MACS 40700 Final Project Paper
Due: June 4, 2017
Guangyu Liu
12145039

## Introduction

In the field of sociology, many research questions are related to one's status among a group of people, and how one's status changes over time. For instance, the classic definition of "inter-generational mobility" explores the relationship between father's education and son's education, father's occupation and son's occupation, father's income and son's income, etc. Scholars interested in "intra-generational mobility" may want to examine one's movement in a social hierarchy during his/her life span and to study how one's current status is related to previous status. Educational researchers may want to learn the discrimination of a certain subject, e.g. the gender gap in math is getting wider or narrower as students move to higher grades. All three examples above concern status at different stages, but define status using different measurements, such as income, occupation, or standardized test score.

This Shinny App aims at providing a tool for sociologist to explore such mobility. In the sidebar, users could choose how status is defined during and after K-12 education, and compare mobility trajectories between which groups based on their research interests. During K-12 education, standardized test scores of four subjects – reading, math, science, and history, as well as a composite score, which is the mean of reading and math scores, are provided. After K-12 education, user could choose one of four most common social status proxies – education, occupational prestige, income and social economic status (SES) score – to measure status.

The purpose of this Shinny App is NOT to compare different measurements. Instead, I assume the user has a research question at hand, and chooses measurements at different stages accordingly. Therefore, the flexibility provide by Shinny allows the user to explore different questions with the same dataset. The “Exploratory Graph” section provides mobility trajectories of different groups, while the “Statistic Analysis” section offers detailed statistic results. This mimics the procedure of a social science study – have a research question, define measurements, graph the data to visualize the trend, and then test statistic hypothesis.

## Data and Methods

For this project, I employed data from NELS88/00 (National Education Longitudinal Survey 1988 – 2000). 12,144 $8^{th}$ graders were first interviewed in 1988, and then followed up in 1990 ($10^{th}$ grade), 1992 ($12^{th}$ grade), 1994 (2 years after $12^{th}$ grade), and 2000 (8 years after $12^{th}$ grade). At the third follow up (1994, 2 years after $12^{th}$ grade), respondents were in a period of transition. Some of them have entered the labor market, while others remain in higher education. It is difficult to find a unitary indicator to measure their status. Therefore, data from the third follow up are excluded from this analysis.

For the occupational prestige option, only the respondent's primary job is included, since people's social status is usually defined by their main occupation. Each occupation is converted into an occupational prestige score based on [Norc 1980 Census Occupational Scores] (http://ibgwww.colorado.edu/~agross/NNSD/prestige%20scores.html). Military

prestige is based on the 2010 Census category. When there are no verbatim correspondent occupation categories, an average of related categories' prestige is used.

For the 4th follow up, SES is defined as an average of education, occupational prestige, and annual income. This is inconsistent with the respondent's original family's SES, which is offered in the NELS data.

## Design and Evaluation Based on Cairo's Five Criteria

### Exploratory Graphs

- Y axis. There are two ways to measure one's position in a social hierarchy – percentile and number of standard deviations from the mean. The problem of the former is that people on the top cannot move upward, thus it cannot show the degree of social inequality. When measured by standard deviation, people on the top can still move upwardly by enlarging their distance from the mean, but the disadvantage of this approach is the difficulty of interpretation, especially when the value of standard deviation varies across stages. For the sake of comparability and easiness of interpretation, I represent respondents' status by their percentile among their contemporaries.
- X axis. ggplot2 maps categorical variables to x axis with equal intervals. However, there are 8 years gap between the 4th and 2nd follow up, while there are only 2 years between the first three waves. Map waves with equal interval will result in a steep change between the last two waves and bias the visualization. Therefore, I mapped the year of each wave (continuous variable) to x axis, and then label them

as categorical variables.

- Confidence interval. It is useful to add confidence intervals to show if the difference between two groups is statistically significant. However, for this dataset, when there are multiple lines in the graph, the confidence interval makes it difficult to differentiate different lines. In addition, except for the neighborhood of intersection points, most confidence intervals do not overlap. To keep the visualization clean and clear, I turned off the confidence interval.

**Statistical Analysis**

- This section shows how current status is correlated with previous status, by group. The upper left is a coefficient plot which allows the user to compare the magnitude of coefficients across groups, while the table on the right side provides detailed statistics.
- Color. I mapped group to both the x axis and the color channel. It is sometime redundant to use multiple channels representing the same information, but since this graph is faceted, the color channel allows the user to easily concentrate on the same group across facets.
- Axis labels. I intentionally put the labels of y axis (“Coefficients”, the coordinate is flipped) on the top. When user chooses a compare variable with many categories (e.g. mother’s education), the graph will extend very long. If the labels were at the bottom of the graph as usual, the user will have to scroll to the end of the graph to figure out the meaning of the y axis, which is inconvenient.