Anita Mehrotra
anitamehrotra@fas.harvard.edu
CS 171-Spring 2014

Visualization Process Book v1.0

Due: April 10, 2014

Collaborators: -

About the Project

Overview, Motivation & Related Work

In his 2012 State of the Union Address, President Obama noted that the future income of U.S. children depended heavily on the quality of their K-12 teachers, so much so that "a good teacher [could] increase the lifetime income of a classroom by over $250,000 [2]." The reference was to a paper that had been published just three months earlier by Harvard economists Raj Chetty and John Friedman, and Columbia Business School professor Jonah Rockoff [1]. Chetty et al had measured the value-add (VA) of teachers - defined in terms of students' test scores - and utilized this information to quantify the impact that VA had on students' college attendance and projected income.

It was from this research that the Equality of Opportunity Project (EOP) was born. Designed to study U.S. educational inequalities and their impact on future leaders, EOP was spearheaded by four economists from Harvard University and the University of California-Berkeley. In January 2014, two new works were published. They explored the effect of education, mean parent income, racial segregation and teenage pregnancy (among other features) on future *relative* and *absolute mobility* [3, 4].

The mobility metrics were couched in terms of children's projected mean income relative to other children in the same birth cohort. To summarize the conditional expectation of a child's rank given her parents' rank, Chetty et al performed a linear regression and defined "relative mobility" as the estimated slope. Then, given a percentile p of the national parent income distribution, the researchers were able to calculate the expected rank of children, i.e. the child's "absolute mobility at percentile p" [3].

I am most interested in the spatial trends in education and mobility that Chetty et al study in Section V of [3]. Two years earlier (in 2012), *The New York Times* produced a visualization to summarize a small subset of the then-computed mobility data. My hope is to create a more comprehensive, better visualization that provides the user with additional information about how the 2014 mobility statistics were computed, as well as the feature inputs that tend to play a significant role in a child's future income.

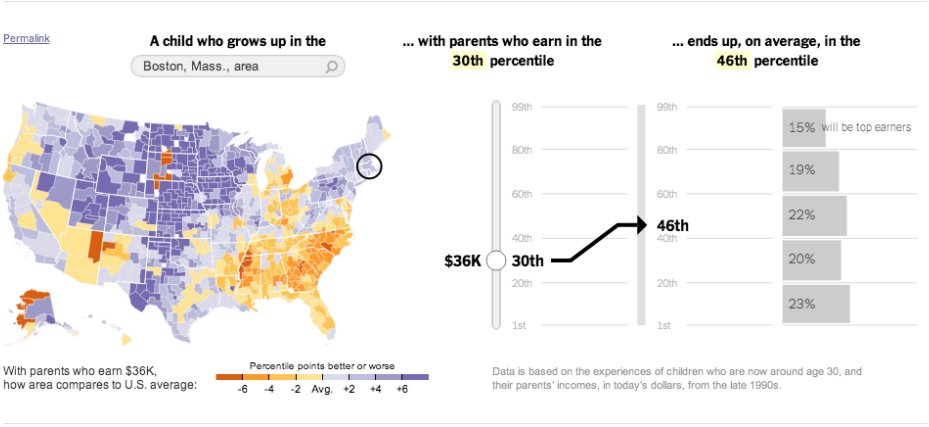

Figure 1: *The New York Times* visualization [6].

Project Obectives

I have five primary objectives:

1. Capture a country-level look at U.S. K-12 education.
2. Succinctly present a clear picture as to the factors that influence a child's future income.
3. Provide users with additional information as to how mobility is computed.
4. Present the data in a more creative way than is being done with *The New York Times* visualization.
5. Tighten my skills in brushing, hover, tool-tips and, importantly, linking visualizations together.

The second type of data required will be data that contains U.S. county coordinates. Fortunately, this data appears to be available via Michael Bostock's Github site in JSON format [7].

Initial Sketch of Visualization

This is an initial sketch of my proposed visualization:

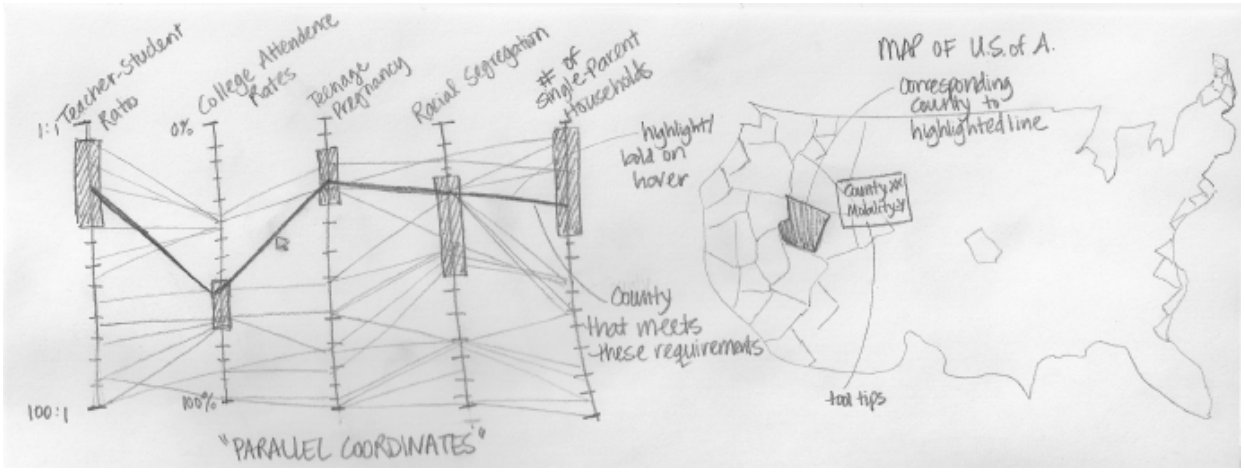

Figure 2: Anita's proposed final project visualization sketch.

Exploratory Analysis

I complete some initial exploratory analysis of the data in `exploration.ipynb`. This is useful because it gives me a feel for how much data exists, what my data of interest looks like, and what data processing and aggregation I will need to do.

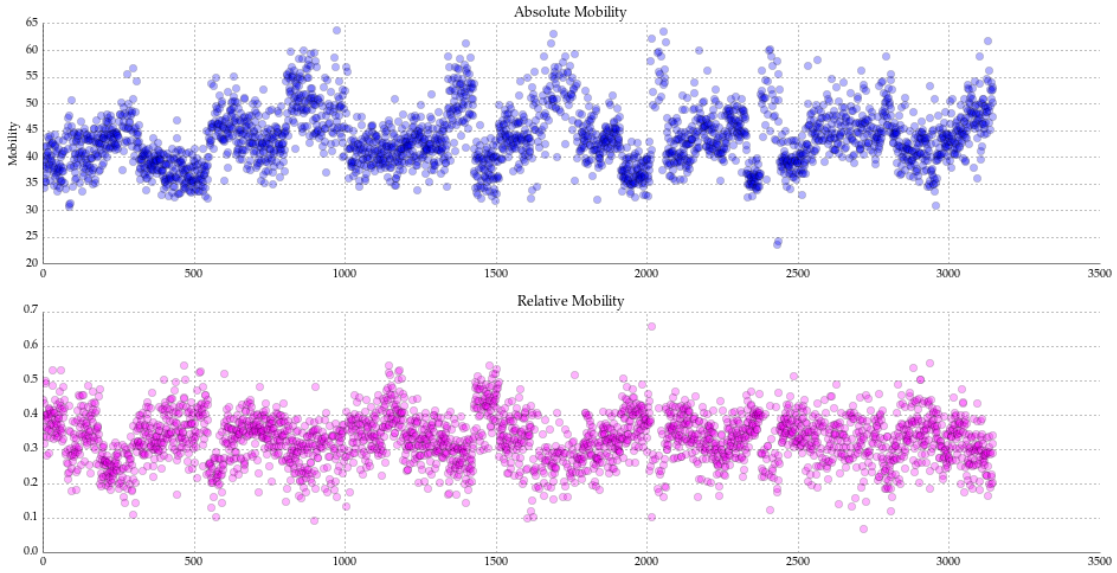

Figure 3: Comparison of absolute and relative mobility.

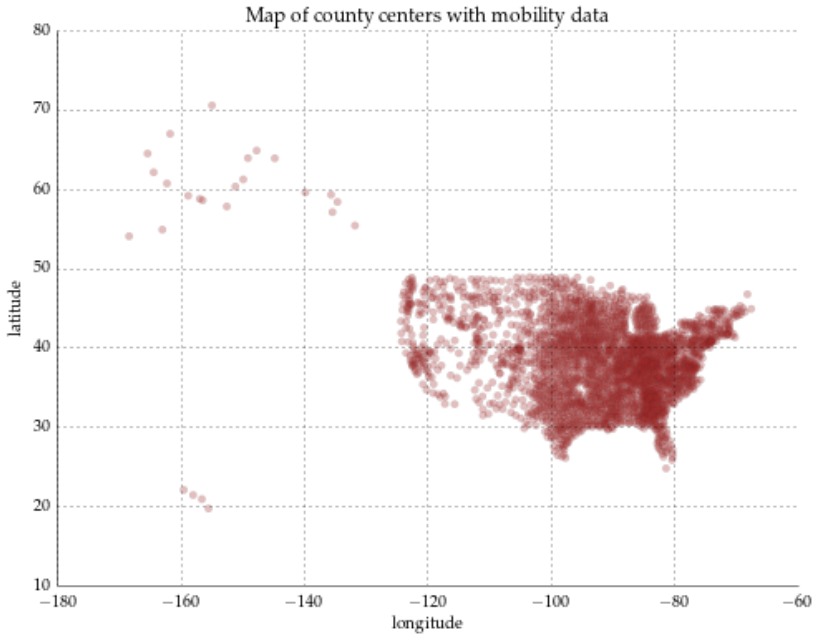

Figure 4: Map of mobility data on county centers (existence).

Implementation & Design Evolution: Timeline of Progress

MARCH 31, 2014

The Data

I use data from Online Tables 3 and 8 from the Equality of Opportunity data download site. Online Table 3 contains intergenerational mobility statistics by county, as computed in [3]. Online Table 8 contains commuting zone characteristics and will be essential to performing exploratory analysis on the education data. All of this data is publicly available [5].

Here are some important definitions and econometric concepts that played a significant role in my design choices:

- A base birth cohort (1980-82) was used to compute the mobility metrics. These are U.S. citizens born between 1980 and 1982 and are the oldest children in the data for whom Chetty et al could reliably identify parents based on information on dependent claiming. These children's incomes are computed in 2011 and 2012, when the children are approximately 30 years old. Additionally, their parents' mean income is measured between 1996 and 2000, when the children are between the ages of 15 and 20.
- There are two different mobility metrics that Chetty et al compute: "relative mobility," also known as "rank-rank specification," and absolute mobility. Let's recall the definitions highlighted in the third paragraph of the "Overview..." subsection of "About the Project:"
    1. "Relative mobility" is defined as the difference in outcomes between children from top- versus bottom-income families, within a county.
    2. "Absolute mobility (at percentile p)" is computed in the following way: by combining the intercept and slope for a commuting zone, Chetty et al calculate the expected rank of children from families at any given percentile p of the national parent income distribution. They claim that measuring absolute mobility is valuable because increases in relative mobility have ambiguous normative implications, i.e. they may be driven by worse outcomes for the rich rather than better outcomes for the poor.

Completed

Today I created a choropeth map of the U.S using the county shape file used by Mike Bostock [7]. I also make a deliberate decision to use *absolute* mobility data for two reasons:

1. First, there is significant variation across the country for absolute mobility. As a result, the visualization is more interesting to study at, and, frankly, is prettier to look at. I tried plotting relative mobility and there is little variation relative to the absolute mobility map.
2. Second, as noted in "The Data" subsection above, [3] and [4] both use absolute mobility in their analysis. Absolute mobility metrics are valuable because the conclusions drawn from studying relative mobility are more ambiguous, making it even more difficult to claim causality.

To ensure that the correct mobility information was used, I needed to merge the FIP county ID's with the county mobility metrics so that the correct mobility measure was mapped to the correct region. This was a bit tricky since there are multiple counties (in different states) with the same name. The final data file can be found in `absolute_mobility.tsv`.

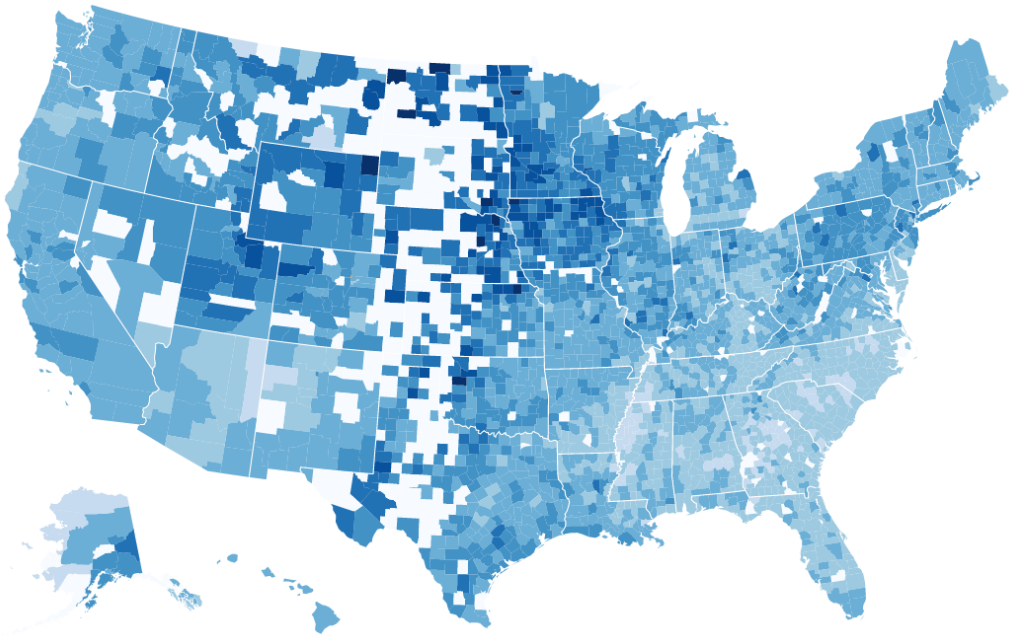

Figure 5: Map of the U.S., colored by mobility.

Notice that counties with very low mobility scores are very, very light blue. I thus may need to change the color scale. Also (and this is more noticeable when you view the `html` code in a webpage), there is one county in the very center of the U.S. that is colored orange. This county has undefined mobility.

APRIL 8, 2014

Design Studio Feedback

Today I met with a group that is creating a network graph visualization using music and preference data from `last.fm`. The members of the group are Tony Blum, Alan Xie and Bryan Kauder.

In speaking with the group, I gathered some useful feedback. They suggested the following:

- Make sure to include a color bar with the geographic map, so that it is clear how the color scale maps to mobility.
- Consider including zoom functionality since the counties are so small.
- Change the state outlines so you can differentiate between counties even if they have the same mobility level.
- An alternative to the previous point is to change your color scale, or change your background color to black.
- Perhaps to handle counties with missing data, you can instead create a heat map with an invisible county map overlay. This way, you don't have to worry about differentiating between those without data, those with very low data, and the rest.

My Thoughts

I agree with most of the group's feedback and plan on implementing the first few points. For example, it was only until I began presenting to them that I realized a color bar will be necessary. However, I do not plan on creating a heat map because I think that this defeats the purpose of what I'd like to convey to the audience. In particular, anomalies in the data (say, pockets of the U.S. with high mobility despite surrounding areas with low mobility) will be smoothed out. But these are the areas that are most interesting to the viewer!

APRIL 9, 2014

More Data Processing

I realize that I will need a file of data containing education "features" that are used to compute mobility in [3]. However, this will require more than just copying and pasting from the raw data files, because the characteristics I am interested in are provided for *commuting zone* and not for county. As a result, I need to merge the two files. The code to do this is currently in an iPython notebook titled `merge-county-cz.ipynb`. I will put this into a complete python script later. The merged file is `feature_vectors.csv`.

A big decision that I made today was to choose the features that I would like to display in the parallel coordinates visualization. I choose those that I think are interesting (and will be interesting to the viewer), are relevant to education, and provide a reasonable, representative subset of features used in computing the mobility measures. Beyond that, however, my decision to include the six features below is somewhat arbitrary. Of course, I could go back through the model that is used in [3] to determine (via e.g. ten-fold cross-validation) which features are most statistically relevant. I defer this in case I have extra time after completing the visualization. The features I ultimately include in my visualization (in no particular order) are:

- Teenage birth rates
- Mean parent income
- School expenditure per student
- Teacher-Student ratio
- Test score percentile

Completed

After processing the data and creating the second data file, I completed the parallel coordinates visualization. This visualization can be found in `edu-lines.html` and `edu-lines.js`. I include a picture of it here:

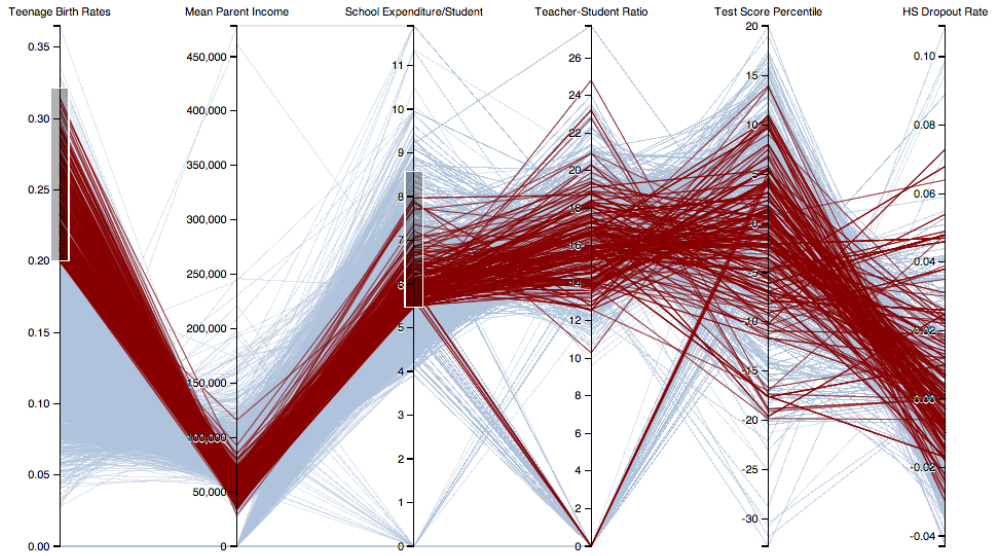

Figure 6: Parallel coordinates for education features.

Next Steps

The next major step is to link the map of the U.S. with the parallel coordinates visualization.

After that, there are some minor - but important - attributes I'd like to add to the U.S. map:

- Hover tool tip to identify county name and mobility percentage
- Color scale
- Zooming capabilities on US map
- Title

Thoughts and Questions

Here are some questions I'd like to discuss during the TF review next week:

- There are over 3000 counties, so the parallel coordinates visualization is kind of slow during brushing. It's also not super aesthetically attractive. What would be a good way to handle all the counties? Or should I leave it as is?
- What is a good ordering of the features in the parallel coordinates visualization?
- The original paper used commuting zones (as discussed above). As a result, there are some areas in the map that have undefined or very levels of mobility. One option is to change the color of these counties to something contrasting the current blue color scale, but I don't want them to get in the way when I highlight selected counties (that come from the linked parallel lines). Do you have any ideas about a good way to handle these?
- I'm concerned that the side-by-side visualizations on a shared screen will be too small. Should I consider a horizontal stack (I'm not super excited about this idea). Thoughts?

References

[1] Chetty, Raj, John N. Friedman, Jonah E. Rockoff. "The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood" NBER Dec 2011. http://www.nber.org/papers/w17699.

[2] Strauss, Valerie. "Obama on education in State of the Union address" 24 Jan 2012. http://www.washingtonpost.com/blogs/answer-sheet/post/obama-on-education-in-state-of-the-union-address/2012/01/24/gIQAVfAwOQ_blog.html.

[3] Chetty, Raj, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez. "Where is the Land of Opportunity? The geography of intergenerational mobility in the United States." Jan 2014. http://obs.rc.fas.harvard.edu/chetty/mobility_geo.pdf

[4] Chetty, Raj, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez. "Is the United States still the Land of Opportunity? Recent trends in intergenerational mobility." NBER Jan 2014. http://www.nber.org/papers/w19844

[5] *Equality of Opportunity* Raw Data Download. http://www.equality-of-opportunity.org/index.php/data

[6] Leonhardt, David. "In Climbing Income Ladder, Location Matters." *New York Times* 22 July 2013. http://www.nytimes.com/2013/07/22/business/in-climbing-income-ladder-location-matters.html?pagewanted=all&_r=3&#map-search

[7] U.S.A. County Coordinates. http://bl.ocks.org/mbostock/7586334.