

Mr. Holland's Networks:
A Brief Review of the Importance of
Statistical Studies of Local Subgraphs

or
One Small Tune in a Large Opus

Stanley Wasserman^a

20 December 2008

Technical Report 08-08
Department of Statistics
Indiana University
Bloomington, IN 47408

^aDepartment of Statistics, and Department of Psychological and Brain Sciences, Indiana University. stanwass@indiana.edu

I enrolled in the Graduate School of Arts and Sciences at Harvard in Fall, 1973, to do graduate work in statistics. I had six classmates in my cohort, four of whom eventually received PhDs. One, Richard Hill, a recent graduate of MIT, was also an administrator at the Computer Research Center (CRC) of the National Bureau Economic Research (NBER). He worked there with his former classmate Mark Eisner, Technical Director of the CRC.

Paul Holland was full time at CRC, as a Senior Research Associate, from 1972–1975, while maintaining a lectureship in the Department of Statistics up the river. After my first year of studies, Richard and Paul were looking for a research assistant, to begin full-time for Summer 1974, and to continue part-time during the academic year. I was interested, and very pleased to be hired. So, my work with Paul began that summer, and continued for the next two years, while Paul and I were both in Cambridge. Paul moved to ETS late in 1975, so I moved on as well, spending my last year in graduate school (1976–1977) at Carnegie-Mellon University, working with his close collaborator at the time, Sam Leinhardt.

Paul directed my thesis, and was my mentor throughout the 1970’s. I owe much to him ... I valued his enthusiasm, enjoyed his humor, and was very grateful that much of what I did back in those early years was regarded by him as a “thing of beauty” (a standard which I never met again). I do regret that we were not more in contact over the last two decades.

I am grateful to be able to write a short note for his *Festschrift*, commenting on the importance of his research to the burgeoning field of network science.

1 Notation

We begin with a graph (or a directed graph), a single set of nodes \mathcal{N} , and a set of lines or arcs \mathcal{L} . It is common to use this mathematical concept to represent a *network*. We use the notation of Wasserman and Faust (1994), especially Chapters 13 and 14. There are extensions of these ideas to a wide range of networks, including multiple relations, affiliation relations, valued relations, and social influence and selection situations (in which information on attributes of the nodes is available); see the chapters of Carrington, Scott, and Wasserman (2005).

The purpose of this short exposition is to discuss the developments in statistical models for networks that have occurred over the past ten years and relate them to Paul’s early statistical network research. Background for much of this is summarized in the statistical chapters (8, 9, 10, and 11) of Carrington, Scott, and Wasserman (which were written in 2002). More of it can be found in the statistical physics literature, for example, the review paper of Newman (2003) or the edited volume of Newman, Barabasi, and Watts (2006). The statistical modeling of social networks is advancing quite quickly. The many exciting new developments include, for instance, longitudinal models for the co-evolution of networks and behavior (Snijders, Steglich & Schweinberger, in press) and latent space models for social networks (Hoff, Raftery & Handcock, 2002; Handcock,

Raftery, & Tantrum, 2007). Here, we review a few developments that are relevant to Paul’s work in the early 1970’s.

2 The Importance of Mr. Holland to Network Science

2.1 Some Past History

One of the most important structural theories in network analysis is *structural balance*, and its many derivatives. The history of structural balance, clusterability, and ranked clusterability began in network science in the 1940’s when a variety of mathematicians “invaded” the structural space occupied by the early sociometricians. The forefront of this research yielded a variety of theorems, rooted in graph theory, that allowed for checks on whether a particular graph was structurally balanced or clusterable. With these clusterability theorems in hand, a number of researchers embarked on empirical investigations. Questions such as how common clusterable signed (di)graphs are, and whether such signed (di)graphs were balanced, needed answers. These investigations required surveying many sociomatrices obtained from diverse sources. Further, the empirical studies had to be accompanied by statistical models that allowed those interested to study whether departures from theoretical models such as clusterability were “statistically large”.

The necessary statistical techniques are a bit too long and tedious for the scope of the current chapter. We give a few details below, and refer the reader to Chapter 14 of Wasserman and Faust (1994) for lots more information. But we can report here how the theorems of clusterability were generalized due to unexpected empirical evidence.

The standard Holland-Leinhardt index for clusterability or transitivity starts with the triad census, a vector of isomorphic triad counts, either of length 4 (for graphs) or 16 (for directed graphs). As usual, we let \mathbf{T} denote the triad census vector. Mathematically, we let \mathbf{l} be a weighting vector, designed to count the frequency of a particular structural tendency. Then, $\mathbf{l}'\mathbf{T}$ is a linear combination of the triad census, using one of the weighting vectors derived from the substantive hypothesis under study.

This linear combination is the number of times that the specific configuration, associated with the chosen weighting vector, occurs in the observed sociomatrix. Under one of the random directed graph distributions, we can calculate the expected value and covariance matrix of \mathbf{T} , and hence the expected number for this configuration and its variance. This expected number is $\mathbf{l}'\boldsymbol{\mu}_T$, and the standard error is $\sqrt{\mathbf{l}'\boldsymbol{\Sigma}_T\mathbf{l}}$, where $\boldsymbol{\mu}_T$ is the mean triad census vector, and $\boldsymbol{\Sigma}_T$ is the 16×16 (or 4×4) covariance matrix of the counts of the triad census.

This standardized index is then used as a test statistic for a variety of substantive null hypotheses. The first two moments of the linear combination of raw counts are calculated under these null hypotheses, which invariably assumes

















some particular random digraph distribution. From substantive hypothesis, to weighting vector, to test statistic, to a statistical evaluation very good science, popularized in the network science methodological toolkit by Holland and Leinhardt. And, they did the data analysis necessary to prove it was good science, as well.

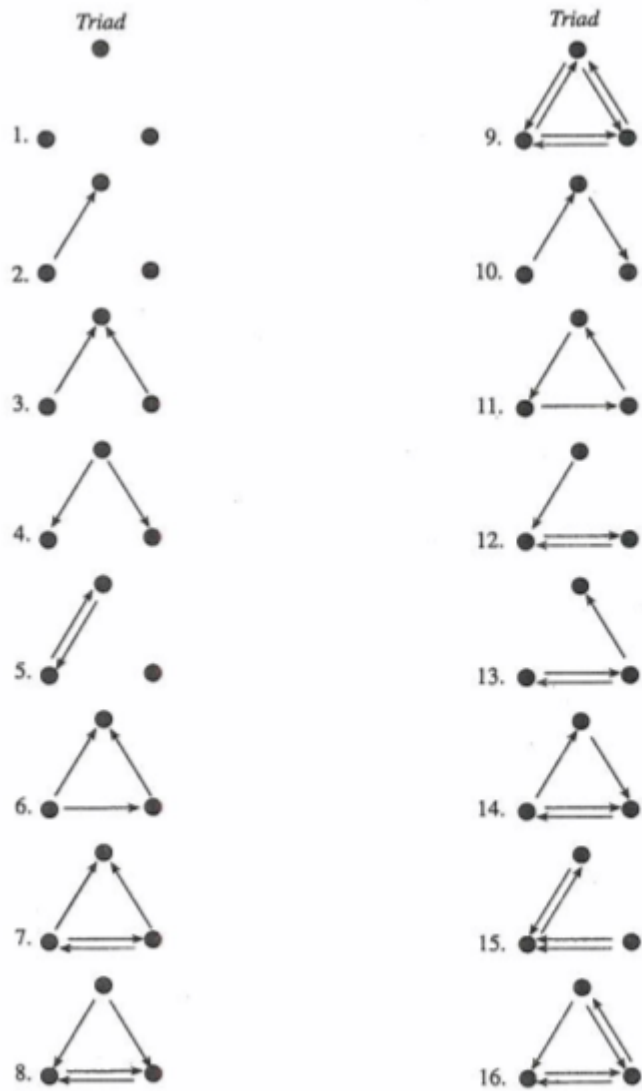
2.2 Empirical Evidence

Leinhardt (1968, 1973), Davis and Leinhardt (1968, 1972), and Davis (1970) gathered nearly 800 sociomatrices from many different sources, and discovered a few interesting facts. First, they found that many relations measured were directional. The old, recommended strategy of focusing on semicycles in such structures was difficult to implement. Secondly, asymmetric dyads, in which one actor chooses another actor, but the choice is not reciprocated, were very common. The ideas of balance and clusterability needed to be modified to take such situations into account (rather than ignoring the directionality of these arcs, which was the current practice when attention is focused on semicycles). Thirdly, they found that signed relations were rather rare. Thus, they decided to modify the theories of balance and clusterability so that they could be applied to signed directional relations. When even these new theories were later found lacking, Holland and Leinhardt (1971) revised them to unsigned directional relations.

Davis and Leinhardt also found that in some digraphs, one subset of actors chose a second, while actors in this second subset chose members of a third subset. The clusters of actors appeared to be *ranked*, or hierarchical in nature, with the actors “on the bottom” choosing those “at the top” (but not vice versa).

Holland and Leinhardt (1970) were the first to suggest the extension of these ideas to nonsigned directional relations. To turn ranked clusterability for complete signed digraphs into an equivalent idea for digraphs without signs is quite simple. We take the idea of ranked clusters for complete signed digraphs, and do not consider arcs with negative signs. Then, any arc with a sign of “-” is removed from the signed digraph. We then drop the positive signs from the remaining arcs. The assumption is that the relation under study is the “positive” part of the signed relation — for example, we study only “like”, “not like”, and “dislike”. Figure 1 shows the triples of Figure 2, without the negative arcs. The triples arising from directional relations are commonly referred to as *triads*, since we consider the threesome of nodes, and all the arcs between them.

1. 
2. 
3. 
4. 
5. 
6. 
7. 
8. 
9. 
10. 
11. 
12. 
13. 
14. 
15. 
16. 



We note that the two problematic triads from ranked clusterability found empirically to be quite common have one and five arcs. These are the triads

numbered 2 and 16 in Figure 1. Holland and Leinhardt showed that ranked clusterability is a special case of a more general set of theorems, which naturally blend balance, clusterability, and ranked clusterability. Their *partially-ordered clusterability* leads naturally to a consideration of the concept of *transitivity*.

Holland and Leinhardt (1971) reviewed the postulates of balance theory, clusterability, and ranked clusterability, as well as transitive tournaments (Landaу, 1951a, 1951b, and 1953; Hempel, 1952), and proposed the very general concept of transitivity to explain social structures. Transitivity includes all the earlier ideas as special cases. From a transitive digraph, one can obtain balanced, clusterable, and ranked clusterable graphs by making various assumptions about reciprocity and asymmetry of choices. During the past two decades, evidence has accumulated that transitivity is indeed a compelling force in the organization of social groups. What is even more remarkable, is that the idea was rediscovered, anew, by the physicists invading the network science world ten years ago. And now, transitivity and clusterability are very *hot*.

3 Some Current History

Early work on distributions for graphs was quite limited, forcing researchers to adopt independence assumptions that were not terribly realistic (see Chapters 13-16 of Wasserman and Faust, 1994). It is hard to accept the standard assumption common in much of the literature, especially in physics, of complete independence and then to adopt the mis-named and overly simplistic “random graph” distribution (there are, of course, an infinite number of random graph distributions). *The* random graph distribution to the physicists, that is usually referred to as a *Bernoulli graph*, assumes no dependencies at all among the random components of a graph. Equally hard to believe as a true representation of social behavior are the many conditional uniform distributions and p_1 , which assumes independent dyads (Holland and Leinhardt 1977, 1981).

The breakthrough in statistical modeling of networks was first expounded by Frank and Strauss (1986), who termed their model a *Markov random graph*. Further developments, especially commentary on estimation of distribution parameters, were given by Strauss and Ikeda (1990). Wasserman and Pattison (1996) elaborated upon the model, describing a more general family of distributions. Pattison and Wasserman (1999), Robins, Pattison, and Wasserman (1999), and Anderson, Wasserman, and Crouch (1999) further developed this family of models, showing how a Markov parametric assumption gives just one, of many, possible sets of parameters. This family, with its variety and extensions, was named p^* , a label by which it has come to be known. The parameters (which are determined by the hypothesized dependence structure) reflect structural concerns, which are assumed to be governing the probabilistic nature of the underlying social and/or behavioral process.

Work continues on this family, pointing out generalizations (Pattison and Robins 2002, Robins, Elliot, and Pattison 2001; Robins, Pattison, and Elliot 2001; Snijders, Pattison, Robins, and Handcock 2006), degeneracies (Handcock

2002), and new estimation strategies (Hunter 2007, Hunter and Handcock 2006, Snijders 2002).

The early work by the first researchers extended p^* in a variety of ways, and laid the foundation for work in this decade on the estimation problems inherent in the early formulations. This research also was an important forerunner of the new parametric specifications that promise wider usage of the family. A more thorough history of this family of distributions, including a discussion of its roots in spatial modeling and statistical physics, can be found in Borner, Sanyal, and Vespignani (2007). Wasserman and Robins (2005) offer a review of p^* circa-2003, while Robins, Pattison, Kalish, and Lusher (2006) review the 2003-2006 period. Other recent thoughts can be found in the May 2007 issue of *Social Networks*, a special issue devoted, in part, to p^* ,

The work of Frank and Strauss (1986) did indeed begin a new era for statistical modeling of networks, although it took ten years for Markov random graphs to be discussed at more length by network methodologists. What is remarkable is how the wheel keeps getting reinvented. Witness, the rebirth of Holland and Leinhardt's transitivity index, as we describe below

4 Clustering Coefficients

The clustering coefficient of a vertex in a graph quantifies how close the vertex and its neighbors are to being a clique (complete graph). Duncan Watts and Steven Strogatz introduced the measure in 1998 to determine whether a graph is a small-world network (hubs, which are part of cliques, but also adjacent to other hubs).

4.1 Formal Definition

A graph $G = (V, E)$ formally consists of a set of vertices V and a set of edges E between them. An edge e_{ij} connects vertex i with vertex j . The neighborhood N for a vertex v_i is defined as its immediately connected neighbors as follows:

$$N_i = \{v_j\} : e_{ij} \in E \text{ or } e_{ji} \in E. \quad (1)$$

The degree k_i of a vertex is defined as the number of vertices, $||N_i||$, in its neighborhood N_i . The clustering coefficient C_i for a vertex v_i is then given by the proportion of links between the vertices within its neighborhood divided by the number of links that could possibly exist between them. For a directed graph, e_{ij} is distinct from e_{ji} , and therefore for each neighborhood N_i there are $k_i(k_i - 1)$ links that could exist among the vertices within the neighborhood (k_i is the total (in + out) degree of the vertex). Thus, the clustering coefficient for directed graphs is given as

$$C_i = \frac{||\{e_{jk}\}||}{k_i(k_i - 1)} : v_j, v_k \in N_i, e_{jk} \in E. \quad (2)$$

An undirected graph has the property that e_{ij} and e_{ji} are equal by definition. Therefore, if a vertex v_i has k_i neighbors, edges could exist among the vertices within the neighborhood. Thus, the clustering coefficient for undirected graphs can be defined as

$$C_i = \frac{2|\{e_{jk}\}|}{k_i(k_i - 1)} : v_j, v_k \in N_i, e_{jk} \in E. \quad (3)$$

Let $\lambda_{G(v)}$ be the number of triangles on $v \in V(G)$ for undirected graph G . That is, $\lambda_{G(v)}$ is the number of subgraphs of G with 3 edges and 3 vertices, one of which is v . Let $\tau_{G(v)}$ be the number of triples on $v \in V(G)$. That is, $\tau_{G(v)}$ is the number of subgraphs (not necessarily induced) with 2 edges and 3 vertices, one of which is v and such that v is incident to both edges. Then we can also define the clustering coefficient as

$$C_i = \frac{\lambda_{G(v)}}{\tau_{G(v)}}. \quad (4)$$

It is simple to show that the two preceding definitions are the same, since

$$\tau_{G(v)} = \frac{1}{2}k_i(k_i - 1). \quad (5)$$

These measures are 1 if every neighbor connected to v_i is also connected to every other vertex within the neighborhood, and 0 if no vertex that is connected to v_i connects to any other vertex that is connected to v_i .

The clustering coefficient for the whole system is usually defined as the average of the clustering coefficient for each vertex:

$$\bar{C} = \frac{1}{n} \sum_{i=1}^n C_i. \quad (6)$$

4.2 Clustering Coefficients in Usage and Practice

In words The clustering coefficient \bar{C} is defined as follows: Suppose that a vertex v has k_v neighbors (or *alters*, a term usually used in the social network literature); then at most $k_v(k_v - 1)/2$ edges can exist between the alters (this occurs when every neighbor of v is connected to every other neighbor of v). Let C_v denote the fraction of the allowable edges that actually exist. Then the clustering coefficient \bar{C} is simply the average of C_v over all v .

This definition, introduced by Watts and Strogatz (1998), has been very important empirically, as featured in Duncan Watts' books (1999, 2003), and much research (see, for example, Robins, Pattison, and Woolcock 2005).

Besides being used over the past decade to check a nondirected graph for transitivity, it has been used extensively to study the small-world nature of a graph. From Matt Jackson's nice text (2008), a graph exhibiting the *small world property* has a small diameter and small average path length (as well illustrated in Watts, 1999). Quantitatively, a graph is considered small-world, if its average

clustering coefficient is significantly higher than a random graph constructed on the same vertex set, and if the graph has a small mean-shortest path length.

The small-world paradigm, introduced by Stan Milgram in the mid-1960's has stormed into our culture. Milgram's study, published in *Psychology Today*, showed that people in the United States seemed to be connected by approximately six acquaintanceship links, on average. From this finding, the notion of "six degrees of separation" was born. Milgram actually never used this now, very well-known phrase; the most likely popularizer of the term "Six Degrees of Separation" would be John Guare, whose Pulitzer Prize-winning and Tony Award play with the same name was published in 1990.

A generalization of the clustering coefficient to directed graphs is obvious and straightforward, thus bringing this idea directly in line with Holland and Leinhardt's τ index for transitivity. The only difference, of course, is the normalization for \bar{C} and the standardization (z -scoring) for τ . The similarity, was not noticed by Watts or Newman. Paul and Sam's research on this is not even mentioned by Matt Jackson.

The moral:

**Paul's work with Sam on indices for triads was replicated,
with very little attribution, by the current generation of
poorly educated physicists doing network science!**

Acknowledgements

This research was supported by a grant from the US Office of Naval Research (#N00014-02-1-0877). Ann McCranie graciously provided research and data analysis assistance.

References

- [1] Anderson, C.J., Wasserman, S., & Crouch, B. (1999). A p^* primer: Logit models for social networks. *Social Networks*. 21, 37–66.
- [2] Borner, K., Sanyal, S., & Vespignani, A. (2007). Network science: A theoretical and practical framework. In Blaise Cronin (Ed.), *Annual Review of Information Science & Technology*, Volume 4. Medford, NJ: Information Today, Inc./American Society for Information Science and Technology, 537–607.
- [3] Carrington, P.J., Scott, J., & Wasserman, S. (eds.) (2005). *Models and Methods in Social Network Analysis*. New York: Cambridge University Press.
- [4] Frank, O., & Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association*. 81, 832–842.

- [5] Handcock, M.S. (2002). Statistical models for social networks: Degeneracy and inference. In Breiger, R., Carley, K., & Pattison, P. (eds.). *Dynamic Social Network Modeling and Analysis* (pp. 229-240). Washington DC: National Academies Press.
- [6] Handcock, M.S., Raftery, A.E., & Tantrum, J.M. (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society, Series A. 170*, . 301 – 354 (with discussion).
- [7] Holland, P.W., & Leinhardt, S. (1977). Notes on the statistical analysis of social network data.
- [8] Holland, P. W., & Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association. 76*, 33–65 (with discussion).
- [9] Hoff, P., Raftery, A., & Handcock, M. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association, 97*, 1090-1098.
- [10] Hunter, D.R. (2007). Curved exponential family models for social networks. *Social Networks. 29*, 216–230.
- [11] Hunter, D. & Handcock, M. (2006). Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics. 15*, 565–583.
- [12] Jackson, M.O. (2008). *Social and Economic Networks*. Princeton: Princeton University Press.
- [13] Newman, M.E.J. (2003). The structure and function of complex networks. *SIAM Review. 45*, 167256.
- [14] Newman, M.E.J., Barabasi, A.-L., & Watts, D. J. (2006). *The Structure and Dynamics of Networks*. Princeton, NJ: Princeton U. Press.
- [15] Pattison, P. E., & Robins, G. L. (2002). neighborhood-based models for social networks. *Sociological Methodology. 32*, 301-337.
- [16] Pattison, P.E., & Wasserman, S. (1999). Logit models and logistic regressions for social networks: II. Multivariate relations. *British Journal of Mathematical and Statistical Psychology. 52*, 169–193.
- [17] Robins, G.L., Elliot, P., & Pattison, P.E. (2001). Network models for social selection processes. *Social Networks. 23*, 1–30.
- [18] Robins, G.L., Pattison, P.E., & Elliott, P. (2001). Network models for social influence processes. *Psychometrika. 66*, 161–190.

- [19] Robins, G.L., Pattison, P.E., Kalish, Y., & Lusher, D. (2007). An introduction to exponential random graph (p^*) models for social networks. *Social Networks*.
- [20] Robins, G.L., Pattison, P.E., & Wasserman, S. (1999). Logit models and logistic regressions for social networks, III. Valued relations. *Psychometrika*. 64, 371–394.
- [21] Robins, G.L., Pattison, P.E., & Woolcock, J. (2005). Social networks and small worlds. *American Journal of Sociology*. 110, 894–936.
- [22] Robins, G.L., Snijders, T.A.B., Wang, P., Handcock, M., & Pattison, P.E. (2007). Recent developments in exponential random graph (p^*) models for social networks. *Social Networks*.
- [23] Snijders, T.A.B. (2002). Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure*. 3, 2.
- [24] Snijders, T.A.B., Pattison, P.E., Robins, G.L., & Handcock, M. (2006). New specifications for exponential random graph models. *Sociological Methodology*.
- [25] Snijders, T.A.B., Steglich, C., & Schweinberger, M. (in press). Modeling the co-evolution of networks and behavior. To appear in van Monfort et al (Eds.), *Longitudinal Models in the Behavioral and Related Sciences*. New York: Erlbaum.
- [26] Strauss, D., & Ikeda, (1990). Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association*. 85, 204–212.
- [27] Wasserman, S., & Faust, K. (1994). *Social Network Analysis: Methods and Applications*. New York: Cambridge University Press.
- [28] Wasserman, S., & Pattison, P.E. (1996). Logit models and logistic regressions for social networks: I. An introduction to Markov random graphs and p^* . *Psychometrika*. 60, 401–426.
- [29] Wasserman, S., & Robins, G.L. (2005). An introduction to random graphs, dependence graphs, and p^* . In Carrington, P.J., Scott, J., and Wasserman, S. (eds.), *Models and Methods in Social Network Analysis*. New York: Cambridge University Press.
- [30] Watts, D.J. (1999). *Small Worlds: The Dynamics of Networks between Order and Randomness*. Princeton, N.J.: Princeton University Press.
- [31] Watts, D.J. (2003). *Six Degrees: The Science of a Connected Age*. New York: W.W. Norton and Company.
- [32] Watts, D.J. (2004). The “new” science of networks. *Annual Review of Sociology*. 30, 240–270.

- [33] Watts, D.J., & Strogatz, S.H. (1998). Collective dynamics of 'small-world' networks. *Nature*. 393, 440–442.