August 29, 2000, revised June 16, 2001

Heterogeneity in Component Membership (IQV) and Network Connectivity

This memo describes two very closely related measures: heterogeneity in component membership and network connectivity measured as the average proportion of actors reachable in the network. Each actor (household) in the village belongs to a component. A component is a maximal connected subgraph (see Wasserman and Faust 1994 Chapter 4). Prior to finding the components, each relation was dichotomized (any value greater than 1 was set equal to 1) and symmetrized (a tie was defined as present if either the tie from household i to household j or the tie from household j to household i was present)

A measure for heterogeneity in component membership is: $1 - \sum p^2$, where p is the proportion in a component. Agresti and Agresti (1977) refer to this measure as the index of diversity (D). In general it is interpreted as the probability that two elements are in different categories of a variable. In our application it is the probability that two households are in different components in the village. According to Agresti and Agresti, the index of qualitative variation (IQV) is $1 - \sum p^2 / (1 - 1/k)$, where k is the number of categories of the variable. The denominator is the maximum possible value of $1 - \sum p^2$, which varies depending on the number of categories of the variable.

It turns out that the index of diversity in component membership ($1 - \sum p^2$) is very closely related to the average proportion of actors (i.e. households) reachable in the network. In fact the average proportion of actors reachable is equal to $\sum p^2$. (To convince yourself that this works, think of a network in which .4 are in component A, .3 in component B, and .3 in component C. To find the average proportion reachable use the weighted mean $(.4 \times .4)+(.3\text{x}.3)+(.3\text{x}.3)= \sum p^2$.)

The "average proportion reachable" measure has an established place in the network literature on random and biased networks (Skvoretz 1985). In that context it is one of the "structure statistics" calculated on a network. The structure statistics are the proportions reachable in 1, 2, ... steps from a randomly selected set of "starter" actors. To quote Skvoretz (1985)

"The network property of interest is connectivity or integration in the sense of how reachable the entire population is from a randomly chosen starting point. It is measured by the network's structure statistics: the sequence P0, P1, P2... where Pi is the proportion of the population newly contacted i steps away from a set of randomly selected "starters" or the related cumulative sequence X0, X1, X2... where $Xi = \sum pi$, the proportion reachable from the starter set in i or fewer steps and where $\gamma = X_\infty$ is the limiting fraction ever reached." (Skvoretz 1985 page 228).

These are clearly related to the exact and cumulative path length statistics.

References:

Agresti, Alan and Barbara F. Agresti 1977. “Statistical analysis of qualitative variation.” Pages 204-237 in K. F. Schuessler (ed.) Sociological Methodology 1978. San Francisco: Jossey-Bass.

Skvoretz, John 1985. “Random and biased networks: Simulations and approximations.” Social Networks 7:225-261.