CS171 PROJECT PROPOSAL

Linghong Chen, Peter Gathua, Suhas Watturkar

Campus Security and Safety Visualization

Background and Motivation

Campus safety is of utmost importance to everyone. There are a number of sources from where safety data is available, but these sources do not necessarily provide an integral view of the data, which can be easily correlated. Also, the magnitude of this data is large and the granularity provided by the data is also very diverse, which makes it very difficult (if not impossible) to make informative conclusions from the data.

One prominent sources of campus safety data is The Office of Postsecondary Education, they provide this data at granular level, with different severity levels right from disciplinary actions to criminal offences, on / off campus locations, etc. This data is available for years 2005 to 2013. The goal of this project is to unify this data into a common data semantics, and provide interactive visualizations for this data.

Project Objectives

The objectives of this project are to provide visual representation of the campus safety and security data in following two major categories:

a) Map View
b) Crime patterns view

The user will be able to interact with the visualization and review the data based on various dimensions, such as location (state, county, city, etc.) or individual college. Also, the crime pattern view will enable the user to visualize the possible correlations with college details such as ranking, acceptance rates, gender ratio, size, etc.

Data

1. Office of Postsecondary Education (OPE)

Campus safety and security data is available from the Office of Postsecondary Education (www.ope.ed.gov) in excel format. The details on how to get to this data set are listed in Appendix A.

From this data source, following three data set files will be used:

File name	**Description**
oncampusarrest	Contains on-campus arrest data for various categories of crime ranging from drugs to weapons related offenses.
oncampuscrime	Contains on-campus crime data for various categories such as murder, robbery, sex offense, burglary, etc.
oncampusdiscipline	Contains discipline data for the categories under on campus arrest dataset.

Each data set covers data from all the United States colleges for a three year period, starting from the earliest (2005-2007) to most recent (2011-2013) periods.

Each dataset has the following common portions: school name, address, and number of students by gender, total student number, school category (public, private, 2 year, 4 year, etc.).

Each dataset also contains security information. The **on-campus discipline** and **on-campus arrest** data files each have weapons, drugs, and liquor crime categories, while the **on-campus crime** file has nine crime categories.

2. College Application Data

Open ICPSR is a vault where research data is warehoused and available to share. College applications data for 2014 is available at this location. This dataset is available in .sav format; it contains data compiled from Peterson's Guide in spring 2014. It includes college name, number of college applicants (male and female), acceptance rate, the college's US NEWS college ranking, and the college's student body size.

Data Processing

The dataset used for the visualization consists of under 30 files, mostly in excel format. The data from these files will be converted to tab delimited or JSON format. For converting to JSON, we can use a simple Excel VBA script, not special tooling will be required. In addition, efforts may be required to correlate school names from different sources of data, with slight variations due to usage of abbreviations / synonyms etc.

The total number of schools included in the dataset would be of the order of ten to eleven thousands. We may need to create aggregated or random samples of this data for various parts of the visualizations.

There are several open source toolset available to parse data in .sav format (Apache POI for example). The college application data from Open ICPSR is in **sav** format, and it will be converted to JSON by using open source toolset.

Google geo coding will be used to correlate the addresses of the schools to the latitude and longitude on the map.

Visualization

Visualization will be implemented in two portions – Security Map and Visual Analysis of the data.

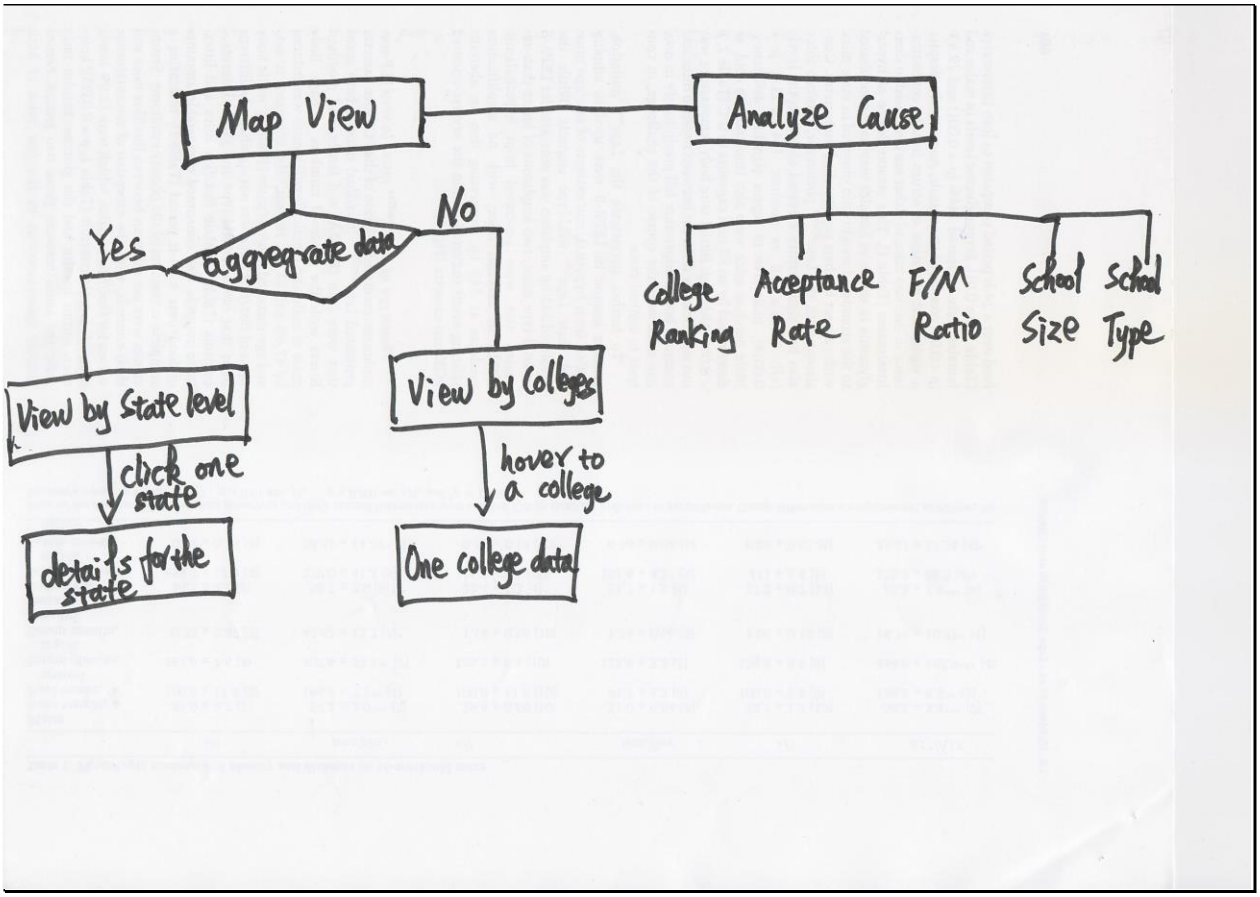

Figure -1: Overall Structure.

1. Security Map

1.1. Aggregating college security data to a state level, and show the average security level in this state on the map.

On the top of the map, will a bar plot showing aggregated data for various categories, which relate to the overall data shown on the map. Hovering on the individual bars will update the map with the data corresponding to the category shown by the bar.

For example, if the user hovers over the *public schools'* ***bar***, the map will show data only pertaining to the public schools.

Similarly, the other radio buttons on the top side of the map will update the map with the category represented by the radio button.

After visualizing the data, we may decide to add more categories to filter the data so that the visualization becomes more effective.

(See Figure 2)

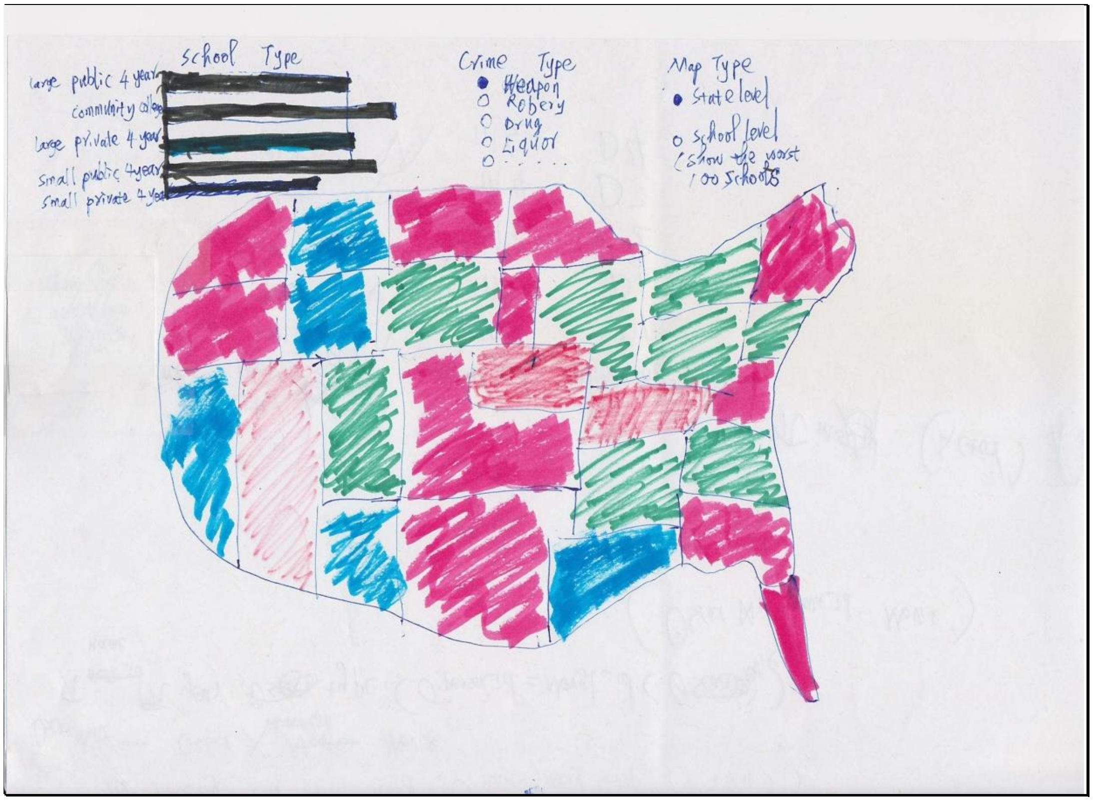

Figure 2: Security Map. Two colors will be used. One color represents the safety zone for the area close to zero violation, and another color represents the security level, color hue is the indicates of the safety level.

1.2. City level view

We will selectively choose certain number of cities/towns (to a number that doesn't not make the map look too crowded) to show their security level. Color will be used to represent the average level of college security for a city/town, and circle size will be used to indicate how many colleges is included.

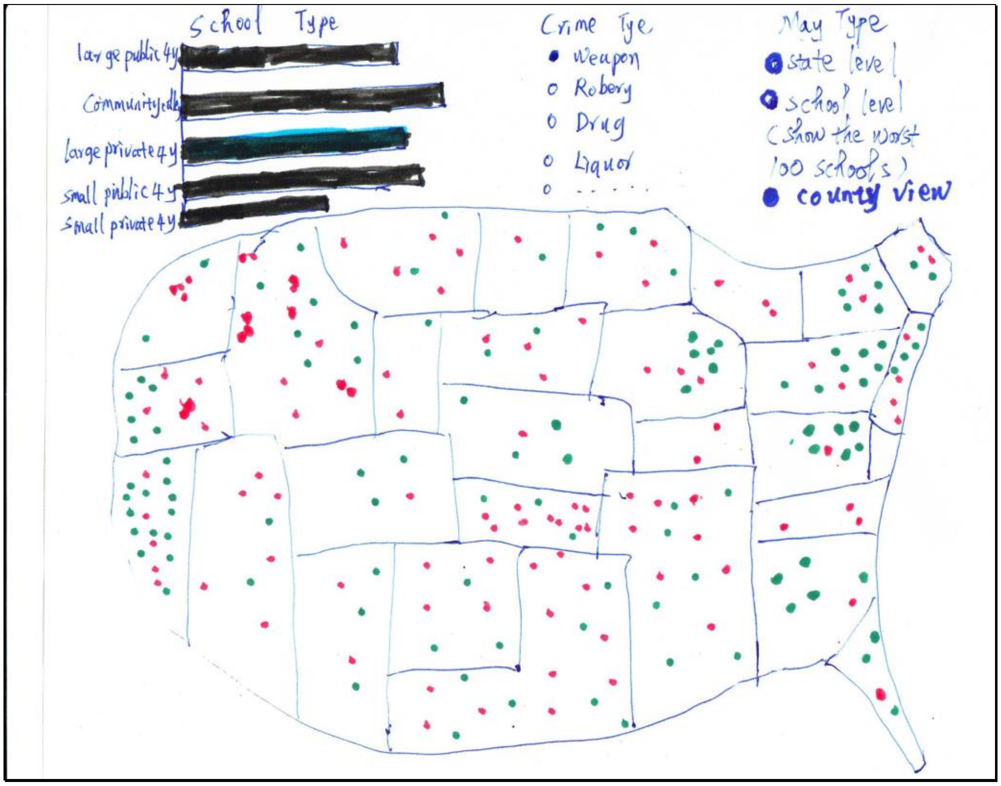

Figure 3: City level view.

1.3. **Colleges that have the worst security level**: As shown in figure 3, each university will be represented by a different color, which will be indicative of the university's security level. Hovering over the circle will show the number.

We may also use other criteria to select colleges to show the security level

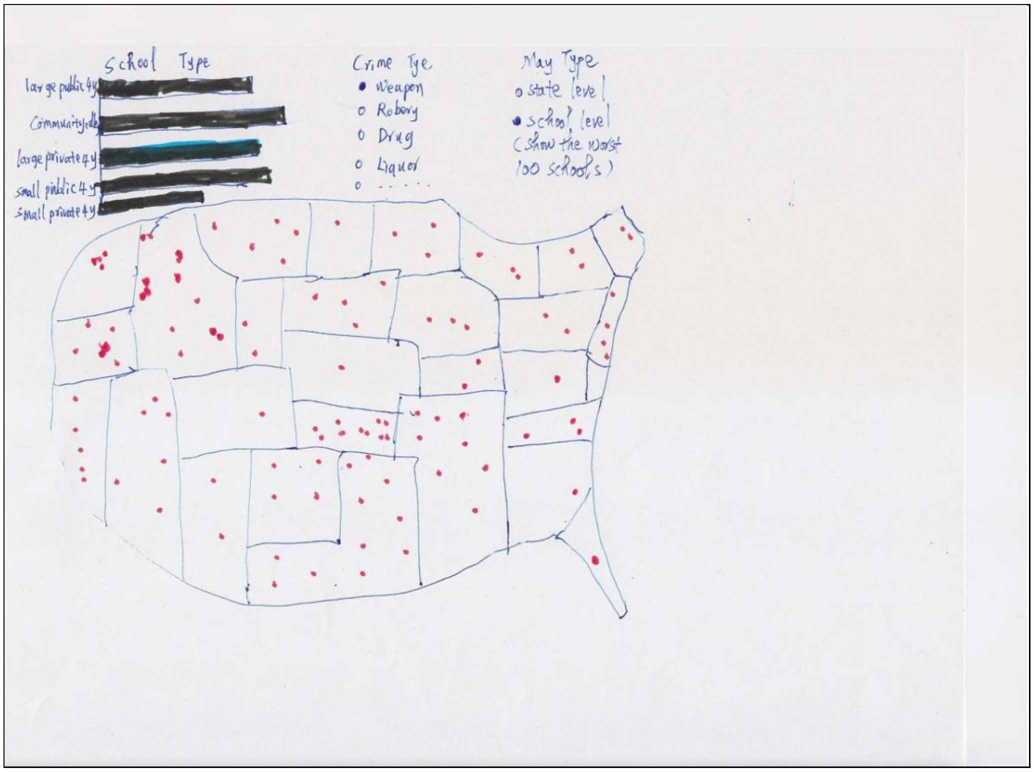

Figure 4: A Map view of the colleges in the United States with the worst crime statistics, color hue indicates the level of crime/violation number.

Clicking on a state will let the user *drill down* to the state level, showing all colleges within the state either using the zooming method (If we are able to achieve it) or use scatter plot (x as longitude and y as latitude) (see Figure 5)

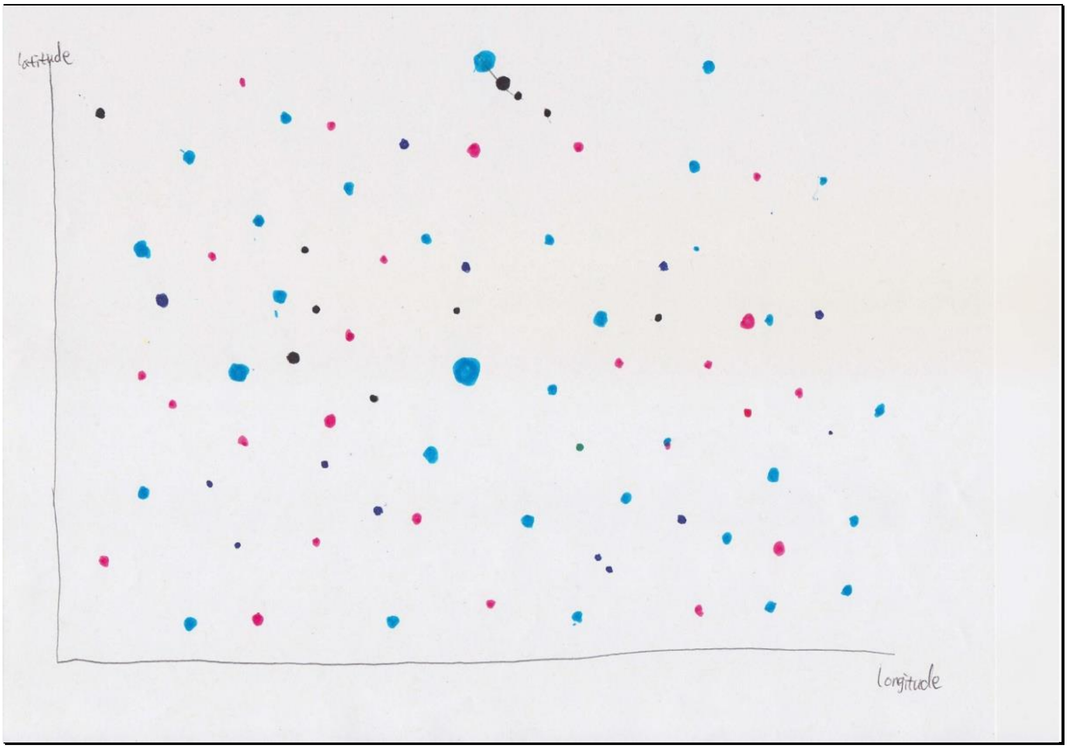

Figure 5: State level details.

2. Visual Analysis

The visualization should answer the following questions:

a) Does college quality (admission rate and ranking) affects campus security?
b) Does one school category (such as two-year colleges) had more violation than another category?
c) Does the gender ratio affect campus security?
d) Does school size affect campus security?

Possible approach 1: (The data aggregated way, preferred approach)

a) Does college quality (admission rate and ranking) affect campus security?

Scatter plots may be used. Data will be grouped by several groups according to admission rate / US News College Ranking. Within each group, random sampling will be used to reduce samples crowd level (see figure 6)

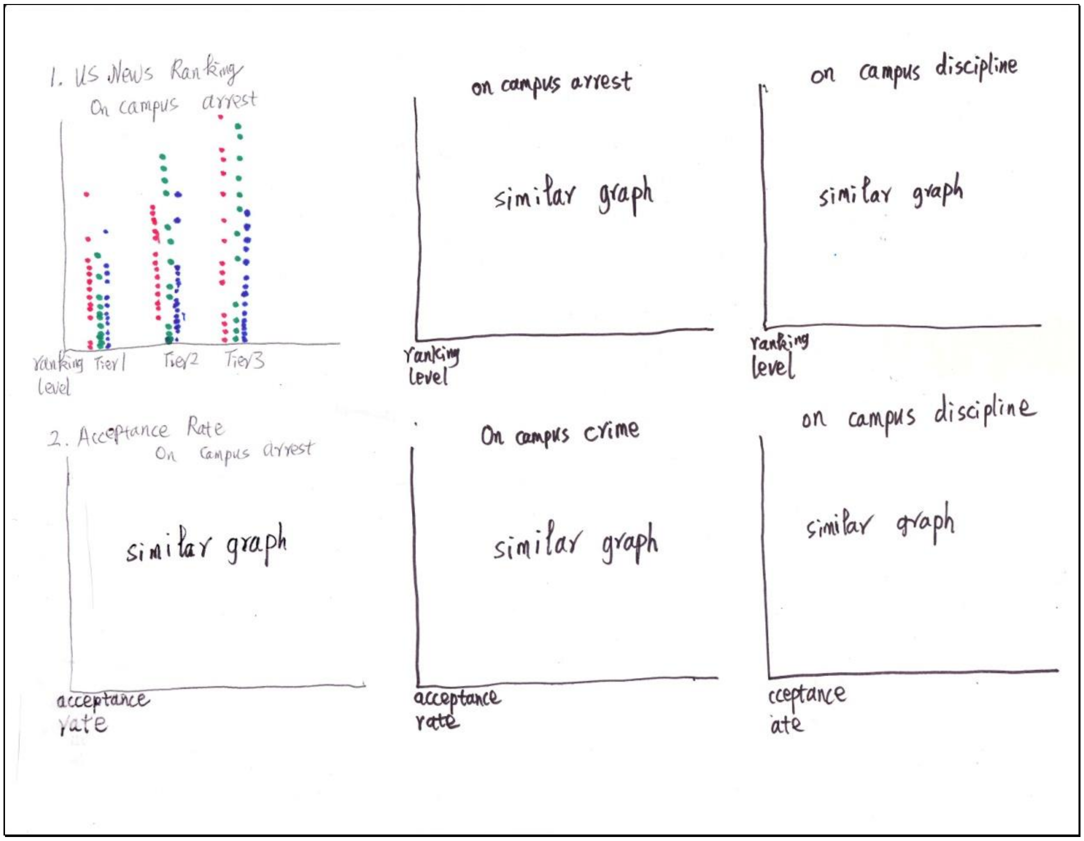

Figure 6: Effect of college quality

b) Does the gender ratio affect campus security?
Scatter plots may be used. Data will be grouped to several groups according to gender ratio. Within each group, random sampling will be used to reduce samples crowd level. We will determine how many groups will be used to best visualize the data. (See figure 7)

c) Does school size affect campus security?

Scatter plots may be used. Data will be grouped by several groups according to college size. We will determine how many groups will be used to best visualize the data. Within each group, random sampling will be used to reduce samples crowd level. (See figure7)

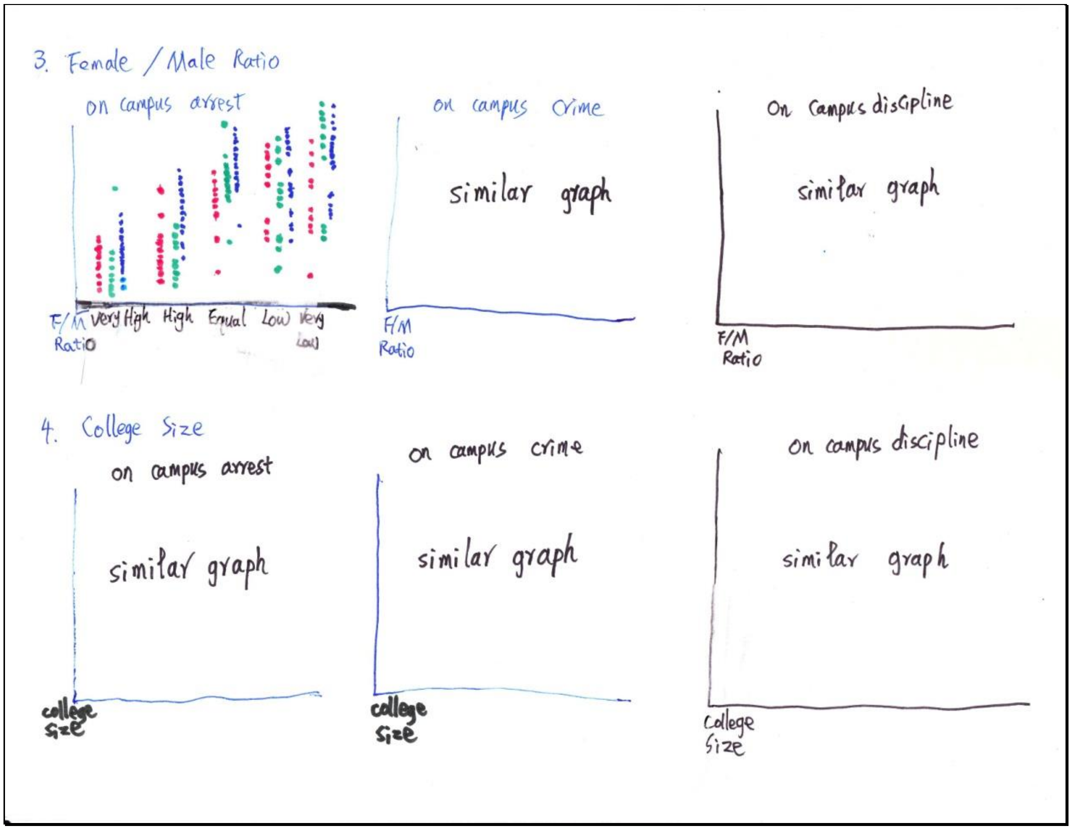

Figure 7: Effect of gender distribution and school category on campus security. Colors are indicated different crime of violation categories

d) Whether one school category (such as two-year colleges) had more violation than another category?

Data will be grouped by different school types. Scatter plots will be used. Within each group, random sampling will be used to reduce samples crowd level (See figure 8)

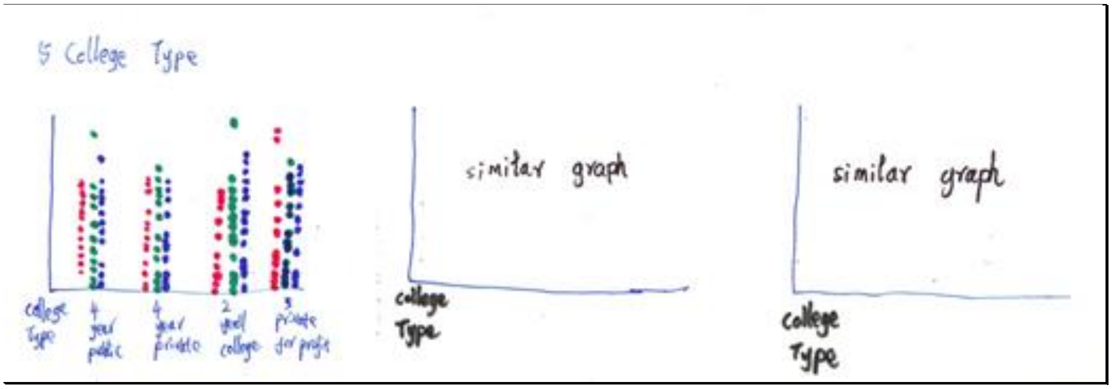

Figure 8: Effect of school size

Possible approach 2. Parallel Coordinates to show the relationship (Optional)

We will try the parallel coordinate method to see whether a clearer relationship among the above factors is found.

(See an example at: http://bl.ocks.org/jasondavies/1341281)

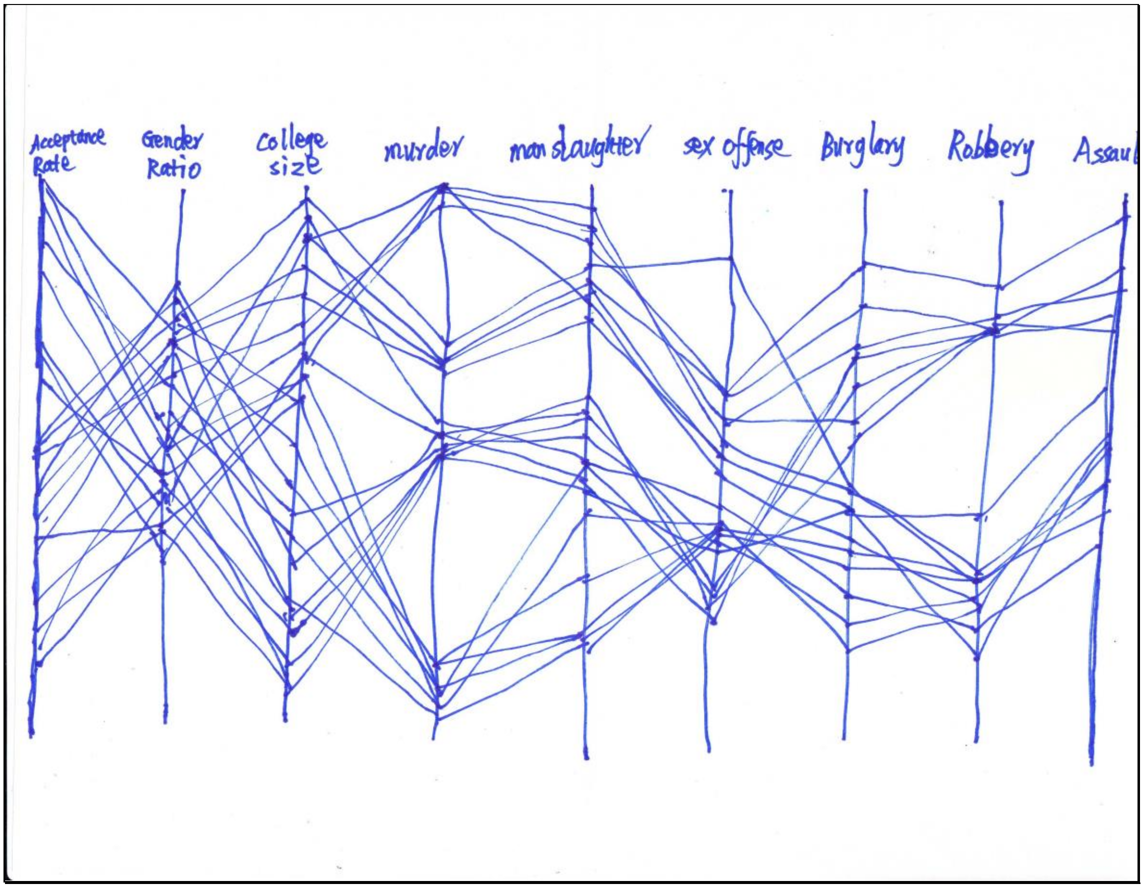

Figure 9: use of parallel coordinates to visualize the relationship

Must-Have Features

We will use a modified agile development approach, thus the details of the features may be refined when we get to know more about our data, and view the visualization results.

At a high-level, following features will be available in the visualization:

1) A map showing the campus security distribution.

2) Tables or graphs for a display to study the effects of internal and external factors on campus security.

3) Rich event based interaction between various parts of the visualization, which will make the data more understandable to the user.

Optional Features

1. Display the security level for individual colleges in the map and implement a zoom feature to drill down into details.
2. Not all the data research results will be displayed, only those give best visualization will be used.

Project Schedule

Date	Milestone
4/3/2015	Project proposal completed and submitted.
4/8/2015	Data to be used cleaned up and converted to JSON, ready to use for the visualization. Design of the visualization finalized.
4/15/2015	An initial working prototype ready working with subset of data.
4/22/2015	Code Merge and integration test.
4/26/2015	Project reviews with TFs completed.
5/1/2015	Beta version released.
5/5/2015	Final project submission.