Longitudinal Study Using Belief Network Analysis on ANES Time series Data

For the project of Sujaya and Devin

This document describes a proposal for the methods that we can use in this project.

# Methods Procedure

There are four general steps to the proposed procedure.

1. Separate data by year, choosing only variables which can be compared across years
2. Create correlation network from each year of data
3. Correlation network path analysis
4. Cluster correlation network to find ideological groups
5. Match groups over time so individual variables can be tracked over time

The clustering/path analysis approaches depend on a fairly static belief network over the years. Depending on the validity of this assumption we may need to take a different route starting at step 3. This document will continue on as if this assumption holds.

Each of these steps will be described in the next sections.

## Data Pre-processing

Pre-processing of the data simply involves the inclusion of variables which can be compared longitudinally and any normalization that may be required to make them comparable.

**Implementation:** The processing of the data could be done using the Python Pandas library’s [dataframe object](http://pandas.pydata.org/pandas-docs/stable/dsintro.html).

## Correlation Network Creation

The creation of the correlation network is fairly simple. Create a vertex to represent each survey question. Create an edge between every two survey questions where the weight represents the correlation coefficient. Pearson’s correlation is appropriate for pairs of continuous scalar variables and Spearman’s Rank correlation is appropriate between two ordinal or one ordinal and one continuous variable.

**Implementation**: This step could be accomplished by using [Python Network X library](https://networkx.github.io/) for working with graphs. Note that this implementation may not be needed unless some kind of visualization is involved. Wait to implement until visualization is necessary.

## Path Analysis on Correlation Network

In order to perform path analysis a transformation to the network weights needs to be performed so that higher correlation have shorter distances (because the variables are more closely related).

One possible way of working with graphs with these weights (as defined in the BNA paper) is to identify the variable with the shortest-path betweeness measure. Additionally, a minimum spanning tree could be created to remove those edges which we are not interested in. Otherwise there are too many connections to feasibly visualize.

**Implementation**: This step could be accomplished using [dataframe correlation calculation](http://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.corr.html) on an entire year’s survey at once. Then the resulting matrix simply needs to be placed into edge weights.

## Ideological Clustering

The next possible step for analysis could be to use either the original correlation network or the minimum spanning tree to cluster the weights. This method would allow us to track the movement of clusters over the years. In addition to the clustering there would need to be an effort to align the clusters from year to year into a smaller time series.

**Implementation**: The clustering could be performed using the [k-means implementation in scikit-learn](http://scikit-learn.org/stable/modules/clustering.html#k-means) where the distances between points are the transformed weight matrix.

# Analysis Methods

These are the proposed methods for analysis. This will be important when presenting and drawing conclusions for the results section.

## Group Tracking Analysis

This section would involve matching clusters of questions across each year and determining which items flow in and out of those clusters each year. If we can come up with a title for each cluster corresponding to some kind of ideological grouping, then we can make some conclusions about the ways our ideologies have changed.

**Implementation**: I can’t think of any library we can use for this, but it may take a few tries to get right.

## Static Visualization

Visualization of an image of the correlation may make some obvious structural properties obvious. Vertex size could be based on some centrality measure of the system. Weaker/larger edges could be removed to make relationships more clear, or even the minimum-spanning edges could be included.

Implementation:

## Time-Dynamic Visualization

This is an extension of the time-dynamic visualization method where we create a video that captures the changes in importance/structure of our correlation network.

Implementation: