Occam's Razor

Volume 4 (2014) Article 9

2014

Social Neworks of Shared Attributes

Sky Hester
Western Washington University, sky.hester@wwu.edu

SOCIAL NETWORKS
of shared attributes

BY SKY HESTER

INTRODUCTION

People are connected. Some of the connections—like friendship or marriage—are easy to see, and some are less obvious. For instance, think of an abstract similarity based on an idea of identity, like ethnicity or gender; these too are connections between people. It is the collection of these relationships that define a person in their social context. A set of individuals connected through chosen relationships defines a *social network*.

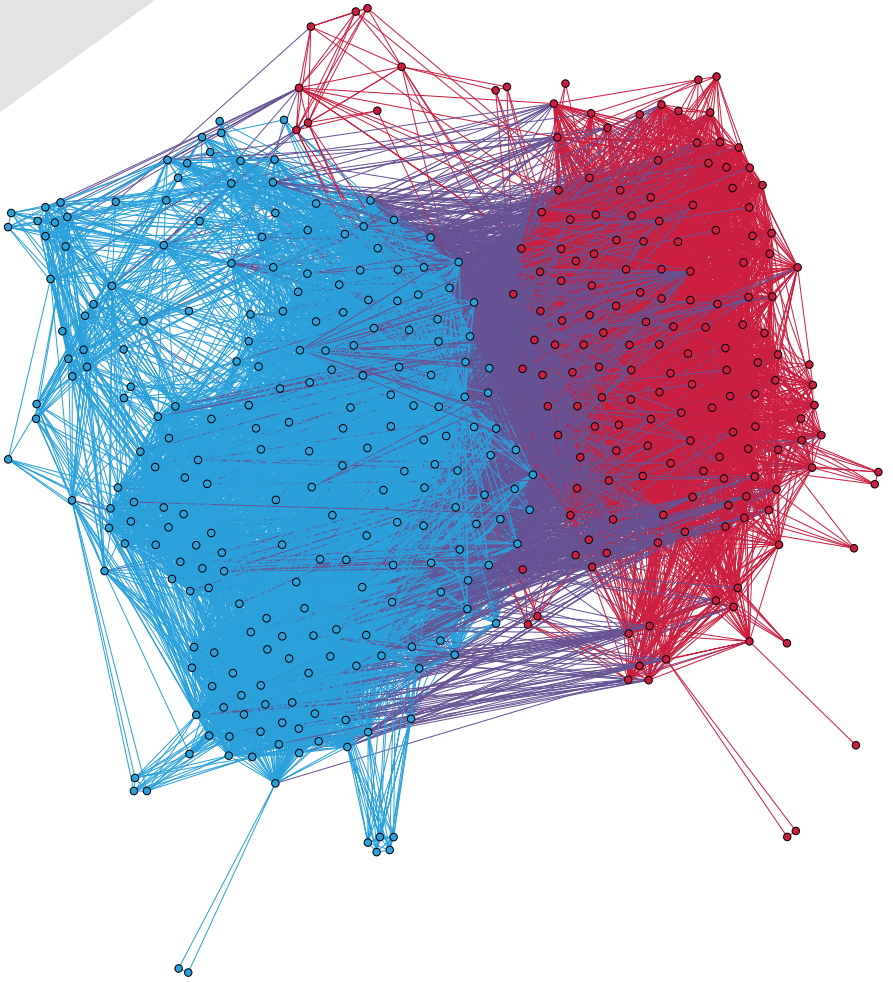
Social network analysis may be understood to deal with questions about social structure and individual or group behavior from the perspective of the relations between actors. Although it deals with people and can be considered a branch of sociology, the methods of social network analysis are highly mathematical. Fortunately, because the abstractions involved in the mathematical methods represent familiar constructs of our experience—people and their connections—they can be discussed in familiar terms. What follows is an effort to give an appreciation of this practice to the non-mathematical reader.

Social networks help to clarify the discussion of important social phenomena by allowing quantitative techniques to approach what is normally a qualitative domain of discourse. In particular, the extent to which educational attainment is socially closed within status groups can be examined by constructing a certain social network, which will be discussed here. This network was derived from a subset of the 2010 General Social Survey (GSS) dataset for the respondent variables described in table 1, restricted to respondents whose ages ranged from 30 to 40. Before treating the analysis itself, it will be necessary to familiarize the reader with certain relevant ideas.

CONCEPTS

A social network is treated mathematically as a *graph*, a structure which consists of objects called *vertices* to represent people, and other objects called *edges* to represent a connection between two people [2]. Visually, a graph is represented by a small dot for each vertex, and a line segment between two dots represents an edge (this is called a *drawing of a graph*, and should not be confused with the sort of charts that are usually called graphs or plots). In the language of social network analysis, we would say that the vertices of the graph represent actors in their social context and the edges their relations.

A *weighted graph* begins as an ordinary graph, where the vertices represent people and the edges represent their relationships, but once the relationships are in place, we assign a *weight* to each edge to represent the strength of the connection it reflects. In the situation where the edge indicates friendship, we may use the number of years the two friends have known each other as a weight, or maybe the number of hikes they have been on together or even the number of times they have given each other a hug.



These may seem like silly examples of the strength of a friendship, but each of these data would yield qualitatively different conclusions about the connection structure of the weighted graph representing friendships in a group of people. This is particularly true when the questions we wish to answer concern the *community structure* of a given network. For instance, the communities of hikers might be very different from communities of friends who attend the same concerts.

But how can we determine the communities in a social network? We would like to be able to do this objectively, according to some mathematical definition of a community. The goal of a community extraction procedure for a graph is to create a partition of the vertex set, meaning a way of assigning a community to each person, so the connections between people within a community are stronger than the connections between people in separate communities. The commonly used measure of quality for a community partition is known as *modularity* [1], which measures exactly how many more connections are in a community than would be expected if the relationships were distributed at random.

Here the phrase "at random" has a certain technical meaning that is usually based on the Newman-Girvan null model. In this model, the network under examination is compared to an artificial approximation where the local strength of relationships from the original network is maintained—that is, the total strength of relationships between a given vertex and the rest of the network is the same—but the relationships are distributed evenly across the network without considering its underlying community structure [5].

The optimal community partition with respect to modularity is determined here by the Louvain method, an iterative

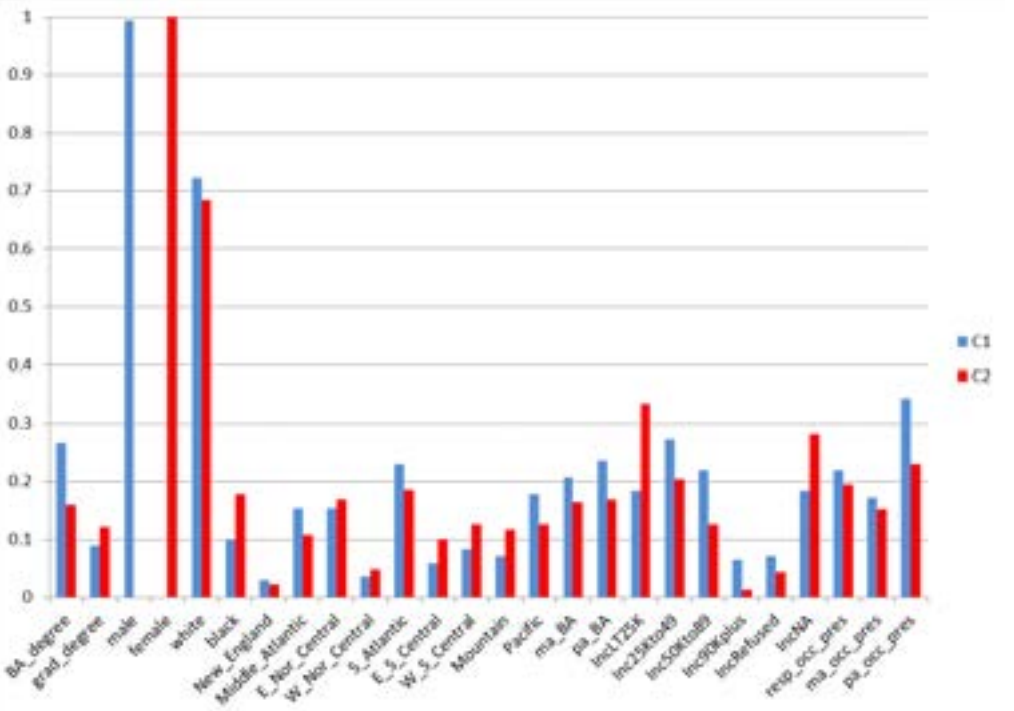


FIGURE 1: Modularity clusters in network of shared attributes.
(Previous page) Graph of shared labels, full set of labels
(Above) Community partition label distributions

algorithm which begins by assigning each person their own community and at each step joins pairs of communities in a randomized order if the larger community formed by the pair yields a higher modularity. In other words, the algorithm builds communities from the bottom up until the highest quality partition has been achieved.

Because we are unaware of the direct social connections between GSS respondents, we will have to satisfy our curiosity by establishing abstract connections between them based on their response variables. In particular, we will construct a weighted graph on which each edge is weighted by the number of common responses shared by the two respondents on either end of the edge in question. This construction will be called a *network of shared attributes*, and it is closely related to the concept of a multi-mode social network. At this point, it is important to note that the communities we extract from the graph no longer represent social communities but abstract *clusters* of respondents, though we will continue to refer to them as communities.

Once a community partition is determined for the network of shared attributes, we can analyze the differences between communities by comparing the relative distributions of their attributes; that is, we will see if there are any qualitative differences between the clusters of respondents by comparing their responses.

METHODS

As mentioned above, we will consider the selected GSS response variables as attributes, and from this we may create the corresponding graph of shared attributes, determine communities within that graph, and observe the distribution of responses within a community. For a given community, responses of high relative probability can be thought of as distinguishing traits, and the differences in response distributions will describe the differences in measured attributes for members of each community.

This analysis was achieved using R [3] and the network analysis software package Gephi [4]. Table 1 summarizes GSS response codes used in the figures.

RESULTS

When all attributes are used to construct a network of shared attributes for the GSS dataset, only two communities are extracted, each entirely distinguished by gender, as shown in figure 1. In all visualizations, edges of weight 1 are not shown in order to make the community separations more clear. When gender labels are removed from the set of attributes, the partition is characterized by the income distribution of each community, as shown in figure 2.

Both of the community partitions in figures 1 and 2 are determined by including edges of all weights. If we were to remove edges of small weight, the network would become extremely disconnected; many of the communities extracted from such a graph would be too small to analyze statistically, hence the motivation for including all edges.

It can be seen from figure 1 that two communities are extracted, each homogeneous with respect to gender response variables. Recall that this partition is constructed based only on the strength and number of connections within a community; as such, it is unbiased with respect to the actual meaning of the responses from which these weights are drawn. This indicates that male respondents had more common responses with other males than with female respondents. However, it is important to note that these gender clusters have one common response guaranteed and that their remaining response variables are only partially correlated with gender. In order to elucidate the finer structure of this data, we remove gender labels to obtain figure 2.

This second network decomposes into four communities, each nearly homogeneous with respect to their income response variables. From the distribution of attributes in figure 2, it is clear that high income (greater than USD 50k/yr) respondents (group C0) have the highest levels of occupational prestige; this is also the case for both their parents. These respondents also have the highest educational attainment and are twice as likely to have a graduate degree as the second highest income group, C2. The parents of high-income respondents are also most likely to hold a bachelor's degree. This community partition demonstrates a positive correlation between status-based (parental and self-occupational prestige, parental education) response variables and attainment-based (bachelor's and graduate degrees, income) response variables for this sample.

EDUCATION	
BA_degree	Respondent attained BA degree
grad_degree	Respondent attained graduate-level degree
PARENTAL EDUCATION	
ma_BA	Respondent's mother attained BA degree
pa_BA	Respondent's father attained BA degree
ETHNICITY	
White	
Black	
REGION	
New_England	
Middle_Atlantic	
E_Nor_Central	
S_Atlantic	
E_S_Central	
Mountain	
Pacific	
ANNUAL INCOME	
IncLT25K	Income < 25,000
Inc25Kto49	25,000 ≤ Income ≤ 49,000
Inc50Kto89	50,000 ≤ Income ≤ 89,000
Inc90Kplus	Income > 90,000
IncRefused	Respondent refused to provide income
IncNA	Unemployed or disabled
OCCUPATIONAL PRESTIGE	
resp_occ_pres	Occ. prestige score > 80, respondent
ma_occ_pres	Occ. prestige score > 80, respondent's mother
pa_occ_pres	Occ. prestige score ≥ 80, respondent's father

TABLE 1: GSS VARIABLES USED FOR LABELLING

Note that C3, the group associated with unknown income, could correspond to unemployed respondents. It is clear visually that they make up the smallest proportion of the data and are for the most part not strongly connected to the largest connected component (at weight 2). While C1, the community known to have income less than USD 25k/yr, makes up a larger proportion of the data, it is also not highly connected.

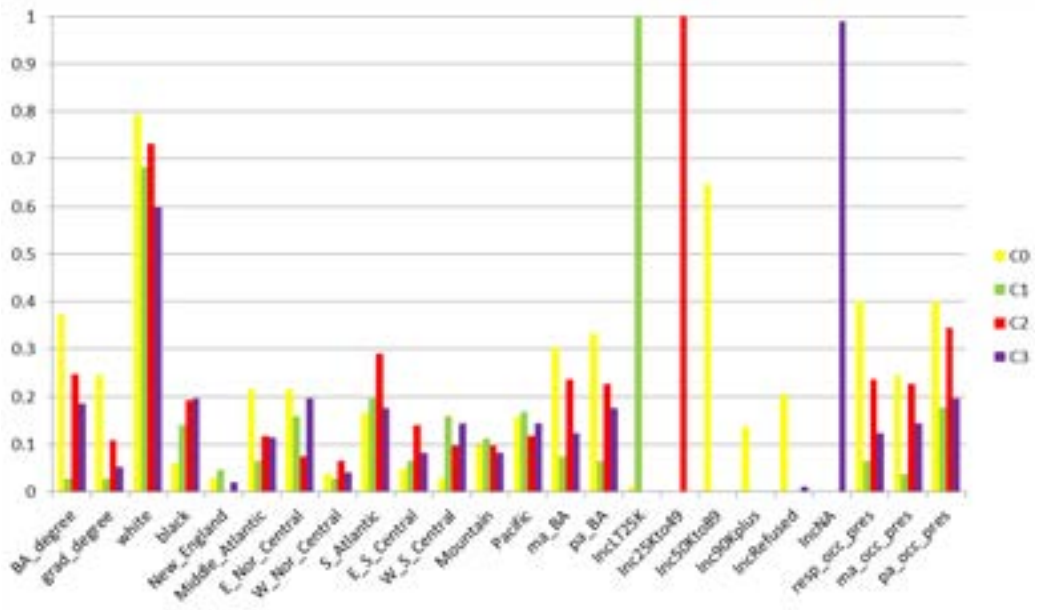
DISCUSSION

Intuitively, partition 2 is perhaps the most appealing. It seems to support the notion of social closure within families for socioeconomic status and education attainment and suggests the expected relationship between parental education and occupational prestige and attainment variables.

It is interesting to note that the lower income communities C1 and C2 appear to have several subcommunities which had only one response in common; this is determined visually by finding those clusters in the drawing which appear to be separated from the remaining vertices of their assigned community, or else only connected by one or two edges. We know this common response must be the income response by the homogeneity of those communities with respect to that response variable. This could indicate that there are relatively many reasons that respondents are experiencing low income and relatively few for high income, but that question can't be answered without a more careful analysis of a greater variety of GSS response variables.

There are other, more standard methods of achieving clustering with categorical datasets like the GSS. These methods have the disadvantage that they are usually dependent upon assumptions about the distribution of the data. In addition, these methods are not easy to visualize. Fortunately, modularity optimization is easy to visualize and makes no assumptions about distribution, but as noted earlier, it is highly sensitive to the choice of weights. A more justifiable use of the technique would be to relate respondents by their acquaintanceship and extract communities from the social network thus formed. Given the restrictions of the data, this could not be achieved, but to do so would be an important next step to verify the interpretation put forward by this analysis.

FIGURE 2: Modularity clusters in network of shared attributes, gender not considered.
(Opposite) Network of shared attributes, gender not considered
(Below) Community partition label distributions



REFERENCES

[1] V. Blondel, J. Guillaume, R. Lambiotte, E. Lefebvre, *Fast Unfolding of Communities in Large Networks*, arXiv:0803.0476 [physics.soc-ph], arXiv.

[2] D. Knoke, S. Yang, *Social Network Analysis* (2nd edition), 2008, Los Angeles: SAGE Publications.

[3] R Core Team (2012). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL www.R-project.org.

[4] M. Bastian, S. Heymann, M. Jacomy (2009). *Gephi: an open source soft- ware for exploring and manipulating networks.* International AAAI Conference on Weblogs and Social Media.

[5] M. E. J. Newman, M. Girvan (2003). *Finding and evaluating community structure in networks*, arXiv:cond-mat/0308217v1 [cond-mat.stat-mech], arXiv.