Jack Deschler

CS 105

Final Project

Expanding the Sample Frame

**Introduction**

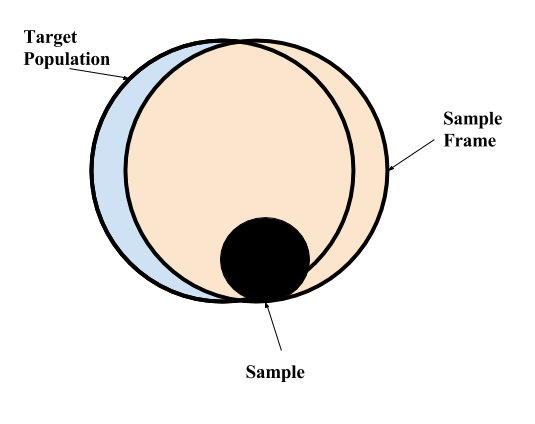
 In late October, the Harvard Undergraduate Council released a survey to the student body meant to measure student opinion on the hot button issue of Unsanctioned Single Gender Social Organizations[[1]](#footnote-1). I have already opined on the untenable sampling method[[2]](#footnote-2), so this project focuses on the UC survey’s other major flaw: it collected virtually no demographic information on those it sampled. The only two pieces of demographic information collected were gender and class year. The lack of demographic information, combined with the eventual low turnout of under 30%[[3]](#footnote-3), mean that virtually no conclusions can be drawn from the survey results. In fact, any “conclusions” drawn may be entirely incorrect.

Figure 1

When administering any opinion survey, there are three important groups: the *target population*, the *sample frame*, and the actual *sample.* The target population is the group that the survey hopes to be able to determine something about; the end result of a survey is hopefully to be able to say something like “the majority of the target population believes X.” The sample frame is the actual set of people that can be reached by the survey, and the sample is the set of people who actually fill the survey out. In typical survey sample paradigms, the three groups are arranged like in Figure 1, where the target population and the sample frame are not exactly the same – the classic example of this is calling cell phones to look for voters, as not all voters have cell phones, and not all with cell phones are voters. Given the methodology of the UC survey – opt-in participation via email – the target population and sample frame were actually the same: all undergraduate students.

The ultimate task for the administrators of a survey is to take the responses that they accrued in their sample, and generalize to the sample frame, and then to the target population. Given that they only had to complete half this process, the UC’s job should have been relatively easy, but they made it impossible by failing to collect any demographic information. The primary way to generalize from the sample to the frame, and then eventually the population, is by demographic weighting, where the responses from different groups in the sample are expanded to the correspond with the percentage of the entire frame that those groups make up. With only gender and class year collected, however, the UC had no way draw any meaningful conclusions about the sample frame through demographic weighting[[4]](#footnote-4). In order to actually make inferences about how the student body actually feels on this, or any issue, more demographic information is needed.

There is a reasonable worry, however, that if too much demographic information is collected in surveys like this, the respondents will lose anonymity. With too much information, the administration could conceivably trace responses back to individual students. Even the thought that the administration could do this could lead to bias in student responses, as students would want to avoid controversial responses for fear of reprisal from the administration. The UC survey protected this to the extreme. The smallest Harvard Houses have around 360 students, so with an even gender split, the UC protected anonymity on the level of about *180*-anonymity. This is far, far more But how much could they have colinions as white women, and that Economics concentrating men held the same opinions as Philosoph anonymous than the normally accepted threshold of *5*-anonymity. The UC could have reasonably collected significantly more demographic information and still maintained a sufficient degree of anonymity. But what should they have collected? What could they have collected?

**Project Goal**

This project is motivated by three principles. First, it is important to maintain at least a level of *5*-anonymity when surveying students at the College. Second, when surveying students, as much demographic information as possible should be collected, in order to draw the strongest inferences possible from the results. Third, the second principle should never be pursued at the cost of the first. Therefore, there may exist combinations of demographic information that may not be able to be collected, and there may be certain pieces of demographic information that must be generalized if they are to be used in combination with others. Any piece of demographic information collected could be used as a part of a quasi-identifier to re-identify students in a survey.

The goal of this project was originally to simply to determine how to generalize concentration data such that *5*-anonymity was maintained when surveying students and asking for concentration as a piece of “demographic” information. However, it is important to note that for different questions, different demographic items will be more relevant than others. For example, if the College wanted to poll students on advising in the Houses, items like House and concentration would be good to know. For questions on sexual health and support, gender would be more salient. Indeed, even within one question, certain pieces of information may be more relevant. It is a rather trivial exercise to come up with questions that the College may be interested in knowing the answer to, and that would require any different subset of demographic items. In the interests of never writing the same piece of code twice, the project has thus become more general. I have built an algorithm that, given a quasi-identifier *Q*, will generalize the fields of *Q*, in order of preference, such that a given level of *k*-anonymity can be assured. The code protects the most valuable pieces of data, and is adaptable to any *Q* that is a subset of the fields that are contained in the data. In this way, I can analyze different quasi-identifiers, as well as the effect that different levels of anonymity have on the generalization process.

**Gathering Data**

Initially, I hoped to gather data by scraping it concentration by concentration from the Harvard Facebook, which would have, ironically, been a violation or the Privacy of Information policy[[5]](#footnote-5). Instead, I emailed the academic coordinators of all the undergraduate Houses, asking them to send me a list of students without names, but with concentrations, class years, and genders. I received usable data from 7 of the 12 Houses, though some were missing gender, and one was missing class year. Importantly, there is much more demographic information that the University may want to collect in surveys, such as race, sexual orientation, family income, and so on. I could not get this data, but the algorithm I hoped to write would still be able to work with these categories if the administrators had that data.

I decided to impute both the missing genders and the missing class years randomly. I am aware that this approach has its issues – the Crimson has recently opined on the lack of women in the Math department[[6]](#footnote-6). That said, I thought having more, if imperfect, data would be better for the overall analysis than missing even more data. Additionally, this project is not intended to present an end-all-be-all solution to collecting demographic information in surveys, but rather to present a method that will allow administrators to determine how they should generalize categories in order to maintain different levels on anonymity. The data I have, even were it to be complete, only represents one snapshot in time. Because of the size of some departments, one student joining the department in a certain year in a certain House could drastically alter the analysis. The process of creating the generalizations is more important than any individual result actually is.

Before beginning the project, I made a few adjustments to the data that I had collected. First, in the case of joint concentrations, I only listed the primary concentration. As a joint concentrator myself, I could be identified simply by giving my House and concentration, and it is conceivable that some could be identified with just their concentration. So, rather than square the number of possible categories, which would drastically increase the runtime of any algorithm, I only used primary concentration. Accordingly, survey administrators should be asking for “primary concentration” not just “concentration.” Second, I added a column to my data with a code[[7]](#footnote-7) for the overarching subject of the concentration: AH for Arts and Humanities, NS for Natural Sciences, SS for Social Sciences, SEAS for Engineering and Applied Sciences, and NO for cases like Special Concentrations and Undeclared students. When generalizing concentration data, we want to keep like concentrations together if possible, as the conclusions we draw would be more valuable[[8]](#footnote-8). It seems more informative, for example, to say that “X percent of Comparative Literature/Folklore and Mythology concentrators believe Y” as opposed to “X percent of Comparative Literature/Mathematics concentrators believe Y.” Finally, I removed subfields of concentrations from my analysis, for the same reasons as working with only primary concentrations. Thus, *Anthro: Archaeology* is treated the same way as *Anthro: Social Anthropology:* simply as *Anthro*.

**Algorithm**

The algorithm I elected to use is a greedy algorithm, meaning that it prioritizes combining the smallest sized categories when generalizing. The algorithm can be summarized as follows, when aiming for *k*-anonymity:

1. Categorize each student according to their value on the quasi-identifier *Q*
2. Count each value of *Q* present in the set of all students
3. If all counts are greater than or equal to *k*, we are finished.
4. Combine the two smallest-count categories of *Q* together, on the axis of the lowest priority piece of demographic information (least necessary to keep separate)
5. Return to 1, and repeat as needed.

The asymptotic runtime of this algorithm will be , where *n* is the number of possible permutations of *Q* and we are aiming for *k*-anonymity, as we will have to do some factor of *n* combinations in the worst case to go up a levels of anonymity. Note that including all possible joint concentrations would cause this to balloon. While this algorithm may not create the ideal solution, it runs in polynomial time for a given *k*, which an ideal algorithm may not, and greedy algorithms often give decent approximations of the ideal solution. This analysis is consistent with the idea that increasing the level of anonymity requires more generalizations, and is thus much more difficult, than anonymizing to lower values of *k*.

**Code Walkthrough**

The algorithm’s code is below in Appendix I[[9]](#footnote-9). The lumper function takes a pandas dataframe as its first argument, where one row represents one student, and each column is a different piece of demographic information collected (or imputed). The columns argument is a list of column names in the dataframe that will be generalized – essentially it is the quasi-identifier. The key piece is that these column names are *in order of priority to generalize*. In other words, columns[0] will be fully generalized before columns[1] is generalized at all. In this way, the code protects more important pieces of demographic information from being smudged. The variable current\_col keeps track of the column that is currently being generalized. The variable k gives the level of *k*-anonymity that must be reached, and defaults to 5. The outfile argument is simply the name of a text file to print the steps in generalizing to.

The inner function new\_lump\_col essentially accomplishes step 1 of the algorithm. It creates a new entry for each row in the dataframe that stores the combined value of the quasi-identifier, which is then counted in step 2. The pandas functions idxmin and nsmallest are used to find the smallest sized value counts, and thus what to generalize each time through the loop[[10]](#footnote-10). The inner function lump takes two values of a column, and turns them into the same value, basically accomplishing step 4. Each time through the loop, there are checks to see if the next column is needed, or if the goal of k-anonymity has been reached. The final while loop near the bottom of the function is what ensures that each concentration is only combined with concentrations from the same overarching distribution.

A second version of this code would contain more error checking, and better behavior on errors. Currently, the code simply breaks and throws an error if the desired level of anonymity cannot be reached. There are also corner cases where the code could get caught inside the last loop when trying to merge concentrations from the same overarching subject area. As these would only manifest themselves with very high degrees of anonymity, I chose to focus my efforts elsewhere.

**Description of Results[[11]](#footnote-11)**

After writing the code, I ran the program on different quasi-identifiers of the data I was able to collect, and different levels of *k*-anonymity. Generally speaking, I found that including more pieces of demographic information in the quasi-identifier and using higher values of *k* caused more generalization to occur, blurring the least protected piece of information. This was expected, and it makes sense given how *k*-anonymity is defined. More importantly, I found that one *can* generalize to a level of *5*-anonymity, and perhaps beyond, and still obtain useful demographic information, that could be used for weighting samples, and drawing more accurate and meaningful conclusions.

When generalizing on the ordered set of demographics consisting of *concentration* and *house* (meaning that concentration is generalized first), reaching *5-*anonymity required about 40 generalization steps. To reach a level of *7*-anonymity, about 5 more steps were required[[12]](#footnote-12). This indicates that in the real world, not the asymptotic world, runtime increases may not be as costly as predicted for higher levels of anonymity. Perhaps more importantly, however, reaching higher levels of anonymity requires larger and larger departments to be blended together in the name of privacy. The biggest “name-brand” department, if you will, that was no longer its own category for *5*-anonymity perhaps was Mathematics, or Earth and Planetary Sciences. Moving to *7*-anonymity requires the generalization of the Social Studies department, and moving much higher may put even larger departments, like Government and Computer Science, at risk. It is important, especially with large departments, to be keep as many quasi-identifier field entries in their own category as possible, in order to be able to use demographic weighting, and draw strong conclusions. Moving to stronger levels of anonymity predictably puts that at risk. So even though higher levels of anonymity may not take much longer to generalize, they perhaps should not be sought once an acceptable threshold is reached.

Another important finding is that the order of generalization matters. When I flipped the order of the *house* and *concentration*, to generalize *house* first, no single House survived as its own category. This indicates that, if we want to collect *concentration*, we should generalize on that piece of information first, as the concentration data makes the categories too fine grained for any other category to be generalized – *house* is the next most general I collected – if we wish to reach *5*-anonymity. Should survey administrators have access to more demographic information than I did, they should take care in the order in which they generalize their fields. The code that I wrote ensures that privacy is maintained by reaching *5*-anonymity, but it does which pieces of information are collected, and how they are generalized, is up to the user.

Finally, in accordance with the issue that originally sparked this project, I ran my code with *class year* and *gender* as the two quasi-identifier fields, the same two pieces of demographic information collected by the UC survey in late October. As I suspected, no generalizations were required. I then added the *house* field to my quasi-identifier, and still, no generalizations were required. Even adding *concentration* only required 40 steps, a small number when one considers what information could have been gleaned from the question. Perhaps Computer Science concentrators feel largely different than Government concentrators do on certain issues. If we do not collect the requisite information, we will not know. The UC could have, and they did not. Hopefully, moving forward, that will be rectified.

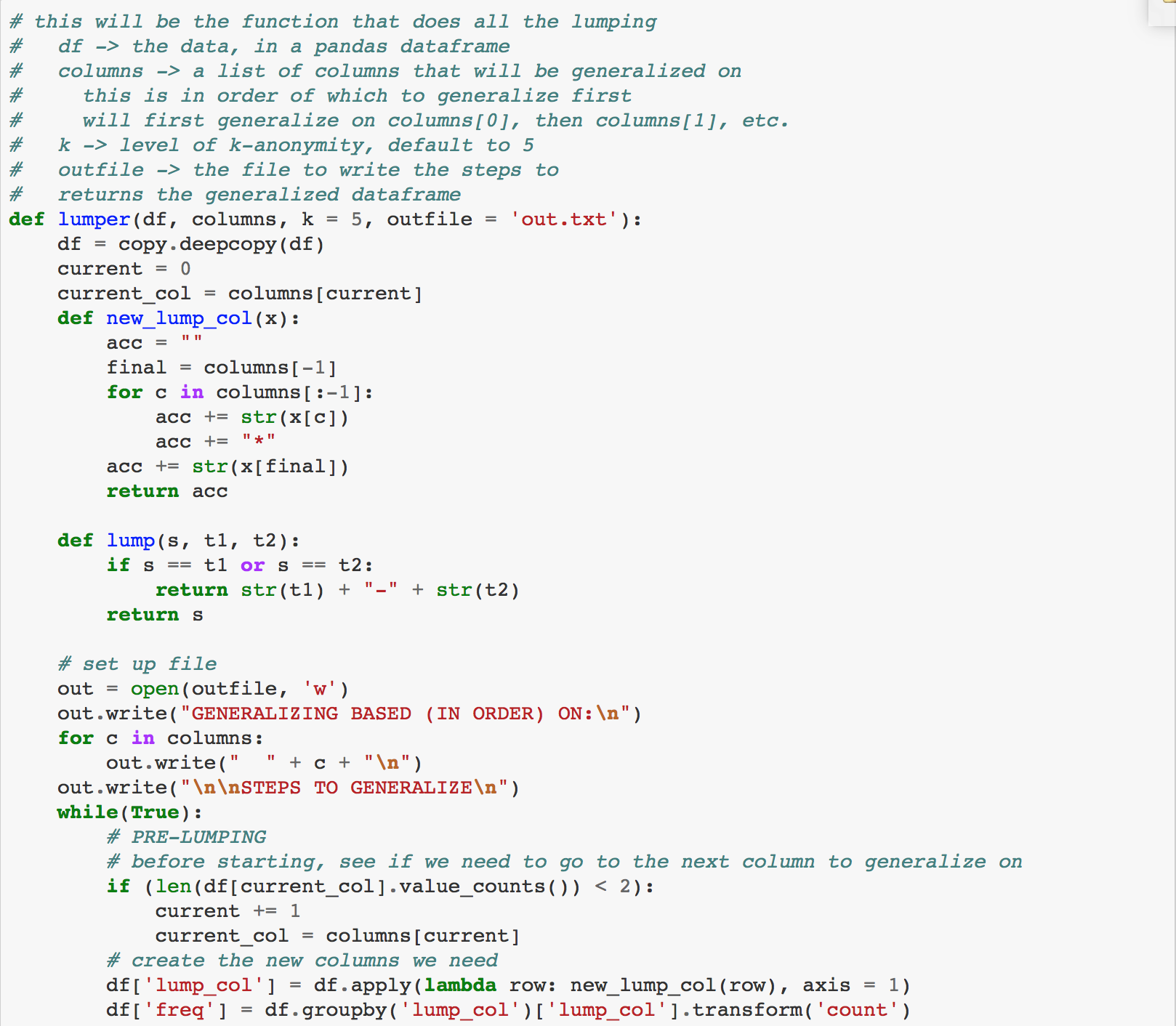
**Addressing Questions**

It is important to note that the work I have done cannot save a poorly conceived survey. Finding a way to generalize demographic information while still maintaining privacy in responses, as I have done, does not mitigate non-response bias, social desirability bias, sample frame error, or any of the multitude of other types of errors that can be present in surveys. Even collecting all demographic information possible would not have saved the UC survey from is abysmal response rate and poor sampling method. Anonymizing data generally makes it less useful for demographic weighting, and thus drawing conclusions. The process of generalization “lumps” together two categories into one, over and over, in order to reach the desired level of anonymity. In other words, this code makes data *worse* for predictions, not better. If the starting point is already bad, biased data, this code will only serve to make that worse. If, however, a survey is backed by sound methodology, then this code demonstrates that we can still collect relevant demographic information and maintain anonymity at the same time. Surveys should still be done via random sampling, not opt-in non-probability sampling, but this project shows that they can, and therefore should, contain questions pertaining to relevant demographics.

**Conclusion**

If we are going to survey the student body at Harvard, we should do the best job we can. In many cases, this will require collecting demographic information in responses. While properly representing student opinion is important, protecting students against potential reprisal by backtracing survey responses is more important. If a student fears that his or her survey response will be traced back to them, they may answer dishonestly, or not answer at all. This cannot stand, and so anonymity must be protected. I offer an algorithm, albeit a simple one, already built, that can offer a way to ensure that anonymity. Through generalization, administrators can collect information and still ensure privacy. Though my code was only tested on *house*, *gender*, *concentration*, and *class year*, it was built to function for any demographic labels that survey administrators may possess. If the inputs to a decision making process are bad, so too will be the final decision. The information going in, especially when making decisions that have long, far reaching effects, must be as sound as possible. This project demonstrates that the information provided from surveys can be better, and not at the cost of the anonymity and privacy of the respondents.

*Appendix I: Code*



1. http://www.thecrimson.com/article/2017/10/27/uc-usgso-survey/ [↑](#footnote-ref-1)
2. http://www.thecrimson.com/article/2017/10/31/deschler-framing-student-voice/ [↑](#footnote-ref-2)
3. http://www.thecrimson.com/article/2017/11/9/uc-survey-sanctions-results/ [↑](#footnote-ref-3)
4. The UC likely could not even make inferences from the information they did collect. Had they tried, for example to draw a conclusion from their gender information, they would have been assuming that black women held the same opinions as white women, and that Economics-concentrating men held the same opinions as Philosophy-concentrating men, for example. [↑](#footnote-ref-4)
5. https://facebook.college.harvard.edu//compilation\_forbidden.html [↑](#footnote-ref-5)
6. http://www.thecrimson.com/article/2017/10/20/everyday-struggle-women-math/ [↑](#footnote-ref-6)
7. This was done by hand, so there may be mistakes in coding. Hopefully Harvard would not make such errors. [↑](#footnote-ref-7)
8. I did not do this for Houses and neighborhoods because the Quad seems to be the only neighborhood with a discernible neighborhood identity. [↑](#footnote-ref-8)
9. I also turned it in on canvas as both a PDF and Jupyter Notebook, along with the data I used in a CSV file. [↑](#footnote-ref-9)
10. I experimented with using different functions, such as both the pandas and numpy versions of argmin, and these gave me the best results. [↑](#footnote-ref-10)
11. I have not included the text files here as they are easily reproducible with the included code. [↑](#footnote-ref-11)
12. I say “about” because I am working with imperfect data for only 7 of the 12 Houses. [↑](#footnote-ref-12)