Ann Tran
May 13, 2016
CS 207: Data Science and Visualization
Sorelle Friedler

The Untold Narrative:
Data Visualizations of Honor Code Trial Data
Haverford College

1. INTRODUCTION.

Data Set

At the beginning of this semester, Mary Glaser, Lucas Franca, Ryan Herlihy and Dylan Emery and I began working with a data set of Haverford College Honor Code abstracts that Honor Council co-chair Brian Guggenheimer came to us with. The data set dated from Spring 1993, and every row in the file represented a catalogued case and every column denoted a detail of the case. Most features contained numerical data, where column values were either "0" or "1". The only columns with categorical data were "Release Semester and Date", "Confrontation Type", "Type of Trial", "URL" and "Abstract Name".

Data Cleaning

Before analyzing any data, missing data must be filled in. I had difficulty learning Python early this semester and could not properly convert categorical data into a CSV as numerical data. More specifically, my program appended rows instead of columns, and all appended rows were filled with "1"s. As a result, I used team member Lucas' cleaned file to run analyses.

In his data set, Lucas turned the category values of "Type of Trial", "Confrontation Type" and "Release Semester and Year" into numerical data of "0"s and "1"s. Additionally, he removed any case with over half of its attributes missing. The remaining trials with missing values were filled with the values of its nearest neighbor (NN). Although there are many ways to fill in missing data (random value within a given column, mean of column values, etc.), the NN algorithm is an excellent predictor of possible values for missing data based on related cases. Additionally, NN limits possible bias since the data used comes from a pre-existing point.

The NN algorithm optimizes finding the point closest to the query point. To measure distance between points, I used the Euclidean distance as the distance metric. The formula is below.

$$\mathrm{d}(\mathbf{p}, \mathbf{q}) = \mathrm{d}(\mathbf{q}, \mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

In this equation, $\mathbf{p} = (p_1, p_2, ..., p_n)$ and $\mathbf{q} = (q_1, q_2, ..., q_n)$ are two points, i.e. Honor Code Trials.[1] Using a kd-tree, the nearest neighbor is determined like so:

1. Compare the distance between the query point and a guess point to the query point's currently NN's distance determined (first set to infinity).

2. If the new distance is less than, set the guess point as the current NN and update the nearest distance.
3. To check if there are any closer points on the other side of the splitting plane determined by the kd-tree, use the current nearest distance as a radius of the hypersphere and see if any other points lie within it.
    a. If no points are within the hypersphere, do not measure any point on the other side of the splitting plane.
    b. Otherwise, traverse the other side of the tree to find any closer points.[2]

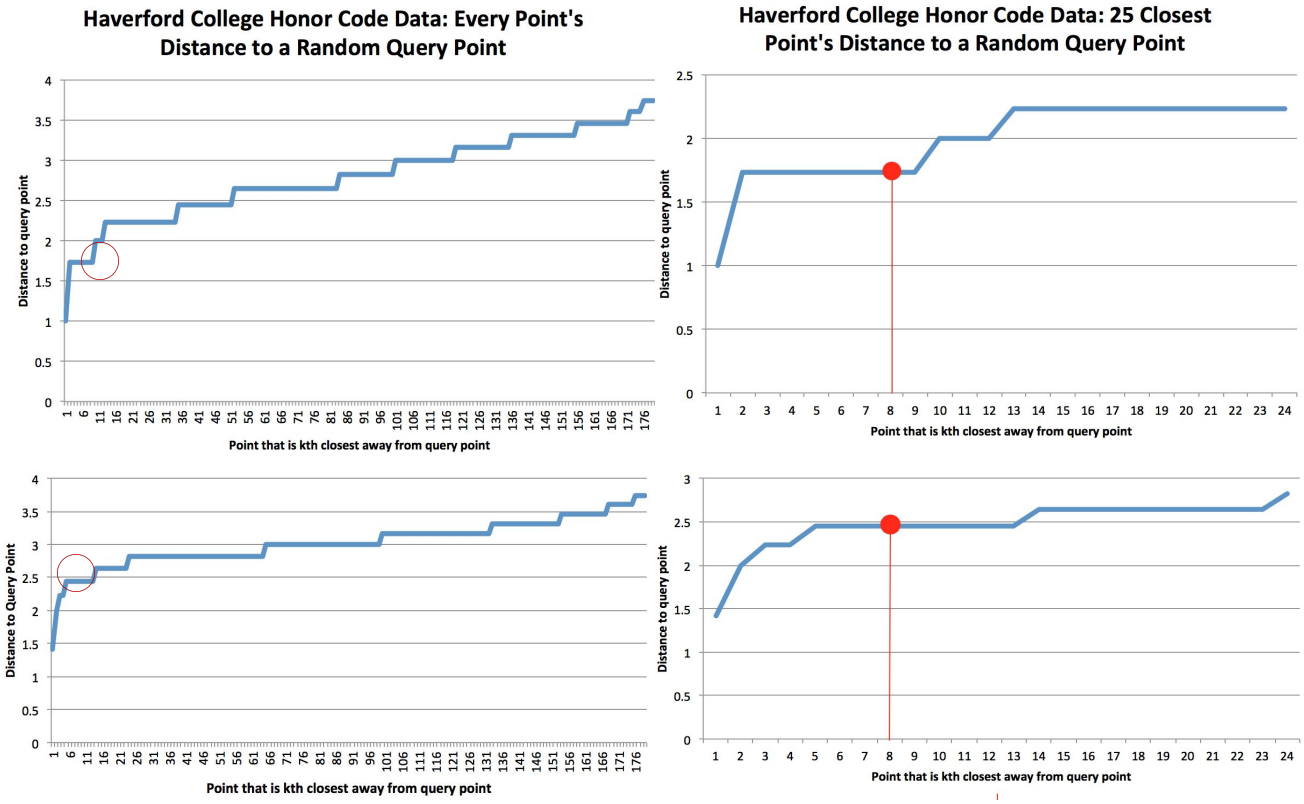
Questions From the Domain Expert
Brian Guggenheimer did not have a specific question in mind for us to answer, but rather hoped that our findings would unravel patterns in the trials. For example, he wondered if any case attributes were surprising indicators of others. Guggenheimer was also interested in using the visualizations on the Haverford College Honor Council website for students to explore Honor Code violation patterns themselves. In this way, students could self-reflect on the tangible role of the Honor Code at the College.

2. NETWORK USING CONFRONTATION TYPE, K-NN, & PAGERANK.

Hypothesis
This network uses the k-nearest neighbors algorithm (k-NN) to connect points, PageRank to size points, and Confrontation Type to color points. I expected to see points of the same Confrontation Type to cluster more closely, for I expected them to contain a greater number of similar case attributes.

Method Description: K-Nearest Neighbors
To create a network from the project data, I first had to determine a) how many points every point should be connected to, and b) how and why points should connect. The strategy I ultimately implemented was k-NN using SciPy, an open source Python library for scientific computing in case my written version of the algorithm did not solve correctly.

When given *k*, the k-NN algorithm finds a point's k-nearest neighbors. I determined the ideal *k* for the data set by choosing a random query point, comparing its distance between every other point in the data set, and finding the *k*th point where the distance value dramatically changed. In theory, the optimal *k* value is the *k*th nearest neighbor whose distance from the query point drastically increases in comparison to previous distances, marking the point on the edge of a cluster. This is why it is of good practice to visualize the results. Below illustrates two of the five total times I sampled points from the data set.



The graphs on the left illustrate the distance between all other points and two random query points. The graphs on the right zoom in on the random query point's twenty-five closest neighbors. k = 8 is at all red circles.

However, there lacked such a finding when any query point was graphed. Each graph had a slow incline of distance values except early on. One reason why may be because the Honor Code data set is comprised entirely of "1"s and "0"s, and since the NN algorithm used Euclidean distance as its distance metric, the calculated distances between points may be different to our expectations of the data's results. For example, both graphs frequently plateau because some points are all equidistant from the query point, though this is unusual for other data sets. For that reason, I looked for a different trend. After discussing with teammates, we agreed that the *k* used should be when the graphs attempted to level off, which is between $k = 3$ and $k = 14$, the first sign of dramatic change in the graph. As a result, I chose $k = 8$, the lower average for *k*, since the more nearest neighbors a given point is connected to, the slower the graph functions and the less distinctive links between points in the visualization become.

Method Description: PageRank

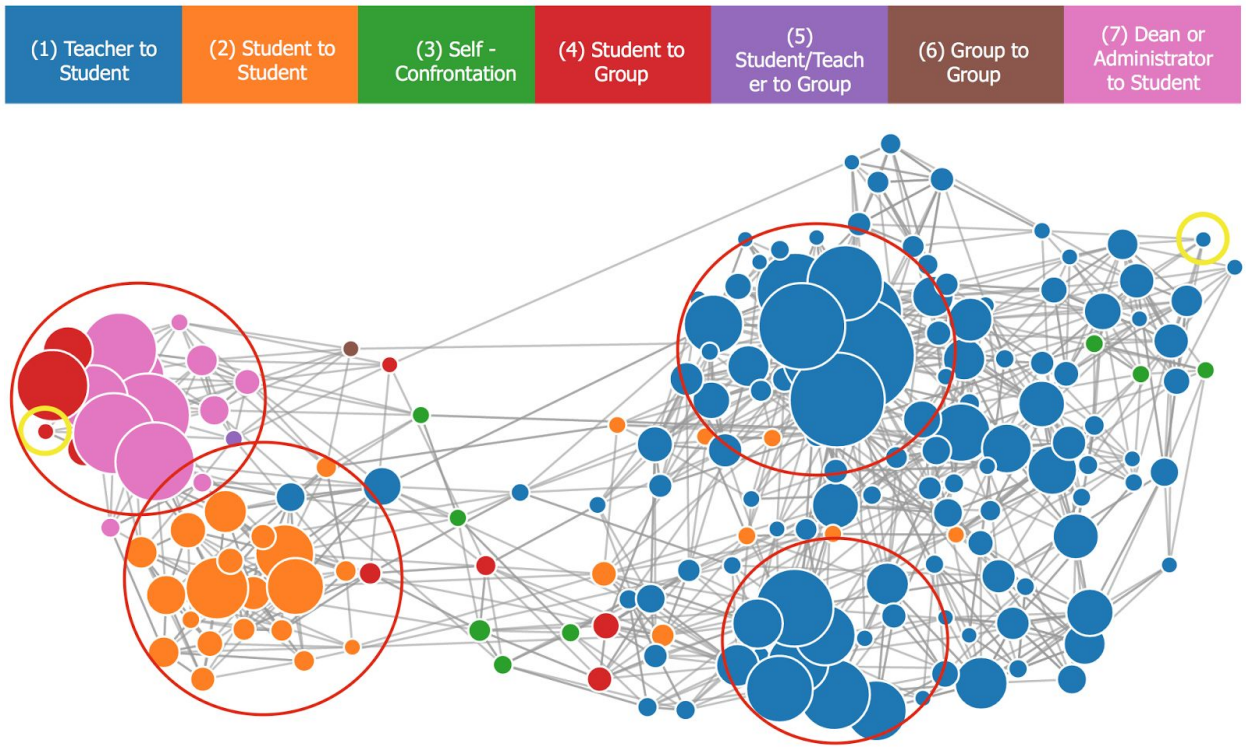
The PageRank algorithm was used to size the graph's points. PageRank was originally based on the likelihood that a random web surfer would land on a given page and is commonly used to determine the "most important point". It is calculated using the following formula.

$$PR(u) = \frac{1-d}{N} + d \sum_{v \in In(u)} \frac{PR(v)}{|Out(v)|}$$

To determine the PageRank of given node (*PR(u)*), for every node *v* that links to *u*, divide its PageRank (*PR(v)*) by the number of links pointing out of *v*. For this data set, the number of links pointing out of a point is equal to the number of times the given point's nearest neighbor is another point's k-nearest neighbor, where $k = 1, 2, \ldots, 8$.

Then, add this sum to $(1 - d) / N$.[3] The damping factor *d* acknowledges the probability that the surfer will continue clicking links and is typically set at .85, and *N* is the total number of points in the collection, and in this case N = 181.[4] PageRank assumes that at first, the random surfer is equally as likely to land on any point. This means that the sum of all point PageRanks should be 1 and that the PageRank of every point is initially 1/181 = .00552486. The algorithm continues until the PageRank stops changing, i.e. the difference between every point's new and previous PageRank is equal to 0. To account for numerical error in the data, the algorithm continues until the difference is less than or equal to .00001.

Results



Visualization of Honor Code network using the k-NN algorithm (k = 8) to link points, PageRank to size points, and Confrontation Type to color points. The circled points/clusters will be discussed later.

Clusters separate nicely through Confrontation Type, even better than hypothesized. The network contains two main clusters (on the right and on the left) and four inner clusters seen through PageRank (circled in red). First, we will look at the right-hand cluster, i.e. the cluster with mostly Teacher to Student Confrontation (1). The bottom-circled cluster with smaller PageRank consistently involved the attributes "Plagiarism" and "Academic Trial Type". The cluster above with larger PageRank points had attributes of "Plagiarism", "Letter", "Statement of Violation", and "Academic Trial Type", suggesting this collection of attributes is more common.

The left-hand cluster did not have as consistent of attributes as the right-hand cluster. Points in the Student to Student Confrontation (2) cluster usually involved "Joint Student-Administration Panel Trial Type" and "Statement of Violation". The mostly Dean or Administrator to Student Confrontation (7) cluster had "Statement of Violation", "Involve Admin/Staff", and "Joint Student-Administration Panel Trial Type". Something to note is that the Student to Group Confrontation (4) points in that cluster also had Confrontation Type (7) as an attribute.

All points in the right-hand cluster not of Confrontation Type (1) contained "Academic Trial Type" and "Statement of Violation", which make sense given that all right-hand cluster points have those features. Additionally, the tightly clustered Self-Confrontation (3) points on the far right share variables "Use of Disallowed Resource", "Academic Trial Type", and "Inappropriate Use of Tool on Exam". Their tightness and location too make sense, as the points are very similar and all relate to academic Honor Code violations. Finally, there is the most Confrontation Type variety between the two clusters, which suggests that outlier Confrontation Type cases have very different features.

Analysis

What is interesting is that the network visualization goes against the narrative students are so commonly told about maintaining the Honor Code by maintaining one's own academic integrity. The Haverford Honor Code describes itself as a "student-run tradition" of "student involvement and self-governance" where "the student voice is valued on the same level as all others."[5] The Code is celebrated for its removal from Haverford College staff and administration, yet what is seen in this visualization is that the majority of trials rely heavily on senior persons at Haverford confronting students. Staffs are not actually removed, but in fact just as responsible for governing the Code as students are.

Future Questions

A quick glance at the visualization suggests two clusters, one of academic and the other of social Honor Code violations. A closer look revealed that most right-hand cluster points were Academic Trial Types and left-hand cluster points had Joint Student-Administration Panels. For that reason, I wanted to explore if Teacher to Student Confrontation (1) trials frequently indicate Academic Trial Type cases, and whether or not Student to Student (2) and Dean or Administrator to Student (7) Confrontations indicate Joint Student-Administration Panel Trial Type cases. This graph would also verify that academic and social Honor Code violations have distinct trial traits.

In addition, I was concerned that this network was biased to cluster through Confrontation Type because k-NN links points based on matching case attributes and Confrontation Type is an attribute. I hoped to see if a new network with Confrontation Type removed from the data set still clustered closely to it. How strongly Confrontation Type correlates to the new graph will point to how well it decides other trial features. Also, if the data set still separates well to Confrontation Type and similar clusters arise from the new visualization, the amount of connectivity between clusters demonstrates the amount of relatedness between academic and social Honor Code cases.

3. CONFRONTATION TYPE AS AN INDICATOR FOR OTHER ATTRIBUTES.

Exploring The Meaning Of A Trial's Confrontation Type

I wanted to explore if Confrontation Type is a deciding factor on other case attributes in order to observe the distinction between academic and social Honor Code trials. I used two different techniques to do so - visualizing case Trial Types with texture, and removing the Confrontation Type in two new networks that use the same k-NN ($k = 8$) to link points.

3.1 Visualizing Trial Types With Textures

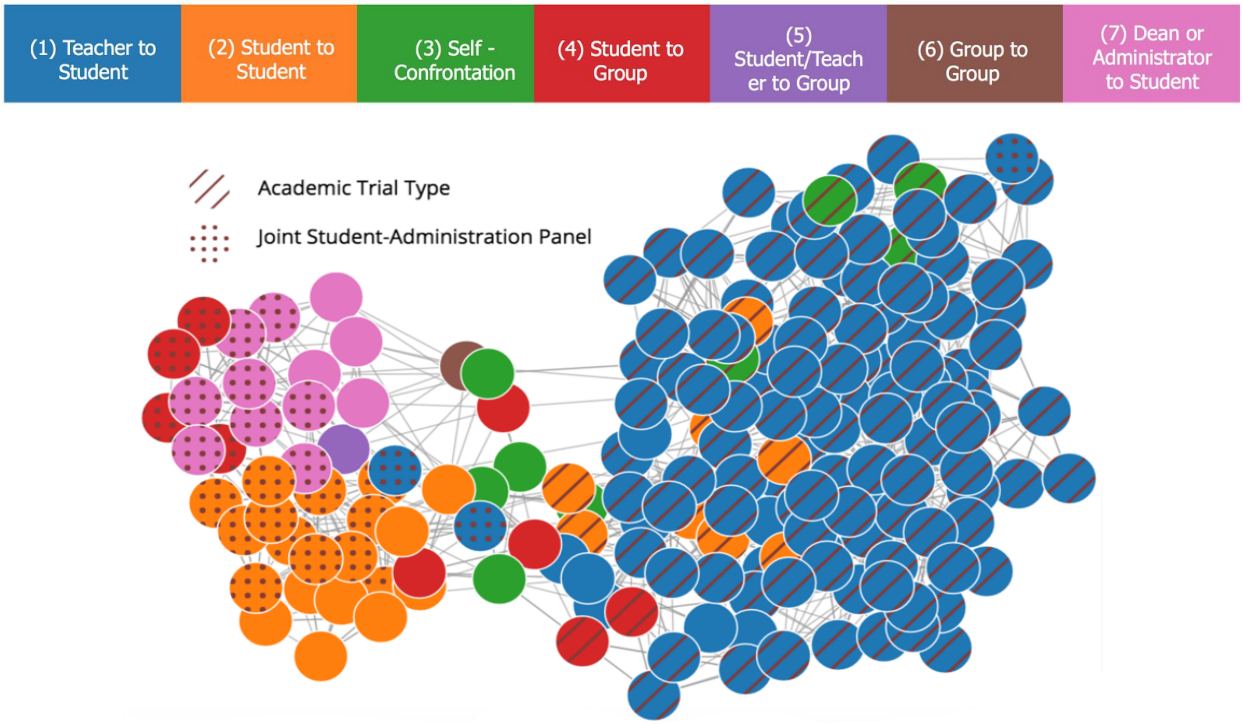
Hypothesis

In the previous network, points of the same cluster seemed to share many attributes, frequently Trial Type. This graph was created to visualize how Trial Type broke down between clusters. I hypothesized that most Confrontation Type (1) trials were Academic Trial Types, and most Confrontation Type (2) and (4) cases were Joint Student-Administration Panel Trial Types.

Method Description

Patterns are useful for the perception of multiple categories in a network. I used an online JavaScript library to texture my points with lines or circles depending on the Trial Type.[6] However, too many textures can clutter visualizations. At a fundamental level, data visualization finds and displays meaningful patterns in a given data set. The pattern that I hoped to find focuses on two Trial Types, so I chose to only include two special textures. This network also applied k-NN to link points but with a fixed point size to better see point color and pattern.

Results



Visualization of Honor Code network using the k-NN algorithm (k = 8) to link points. Point colors are based on Confrontation Types and point textures are based on only two Trial Types. Otherwise, points are not textured.

As expected, the right-hand cluster points almost all contain Academic Trial Types, while the left cluster has Joint Student-Administration Panel Trial Types, albeit with less uniformity. Nearly all Confrontation Type (1) cases involve Academic Trials. Most points in the graph's center do not have either trial of interest associated.

Analysis

Very easily, we see that most cases are either Academic or Joint Student-Administration Panel Trial Types. What is interesting is that by the Honor Council's definition, both Trial Types recognize administrative concerns in relation to student actions, which suggests that these Trial Types consistently involve administration-student collaboration, in contrast to the Code's previously outlined values.[7] Again, this counters the idea that only students enforce the Honor Code. Rather, both students and Haverford administration play active roles in the Code's establishment.

Additionally, the fact that most trials towards the network's center were unassociated with either Trial Type of interest reinforces that their attributes are not common to Honor Code cases and implies that they do not fit the academic or social Honor Code violation binary.

3.2 Confrontation Type Attribute Removal

Hypothesis

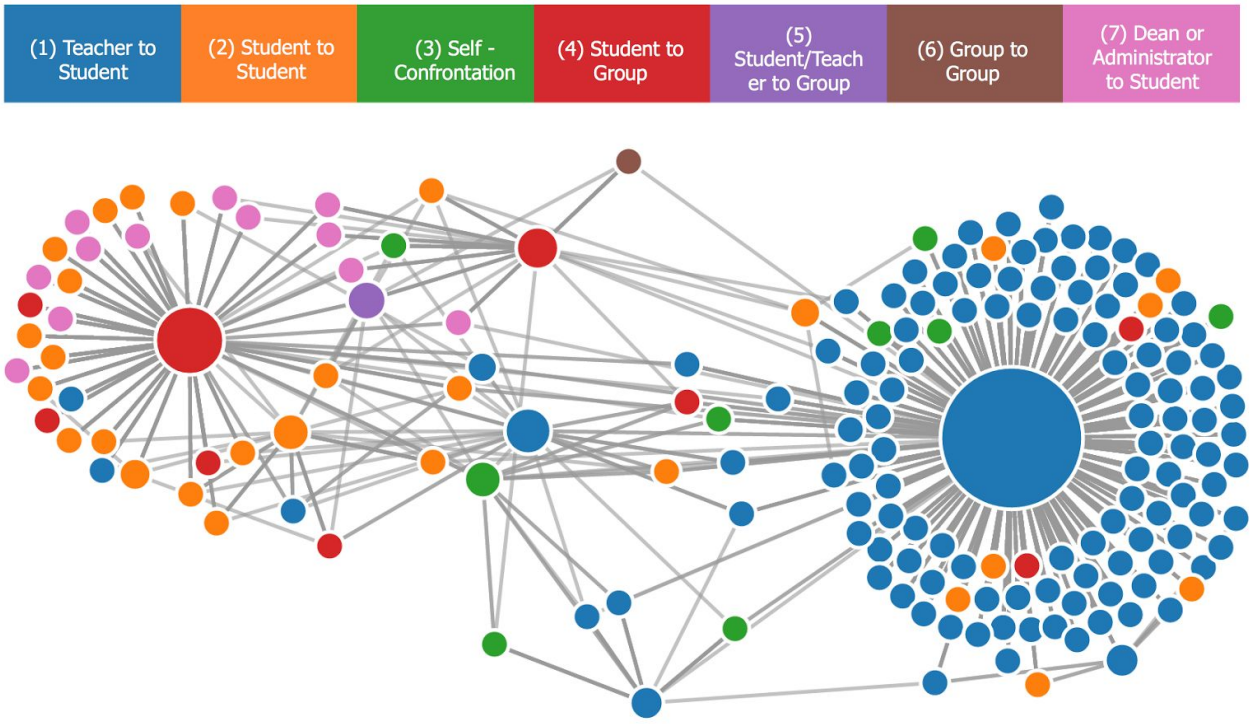
I wanted to see if the data set still clustered by Confrontation Type when the attribute was removed. I hypothesized that in the new network, points of the same Confrontation Type would still cluster more closely together, and that there would be certain clusters of points with higher PageRank like in the original network. I also expected that points originally in the center of the graph would now either link more closely to a cluster's center, since Confrontation Type may have been a dividing factor, or stretch far away from other cases if they were outliers.

Method Description

I used two different techniques to ensure that I properly cleaned the data set of Confrontation Type. In the first, I wrote code in Python to ignore any columns that involved this feature. Based on the surprising network visualization, I tried cleaning out the attribute again, this time creating a new file without any Confrontation Type columns in case I had removed the variable incorrectly in my previous attempt. After coming up with the same visualization, it was determined that the shape of the clustering was actually an artifact of using Euclidean distance as a distance metric for NN (explained more in the *Results* section).

I used the same k-NN algorithm to link points, choosing to keep $k = 8$ to more easily compare the original and new network, and recalculated PageRank to size the points.

Results



Visualization of Honor Code network with "Confrontation Type" attribute removed. It uses the k-NN algorithm (k = 8) to link points, PageRank to size points, and Confrontation Type to color points.

Although the graph's cluster shape is severely different from its original, it still has two clusters, one of mainly Confrontation Type (1) points, and the other with Type (2), (4), and (7) points. Half of the points closer to the middle of the original network are now strictly associated with one of the clusters. Also in the middle are two large Confrontation Type (1) and (4) points that, upon looking closer, share many attributes of the right and left-hand cluster. So, their location and point size reasonably suggests their roles as points that bridge the clusters.

What makes these clusters look so different is because each is centralized around a single point. These points contain the greatest number of attributes compared to those that link to them. In the original network, the center points are within but on the outskirts of the clusters, likely because they had so many features (circled in yellow on the original network). Again, the shape of the clustering is an artifact of using Euclidean distance as a distance metric for NN. Given that the clusters' center points contain more features with value "1" than any nearby points, the center points technically contain the most attributes in common with any given point. As a result, many points link only to the center points, as illustrated in their PageRank/point size.

Finally, when looking at the shape of the visualization, we see that a bigger percentage of points in the right cluster are isolated. That is, they connect to none of their neighbors. The left-hand cluster's shape is similar, but more points connect to their neighbors.

Analysis

The removal of Confrontation Type from the data may determine how well of a deciding factor it is for trial attributes. The fact that many points with varied Confrontation Type remained in the middle of the original and new graph indicates that Confrontation Type is not a clean dividing factor between academic and social Honor Code violations for some cases. However, the overarching trends remain. For example, how often cluster points are interlinked suggests how many common features they have. The right-hand cluster's points (mostly Confrontation Type (1)) only connect to the center point. On the other hand, the left-hand cluster's points are more interconnected, i.e. share more features. So, the graph's structure suggests that there is less procedural and contextual variety in Teacher to Student Confrontation trials. To continue, we see in the textured graph that nearly all Teacher to Student Confrontation cases involved Academic Trials, which even further suggests Confrontation Type (1) may have a regularized protocol.

4. ANALYSIS OF CONFRONTATION TYPE AS AN INDICATOR.

Exploring The Meaning Of A Trial's Confrontation Type

We see in both new networks that a case's attributes rely frequently on the its Confrontation Type. In the textured network, the greatest takeaway is that often, Confrontation Type correlates to Trial Type. In the network with Confrontation Type removed, the greatest takeaway is that points of the same Confrontation Type usually contained common variables. Both networks affirm what was suspected, which is that there is a pattern to attributes depending on a trial's Confrontation Type.

Both new graphs reinforce that the data set breaks down into two clusters, and these clusters are categories: academic or social Honor Code violations. Furthemore, they also only corroborated the idea that there are common contexts and protocols in those clusters. Sadly, the notion that underlying patterns in the trials exist undermines the hope that Honor Code infractions are anomalies of Haverford's long-held tradition of principles.

5. CONCLUSION.

Reflection

Though I am not a Haverford College student, Bryn Mawr College abides by a similar Honor Code. Before Finals Week, every student was sent an email from the Bryn Mawr Honor Board. The first paragraph reads:

> *... As the semester comes to a close, the members of the Honor Board would like to remind you to complete your work with integrity. Wanting to be done is not an excuse to submit work that is not your own.*[8]

It is safe to say that students of the Bi-Co think about the Honor Code mostly during Exam Week, when actively reminded about how it trusts students to take responsibility for their individual academic learning.

Yet, the original PageRank k-NN network shows that the most frequent Confrontation Types in cases include Haverford College administration. What is expected of the Honor Code and what the data shows are two different narratives. The data shows that Haverford administration may play a greater role in administering the Honor Code than expected. This goes against the average student's perception of the Code as a set of social and academic expectations run by, administered by, and agreed upon by students.

To research the role of Confrontation Type as an indicator for trial details, I generated two new networks where points were still linked using k-NN (k = 8). The first network textured points based on the Trial Type they were related to, and the second was generated without the "Confrontation Type" attribute, yet colored according to it. Both networks produced two clusters, one with mainly Confrontation Type (1), and the other with Confrontation Types (2), (4), and (7). The textured network revealed Trial Types as a common attribute to points within a cluster. The network that removed Confrontation Type revealed that when comparing all attributes, points with the same Confrontation Type usually still shared features. This suggests an underlying structure within case protocol and context, which juxtaposes the principles of the Code that say that regardless of either, students are expected to follow the system.

Future Questions

In terms of data analysis, for one, I wish some levels of confidentiality were waived for our team to analyze the data set. Although it is important to respect student privacy, we could have found more interesting patterns in the trial's accused while still keeping student identities classified. Attributes I would have been interested in exploring are "Gender of Accused", "Class Year", "Ethnicity", etc. If we found a pattern within the accused, maybe the College could determine why the pattern existed and how to address the trend in the future.

Second, given more time, implementing a decision tree for Confrontation and Trial Type using Weka (Waikato Environment for Knowledge Analysis), a software of machine learning algorithms, could have even more confirmed these attributes as deciding factors for trials.

Presenting Code violations to students as visualizations may shed light on the reality of the nature of the Honor Code, where most cases actually rely on administrative leads and all cases link to common features. Recently, a friend of mine was on trial for plagiarizing, and she admitted to me that the Honor Code felt intangible compared to the reality of her grades. I agreed that the Code is viewed in an idyllic light rather than a concrete one, as seen in the email's self-disciplinary expectation. Essentially, all of my networks show that other students have made mistakes at Haverford College. Visualizing these mistakes may guide students to see the students put on trial, trial repercussions, and the restorative processes for students as concrete and effective.

References

1. "Euclidean Distance." *Wikipedia*. Wikimedia Foundation, 21 Feb. 2016. Web. 03 May 2016.
2. *Assignment 3: KDTree*. N.p.: Stanford University, 15 May 2014. PDF.
3. Friedler, Sorelle. *PageRank*. N.p.: n.p., 4 Apr. 2016. PPT.
4. Chebolu, Prasad, and Páll Melsted. "PageRank and the random surfer model." *Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 2008.
5. "What Is the Code?" *Haverford College Honor Council*. Haverford College Honor Council, 2016. Web. 08 May 2016. <http://honorcouncil.haverford.edu/the-code/what-is-the-code/>.
6. Stalco, Riccardo. "Textures.js: SVG Patterns for Data Visualization." *Textures*. GitHub, n.d. Web. 03 May 2016. <https://riccardoscalco.github.io/textures/>.
7. "Article VII Section 1. Types of Trials." *Haverford College Honor Council*. Haverford College Honor Council, Spring 2015. Web. 09 May 2016. <http://honorcouncil.haverford.edu/procedures/types-of-trials/>.
8. Honor Board. "Finals Week Reminders from the ~Honor Board~." Message to Bryn Mawr College Undergraduates. 25 Apr. 2016. E-mail.