Resampling Methods for Two Mode Data

Lorien Jasny `ljasny@uci.edu`

Department of Sociology; University of California, Irvine

Modes in Network Data

Each ***mode*** is a distinct set of entities. Ties are defined as relationships between modes and not among.

- m sets of vertices $V_1, V_2, \ldots V_m$ with sizes $n_1, n_2, \ldots n_m$
- with edges defined as an ordered (directed) or unordered (undirected) pair on vertices (u, v) where $u \in V_a, v \in V_b$
- examples include people attending events, congress member voting roll calls, board members and corporate boards

Resampling Tests

A ***resampling test*** pits the empirical value of a test statistic against a permuted version of the original data (null model). In the network case, we usually condition on number of edges, the dyad census, or the degree distribution.

- useful in cases where we don't want to assume edge independence
- answers questions of the form: "given the network structure, is this feature significant with respect to the null model"

The Problem

Multi-mode resampling tests are extremely useful but infrequently used. While research into two-mode data is exploding, this work generally lacks rigorous statistical tests. We show how multi-mode resampling tests can be performed, how these tests can address a broader class of hypotheses than the one mode case, can result in different reference distributions than the one mode resampling test, and subsequently different conclusions.

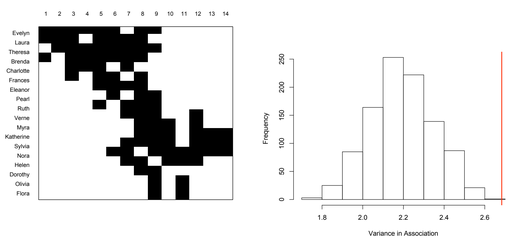
Example: a two mode hypothesis

- Example dataset: Davis, Gardner, Gardner 1941 Southern Women [2]
- Two mode hypothesis: is the number of women at each party (column marginal) and the number of parties each woman attends (row marginal) sufficient to explain the variation in association levels of the women?

Method

- The two mode resampling test conditions on the row and column marginals of the two mode structure.
- Permuting the structure by rotating 0-1 matrices controls for the marginals
- Done using an MCMC with samples taken at intervals the size of the two mode structure [3].
- Verified with the bipartite simulation routines in `statnet networksis` library [1].
- Two mode matrices converted to one mode to address hypothesis.

Results



(a) Raw Data (b) Permuting Two Mode Structure

- The red line shows the empirical value for the variance in the raw counts of parties each pair of women attended together
- The higher empirical variance shows that the women are more variegated in their dyadic association levels than expected when conditioning on the marginals.

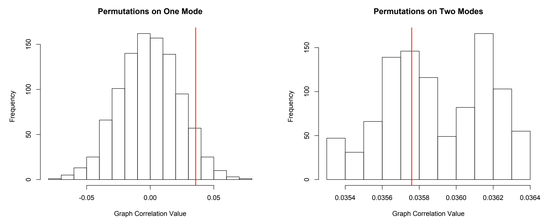
Example: differing results

- Comparing the structure of two graphs
- Example dataset: 2010 Congressional voting and donations to House members in the 2009-2010 election cycle
- Hypothesis: these structures are more similar than conditioning (a) on the structure of similarities in votes and donations (one mode) or (b) the row and column marginals of the voting and donation structures (two mode)

Method

- For the two mode test, permute one of the structures using the 01 matrix MCMC method
- Matrices too large for efficient use of `networksis` routines
- For the one mode test, permute one of the one mode matrices using the rmperm routine in SNA
- Calculate graph correlations on the one mode projection

Results



(c) One Mode Null Model (d) Two Mode Null Model

- The red line indicates empirical correlation
- Conditioning on the one mode projection shows the graphs are more correlated than expected
- Conditioning on the row and column marginals of the two mode structures shows that the level of correlation is not outside what would be expected under this null model.

Interpretation

Two mode resampling tests can answer useful questions in data analysis, and possibly should replace one mode tests in some cases resulting in very different conclusions. Where we consider the two mode marginal structure much more likely to occur than the one mode similarity structure, the two mode resampling test should be used.

Future Extensions

- Larger data (in progress)
  - RFID sensor data from a conference
  - Additional years of Congressional votes and donatiosn
  - Composition of Norwegian Corporate boards
- More complex hypotheses
  - Include more modes
  - Include more matrices – model temporal dynamics
- Faster Algorithms
  - Current algorithms for MCMC and bipartite simulation are fine for small networks but scale incredibly badly

References

[1] R. Admiraal, M. Handcock. `networksis`: A Package to Simulate Bipartite Graphs with Fixed Marginals Through Sequential Importance Sampling. *Journal of Statistical Software* **24**, (2008).

[2] S. Borgatti, M. Everett. Network analysis of 2-mode data. *Social Networks* **19**, (1997).

[3] A. Davis, B. Gardner, M. Gardner. *A Social Anthropological Study of Caste and Class*. Chicago: University of Chicago Press (1941).

[4] L. Hubert. *Assignment methods in combinatorial data analysis*. New York: Marcel Dekker, Inc. (1987).

[5] R. Rao, R. Jana, S. Bandyopadhyay. A Markov Chain Monte Carlo Method for Generating Random (0,1)-Matrices with Given Marginals. *The Indian Journal of Statistics, Series A* **58** (1996).

[6] T. Snijders. Enumeration and Simulation Methods for 0-1 Matrices with Given Marginals. *Psychometrika* **56**, (1991).

UCIrvine | Department of Sociology

NCASD Networks, Computation, and Social Dynamics

This material is based on research supported by the Office of Naval Research under award N00014-08-1-1015.

Scalable Methods for the Analysis of Network-Based Data