– Independent Work Report Spring, 2017 –

Mining FERPA Notices for Textual Analysis of Education Privacy Policy

David M. Liu
In Collaboration with Sejin Park
Advisers: Arvind Narayanan and Elana Zeide

Abstract

The growth of technology has threatened the ability of schools to safeguard their student's personal information as such data becomes harder to contain and regulate. Moreover, within school systems, general lack of knowledge regarding the laws has further compounded the issue. Building on existing initiatives to access the current state of student privacy in the United States, The goal of the project is to use computer science tools to collect and analyze FERPA statements. We scraped 217 HTML and PDF based FERPA notices from the Bing search engine. Our repository of notices included a geographically representative sample of the country. Using a paragraph-level SVM text classifier, we extracted the notices from large handbooks, guides, and catalogs, which are common vesicles for FERPA notices. From analyzing the lengths of the notices and calculating similarity scores with the model notice, we conclude that a high percentage of schools publish annual notices similar to the model one. An N-gram based analysis also shows that many schools do not include optional sections such as contact information in their notices. Looking forward, legal analysts can use the data from this project to draw greater policy claims regarding the current state of student-data privacy.

1. Introduction

1.1. Motivation and Goals

The Family Education Rights and Privacy Act (FERPA) was created in 1974 to protect the privacy rights of students. In the decades since the federal law's inception, the student data privacy landscape has transformed with technology. Today, with the rise of the Internet and third-party education technology vendors, student data are collected and distributed through many more media than the creators could have imagined. As the policies and legalities of education privacy play catch up to the pace of technology, schools have sunk into a disarray of occlusion when it comes to student data management [11].

FERPA, however, does require schools to publish an "Annual Notification of Student Rights" document on a yearly basis. These so called FERPA notices are the primary mechanisms by which schools communicate their policies to students and families. With a lack of cohesion and standardization in the student privacy section, it becomes increasingly important for schools to disclose their privacy policies. Yet, on a macro level, there is little understanding among legal scholars regarding the contents of these FERPA notices [11]. Many in the field of education privacy are keen to discover whether schools are consistent in their definitions of major terms and how they handle sensitive topics such as immigration, among others.

The purpose of this technical study was to amass a holistic collection of FERPA notices from across the country. By compiling a survey of FERPA notices, we provide the resources for legal experts to gage the state of education privacy in America.

1.2. Summary of Results

By scraping a popular search engine in Bing, we were able to collect and clean over 200 FERPA notices from universities across the country. A major accomplishment of this project involved extracting the FERPA notices from larger guides, documents, and catalogs, which are the origins of many FERPA notices. To isolate the FERPA notices, we devised an SVM-based paragraph classifier

which we trained on a set of example FERPA and non-FERPA paragraphs. We represented all of the documents as vectors of term frequencies. PCA analysis showed that the data was separable by label. The classifier passed our testing set with high accuracy (89%). Our analysis of the FERPA notices shows that those collected closely resemble the model FERPA from the Department of Education. We also analyzed the prevalence of specific keywords in the dataset and represented the corpus in a dendrogram to visualize the similarities among the notices. Using our technical analysis and compiled data set, we hope for experts in the field of education privacy to extend our conclusions.

2. Problem Background and Related Work

2.1. FERPA and Education Policy Background

The field of education privacy has struggled to keep pace with the rate of technological growth. Many of the core policies centered around protecting student data, including FERPA, were enacted before the modern age of the Internet. As a result, schools have diverged into a disarray of various student-privacy practices, with very little checks for transparency and accountability.

The policies specified in FERPA are no exception to this trend. Under the 1974 law, each school that receives federal funding, including public k-12 schools and universities alike, is required to publish an "Annual Notification of Rights". Because FERPA leaves many of the decisions relating to student privacy practices to the schools themselves, there is little consistency to be expected among the notices [11].

In reality, there is little comprehensive knowledge of the contents of these notices. Like student-privacy itself, there is no existing consensus. The United States Department of Education does provide a model FERPA notice for schools to follow. However, prior studies have not examined to what degree schools follow the model FERPA notice or examined the changes that schools make to the example. Understanding the language and contents of these notices will be very useful for education-privacy legal scholars in the modern age.

2.2. Information Mining of Policy Documents

The primary contribution of this project is a repository of cleaned FERPA notices; yet the project fits into a larger initiative of information retrieval and document collection in the digital age. Previous studies, such as an attempt to scrape the ACM digital library [1], have also undergone large-scale querying and subsequent scraping. When document collection is applied to the legal domain, many of the same principles of information retrieval are still relevant and practical. For example, when scraping the ACM library, Bergmark converted the collected PDF documents into a more readable and processable format, a technique used in this study as well [1]. Other studies aimed at scraping legal documents have place greater focus on providing multilingual results, as in [5] and [8]. Moreover, to the best of our knowledge, this study is the first effort to scrape and collect a representative sample of FERPA notices.

The methodology used in this project relates closely to that used in standard coverage testing of search engines. Dasdan et al. framed the coverage testing problem as such: given a sample document, such as a model FERPA notice, query a search engine for copies of the sample or near duplicates and evaluate the similarities [4]. The procedure for coverage testing begins with generating a query signature from the original document, with the goal of extracting representative terms from the source document. Dasdan et al. then queried the search engine with the signature and scraped the results, similar to our approach. The results were then compared with the source document and evaluated by similarity metrics.

Previous studies have measured the coverage extent of Bing, our search engine of choice. Lewandowski tested Bing's retrieval effectiveness by querying the search engine for 1,000 informational searches [7]. Each search had a list of correct, or expected results. These searches were termed informational because they sought specific documents and topics. The results show that Bing performed with 79.3% accuracy. The results from Lewandowski study place the extent of our data collection in context and help us to understand the limits of information retrieval on search engines.

2.3. Automated Analysis of Policy

Prior work in the field of privacy policy technical analysis has centered around summarization and categorization, with the goal of consolidating long, verbose policy corpra into digestible segments. Back in 2006, IBM researcher Carolyn Brodie, et. al. created a document processing tool specifically designed to parse privacy policy documents [2]. The goal of IBM's research was to simplify overly complex and verbose organizational privacy statements, a common trend in the field of policy. Additionally, machine learning has even been applied to privacy documents. Research conducted at DePaul University by Tomuro, Lytinen, and Hornsburg successfully implemented classification on privacy-policy text [9]. The group created a tool to classify each sentence of an input policy into any of five major categories of sentences. At the same time, the categorization study reminds us that text classificatioin techniques are stil in need of improvement; in the case of Tomuro, the group was only able to successfully classify three in every four sentences. Nevertheless, the research from Tomuro and Brodie confirms the necessity of parsing privacy policies through computational techniques.

Similar to Tomuro, this project also developed a text classification system. The goal of the classifier in this project was to identify FERPA paragraphs from non-FERPA paragarphs. As with Tomuro, classification implied a need to vectorize the textual data into a numerical form. We based our vectorization technique on a study conducted by Conrad et. al, which also applied language processing tools to legal data. The components in Conrad's vector were term frequency counts. More formally, we define the vector representation of a paragraph $< V >$ as:

$$< V >=< v_1, ..., v_n > | v_i = count(w_i)$$

Where w_i is the i^{th} word in the vocabulary and $count(w_i)$ is the frequency of the word in the paragraph. While Tomuro classified sentences, for segmenting documents, prior studies have shown that paragraphs are appropriate units of text for classification [10]. In terms of performing the classification itself, we utilized Scikit-Learn's SVM classifier [1]. The basis of Support Vector

[1] http://scikit-learn.org/stable/modules/svm.html

Machines (SVM) is to identify geometric boundaries that best divide the data. SVM-based classifiers are used in contrast to Naive Bayes classifiers.

Building on vectorization, many prior studies have used the numerical representations to perform clustering analysis on the entire corpus of text. Iwayama and Tokunaga introduced the technique of probabilistic clustering of documents and visualization via dendogram [6], which we have extended to our project. Dendogram analysis provides an intuitive method for capturing the macro clustering in a textual dataset, such as FERPA notices.

3. Approach

3.1. Defining the Scope of Data Collection

Early in the project, we decided to restrict ourselves to collecting FERPA notices just for colleges and universities. In an unrealistic, ideal world, we would collect FERPA notices for all educational institutions, regardless of education level. However, to evaluate the extent of our data collection, we need to clearly delineate the scope of notices sought. We began the project collecting FERPA notices for K-12 schools, but quickly realized the dataset was not well-defined. In certain states, there was a FERPA notice per school district, while in other states FERPA notices were organized by locality or county. Furthermore, even our preliminary searches revealed that colleges and universities were more diligent in maintaining up-to-date FERPA notices.

Collecting FERPA notices for colleges and universities also afforded a few practical advantages, which will be elaborated upon in the implementation section, below. First, every site scraped had a ".edu" domain, decreasing the likelihood of false positives in our search results. For the K-12 schools, we came across a much wide range of sites, many of which did not contain FERPA notices. Secondly, the Department of Education maintains two versions of FERPA's, one for K-12 and a second, slightly different version for post-secondary institutions. Focusing on colleges and universities allowed us to simply analyze one of the two versions. This design choice becomes extremely relevant when we compare FERPA notices later in the project.

3.2. Data Collection Pipeline

Similar to most information retrieval pipelines, our project followed a three step process of raw data collection, followed by cleaning and then analysis, as illustrated in Figure 1 below. As soon as we began data collection, we realized that online FERPA notices are generally stored on a webpage (HTML) or PDF document. In terms of information retrieval, we created two separate data collection

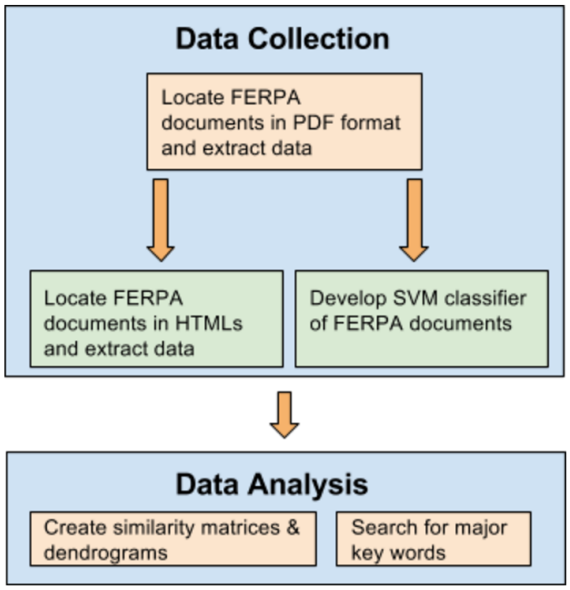

Figure 1: We began with a set of PDF-based FERPA notices. This project focused on cleaning the original dataset while a parallel project scraped HTML-based projects. We then combined the two datasets into a single repository and performed a joint analysis.

pipelines, one for the HTML FERPA notices and a second for the PDF notices. This division was necessary because each format required its own scraping and cleaning specifications. We were able to execute both pipelines in a parallel, and of course, at the conclusion of data collection, we combined notices from both formats into a single repository. This project focused primarily on collecting PDF FERPA notices - the details of which are introduced below and continued in the Implementation section.

3.2.1. PDF scraping Compared to HTML-based documents, PDFs offered several advantages in terms of data collection. First, the entirety of the document is stored in a single file and can be easily downloaded onto a local machine. Additionally, PDFs of legal policies generally contain fewer

miscellaneous images and formatting irregularities, often featuring the raw text itself. However, we encountered a major challenge in converting the PDFs to a processable text. While well-tested PDF converters are readily available, we cannot reasonably expect this conversion process to be fully accurate. At the same time, a holistic retrieval of FERPA notices cannot simply disregard PDF documents as many schools prefer to publish their legal documents PDF format. As a result, we had to account for the possibility of conversion error in our analysis. Many of our later project design choices involving similarity measurements were based on the fact that conversion, though necessary, introduced formatting miscellaneous characters. Figure 2 shows an example of minor newline characters added into the text.

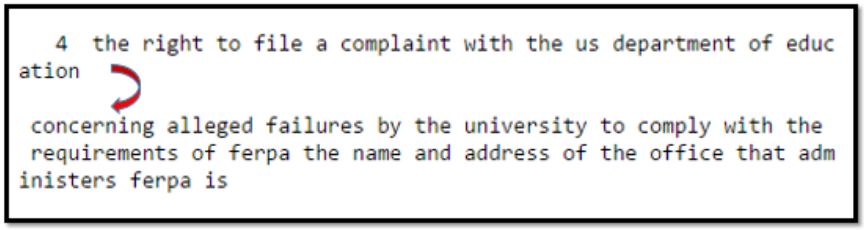

Figure 2: An example of how additional newline characters maybe inserted into the text during conversion. These formatting artifacts do not affect the semantic meaning of the text but can affect similarity metrics.

To mitigate the potential biases of conversion errors, we represented each notice as a vector, whose components were word frequencies. By recording the frequencies of specific words, we ignore the presence of any miscellaneous formatting anomalies. Additionally, by representing the documents as vectors, we open the possibility for greater analysis. As previous studies have done, we can use the vector representation of documents to analyze the similarities among notice as well as study the hierarchical clustering of the text [3].

3.2.2. Text Extraction Early in the PDF scraping phase, we observed that many of the downloaded documents were not isolated FERPA notices but rather larger college catalogs and handbooks that contained FERPA notices. Figure 3 shows that while many of the notices were under 10,000 characters in length, the rough expected length of a notice, others were well over the limit. Because FERPA documents are generally sections of larger documents and handbooks, we needed to isolate

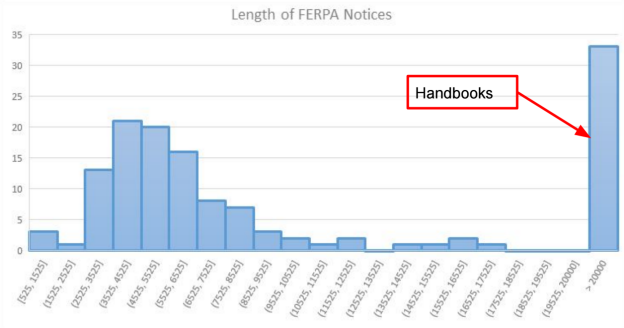

Figure 3: In terms of length, the FERPA notices fell into two major groups: a "short" group that contained isolated FERPA notices and a "long" group that consisted of multi-section handbooks, guides, and catalogs.

just the paragraphs and sections of interest. If we do not do so, our later analysis and conclusions may not be specific to FERPA notices. At the same time, there are no clear guidelines for what constitutes a FERPA document as opposed to a more general policy statement on education privacy and student rights. Our approach for tackling this general problem was to utilize machine-learning based techniques to identify the sections of interest inside these larger handbooks. At a high level, we provide examples of FERPA text segments to the classifier as well as miscellaneous paragraphs and sentences that are similar in style to the FERPA notices yet semantically different. The advantage of the machine learning approach is that we do not need to specify explicit guidelines regarding FERPA notices, we can simply rely on the algorithm to solve for the differences between FERPA and Non-FERPA paragraphs.

3.3. Objective for Analysis

Having cleaned and consolidated the notices, there were many possible directions for analysis; in this project, we focused, for the most part, on conducting technical analysis, saving the legal analysis for the policy experts. In terms of technical analysis, our approach was to utilize much of the framework we had developed in the data cleaning portion of the project. Specifically, the technique of feature extraction and vectorization lent itself to document similarity and clustering analysis. The technical findings, which will be discussed more extensively in the results section, are designed to supplement the legal conclusions.

4. Implementation

4.1. Information Mining

To collect the FERPA notices, we scraped the documents from Bing. In crafting the Bing search query, we sought to maximize the number of FERPA notices returned while also limiting false positives. Following a series of trial and error experiments, we required the exact starting sentence from the model FERPA notice ("The Family Educational Rights and Privacy Act (FERPA) affords students certain rights with respect to their education records") and specified the filetype to be PDF. By supplying a specific sentence, we reduced the possibility of rogue documents arising in the search results. At the same time, the sentence is not so niche that it would filter out legitimate notices. To scrape the notices from Bing, we utilized the BeautifulSoup[2] scraping package in Python. A study of the Bing source DOM revealed that all search result links are stored in a div of class "b_title". Thereafter, we made HTTP GET request calls to each of the PDF links in the search output. To improve accuracy, we downloaded the notices in discrete chunks.We repeated the above procedure for each page in the Bing results.

Using this method, we collected 136 PDF files. The original Bing search had returned 245 results, however, we only collected one notice per school, which eliminated duplicates.To transform the downloaded PDF dataset into a analyzable form, we passed each raw document through a PDF to text converter (PDF2Text[3]). We then stored all of the text outputs in a single JSON object for further cleaning and analysis.

4.2. Data Processing and Cleaning for PDFs

The goal of the cleaning phase was to trim off the portions of the document surrounding the FERPA notice. We split the task into two parts: trimming the text before the start of the FERPA and then trimming the portion following the end of the FERPA notice. The former proved to be trivial. Because we required all of the notices in the dataset to begin with a specified sentence - the first

[2]https://www.crummy.com/software/BeautifulSoup/
[3]https://pypi.python.org/pypi/pdf2text/

sentence of the model FERPA notice - we simply truncated all of the text prior to the specified sentence.

However, identifying the end of the FERPA notice was a much more general and complex problem. It is important to remember that the Department of Education does not provide specific guidelines for the length of FERPA notices or how they conclude. Furthermore, a manual examination of a few notices confirmed that FERPA notices vary in their conclusions. We visualized the length of the collected documents to get a better sense of where the end of the notice may be. As stated previously, there were two categories of FERPA notices. The first group contains notices that are under 10,000 characters in length. Because the model FERPA itself is about this length, we can reasonably claim that these documents consist solely of a FERPA notice as there simply aren't enough characters for other policies. On the other hand, the graph does also show a group of notices who length is well over the 10,000 character mark. In fact, there were very few notices that were in between these two extremes. The longer documents were interpreted as large handbooks, guides, and booklets. While it was not necessary to locate a specific ending point for the short notices, the longer documents necessitated a more thorough solution.

4.3. Paragraph Exraction

In order to extract the FERPA-portions of the large handbooks we scraped, we broke each document down into separate paragraphs. We defined a paragraph break as the longest sequence of line breaks found in each document, which varies from document to document. The goal of the machine learning portion of the project is to distinguish between paragraphs which are part of FERPA-documents and those which are not. Below, we will discuss the methodology employed to train and test the SVM-based paragraph classifier.

4.3.1. Feature extraction To vectorize our text, we represented each paragraph as a term-frequency vector. The terms we chose were all of the words that are used in the model post-secondary FERPA that the Department of Education publishes. For the purposes of this study, the model FERPA serves as our definitive example of a true FERPA document. To improve our feature selection, we only

included words that were longer than three characters and enabled stemming to avoid duplication of similar features. We used the common Lancaster stemming method [4]. Our final feature list contained 163 such stems.

For each paragraph in the training and testing set, we computed the paragraph's vector representation by running a substring search for each stem in our feature list. Because we were running a substring search over the entire paragraph instead of iterating through individual words, it was theoretically possible for there to be false positive frequency counts. However, because the stems were at least four characters in length, the false positive probability was negligible. We also normalized each vector to be of length 30. At first we normalized the vectors to unit length, however, the smaller magnitude proved to hinder our later clustering efforts, while the larger magnitude increased the separation between FERPA and non-FERPA paragraphs.

4.3.2. Training Data To train the SVM classifier, we needed a large training set of labeled paragraphs. Because our vectors had dimensionality greater than 100, we sought training samples on the order of a thousand. Manually reading and labeling individual paragraphs was not feasible, so we leveraged our knowledge of the dataset itself. To assemble the training set, we divided the "short" group into two, and used half of the notices to serve as examples of true FERPA paragraphs, reserving the later half for testing. Of course, we also added the paragraphs from the model FERPA for a total of 661 example FERPA paragraphs. To help the reader better understand the difference between the two paragraphs, an example from each category is shown in Figure 4.

Next, we enumerated a list of non-FERPA paragraphs. To increase the accuracy and sensitivity of the classifier, we sought examples that were definitely not FERPA paragraphs themselves but similar in nature. With some manual manipulation, we converted a few of the "long" notices into examples of non-FERPA paragarphs. We, specifically, deleted the FERPA notice from the handbook for the White Mountain Community College, leaving us with 474 non-FERPA paragraphs. FERPA and non-FERPA paragraphs together, we trained the SVM classifier on 1135 labeled paragraph vectors.

[4] http://www.nltk.org/api/nltk.stem.html

Non-FERPA Paragraph

```
begin the grade appeal process with the office of Academic Affa
irs.  Every attempt will be made to have
the faculty member contact and meet with the student within the
specified time.  On occasion, however,
these times may need to be adjusted.  For more detailed informa
tion go to the Academic Policies
section in the Catalog.
```

FERPA Paragraph

```
 education records may be released in person or in writing to an i
nquirer and only with the written and signed consent of the studen
t except when ferpa authorizes disclosure without consent as indic
ated below
```

Figure 4: We include an example of both a FERPA paragraph as well as a non-FERPA one. It is still important that the non-FERPA paragarph could be potentially confused for a FERPA paragraph as the topic matter is similar.

4.3.3. Testing and Validation To measure the accuracy of our SVM Classifier, we tested the classifier on the remaining labeled data. The testing set consisted of the "short" FERPA notices that were not used for training as well as on the "long" FERPA notices, with the FERPA portion removed. If the classifier performed perfectly, it would classify all of the "short"-notice paragraphs as FERAPA and all of the "long"-notice paragraphs as non-FERPA.

We also performed a PCA analysis of the classifier to better understand the inner workings. Our goal in performing the PCA analysis was to project all of the test vectors onto a human-readable coordinate system. To justify our use of the SVM classifier, we expect the points for the FERPA and non-FERPA paragraphs to form two distinct clusters. Only then can the SVM classifier effectively model the separation between the groups. To perform the PCA, we used Scikit Learn's Decomposition package, and projected each testing vector onto the three first principle components. In our final visualization, we distinguished between the FERPA and non-FERPA paragraphs to clearly show the separation.

4.4. Combining Datasets and Analysis

Having developed the classifier we extracted the FERPA notice from the long handbooks and guides. We extracted the notices using the following algorithm:

- For each "long" handbook or guide:
  1. Step 1: Using the classifier, predict whether each paragraph is a FERPA paragraph or a non-FERPA paragraph.
  2. Step 2: Starting at the beginning of the document, identify the first sequence of two consecutive non-FERPA paragraphs.

Because we had already trimmed all of the text prior to the notice, we knew the FERPA notice resided at the beginning of the remaining document. To mitigate potential errors from the classifier, we mark the end of the FERPA notice not as the first occurrence of a non-FERPA paragraph but when two have been encountered consecutively.

Once all of the PDF-based FERPA notices were cleaned, we joined in the HTML notices. In the merging process, we permitted each school to contribute a single notice. In the event of a collision, we retained the PDF notice. Thereafter, we proceeded to perform language-processing based analysis on the dataset. To reiterate, we focused on the technical aspect of language analysis, leaving the majority of the policy analysis for legal experts. The procedure for language analysis was as follows:

4.4.1. Word Frequencies and Length Our unigram and length analysis of the notices consisted simply of iterating through the dictionary of notices and tallying frequency counts. We then visualized all of the results using Python's matplotlib package [5].

4.4.2. Similarity and Dendrogram To calculate the similarity scores between individual notices in the repository and the model notice, we used a vectorization technique extremely similar to the one used for classification. Namely, the vectors consisted of word frequency counts. The notable differences were first, we created vectors for entire documents, not just paragraphs and second, we

[5] https://matplotlib.org/

analyzed all of the notices, both HTML and PDF. The similarity score was as follows:

$$s_{ij} = v_i \bullet v_j, ||v_i|| = ||v_j|| = 1$$

A score of 1.0 would suggest the two documents are very similar while a score of zero would suggest the opposite.

For the dendrogram, we calculated the similarity among notices and used the edit distance metric as opposed to cosine similarity. The former is advantageous because it captures the ordering and sequence of words in the text; when comparing with the model notice, cosine similarity was useful to mitigate the potential formatting artifacts present in our notices and not present in the model notice. Because we did not include the model FERPA notice in the dendrogram, all of the notices had passed through the converter. A closer justification of this decision is included in the Results section.

To generate the dendrogram, we created a list of similarity vectors v_i. Each component v_{ij} in vector v_i represents the edit distance between the i^{th} and j^{th} notices. We then passed the list of similarity vectors into a Scikit-Learn's dendrogram generator [6].

5. Results and Evaluation

5.1. Data Collection

With the goal of collecting a holistic repository of FERPA notices, we were able to amass 217 notices in total using our scraping methods. Of these 217 notices, we collected 115 PDF notices and 127 HTML notices with 23 schools represented in both sets (for these institutions, we used their PDF notices). From these numbers, we can conclude that mining from multiple formats doubled the size of our dataset.

The question now arises: Were we successful in collecting a holistic repository of FERPA notices? There are numerous indicators that we were indeed successful. First, we based our approach on

[6] https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.cluster.hierarchy.dendrogram.html

scraping search engines, and thus can be confident that for the search parameters we specified, the results were exhaustive. It would obviously be possible for us to broaden the search parameters to allow for more notices, but these may also decrease the accuracy of the search and introduce many false positives. Instead of scraping search engines, we briefly considered writing a web crawler that would actively seek FERPA notices, however, it is doubtful that our web crawler would be able to compete again a major search engine. Secondly, a geographic visualization of our data suggests that our 217 FERPA notices samples the entire nation. In Figure 5 below, the location of each school scraped is plotted on a heat map, with warmer colors suggesting greater density. The distribution shown below matches the distribution of colleges and universities across the country, so geographically we collected a representative sample. Lastly, our dataset represents a wide range of

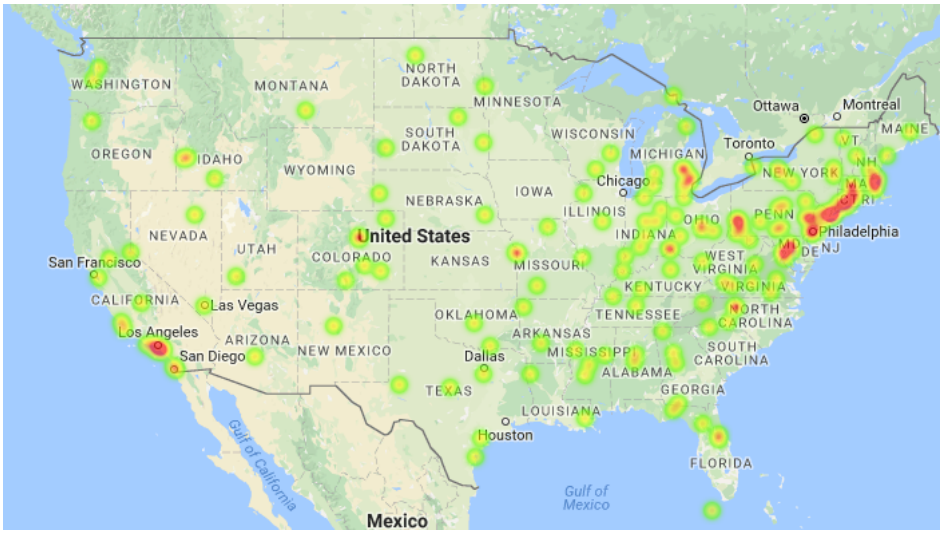

Figure 5: The geographic distribution of schools included in the repository mirrors the overall distribution of colleges and universities in the country, suggesting a representative set.

post-secondary institutions, including large public schools such as the University of Connecticut as well as little known schools such as Minot State College in Montana. Having a representative dataset will be integral in justifying the claims made in the conclusion section.

On the other hand, it would be misleading to state that our dataset is comprehensive by any means. According to the Nationals Center for Education Statistics[7], there were approximately 4,700 degree-granting post-secondary institutions in 2014. Under this statistic, our data set represents just

[7]https://nces.ed.gov/fastfacts/display.asp?id=84

under 5% of post-secondary institutions in the nation. From a technical standpoint, however, our modest dataset remains respectable. It is important to remember that FERPA notices are a priori disorganized and poorly coordinated in nature. When scripting a generic algorithm to maximize data collection, it is to be expected that only a subset of the data available will be collected. Thus, when evaluating our dataset and its fulfillment of project goals, we claim that the dataset is not comprehensive but is is representative and apt for analysis.

5.2. FERPA Classifier

Following up on the implementation of the SVM paragraph classifier, we will now discuss its accuracy when applied to the testing data. The classifier performed with an overall accuracy of 89.3% on a total of 1364 testing paragraphs. Of the 1364 paragraphs, 738 were FERPA paragraphs while the remaining were rogue examples. The accuracy of the classifier increased by over 10% when we introduced stemming into the feature extraction. Furthermore, normalizing the vectors also increased the accuracy. Both of these techniques increased the vectorization precision, placing greater weight in stems with greater frequency proportion. Additionally, the PCA analysis supported the legitimacy of the classifier. In Figure 6, below, we can see that the non-FERPA paragraphs (blue) and FERPA paragraphs (red) were largely geometrically separable.

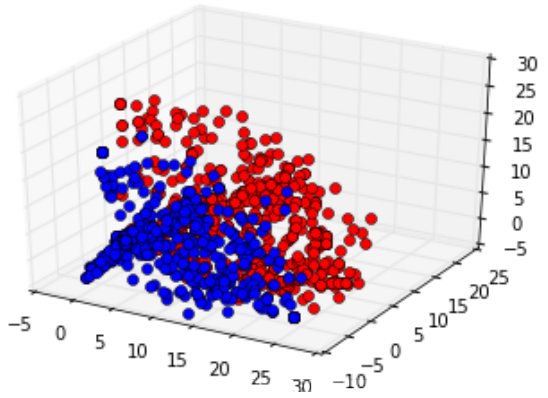

Figure 6: When the testing paragraphs were projected onto the three principle components, the non-FERPA paragraphs (blue) and FERPA paragraphs (red) demonstrated separation, justifying the use of classification.

In addition to directly measuring the accuracy of the classifier through testing data and PCA

analysis, supplementary observations also confirmed the effectiveness of the classifier. For example, looking at the histogram of notice lengths following extraction shows that many of the extracted notices had lengths that were similar to the lengths of notices that did not require extraction. It is important to remember that when we used the classifier to extract the notices from the long handbooks and guides, we did not specify a length requirement, relying instead on the classifier's decisions on a paragraph-by-paragraph level. The fact that the output notices have lengths similar to true FERPA notices suggests the classifier was accurate in making predictions.

5.3. FERPA Trends and Differences

To develop a better understanding of the notice repository as a whole, we studied the lengths of the FERPA notices. In Figure 7 below, we plot a histogram of notice lengths across our entire dataset of both HTML and PDF notices. The results agree with our expectations for the length of

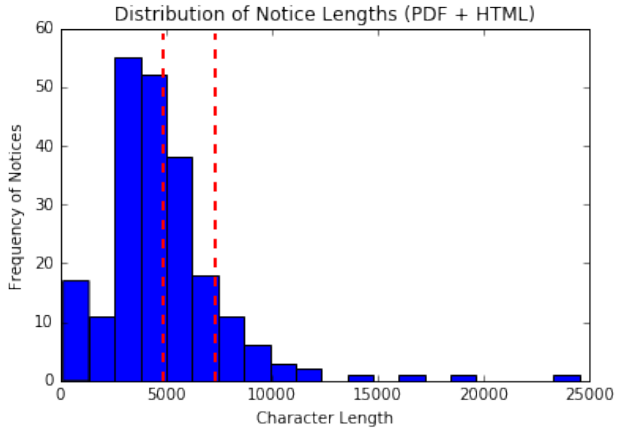

Figure 7: The histogram of FERPA Notice lengths shows that the mean length (left line) was close to the length of the model notice (right line).

the notices. First, the average length is just above 5000 characters, which is similar to the length of the model FERPA notice: 7300 words. The gap between the two metrics may even be closer because the model FERPA contains a few optional clauses and supplementary points which are not required. Outside of the average, the histogram does show a few outliers on both ends of the data. The fact that there are notices under 1000 characters and over 25,000 character in length attests to the variance of FERPA notices. Taking a look at the short outliers, we gain deeper insight

into the content of FERPA notices. For example, institutions such as Clarkamas Commmunity College document their FERPA policy in multiple locations, with one source referencing the other, explaining the terse nature of our scraped notice. On the other hand, institutions such as Missouri State University, which possessed the longest FERPA notice in our dataset, maintain comprehensive webpages that document their FERPA practices. The longer length corresponds with greater detail in this instance.

Along the lines of notice length, we also studied the frequencies of several key words. The frequencies of FERPA keywords are recorded below:

Keyword	Percentage of Notices
Directory	71.89%
School Official	76.5%
Legitimate School Interest	79.7%
Amend	90.3%
Challenge	7.8%
Opt Out	7.37%
Email Address (@)	4.6%

Table 1: Inclusion of Key Terms in Collected Notices

The first three entries in the table are standard topics mentioned in FERPA notices and so we expect high proportion of occurrences. In fact, the presence of these terms, in part, measures the quality of the FERPA notice as one of the key purposes of a FERPA notice is defining significant yet vague legal terms. The notices in our dataset, with high probability (>70%) discuss major terms that legal experts look for in FERPA notices.

The later terms in the table are more optional yet their inclusion can have significant policy implications, For instance, it is important to note whether a school specifies the procedures for opting out of the institution's information sharing practices. In the collected data set,only one in 15 schools even mention an opt out section.

One of the major takewawys from the analysis of length and keywords was the strong similarities between the notices in our repository and the model FERPA notice. To investigate the similarity

more precisely, we calculated similarity scores between the collected notices and the model notice. A histogram of these scores is shown in Figure 8. The vast majority of the notices had cosine similarity scores just under 0.9, confirming our hypothesis that the collected notices are highly similar to the model one. As a note, we made a concious design choice to use cosine similarity over

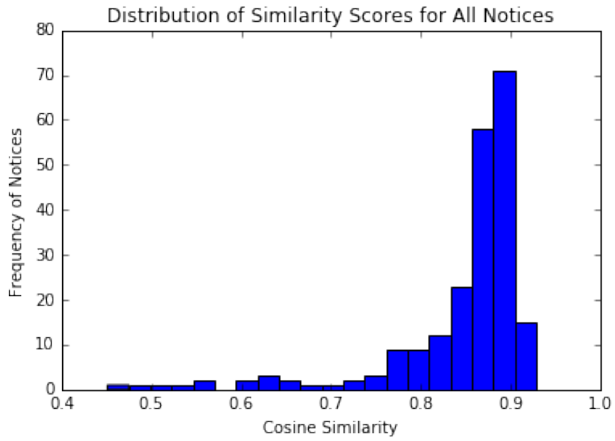

Figure 8: A historgram of cosine similarities shows that the vast majority of all notices were very similar (> 0.8) to the model notice.

other similarity metrics such as edit distance, which we used later for the dendrogram analysis. In the case of comparing with the model FERPA notice, the conversion artifacts from PDF conversion could bias the reliability of the edit distance metric. For example, the PDF to text converter often introduced many newline characters and spaces. These conversion artifacts would all inflate the edit distance, while the cosine similarity is more resistant to these conversion artifacts.

To understand the similarities between the notices from a different angle, we generated dendrogram visualizations. Dendograms illustrate the hierarchical clusterings of data point, which were entire notices in our visualization. Figure 9 shows a truncated version of the dendrogram for the PDF notices. We included PDF notices only primarily for practical reasons. First, including all 217 notices would render the visualization nearly illegible. Secondly, the similarities of PDF and HTML notices cannot be easily compared because the two were generated through different processes. In particular, the PDF notices all passed through a single converter which may have introduced systematic formatting artifacts that could have biased the similarity calculations, as described earlier. From the dendrogram, it is clear that there was little hierarchical structure among the notices, as

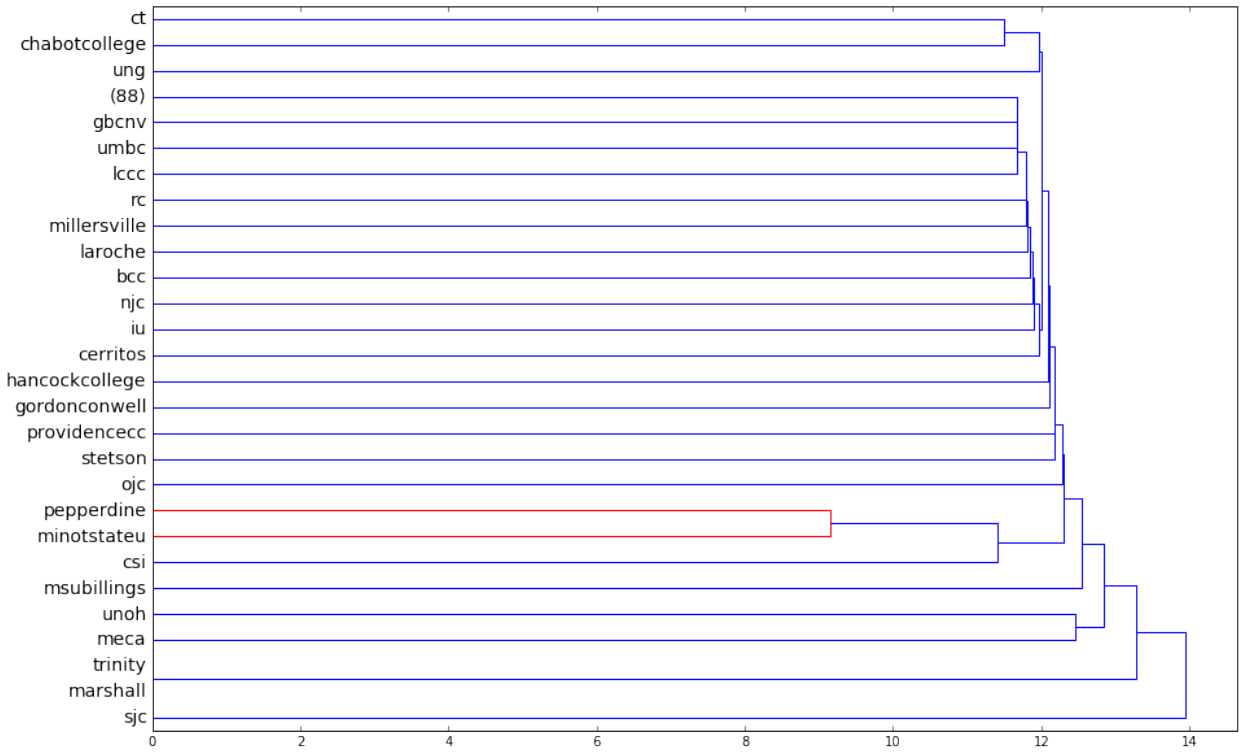

Figure 9: This truncated dendrogram of all PDF-based FERPA notices shows that many of the notices were indeed highly similar with limited hierarchical separating the 115 PDF notices. A more complete dendrogram is included in the Appendix.

suggested by the flatness of the tree.

6. Conclusion

Stepping back from the technical details of our implementation and analysis, we conclude with the key policy takeaways of our findings. These points are meant to spur later legal research into the state of FERPA and education privacy.

6.1. Policy Conclusions

Because our query in a major search engine returned only 5% of the notices in existence, we conclude that at a nationwide level, FERPA notices are not documented in an organized and easily search-able manner. Furthermore, our dataset demonstrated that a sizeable fraction of FERPA notices are embedded in longer documents, which increases the difficulty of tracking down individual notices.

Next, we observe many schools publish FEPRA notices very similar to the one from the De-

partment of Education. Heading into the study, there was little prior knowledge regarding the similarities of FERPA notices as the federal government does not require schools to follow the model notice. Because our study shows that many schools do simply re-purpose the model FERPA, the Department of Education should take greater effort to ensure the model notice is in fact accurate, thorough, and clear.

Finally, we find that the notices are very effective in mentioning the required terms, such as "Directory Information" and "Legitamate Educational Interest". But, very few notices elaborate on the optional components of FERPA notices, including an "Opt Out" section or contact information. This suggests that the notices are simply stating the bare minimum and may not give student and parents sufficient information to translate their privacy concerns into action.

6.2. Ethics

As a note on the ethics of our policy conclusions, it should be noted that our results may suggest misleading claims against the schools represented in this study. As any classification model that avoids overfitting will contain some testing error, it is possible that our FERPA extractions are not fully accurate. As such, the notices in our repository may not precisely reflect true notices. In response, we have denoted the accuracy of our classifier. Additionally, we did not validate the recency of the notices, which we directly scraped from Bing. It is possible that some of these notices have since been updated and do not reflect individual school's current polices. In short, one should interpret our data collection and analysis not as a listing of notices for specific schools but rather a general survey of the state of education privacy in the United States.

7. Acknowledgments

First and foremost, this project would not have been possible without the wise and gracious help of my adviser Dr. Arvind Narayanan, who corrected mistakes, suggested new directions, and challenged assumptions on a weekly basis. In addition, our co-advisor Elana Zeide of the Center for Information Technology and Policy (CITP) provided many of the initial motivations for the project and followed with expert background knowledge. A final thanks to Dr. Christiane Fellbaum for providing advise on the NLP analysis of this project.

7.1. Note on Partner Collaboration

This written report was written in collaboration with my partner, Sejin Park, although we wrote the majority of our papers separately. Specifically, parts of the Abstract, Introduction, and Problem Background were written in collaboration. In addition, we shared many of the same figures, which depended on our combined efforts. Content-wise, our research (in ideation as well as execution) was in collaboration with one another. However, the problems that we tackled independently, which were developing SVM-based classifiers for myself and content extraction heuristics from the web for Sejin, constituted a big bulk of our independent work and thus our reports.

8. Honor Code

I pledge my honor that this paper represents my own work in accordance with University regulations.

David M. Liu

References

[1] D. Bergmark, P. Phempoonpanich, and S. Zhao, "Scraping the acm digital library," *SIGIR Forum*, vol. 35, no. 2, pp. 1–7, Sep. 2001. [Online]. Available: http://doi.acm.org/10.1145/511144.511146

[2] C. A. Brodie, C.-M. Karat, and J. Karat, "An empirical study of natural language parsing of privacy policy rules using the sparcle policy workbench," in *Proceedings of the second symposium on Usable privacy and security*. ACM, 2006, pp. 8–19.

[3] J. G. Conrad, K. Al-Kofahi, Y. Zhao, and G. Karypis, "Effective document clustering for large heterogeneous law firm collections," in *Proceedings of the 10th International Conference on Artificial Intelligence and Law*, ser. ICAIL '05. New York, NY, USA: ACM, 2005, pp. 177–187. [Online]. Available: http://doi.acm.org/10.1145/1165485.1165513

[4] A. Dasdan, P. D'Alberto, S. Kolay, and C. Drome, "Automatic retrieval of similar content using search engine query interface," in *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, ser. CIKM '09. New York, NY, USA: ACM, 2009, pp. 701–710. [Online]. Available: http://doi.acm.org/10.1145/1645953.1646043

[5] M. W. Davis and W. C. Ogden, "Implementing cross-language text retrieval systems for large-scale text collections and the world wide web," in *AAAI Symposium on Cross-Language Text and Speech Retrieval. American Association for Artificial Intelligence*, 1997.

[6] M. Iwayama and T. Tokunaga, "Cluster-based text categorization: a comparison of category search strategies," in *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1995, pp. 273–280.

[7] D. Lewandowski, "Evaluating the retrieval effectiveness of web search engines using a representative query sample," *Journal of the Association for Information Science and Technology*, vol. 66, no. 9, pp. 1763–1775, 2015. [Online]. Available: http://dx.doi.org/10.1002/asi.23304

[8] P. Sheridan, M. Braschlert, and P. Schäuble, *Cross-language information retrieval in a Multilingual Legal Domain*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1997, pp. 253–268. [Online]. Available: http://dx.doi.org/10.1007/BFb0026732

[9] N. Tomuro, S. Lytinen, and K. Hornsburg, "Automatic summarization of privacy policies using ensemble learning," in *Proceedings of the Sixth ACM Conference on Data and Application Security and Privacy*, ser. CODASPY '16. New York, NY, USA: ACM, 2016, pp. 133–135. [Online]. Available: http://doi.acm.org/10.1145/2857705.2857741

[10] Y. Yaari, "Segmentation of expository texts by hierarchical agglomerative clustering," *arXiv preprint cmp-lg/9709015*, 1997.

[11] E. Zeide, "Student data privacy: Going beyond compliance." *State Education Standard*, vol. 16, no. 2, p. 21, 2016.

9. Appendix

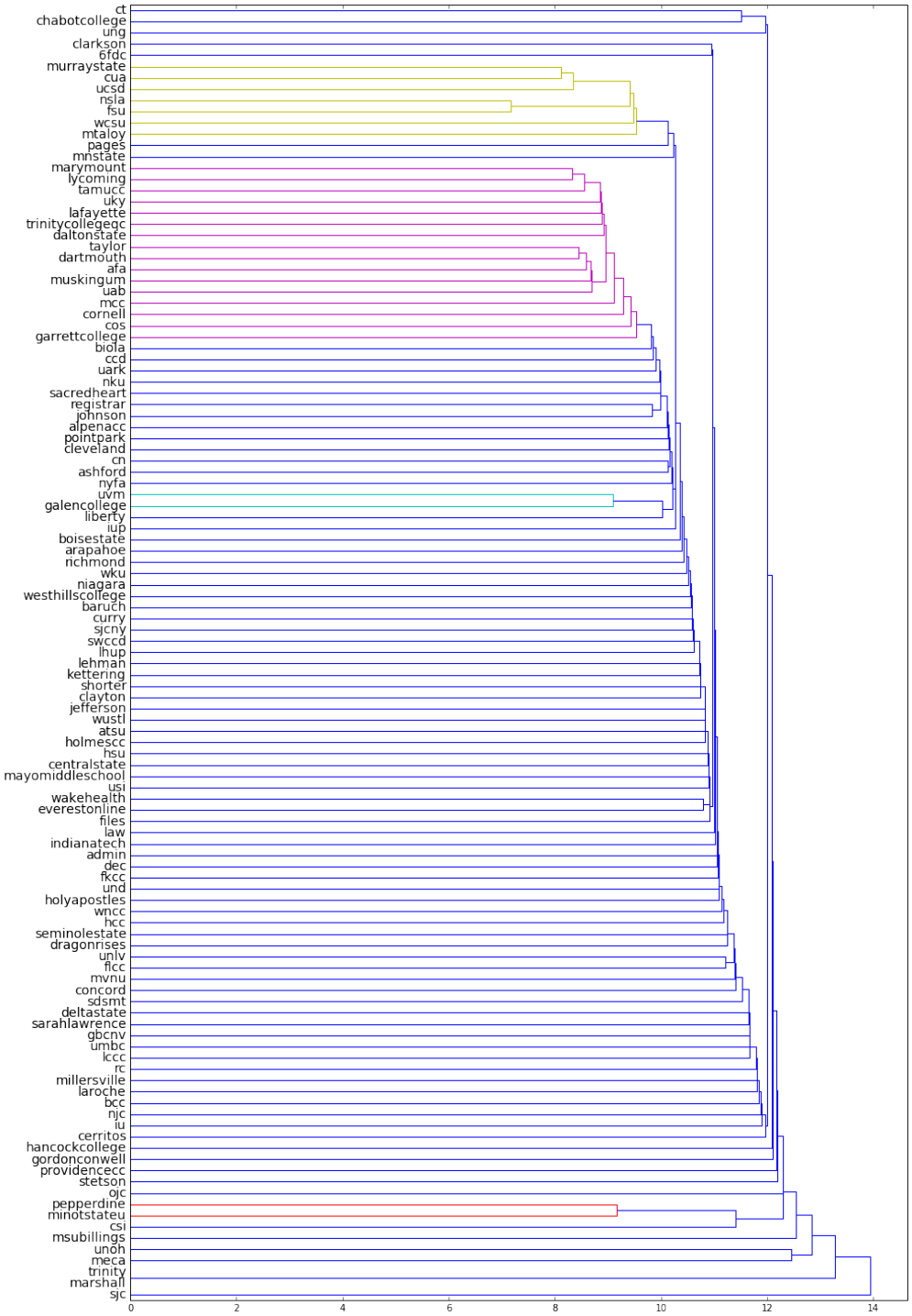

Figure 10: The full dendrogram for all PDF notices shows that while the majority of notices were highly similar there were two smaller clusters of distinguished notices shown in purple and yellow.

The complete source code for this project is available at the following private GitHub repository:

https://github.com/citp/ferpa-analysis