NBER WORKING PAPER SERIES

METHODS OF IDENTIFICATION IN SOCIAL NETWORKS

Bryan S. Graham

Working Paper 20414
http://www.nber.org/papers/w20414

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 2014

Forthcoming in Annual Review of Economics Doi: 10.1146/annurev-economics-080614-115611.
I thank Guido Imbens for reading an initial draft and Joachim De Weerdt for generously sharing his Nyakatoke network data. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

Methods of Identification in Social Networks
Bryan S. Graham
NBER Working Paper No. 20414
August 2014
JEL No. C23,C25,D85

ABSTRACT

Social and economic networks are ubiquitous, serving as contexts for job search, technology diffusion, the accumulation of human capital and even the formulation of norms and values. The systematic empirical study of network formation – the process by which agents form, maintain and dissolve links – within economics is recent, is associated with extraordinarily challenging modeling and identification issues, and is an area of exciting new developments, with many open questions. This article reviews prominent research on the empirical analysis of network formation, with an emphasis on contributions made by economists.

Bryan S. Graham
University of California - Berkeley
530 Evans Hall #3880
Berkeley, CA 94720-3880
and NBER
bgraham@econ.berkeley.edu

Job-seekers often receive help from family and acquaintances when conducting searches (e.g., Loury, 2006). Likewise individuals learn about new products and technologies from friends and colleagues (e.g., Banerjee, Chandrasekhar, Duflo and Jackson, 2013). The actions and attributes of an adolescent's peer group predict her initiation of sexual activity, drug use and academic performance among other behaviors (Case and Katz, 1991; Gaviria and Raphael, 2001). Even the exchange of goods and services may occur within a network. For example, electronic producers may utilize different, but overlapping, sets of manufacturers to assemble finished products, sharing valuable technology and know-how with each (e.g., Kranton and Minehart, 2001).

The ubiquitousness of networks, along with their ability to predict many social and economic behaviors, motivates their academic study. In particular, the correlation between the actions of individuals (firms) and the attributes and actions of those with whom they are connected raises at least two questions. First, how do networks form and evolve? Second, do the actions and attributes of one's peers – the set of agents to which one is connected – influence one's own actions? There is ample evidence that measurable features of an individual's social network predict important social and economic outcomes. Whether these associations are causal is unclear.

The processes of network formation and network influence are interconnected. If one's network of friends and acquaintances facilitates job search, it may be that individual's seek out connections, at least in part, because of these benefits. In such situations unobserved drivers of network structure may covary with those of the outcome of interest, rendering any observed relationship between network structure and outcomes, at least partly, spurious. Biases of these type are common in many areas of microeconometric research, addressing them in the context of social network research is difficult.

The interconnected nature of network formation and their effects on outcomes of interest to decision-makers has important policy implications. If, for example, an adolescent's rate of learning is meaningfully influenced by the characteristics and behaviors of her friends, then

interventions acting on an existing network of friendships, as well as those designed to change the structure of friendships, may both influence observed achievement (e.g., Goldsmith-Pinkham and Imbens, 2013).

While social network analysis has a long history in several disciplines, especially sociology (cf., Wasserman and Faust, 1994; Bonacich and Lu, 2012), the study of networks in economics, both theoretically and, even more so, empirically is considerably more recent.

Within economics empirical research on social networks generally concerns itself with one of the two questions posed above (but rarely both). Manski (1993), in a seminal paper, studied the second question: under what conditions can a researcher infer that the actions and/or characteristics of one's peers influence one's own actions? Manski's paper has been extraordinarily influential; both sparking further methodological research (e.g., Brock and Durlauf 2001a,b; Graham, 2008, 2011; Bramoulle, Djebbari and Fortin, 2009) as well as deeply influencing empirical practice (e.g., Angrist and Lang, 2004; Card and Rothstein, 2007; Sacerdote, 2014).

Empirical research on the first question – how do networks form? – is comparatively more recent. Jackson and Wolisky (1996) introduced the notion of a strategic model of network formation, where pairs of agents form, maintain or sever links in a decentralized way in order to maximize utility. Choices are interdependent, since the utility an agent attaches to a particular link may vary with the presence or absence of other links in the network. This approach to network formation, with agents maximizing utility in a decentralized way, is a natural one for economists. Formulating an empirical model with these features is difficult.

Since McFadden (1973) and Manski (1975) economists have modeled single agent discrete choice problems using random utility models (RUMs). These models provide a principled way of inferring the distribution of preferences from the observed distribution of choice. Unfortunately, as is familiar from the literature on games (e.g., Bresnahan and Reiss, 1991; Tamer, 2003), when agents' choices are interdependent, as is the case in network formation, a number of econometric challenges arise. These challenges are compounded by the scale of

the network formation problem. In an undirected network with N agents, a total of $2^{\binom{N}{2}}$ configurations of links are possible.

This review will primarily focus on empirical models of network formation (i.e, on the first of the two questions posed above). The emphasis will further be on models of strategic network formation (i.e., those with coherent random utility foundations). Other approaches to network modeling, such as exponential random graph models (ERGMs), will only be touched upon.

Research on the identification of peer groups effects (i.e, on the second of the two questions posed above) is comparatively more mature than that on how networks (or peer groups) form. Several high quality reviews of work in this area are currently available (e.g., Blume, Brock, Durlauf and Ioannides, 2011). Work in this area will not feature prominently in this review, except insofar as it connects to questions of network formation.

The partition of research on networks into (i) the study of agent behavior conditional on friendship (i.e., network) structure and (ii) the study of how networks form is convenient but not intellectually desirable. In single agent problems the analogs of these two questions are generally considered jointly. For example, labor economist study the distribution of wages jointly with schooling and/or labor force participation decisions (e.g., Heckman 1977, Card, 1995).

Brock and Durlauf (2001b) discuss the joint modeling of neighborhood choice and neighborhood influences. Ioannides and Zabel (2008) provide an empirical illustration. Goldsmith-Pinkham and Imbens (2013) jointly model peer selection and peer influences on academic achievement. Their paper also includes a fully worked empirical application. Work of this type, however, remains very much an exception. Some of the strongest empirical work on peer group effects exploits research designs where group membership is (quasi-) randomly determined (see Sacerdote (2014) for a review). While the study of social interactions in settings where peer networks are plausibly exogenous is attractive, limiting research to such settings substantially narrows the set of behaviors amenable to study.

Clearly econometric models of strategic network formation are needed to successfully "control for" network endogeneity when studying peer group effects. While their use in this way is largely aspirational at the present time, this goal motivates their development. Of course the study of network formation is scientifically interesting in its own right and, further, policy-makers may legitimately have preferences over the structure of links in a network, irrespective of any relationship between these links and other outcomes. For example, a school administrator may wish to structure instruction in her school so as to encourage cross-social-class and/or cross-race friendships, reduce the presence of cliques and/or isolated students and so on. She may wish to do this irrespective of any relationship between students' network structures and their academic performance.

Section 1, which follows next, describes methods for summarizing network data. Just as analysis of the distribution of a single random variable typically begins with the calculation of a sample mean, or one on the association between two random variables with that of a correlation coefficient, the analysis of network data generally begins with a summary of various features of a network's architecture. This material also serves as a vehicle to establish some basic notation and to review some 'stylized facts' on social networks.

Section 2 selectively reviews empirical models of network formation. Here I focus on strategic models of network formation. I exclude from my discussion so called "network evolution models" (NEMs) (Toivonen et al. , 2009). Examples of NEMs include the scale free model of Barabási and Albert (1999) and the small world model of Watts and Strogatz (1998). NEMs typically begin with a small randomly-generated seed network and then specify simple stochastic rules for adding and deleting links. This literature focuses on replicating key features of real world networks (e.g., degree distribution, average path length etc.). I also exclude exponential random graph models (ERGMs) from my discussion. This approach directly specifies a probability distribution for the entire graph/network. While EGRMs have featured prominently in the empirical literature on networks, they are generally not consistent under sampling (Shalizi and Rinaldo, 2013). Specifically the parameters associated with an

ERGM fit to a sub-network do not coincide with those associated with the full network. This makes structural interpretation of model parameters problematic (Christakis, Fowler, Imbens and Kalyanaraman, 2010). Relatedly, estimation of EGRMs is also challenging (Snijders, 2002). In an interesting working paper, Chandrasekhar and Jackson (2014) propose an alternative to ERGMs that addresses many of these limitations.

Section 3 discusses methods of network simulation. The inclusion of this material is motivated by the need to have interesting but tractable null models for network data. Such models can serve a number of purposes. First, they can be used to assess whether a certain feature of a network is unusual among the set of all networks that share certain other features in common. Second, some data collection protocols may only partially reveal network structure. For example public health surveys may collect information on the number of concurrent sexual partners, but not their identity (Morris, Kurth, Hamilton, Moody and Wakefield, 2009). The General Social Survey (GSS) collects ego-centered network data (Burt, 1984; McPherson, Smith-Lovin and Brashears, 2006). Specifically information on each GSS respondent's direct links are collected, as well as information on any links among those links. This sampling scheme reveals the network's degree distribution (i.e., distribution number of links across agents), as well as the frequency with which one's friends are friends themselves. Simulation methods can allow researchers to study the properties of the class of all networks that are consistent with the sample information available (which may not be the complete network). Finally, constructing a binary matrix which satisfies a set of side constraints is a well-defined discrete math problem. In recent years computer scientists have studied these types of problems (e.g., Blitzstein and Diaconis, 2011; Stanton and Pinar, 2012). Familiarity with this work is valuable to researchers interested in networks.

Section 4 ends with some thoughts about future directions for research.

1 Describing networks

Figure 1 provides a visual representation of a set of risk-sharing links, measured in the year 2000, between 119 households residing in Nyakatoke, a small village in Tanzania. These data are described and analyzed by de Weerdt (2004) and de Weerdt and Fafchamps (2011). Individuals were asked for lists of people that they could "personally rely on for help". A list of undirected links between all households was constructed using responses to this question. Each point in the figure represents a household, with the size of the point proportional to the number of risk sharing links to which the household is party. Yellow, orange, green and blue households correspond to categories of increasing land and livestock wealth (see the notes to Figure 1).

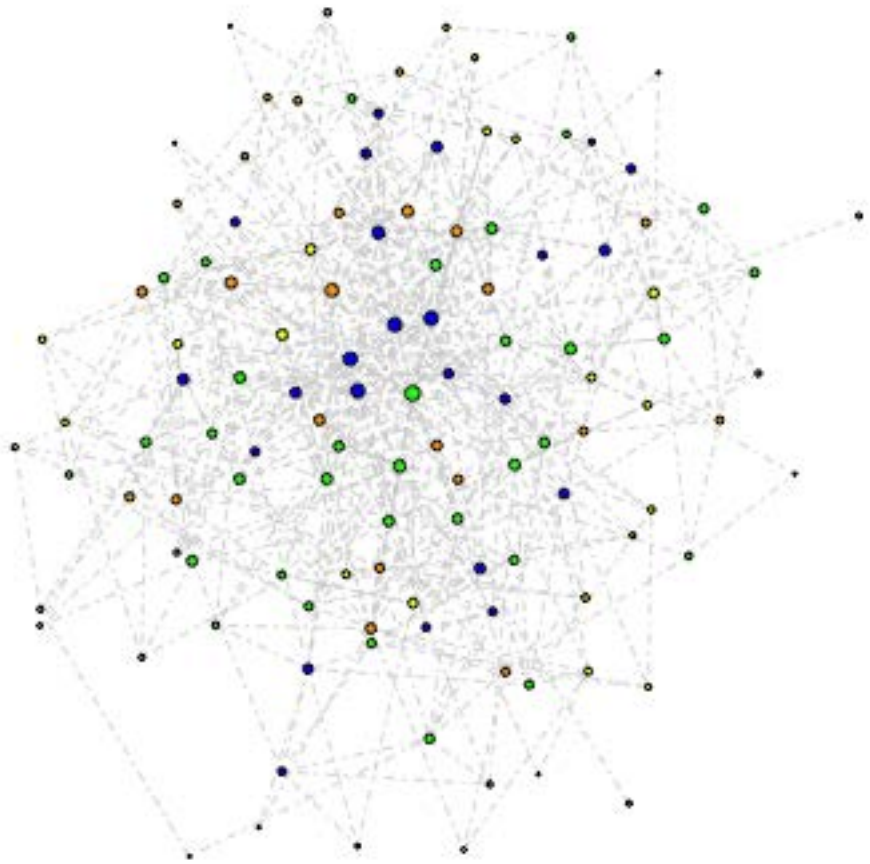
Graphical representations of network data like Figure 1 have historically played an important role in empirical analysis and continue to do so (Freeman, 2000). While certain features of a network can often be intuited from a visual representation, it is also valuable to have a suite of standard network summary statistics. This section describes methods for summarizing network data. There are many basic references for the material surveyed here, including Wasserman and Faust (1994), Newman (2003), Jackson (2008) and Kolaczyk (2009). A few minor results presented below, mostly of pedagogical significance, are new.

The mathematical language of networks is that of discrete math and, specifically, graph theory. Bollobas (2013) is a standard graph theory reference. Rosen (2006) is a more introductory discrete math reference. An undirected graph $G(\mathcal{N}, \mathcal{E})$ consists of a set of nodes $\mathcal{N} = \{1, \ldots, N\}$ and a list of unordered pairs of nodes called edges $\mathcal{E} = \{\{i, j\}, \{k, l\}, \ldots\}$ for $i, j, k, l \in \mathcal{N}$. A graph is conveniently represented by its adjacency matrix $\mathbf{D} = [D_{ij}]$ where

$$D_{ij} = \begin{cases} 1 & \text{if } \{i, j\} \in \mathcal{E} \\ 0 & \text{otherwise} \end{cases}. \tag{1}$$

A node, depending on the context, may be called a vertex, agent or player. Likewise edges

Figure 1: Nyakatoke risk-sharing network



Source: de Weerdt (2004) and author's calculations.
Notes: Node size proportional to household degree. Yellow nodes represent households with land and livestock wealth below 150,000 Tanzanian Shillings, orange those with wealth between 150,000 and 300,000 Shillings, green those with wealth between 300,000 and 600,000 Shillings and blue those with wealth of 600,000 Shillings and above. Following Comola and Fafchamps (forthcoming) land was valued at 300,000 shillings per acre. Network plotted using igraph package in R (see `http://igraph.org/r/`).

may be called links, friendships, connections or ties. Since self-ties are ruled-out, and the nodes in edges are unordered, the adjacency matrix is a symmetric binary matrix with a diagonal of so-called structural zeros (i.e., $D_{ij} = D_{ji}$ and $D_{ii} = 0$).

Networks may also be directed, such that each link has an ego (sender) and alter (receiver) ordering. Indeed, the analysis of directed networks in sociology probably predominates. The focus on undirected networks here is soley for pedagogical reasons.

A social network consists of a set of agents (nodes) and ties (edges) between them. A social network can be conveniently represented by it node and edge list or by its adjacency matrix. I will utilize the adjacency matrix representation in most of what follows. Two examples of undirected network adjacency matrices are

$$\mathbf{D}_{\text{ex1}} = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}, \mathbf{D}_{\text{ex2}} = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{pmatrix}.$$

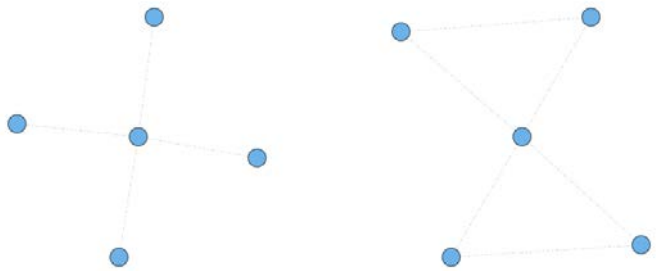
These two networks are graphically depicted in Figure 2. The first network, $\mathbf{D}_{\text{ex1}}$, takes a so called 'star' configuration, in which a central agent linked to all other agents. The second network, $\mathbf{D}_{\text{ex2}}$, consists of two triangles, which share a single agent in common.

In summarizing the structure of a social network it is convenient to define network statistics at the level of individual agents, at the level of pairs of agents or dyads, and at the level of triples of agents or triads.

Network statistics involving single agents and paths through the network

The total number of links belonging to agent i, or her degree is $D_{i+} = \sum_j D_{ij}$. The degree frequency distribution of a network, or degree distribution for short, consists of the frequency

Figure 2: Two simple networks



of each possible agent-level degree count $\{0, 1, \ldots, N\}$ in the network. A important component of the literature on networks takes the degree distribution as its primitive object of interest (e.g., Barabási and Albert (1999) and Albert and Barabási (2002)). This focus is motivated by the fact that many other topological features of a network are fundamentally constrained by its degree distribution (see Faust, 2007). I will have more to stay about the connection between a network's degree sequence and its other topological features below.

The density of a network equals the frequency with which any randomly drawn dyad is linked:

$$P_N = \binom{N}{2}^{-1} \sum_{i=1}^{N} \sum_{j<i} D_{ij}. \tag{2}$$

Note that $(N-1)\,P_N$ coincides with average degree. The density of the Nyakatoke network is 0.0698. The density of $\mathbf{D}_{ex1}$ is 0.4, that of $\mathbf{D}_{ex2}$ is 0.6.

Consider the matrix product

$$\mathbf{D}^2 = \begin{pmatrix} D_{1+} & \sum_i D_{1i}D_{2i} & \cdots & \sum_i D_{1i}D_{Ni} \\ \sum_i D_{1i}D_{2i} & D_{2+} & \cdots & \sum_i D_{2i}D_{Ni} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_i D_{1i}D_{Ni} & \sum_i D_{2i}D_{Ni} & \cdots & D_{N+} \end{pmatrix}.$$

The i^{th} diagonal element of $\mathbf{D}^2$ equals the number of agent i's links or her degree**.** The

$\{i,j\}^{th}$ element of $\mathbf{D}^2$ gives the number of links agent i has in common with agent j (i.e., the number of "friends in common"). In the language of graph theory the $\{i,j\}^{th}$ element of $\mathbf{D}^2$ gives the number of paths of length two from agent i to agent j. For example, if i and j share the common friend k, then a length two path from i to j is given by $i \to k \to j$. The diagonal elements of $\mathbf{D}^2$ correspond to the number of length two paths from an agent back to herself. For example if i is connected to k, then one such a path is $i \to k \to i$. The number of such paths coincides with an agent's degree.

Calculating $\mathbf{D}^3$ yields

$$\mathbf{D}^3 = \begin{pmatrix} \sum_{i,j} D_{1i}D_{ij}D_{j1} & \sum_{i,j} D_{1i}D_{ij}D_{j2} & \cdots & \sum_{i,j} D_{1i}D_{ij}D_{jN} \\ \sum_{i,j} D_{1i}D_{ij}D_{j2} & \sum_{i,j} D_{2i}D_{ij}D_{j2} & \cdots & \sum_{i,j} D_{2i}D_{ij}D_{jN} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i,j} D_{1i}D_{ij}D_{jN} & \sum_{i,j} D_{2i}D_{ij}D_{jN} & & \sum_{i,j} D_{Ni}D_{ij}D_{jN} \end{pmatrix},$$

whose $\{i,j\}^{th}$ element gives the number of paths of length 3 from i to j.

The diagonal elements of $\mathbf{D}^3$ are counts of the number of transitive triads or triangles in the network. If both i and j are connected to k as well as to each other, then the $\{i,j,k\}$ triad is closed (i.e., "the friend of my friend is also my friend"). Note that if $\{i,j,k\}$ is a closed triad it is counted twice each in the i^{th}, j^{th} and k^{th} diagonal elements of $\mathbf{D}^3$. Therefore $\mathrm{Tr}\left(\mathbf{D}^3\right)/6$ equals the number of unique triangles in the network.

Proceeding inductively it is easy to show that the $\{i,j\}^{th}$ element of $\mathbf{D}^K$ gives the number of paths of length K from agent i to agent j.

Measuring agent centrality

One preoccupation of network researchers has been the identification of "central" or "important" agents. This is sometimes called the key player problem (Ballester, Calvó-Armengol and Zenou, 2006; Ballester and Zenou, forthcoming). Consider, as just one example, a policy-

maker who wishes to facilitate the spread of a new technology or idea. She is constrained to introduce the technology to just one agent in the network. Which agent should she choose? The answer to this question will depend, of course, on assumptions about how information moves from one agent to another, as well as the precise objective of the policy-maker. Banerjee, Chandrasekhar, Duflo and Jackson (2013) use the above question to motivate a specific measure of centrality.

Many measures of agent centrality have been proposed (see Wasserman and Faust (1994), Jackson (2008) and Bonacich and Lu (2012) for textbook expositions of some key examples). Centrality measures are often heuristically motivated, but have been, in some cases, micro-founded ex post (e.g., Ballester, Calvó-Armengol and Zenou, 2006).

In turns out that one notion of agent centrality can be defined via an interesting connection with the "social multiplier". The concept of a social multiplier has been a key theme in empirical work on social interactions in economics since the publication of Manski (1993). It features in, for example, Brock and Durlauf, (2001b), Glaeser and Scheinkman (2001, 2003), Graham (2008) and Angrist (forthcoming). In the presence of social multiplier effects, the full impact of an intervention exceeds the initial impact due to feedback effects across agents. When interactions occur on a non-trivial network, the magnitude of any multiplier effect will also depend upon exactly which agent is initially acted upon by the policy-maker. This is the intuition behind social multiplier centrality.

Let Y_i be some continuously-valued action chosen by network member $i = 1, \ldots, N$. Let $\mathbf{Y}$ be $N \times 1$ vector of all agents actions. Let ι_N be an $N \times 1$ vector of ones and $\mathbf{G} = \text{diag}\left(\mathbf{D}\iota_{\text{N}}\right)^{-1}\mathbf{D}$ be the row-normalized network adjacency matrix (i.e., the network adjacency matrix where each element of the i^{th} row is divided by D_{i+}, the i^{th} agent's degree). Note that all rows of this matrix sum to 1 by construction. The matrix is row-stochastic.

Let

$$\mathbf{G}_i\mathbf{y} = \sum_{j \neq i} G_{ij} y_j \overset{def}{\equiv} \bar{y}_{n(i)}$$

equal the average action of player i's peers under the (perhaps hypothetical) action profile $\mathbf{y}$. Here $\mathbf{G}_i$ denotes the i^{th} row of $\mathbf{G}$.

Following Blume, Brock, Durlauf and Jayaraman (2013), among others, assume that the utility agent i receives from action profile $\mathbf{y}$ given network structure ($\mathbf{D}$) is

$$\begin{aligned} u_i(\mathbf{y};\mathbf{D}) &= (\alpha_0 + U_i)\, y_i - \frac{1}{2} y_i^2 + \beta_0 \bar{y}_{n(i)} y_i \\ &= (\alpha_0 + U_i)\, y_i - \frac{1}{2} y_i^2 + \beta_0 \mathbf{G}_i \mathbf{y} y_i \end{aligned} \tag{3}$$

with $0 < |\beta_0| < 1$ and $\mathbb{E}[U_i] = 0$. Here U_i captures heterogeneity in agents' preferences for action.

The marginal utility associated with an increase in y_i is increasing in the average action of one's peers, $\bar{y}_{n(i)}$. Specifically,

$$\frac{\partial^2 u_i(\mathbf{y},\mathbf{D})}{\partial y_i \partial \bar{y}_{n(i)}} = \beta_0.$$

That is, own- and peer-effort are complements. In the terminology of Manski (1993), the magnitude of β_0 indexes the strength of any endogenous social interactions.

Assume that the observed action $\mathbf{Y}$ corresponds to a Nash equilibrium where no agent can increase her utility by changing her action given the actions of all other agents in the network. Agents observe $\mathbf{D}$ and $\mathbf{U}$, the $N \times 1$ vector of individual-level heterogeneity terms.

The first order condition for optimal behavior associated with (3) generates the following best response function:

$$Y_i = \alpha_0 + \beta_0 \bar{Y}_{n(i)} + U_i \tag{4}$$

for $i = 1, \ldots, N$. Equation (4) is a special case of what is called the linear-in-means model of social interactions (e.g., Brock and Durlauf, 2001b). An agent's best reply varies with the average action of those to whom she is directly connected ($\bar{Y}_{n(i)}$) and unobserved own attributes (U_i).

Equation (4) defines an $N \times 1$ system of simultaneous equations. It is convenient, for what

follows, to write the system defined by (4) in matrix form:

$$\mathbf{Y} \;=\; \alpha_0 \iota_N + \beta_0 \mathbf{G}\mathbf{Y} + \mathbf{U}. \tag{5}$$

For $|\beta_0| < 1$ the matrix $I_N - \beta_0 \mathbf{G}$ is strictly (row) diagonally dominant (I_N is the $N \times N$ identity matrix). By the Levy-Desplanques Theorem (cf., Horn and Johnson, 2013) it is therefore non-singular. Non-singularity of $(I_N - \beta_0 \mathbf{G})$ allows us to solve for the equilibrium action vector as a function of $\mathbf{D}$ and $\mathbf{U}$ alone.

Solving (5) for $\mathbf{Y}$ yields the reduced form

$$\mathbf{Y} = \alpha_0 \left(I_N - \beta_0 \mathbf{G}\right)^{-1} \iota_N + \left(I_N - \beta_0 \mathbf{G}\right)^{-1} \mathbf{U}. \tag{6}$$

It is helpful to simplify (6) in a number of ways. First, using the series expansion

$$\left(I_N - \beta_0 \mathbf{G}\right)^{-1} = \sum_{k=0}^{\infty} \beta_0^k \mathbf{G}^k,$$

as well as the fact that $\mathbf{G}\iota_N = \iota_N$ (and hence that $\mathbf{G}^k \iota_N = \iota_N$ for $k \geq 1$) we get the simplification $\alpha_0 \left(I_N - \beta_0 \mathbf{G}\right)^{-1} \iota_N = \alpha_0 \left(1 - \beta_0\right)^{-1} \iota_N$. Using this result and re-arranging (6) yields

$$\mathbf{Y} \;=\; \frac{\alpha_0}{1 - \beta_0} \iota_N + \left[\sum_{k=0}^{\infty} \beta_0^k \mathbf{G}^k\right] \mathbf{U}. \tag{7}$$

Equation (7) provides some insight into what various researchers have called the social multiplier. Consider a policy which increases the i^{th} agent's value of U_i by Δ. We can conceptualize the full effect of this increase on the network's distribution of outcomes as occurring in "waves". In the initial wave only agent i's outcome increases. The change in the entire action vector is therefore

$$\triangle \mathbf{c}_i,$$

where $\mathbf{c}_i$ is an N-vector with a one in its i^{th} element and zeros elsewhere.

In the second wave all of agent i's friends experience outcome increases. This is because their best reply actions change in response to the increase in agent $i's$ action in the initial wave. The action vector in wave two therefore changes by

$$\triangle\beta_0\mathbf{G}\mathbf{c}_i.$$

In the third wave the outcomes of agent i's friends' friends change (this includes a direct feedback effect back onto agent i). In wave three we get a further change in the action vector of

$$\Delta\beta_0^2\mathbf{G}^2\mathbf{c}_i.$$

In the k^{th} wave we have a change in the action vector of

$$\triangle\beta_0^{k-1}\mathbf{G}^{k-1}\mathbf{c}_i.$$

Observing the pattern of geometric decay we see that the "long-run" or full effect of a Δ change in U_i on the entire distribution of outcomes is given by

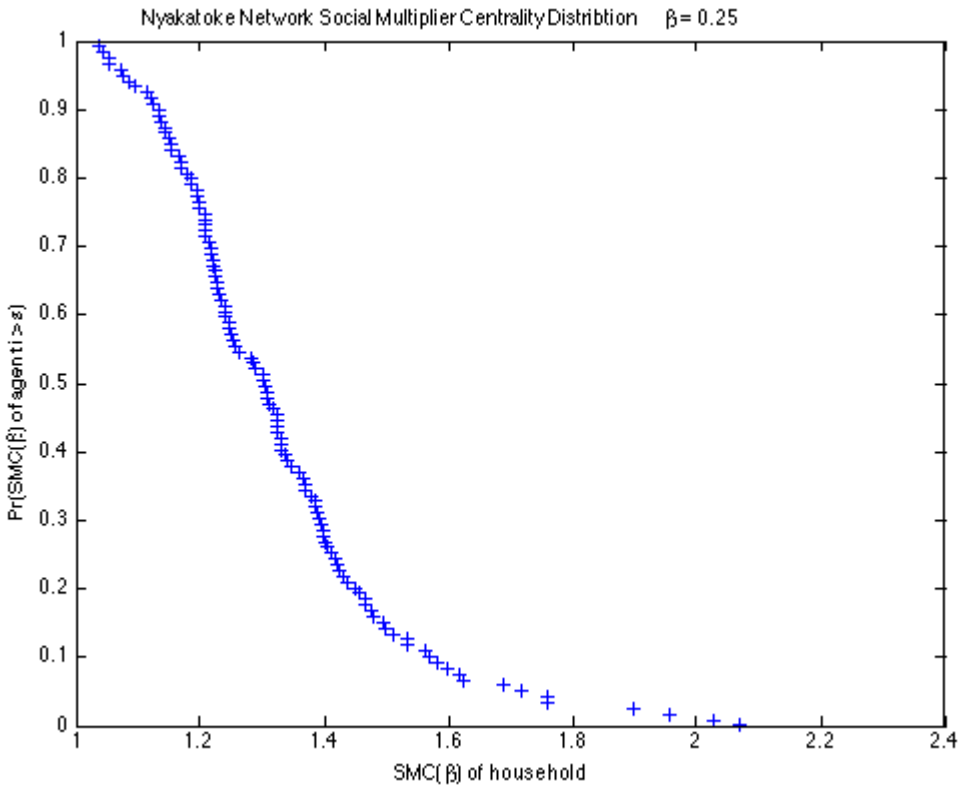
$$\triangle\left(I_N-\beta_0\mathbf{G}\right)^{-1}\mathbf{c}_i. \tag{8}$$

Equation (8) indicates the effect of perturbing U_i by $\triangle$ on the equilibrium action vector coincides with the i^{th} column of the matrix $\triangle\left(I_N-\beta_0\mathbf{G}\right)^{-1}$. The total effect on aggregate action is therefore given by the sum of the i^{th} column of this matrix. Therefore

$$\text{SMC}\left(\beta_0\right)=\iota_N'\left(I_N-\beta_0\mathbf{G}\right)^{-1}$$

gives a row vector of aggregate effects associated with unit perturbations of U_i for each agent in the network. If the cost of perturbing U_i does not vary with i, the planner can use

Figure 3: Social multiplier centrality at $\beta = 0.25$ in the Nyakatoke network



Source: de Weerdt (2004) and author's calculations.

this vector to efficiently target interventions. Call the i^{th} element of this vector agent $i's$ *social multiplier centrality* at parameter β_0. Specifically if the planner seeks to maximize the average action, she will target her intervention toward an agent with high social multiplier centrality.

Bonacich (1987) develops a closely related, and widely used measure of centrality, Katz-Bonacich Centrality. Ballester, Calvó-Armengol and Zenou (2006) provide a game-theoretic foundation for Katz-Bonacich Centrality. Blume, Brock, Durlauf and Jayaraman (2013) discuss the relationship between the linear-in-means interaction game exposited above and the one studied by Ballester, Calvó-Armengol and Zenou (2006). See also Jackson and Zenou (forthcoming) and Ballester and Zenou (forthcoming).

Figure 3 plots the distribution of $\text{SMC}(\beta) = \iota_N' (I_N - \beta \mathbf{G})^{-1}$ at $\beta = 0.25$ for the Nyakatoke

risk-sharing network. Under linear-in-means interaction, an intervention initiated on many households in the Nyakatoke Network would barely increase the aggregate action beyond the direct effect on the intervened household. For the most central households, however, the full effect on the aggregate outcome would exceed twice that of the "initial" (direct) effect on the intervened agent. In the presence of linear-in-means interaction, knowledge of network structure can lead to substantial improvements in policy-targeting.

Network statistics involving pairs of agents or dyads

The distance between agents i and j corresponds to the minimum length path connecting them. If there is no path connecting i to j, then the distance between them is infinite. We can use powers of the adjacency matrix to calculate these distances. Specifically,

$$M_{ij} = \min_{k \in \{1,2,3,\ldots\}} \left\{ k : D_{ij}^{(k)} > 0 \right\}$$

equals the distance from i to j (if it is finite). Here $D_{ij}^{(k)}$ denotes the ij^{th} element of $\mathbf{D}^k$. For modest sized networks M_{ij} can be calculated by taking successive powers of the adjacency matrix.

If the network consists of a single, giant, connected component, such that the minimum length path between any two agents is finite, we can compute average path length as

$$\overline{M} = \binom{N}{2}^{-1} \sum_{i=1}^{N} \sum_{j<i} M_{ij}. \tag{9}$$

If the network consists of multiple connected components, standard practice is to compute average path length within the largest one. Alternatively, following Newman (2003), we can calculate average distance or path length in the network as

$$\overline{M}^{\text{alt}} = \left[\binom{N}{2}^{-1} \sum_{i=1}^{N} \sum_{j<i} M_{ij}^{-1} \right]^{-1}. \tag{10}$$

Table 1: Frequency of degrees of separation in the Nyakatoke network

	1	2	3	4	5
Count	490	2666	3298	557	10
Frequency	0.0698	0.3797	0.4697	0.0793	0.0014

Source: de Weerdt (2004) and author's calculations.

By taking the reciprocal of an average of reciprocal distances, we neatly handle the 'problem' of infinite paths.

The diameter of a network is the largest distance between two agents in it. It will be finite if the network consists of a single connected component (in which case all agents are "reachable" starting from any given agent) and infinite in networks consisting of multiple components (in which case there are no paths connecting some pairs of agents).

Table 1 gives the frequency of minimum path lengths in the Nykatoke network. There are 490 direct ties in the network (paths of length one). Just under 7 percent of all *pairs* of households are directly connected in Nykatoke. Another 2,666 dyads are only two degrees apart. That is, although they are not connected directly, they share a tie in common. About 80 percent of dyads are separated by three or fewer degrees. The diameter of the Nyakatoke network is 5. The juxtaposition of low density (i.e., only a small fraction of all possible ties exists), with few degrees of separation (i.e., low average degree and/or diameter) is a feature of many real world social networks.

The analysis of distances and diameter has a long history in social network analysis and falls under the rubric of the "small-world problem". Stanley Milgram (1967) popularized this phrase and, through a series of postal experiments in the 1960s, showed that two random individuals in the United States could be often be connected through a short chain of acquaintances ("six degrees of separation").

Figure 4: Types of triads in undirected networks

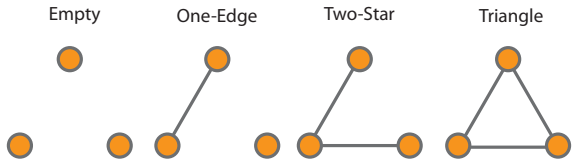


Table 2: Nyakatoke risk-sharing network triad census

	empty	**one-edge**	**two-star**	**triangle**
Count	221,189	48,245	4,070	315
Proportion	0.8078	0.1762	0.0149	0.0012
Random Graph Proportion	0.8049	0.1812	0.0136	0.0003

Source: de Weerdt (2004) and author's calculations.
Notes: The Nyakatoke network includes $N = 119$ households, corresponding to $\binom{N}{2} = 7,021$ unique dyads and $\binom{N}{3} = 273,819$ unique triads.

Network statistics involving triples of agents of triads

Triads, a set of three unique agents, come in four types: no connections, one connection, two connections, or three connections between them. These triad types are called empties, one-edges, two-stars and triangles respectively. There are $\binom{N}{3} = \frac{N(N-1)(N-2)}{6}$ unique triads in a network of size N. A complete enumeration of them into their four possible types constitutes a triad census.

Each agent can belong to as many as $(N-1)(N-2)$ triangles. The counts of these triangles are contained in the N diagonal elements of $\mathbf{D}^3$. However each such triangle appears 6 times in these counts: as $\{i,j,k\}$, $\{i,k,j\}$, $\{j,i,k\}$, $\{j,k,i\}$, $\{k,i,j\}$ and $\{k,j,i\}$. Thus

$$\#\text{ of triangles} = T_T = \frac{\text{Tr}\left(\mathbf{D}^3\right)}{6} \tag{11}$$

equals the number of unique triangles in the network.

Each pair of agents $\{i,j\}$ can share of up to $N-2$ links in common. If $\{i,j\}$ are not linked themselves, then they may belong to two stars with their links in common as star centers. Since each dyad can create up to $N-2$ closed triangles by forming a link between themselves, there may be up to $\binom{N}{2}(N-2) = \frac{N(N-1)(N-2)}{2}$ (actual) triangles or two stars (i.e., potential triangles) in the network. These counts are contained in the lower (or upper) off-diagonal elements of $\mathbf{D}^2$. Each triad appears three times in these counts: as $\{i,j,k\}$, $\{i,k,j\}$ and $\{j,k,i\}$. If the triad is a two star, then only one of $D_{ji}D_{ki}$, $D_{ij}D_{kj}$, or $D_{ik}D_{jk}$ quantities will equal one (i.e., contribute). If it is a triangle, then all three will equal one. Therefore $\text{vech}\left(\mathbf{D}^2\right)'\iota$ gives the network count of *three times* the number triangles *plus* the number of two-stars, with the count of the latter alone equal to

$$\#\text{ of two stars} = T_{TS} = \text{vech}\left(\mathbf{D}^2\right)'\iota - \frac{\text{Tr}\left(\mathbf{D}^3\right)}{2}. \tag{12}$$

We can use a similar logic to calculate the number of one-edge triads. Each agent belongs to $N-2$ triads. If all triads are empty or have only one edge, then there will be $(N-2)\,\text{vech}\left(\mathbf{D}\right)\iota$ one edge triads. However if some triads are two-stars or triangles this count will be incorrect. It turns out that subtracting twice the number of two stars and three times the number of triangles gives the correct answer.

$$\#\text{ of one edges} = T_{OE} = (N-2)\,\text{vech}\left(\mathbf{D}\right)'\iota - 2\text{vech}\left(\mathbf{D}^2\right)'\iota + \frac{\text{Tr}\left(\mathbf{D}^3\right)}{2} \tag{13}$$

The number of empty triads, T_E, equals $\binom{N}{3}$ minus the sum of (11), (12) and (13). Note that (11), (12) and (13) collectively imply that

$$\begin{aligned} T_{OE} + 2T_{TS} + 3T_T &= (N-2)\,\text{vech}\left(\mathbf{D}\right)'\iota, \\ &= \frac{1}{4}N\left(N-1\right)\left(N-2\right)P_N \end{aligned}$$

suggesting that network density can be computed from the triad census according to

$$P_N = \left(\frac{4T_{OE} + 8T_{TS} + 12T_T}{N\left(N-1\right)\left(N-2\right)} \right). \tag{14}$$

The triad census for the Nyakatoke network is given in Table 2. As a point of comparison the proportion of each type of triad that we would expect to see in a random graph, where the probability of a link between any two agents coincides with the observed density of the Nyakatoke network (0.0698), is given in the last row of the table.

A measure of network transitivity is given by three times the number of transitive triads in the network relative to three times the number of transitive triads *plus* those triads which could become transitive with the addition of a single link (i.e., two stars). The Transitivity Index, sometimes called the clustering coefficient, is

$$\begin{aligned} \text{Transitivity Index} &= \frac{3T_T}{T_{TS} + 3T_T} \\ &= \frac{1}{2} \frac{\text{Tr}\left(\mathbf{D}^3\right)}{\text{vech}\left(\mathbf{D}^2\right)' \iota} \\ &= R_N. \end{aligned}$$

In random graphs R_N should be close to network density. For the Nyakatoke network the transitivity index is 0.1884, which substantially exceeds the density of the network (0.0698). We will explore how to assess the statistical significance of this difference in Section 3 below. Transitivity has been hypothesized to facilitate risk sharing and other activities where monitoring may be helpful. If the (i, j, k) triad is transitive, then agent k may be able to monitor actions involving i and j. See Jackson (2014) for additional discussion. Faust (2007) surveys the extensive sociological literature on triad configurations.

Degree distributions and triad counts

A reoccurring theme in social network analysis concerns whether observed network structures can be explained through a series of dyadic decisions, or whether interactions among larger groups of agents, most often a triads, need to be considered (see Faust (2007) for a recent statement and references to earlier work). Most economic models of network formation model links as forming via pairwise interactions, albeit interdependent ones.

While network transitivity, and the triad census, has often been a focus of sociologists, other network researchers have made a network's degree distribution

$$F\left(d_{+}\right)=\Pr\left(D_{i+}\leq d_{+}\right)$$

their primary object of study (e.g., Barabási and Albert, 1999). Figure 5 plots the Nyakatoke network's degree distribution. A small number of households in the Nyakatoke network have many links (over 20), while the vast majority have only a small number of links (less than 10).

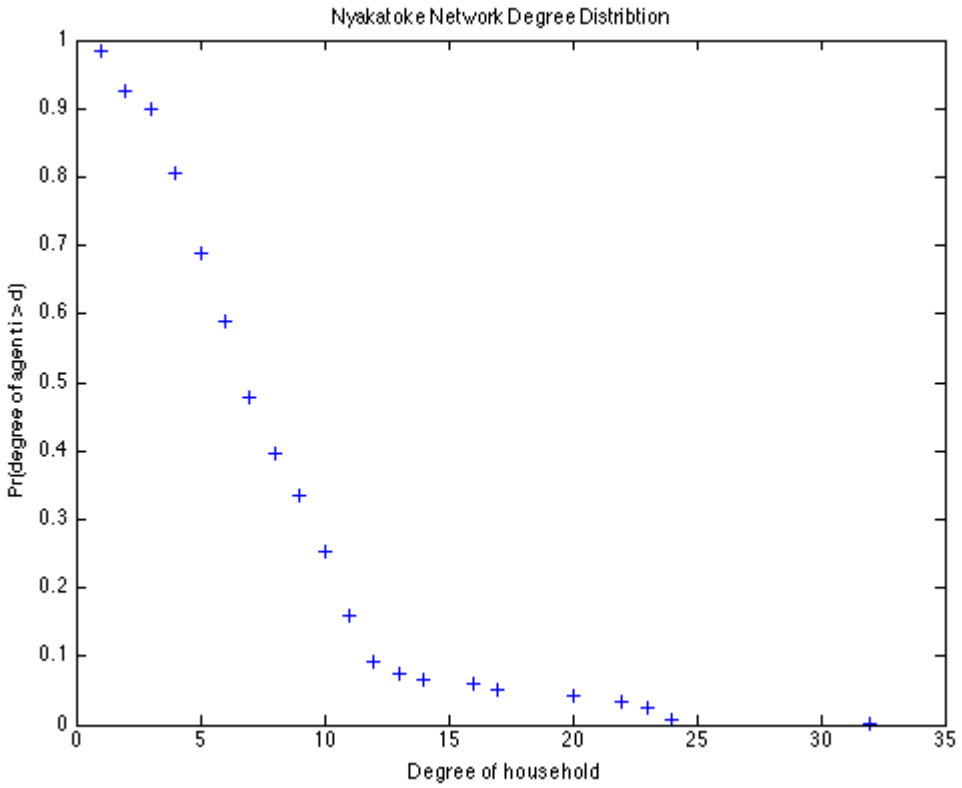
Faust (2007) argues, via a collection of empirical examples, that the distribution of triad configurations within networks are well-predicted by network statistics defined on lower order sub-graphs (i.e., dyads). Some additional insight in this finding can be developed via some basic algebra.

Some tedious manipulations give a variance of the degree distribution equal to

$$S_{N}^{2}=\frac{2}{N}\left(T_{TS}+3T_{T}\right)-\left(N-1\right)P_{N}\left[1-\left(N-1\right)P_{N}\right]. \tag{15}$$

Consider the effect of inducing a mean preserving spread in a network's degree distribution. That is, we seek manipulations which keep network density fixed, while increasing the variance of the degree distribution. Jackson and Rogers (2007a), in the context of a technology diffusion model, provide an interesting motivation for considering this thought experiment.

Figure 5: Nyakatoke risk-sharing network degree distribution



Source: de Weerdt (2004) and author's calculations.

Using (14) and (15) we get

$$\begin{aligned} S_N^2 &= \frac{2}{N}\left(T_{TS}+3T_T\right) \\ &\quad -\left(N-1\right)\left(\frac{4T_{OE}+8T_{TS}+12T_T}{N\left(N-1\right)\left(N-2\right)}\right)\left[1-\left(N-1\right)\left(\frac{4T_{OE}+8T_{TS}+12T_T}{N\left(N-1\right)\left(N-2\right)}\right)\right] \end{aligned}$$

Inducing a mean-preserving spread requires triad manipulations that (i) increases the first term in the expression above, while (ii) leaving the second term unchanged. Table 3 list several mean-preserving triad manipulations. A triad is the smallest subgraph rearrangement we can use to induce a mean-preserving spread in the degree distribution.

To increase S_N^2 a two-star or triangle must be added to the network (accommodated by changes in the number of empties and one edges). Alternatively we can convert a two-star into a triangle (again accommodated by changes in the number of empties and one edges). These correspond to manipulations 2 to 5 and manipulation 6 in Table 3. Note that manipulations 2 and 4 and 3 and 5 are isomorphic, while manipulation 1 does not increase S_N^2. This leaves 4, 5 and 6 as unique triad manipulations which induce mean preserving spreads in a network's degree distribution.

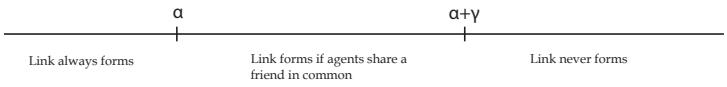
Each of manipulations 4 to 6 involve net increases in $T_{TS}+3T_T$, accommodated by decreases in the number of one edges and increases in the number of empties. This is an example of how a network's degree distribution fundamentally constrains other aspects of its topology. In this case higher variance degree sequences imply networks with more "hubs" (nodes with many links emanating outwards from them) *as well as* more isolated nodes. Jackson and Rogers (2007a) show that the first effect tends to facilitate the spread of infections (ideas, new technology, etc.) in a network, while the second acts as a break to diffusion.

Note that the effect of the triad manipulations listed in Table 3 on transitivity is not one-directional. Manipulation 4 reduces transitivity, while manipulations 5 and 6 increase it. Nevertheless, Barabási and Albert's (1999) focus on degree distributions may not be misplaced. Evidently the form of a network's degree distribution is strongly connected to the

Table 3: Mean-preserving spreads via triad manipulations

		Net final change in triad type				Change in S_N^2	Change in R_N
#	Initial triad manipulation	empty	one edge	two star	triangle		
1	change empty to one edge	0	0	0	0	0	0
2	change empty to two star	+1	-2	+1	0	$\frac{2}{N}$	-
3	change empty to triangle	+2	-3	0	+1	$\frac{6}{N}$	+
4	change one edge to two star	+1	-2	+1	0	$\frac{2}{N}$	-
5	change one edge to triangle	+2	-3	0	+1	$\frac{6}{N}$	+
6	change two star to triangle	+1	-1	-1	+1	$\frac{4}{N}$	+

Figure 6: Realized values of U_{ij}



distribution of other, high order, subgraph features.

Faust (2007) finds that, as a practical/empirical matter, the distribution of triad configurations across a wide range of networks can be largely explained by lower order subgraph features (e.g., moments of the degree distribution).

2 Modeling network formation

To characterize some of the issues that arise when empirically modeling network formation it is helpful to initially consider a very simple model. Assume that directly-linked agents may make transfers to one another. Therefore agents i and j will form a link if the net surplus from doing so is positive, conditional on the link behavior of all other agents in the network. This corresponds to a variant of the direct-transfer network formation game, under pairwise equilibrium, studied by Bloch and Jackson (2007). Let $F_{ij}(\mathbf{D}) = \left(\sum_{k=1}^{N} D_{ik}D_{jk}\right)$ denote the number of friends agents i and j share in common. Links form according to the rule

$$D_{ij} = \mathbf{1}\left(\alpha_0 + \gamma_0 F_{ij}(\mathbf{D}) - U_{ij} \geq 0\right) \tag{16}$$

for $i = 1, \ldots, N$ and $j < i$. Here U_{ij} is an unobserved component of link surplus; independently and identically distributed across dyads according to a known distribution:

$$U_{ij} \overset{iid}{\sim} F_U,\ i = 1, \ldots, N,\ j < i\ U_{ij} \in \mathbb{U}. \tag{17}$$

Rule (16) implies that agents form links if (i) they share many friends in common ($F_{ij}(\mathbf{D})$) and/or (ii) the unobserved idiosyncratic utility from doing so is high ($-U_{ij}$). The magnitude of $\gamma_0 > 0$ captures the strength of agents' preferences for triadic closure in links. The dependence of the surplus generated by an i-to-j link on the presence or absence of links across other pairs of agents constitutes a *network externality*. Network externalities generate complex interdependencies across the choices of different agents, a modeling challenge not present in textbook single agent models.

As noted earlier, in real world social networks linked agents often share additional links in common, generating a clustering of ties. Rule (16) generates such clustering by positing a structural taste for link transitivity – the returns to a relationship are higher if two individuals share a friend in common. A preference for transitive links may be micro-founded in a variety of ways. For example, actions between dyad partners can be monitored or refereed by a shared friend; this may be valuable in the context of a risk-sharing network. Alternatively it may be more enjoyable to socialize with two friends, if they are also friends with each other.

An alternative explanation for clustering is that agents assortatively match on some unobserved attribute, a process called homophily. Homophily on observed attributes is a feature of many real-world networks (McPherson, Smith-Lovin and Cook, 2001). Rule (16) and assumption (17) rules out homophily a priori.

As an alternative to rule (16) Handcock, Raftery and Tantrum (2007), Krivitsky, Handcock, Raftery and Hoff (2009) and Graham (2014), consider link formation rules like

$$D_{ij} = \mathbf{1}\left(Z_{ij}'\eta_0 + \nu_i + \nu_j - g\left(\xi_i, \xi_j, \delta_0\right) - U_{ij} \geq 0\right), \tag{18}$$

where Z_{ij} is an observed $K \times 1$ vector of dyad attributes, ν_i and ξ_i are unobserved agent-level heterogeneity, and U_{ij} is an idiosyncratic dyad level surplus component; $g\left(\xi_i, \xi_j, \delta_0\right)$ is a known symmetric distance function which (i) takes a value of zero at $\xi_i = \xi_j$ and (ii) is increasing in $\left|\xi_i - \xi_j\right|$. The goal is to learn about η_0, δ_0 and features of the conditional

distribution of (ν_i, ξ_i) given $\mathbf{Z}$.

Relative to rule (16), rule (18) introduces a much richer form of unobserved agent-level heterogeneity. First, agents are heterogeneous in the amount of link surplus they generate. Agents with high values ν_i generically generate more surplus. Such agents will have more links, giving rise to degree heterogeneity; an important feature of real work networks (see Figure 5). Second, the model allows for assortative matching on ξ_i. Agents which are similar in terms of the unobserved characteristic ξ_i generate more surplus from linking. This feature of the model induces clustering in links. Unlike rule (16), rule (18) does not include any network externalities. The presence or absence of a link elsewhere in the network does not change the returns to an i-to-j link.

In practice link rules with network externalities and those with rich forms of agent-level heterogeneity can generate very similar networks. This makes discriminating between, for example, structural transitivity and homophily on unobservables difficult. Nevertheless, distinguishing between them is scientifically interesting and policy-relevant. Transitivity is associated with an externality in link formation. In the presence of externalities a local manipulation of network structure can influence link formation elsewhere in the network. If clustering is due soley to homophily, local manipulations do not have effects that cascade through the network.

Below I discuss how panel data may be used to model both a structural taste for transitivity and assortative matching on unobserved attributes simultaneously. Initially, however, I focus on cross-sectional models that include either network externalities or heterogeneity, but not both.

A simple cross-sectional model with structural transitivity

Returning to link rule (16), assume that the econometrician bases her inferences on a random sample of networks from some well-defined population (of networks). For example networks of food sharing among households across a population of indigenous communities

(e.g., Koster and Leckie, 2014). For each sampled network (community) the entire adjacency matrix is observed. This sampling process asymptotically reveals $F(\mathbf{D}|\, N = n)$ for $n \in \mathbb{N} = \{2, 3, 4, \ldots\}$. Implicit in (6) is the assumption that the distribution of U_{ij} is independent of network size. The notation D_{ij} corresponds to the link status of the generic, randomly drawn, (i, j) dyad, itself sampled from a randomly drawn network. To economize on notation there is no explicit network subscript in what follows.

Equation (16) defines a system of $\binom{N}{2}$ simultaneous discrete choices. Viewed in this way, two questions naturally arise. First, for a given $\theta_0 = (\alpha_0, \gamma_0)'$ does (16) have a solution for all $\mathbf{U} \in \mathbb{U}^N$? This is a question of equilibrium existence or model coherence. Demonstrating existence can be non-trivial for some models of network formation (cf., Jackson, 2008; Chapter 11; Hellmann, 2013). Second, if an equilibria does exist, is it unique (again for all $\mathbf{U} \in \mathbb{U}^N$)? This is a question about model completeness: given a particular draw of the model's underlying latent variable $\mathbf{U}$, does it deliver a unique prediction for the observed network, $\mathbf{D}$? Multiplicity of equilibrium network configurations is a common feature of many models with network externalities.

The study of models with qualitative features similar to those of (16) has a long history in econometrics (e.g., Heckman, 1978a). Important recent contributions include those of Bresnahan and Reiss (1991), Tamer (2003) and Ciliberto and Tamer (2009) among others. Unfortunately the combinatoric complexity of networks, with $2^{\binom{N}{2}}$ link configurations possible in a network with N agents, makes the direct application of insights from prior work difficult.

To keep the discussion simple assume that $N = 3$. In this case there are four possible non-isomorphic network configurations corresponding to the four types of triads depicted in Figure 4 above. The heterogeneity draw is given by the triple $\mathbf{U} = (U_{12}, U_{13}, U_{23})' \in \mathbb{U}^3$. For any given draw of $\mathbf{U}$ one of these four configurations will be observed.

Call draws of U_{ij} below α, between α and $\alpha+\gamma$, and above $\alpha+\gamma$ respectively low (L), medium (M) and high (H) draws (see Figure 6). Let $p_{LLL}(\theta, F_U) = F_U(\alpha)^3$ denote the probability of

three 'low' draws; $p_{LMH}(\theta, F_U) = F_U(\alpha) F_U(\alpha+\gamma)[1 - F_U(\alpha+\gamma)]$ the probability of one low, one medium and one high draw and so on. Observe that low draws of U_{ij} correspond to higher link surplus.

If U_{12} falls in the 'low' region, then agents 1 and 2 will form a link regardless of whether they share a friend in common (i.e., $D_{13}D_{23}$ may equal zero or one). In contrast if U_{12} falls in the 'medium' region, then agents 1 and 2 will form a link only if they share a friend in common (i.e., if $D_{13}D_{23} = 1$). If U_{12} falls in the 'high' region, then they never form a link.

The contingent behavior associated with a 'medium' idiosyncratic surplus component is what generates the possibility of multiple equilibria. Consider the case where all three elements of $\mathbf{U}$ fall into the 'medium' range. In that case two network configurations are consistent with (16): (i) the empty triad and (ii) a triangle. The model, as specified, is silent on which of these two networks is chosen.

Let $\underline{\pi}_T(\theta, F_U)$ denote the minimum probability the model defined by (16) and (17) logically attaches to observing a triangle for a particular θ and F_U. This probability coincides with the probability mass attached to the region of $\mathbb{U}^3$ where the model uniquely predicts a triangle network. Let $\bar{\pi}_T(\theta, F_U)$ denote the maximal probability the model logical attaches to observing a triangle. This probability coincides with the probability mass attached to the region of $\mathbb{U}^3$ where a triangle network is either the unique network configuration, or among the set of multiple configurations, consistent with (16).

Recalling the notation of 'T' for 'triangle', 'TS' for 'two-star', 'OE' for 'one-edge' and 'E' for 'empty', the above logic yields the following probability bounds on the four non-isomorphic

network configurations:

$$\begin{aligned}
\underline{\pi}_T(\theta, F_U) &= p_{LLL}(\theta, F_U) + p_{LLM}(\theta, F_U) \\
\bar{\pi}_T(\theta, F_U) &= p_{LLL}(\theta, F_U) + p_{LLM}(\theta, F_U) + p_{LMM}(\theta, F_U) + p_{MMM}(\theta, F_U) \\
\pi_{TS}(\theta, F_U) &= p_{LLH}(\theta, F_U) \\
\underline{\pi}_{OE}(\theta, F_U) &= p_{LMH}(\theta, F_U) + p_{LHH}(\theta, F_U) \\
\bar{\pi}_{OE}(\theta, F_U) &= p_{LMM}(\theta, F_U) + p_{LMH}(\theta, F_U) + p_{LHH}(\theta, F_U) \\
\underline{\pi}_E(\theta, F_U) &= p_{MMH}(\theta, F_U) + p_{MHH}(\theta, F_U) + p_{HHH}(\theta, F_U) \\
\bar{\pi}_E(\theta, F_U) &= p_{MMM}(\theta, F_U) + p_{MMH}(\theta, F_U) + p_{MHH}(\theta, F_U) + p_{HHH}(\theta, F_U).
\end{aligned}$$

Let π_T denote the *population* frequency of triangle networks, etc. Rule (16) therefore delivers the following inequality restrictions

$$\begin{aligned}
\underline{\pi}_T(\theta, F_U) \leq \pi_T &\leq \bar{\pi}_T(\theta, F_U) \\
\pi_{TS} &= \pi_{TS}(\theta, F_U) \\
\underline{\pi}_{OE}(\theta, F_U) \leq \pi_{OE} &\leq \bar{\pi}_{OE}(\theta, F_U) \\
\underline{\pi}_E(\theta, F_U) \leq \pi_E &\leq \bar{\pi}_E(\theta, F_U).
\end{aligned} \tag{19}$$

The model also generates the equalities

$$\pi_T + \pi_{OE} + \pi_E = \bar{\pi}_T(\theta, F_U) + \underline{\pi}_{OE}(\theta, F_U) + \underline{\pi}_E(\theta, F_U) = \underline{\pi}_T(\theta, F_U) + \bar{\pi}_{OE}(\theta, F_U) + \bar{\pi}_E(\theta, F_U). \tag{20}$$

The identified set, Θ_I, is the set of all $\theta \in \Theta$ such that (19) and (20) are satisfied. Ciliberto and Tamer (2009), among others, discuss methods of estimating Θ_I and conducting inference on it and/or on θ_0.

The observation that link formation rule (16) is a system of simultaneous discrete choices and, further, that this system generates a set of moment inequalities which may be used as a

basis for inference on θ_0, appears promising. Unfortunately, as noted above, this observation may be of limited practical importance. In a network with N agents, there are $2^{\binom{N}{2}}$ possible configurations of links. For each $\mathbf{U}$ in $\mathbb{U}^N$ and $\theta \in \Theta$, the consistency of a given network with (16) must be checked. In practice this is not feasible in real time for all but very small networks. Even showing that two networks are isomorphic is a non-trivial problem (e.g., Read and Corneil, 1977).

While fully exploiting the identifying power of (16) and (17) may be infeasible in even modest-sized networks, exploiting some of its identifying content is straightforward. Assume that networks vary in size with $N \in \mathbb{N} = \{2, 3, 4, \ldots\}$ and recall that the distribution of U_{ij} is constant in N. Under (16) and (17) the probability that a randomly drawn dyad from a network of size N is linked (i.e., density in networks of size N) satisfies the inequalities

$$F_U(\alpha_0) \leq \Pr(D_{ij} = 1 | N) \leq F_U(\alpha_0 + \gamma_0(N-2))$$

for all $N \in \mathbb{N}$. The lower bound occurs when the randomly drawn dyad share no friends in common, the upper bound when the dyad is linked to all other members of the network (except possibly each other).

These upper and lower bounds coincide at $N = 2$ so that α_0 pointed identified by the density of links across networks consisting of a single dyad:

$$\alpha_0 = F_U^{-1}(\Pr(D_{ij} = 1 | N = 2)).$$

A lower bound on γ_0 is then given by

$$\underline{\gamma} = \sup \left\{ \frac{F_U^{-1}(\Pr(D_{ij} = 1 | N)) - \alpha_0}{N-2} \;\middle|\; N \in \mathbb{N} \right\}.$$

Here an informative lower bound on γ_0 is generated by observing a higher density of link formation in networks with $N > 2$, than across networks consisting of single dyads. This

does not strike me as an especially attractive approach to inferring the presence of a taste for transitivity, but it is illustrative of how some identifying implications of a network formation model can be easy to exploit (even if utilizing all implications is impractical).

Another, and more interesting, example of this type of approach is provided by Sheng (2012), who explores the the identifying content of (non-trivial) subnetwork configurations. Assume networks consist of N agents and consider the probability that, for a randomly drawn triad, itself drawn from a randomly sampled network, we observe a particular triad configuration (see Figure 4). This probability will depend on the degree to which members of the sampled triad are connected to the rest of the network. Maximal connection occurs when all members of the sampled triad are connected to all other agents in the network. Isolation occurs when no member of the triad is linked to other agents in the network.

Now imagine repeating the thought experiment used to derive (19) above, but doing so conditional on different assumptions about the triad's connectivity to the rest of the network. For example conditional on the three dyads forming the triad having, say, no, two and two friends (outside the triad) in common, the model provides upper and lower bounds on the probability of observing, say, a triangle configuration. An identification region for θ_0 can be computed using the union of these conditional bounds on each triad configuration (computed for all possible degrees of triad connectivity).

Christakis, Fowler, Imbens and Kalyanaraman (2010) suggest an alternative approach to dealing with the inferential challenges posed by multiplicity. They posit that the network forms sequentially. Agents form, maintain or dissolve, links in a specific order and do so myopically. Specifically they do not anticipate how the links they choose to form today change the incentives for link formation faced by subsequent agents.

Returning to the $N = 3$ case, assume that U_{12}, U_{13} and U_{23} are respectively low, low and medium draws (see Figure 6). Assume that agent 1 forms links first, followed by agents two and three. Under this ordering, agent 1 will immediately form links with both agents 2 and 3. Agent 2 will then form a link with agent 3. Although the idiosyncratic utility from this

link is only 'medium', the link forms to reap the benefits of triadic closure, since both agents 2 and 3 already share 1 as a friend. Finally, agent 3 maintains all links formed earlier. The triangle configuration emerges from this ordering (and draw of $\mathbf{U}$).

Now consider the alternative ordering where agent 3 forms links first, followed by agents 2 and 1. In this case agent 3 will form a link with agent 1, but not 2. The absence of the utility gain associated with triadic closure means the 2-to-3 link does not form. Agent 2 then forms a link with 1. Finally, agent 1 maintains her links with agents 2 and 3. A two-star configuration emerges from this ordering.

As the above examples indicate, if the ordering of link formation opportunities were observed, likelihood-based inference would be straightforward. Christakis, Fowler, Imbens and Kalyanaraman (2010) address the unobservability of the posited sequential network formation process by assigning a probability distribution to agents' ordering, and then working with the resulting integrated likelihood. In the the simple example discussed here, there are $N! = 3! = 6$ possible orderings. If each ordering is a priori assumed equally likely the likelihood is easily written down. Christakis, Fowler, Imbens and Kalyanaraman (2010) approach to inference is Bayesian (and based on the observation of a single network). An important contribution of their paper is to make the simple idea sketched above computationally operational for realistically-sized networks. Specifically, they use Markov Chain Monte Carlo (MCMC) methods to take draws from a posterior distribution for the model parameters.

An unattractive feature of assuming the network is formed sequentially, is that the resulting likelihood will, for certain values of $\mathbf{U}$, place positive probability on network configurations that do not correspond to an equilibrium of the simultaneous-move static game. This is again illustrated by the example above. In the static game a low, low, medium draw of $\mathbf{U}$ uniquely predicts a triangle network. For the same draw of $\mathbf{U}$ the sequential game places a probability of two-thirds on the triangle network, and a probability of one-third on the two-star network. If, in reality, agents have the opportunity to continually revise their links, a two-star configuration would not emerge conditional on a low, low and medium draw of

idiosyncratic link surpluses.

Mele (2013) develops a related approach to empirically modeling network formation. He posits a process whereby in each 'period' a randomly drawn dyad is given the opportunity to form, maintain or dissolve a link. For a specific specification of link surplus and meeting probabilities, he shows that the sequence of networks generated by the model is a stationary ergodic process. The long-run probabilities attached to specific network configurations are used to formulate a likelihood. Like, Christakis, Fowler, Imbens and Kalyanaraman (2010), Mele's (2013) approach to inference is Bayesian. He develops an MCMC algorithm for generating draws from a posterior distribution for the model parameters. His approach also places positivity probability on network configurations that are not equilibria of the corresponding simultaneous-move static game.

The Sheng (2012), Christakis, Fowler, Imbens and Kalyanaraman (2010) and Mele (2013) papers all provide operational methods for inferring the distribution of link surplus from observed network structure. Sheng's (2012) approach provides a computationally feasible (albeit difficult) way to harness the identifying content of pairwise stability. Her approach to inference requires the observation of many independent networks (see also Miyaichi, 2013). Christakis, Fowler, Imbens and Kalyanaraman (2010) and Mele (2013) show the identifying power of moving from a simultaneous to sequential network formation process. All three methods are computationally intensive.

A simple cross-sectional model with heterogeneity

Now return to the link formation rule (18). This model has a rich heterogeneity structure, complicating its analysis relative to rule (16). However rule (18) also excludes externalities in link formation a priori, side-stepping the coherence and completeness issues associated with rule (16).

Graham (2014) studies (18) with $g\left(\xi_i, \xi_j, \delta_0\right)$ empty; that is a model with unobserved degree heterogeneity, but no homophily on unobservables. He derives the joint maximum likelihood

estimator where both the common parameter η_0 and the incidental parameters $\{\nu_i\}_{i=1}^{\infty}$ are estimated simultaneously. He further assume U_i is a logistic random variable. Graham (2014) derives the limiting distribution of the common parameter as the network grows large. This limit distribution is normal, but includes a bias term.

Graham also proposes an estimator which conditions on a sufficient statistic for the degree heterogeneity parameters. Charbonneau (2014), in independent work, develops a closely related procedure in the context of studying gravity trade models. Random effects estimation of (18) is pursued in Krivitsky, Handcock, Raftery and Hoff (2009) using MCMC methods.

One advantage of a fixed effects treatment of degree heterogeneity is that the resulting model of tie formation will be able to perfectly match any observed degree sequence (cf., Chatterjee, Diaconis and Sly, 2011). As argued above algebraically, and shown by Faust (2007) empirically, a network's degree distribution often does a reasonably good job of explaining (i.e., predicting) other "higher order" aspects of network architecture (e.g., the frequency of different triad configurations). For this reason analyses based on (18) are likely to provide good fits, even if the true link formation process includes network externalities.

Dynamic models of network formation

If the econometrician observes the structure of links within a network evolve over time, a number of new modeling opportunities arise. In particular is becomes possible to meaningful incorporate both network externalities and rich forms of agent level heterogeneity into a single model of link formation. Let $t = 0, 1, 2, 3$ index the periods in which each network is observed and assume that links form in period t according to the rule

$$D_{ijt} = \mathbf{1}\left(\beta_0 D_{ijt-1} + \gamma_0 F_{ijt-t}\left(\mathbf{D}_{t-1}\right) + A_{ij} - U_{ij} \geq 0\right) \tag{21}$$

with, for example,

$$A_{ij} = \nu_i + \nu_j - g\left(\xi_i, \xi_j\right), \tag{22}$$

where all notation is as previously defined. Model (21) combines features of the two static models discussed above (rules (16) and (18)). It incorporates key network dependencies emphasized in prior work (cf., Snijders, 2011). First, links are persistent. If agents i and j are linked in period t, they are more likely to be linked in subsequent periods ($\beta_0 > 0$). Second, as in the first static model discussed above, there are returns to 'triadic closure' ($\gamma_0 > 0$). The net surplus associated with an i-to-j link is increasing in the number of friends i and j shared in common during the prior period. Third and fourth, as in the second static model discussed above, both degree heterogeneity and assortative matching on unobservables are incorporated.

As in Christakis, Fowler, Imbens and Kalyanaraman (2010) and Mele (2013), model (21) implies that agents form links myopically. At the beginning of each period agents form, maintain and dissolve thinks 'as if' all other features of the network will remain fixed. This is analogous to a best-reply dynamic. Assuming a best-reply type dynamic eliminates the contemporaneous feedback which generated multiple equilibria, and its associated inferential challenges, in the static model discussed above. At the same time, by allowing link surplus to vary with the structure of the network in the prior period, network dependencies, such as a taste for triadic closure, are incorporated into (21).

Most theoretical models of network formation assume agents form links according to some variant of naive best-reply dynamics (e.g., Jackson and Wolinsky, 1996; Jackson and Watts, 2002; Bala and Goyal, 2000; Watts, 2001; Jackson and Rogers, 2007b), although some scholars have studied models with forward-looking agents (e.g., Dutta, Ghosal and Ray, 2005). The dynamics of link formation implied by (21) are closely aligned with the types of dynamics assumed by theorists. Although the myopic nature of link formation may not be of particular concern, a more mundane, but nevertheless important, concern may arise in empirical work. It may be that the frequency at which the network is sampled, and the structure of links recorded, does not correspond naturally with the timing at which agents actually make link decisions. Similar concerns arise in single agent discrete choice analyses

(cf., Chamberlain, 1985). When formulating a social network data collection protocol, the timing of link decisions and the timing of data collection should be aligned.

In the first static model discussed above, the clustering of ties was explained soley by a taste for triadic closure. In practice tie clustering might also arise because agents assortatively match on attributes unobserved by the econometrician (homophily), as was assumed in the second model. The dynamic model introduced here allows for both sources of clustering.

Goldsmith-Pinkham and Imbens (2013) take random-effects approach to model (21). If the density of U_{ij} is known (e.g., Standard Normal or Logistic), and the joint distribution of $(\mathbf{D}_0, \mathbf{A})$ belongs to a parametric family, then inferences on $\theta_0 = (\beta_0, \gamma_0)'$ may be based on an integrated or random effects likelihood. In principle this is very much analogous to random effects approaches to single-agent dynamic panel data models (Heckman, 1981a-c; Chamberlain, 1985). In reality both the specification and maximization of an integrated likelihood in this setting is non-trivial.

Ideally the specified joint distribution for $(\mathbf{D}_0, \mathbf{A})$ should allow for dependence between $\mathbf{D}_0$ and $\mathbf{A}$. Since the model implies that $\mathbf{D}_1$ varies with $\mathbf{A}$, it seems 'natural' to allow the initial network configuration, $\mathbf{D}_0$, to also vary with $\mathbf{A}$. This is a complicated version of the initial conditions problem which arises in single-agent dynamic panel data models (Wooldridge, 2005).

To get a sense of the modeling issues involved, assume that A_{ij} takes the form given in (22) with (ν_i, ξ_i) bivariate normal with an unknown location vector and scale matrix. Assume that $g(\cdot, \cdot)$ is a known function, that U_{ijt} is a standard normal random variable, and that $D_{ij0} = 1(A_{ij} - U_{ij0} \geq 0)$. These assumptions are sufficient to write down the integrated-likelihood. Evaluating that likelihood, however, would be challenging. Doing so would involve calculating a $2N$-dimensional integral. This integral does not obviously factor into a set of lower dimensional integrals (since A_{ij} and A_{kl} will share components in common whenever $i = k$ or $j = l$).

Motivated by these computational challenges Goldsmith-Pinkham and Imbens (2013) instead

work with a highly stylized model. They rule out degree heterogeneity, set $g(\xi_i, \xi_j) = |\xi_i - \xi_j|$, and assume that ξ_i is binary-valued with $\Pr(\xi_i = \alpha_\xi | \mathbf{D}_0) = \Pr(\xi_i = 0 | \mathbf{D}_0) = \frac{1}{2}$. Note that this last assumption assumes, unattractively, independence between $\mathbf{D}_0$ and $\mathbf{A}$. Under these assumptions Goldsmith-Pinkham and Imbens (2013) develop an algorithm for taking draws from the posterior distribution for the model's parameters.

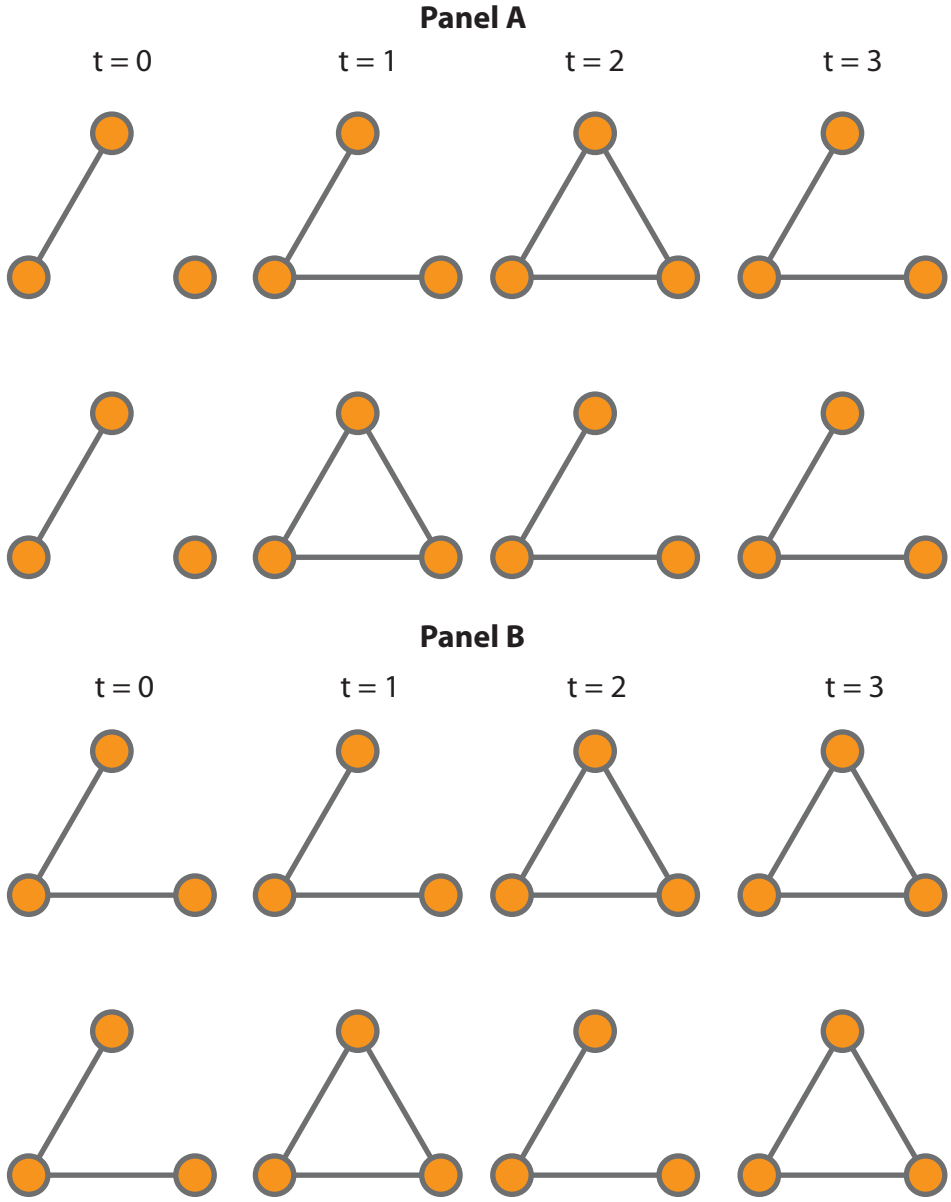
Graham (2013) approaches model (21) from a fixed effects perspective, asking if it contains implications that are invariant to $\mathbf{A}$, but useful for identifying θ_0. This approach leaves the distribution of $(\mathbf{D}_0, \mathbf{A})$ unspecified and unrestricted. Perhaps surprisingly, fixed effects identification results can be derived.

Consider a dyad that is embedded in a stable neighborhood. A stable neighborhood has two features. First, with the exception of possible link formation and dissolution between themselves, the set of links maintained by agents i and j is the same across periods 1, 2 and 3. Agents i and j may add, maintain or delete links between periods 0 and 1. Second, the links maintained by *friends* of players i and j do not change between periods 1 and 2. Dyads in stable neighborhoods are embedded in local networks with link structures that are largely fixed up to two degrees away across periods 1, 2, and 3.

Panel A of Figure 7 visually depicts two network sequences in a network consisting of three agents. Number agents 1, 2 and 3 counter-clockwise from the top in each network. Observe that Agents 1 and 3 are embedded in a stable neighborhood. Agent 1 is linked to agent 2, and agent 3 to agent 2, in periods 1, 2 and 3 in both sequences depicted in Panel A.

The only difference between the two network sequences is that in the upper one agents 1 and 3 are linked in period 2, but not in period 1; while in the lower sequence they are linked in period 1, but not in period 2. In the presence of a taste for triadic closure the net surplus associated with a 1-to-3 link will be, in expectation, higher in period 2 than it is in period 1. Since agents 1 and 3 share a common friend in period 1, a 1-to-3 link in the next period will generate additional utility from ensuring triadic closure. Agents 1 and 3 do not share a common friend in period 0. Therefore forming a link in period 1 generates no extra utility

Figure 7: Fixed effects identification of transitivity versus homophily (Panel A) and state dependence versus heterogeneity (Panel B)



Notes: Number agents 1, 2 and 3 counter-clockwise from the top in each network. In panel A $d_{120}d_{230} = 0$ but $d_{121}d_{231} = 1$ so that $(1, 3)$ forming a link has a higher return in period 2 than in period 1. In period 2 the link generates utility from ensuring 'triadic closure', no such utility gain is generated by a period 1 link. Consequently, the first network sequence in panel A arises more frequently than the second in the presence of a structural taste for transitivity in links. Observe that $(1, 3)$ are embedded in a stable neighborhood since $d_{121} = d_{122} = d_{123} = 1$ and $d_{231} = d_{232} = d_{233} = 1$. While the two panel A sequences are uninformative about the presence of true state dependence in ties, this is not the case for the two sequences in panel B. In panel B, the first sequence arises more frequently relative to the second in the presence of true state dependence. Here the intuition is very much analogous to that in Cox (1958).

from ensuring triadic closure. In the presence of a genuine taste for transitivity in links, as embodied in link rule (21), the upper sequence should be observed more frequently than the lower sequence.

Panel B of Figure 7 presents an example of how the relative frequency of different sequences of dyad links, when embedded in a different stable neighborhood from the one depicted in Panel A, provides information about β_0 or state-dependence in links.

In single agent models, fixed effects identification of true state dependence in the presence of unobserved heterogeneity is based on the frequency of observing certain sequences of choices relative to other sequences (e.g., Cox, 1958; Heckman, 1978b; Chamberlain, 1985; Honoré and Kyriazidou, 2000). For example, in the absence of state dependence the binary sequences 0101 and 0011 are equally likely. In the presence of state dependence, the relative frequency of the latter sequence will be greater.

The identification of transitivity versus homophily involves a similar intuition. Conditional on a dyad being embedded in a certain type of local network architecture, certain orderings of link histories should be more frequent than others. This approach involves making comparisons 'holding other features of the network fixed'. This is not straightforward to do. The likelihood associated with a single network sequence includes $3 \times \frac{1}{2}N(N-1)$ distinct components plus the initial condition (itself 'high dimensional'). The challenge is that the likelihood functions associated with the two network histories, even though they are identical in all respects except that the (i,j) friendship history in one is a permutation of that in the other, may be very different. This is because the presence or absence of a link in a given period can affect the likelihood contribution of many other pairs in subsequent periods. For example if (i,k) are linked in period t, then the addition of an (i,j) link increases the probability of a (j,k) link in period $t+1$. Local changes in the network can have widespread effects on the structure of the network likelihood in subsequent periods.

If (i,j) are embedded in a stable neighborhood, the two likelihoods will be nominally quite different, however many contributions in the first likelihood will be permutations of contri-

butions which also appear in the second. As a result, the number of distinct terms in the two likelihoods is small. Exploiting this simplification then allows for the application of identification ideas used in prior work on binary choice (e.g., Manski, 1987; Honoré and Kyriazidou, 2000). Graham (2012), extending earlier work published in Graham (2013), show this type of intuition can be made rigorous.

The relative strengths and weakness of fixed versus correlated random effects approaches to dynamic network analysis, closely mirror those in single agent dynamic discrete choice analysis (cf., Chamberlain, 1984). The computational complexity of these approaches when applied to network models substantially exceeds their single-agent counterparts. The Goldsmith-Pinkham and Imbens (2013) paper provides a valuable template for undertaking a correlated random effects analysis. While some of their modeling assumptions are unattractive, it is one the few coherent likelihood-based empirical models of dynamic network formation and will no doubt be the building block for future research. The fixed-effects results in Graham (2012, 2013) indicate that some features of the distribution of link surplus may be identified without making assumptions about the initial network condition and/or the distribution of unobserved dyad-level heterogeneity. A fixed effects analysis can provide evidence of a structural taste for transitivity under weak assumptions and/or be used to validate specific correlated random effects specifications.

3 Simulating networks

Consider a network with adjacency matrix $\mathbf{D} = \mathbf{d}$ and corresponding degree sequence $\mathbf{D}_+ \stackrel{def}{=} (D_{1+}, \ldots, D_{N+}) = (d_{1+}, \ldots, d_{N+}) \stackrel{def}{=} \mathbf{d}_+$. Let $\mathbb{D}_{N,\mathbf{d}_+}$ denote the set of all undirected $N \times N$ adjacency matrices with degree counts equal to $\mathbf{d}_+$. This section describes an algorithm for sampling uniformly from the set $\mathbb{D}_{N,\mathbf{d}_+}$. Recently, Del Genio, Kim, Toroczkai and Bassler (2010), Blitzstein and Diaconis (2011), Zhang and Chen (2012) and others have constructed (reasonably) efficient procedures for sampling uniformly from the set $\mathbb{D}_{N,\mathbf{d}_+}$. Here I outline

the importance sampling algorithm of Blitzstein and Diaconis (2011), whose exposition I also follow closely.

Sampling uniformly from $\mathbb{D}_{N,\mathbf{d}_+}$ has a number of uses. First, it may be used to determine how unusual a certain graph feature is among the set of all graphs with the same degree sequence. In the Nyakatoke network the transitivity index is almost three times the magnitude of link density. Among networks of the same size, with identical degree sequences, how unusual is it to observe a network with transitivity this high? If there exist very few graphs in this set with such high transitivity, then the researcher might conclude that modeling transitivity is worthwhile.

To be specific Let $f(\mathbf{D})$ be some function of the observed adjacency matrix, say its transitivity index. Among all undirected networks with degree sequences coinciding with $\mathbf{D}$'s what fraction have a transitivity index less than the one observed in the network in hand? Let $\left|\mathbb{D}_{N,\mathbf{d}_+}\right|$ denote the size, or cardinality, of $\mathbb{D}_{N,\mathbf{d}_+}$. We seek to evaluate

$$\Pr\left(f(\mathbf{D}) \leq c\right) = \frac{\sum_{\mathbf{v}\in\mathbb{D}_{\mathbf{N},\mathbf{d}_+}} \mathbf{1}\left(f(\mathbf{v}) \leq c\right)}{\left|\mathbb{D}_{N,\mathbf{d}_+}\right|}. \tag{23}$$

A second motivation for studying $\mathbb{D}_{N,\mathbf{d}_+}$, and specifically how to take random draws from it, is that some sampling schemes do not fully reveal the network adjacency matrix. As noted in the introduction many surveys collect information on an agent's degree (e.g., number of friends, co-workers, sexual partners etc.), but not information on the precise identity of link partners. In such a situation the sampling process reveals that the true network belongs to the set $\mathbb{D}_{N,\mathbf{d}_+}$, but is agnostic about which network within this set is the actual one.

Consider a family of network formation models indexed by the parameter θ (in what follows I ignore observed agent characteristics, but incorporating them would be straightforward). Assume that θ could be consistently estimated if the entire adjacency matrix were observed. Let $\mathbf{D}_b$ denote a uniform random draw from the set $\mathbb{D}_{N,\mathbf{d}_+}$ and let $\hat{\theta}_b$ denote an estimate of θ using $\mathbf{D}_b$. The identified set may be estimated by the convex hull of $\hat{\theta}_1, \ldots, \hat{\theta}_B$ for some

large value of B.

In some situations a researcher may wish to condition on other network features beyonds the degree sequence. This may because it is more appropriate to study the "unusualness" of a given network feature, conditional on additional aspects of network architecture. Alternatively, if estimation of a network formation model is the goal, some sampling schemes reveal more than an agent's degree. For example the General Social Survey (GSS) collects information of which friends are friends themselves (i.e., on triad configurations).

Stanton and Pinar (2012) develop a procedure for sampling from the set of networks with given marginal *and* joint degree distributions. The joint degree distribution captures, for example, the frequency with which an agent with three links is matched to an agent with five links. In very recent work, Goyal, Blitzstein and de Gruttola (2014), show how to construct networks which satisfy a variety of constraints. The focus on drawing from $\mathbb{D}_{N,\mathbf{d}_+}$ in this section is pedagogical. This is the simplest interesting case, illustrates the difficulty of the larger problem, and provides a stepping stone into a growing mathematics and computer science literature on network simulation.

Determining whether a candidate degree sequence is graphical

Direction enumeration of all the elements of $\mathbb{D}_{N,\mathbf{d}_+}$ is generally not feasible. Even for networks that includes as few as 10 agents, this set may have millions of elements. We therefore require a method of sampling from $\mathbb{D}_{N,\mathbf{d}_+}$ uniformly and also estimating its size.

Two complications arise. First, it is not straightforward to construct a random draw from $\mathbb{D}_{N,\mathbf{d}_+}$. Second, we must draw uniformly from this set. Fortunately the first challenge is solvable using ideas from the discrete math literature. To ensure our draws are uniform we use importance sampling (e.g., Owen, 2013).

A sequential network construction algorithm begins with a matrix of zeros and sequentially adds links to it until its rows and columns sum to the desired degree sequence. Unfortunately, unless the links are added appropriately, it is easy to get "stuck" (in the sense that a certain

point in the process it becomes impossible to reach a graph with the desired degree and the researcher must restart the process (e.g., Snijders, 1991)).

Researchers in graph theory and discrete math have studied the construction of graphs with fixed degrees and, in particular, provided conditions for checking whether a particular degree sequence is graphical (e.g., Sierksma and Hoogeveen, 1991). We say that $\mathbf{D}_+$ is graphical if there is feasible undirected network with degree sequence $\mathbf{D}_+$. Not all integer sequences are graphical. For example, there is no feasible undirected network of three agents with degree sequence $\mathbf{D}_+ = (3, 2, 1)$.

Blitzstein and Diaconis' (2011) algorithm is *guaranteed* to produce a matrix from the set $\mathbb{D}_{\mathbf{d}_+,N}$. This is accomplished by cleverly using checks for whether an integer sequence is graphic when adding links. To get a sense how this works in practice it is helpful to begin with a check due to Erdos and Gallai (1961). Let $\mathbf{D}_+ = (D_{1+}, \ldots, D_{N+})$ be a sequence of candidate degrees for each of N agents in a network. Without loss of generality assume that the elements of $\mathbf{D}_+$ are arranged in descending order so that $D_{1+} \geq D_{2+} \geq \cdots \geq D_{N+}$. Erdos and Gallai (1961) showed that $\mathbf{D}_+$ is graphical if and only if $\sum_{i=1}^{N} D_{i+}$ is even and

$$\sum_{i=1}^{k} D_{i+} \leq k\,(k-1) + \sum_{i=k+1}^{N} \min\,(k, D_{i+}) \text{ for each } k \in \{1, \ldots, N\}\,.$$

To show necessity of the condition observe that for any set S of k agents in the network there can be at most $\binom{k}{2} = \frac{1}{2}k\,(k-1)$ links between them. For the remaining $N-k$ agents with $i \notin S$ there can be at most $\min\,(k, D_{i+})$ links from i to agents in S.

The study of graphic integer sequences has a long history in discrete math. Sierksma and Hoogeveen (1991) summarize several criteria that can be used to check whether $\mathbf{D}_+$ is graphical. Blitzstein and Diaconis (2011) base their sampling algorithm on a simple recursive test for whether D_+ is graphical due to Havel (1955) and Hakimi (1962). In what follows D_{i+} denotes the i^{th} element of $\mathbf{D}_+$.

Theorem 1. *(Havel-Hakimi) Let $D_{i+} > 0$, if $\mathbf{D}_+$ does not have at least D_{i+} positive entries*

other than i it is not graphical. Assume this condition holds. Let $\tilde{\mathbf{D}}_+$ be a degree sequence of length $N-1$ obtained by

[i] deleting the i^{th} entry of $\mathbf{D}_+$ and

[ii] subtracting 1 from each of the D_{i+} highest elements in $\mathbf{D}_+$ (aside from the i^{th} one).

$\mathbf{D}_+$ is graphical if and only if $\tilde{\mathbf{D}}_+$ is graphical. If $\mathbf{D}_+$ is graphical, then it has a realization where agent i is connected to any of the D_{i+} highest degree agents (other than i).

Proof. See Blitzstein and Diaconis (2011). □

Theorem 1 is suggestive of a sequential approach to building an undirected network with degree sequence $\mathbf{D}_+$. The procedure begins with a target degree sequence $\mathbf{D}_+$. It starts by choosing a link partner for the lowest degree agent (with at least one link). It chooses a partner for this agent from among those with higher degree. A one is then subtracted from the lowest degree agent and her chosen partner's degrees. This procedure continues until the residual degree sequence (the sequence of links that remain to be chosen for each agent) is zero.

To describe the method proposed Blitzstein and Diaconis (2011) we require some additional notation. Let $(\oplus_{i_1,\ldots,i_k}\mathbf{D}_+)$ be the vector obtained by adding a one to the $i_1,\ldots,i_k$ elements of $\mathbf{D}_+$:

$$(\oplus_{i_1,\ldots,i_k}\mathbf{D}_+)_j = \begin{cases} D_{j+}+1 & \text{for } j \in \{i_1,\ldots,i_k\} \\ D_{j+} & \text{otherwise} \end{cases}$$

Let $(\ominus_{i_1,\ldots,i_k}\mathbf{D}_+)$ be the vector obtained by subtracting one from the $i_1,\ldots,i_k$ elements of $\mathbf{D}_+$:

$$(\ominus_{i_1,\ldots,i_k}\mathbf{D}_+)_j = \begin{cases} D_{j+}-1 & \text{for } j \in \{i_1,\ldots,i_k\} \\ D_{j+} & \text{otherwise} \end{cases}$$

Algorithm 1. *(Blitzstein and Diaconis) A sequential algorithm for constructing a random graph with degree sequence $\mathbf{D}_+ = (D_{1+},\ldots,D_{N+})'$ is*

1. Let $\mathbf{G}$ be an empty adjacency matrix.

2. If $\mathbf{D}_+ = \mathbf{0}$ terminate with output $\mathbf{G}$

3. Choose the agent i with minimal positive degree D_{i+}.

4. Construct a list of candidate partners $J = \{j \neq i \,:\, \mathbf{G}_{ij} = \mathbf{G}_{ji} = 0 \text{ and } \ominus_{i,j} \mathbf{D}_+ \text{ graphical}\}$.

5. Pick a partner $j \in J$ with probability proportional to its degree in $\mathbf{D}_+$.

6. Set $\mathbf{G}_{ij} = \mathbf{G}_{ji} = 1$ and update $\mathbf{D}_+$ to $\ominus_{i,j}\mathbf{D}_+$.

7. Repeat steps 4 to 6 until the degree of agent i is zero.

8. Return to step 2.

The input for Algorithm 1 is the target degree sequence $\mathbf{D}_+$ and the output is an undirected adjacency matrix $\mathbf{G}$ with $\mathbf{G}'\iota = \mathbf{D}_+$.

An example of how this algorithm might work, adapted from by Blitzstein and Diaconis (2011), is:

$$(3,2,2,2,1) \to (3,1,2,2,0) \to (2,0,2,2,0) \to (1,0,2,1,0) \to (0,0,1,1,0) \to (0,0,0,0,0)$$

The goal is to construct a five agent network with degree sequence $(3,2,2,2,1)$. The algorithm begins by choosing the 'last' agent, who has only a single link. The second agent is randomly chosen as her partner, generating the residual degree sequence $(3,1,2,2,0)$. Now the second agent has the lowest non-zero residual degree. She is selected and matched with the first agent. This leads to a new residual degree sequence of $(2,0,2,2,0)$. Now all three remaining agents who need links have the same residual degree. The algorithm randomly picks the first agent and matches her with the fourth agent. This leads to residual degree sequence $(1,0,2,1,0)$. The procedure concludes by matching agents one and three and finally three and four.

Note the check for graphicality has real bite. Say that in step 2, agent two was matched instead with agent three, leading to the residual degree sequence of $(3,0,1,2,0)$. This degree

sequence is not graphic. Indeed on the second step agent two is matched with agent one with probability one.

Blitzstein and Diaconis (2011) discuss how the recursive formulation of Havel-Hakimi check can be used to speed up the algorithm.

Importance sampling

Let $\mathcal{Y}_{N,\mathbf{d}_+}$ denote the set of all possible sequences of links outputted by Algorithm 1 given input $\mathbf{D}_+ = \mathbf{d}_+$. Let $\mathcal{G}(Y)$ be the adjacency matrix induced by link sequence Y. Let Y and Y' be two different sequences produced by the algorithm. These sequences are equivalent if their "end point" adjacency matrices coincide (i.e., if $\mathcal{G}(Y) = \mathcal{G}(Y')$). We can partition $\mathcal{Y}_{N,\mathbf{d}_+}$ into a set of equivalence classes, the number of such classes coincides with the number of feasible networks with degree distribution $\mathbf{D}_+$ (i.e., with the cardinality of $\mathbb{D}_{N,\mathbf{d}_+}$). Let $c(Y)$ denote the number of possible link sequences produced by Algorithm 1 that produce Y's end point adjacency matrix (i.e., the number of different ways in which Algorithm 1 can generate a given adjacency matrix).

Let $i_1, i_2, \ldots, i_M$ be the sequence of agents chosen in step 3 of Algorithm 1 in which Y is the output. Let $a_1, \ldots, a_m$ be the degree sequences of $i_1, \ldots, i_M$ at the time when each agent was first selected in step 3, then

$$c(Y) = \prod_{k=1}^{M} a_k!$$

Let $\sigma(Y)$ be the probability that Algorithm 1 produces link sequence Y. Note that $\sigma(Y)$ is easy to compute. Each time the algorithm choose a link in step 5 record the probability with which it was chosen (i.e., the residual degree of the chosen agent divided by the sum of the residual degrees of all agents in the choice set). The product of all these probabilities equals $\sigma(Y)$.

Let $f(\mathbf{G})$ be some function of the adjacency matrix and consider the expected value (using

the Law of Iterated Expectations)

$$\begin{aligned}
\mathbb{E}\left[\frac{\pi\left(\mathcal{G}\left(Y\right)\right)}{c\left(Y\right)\sigma\left(Y\right)}f\left(\mathcal{G}\left(Y\right)\right)\right] &= \sum_{y\in\mathbb{Y}_{N,\mathbf{d}}}\frac{\pi\left(\mathcal{G}\left(y\right)\right)}{c\left(y\right)\sigma\left(y\right)}f\left(\mathcal{G}\left(y\right)\right)\sigma\left(y\right)\\
&= \sum_{y\in Y_{N,\mathbf{d}}}\frac{\pi\left(\mathcal{G}\left(y\right)\right)}{c\left(y\right)}f\left(\mathcal{G}\left(y\right)\right)\\
&= \sum_{\mathbf{g}\in\mathbb{D}_{N,\mathbf{d}}}\sum_{\{y:G(y)=\mathbf{g}\}}\frac{\pi\left(\mathbf{g}\right)}{c\left(y\right)}f\left(\mathbf{g}\right)\\
&= \sum_{\mathbf{g}\in\mathbb{D}_{N,\mathbf{d}}}\pi\left(\mathbf{g}\right)f\left(\mathbf{g}\right)\\
&= \mathbb{E}_{\pi}\left[f\left(\mathbf{G}\right)\right].
\end{aligned}$$

The ratio $\pi\left(\mathbf{G}\left(Y_t\right)\right)/c\left(Y_t\right)\sigma\left(Y_t\right)$ is called the likelihood ratio or the importance weight. If we set $f\left(\mathbf{G}\right)$ to the constant function we see that the expected value of this weight is one. This suggests the analog estimator

$$\hat{\mu}_{f(\mathbf{G})} = \left[\sum_{t=1}^{T}\frac{\pi\left(\mathbf{G}\left(Y_t\right)\right)}{c\left(Y_t\right)\sigma\left(Y_t\right)}\right]^{-1}\times\sum_{t=1}^{T}\frac{\pi\left(\mathbf{G}\left(Y_t\right)\right)}{c\left(Y_t\right)\sigma\left(Y_t\right)}f\left(\mathbf{G}\left(Y_t\right)\right).$$

Setting $\pi\left(\mathbf{G}\right)=1$ we get a procedure for estimating the expectation of $f\left(\mathbf{G}\right)$ when $\mathbf{G}$ is drawn uniformly from $\mathbb{D}_{N,\mathbf{d}_+}$.

4 Future research directions

The analysis of networks has always been a multi-disciplinary endeavor. Economists are relative latecomers to this project. This survey has been deliberately eclectic and biased toward recent work done by economists. This work has not been undertaken in a vacuum. Economists interested in studying networks would be well-advised to read widely. Goldenberg, Zheng, Fienberg and Airoldi (2009) provide a monograph-length review of the literature from the perspective of statistics and machine learning. Snijders (2011) surveys the quantitative sociology literature.

At the same time there is tremendous latitude to approach network data from first principles or so-called "fresh eyes". In my view there is not one obvious "correct" way to formulate a network formation model and much work remains to be done. At this stage it seems apparent that a sizable component of empirical research on networks will be computationally complex. Ideas from discrete math, computer science, and Bayesian MCMC estimation have all proved to be very useful in work done thus far.

In thinking about identification, ideas from the recent literature on games, as well as the more established literature on dynamic panel data, have led to valuable insights. In both case the combinatoric complexity of networks precludes a direct application of methods from these literatures in all but the very simplest of cases. At the same time clever exploitation of various peculiarities and symmetries in the network formation problem can lead to tractable procedures.

This survey has not emphasized special purpose models (e.g., Currarini, Jackson and Pin, 2010). In some settings, for example those often encountered in industrial organization, substantial additional information may be available about the form of agents' objective functions, the timing of decisions and so on. Building empirical models that fully exploit all this extra information can be fruitful, both for expanding subject area knowledge and for methodological advancement. Indeed an important component of research by economists should involve modeling real-world datasets coherently, even if realistic models are only aspirational at the present time. Simple models are okay.

References

[1] Albert, Reka and Albert-László Barabási. (2002). "Statistical mechanics of complex networks," *Review of Modern Physics* 74 (1): 47 - 97.

[2] Angrist, Joshua D. and Kevin Lang. (2004). "Does school integration generate peer effects? Evidence from Boston's Metco program," *American Economic Review* 94 (5): 1613 - 1634.

[3] Angrist, Joshua D. (forthcoming). "The perils of peer effects," *Labour Economics.*

[4] Bala, Venkatesh and Sanjeev Goyal. (2000). "A noncooperative model of network formation," *Econometrica* 68 (5): 1181 - 1229.

[5] Ballester, Coralio and Yves Zenou. (forthcoming). "Key player policies when contextual effects matter," *Journal of Mathematical Sociology.*

[6] Ballester, Coralio., Antoni Calvó-Armengol and Yves Zenou. (2006). "Who's who in networks. wanted: the key player," *Econometrica* 74 (5): 1403 – 1417.

[7] Banerjee, Abhijit, Arun G. Chandrasekhar, Esther Duflo, Matthew O. Jackson. (2013). "The diffusion of microfinance," *Science* 341 (6144): 363 - 370.

[8] Barabási, Albert-László and Réka Albert. (1999). "Emergence of scaling in random networks," *Science* 286 (5439): 509 - 512.

[9] Blitzstein, Joseph and Persi Diaconis. (2011). "A sequential importance sampling algorithm for generating random graphs with prescribed degrees," *Internet Mathematics* 6 (4): 489 - 522.

[10] Bollobas, Bela. (2013). *Modern Graph Theory.* Springer: New York.

[11] Bonacich, Phillip. (1987). "Power and centrality: a family of measures," American Journal of Sociology 92 (5): 1170 - 1182.

[12] Bonacich, Phillip and Philip Lu. (2012). *Introduction to Mathematical Sociology.* Princeton, N.J.: Princeton University Press.

[13] Bloch, Francis and Matthew O. Jackson. (2007). "The formation of networks with transfers among players," *Journal of Economic Theory* 113 (1): 83 -110.

[14] Blume, Lawrence E., William A. Block, Steven N. Durlauf and Yannis M. Ioannides. (2011). "Identification of social interactions," *Handbook of Social Economics* 1B: 853 - 964 (J. Benhabib, A. Bisin, & M. Jackson, Eds.). Amsterdam: North-Holland.

[15] Blume, Lawrence E., William A. Brock, Steven N. Durlauf and Rajshri Jayaraman (2013). "Linear social interaction models," *NBER Working Paper No. 19212.*

[16] Bramoulle, Yann, Habiba Djebbari and Bernard Fortin. (2009). "Identification of peer effects through social networks," *Journal of Econometrics* 150 (1): 41 - 55.

[17] Bresnahan, Timothy F. and Peter C. Reiss. (1991). "Empirical models of discrete games," *Journal of Econometrics* 48 (1-2): 57 - 81.

[18] Brock, William A. and Steven N. Durlauf. (2001a). "Discrete choice with social interactions," *Review of Economic Studies* 68 (2): 235 - 260.

[19] Brock, William A. and Steven N. Durlauf. (2001b). "Interactions-based models," *Handbook of Econometrics* 5: 3297 - 3380 (J.J. Heckman & E. Leamer, Eds.). Amsterdam: North-Holland.

[20] Burt, Ronald S. (1984). "Network items and the general social survey," *Social Networks* 6 (4): 293 - 339.

[21] Card, David. (1995). "Earnings, school, and ability revisited," *Research in Labor Economics* 14(1): 23 - 48 (S.W. Polachek, Ed.). Greenwich, CT: JAI Press Inc.

[22] Card, David and Jesse Rothstein. (2007). "Racial segregation and the black–white test score gap," *Journal of Public Economics* 91 (11–12): 2158 – 2184.

[23] Case, Anne C. and Lawrence F. Katz. (1991). "The company you keep: the effects of family and neighborhood on disadvantaged youths," *NBER Working Paper No. 3705.*

[24] Chamberlain, Gary. (1984). "Panel data," *Handbook of Econometrics* 2: 1247 - 1318 (Z. Griliches & M. D. Intriligator, Eds.). Amsterdam: North-Holland.

[25] Chamberlain, Gary. (1985). "Heterogeneity, omitted variable bias, and duration dependence," *Longitudinal Analysis of Labor Market Data*: 3 - 38 (J.J. Heckman & B. Singer, Eds.). Cambridge: Cambridge University Press.

[26] Chandrasekhar, Arun G. and Matthew O. Jackson. (2014). "Tractable and consistent random graph models," *NBER Working Paper No. 20276.*

[27] Charbonneau, Karyne B. (2014). "Multiple fixed effects in binary response panel data models," *Bank of Canada Working Paper 2014-17.*

[28] Chatterjee, Sourav, Persi Diaconis and Allan Sly. (2011). "Random graphs with a given degree sequence," *Annals of Applied Probability* 21 (4): 1400 - 1435.

[29] Christakis, Nicholas A., James H. Fowler, Guido W. Imbens, Karthik Kalyanaraman. (2010). "An empirical model of strategic network formation," *NBER Working Paper No. 16039.*

[30] Ciliberto, Federico and Elie Tamer. (2009). "Market structure and multiple equilibria in airline markets," *Econometrica* 77 (6): 1791 - 1828.

[31] Comola, Marghertia and Marcel Fafchamps. (forthcoming). "Testing unilateral and bilateral link formation," *Economic Journal.*

[32] Cox, D. R. (1958). "The regression analysis of binary sequences," *Journal of the Royal Statistical Society B* 20 (2): 215 - 241.

[33] Currarini, Sergio, Matthew O. Jackson and Paolo Pin. (2010). "Identifying the roles of race-based choice and chance in high school friendship network formation," *Proceedings of the National Academy of Sciences* 107 (11): 4857 - 4861.

[34] Del Genio, Charo I., Hyunju Kim, Zoltan Toroczkai and Kevin Bassler. (2010). "Efficient and exact sampling of simple graphs with given arbitrary degree sequence," *Plos One* 5 (4): e100012.

[35] De Weerdt, Joachim. (2004). "Risk-sharing and endogenous network formation," *Insurance Against Poverty*: 197 - 216 (Dercon, Stefan, Ed.). Oxford: Oxford University Press.

[36] De Weerdt, Joachim and Marcel Fafchamps. (2011). "Social identity and the formation of health insurance networks," *Journal of Development Studies* 47 (8): 1152 - 1177.

[37] Dutta, Bhaskar, Sayantan Ghosal and Debraj Ray. (2005). "Farsighted network formation," *Journal of Economic Theory* 122 (2): 143 - 164.

[38] Erdös, Paul and Tibor Gallai. (1961). "Graphen mit punkten vorgeschriebenen grades," *Matematikai Lapok* 11: 264 - 274.

[39] Faust, Katherine. (2007). "Very local structure in social networks," *Sociological Methodology* 37 (1): 209 - 256.

[40] Freeman, Linton. (2000). "Visualizing social networks," *Journal of Social Structure* 1 (1).

[41] Gaviria, Alejandro and Steven Raphael. (2001). "School-based peer effects and juvenile behavior," *Review of Economics and Statistics* 83 (2): 257 - 268.

[42] Glaeser, Edward L. and Jose A. Scheinkman. (2001). "Measuring social interactions," *Social Dynamics*: 83 - 132 (S.N. Durlauf & H.P. Young, Eds.). Cambridge, MA: The MIT Press.

[43] Glaeser, Edward L. and Jose A. Scheinkman. (2003). "Non-market interactions," *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress* 1: 339 - 369 (M. Dewatripont et al., Eds.). Cambridge: Cambridge University Press.

[44] Goldenberg, Anna, Alice X. Zheng, Stephe E. Fienberg and Edoardo M. Airoldi. (2009). "A survey of statistical network models," *Foundations and Trends in Machine Learning* 2 (2): 129 - 233.

[45] Goldsmith-Pinkham, Paul and Guido W. Imbens. (2013). "Social networks and the identification of peer effects," *Journal of Business and Economic Statistics* 31 (3): 253 - 264.

[46] Goyal, Ravi, Joseph Blitzstein and Victor de Gruttola. (2014). "Sampling networks from their posterior predictive distribution," *Network Science* 2 (1): 107 - 131.

[47] Graham, Bryan S. (2008). "Identifying social interactions through conditional variance restrictions," *Econometrica* 76 (3): 643 - 660.

[48] Graham, Bryan S. (2011). "Econometric methods for the analysis of assignment problems in the presence of complementarity and social spillovers," *Handbook of Social Economics* 1B: 965 - 1052 (J. Benhabib, A. Bisin, & M. Jackson, Eds.). Amsterdam: North-Holland.

[49] Graham, Bryan S. (2012). "Homophily and transitivity in dynamic network formation," *In Progress*, UC - Berkeley.

[50] Graham, Bryan S. (2013). "Comment on "Social networks and the identification of peer effects" by Paul Goldsmith-Pinkham and Guido W. Imbens," *Journal of Business and Economic Statistics* 31 (3): 266 - 270, 2013.

[51] Graham, Bryan S. (2014). "An empirical model of network formation: detecting homophily when agents are heterogenous," *NBER Working Paper w20341.*

[52] Hakimi, S. L. (1962). "On realizability of a set of integers as degrees of the vertices of a linear graph. I," *Journal of the Society for Industrial and Applied Mathematics* 10 (3): 496 – 506.

[53] Handcock, Mark S., Adrian Raftery and Jeremy M. Tantrum. (2007). "Model-based clustering for social networks," *Journal of the Royal Statistical Society A* 170 (2): 301 - 354.

[54] Havel, Václav J. (1955). "A remark on the existence of finite graph," *Časopis Pro Pěstování Matematiky* 80: 477 - 480.

[55] Heckman, James J. (1977). "Sample selection bias as a specification error," *Econometrica* 47 (1): 153 - 161.

[56] Heckman, James J. (1978a). "Dummy endogenous variables in a simultaneous equation system," *Econometrica* 46 (4): 931 - 959.

[57] Heckman, James. J. (1978b). "Simple statistical models for discrete panel data developed and applied to test the hypothesis of true state dependence against the hypothesis of spurious state dependence," *Annales de l'inséé* 30-31: 227 - 270.

[58] Heckman, James. J. (1981a) "Heterogeneity and state dependence," *Studies in Labor Markets*: 91 - 139 (S. Rosen, Ed.). Chicago: University of Chicago Press.

[59] Heckman, James. J. (1981b). "Statistical models for discrete panel data," *Structural Analysis of Discrete Data and Econometric Applications*: 114 - 178 (C.F. Manski & D.L. McFadden, Eds.). Cambridge, MA: The MIT Press.

[60] Heckman, James. J. (1981c). "The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process,"

Structural Analysis of Discrete Data and Econometric Applications: 179 - 195 (C.F. Manski & D.L. McFadden, Eds.). Cambridge, MA: The MIT Press.

[61] Hellmann, Tim. (2013). "On the existence and uniqueness of pairwise stable networks," *International Journal of Game Theory* 42 (1): 2111 - 237.

[62] Honoré, Bo E. and Ekaterini Kyriazidou. (2000). "Panel data discrete choice models with lagged dependent variables," *Econometrica* 68 (4): 839 - 874.

[63] Horn, Roger A. and Charles R. Johnson. (2013). *Matrix Analysis, 2nd. Ed.* Cambridge: Cambridge University Press.

[64] Ioannides, Yannis M. and Jeffrey E. Zabel. (2008). "Interactions, neighborhood selection and housing demand," *Journal of Urban Economics* 63 (1): 229 - 252.

[65] Jackson, Matthew O. (2008). *Social and Economic Networks.* Princeton, NJ: Princeton University Press.

[66] Jackson, Matthew O. (2014). "Networks and the identification of economic behaviors," *Mimeo*, Stanford University.

[67] Jackson, Matthew O. and Brian W. Rogers. (2007a). "Relating network structure to diffusion properties through stochastic dominance," *B.E. Journal of Theoretical Economics* 7 (1) (Advances), Article 6.

[68] Jackson, Matthew O. and Brian W. Rogers. (2007b). "Meeting strangers and friends of friends: how random are social networks?" *American Economic Review* 97 (3): 890 - 915.

[69] Jackson, Matthew O. and Alison Watts. (2002). "The evolution of social and economic networks," *Journal of Economic Theory* 106 (2): 265 - 295.

[70] Jackson, Matthew O. and Asher Wolinsky. (1996). "A strategic model of social and economic networks," *Journal of Economic Theory* 71 (1): 44 - 74.

[71] Jackson, Matthew O. and Yves Zenou (forthcoming). "Games on networks," *Handbook of Game Theory* 4 (P. Young & S. Zamir, Eds.). Amsterdam: North-Holland.

[72] Kolaczyk, Eric D. (2009). *Statistical Analysis of Network Data.* New York: Springer.

[73] Koster, Jeremy M. and George Leckie. (2014). "Food sharing networks in lowland Nicaragua: an application of the social relations model to count data," *Social Networks* 38: 100 - 110.

[74] Kranton, Rachel E. and Deborah F. Minehart. (2001). "A theory of buyer-seller networks," *American Economic Review* 91 (3): 485 - 508.

[75] Krivitsky, Pavel N., Mark S. Handcock, Adrian E. Raftery, and Peter D. Hoff. (2009). "Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models," *Social Networks* 31 (3): 204 - 213.

[76] Loury, Linda Datcher. (2006). "Some contacts are more equal than others: informal networks, job tenure, and wages," *Journal of Labor Economics* 24 (2): 299 - 318.

[77] Manski, Charles F. (1975). "Maximum score estimation of the stochastic utility model of choice," *Journal of Econometrics* 3 (3): 205 - 228.

[78] Manski, Charles F. (1987). "Semiparametric analysis of random effects linear models from binary panel data," *Econometrica* 55 (2): 357 - 362.

[79] Manski, Charles F. (1993). "Identification of endogenous social effects: the reflection problem," *Review of Economic Studies* 60 (3): 531 - 542.

[80] McFadden, Daniel L. (1973). "Conditional logit analysis of qualitative choice behavior," *Frontiers in Econometrics*: 105 - 142 (P. Zarembka, Ed.). New York: Academic Press.

[81] McPherson, Miller, Lynn Smith-Lovin and James M. Cook. (2001). "Birds of a feather: homophily in social networks," *Annual Review of Sociology* 27 (1): 415 - 444.

[82] McPherson, Miller, Lynn Smith-Lovin and Matthew E. Brashears. (2006). "Social isolation in America: changes in core discussion networks over two decades," American Sociological Review 71 (3): 353 - 375.

[83] Mele, Angelo. (2011). "A structural model of segregation in social networks," *Mimeo,* John Hopkins University.

[84] Milgram, Stanley (1967). "The small-world problem," *Psychology Today* 1 (1): 61 - 67.

[85] Miyaichi, Yuhei. (2013). "Structural estimation of a pairwise stable network with non-negative externality." *Mimeo*, Massachusetts Institute of Technology.

[86] Morris, Martina, Ann E. Kurth, Deven T. Hamilton, James Moody and Steve Wakefield. (2009). "Concurrent partnerships and HIV prevalence disparities by race: linking science and public health practice," *American Journal of Public Health* 99 (6): 1023 - 1031.

[87] Owen, Art. B. (2013). *Monte Carlo Theory, Methods and Examples* available online at `http://statweb.stanford.edu/ owen/mc/`.

[88] Read, Ronald C. and Corneil, Derek. G. (1977). "The graph isomorphism disease," *Journal of Graph Theory* 1 (4): 339 – 363.

[89] Rosen, Kenneth H. (2006). *Discrete Math and its Applications, 6th Ed.* Blacklick, OH: Mcgraw-Hill.

[90] Sacerdote, Bruce. (2014). "Experimental and quasi-experimental analysis of peer effects: two steps forward?" *Annual Review of Economics* 6: 253 – 272.

[91] Shalizi, Cosma Rohilla and Alessandro Rinaldo. (2013). "Consistency under sampling of exponential random graph models," *Annals of Statistics* 41 (2): 508 - 535.

[92] Sheng, Shuyang. (2012). "Identification and estimation of network formation games," *Mimeo*, University of Southern California.

[93] Sierksma, Gerard and Han Hoogeveen. (1991). "Seven criteria for integer sequences being graphic," *Journal of Graph Theory* 15 (2): 223 – 231.

[94] Snijders, Tom A. B. (1991). "Enumeration and simulation methods for 0-1 matrices with given marginals," *Psychometrika* 56 (3): 397 - 417.

[95] Snijders, Tom A.B. (2002). "Markov chain Monte Carlo estimation of exponential random graph models," *Journal of Social Structure* 3 (2).

[96] Snijders, Tom A.B. (2011). "Statistical models for social networks," *Annual Review of Sociology* 37 (1): 131 - 153.

[97] Stanton, Isabelle and Ali Pinar. (2012). "Constructing and sampling graphs with a prescribed joint degree distribution," *Journal of Experimental Algorithmics* 17.

[98] Tamer, Elie. (2003). "Incomplete simultaneous discrete response model with multiple equilibria," *Review of Economic Studies* 70 (1): 147 - 167.

[99] Toivonen, Riitta, Lauri Kovanen, Mikko Kivela, Jukka-Pekka Onnela, Jari Saramaki, Kimmo Kaski. (2009). "A comparative study of social network models: network evolution models and nodal attribute models," *Social Networks* 31 (4): 240 - 254.

[100] Wasserman, Stanley and Katherine Faust. (1994). *Social Network Analysis.* Cambridge: Cambridge University Press.

[101] Watts, Alison. (2001). "A dynamic model of network formation," *Games and Economic Behavior* 34 (2): 331 - 341.

[102] Watts, Duncan J and Steven H. Strogatz. (1998). "Collective dynamics of 'small-world' networks," *Nature* 393 (6684): 440 – 442.

[103] Wooldridge, Jeffrey M. (2005). "Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity," *Journal of Applied Econometrics* 20 (1): 39 – 5.

[104] Zhang, Jingfei and Yuguo Chen. (2013). "Sampling for conditional inference on network data," *Journal of the American Statistical Association* 108 (504): 1295 - 1307.