to them.

Question 3:

Would you recommend that Harvard University Information Technology (HUIT) send out the suggestions you made for anonymizing surveys? Or is not really worth it?

Answer: We would recommend that HUIT send out the suggestions we have made for anonymizing

surveys as what students should default to when creating them. This could be done by providing students with an easy-to-use template to increase uptake. We believe that our suggestions preserve the valuable information surveyors need on respondents' backgrounds, while providing strong anonymity for respondents, as explained in Results (Section 3) and Discussion (Section 4). We say that these suggestions are what students should default to, as there may be some exceptional cases where surveyors really need a student's exact house or concentration. One such example is the semesterly survey on a student's experience in residential life. In that case, we suggest eliminating the inclusion of other fields (from house, concentration, year, etc.) that are not necessary. This will help increase anonymity. However, if this is not feasible, and there is good reason for this, the surveyor should at least provide a binding promise not to re-identify individuals. They should also ensure careful treatment of data so it is not put into the hands of one who would try to use it to re-identify and learn information about individuals.

References

[1] Latanya Sweeney. "k-anonymity: A model for protecting privacy". In: *International journal of uncertainty, fuzziness and knowledge-based systems* 10.05 (2002), pp. 557–570.

[2] S De Capitani di Vimercati, Sara Foresti, et al. "Quasi-identifier". In: *Encyclopedia of cryptography and security*. Springer, 2011, pp. 1010–1011.