CS224W Project
Using Network Features to Identify Stratified Student Behavior During College Admissions

Noah Arthurs (narthurs@cs.stanford.edu)

December 11, 2019

1 Introduction and Background

College admissions in the US Is a form of high-stakes assessment in which admissions officers use a process known as "individualized holistic review" (IHR) to make admissions decisions on students. IHR is a human-centered approach that combines quantitative aspects of a student's application (i.e. grades, test scores) with qualitative aspects (i.e. admissions essays). IHR has a fraught history, as the inclusion of personal essays in applications was originally motivated by the desire to reject Jewish applicants with high test scores [21, 3]. Since then, the narrative has shifted, as IHR is currently used as a vehicle for affirmative action [16], but its legitimacy has been tested in several recent supreme court cases.

Central to these debates around IHR is the question of whether the qualitative aspects of holistic review create a fairer or less fair process. Proponents argue that personal essays can help offset the fact that quantitative aspects are strongly influenced by demographic features like race, gender, and income. This argument is rooted in the idea that in order for college admissions to be fair, they need to actively resist the inequality that exists in society. In order to accomplish this, educators need to have an understanding of how demographic variables influence the approach that students take to the admissions process. Due to the lack of availability of large-scale datasets, our current understanding of this relationship is based primarily on qualitative studies. This study seeks to remedy the lack of quantitative results by using machine learning techniques to analyze a dataset of almost 60,000 college applications. Specifically, we will be trying to quantify the relationship between a student's background and their approach to the admissions process.

Our dataset includes three types of features for each student:

- **Background Features**, including gender, reported household income, and high school.
- **Performance Features**, including SAT/ACT scores and GPA.
- **Admissions Process Features** including schools applied to, essay prompts chosen, and the text of the essays themselves (around 240,000 total).

We focus on the first and third sets of features in order to answer two categories of questions. First, we are interested in describing how the demographic variables of gender and income influence the admissions process features. Second, we are interested in how school-level effects interact with these demographic effects. Specifically, we would like to know whether a student's high school or background has more influence on their approach to their college applications.

To tackle both sets of questions, we set up graphs with student nodes, high school nodes, and college nodes, and use graph-based algorithms to extract additional features for each student. Because these features

take into account not only applications, but also each student's role in a complex network of applications, they give a richer representation of a student's position within the college admissions process. We then use supervised machine learning algorithms to describe the relationship between node features and demographic labels (specifically gender and income).

2 Related Work

2.1 College Admissions

As mentioned above, the majority of related research in the college admissions space is qualitative. Kirkland and Hansen [8], and Vidali [4] discuss the personal essay as a vehicle for performing race and disability respectively from the student's perspective, while Lewis [6] explores it on a larger scale. In addition, Early and DeCosta-Smith [5] tested the ability of different types of interventions to improve the admissions essays of underserved students. The throughline of these studies is that they explore the complicated relationship between identity and the college application. Specifically, we must bear in mind that a student's self-presentation [1] is always influenced by the pressure to perform identity as part of the admissions process.

The quantitative results that do exist tend to focus on predicting college performance. Notably Pennebaker et al. [15] showed that the usage of particular words in admissions essays can predict college grades. Alarmingly, Jones [11] shows that privately-educated students are more likely to use cultural signals in their applications, indicating that they are more prepared to "play the admissions game" than their publicly-educated counterparts. His findings suggest that admissions coaching plays a large role in creating this disparity, which adds another wrinkle to the already complex relationship between income and admissions behavior.

2.2 NLP Approaches

One feature that makes our dataset so powerful is the presence of the raw text of over 800,000 admissions essays. Although there has been little to no NLP work done on admissions essays, there is research that identifies connections between writing and demographic features. Hovy [17] found that age and gender features improve accuracy in a variety of text classification tasks. Furthermore, it is well-known that demographic information can be inferred from natural language and other human behaviors [20, 18]. In addition, there is a recent push to incorporate author-information into word embeddings [19, 22]. Overall, it appears that the relationship between natural language and demographic is an active subfield of NLP.

2.3 Graph-Based Approaches

In addition, it has been shown many times that graph-based features are useful for predicting people's membership in particular groups or demographics. Agrawal et al. [2] found that graph-based behavioral features can be even more predictive of group-membership than text-based features. More recently, Brea et al. [13] successfully predicted demographic information from a mobile-phone based social graph. In addition, as Filippova [9] demonstrated, it can be extremely powerful to combine graph-based and NLP features.

3 Data

Our dataset contains 58,405 applications submitted to a large, selective, state university system by Latinx students in 2017. This dataset includes all US-based applicants who self-identified as latinx in 2017. Students for whom we are missing important information were then filtered out. We focus on the following features that are available for all remaining students:

- Previous school attended (these are mostly high schools, but 20% of students were applying to transfer from another college). There are 4,604 total previous schools in our dataset.
- Four admissions essays written (chosen from among 8 prompts)
- Colleges applied to (within this state university system), of which there are 9.

In addition, we choose to use reported household income and reported gender as our labels for classification. As indicated in Table 1, most but not all applicants chose to report gender and household income:

Students	Reported Gender	RG: Female	RG: Male	Reported RHI	Median RHI
58,405	58,081	34,710	23,371	52,670	$43,000

Table 1: Breakdown of students by year. Not every student reported gender or an RHI of at least $10,000.

3.1 Reported Household Income

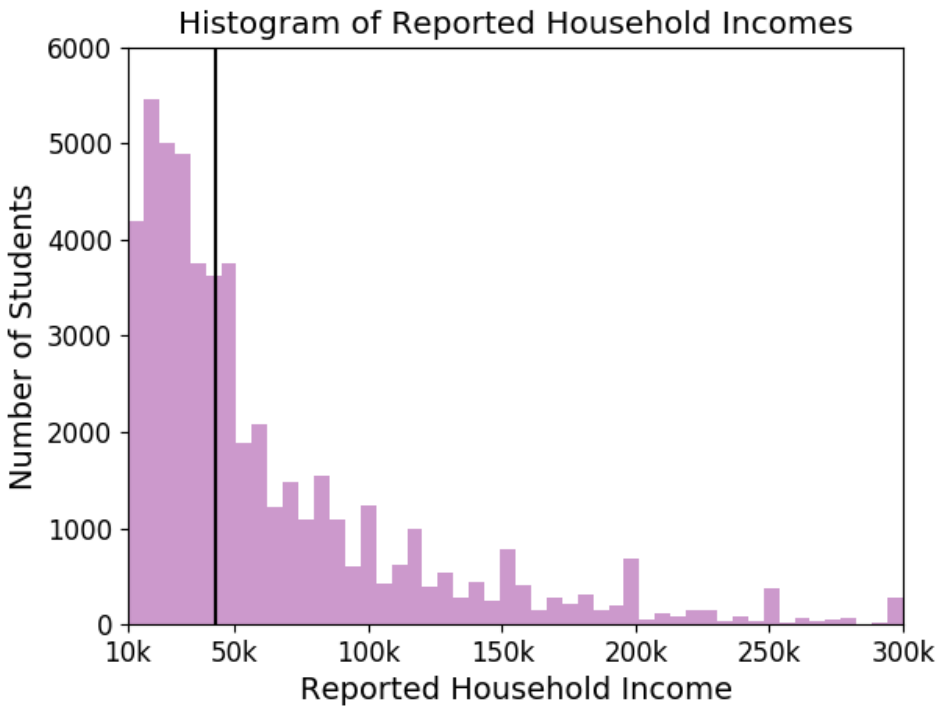

Figure 1: RHI histogram. The black vertical line represents the median income. 1,253 students whose RHI was over $300,000 were cut off.

Our socioeconomic status information comes from the reported household income (RHI) provided by each applicant. It is important to note that RHI is not an objective measure of a family's household income. Students may not be fully aware of how much money their family makes, and they also may not fully understand the question. For example, some students seem to have reported their income in thousands of dollars rather than dollars. We take two precautions to mitigate the noise of this feature. First, we focus on classifying students as being above or below median income with the thought that these binary labels will be more stable than the exact values reported by the students. Second, when predicting income, we filter out any student whose RHI is below $10,000. While it is possible that there are students whose household income is this low, we have found that our ability to classify students goes up when we remove these students, indicating that the signal coming from these RHI values is unhelpful. Figure 1 shows the RHI distribution for each year of our dataset.

3.2 Reported Gender

We used applicant's reported gender (RG) as our other classification outcome. If students chose to report their gender (not everyone did), they were limited to "Male", "Female". We recognize that the traditional gender binary is not an accurate measure of gender identity, and we must keep in mind that this study is limited by the fact that students could only sort themselves into these two categories. Table 1 shows the RG breakdown by year.

4 Methods

Our hypothesis is that a student's background (race, gender, geography, and income) has a strong influence on that student's college application. Through this project, we hope to quantify that influence and begin to describe how it manifests itself. In order to do this, we will try to create interpretable graph-based models that predict background information based on a student's college application. Our methodology contains 4 steps:

1. **Graph Creation**: first, we need to choose a graph representation for our data.
2. **Student Node Featurization**: Next, we use graph-based algorithms to generate features or labels for each node. It is best for these features to be as interpretable as possible.
3. **Classification**: We can then use the detected features to classify students into demographic groups. We are most interested in classifying students by gender, income, and geography (high school). We will err on the side of simpler, more interpretable models.

With this methodology, we can measure our success based on (1) how accurate our classification is in step 3, and (2) how well we can interpret those results.

In the following three sections, we discuss the methodologies we have tried for each step.

4.1 Network Construction

We build a tri-partite graph where students, previous schools, and colleges are represented by nodes. Edges connect students to the high schools they attended and to the colleges they applied to. This network setup ties a student's educational background (previous school) to their decision making during the admissions process (schools applied to).

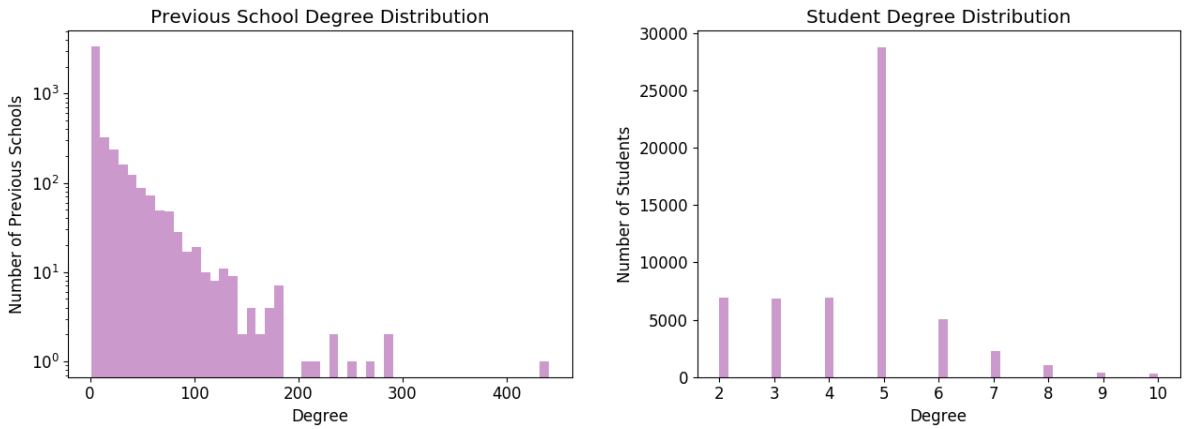

Figure 2: Degree distributions for previous school and student nodes.

School Index	1	2	3	4	5	6	7	8	9
Degree	21157	20709	30008	30001	13140	22828	25651	20408	23801

Table 2: Degrees for the 9 colleges applied to nodes.

Figure 2 and table 2 show the degree distributions for the 3 kinds of nodes. Note that the degree distribution curve for high school nodes is roughly exponential (decreasing), with a large number of schools having a small number of applicants and a small number of schools having a large number of applicants. Also note that the degree of student nodes is 2 at minimum, since each student must report their previously attended school and must apply to at least one college.

This graph has a total of 63,018 nodes, including 58,405 student nodes, 4,604 previous school nodes, and 9 colleges applied to nodes. The entire graph forms a single strongly connected component with a diameter of 6. and a 90% effective diameter of 3.67.

4.2 Student Featurization Methods

Because we are working in a novel domain, we have decided to use well-established techniques to extract node features out of the network and student essays.

4.2.1 Doc2Vec Features (D2V)

In order to featurize a given student's writing, we first concatenate all four admissions essays into a single document. Next, we use Doc2Vec [14] to learn 1000-dimensional embeddings to be used as node features. Doc2Vec involves training a Word2Vec model [12] where the features used to predict the center word are augmented with a document embedding.

Our Doc2Vec model was trained for 10 epochs using negative sampling [12] and a single 1000-dimensional hidden layer, which is concatenated with our 1000-dimensional document embedding for the purposes of making predictions.

4.2.2 Neighbors

The simplest way to incorporate a node's position in the graph into our model is to take into account its list of neighbors. In our case, a student-node's neighbors consist of:

1. schools applied to, which we encode as a 9 dimensional vector whose ith entry is a 1 if the student applied to school i and a 0 otherwise.
2. previous school attended, which we encode as a 4,604-dimensional one-hot vector.

It is important to establish how predictive a node's neighbors are, because we need to know if the recursive techniques below are improving over just using neighbors or not.

4.2.3 Recursive Features

We use a modified ReFeX [7] to perform recursive feature extraction on our nodes. Let $\mathbf{v}_u^{(i)}$ refer to node u's feature vector after i recursive steps, with $\mathbf{v}_u^{(0)}$ referring to the vector of basic features.

In ReFeX, $\mathbf{v}_u^{(0)}$ involves the node's degree, the number of edges leaving the node's egonet, and the number of edges within the node's egonet. These first two features still work for us, but because our graph is bipartite (we can view the previous school nodes and colleges applied to nodes as a single partition), the third feature would always be 0. The spirit of this third egonet feature is to capture the connected-ness of the nodes in the egonet. In order to capture this spirit, we will instead use the expected edge weight between u's neighbors in their one-mode projection. Mathematically, if we let n_1 and n_2 be randomly selected neighbors of u, then we want:

$$(\mathbf{v}_u^{(0)})_3 = E\left[|N_{n_1} \cap N_{n_2}|\right]$$

For non-student nodes this is slow to compute, so we approximate it through sampling. If u is a non-student node and $(n_1^{(1)}, n_2^{(1)}), ..., (n_1^{(100)}, n_2^{(100)})$ are randomly sampled (without replacement) pairs of nodes from u's egonet, then:

$$(\mathbf{v}_u^{(0)})_3 = \frac{1}{100}\sum_{i=1}^{100} |N_{n_1^{(i)}} \cap N_{n_2^{(i)}}| \approx E\left[|N_{n_1} \cap N_{n_2}|\right]$$

When performing recursive steps, we use mean and sum aggregation as in [7]:

$$\mathbf{v}_u^{i+1} = \left[\mathbf{v}_u; \frac{1}{|N(u)|}\sum_{w\in N(u)} \mathbf{v}_w; \sum_{w\in N(u)} \mathbf{v}_w\right]$$

Since the 90% effective diameter of our dataset is 3.67, we only run our algorithm for 4 iterations, as this should allow all nodes to receive information from most other nodes. Since this results in a very reasonable number of features, we do not perform feature pruning, and instead allow our classifiers decide for themselves what features are useful.

To verify that our ReFeX features are useful, we subjected them to a brief Rolx [10] analysis. In order to accomplish this analysis, we performed non-negative matrix factorization (NMF) to find 4 features (soft classifications) for every student node. Below are the Pearson correlations between those features both RHI (continuous) and RG (binary labels) :

Feature Index	1	2	3	4	5
RHI Correlation	-0.54	0.18	0.05	-0.13	0.006
RG Correlation	-0.06	0.01	-0.05	-0.015	-0.05

Table 3: Correlations between Rolx features and RHI, RG.

We find that a few features have strong to medium correlations with RHI, while no features have strong correlations with RG. This suggests that our recursive features might be more predictive of RHI than of RG.

4.3 Classification

Let X represent some combination of the above features, and let $Y \in \{1, 0\}$ represent the label we are trying to predict for X. In the case of RG, $Y = 0$ represents "Male" RG and $Y = 1$ represents "Female" RG. In the case of RHI, $Y = 0$ indicates median or below RHI and $Y = 1$ indicates above-median RHI.

For classification, we use Logistic Regression, a simple discriminative model that learns a linear decision boundary. For a given x, we have:

$$P(Y = 1|X = x) = \sigma(\theta^T x + \theta_0)$$

where θ, θ_0 are the parameters of our model and σ is the sigmoid function, $\sigma(z) = \frac{1}{1+e^{-z}}$. We then classify our example as a 1 if $P(Y = 1|X = x) > 0.5$ and as a 0 otherwise. We then train our parameters to minimize cross entropy loss:

$$\hat{\theta}, \hat{\theta_0} = \underset{\theta,\theta_0}{\operatorname{argmin}} \left(-\sum_{i=1}^{n} y^{(i)} \log(P(Y = 1|X = x^{(i)})) + (1 - y^{(i)}) \log(1 - P(Y = 1|X = x^{(i)})) \right)$$

where $x^{(1)}, ..., x^{(n)}$ are the training examples and $y^{(1)}, ..., y^{(n)}$ are their corresponding labels.

Although this is a very simple model, it has proven to be powerful in our experiments. It is very stable to train (the objective is convex), and in early experiments, it performed as well as feed forward neural networks with three hidden layers. Although deeper networks might be marginally better at predicting the training labels than logistic regression, logistic regression is a principled approach with easily interpretable parameters.

5 Results

During training, we use an 80/20 trian/test split. All reported accuracies are test accuracies. We trained our logistic regression classifier with many different combinations of features. The baseline classifier gets no information and always predicts the most common label:

Features Used	RHI Test Accuracy	RG Test Accuracy
None (Baseline)	50.30%	59.76%
D2V	67.14%	81.72%
Colleges Applied	61.72%	59.76%
Previous School	65.24%	60.68%
Colleges Applied, Previous School	68.12%	60.30%
Recursive Features	72.27%	61.12%
D2V, Colleges Applied, Previous School	70.37%	82.44%
D2V, Recursive Features	75.45%	82.88%
All Features	76.16%	83.02%

Table 4: Test Accuracy of Logistic Regression for predicting above/below median reported household income and reported gender for each year on different combinations of features.

By comparing the increases in accuracy that come from adding different sets of features to the model, we learn the following:

- D2V features contain a large amount of information about both labels, but more information about RG (as there is a higher increase in accuracy from the baseline). This indicates that a student's background has a strong influence on their word choices.
- Colleges Applied and Previous School (the neighbor connections) both contain a good amount of information about RHI but no information about RG.
- The recursive features contain even more information about RHI than the neighbor connections, and a very small amount of information about gender.

Overall what we have learned is that while student writing varies according to both income and gender, a student's participation in the admissions graph mainly varies according to income. This is consistent with our correlation analysis of the Rolx features in section 4.2.3.

6 Conclusion

Our most significant result is that a student's position/role in the admissions graph we have created is a strong signal for their socioeconomic status. This indicates that students of different backgrounds take very different approaches to the admissions process. On the one hand, this is not surprising: it is well known that higher-income students have more access to guidance couselling and other assistance with the admissions process than lower-income students. On the other hand, it is surprising that the difference is so stark that we can predict whether a student is above or below RHI with more than 70% accuracy from the admissions graph alone. This speaks to the need for more equity in the college admissions process. Of course, when making their decisions admissions offices try to take RHI into account and correct for inequality, but nevertheless lower income students cannot set themselves up for success if they are not applying to the right schools.

Also troubling is the fact that the student writing on its own is so stratified according to income. This indicates that there is the potential for implicit bias on the part of the admissions officers when reading these essays.

We also found that students of different genders write significantly differently, but our findings suggest that the differences stop there and there are not significant differences between men and women when it comes to their participation in our admissions graph.

In the future (or if this author had not misjudged how quickly the deadline was approaching), it would be worth doing a deeper interpretation of the logistic regression weights and Rolx features in order to better understand how these differences manifest themselves.

7 Acknowledgements

Thanks to the Student Narrative Lab in the GSE for allowing me to work with this amazing dataset and helping me get my education-related citations right.

References

[1] Erving Goffman et al. *The presentation of self in everyday life*. Harmondsworth London, 1978.

[2] Rakesh Agrawal et al. "Mining newsgroups using networks arising from social behavior". In: *Proceedings of the 12th international conference on World Wide Web*. ACM. 2003, pp. 529–535.

[3] Jerome Karabel. *The chosen: The hidden history of admission and exclusion at Harvard, Yale, and Princeton*. Houghton Mifflin Harcourt, 2006.

[4] Amy Vidali. "Performing the rhetorical freak show: Disability, student writing, and college admissions". In: *College English* 69.6 (2007), pp. 615–641.

[5] Jessica Singer Early and Meredith DeCosta-Smith. "Making a case for college: A genre-based college admission essay intervention for underserved high school students." In: *Journal of Writing Research* 2.3 (2010).

[6] Rachel Devorah Lewis. "The Rhetorical Legacies of Affirmative Action: Bootstrap Genres from College Admissions through First-Year Composition". In: (2010).

[7] Keith Henderson et al. "It's who you know: graph mining using recursive structural features". In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2011, pp. 663–671.

[8] Anna Kirkland and Ben B Hansen. ""How Do I Bring Diversity?" Race and Class in the College Admissions Essay". In: *Law & Society Review* 45.1 (2011), pp. 103–138.

[9] Katja Filippova. "User demographics and language in an implicit social network". In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics. 2012, pp. 1478–1488.

[10] Keith Henderson et al. "Rolx: structural role extraction & mining in large graphs". In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2012, pp. 1231–1239.

[11] Steven Jones. ""Ensure that you stand out from the crowd": A corpus-based analysis of personal statements according to applicants' school type". In: *Comparative Education Review* 57.3 (2013), pp. 397–423.

[12] Tomas Mikolov et al. "Distributed representations of words and phrases and their compositionality". In: *Advances in neural information processing systems*. 2013, pp. 3111–3119.

[13] Jorge Brea et al. "Harnessing mobile phone social network topology to infer users demographic attributes". In: *Proceedings of the 8th Workshop on Social Network Mining and Analysis*. ACM. 2014, p. 1.

[14] Quoc Le and Tomas Mikolov. "Distributed representations of sentences and documents". In: *International conference on machine learning*. 2014, pp. 1188–1196.

[15] James W Pennebaker et al. "When small words foretell academic success: The case of college admissions essays". In: *PloS one* 9.12 (2014), e115844.

[16] Lorelle L Espinosa, Gary Orfield, and Matthew N Gaertner. "Race, class, and college access: Achieving diversity in a shifting legal landscape". In: (2015).

[17] Dirk Hovy. "Demographic factors improve classification performance". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2015, pp. 752–762.

[18] Elena Tutubalina and Sergey Nikolenko. "Automated prediction of demographic information from medical user reviews". In: *International Conference on Mining Intelligence and Knowledge Exploration*. Springer. 2016, pp. 174–184.

[19] Aparna Garimella, Carmen Banea, and Rada Mihalcea. "Demographic-aware word associations". In: *Proceedings of the 2017 conference on empirical methods in natural language processing*. 2017, pp. 2285–2295.

[20] L Podoynitsina et al. "Demographic prediction based on mobile user data". In: *Electronic Imaging* 2017.6 (2017), pp. 44–47.

[21] Marcia Synnott. *The half-opened door: Discrimination and admissions at Harvard, Yale, and Princeton, 1900-1970*. Routledge, 2017.

[22] Kevin Tian, Teng Zhang, and James Zou. "CoVeR: Learning Covariate-Specific Vector Representations with Tensor Decompositions". In: *arXiv preprint arXiv:1802.07839* (2018).