

Improving Prediction and Causal Inference with Graphical Methods and Models

Langche Zeng*

The George Washington University

(Working paper version: February 5, 2004)

*Department of Political Science, George Washington University. Washington, DC 20052. lzeng@gwu.edu. The author thanks Steve Finkle, Gary King, Eric Lawrence, Jim Lebovic, Carey Priebe, Edward Scheinerman, and Lee Sigelman for helpful comments and suggestions; and the National Science Foundation (SES-0318275) for research support. An earlier version of the paper was presented at the 2003 Annual Meeting of the American Political Science Association, August 2003, Philadelphia.

Abstract

Critical to reliable prediction and causal inference is understanding structural relationships in the social and political systems under study. Graphical models are naturally suited for conceptualizing and representing relationships. This paper introduces and synthesizes a large and disparate literature on different types of graphical models, with particular attention on recent developments in theories of causal graphs and models of random graphs for relational data, and discusses the adaptation, application, and extension of these graphical methods and models in political data analysis in general, and in the modeling of structural properties of international relations data in particular. Graphical models can improve prediction and causal inference in these data by facilitating systematic examination and modeling of characteristics of the international network, by guiding the identification of the causal structure in the system, and by adding flexibility in functional form approximation. Initial results from analyzing the 1947-1989 MID data provide strong evidence that properties of the system as a whole and that of individual states/dyads embedded in the system hold important explanatory and predictive power, and clearly reveal the interdependence among dyads sharing a common member.

1 Introduction

Graphical models are excellent tools for conceptualizing and representing relationships, such as relationships among the individual actors (persons, organizations, states, etc.) under study, or relationships among the variables in a model. Understanding the first type of relationship is critical particularly in the study of relational data, such as data on international conflict, and understanding the second type is at the core of causal inference aiming at clarifying the structural relationships among various quantities of interest in the system under study. Graphs also facilitate representation and interpretation of flexible statistical models such as neural networks that accommodate complex functional relationships in social science data. Recent developments in random graphs and related estimation strategies have provided promising new tools for improving statistical analysis of relational data, and recent advances in causal graph theories have offered a formal language for communicating and processing causal information in statistical analysis that greatly facilitates the identification of causal structures and the assessment of causal effects. Relational data abound in political science, especially in the area of international relations, and causal inference has always been a central goal of empirical political analysis. These new tools, still unexplored and largely unknown in political science, therefore hold great promise for the improvement of empirical analysis in the field.

The literature on different types of graphical methods and models is large and diverse, with different branches relatively disconnected. This paper provides a synthesis and a technically accessible introduction to the literature, and discusses the adaptation, application, and extension of these graphical methods and models in political data analysis in general, and in the modeling of structural properties of international relations data in particular. In studying international conflict, for example, the international network can be modeled as a random graph. Structural characteristics and tendencies of the network, which potentially hold important additional explanatory and predictive power beyond the usual individual/dyad level attribute variables, may then be represented by relevant graph theoretic measures and included in the statistical analysis. Theoretically important questions such as whether there is dependence among dyads sharing a common member can then be answered through formal statistical tests. Systematic measurement of the characteristics and depen-

dence structure of the network have largely eluded previous research. Initial empirical results from analyzing the 1947-1989 militarized international dispute (MID) data clearly demonstrate the utility of graphical methods and models, and offer findings of profound substantive significance. In causal inference, correct specification of both structural equation models and single equation models, some version of which is employed in virtually all empirical studies of political science, is aided by the causal graph, without which conditions ensuring unbiased estimation of causal effects are virtually impossible to check and must remain assumptions.

In what follows, Section 2 introduces some basic graph theory terminologies and discusses several types of important graphical models, in particular social networks and random graphs for relational data, and causal graphs for representing causal structures; Section 3 discusses graphs and random graphs for relational data in some detail, proposing the adaptation and extension of relevant concepts and techniques to the study of international relations data, and report initial results from application to the analysis of the MID data; Section 4 distills some key aspects of causal graph theory that have immediate and profound implications for causal inference with non-experimental data, showing how they may change the way studies are designed and data collected using the example of international conflict data; Section 5 discusses methodological extensions that accommodate special features of political data such as functional complexity and rareness of events in conflict data, and gives some preliminary results; Section 6 concludes.

2 Graphs and Graphical Models

2.1 Graphs

I begin by introducing some basic graphical terminologies and notations relevant to modeling relational data, causal structure, and functional mapping. Excellent introduction to graph theory can be found in, for example, Scheinerman (2000). A *graph* G is a pair $G = (V, E)$ where V is a finite set of *vertices* or *nodes*, and E is a set of *edges*, or 2-element subsets of V , representing a *relation* on V . If the relation is symmetric, such as geographic contiguity or involvement in dyadic conflict, then $\{u, v\} \in E$ implies $\{v, u\} \in E$, and the edge uv is undirected. If the relation is asymmetric,

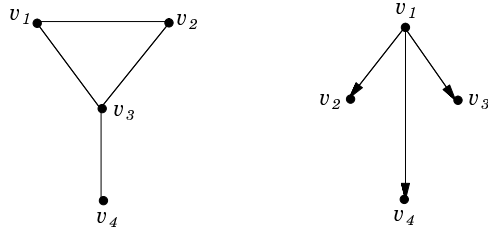


Figure 1: A graph (left) and a digraph (right)

such as initiation of conflict or causation, then $\{u, v\} \in E$ does not imply $\{v, u\} \in E$, and the edge uv is directed. When uv is an edge of G , we say u and v are *adjacent*, and write $u \sim v$ if the edge is undirected, or $u \rightarrow v$ if it is directed. If all edges of G are directed, G is called a directed graph or a *digraph*. Figure 1 depicts a graph and a digraph.¹ The World Wide Web, for example, can be viewed as a (di)graph, with files as vertices and links among files as directed edges. Obviously, plotting a graph is feasible or useful only when the node set is small. A matrix representation is generally more convenient. A graph $G = (V, E)$ can be represented by a square matrix X , such that for $i \in V$ and $j \in V$, $X_{ij} = 1$ if $(i, j) \in E$, and $X_{ij} = 0$ otherwise. That is, the ij th element of X is 1 if there is a link from node i to j . A graph can also have valued edges, in which case X_{ij} will take the value of the edge instead of simply indicating presence or absence of it. For a binary relation, X is also called the *adjacency matrix* of G .

When $u \sim v$, u and v are *neighbors* of each other. When $u \rightarrow v$, u is a *parent* of v and v is a *child* of u . Other terminologies of kinship, such as *ancestors* and *descendants*, is similarly used. A node in a digraph is a *root* if it has no parent. The set of all neighbors of a vertex u is denoted by $ne(u)$, the set of all its parents by $pa(u)$ and the set of all its children by $ch(u)$. For a subset A of V , $pa(A)$ denotes the collection of parents of all elements in A excluding A itself, and $ch(A)$ and $ne(A)$ are similarly defined.

A graph is *complete* if all pairs of nodes are adjacent. A *subgraph* of G induced by a subset of nodes $A \subseteq V$ has A as the node set and the subset of edges in E linking pairs of nodes in A as the edge set. A *clique* of G is a complete subgraph. A *path* is a sequence of edges such that each edge starts with the node ending the preceding edge. A pair of nodes are *connected* if there exists a path

¹I shall refer either as a “graph” where there is no confusion.

between them. A path is directed if it traces out a set of directed edges along the direction of the arrows. A *cycle* is a path that starts and ends on the same node. A directed graph that contains no directed cycles is a directed acyclic graph, or a *DAG*, such as the digraph in Figure 1.

2.2 Graphical Models

In the general definition of a “graph,” “node” and “relation” are completely abstract concepts. Various specific graphical models arise from combinations of particular types of nodes and relations. Three classes of models are considered in this paper.

When the nodes of G denote social actors and the edges of G denote social relations, G represents a *social network*, to which a vast literature is devoted, with most applications in sociology and psychology (Wasserman and Faust 1994; Wasserman and Galaskiewicz 1994.) Traditionally, the literature has focused largely on descriptive measures of graphical properties such as centrality, density, cohesiveness etc. Recent development has focused on the statistical analysis of *random graphs* in which the presence or absence of edges is assumed to be probabilistic. Much of this literature can shed light on relational data analysis in political science. For example, international conflict is a “relation” on the network of nation-states. In section 3 (and section 5.1) I shall discuss the application and extension of some of the important ideas from the social network literature, especially the recent literature on random graphs, to the analysis of international relations data. These ideas will allow us to identify and measure new explanatory variables that represent system-level characteristics and tendencies, will enable us to refine measurement of some existing variables that have been proven important, and will make the systematic modeling of dependence structures possible.

When the nodes of G denote variables in the model of some system under study, and the edges denote *causal* relations among the variables, G is a *causal graph* (Pearl 2000). The causal graph depicts the causal structure of the system under study, knowledge of which is critical to reliable causal inference. Model specification in most empirical work implicitly assumes knowledge of the pertinent part of this structure, but the assumptions are rarely made clear and justified. Indeed, in most situations the causal graph is really unknown. One widespread symptom of causal inference without adequate knowledge of the underlying causal structure is the seeming arbitrariness in the

selection of “control variables” or “covariates” in a regression model. In the large empirical literature studying the causal effect of democracy on international conflict, for example, it is difficult to find any two papers that use an identical set of control variables. In section 4, I discuss recent developments in causal graph theory pertinent to the discovery of the underlying causal structure and to the correct specification of control variables. These new techniques should greatly improve the quality of causal inference. I also show, in light of the causal graph, how a model aimed at causal inference can be fundamentally different from one aimed at forecasting/prediction, a distinction not made in the political science empirical literature and a source of deep confusion and model specification error.

When the nodes of G denote the variables, observed or latent, in a statistical model, and the edges denote input-output relations, G can represent complex functional mappings, such as in neural networks. Neural networks are capable of approximating arbitrary functional forms, and are useful tools for the analysis of social science data for which the underlying data generating functions are usually unknown and are likely complex (Zeng 1999, 2000a; Beck, King, and Zeng 2000, 2004; King and Zeng 2001a.) In section 5, I give suggestions on the use of neural networks in improving statistical modeling of random graphs of relational data, and in improving causal inference with the increasingly popular non-parametric approach of propensity score matching.

3 Modeling Structural Properties in International Relations Data

Structural and system-level analysis is critical to the study of international relations, especially in a global age (Waltz 1979, James 2002, Russett 2003a and 2003b). There is little theoretical doubt that structural properties of the international system can hold important explanatory and predictive power for the behavior of individual nation states and state dyads. Testing the theories with empirical data, however, has been impeded by difficulties in operationalizing, measuring, and modeling structural and system-level characteristics in a rigorous and systematic fashion. Quantitative measures of “system structure” are few and vary greatly from one study to another in conceptualization and construction, and capturing endogenous dependence structures systematically in statistical models has largely eluded previous researchers.

Graphical methods and models can help improve this situation in two ways: by providing unambiguous, precise definitions of an array of structural characteristics of the system that are easy to measure, and by allowing systematic modeling of endogenous dependence structures in random graphs representing statistical relational data. I discuss these possibilities in the context of studying international and civil conflict, and report initial results from analyzing the 1947-1989 MID data.

3.1 Measuring Structural Characteristics

Let the $N \times N$ square matrix X denote a graph of the international network, where N is the number of states in the network, and $X_{ij} = 1$ if there is a given relation, or a tie, from country i to j , 0 otherwise. Obviously, $X_{ij} = X_{ji}$ if the relation in question is symmetric. When more than one relation is considered, we have a *multi-graph*, and X may denote a three dimensional array $X = [X_{ijm}]$ such that $X_{ijm} = 1$ if there is a type m relation from country i to j , where $m \in R$, the set of relations.² Graph characteristics are functions of elements of X .

Among other things, graph theory provides ideas on measuring such characteristics as centrality, the prestige and prominence of actors, and the density and centralization tendencies of the graph as a whole; on defining cohesive subgroups; and on measuring the equivalence of actors and grouping them into blocks occupying different roles and structural positions in the network. Some of the measures, such as individual centrality, are actor-level measures, but they measure properties of actors not as independent units but embedded in the international network. In this sense they are “structural” measures. Most actor-level variables used in conflict studies, such as levels and growth rates of national demographic and economic data, characteristics of political systems, and military capabilities, do not have this feature. Perhaps this is partly the reason why these variables have not proved very useful in explaining conflict (Russett 2003a).³ It is likely that what matters is the relative position of a state in the international system, rather than attributes of the state as an independent entity.

²Where there is no confusion, I omit the subscript m for notational simplicity.

³One exception is “major power” status. But this is hardly a truly “explanatory” variable. It simply states that a few named states are different from the rest.

3.1.1 Structural Characteristics

Degree and actor centrality: In a undirected graph, the *degree* of a node i is the number of nodes adjacent to it: $\sum_j X_{ij} = \sum_j X_{ji}$. In a digraph, the *in-degree* and *out-degree* of a node i are the numbers of nodes adjacent to and from it respectively: $\sum_j X_{ji}$ and $\sum_j X_{ij}$. For a valued graph, the degree is the average value of the edges incident to the node. Degrees also provide a basis for one type of *centrality* measure, $\sum_j X_{ij}/N$. An actor with more ties with other actors is more active and more visible.⁴

If the relation is conflict involvement, a country with a high degree/centrality is more conflict-prone. If the relation is conflict initiation, then out-degree/centrality reflects tendencies of aggression and risk taking behavior. A potentially useful indicator for conflict prediction then can be constructed as the average of an actor's centrality scores, for example averaged over a specified number of years.

Appropriate centrality scores from such international networks as trade flows, cultural exchange, diplomatic relations, military interventions, and conjoint treaty and international organization memberships can be used, together with the usual non-network embedded country level data, in constructing measures of national "power," a central concept in international relations that has eluded easy measurement. Because power is a relative concept, a power index constructed using information on the relative importance of a country in the international system is more meaningful than the usual measures.

In a similar fashion, centrality scores can also be used to refine the measurement of other variables such as "trade openness" or "open economy," a key variable used in the study of state failures (e.g., King and Zeng 2001a). The standard measure of trade openness is the country's total trade as a percentage of its GDP. Thus two countries could have the same openness score even though one may have only one trading partner, and the other one hundred. The refined scores would meaningfully differ in such cases.

Degree/centrality measures are also useful in handling spatial dependence. Having states expe-

⁴Centrality measure can also be constructed based on other concepts such as closeness, betweenness, and information flow (Wasserman and Faust 1994, Bonacich 1987).

riencing conflicts as neighbors may increase the likelihood of conflict involvement due to spill-over effects, and having democratic states as neighbors seems to help prevent wars (Russett 2003b; Gleditsch and Ward 2000, 2001). In their effort to model this effect, Ward and Gleditsch (N.D.) use two new variables, the average level of democracy in proximate countries and the number of neighboring countries experiencing conflict. Both are in fact degree measures in two graphs of the international system: one is a digraph with the relation “neighboring a state in conflict,” the other a valued digraph of geographic proximity, the value of the link being the democracy level of the neighboring state. Thinking in terms of the graphs immediately suggests other potentially useful measures such as closeness-based centrality scores.

Density and centralization: The density of a graph is the ratio of the number of edges present to the total number of edges possible: $\sum_{i,j} X_{ij}/N(N-1)$. For conflict data it is a measure of rareness of events. The centralization of a graph measures the extent to which the system is centralized. One measure of centralization is simply the variance of the actor-level degrees or centrality scores. This records the variability and spread of actor-level centrality scores. In a highly centralized system, some actors are much more central and important than others who may be viewed as residing in the periphery of the system. Many other measures of graph centralization are possible (Wasserman and Faust, Chapter 5.)

System-level characteristics play an important role in explaining and forecasting conflict behavior (Bueno de Mesquita 1975). The most widely used measure of “structure” of the international system in existing studies simply indicates whether the system is “bipolar” or “multi-polar,” that is, whether there are two or more “great powers,” a status designated to certain countries or clusters of countries. One problem of this measure is that it is gross and insensitive to finer features of the international system. Most international systems in the modern era have been multi-polar, with the cold war era being bipolar. There is also lack of theoretical consensus on the operationalization of “polarity” (Russett 2003a). Making use of centralization scores from pertinent international networks (such as those considered in constructing measures of “power”) can greatly enrich the measurement of the centralization structure of the international system. Another advantage of these measures is that they are also defined for any subgraph, so we could for example construct regional measures

of density and centralization to test the presence of any regional effect. With cross-sectional time series data commonly used in conflict studies, we can construct measures that vary over both time and subgroups of states, however defined.

In a similar fashion, density measures on networks or subnetworks of such relations as trade, co-membership in international organizations, and neighboring a democratic state can provide informative measures, at the system level, of economic interdependence, growth of international norms and institutions, and degree of democratization. Current measures of these three key suppressors of violence are largely restricted to the dyadic level (Russett and Oneal 2001, Russett 2003b).

Cohesive subgroups: Cohesive subgroups can be defined based on mutuality of ties, closeness of subgroup members, frequency of ties among members, or the frequency of within-group ties relative to between-group ties (Wasserman and Faust 1994, Chapter 7). For example, cliques are cohesive subgroups in the sense of having complete mutuality of ties among group members. Different definitions capture different specific properties of cohesive subgroups. I use a definition based on nodal degrees, or frequency of ties among members, to fix ideas. Specifically, cohesive subgroups are defined as *k*-cores of the graph. A *k*-core is a subgraph in which each node is adjacent to at least *k* other nodes in the subgraph. The appropriate value of *k* depends on the substantive context.

Cohesive subgroups in the network of treaty co-membership, for example, have direct relevance to the study of coalition formation and alliance configuration in the international system. The substantive meaning of cohesive subgroups in other pertinent networks such as trading is similarly clear. Identification of cohesive subgroups is useful in several ways. It can distinguish actors who belong to the same cohesive subset, and those who do not. Nations in the same coalition are less likely to fight with each other. It allows further comparison of frequency of ties among different subgroups to see which subgroups are more likely to see their members interact. It allows comparison of attributes of states in different cohesive subgroups to see how the groups differ. And finally, it can foster structural knowledge of the whole network: if the cohesive groups are largely overlapping and contain most of the actors, the network is more cohesive, otherwise the network is more fragmented.

Roles and subgroup positions: Another approach to grouping actors in a network is according to

the *roles* they play and *positions* they occupy in the network. Actors in different positions have different patterns of ties with other actors. Those in the same positions are “equivalent” in some sense. One definition of “equivalence” is *structural equivalence*. Two actors i and j are structurally equivalent if they have identical ties to and from all other actors on all relations under consideration, that is, $X_{ikr} = X_{jkr}$ and $X_{kir} = X_{kjr}$ for any k in the node set and any r in the relation set. In practice it is unlikely to observe exactly equivalent actors, but we can seek to measure the degree to which they approach equivalence by comparing the similarity of the entries in X for the two actors. The comparison is usually based on measures of either correlation or distance. Once pairwise similarity measures are obtained, they can be used to partition the actors into groups/positions and to construct spatial representations of their structural equivalence, using standard data analysis techniques such as hierarchical clustering and multidimensional scaling. Once positions are identified, patterns of within and between group ties can be examined. A *density table* for example can be constructed, which has positions as its rows and columns, with values in the table recording the proportions of ties that are present from actors in the row position to actors in the column position (e.g., Snyder and Kick 1979).

In an influential paper studying the structural properties of the international system, Bueno de Mesquita (1975) proposes a correlation type measure, τ_b , of alliance portfolio similarity and uses the measure to group states into “poles.” He then tests the impact of such factors as number of poles, tightness of poles, and power distribution among poles on amount of war. “ τ_b ” has since been widely used in the international relations literature. Recently Signorino and Ritter (1999) discuss the problems of this measure and propose an alternative measure “ S ” based on distances. It is easy to see that τ_b and S are in fact measures of structural equivalence in the graph of the alliance network. Casting them into this standard framework immediately makes available a whole array of other relevant concepts and techniques developed which may prove productive for the study of the international system. For example, alternative measures could be constructed using other definitions of “equivalence,” such as regular equivalence and local role equivalence (see Wasserman and Faust 1994, Chapter 12 for an overview). Patterns of interactions among the poles can be studied using the density table, and a variable reflecting block effects can be constructed using the densities. This

variable may have significant explanatory/predictive power in the analysis of dyadic conflict data.

3.2 Modeling Endogenous Dependence Structure

Ideas discussed above lead to the identification and measurement of potentially important explanatory variables reflecting structural properties of the international system. I now turn to the statistical modeling of endogenous dependence structures in observed relational data. Interdependence among actors and actions in international relations is widely recognized (e.g., Beck and Tucker 1996, Beck et al. 1998, Signorino 1999, Beck and Katz 2001, Green et al. 2001, Gleditsch and Ward 2000, King 2001, Oneal and Russett 2001, Russett 2003a and 2003b, Ward and Gleditsch n.d.) The recent symposium in *International Organization* (2001, 55(2)) is largely centered on this issue. As King (2001) explains, for the prominent case of dyadic conflict data:

“Unlike, say, simple random survey sampling, dyadic observations in international conflict data have *complex dependence structure*. In a survey, observations 1 and 2 are two people who almost surely have never met and have no relationship. In contrast, in dyadic data, observation 1 may be U.S.-Iraq; observation 2, U.S.-Iran; and observation 3, Iraq-Iran. the dependence among these separate observations is complicated, central to our theories about the international system, critical for our methodological analysis, and ignored by most previous researchers.” (p.498)

In concluding the paper, he points out that “an approach that extracts the most information will likely be one that directly models the unique structure of dyadic data. Unfortunately, no off-the-shelf model is available for these data.” (p.506)

Fortunately, recent developments in models for random graphs provide useful tools for modeling such complex dependence structure in relational data (Frank and Strauss 1986, Strauss and Ikeda 1990, Wasserman and Pattison 1996, Anderson et al. 1999, Wasserman and Pattison 2000.) A random graph is a graph with random edges, so that each element in X is a random variable.⁵ Observed dyadic conflict data can be viewed as realizations of the underlying random graph. Denote

⁵The randomness of a graph can also extend to the node set, but I do not consider this case as it is largely irrelevant for international relations data.

by $x = [x_{ij}]$ a realization of the graph X . Dependences among elements of X mean that $Pr(X = x) \neq \prod_{i \neq j} Pr(X_{ij} = x_{ij})$. However, this condition is rarely recognized in empirical models of international conflict. The most widely used standard logit model, for example, implicitly assumes independence, i.e., $Pr(X = x) = \prod_{i \neq j} Pr(X_{ij} = x_{ij})$.

So we need instead a general expression for $Pr(X = x)$ that allows any dependence structure among the elements of X , i.e., the dyads.⁶ A class of random graph models, known as “p*” models in the social networks literature, postulates a general log-linear model:

$$Pr(X = x) = \frac{\exp(\theta' z(x))}{c(\theta)} \quad (1)$$

where $z(x)$ is a vector of network statistics that are functions of elements of x , θ the vector of model parameters, and $c(\theta)$ a normalization constant that ensures that the probability distribution sums to 1.⁷ Note that $z(x)$ can contain *any* network characteristics and hence the model allows general patterns of interdependence among elements of X , or the dyads.

The choice of $z(x)$ is guided by substantive theory about the structure of the interdependence among dyads, and is greatly facilitated by the use of the *dependence graph*, D , that depicts this dependence structure, and the *Hammersley-Clifford theorem* (Besage 1974; Frank and Strauss 1986) that identifies the sufficient network statistics based on the dependence graph. In the dependence graph, all possible dyads constitute the node set, and there is a tie between two dyads if they are conditionally dependent given the remaining dyads. Different patterns of interdependence among the dyads lead to different connection patterns in the dependence graph. For example, if all pairs of dyads are conditionally independent, the dependence graph would be empty and have no ties. If the relation is directed (such as conflict initiation), dyadic independence would mean that only dyads sharing the same two members are conditionally dependent. The most commonly assumed dependence structure is *Markov dependence*, in which dyads sharing at least one member (i.e., “incident edges” in the original graph) are conditionally dependent. So, for example, U.S.-Iraq and

⁶I consider a single binary relation here. Extensions to multiple relations and valued relations are straightforward (e.g., Robins, et al. 1999; Pattison and Wasserman 1999).

⁷As written model (1) does not involve variables exogenous to the network, such as dyad attributes and statistics from networks of other relations, but these can be used along with $z(x)$. I omit them for notational simplicity.

U.S.-Iran would be conditionally dependent, as would be U.S.-Iraq and Iraq-Iran, and U.S.-Iran and Iraq-Iran.

The Hammersley-Clifford theorem states that cliques (single nodes or complete subgraphs) of the dependence graph are the *sufficient subgraphs* for the representation of $Pr(X = x)$, so that $z(x)$ only needs to contain indicators of these cliques. For example, when there is dyadic independence in an undirected graph and hence the dependence graph is empty, the cliques of D are just its single nodes, which are single dyads in the original graph G . For Markov dependence, it is easy to see that the cliques of D are just the *triangles* and *k-stars*, $k = 1, 2, \dots, N - 1$, of G . A triangle T_{ijk} is the set of the three edges (ij, jk, ki) . For example, U.S.-Iraq, Iraq-Iran, and Iran-U.S. form a triangle. A k -star $S_{i_0 i_1 \dots i_k}$ is the set of k edges $(i_0 i_1, i_0 i_2, \dots, i_0 i_k)$. For example, U.S.-Iraq is a 1-star, and U.S.-Iraq and U.S.-Iran form a 2-star. Assuming *homogeneity* in the sense that isomorphic graphs have the same probability,⁸ the theorem tells us that the pertinent network statistics in a dyadic independent network are just the number of dyads (so $z(x)$ has only one element). In a Markov dependence network, they are just the number of triangles and k -stars,⁹ so $z(x)$ has N elements. No tetrads or other more complicated subgraphs are needed. For the international conflict network high order stars are unlikely and many of the $N - 1$ terms on the stars will be irrelevant. For example, a simple model capturing both transitivity and clustering can include just the number of triangles and the number of 1-stars and 2-stars. This not only greatly simplifies the model but also eliminates the problem of number of parameters increasing with data, which leads to inconsistency of estimated parameters.

Having identified the elements in $z(x)$, I now turn to the estimation of model (1). Standard likelihood techniques are difficult to apply due to the presence of the normalization term. However, pseudo-likelihood approaches based on conditional probabilities are easy to implement (Strauss and

⁸This means that nodes of G are a priori indistinguishable so that only the structure of G , and not the labeling of its nodes, matters to $Pr(G)$. In the study of international conflict, this assumes that after taking into consideration of dependence structures and all other relevant actor, dyad and global level variables, country names per se do not contain information about the probability of conflict.

⁹It is intuitively appealing to replace, through re-parameterization and without loss of information, the number of k -stars with the number of nodes in G that are of degree k (Frank and Strauss 1986, p.836.)

Ikeda (1990).¹⁰ Denote X_{ij}^+ the graph with the edge ij forced to be present, X_{ij}^- the graph with the edge ij forced to be absent, and X_{ij}^C the graph with the edge ij “missing”. We have:

$$\begin{aligned}
Pr(X_{ij} = 1 | X_{ij}^C) &= \frac{Pr(X_{ij}^+)}{Pr(X_{ij}^+) + Pr(X_{ij}^-)} \\
&= \frac{\exp(\theta' z(x_{ij}^+))}{\exp(\theta' z(x_{ij}^+)) + \exp(\theta' z(x_{ij}^-))} \\
&= \frac{1}{1 + \exp(-\theta'(z(x_{ij}^+) - z(x_{ij}^-)))} \tag{2}
\end{aligned}$$

$$\begin{aligned}
Pr(X_{ij} = 1 | X_{ij}^C) &= \frac{Pr(X_{ij}^+)}{Pr(X_{ij}^+) + Pr(X_{ij}^-)} = \frac{\exp(\theta' z(x_{ij}^+))}{\exp(\theta' z(x_{ij}^+)) + \exp(\theta' z(x_{ij}^-))} \\
&= \frac{1}{1 + \exp(-\theta'(z(x_{ij}^+) - z(x_{ij}^-)))} \tag{3}
\end{aligned}$$

Expression (3) is identical to a logit model probability with $z(x_{ij}^+) - z(x_{ij}^-) = \delta(x_{ij})$ as the explanatory variables. $\delta(x_{ij})$ is constructed as the difference in network statistics when the edge ij changes from absent to present (for example, the difference in the number of triangles when dyad ij changes from peaceful relation to conflict involvement.) Strauss and Ikeda (1990) show that estimating such a logit model for (as if) independent x_{ij} 's through maximum likelihood gives identical parameters to maximizing the pseudo-likelihood function:

$$PL(\theta) = \prod_{i \neq j} Pr(X_{ij} = 1 | X_{ij}^C)^{x_{ij}} Pr(X_{ij} = 0 | X_{ij}^C)^{(1-x_{ij})} \tag{4}$$

This result makes the model extremely easy to estimate as logit routines are widely available in statistical packages. We only need to add the network statistics $\delta(x_{ij})$ capturing structural dependence to the usual list of explanatory variables in existing logit models. Interpretation of estimation results is intuitive. For example, a large positive parameter of the 2-star indicator would mean that a tie between a dyad is more likely to be present, if its presence increases the number of 2-stars in the network. In other words, network configurations with 2-star patterns are more likely to occur.

¹⁰This approach is similar in spirit to the use of conditional probabilities in the estimation of, for example, the Cox proportional hazard model, the fixed effect logit model, and auto-logistic models. In each case intractable terms are eliminated through the conditioning.

The statistical test of the parameter is usually based on the comparison of the likelihoods with and without the parameter (e.g., Wasserman and Pattison 1996).¹¹

3.3 Empirical Results

In this section, I apply some of the idea discussed above to the analysis of militarized international dispute (MID) data, taking the logit model used in Beck, King and Zeng (2000) as the “standard logit” model for comparison.¹² The data contain 23,529 interstate dyad-years between 1947 and 1989, with 976 (4.1%) of the cases being MIDs (coded 1.) The explanatory variables used in BKZ (2000) include “Contiguity” (whether the two members of the dyad are geographically contiguous), “Ally” (whether they are allies), “Similarity” (degree of similarity in the dyad’s foreign policy portfolios), “Asymmetry” (the degree of balance of power within the dyad), “Dema” and “Demb” (the degree of democratization of the dyad), and “Peace Years” (the number of years since the last conflict in the dyad). The “standard logit” model in table 1 uses this same set of explanatory variables, except that for the democracy variables, I use the minimum of the two (“min-dem”) instead of entering both, a practice that is common in the literature and that, in the linear logit model, makes better use of the information contained in the democracy variables.¹³

I first test the conjecture that properties of individual states/dyads embedded in the international network, in addition to attributes of them as independent entities, are important factors in explaining and predicting conflict behavior. In particular, I examine how past conflict tendencies of the dyad as members of the international community influence the current conflict behavior within the dyad. I measure conflict tendency of a country in the past by the *degree* of the country in the conflict graph for the international network, averaged over the past years. Thus, for a dyad in year 1980, for example, the “past” is up to 1979, and the degree measures for years up to 1979 are averaged to give

¹¹I note that statistical properties for the maximum pseudo-likelihood estimators are less well studied than the standard maximum likelihood estimators. Alternatively we can, at a much higher computational cost, employ Markov Chain Monte Carlo methods to simulate the network over the space of all possible graphs (e.g., Handcock 2000).

¹²In this initial analysis I employ the logit model as the first cut, assuming nearly linear effects of variables. Future work can use more refined models such as neural networks that accommodate complex functional forms.

¹³BKZ (2000)’s focus is on neural network models, for which it is better to use the two original variables and let the neural network to “discover” the true functional form.

the measure of conflict tendency in the past. This measure takes missing values for observations in the first year, 1947, of the data set, with a resulting valid observations of 23255 cases.

The results of adding the minimum of the two past conflict tendency measures for the dyad (“Min-Conf”) to the “standard logit” model are found in table 1, in the two columns to the right of “standard logit” model. This “Min-Conf” variable has a positive coefficient that is highly significant, and its addition to the model increases the log-likelihood at convergence from -3178.65 to -3066.48. In-sample test statistics can be unreliable, however, since it could be a result of overfitting. To confirm that the new measure really adds to the explanatory and predictive power of the model, I randomly split the data into two sets, one containing about 80% of the original data, used to estimate the models, and the other, about 20%, used to check out-of-sample performance. To compare performance, I use the ROC curve area as in Beck, King, and Zeng (2004). And as in BKZ (2004), I also estimate a model with only the “Peace Year” variable, found in table 1 in the two columns to the left of “standard logit” model, to serve as the base and to put the comparison in context. As the table shows, the standard logit model, by adding all the variables not in the base model, improves the ROC area on out-of-sample test data by .008; adding the past conflict tendency variable doubles this improvement to .017.

Figure 2 plots the marginal effects of several variables from this improved model. In each graph, the plot shows the probability of conflict as a function of one explanatory variable, holding constant all the others at values that give a relatively high ex ante probability of conflict.¹⁴ Contiguity, Democracy, and Peace-Years are among the variable that are shown to have the largest and most stable effects in the existing literature, but Figure 2 reveals that none of them has an impact nearly comparable to that of the new conflict tendency variable. While the standard variables can change the probability of conflict by at most about 20% as they vary over the range of their possible values, the new variable can lead to an increase in the probability of conflict from below 20% to close to 90% as it moves from its minimum value (0) to the highest (about 4) observed in the data.

I now turn to the test of the conjecture that there is Markov dependence in the graph of inter-

¹⁴BKZ (2000) shows that the effects of variables are larger and more stable when the ex ante probability of war is higher. As in BKZ (2000), I set the values of the variables at the median of each explanatory variable among observations with $Y = 1$.

Table 1: Models of International Conflict

| Variables | P.Y.-Base | | Standard Logit | | +Past Degree | | +Network Stats | |
|--------------------|-----------|---------|----------------|---------|--------------|---------|----------------|---------|
| | Coeff | P-value | Coeff | P-value | Coeff | P-value | Coeff | P-value |
| Peace Years | -.181 | 0.000 | -.153 | 0.000 | -.138 | 0.000 | -.135 | 0.000 |
| Contiguity | — | — | 1.046 | 0.000 | 1.179 | 0.000 | 1.411 | 0.000 |
| Ally | — | — | -.325 | 0.001 | -.227 | 0.020 | -.205 | 0.037 |
| Similarity | — | — | -.282 | 0.000 | -.084 | 0.336 | -.006 | 0.949 |
| Asymmetry | — | — | -.882 | 0.000 | -.329 | 0.017 | -.657 | 0.000 |
| Min-Dem. | — | — | -.045 | 0.000 | -.038 | 0.000 | -.036 | 0.000 |
| Min-Conf. | — | — | — | — | .868 | 0.000 | .5212 | 0.000 |
| D1star | — | — | — | — | — | — | -.675 | 0.000 |
| D2star | — | — | — | — | — | — | -.903 | 0.000 |
| D3star | — | — | — | — | — | — | -.910 | 0.000 |
| D4star | — | — | — | — | — | — | -.434 | 0.002 |
| D5star | — | — | — | — | — | — | -.472 | 0.000 |
| Constant | -1.619 | 0.000 | -1.786 | 0.000 | -2.698 | 0.000 | -1.85 | 0.000 |
| N | 23255 | | 23255 | | 23255 | | 23255 | |
| Log-likelihood | -3379.83 | | -3178.65 | | -3066.48 | | -2990.036 | |
| Out-of-sample ROC | .826 | | .834 | | .843 | | .859 | |
| ROC gain over base | 0 | | .008 | | .017 | | .033 | |

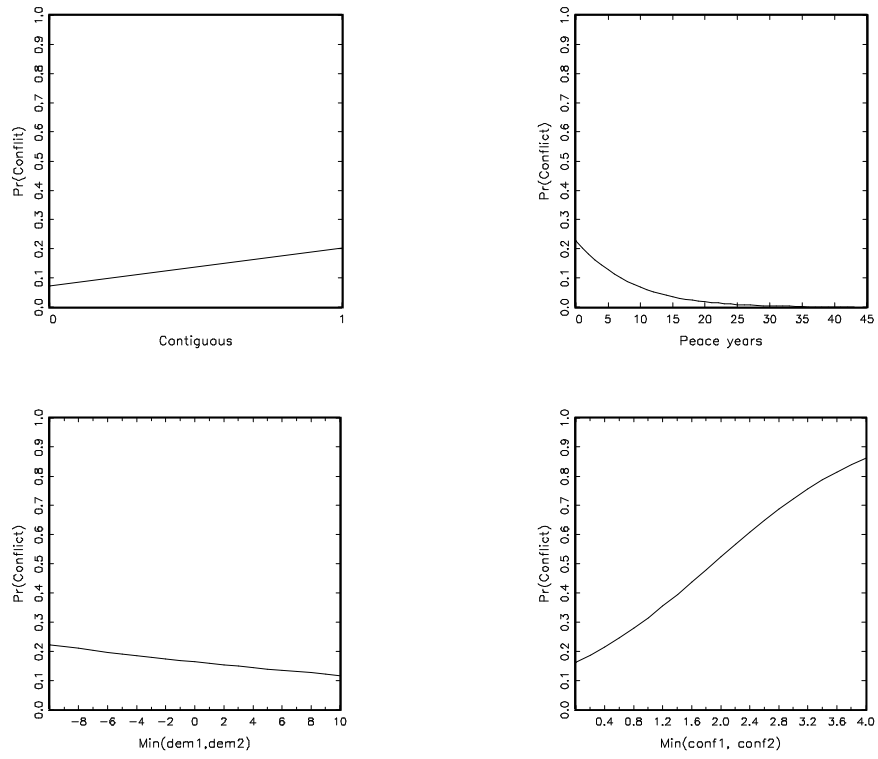


Figure 2: Marginal effects of input variables in the MID model.

national conflict, so that dyads sharing a common member are conditionally dependent. From the discussion in section 3, we see that with Markov dependence, changes in the number of triangles and k -stars are relevant graph statistics that enter the pseudo likelihood function of conditional probabilities. Since the meaning of transitivity in a conflict relation is unclear, I focus on the k -stars in the test.

The last two columns in table 1 report the estimation results with the k -star change statistics added to the model. The first five star statistics have negative and highly significant coefficients. The negative signs mean that graphs with less clustering are more likely to realize, and that higher order clustering is less likely than lower level clustering to occur. This appears consistent with the data, which sees very rare events, and even rarer are cases where one country is involved in multiple conflicts at the same time.

To confirm the results with out-of-sample test data, I estimate the model with the 80% subset, and examine generalization to the 20% test set. The network statistics capturing Markov dependence structure improve the ROC area over the base model by .033, four times as much as the standard logit model improves over the base model. This is clear evidence that the assumption of independence among dyads implicit in standard models is incorrect.

4 Causal Structure and Causal Inference with Observational Data

Causal inference is a central goal of empirical political science and the other social sciences. The vast majority of empirical work explicitly or implicitly involves clarifying causal relationships among a set of variables of interest and assessing the causal effects of one set of variables on another. While randomized experiments are ideally suited for the purpose of causal inference, social experiments are costly and subject to various constraints (such as noncompliance, issues of internal/external validity, ethical considerations, etc.) that make them difficult or infeasible in many cases. Empirical researchers have therefore largely relied upon non-experimental data and on the tools of probability theory and statistical inference. The language of probability calculus, however, is not designed for handling causality, but rather statistical association. This inherent difficulty is responsible for the blurring of the causal and statistical foundations of empirical studies and for

a variety of related practical issues, ranging from an unambiguous definition/notation for “causal effects” to the accurate identification of control variables (or “covariates”, or “confounders”) to be measured and adjusted in estimating causal effects.

To assess the effects of a variable x on another variable y using observational data, the most common practice in political science (and much of the other social sciences) is to estimate a regression type model (such as a logit or probit model when y is binary) with y as the dependent variable, and x and some other “control variables,” denoted z , as independent variables. In theory, the control variables are chosen to eliminate confounding of the relationship between x and y to ensure unbiased estimation. In practice unbiasedness conditions from standard statistical/econometric theories are difficult or impossible to check,¹⁵ and often z 's are chosen simply because substantive theories indicate that they may directly or indirectly have some causal effects on y . After estimation, the marginal effects of x on y , i.e., the change in y following a change in x computed from the model, with all other independent variables taking certain fixed values, are often interpreted as the causal effects of x on y .

Confusion reigns in this practice. First, different scholars studying the causal effects of the same x on the same y typically use different sets of control variables, guided by even slightly different substantive theories about what other variables may affect y . As Pearl (1997a) observes, “whether an adjustment for a given covariate z is appropriate in any given study continues to be decided informally, on a case-by-case basis, with the decision resting on folklore and intuition rather than on hard mathematics.” This practice in turn has led not only to changes in magnitude but often even to reversal of signs in estimated key relationships across different studies, a phenomenon known as *Simpson's Paradox*. This makes finding the “right” set of control variables the “Achilles heel of all social science inference because it is so very hard to do” (Brady 2002, p.1). Second, it is not clear whether the marginal effects are indeed measures of causal effects, since it is not clear whether the changes in x are *observed*, which could be a result of x being determined by its own causes, or are

¹⁵These conditions are part of the “specification assumptions” in econometrics which specify that x and the error terms are independent given the other control variables. In statistics they are called “conditional independence” conditions, or “strong ignorability” conditions in the potential response framework (Rosenbaum and Rubin 1983). These conditions do not provide a working test for the validation of control variable selection, thus essentially remain *assumptions*.

the results of external *interventions* on x . The two are not only qualitatively different but can differ greatly in quantity as well. Similarly, it is not clear whether the control variables are *observed* to take the specified values, or are *controlled* in the sense of intervention. This problem is particularly obvious in multiple equation models, where x can be literally endogenous. Third, do the marginal effects of the control variables z on y , which can be computed just the same way as that for x , measure the causal effects of z on y ? In other words, can the same model be used to infer about more than one causal relationship? It typically is, even though if the issue is raised explicitly most scholars would be intuitively wary. The common practice of controlling for all potential causes of y , however, does not mathematically differentiate between the key causal variable(s) x and the “controls”. Finally, a model for causal inference is often used for the purpose of forecasting y as well (or vice versa), but is a causal model (even if an optimal one) necessarily an optimal forecasting model, and vice versa? Intuitively the answer is no, since for the purpose of forecasting y , one would want to include all direct causes of y , a condition different from controlling for variables that eliminate confounding of the relationship between x and y for the purpose of causal inference. But without effective tools for identifying the latter, the common practice again confuses the two.

Fortunately, recent developments in causal graph theory exemplified by Pearl (2000) and Spirtes et al. (2000) shed light on these and related issues. That work provides a new language and a set of new tools naturally suited to causal inference, and promises to make possible the leap “from a century of statistics to an age of causation” (Pearl 1997a). In this section, I distill some of the key aspects of this development that have immediate relevance for empirical political and social science research in general, and the study of international/civil conflict in particular, laying the theoretical foundation for empirical work that applies these ideas. Specifically, I discuss the theoretical possibility of causal inference based on observational data (Section 4.1); the causal graph (or causal diagram) that allows the explication of qualitative causal assumptions, the determination of control variables, and the assessment of the sufficiency of measured variables for consistent estimation of specified causal effects (Section 4.2); and ways of learning about the underlying causal graph from observed data (Section 4.3). I fix ideas using the example of the causal effects of regime type on militarized international conflict, a topic central to the literature on democratic peace.

4.1 Causal Inference with Observational Data

Unlike an associational model, which is essentially represented by a joint probability distribution that tells us how probable various events are and how the probabilities would change conditioning on the *observation* of some of the variables in the system, a causal model also tells us how these probabilities would change as a result of *external intervention* in the system. A simple example illustrates the differences between the two: the probability of it having rained if we *see* the grass wet is clearly different from the probability of rain if we *make* the grass wet using the sprinkler—the latter is of course the *unconditional* probability of it having rained. Applying the standard probability calculus, $P(\text{rain}|\text{see}(\text{wet})) = P(\text{rain}|\text{wet}) = \frac{P(\text{rain} \cap \text{wet})}{P(\text{wet})} = P(\text{rain}) \frac{P(\text{wet}|\text{rain})}{P(\text{wet})}$, a quantity different from $P(\text{rain}|\text{do}(\text{wet})) = P(\text{rain})$. (Pearl 2000).

In causal inference what we are really interested in is *not* quantities in the form of $P(y|x)$, but $P(y|\text{do}(x))$, since the latter would allow us to infer the consequences of, for example, a policy intervention (such as changing the regime type of a state), instead of passively observing how things happen of their own accord. What observational data naturally supply us, however, are of course in the form of the former, which explains why it seems so hard or even impossible to make true causal inferences using observational data. One important contribution of the new causal theory is to show that, under certain easily identifiable conditions, true causal inference using observational data is in fact possible, that is, $P(y|\text{do}(x))$ type of quantities *can* be expressed in terms of observable quantities of the $P(y|x)$ type.

To see why, consider the decomposition of the (observational) joint probability distribution according to the chain rule of probability calculus. Assume the system under study has n variables (or sets of variables), denoted x_i , $i = 1, 2, \dots, n$. We can write

$$\begin{aligned} P(x_1, x_2, \dots, x_n) &= P(x_1)P(x_2|x_1)P(x_3|x_1, x_2) \dots P(x_n|x_1, x_2, \dots, x_{n-1}) \\ &= \prod_j P(x_j|x_1, x_2, \dots, x_{j-1}) \end{aligned} \quad (5)$$

Since any given x_j may be independent from some of the variables in the set x_1, x_2, \dots, x_{j-1} , denote pa_j the minimal subset of x_1, x_2, \dots, x_{j-1} such that $P(x_j|x_1, x_2, \dots, x_{j-1}) = P(x_j|pa_j)$,

we can rewrite (5) as:

$$P(x_1, x_2, \dots, x_n) = \prod_j P(x_j | pa_j) \quad (6)$$

where pa_j are called the ‘‘Markovian parents’’ of x_j . Expression (6) holds for *any* ordering of the x ’s, in particular an ordering that is *causal*, so that pa_j denotes the set of direct causes of x_j .¹⁶ Since this ordering is most natural, and most meaningful for causal inference, assume that (6) is based on such an ordering. This decomposition has a one-to-one correspondence with a *causal Bayesian network*, represented by a DAG in which x ’s are the nodes, and there is a directed link from x_i to x_j if x_i is a possible direct cause of x_j , i.e., $x_i \in pa_j$. For example, assuming the DAG in figure 1 represents a causal Bayesian network, then v_1 is the parent of the other variables, and the joint probability of the system can be expressed as $P(v_1, v_2, v_3, v_4) = P(v_1)P(v_2|v_1)P(v_3|v_1)P(v_4|v_1)$. We call the DAG representing a causal Bayesian network a *causal graph* (or a causal diagram).

How does this decomposition help us to see that causal inference is possible from observational data? In other words, how can $P(y|do(x))$ type of quantities be expressed in terms of the usual joint or conditional probabilities that are observable? The answer is simple if we realize that an intervention, say $do(x_i = x_i^*)$, means that the causal mechanism leading from pa_i to x_i is removed through external control, so that in the causal graph the links from pa_i to x_i are removed, and $P(x_i = x_i^* | pa_i) = 1$ and $P(x_i = x_i^0 | pa_i) = 0 \forall x_i^0 \neq x_i^*$. Thus, from (6) we have:

$$\begin{aligned} P(x_1, x_2, \dots, x_n | do(x_i = x_i^*)) &= \prod_{j \neq i} P(x_j | pa_j) \\ &= P(x_1, x_2, \dots, x_n) / P(x_i^* | pa_i); \text{ for } x_i = x_i^* \end{aligned} \quad (7)$$

and for all other values of x_i the joint probability is 0. Applying conditional probability operations, (7) is equivalent to:

$$P(x_1, x_2, \dots, x_n | do(x_i = x_i^*)) = P(x_1, x_2, \dots, x_n | x_i^*, pa_i) P(pa_i); \text{ for } x_i = x_i^* \quad (8)$$

which in turn leads to the expression for causal effects of x_i on a subset of variables in the system that are not in pa_i , call it y :

$$P(y | do(x_i = x_i^*)) = \sum_{pa_i} P(y, pa_i | x_i^*, pa_i) P(pa_i) = \sum_{pa_i} P(y | x_i^*, pa_i) P(pa_i) \quad (9)$$

¹⁶This reflects the notion of Markovian causality in the sense that conditioning on the direct causes renders a variable independent from all its non-descendants.

by summing (8) over all other variables in the system that are not y . Expressions (7)-(9) are remarkable in that they establish the link between quantities that are causal (left hand side), and that are observable, involving no “do” operations (right hand side). From (9), we also see *when* the causal effects are identifiable from observational data, and how: if we observe all the direct causes of x_i , pa_i , then the causal effects of x_i on y can be obtained just by adjusting for pa_i .¹⁷

Of course, (9) requires that we can identify and measure all the direct causes of x_i , an assumption that may not be met in practice. Fortunately, there are conditions under which the causal effects can be obtained even when some variables in pa_i are not measured, when adjusting for a different set of covariates (or control variables) would be sufficient. Moreover, simple graphical criteria would allow us to identify such control variables (or the need to measure them, if they are latent). I discuss this below.

4.2 Causal Graphs and Covariate Selection

The graphical criteria can be applied when the causal graph is known, which I assume here. I discuss inference on the causal graph itself later. An example of a more complex causal graph is shown in figure 3, where the nodes denote variables or sets of variables, and the directed edges denote the possible flow of causation, with a missing link between any two nodes encoding the assumption that the two are causally unrelated. A causal graph is assumed to be complete in the sense that all common causes of specified variables are included in the graph, and that all causal relations among the specified variables are included (e.g, the absence of an edge from z_1 to z_3 is an accurate representation only if z_1 does not cause z_3 .) A complete causal graph does not, however, need to include *all* causes of all specified variables, or to note all intermediate variables on a causal pathway. And a causal graph can have latent or unmeasured nodes, which are usually denoted by empty circles. As a concrete example, consider the causal structure that this graph may represent for a set of key variables most frequently used in empirical studies of international conflict and in the test of the democratic peace theory: militarized international conflict (y), regime type (x), economic conditions (z_1), trade (z_2), geographic proximity (z_3), and a set of other variables that may have a

¹⁷We also see that, if conditioning on x_i , y is independent from pa_i , then adjustment for pa_i would be unnecessary, and $P(y|do(x_i = x_i^*)) = P(y|x_i^*)$.

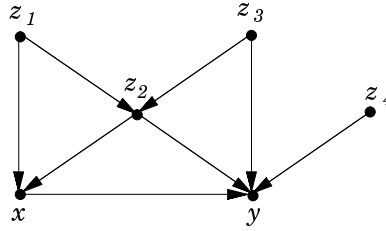


Figure 3: A DAG representing causal structure

causal relationship with conflict, such as balance of power and international treaty and organization co-membership (z_4) (e.g., Oneal and Russett 1999, Beck, King, and Zeng 2000, Gleditsch and Ward 2000, Russett 2003b). Whether this structure is compatible with observed data is an issue I will turn to later, when I discuss the search for causal structures using empirical data. Here I assume the graph is an accurate representation of the causal structure, and use it to illustrate the application of one of the most useful graphical criteria in finding control variables. The substantive interest here is to assess the causal effects of regime type (x) on the probability of international conflict (y). Without applying such criteria, standard practice simply controls for all the z variables in the graph.

Before stating the graphical tool, we need to introduce the notion of “d-separation”, or “blocking”, in a directed graph. If x , y , and z are three disjoint sets of nodes in a digraph, then z is said to d-separate x from y if and only if every path from any node in x to any node in y is blocked by z . A path is blocked by z if and only if: 1. the path contains a chain $i \rightarrow m \rightarrow j$ or a fork $i \leftarrow m \rightarrow j$ such that m is in z , or 2. the path contains an inverted fork $i \rightarrow m \leftarrow j$ such that neither the middle node m nor any of its descendants is in z . It is intuitive to see that if z d-separates x from y , then x is independent of y conditioning on z . This is so since in a causal graph, conditioning on the middle node in a chain or a fork blocks the causal information flow, while conditioning on a common “child” (or its descendants) renders the “parents” dependent.

The theorem of *back-door adjustment* (Pearl 2000, p.79) tells us that the (total) causal effect of x on y is identifiable if there exists a set of nodes z that satisfies the so called “back-door criterion,” so that that no node in z is a descendant of x and that z blocks every path between a node in x and a node in y that contains an arrow into x (i.e., enters x through the “back-door”). The causal effect is

obtained by adjusting for z (such as using z as “control variables” in a regression model, or through direct adjustment in non-parametric matching).

The intuition behind the theorem is that when z blocks all back-door paths from x to y , the “front-door” paths leading from x to y become the only channels through which the effect of x to y manifests, thereby eliminating confounding of the relationship by variables lying on a back-door path (such as spurious correlation due to a common cause). And conditioning on such z , x would be effectively like “root nodes” so that there is no difference between $do(x)$ and $see(x)$, thus permitting the leap from observational data (information from “seeing”) to inference on causal effect (what results from “doing.”)

Whether a given set of nodes z satisfies the back-door criterion can be easily read off the causal graph, so the criterion provides a working test of the sufficiency of z as control variables. Applying the back door criterion to figure 3, it is easy to verify that any set of nodes (excluding x and y) that contains either $\{z_1, z_2\}$ or $\{z_2, z_3\}$ satisfies the back-door criterion and could be used as control variables. There are 4 back-door paths from x to y : xz_1z_2y , xz_2y , xz_2z_3y , and $xz_1z_2z_3y$. The direct common cause of x and y , z_2 , could block the first three, and the fourth is blocked only if either z_1 or z_3 is controlled.

From this example, we can learn a series of lessons, some of which directly address the confusion in standard practice discussed earlier:

(1) More than one set of nodes may satisfy the back door criterion, hence different researchers using different sets of control variables in assessing the same causal effect *could* all be right at the same time—provided each of the control variable set is “right” in the sense of allowing consistent estimation of the causal effect (by satisfying the graphical criterion, for example).

(2) However, not all workable sets of control variables are optimal, and we should choose what is optimal, such as one that minimizes data collection costs and/or maximizes quality of data. In the example, any set larger than $\{z_1, z_2\}$ or $\{z_2, z_3\}$ would be suboptimal in that we would be using too many control variables, incurring unnecessary costs of data collection (if some of the variables were not already measured), risking more data quality problems, and sacrificing efficiency by using too many parameters in estimation, etc. Between the two minimal sets, choice can be made with similar

considerations. For example, if data on economic conditions were more expensive to obtain and/or risked more errors than data on geographic proximity, then $\{z_2, z_3\}$ would be better than $\{z_1, z_2\}$.

(3) Contrary to common belief, just including common causes of x and y in the control variable set may not be adequate.

(4) The marginal effect computed from a model can be interpreted as a causal effect if and only if a sufficient set of control variables is used.

(5) In general, the same model can *not* be used to infer more than one causal effect. In our example, if we want to infer the effect of geographical proximity (z_3) on international conflict (y), we should *not* control for any other variables, since there are no back-door paths between z_3 and y at all. A model that is good for studying the causal effect of x on y , say one that has $\{x, z_2, z_3\}$ as the set of independent variables, is not good for the purpose of causal inference on z_3 . Indeed, it would be a very bad one, since effectively the “control variables” for z_3 are $\{x, z_2\}$, which are descendants of z_3 . Both the back-door criterion and common sense (and common practice, too, where the issue is obvious) tell us that we should *not* control for consequences of a variable in assessing its total effects.¹⁸ In this example, controlling for the (positive) consequences of geographic proximity would bias upwardly the magnitude of the estimated negative effect of the variable on international conflict.

Attempting to infer more than one causal effect with the same model is, however, widespread practice. In the influential work of Russett et al. (1998) for example, the causal effects of three variables of the “Kantian tripod for peace”—democracy, trade, and joint membership in international organizations—are assessed using the same model. The insight from causal graph theory would suggest reevaluation, especially given their finding that democracies are more likely to join IGOs, which means that in the causal graph IGO could be a descendant of democracy and hence should *not* be controlled in estimating the causal effects of democracy on conflict.

(6) In general, a good causal model is not a good forecasting model, and vice versa. In our example, if we want to build a model that is good for forecasting international conflict, y , we would want

¹⁸We may need to control such variables in assessing *direct* effects, as we shall see shortly.

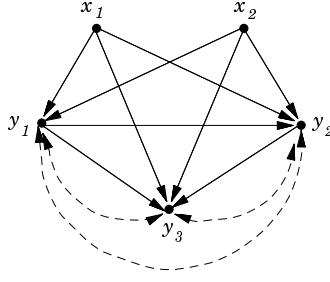


Figure 4: A *Seemingly Recursive System*

to include all the direct causes of y , that is, the set $\{x, z_2, z_3, z_4\}$. This would be a sub-optimal model for inferring the effect of regime type (x), for which $\{x, z_2, z_3\}$ is sufficient as independent variables, and even worse a model for causal inference on geographic proximity (z_3), for which no other variables should be used. Conversely, an optimal causal model for either x or z_3 would be a bad forecasting model for y , as it would omit some of the direct causes of y .

Similar graphical tools as those given in the back-door adjustment theorem shed light on identification problems in structural equation models, too, where the causal relationships in the graph assume linear functional forms, and the causal graph is effectively the same as the path diagram. Figure 4 is an example graph corresponding to the following system of structural equations:

$$\begin{aligned}
 y_1 &= \alpha_1 + && \gamma_{11}x_1 + \gamma_{12}x_2 + \epsilon_1 \\
 y_2 &= \alpha_2 + \beta_{21}y_1 + && \gamma_{21}x_1 + \gamma_{22}x_2 + \epsilon_2 \\
 y_3 &= \alpha_3 + \beta_{31}y_1 + \beta_{32}y_2 + && \gamma_{31}x_1 + \gamma_{32}x_2 + \epsilon_3 \\
 \\
 y_1 &= \alpha_1 + && \gamma_{11}x_1 + \gamma_{12}x_2 + \epsilon_1 \\
 y_2 &= \alpha_2 + \beta_{21}y_1 + && \gamma_{21}x_1 + \gamma_{22}x_2 + \epsilon_2 \\
 y_3 &= \alpha_3 + \beta_{31}y_1 + \beta_{32}y_2 + && \gamma_{31}x_1 + \gamma_{32}x_2 + \epsilon_3
 \end{aligned}$$

where the presence of the bidirectional, broken-line arcs reflects the assumption that the error terms in the equations are correlated. If these bidirectional arcs were absent, this would be a fully recursive system and ordinary least square estimation of the coefficients (which can be interpreted as direct causal effects) would be consistent. But are the direct effects still identifiable (i.e., can the parameters still be consistently estimated), if the errors are correlated, and the system is only “seemingly recursive”? This question has generated much confusion, as witnessed by the recent

debate among leading econometricians and political methodologists.¹⁹ Two popular econometrics textbooks widely adopted in political methodology courses as well give opposing answers. Gujarati (2003) suggests that the system can be estimated through methods for seemingly unrelated regression models (SUR)(footnote 7, p.766) while Greene (2000) disagrees, arguing that identification in this system is not guaranteed without additional assumptions (p.673).

Which is right? In the absence of easy to apply and reliable tests, it is a matter of lengthy debate.²⁰ Fortunately, new graphical methods provide some simple tools to answer the question. Theorem 5.3.1 of Pearl (2000) tells us that the direct effect of a variable x on another y is identified if (and, for all practical purposes, only if) in the causal graph with the edge between x and y removed, there exists a set of variables z which are non-descendants of y such that z d-separates x from y . Such set z is said to satisfy the “single-door criterion” (p.150).

Applying this theorem, it’s easy to see that all parameters in the fully recursive system (that has the bidirectional arcs removed) are identified. For example, in the recursive model, the direct effect of x_1 on y_2 is identified since with the edge $x_1 \rightarrow y_2$ removed, all other paths between the two are d-separated by $\{y_1, x_2\}$: any path that goes through y_3 , such as $x_1 y_3 y_2$, is naturally blocked, $x_1 y_1 y_2$ is blocked conditioning on y_1 , and $x_1 y_1 x_2 y_2$ is blocked conditioning on x_2 .

What about the seemingly recursive system with correlated errors? The direct effect of x_1 on y_2 for example is no longer identified. One of the two paths, $x_1 \rightarrow y_1 \rightarrow y_2$ and $x_1 \rightarrow y_1 \longleftrightarrow y_2$, is always active. Conditioning on y_1 would block the first but activate the second, while not conditioning on y_1 would leave the first connected. In general, then, the parameters in a seemingly recursive system are not identified without additional assumptions.

¹⁹See the discussion on the H-Net PolMeth Listserv in February 2002, under the thread “recursive models and identification” and “dialog with Greene on systems of equations”, <http://web.polmeth.ufl.edu/hnet.html>.

²⁰Though we should point out that Gujarati’s answer is obviously flawed since the system is simply *not* a SUR system (which does not have endogenous variables on the right hand side). SUR systems are of course identified, even estimated just by OLS, since OLS estimators are consistent. The GLS estimator usually employed for SUR is to improve efficiency, which is not an identification issue.

4.3 Causal Structure Inference

Causal graphs not only aid in causal effect identification and control variable selection, they also in themselves facilitate the explication of qualitative assumptions about the underlying causal structure. They provide an unambiguous and efficient language for summarizing causal assumptions, and make the implications of such assumptions clear. Some causal assumptions based on substantive theory are explicit or implicit in any empirical study aiming at causal inference, but they are rarely explicated, and without the aid of the causal graph one hardly knows whether they are adequate. If they are inadequate, then subsequent modeling decisions that implicitly rely on the causal graph such as control variable selection would obviously be ungrounded and arbitrary. Even if they are sufficient, without translating them into a causal graph their modeling implications would be unclear. As we have seen in the example above, common practices of control variable selection such as including all common causes of x and y and/or including all causes of y can lead us astray.

So the causal graph plays a critical role in improving causal inference. Although the language of causal graph is new to the discipline, political scientists intuitively recognize the importance of understanding the structure and mechanisms of the system under study. In offering a series of suggestions for improving causal inference with observational data, Brady (2002) makes having better theory that “provides mechanisms and explanations” the top priority. As he convincingly argues, “researchers should *not* be happy with regression ‘models’ that simply throw variables into a regression...researchers must seek to understand the exact mechanisms...should seek to explain social phenomena in the same way that the Maxwell-Boltzmann theory of gases explains the regularities of the gas laws...” (p.29). Ideally, we should have such good substantive theory that the underlying causal graph is readily available.

Unlike the physical sciences, however, the causal structure in a social system is much less clear, and no matter how hard we try, there will be situations where sufficient substantive theory is unavailable to provide us with an indisputable causal graph to work on.²¹ In an overview of the literature on violence studies (Russett 2003b), for example, “the direction of causality” tops the list

²¹Indeed most, if not all, endeavors of causal inference in the social sciences aim to “test theory”, i.e., to see whether certain causal link exists at all, and the estimation of the magnitude of the effect, if any, is of secondary importance.

of “analytical problems and research directions.” Does alliance formation reduce the risk of conflict, or do countries ally with each other *because* they are at peace? Democracy may decrease the risk of conflict, but would peace-time make preserving democratic government easier? (p.21).

Fortunately, causal graph theory tells us that we do not need to rely completely on substantive theory in learning the causal structure. Under some general assumptions we can learn a great deal about the causal graph from observed data (Spirtes et al, 2000; Pearl and Verma 1991). Software tools that “discover” causal structure from data have been developed and are in widespread use across various fields (Scheines et al., 1994.) These tools can be fruitfully employed in the analysis of political data where causal structure is far from clear from substantive theory alone, such as data on international/civil conflict. Below I briefly discuss the intuition for why such “causal structure discovery” is possible, and explain the common assumptions used in the search.

From the discussion in section 4.1, we have seen why and when causal effect estimation from non-experimental data is possible. The intuition behind causal structure discovery from observational data is essentially similar, and rests in the fact that, under the causal Markov condition, observed patterns of statistical independence may imply constraints on patterns of causation, effectively limiting the number of possible causal graphs compatible with observed data. A simple example illustrates the idea. Suppose a system under study contains just three variables, x_i , $i = 1, 2, 3$, and suppose we observe from the data that x_1 and x_3 are independent conditioning on x_2 . This observation, strictly statistical and not causal in itself, nevertheless implies that the causal graph $x_1 \rightarrow x_2 \leftarrow x_3$ is incompatible with the data, since if x_1 and x_3 were both causes of x_2 , then conditioning on x_2 would render the parents statistically dependent.²²

Observed independence patterns rarely imply a *unique* causal graph compatible with data. In the simple example here, either $x_1 \rightarrow x_2 \rightarrow x_3$, or $x_1 \leftarrow x_2 \leftarrow x_3$, or $x_1 \leftarrow x_2 \rightarrow x_3$ would be compatible with the observation. Software tools such as Tetrad take observed data (and independence conditions they embody) as input, and use effective algorithms to search for all compatible graphs. The number of such graphs can be greatly reduced, possibly to only one, with added as-

²²For example, if the value of x_2 is determined by the other two through $x_2 = x_1 + x_3$, then conditioning on the value of x_2 would mean that x_1 and x_3 cannot take arbitrary, unrelated values. In particular, if x_2 is 0, then it must be that $x_1 = -x_3$.

sumptions, based on substantive knowledge of the problem or knowledge of the temporal order of variables, (such as regime type does not cause geographic proximity, or the GDP variable measured last year cannot be a cause of the population density variable measured five years ago) or theory, (such as conflict involvement has no direct causal effect on regime type) or general assumptions of axiomatic nature. The last type includes:²³

1. Faithfulness (stability): This condition assumes that independence patterns in the probability distribution of data arise not from coincidence but rather from structure. In figure 3, for example, suppose geographic proximity (z_3) increases the risk of conflict (y), but decreases it through facilitating trade (z_2) which is known to inhibit conflict. Suppose the positive effect and negative effect are such that they exactly balance and cancel, then z_3 and y might be independent in the observed data. In this case, the distribution is said to be *unfaithful* to the true causal structure.

2. Causal sufficiency: All common causes of measured variables are measured.

3. Model minimality: a parsimony condition similar in spirit to that considered in standard model selection procedures, which guarantees that any alternative structure compatible with the data is necessarily less specific, less falsifiable, and less trustworthy than the inferred structure(s).

It is not necessary to adopt these conditions for the Tetrad tool to operate, although additional assumptions can dramatically increase inferential power. Compared with pet theories that may often be whimsical, the axiomatic type assumptions are much more general, and much easier to communicate across different types of problems or indeed different disciplines. Substantive assumptions based on theory can still be used as input, and the implications of them are clearly seen from the resulting causal graphs inferred, which in its own provides the researcher with a powerful means to understand the ramification of any theoretical assumption, and may change the way studies are designed and data collected.

Causal graph theory discussed above is still under development. For example, existing methods largely focus on acyclic graphs with no feedback cycles or undirected edges. In the social sciences, however, cyclic graphs and/or undirected edges are often encountered due to unrefined measurement that permits feedback, which blurs the direction of causation. Extending the results for DAG's to

²³See Pearl and Verma 1991, Scheines 1997, Pearl 1997b, Freedman n.d., Freedman and Humphreys 1999, and Spirtes et al. 1997 and 2000 for some detailed discussion and debate on the reasonableness and consequences of these conditions.

such graphs are therefore particularly useful for social science applications.

5 Rare Events and Functional Complexity

This section discusses some extensions of existing methods to accommodate special features of political data, such as functional complexity and rareness of certain events. Rareness of events is a key feature of international and civil conflict data, and functional complexity characterizes most political and social relationships. Section 5.1 gives some results on recovering measures of network characteristics in relational data discussed in section 3.1, when the data result from more efficient “case-control” sampling design for rare events. Section 5.2 discusses improvement of the random graph models discussed in section 3.2, as well as models for the so-called “propensity score” increasingly used in the estimation of causal effects, with flexible functions such as neural networks.

5.1 Rare Events Data and Network Characteristics

Some events, such as international and civil conflict, presidential vetoes, coups, or rare diseases, rarely happen. To study such events, using the full sample data is grossly inefficient since most of the non-events contain little information. Alternative sampling plans that retain only the events and a small fraction of the non-events are much more efficient in terms of both data collection cost and subsequent data storage and analysis. Inferential implications of and necessary corrections for using such data in standard models such as logit are well developed (King and Zeng 2001b, 2001c, and 2001d), but no work has been done analyzing the issue in the context of social networks, where alternative sampling plans previously studied typically select on players (or nodes in the graph), rather than types of relationships (existence/absence of edges). Questions naturally arise, such as whether/which network characteristics may be recovered from such sub-sampled data, and how. The preliminary results below show that essentially *all* measures discussed above are obtainable through simple adjustment assuming knowledge of the size of the full network (N) and the ratio of the number of null dyads (e.g., dyads not involved in conflict) in the full data (N_0) to that in the sub-sampled data (n_0).

The estimators we give are consistent since they are continuous functions of consistent estimators for the sub-sampled data.

Degree and actor centrality: The degree of a node (player) is the number of edges adjacent to it. In the conflict network, for example, it would be the number of countries with which one is in conflict. Since sub-sampling the “0”s, i.e., dyads not in conflicts, leaves all existing edges (conflict) intact, it is obvious that the degree measures for the conflict network are not affected by the sampling process. Hence degrees from the sampled data are also degrees for the original network.²⁴

This is so only for the network of the relation that is also the basis of the sub-sampling, however. We may need to recover degree measures for networks of other relations, say trade. Obviously a null dyad for the “conflict” relation may not be null for the “trade” relation. Thus some correction is needed. Let y denote the relation based on which data are sampled, and let $w = N_0/n_0$. It is easy to see that the corrected measure is simply $d = d_1 + d_0w$, where d_i is the degree measure corresponding to $y = i$, $i = 0, 1$. The centrality index, d/N , is then just the recovered degree measure over the size of the *full* network.

Density and centralization: Similar reasoning leads to the formula for density measures: $D = \frac{D_1 + D_0w}{N(N-1)}$, where D_i is the total number of edges present when $y = i$, $i = 0, 1$. Indices of centralization, measured as variances of actor level degree or centrality measures, are available based on the corrected actor level measures.

Cohesive subgroups: Using the definition of k – *cores* for cohesive subgroups, sub-sampling does not change cohesive subgroup status in the same relation y (that is the basis for sampling), since existing edges are not affected. On a different relation, a cohesive subgroup in the sampled data is obviously also a cohesive subgroup in the full network. However, a subgroup that is not cohesive in the sampled data *may* be cohesive in the full data. One way to estimate the true status would be to fill missing data for $y = 0$ cases using sample density information: if for $y = 0$, the density of edges is τ , then turn the missing links into edges with probability τ . Once all missing links are filled, cohesive subgroup status is easily checked with the k – *core* concept.

²⁴Other characteristics of the graph based on edges present, such as number of triangles and k –stars used in random graph models, are similarly unaffected.

Structural equivalence: The measures I discussed are correlation or distance measures, and as such, proper weighting should give consistent estimates. All data corresponding to $y = 0$ should be weighted by $w = N_0/n_0$. For example, the corrected Euclidean distance measure between nodes i and j , assuming a symmetric relation, would be

$$\sqrt{w \sum_{k:y=0} (x_{ik} - x_{jk})^2 + \sum_{k:y=1} (x_{ik} - x_{jk})^2}$$

where x is the relation on which the measures are constructed. It could be the same as y or could be a different relation. Weighting in correlations is similar. Such weighting in computing correlation coefficients or distance measures can be implemented in standard statistical packages such as Stata and so poses no practical inconvenience.

5.2 Functional Complexity: Improving Random Graph Models for Relational Data and Propensity Score Models for Causal Effect Estimation

Functional complexity characterizes most, if not all, political and social relationships, and the exact functional forms are almost never known. Flexible models such as neural networks that better accommodate such complexity than most standard models are receiving increasing attention in political science. As in a wide range of other fields, neural networks have found successful applications in the modeling of political data (e.g., Zeng 1999, 2000a; Beck, King, and Zeng 2000, 2004; King and Zeng 2001a, Lagazio and Russett 2003.) In this section I discuss their use in improving random graph models for the analysis of relational data and propensity score models for the estimation of causal effects. Standard practice in estimating these models relies on simple models like logit that *assume* certain functional form, usually linear in one part or another, and is likely inadequate for complex data like international conflict.

A neural network model is capable of approximating arbitrary functional forms through the use of “hidden neurons”. Figure 5 depicts a single hidden layer feed forward neural network, where the hidden neurons z are functions of the input variables x , and the output variable y is in turn a function of these hidden neurons. Provided the functions for z satisfy some general conditions (such as being bounded, non-constant, and continuous), and that sufficient number of hidden neurons are used, the function that relates y to x through z is proven to be able to approximate the “true” function

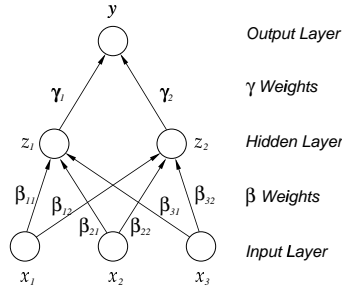


Figure 5: A one hidden layer feed forward neural network

$y = f(x)$ to arbitrary degree of accuracy, whatever form $f(\cdot)$ takes. This explains the theoretical appeal of these models. The logistic form is typically used for z . Output function for y is typically logit if y is binary, and linear if y is continuous.

Neural networks may be superior to standard models like logit in any situation where the underlying “true” functional form is unknown and is likely complex. In the case of random graph models for relational data, improvement over standard models is likely in analyzing complex relationships such as international conflict, especially given previous success of neural networks over logit on such data (e.g., Beck, King, and Zeng 2000, 2004; King and Zeng 2001a). In section 3.2 we have discussed random graph models in some detail, and have seen that, in model 1, the network characteristics $z(x)$ enter the model linearly, as do $\delta(x_{ij})$ in model 3. The linearity is part of the model *assumption*, but is rarely challenged at all in existing work, which also evaluates model performance by goodness of fit rather than out of sample generalization, a practice that can be seriously misleading. In applying random graph model to complex political data, improvement can be made in these areas. Flexible models that can approximate unknown complex functional forms well, such as neural networks, can be used in place of the linear specifications. And model evaluation should rely on the gold standard of out-of sample generalization.

5.2.1 Improving Propensity Score Estimation

We now turn to the subject of estimating causal effects using the propensity score approach. Graphical methods discussed in Section 4 allow us to determine whether certain causal effects are identifiable from observational data, and if so which variables should be controlled/adjusted in estimating

these effects. The remaining task is actually carrying out the estimation. Typically a parametric model (such as a generalized linear regression model) would be run for this purpose. A parametric model may suffer from two major problems, even given the correct set of control variables. One is incorrect functional form assumption, the other is the problem of incomparability of “experimental” and “control” groups in the data. The second problem exists when the distribution of control variables for one value of the causal variable does not completely overlap with that for another value. Observations in the non-overlap region in effect do not have matches and are therefore useless for causal inference. Not excluding them would contaminate the data and results, but checking for non-overlap directly on the control variables is difficult or even infeasible, since with typical size of the control variable set it would involve high dimensional density estimation (King and Zeng 2004).

Conceptually, non-parametric matching is superior to parametric models in that they do not assume specific functional forms and the matching process itself would guarantee that only observations with matches are used. In practice, matching literally is rarely feasible when there are more than just a couple of control variables and/or when the control variables are continuous, because of the “curse of dimensionality”. Recent development in statistics has offered a promising approach that avoids the curse of dimensionality and therefore “solves” the problem of control variable adjustment. The seminal work of Rosenbaum and Rubin (1983) proves that adjustment can be made by matching on just *one* scalar variable, the so called “propensity score”. Let x denote the key causal variable, and z denote the set of (correct) control variables. The propensity score is the probability of being exposed to the cause (receiving a treatment, participating in an experiment, etc.) conditioning on the control variables: $P(x = 1|z)$.²⁵ Rosenbaum and Rubin (1983) proves that the propensity score is a “balancing score” in that conditioning on it the distribution of z is balanced across the “treatment” and “control” groups. This means that adjusting for the propensity score alone is equivalent to adjusting the whole set of control variables. It also means that the propensity score can be used to identify non-overlap region in data without the need for high dimensional

²⁵The propensity score method is originally developed for binary x , which we assume here. It is recently generalized to “propensity function” that works for other types of causes as well, such as multinomial, ordinal, or continuous x (Imai and Dyk 2002). The following discussion applies to propensity function in general, but we use the binary case for illustration as “propensity score” is a more familiar term.

density estimation (e.g., King and Zeng 2004). This is an exciting development, and naturally the propensity score approach is receiving increasing attention in statistics and all of the social sciences. As Brady (2002) argues, “every empirical researcher should become familiar with this framework.” (p.30).

One important issue remains, however. The “true” propensity score is rarely, if ever, known in observational data, and hence itself must be estimated from data before it can be used to estimate causal effects by matching or related methods and/or to identify density non-overlap. Building models that give valid propensity score estimates are critical to subsequent analysis, for the obvious reason that good properties of propensity scores do not apply to non-propensity scores. Although studies have shown that the consequences of misspecification in the propensity score model tend to be milder than that in the outcome model, misspecification of the propensity score model can severely bias causal effect estimation nevertheless (e.g., Imai and Dyk 2002, Table 1, p.14).

Despite this seemingly obvious fact, with few exceptions (Heckman et al. 1998b) applied work using the propensity score approach have largely relied on simple logit models to estimate the propensity score, often without conducting any tests for the validity of the estimated propensity scores. The possibility of invalid propensity scores from such studies is a potential reason why some studies find that estimated causal effects from adjusting for the “propensity score” are biased compared with the bench mark of experimental results (e.g., Agodin and Dynarski 2001; Bloom et al 2002; Ty Wilde and Hollister 2002). Some studies try to alleviate the problem by including higher order terms in the logit model, but polynomial approximation to unknown nonlinear functions is inefficient and often suffers from numerical instability problems.

For reasons discussed earlier, neural networks provide a natural candidate for modeling propensity scores. Below we give some preliminary results comparing performance of a neural network and some logit models in studying the causal effect of regime type on the probability of state failure (King and Zeng 2001a, 2004). There are five control variables: military population, population density, legislative effectiveness, infant mortality, and trade openness. The “treatment” variable is binary indicating whether the state is a partial democracy or autocracy. A single hidden layer feed forward neural network model with 15 hidden neurons for the propensity score is compared with

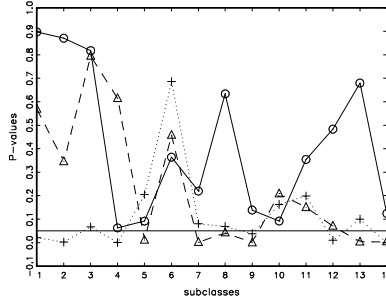


Figure 6: *Propensity score validity test*. “o”: neural network; “+”: logit; “Δ”: logit with up to third order terms

a simple logit (with just the original control variables) and a logit that includes many higher order terms (all terms up to the third order). If the estimated propensity score is valid, then it should possess the balancing property so that, for example, the means, variances, and covariances of the control variables should be the same for the democracies and the autocracies conditioning on the propensity score. Figure 6 shows the results of testing the null hypothesis that the distributions of the control variables are not statistically different across democracies and autocracies, within subclasses of similar propensity score values. The table reports the p -values from the Hotelling’s T^2 test suitable for the purpose. We test the means, variances and covariances by including all first order and second order terms of the control variables in the group for comparison. As in usual hypothesis tests, high p -values lead to failure to reject the null hypothesis, while very low p -values leads to findings of “significance”. Unlike a typical test, “significant results” in this context are “bad” since it would mean that the estimated p -scores are invalid.

The results from the figure are clear. The p -values for the neural network propensity score model (circles) are in general much higher than those from either of the logit models (crosses and triangles), and all are above the standard 0.05 significance level. The many crosses and triangles below 0.05 means that the logit models, even the one with so many higher order terms, have failed to produce valid propensity scores.

6 Conclusion

Prediction and causal inference are the central pursuits of empirical political science. Graphical methods and models provide excellent tools for improving structural and relational modeling that is vital for reliable prediction and causal inference. The importance of structural and relational modeling is particularly evident in the analysis of international relations data. International events take place not in isolation, but within a network of intricate interdependence. The structural characteristics of the international network, in addition to individual state or dyad level attributes, are therefore likely to hold important explanatory and predictive power. How to identify and measure the network characteristics and model the dependence structure systematically has largely eluded previous research. Graphical methods and models greatly facilitate this task by providing a rich array of well defined graph theoretic measures capturing the structure of the entire network, and by allowing the systematic modeling of dependence structure of the network through the use of the dependence graph. In causal inference, graphical methods provide useful tools that help the identification of causal structure from observed associational data, and that improve the assessment of causal effects by ensuring correct specification of control variables. Theoretical unbiasedness conditions such as strongly ignorable treatment assignment are otherwise hard or impossible to check but can only be assumed.

This paper brings together the large and diverse literature on different types of graphical models that are of particular importance to our discipline, integrating the techniques within the unifying framework of graph theory and offers a technically accessible treatment of the methods. The paper discusses the adaptation, application and extension of these graphical methods and models, laying the theoretical foundation for broad scale empirical research employing these novel tools. Initial results from analyzing the 1947-1989 MID data provide strong evidence that properties of the system as a whole and that of individual states/dyads embedded in the system hold important explanatory and predictive power, and clearly reveal the interdependence among dyads sharing a common member. Although the discussion largely focuses on the case of analysis of international relations data, many of the ideas offered readily apply to other subfields of political science and cognate disciplines as well.

References

- Agodin, R., & Dynarski, M. 2001. "Are experiments the only option? A look at dropout prevention programs". Washington, DC: Mathematica Policy Research, Inc.
- Anderson, C., Wasserman, S., and Crouch, B. 1999. "A p* primer: Logit models for social networks." *Social Networks*. 21,37-66.
- Angrist, J. D., Imbens, G. W., and Rubin, D.B. 1996. "Identification of causal effects using instrumental variables," *Journal of the American Statistical Association*, 91, 444-72.
- Beck, Nathaniel, Gary King, and Langche Zeng. 2000. "Improving Quantitative Studies of International Conflict: A Conjecture," *American Political Science Review*, Vol. 94, No. 1: 21-36.
- Beck, Nathaniel, Gary King, and Langche Zeng. 2004. "Theory and Evidence in International Conflict: A Response to de Marchi, Gelpi, and Grynavisk," *American Political Science Review*, forthcoming.
- Beck, N. and R. Tucker. 1996. "Conflict in Space and Time: Time-Series Cross-Sectional Analysis with a Binary Dependent Variable". Paper presented at the APSA Annual Meeting, San Francisco.
- Beck, N., J. Katz and R. Tucker. 1998. "Taking Time Seriously: Time-Series Cross-Section Analysis with a Binary Dependent Variable." *American Journal of Political Science*. 42. 1260-1288.
- Beck, N. and Katz, J. N. 2001. "Throwing out the Baby with the Bath Water: a Comment on Green, Kim and Yoon". *International Organization*. Vol. 55, No.2. 487-95.
- Bennett, D.S. 1996. "Security, Bargaining, and the End of Interstate Rivalry." *International Studies Quarterly*. 40, 157-184.
- Berkowitz, S.D. 1982. *An introduction to structural analysis: The network approach to social research*. Toronto: Butterworths
- Besage, J.E. 1974. "Spatial Interaction and the Statistical Analysis of Lattice Systems". *Journal of the Royal Statistics Society*. Series B. 36, 192-236.
- Bloom, Howard S., Charles Michalopoulos, Carolyn J. Hill, Ying Lei. 2002. "Can Non-experimental Comparison Group Methods Match the Findings from a Random Assignment Evaluation of Mandatory Welfare-to-Work Programs?" MDRC working paper.

- Bonacich, P. 1987. "Power and centrality: A family of measures". *American Journal of Sociology*. 92, 1170-1182.
- Brady, Henry. 2002. "Studying the Causes of Human Variability: The Role of Conditional Independence." in "Models of Causal Inference: Going Beyond the Neyman-Rubin-Holland Theory," Paper presented at the Political Methodology Society Summer Meeting, University of Washington. Seattle, WA. July 2002.
- Bueno de Mesquita, Bruce, 1975. "Measuring Systemic Polarity". *Journal of Conflict Resolution*, 19(2):187-216.
- Cox, D.R. and Wermuth, N. 1996: *Multivariate Dependences. Models, Analysis and Interpretation*. London: Chapman and Hall.
- Doreian, P. and Stokman, F.N. 1997. *Evolution of Social Networks*. Amsterdam: Gordon and Breach.
- Faust, K., and Skvoretz, J. 1999. "Logit models for affiliation networks." In *Sociological Methodology 1999*. Edited by Michael Sobel and Mark Becker. New York: Blackwell.
- Frank, O. and D. Strauss. 1986. *Markov Graphs*. *Journal of American Statistical Association*. 81: 832-842.
- Freedman, D.A., n.d., "On Specifying Graphical Models for Causation, and the Identification Problem." Technical Report No.601, UC Berkeley.
- Freedman, D.A. and P. Humphreys. 1999. "Are There Algorithms that Discover Causal Structure?" *Syntheses*, vol. 121. pp. 29-54
- Freeman, Linton C. 1984. "Turning a profit from mathematics: The case of social networks," *Journal of Mathematical Sociology*, 10: 343-360
- Gleditsch, Kristian S. and Michael D. Ward. 2001. "Measuring Space: A Minimum Distance Database and Applications to International Studies." *Journal of Peace Research*, 38(6):749-768.
- Gleditsch, Kristian S. and Michael D. Ward. 2000. "War and Peace in Space and Time: the Role of Democratization." *International Studies Quarterly*, 44(1):1-29.
- Green, Donald P., Soo Yeon Kim, and David Yoon. 2001. "Dirty Pool." *International Organization*.

- 55:441-68.
- Greene, William H. 2000. *Econometric Analysis*, 4th ed., Upper Saddle River, NJ: Prentice-Hall.
- Gujarati, Damodar N. 2003. *Basic Econometrics*, 4th ed., New York: McGraw-Hill.
- Handcock, Mark. 2000. "Progress in Statistical Modeling of Drug User and Sexual Networks". Manuscript, Center for Statistics and the Social Sciences, University of Washington.
- Heckerman, D., Meek, C., and Cooper, G. 1999. "A Bayesian approach to causal discovery". In *Computation, Causation, and Discovery*, (ed. C. Glymour and G. F. Cooper), pp. 141-65. MIT Press, Cambridge, MA.
- Heckman, James, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. 1998a. "Characterizing Selection Bias Using Experimental Data," *Econometrica*, Volume 66, Number 5, pp. 1017-1098.
- Heckman, James J., Hidehiko Ichimura and Petra Todd. 1998b. "Matching As An Econometric Evaluation Estimator," *Review of Economic Studies*, Volume 65, pp. 261- 294
- Hoff, Peter, Adrian Raftery, and Mark Handcock. 2002. "Latent Space Approaches to Social Network Analysis". Paper presented at the Political Methodology Society Summer Meeting, University of Washington. Seattle, WA. July 2002. *Journal of the American Statistical Association*, 97 (460), 1090-1098
- Holland, P. 1986. "Statistics and causal inference." *Journal of the American Statistical Association*, 81, 945-60.
- Imai, Kosuke and David A. van Dyk, 2002. "Causal inference with general treatment regimes: Generalizing the propensity score". Working paper, Harvard University.
<http://web.polmeth.ufl.edu/papers/02/imai02c.pdf>
- James, Patrick. 2002. *International Relations and Scientific Progress: Structural Realism Reconsidered*. Ohio State University Press.
- King, Gary. 2001. "Proper Nouns and Methodological Propriety: Pooling Dyads in International Relations Data". *International Organization*. 55(2):497-507.
- King, Gary and Langche Zeng. 2001a. "Improving Forecasts of State Failure," *World Politics*, Vol. 53, No. 4: 623-58.
- King, Gary and Langche Zeng. 2001b. "Estimating Risk and Rate Levels, Ratios, and Differences

- in Case-Control Studies,” *Statistics in Medicine*, Vol. 21: 1409-1427.
- King, Gary and Langche Zeng. 2001c. “Logistic Regression in Rare Events Data,” *Political Analysis*, Vol. 9, No. 2: Pp. 137-163.
- King, Gary and Langche Zeng. 2001d. “Explaining Rare Events in International Relations,” *International Organization*, Vol. 55, No. 3: Pp. 693-715.
- King, Gary and Langche Zeng, 2004. “When Can History be Our Guide? The Pitfalls of Counterfactual Inference,” Manuscript.
- Lagazio, Monica and Bruce Russett. 2003. “A Neural Network Analysis of Militarized Disputes, 1885-1992: Temporal Stability and Causal Complexity”, in Paul Diehl, ed., *Toward a Scientific Understanding of War: Studies in Honor of J. David Singer*. Ann Arbor: University of Michigan Press.
- Lauritzen, S. L. 1996. *Graphical Models*. Clarendon Press, Oxford, United Kingdom.
- Oneal, John and Bruce Russett. 1999. “Assessing the Liberal Peace with Alternative Specifications: Trade Still Reduces Conflict.” *Journal of Peace Research*. 36 (July).
- Oneal, John and Bruce Russett. 1999. “The Kantian Peace: The Pacific Benefits of Democracy, Interdependence, and International Organizations,” *World Politics*. v. 52 (1): 1-37.
- Oneal, J. R. and Russett, B. 2001. “Clear and Clean: the Fixed Effects of the Liberal Peace”. *International Organization*. Volume 55, Number 2. 469-85.
- Pattison, P., and Wasserman, S. 1999. “Logit models and logistic regressions for social networks: II. Multivariate relations.” *British Journal of Mathematical and Statistical Psychology*. 52, 169-193.
- Pearl, J., 1993b. “Aspects of graphical models connected with causality,” In *Proceedings of the 49th Session of the International Statistical Institute*, Tome IV, Book 1, Florence, Italy, 391-401.
- Pearl, J. 1995. “Causal diagrams for empirical research.” *Biometrika*, 82, 669-710.
- Pearl, J. 1997a. “The new challenge: From a century of statistics to the age of causation.” *Computing Science and Statistics*, 29(2):415–423.
- Pearl, J., 1997b. “TETRAD and SEM,” (Commentary on The TETRAD Paper). *Multivariate Behavioral Research*, Volume 33(1).

- Pearl, J. 1998. "Graphs, causality, and structural equation models." *Sociological Methods and Research*, 27, 226-84.
- Pearl, J. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, UK.
- Pearl, J. and T.Verma. 1991. "A theory of inferred causation." In J.A. Allen, R. Fikes, and E. Sandewall, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, pages 441–452. Morgan and Kaufmann, San Mateo, 1991.
- Priebe, Carey E. 2002. "Vector Quantization and Classification through the Dominating Set of a Digraph." Technical Report No. 613. Department of Mathematical Sciences, Johns Hopkins University.
- Robins, G., Pattison, P., Wasserman, S. 1999. "Logit models and logistic regressions for social networks, III. Valued relations." *Psychometrika*. 64,371-394.
- Robins, G. and Patterson, P. 2000. "*P** Models for Temporal Processes in Social Networks," Technical Report, Department of Psychology, University of Melbourne.
- Rosenbaum, Paul. 1984. "The Consequences of Adjusting for a Concomitant Variable That Has been Affected by the Treatment." *Journal of the Royal Statistical Society, A*, 147:656-666.
- Rosenbaum, P. and Rubin, D. 1983. "The central role of propensity score in observational studies for causal effects." *Biometrika*, 70:41–55, 1983.
- Rosenbaum, P.R. and Rubin, D.B. 1984. "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score." *Journal of the American Statistical Association*, 79, 516-524.
- Russett, Bruce. 2003a. "International Relations", in Kimberly Kempf-Leonard, ed., *Encyclopedia of Social Measurement*. San Diego, CA: Academic Press.
- Russett, Bruce. 2003b. "Violence Prediction", in Christopher Murray, ed., *Encyclopedia of Public Health*. San Diego, CA: Academic Press.
- Russett, Bruce. 1993. *Grasping the Democratic Peace*. Princeton, NJ: Princeton University Press.
- Russett, Bruce and Zeev Maoz, 1993. "Normative and Structural Causes of the Democratic Peace, 1946-1986," *American Political Science Review*, 87, 3:624-638.

- Russett, Bruce, John R. Oneal, David R. Davis. 1998. "The third leg of the Kantian tripod for peace: international organizations and militarized disputes, 1950-1985." *International Organization*. Summer. v52 n3 p441(27).
- Russett, Bruce and John R. Oneal. 2001. *Triangulating Peace: Democracy, Interdependence, and International Organizations*, W.W. Norton & Company.
- Rubin, D. B. 1974. "Estimating causal effects of treatments in randomized and non-randomized studies." *Journal of Educational Psychology*, 66, 688-701.
- Scheinerman, Edward. 2000. *Mathematics: A Discrete Introduction*, Brooks/Cole.
- Scheinerman, Edward and Daniel Ullman, 1997. *Fractional Graph Theory: A Rational Approach to the Theory of Graphs*. Wiley.
- Scheines, R. 1997. "An Introduction to Causal Inference," in *Causality in Crisis*, ed. by Steven Turner and Vaughan McKim, University of Notre Dame Press.
- Scheines, Richard, Peter Spirtes, Clark Glymour, and Christopher Meek, 1994. *TETRAD II: Tools for Discovery*. Lawrence Erlbaum Associates, Hillsdale, NJ. Also see TETRAD webpage: <http://www.phil.cmu.edu/projects/tetrad/>
- Signorino, Curt. 1999. "Strategic Interaction and the Statistical Analysis of International Conflict." *American Political Science Review*. 93(2):279-298.
- Signorino, Curtis S., and Jeffrey M. Ritter. 1999. "Tau-b or Not Tau-b: Measuring the Similarity of Foreign Policy Positions." *International Studies Quarterly*. 43(1): 115-44.
- Snyder, David and Edward L. Kick. 1979. "Structural position in the world system and economic growth 1955-1970: A multiple-network analysis of transnational interactions," *American Journal of Sociology*, 84: 1096-1126
- Spirtes, P., Glymour, C., and Scheines, R. 2000. *Causality, Prediction and Search*. 2nd ed., The MIT Press.
- Spirtes, P., Glymour, C., and Scheines, R. 1997. "Reply to Humphreys and Freedman's Review of 'Causality, Prediction and Search', *British Journal for the Philosophy of Science*. 48:555-568.
- Spirtes, P., Richardson, T., Meek, C., Scheines, R., and Glymour, C. 1998. "Using path diagrams as a structural modeling tool." *Sociological Methods and Research*, 27, 182-225.

- Strauss, D. and M. Ikeda 1990, "Pseudo likelihood Estimation for Social Networks." *Journal of American Statistical Association*. 85, 204-212.
- Ty Wilde, E. and Hollister, R. 2002. "How Close Is Close Enough? Testing Non-experimental Estimates of Impact against Experimental Estimates of Impact with Education Test Scores as Outcomes." IRP Discussion Paper.
- Waltz, Kenneth N., 1979. *Theory of International Politics*. McGraw-Hill.
- Ward, Michael D. and Kristian S. Gleditsch. N.D. "Location, location, location: An MCMC approach to modeling spatial context with categorical variables." forthcoming, *Political Analysis*.
- Wasserman, Stanley, and Katherine Faust. 1994. *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press.
- Wasserman, S., and Galaskiewicz, J. 1994. *Advances in Social Network Analysis: Research from the Social and Behavioral Sciences*. Newbury Park, CA: Sage.
- Wasserman, S., and Pattison, P. 1996. "Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p^* ." *Psychometrika*, 60, 401-426.
- Wasserman, S., and Pattison, P. 2000. *Multivariate Random Graph Distributions*. Springer Lecture Note Series in Statistics. New York: Springer-Verlag.
- Wellman, Barry and S.D. Berkowitz (eds.) 1988. *Social structures: A network approach*. Cambridge: Cambridge University Press.
- Winship, Christopher and Michael Mandel. 1983. "Roles and positions: A critique and extension of the block modeling approach," pp. 314-344 in Samuel Leinhardt (ed.) *Sociological Methodology 1983-1984*. San Francisco: Jossey-Bass
- Winship, C., & Morgan, S.L. 1999. "The estimation of causal effects from observational data". *Annual Review of Sociology*, 25, 659-707.
- Zeng, L. 1999. "Classification and Prediction with Neural Network Models", *Sociological Methods and Research*. Vol. 27, No. 4, 499-524.
- Zeng, L. 2000a. "Neural Network Models for Political Analysis", in Diana Richards (ed.) *Political Complexity: Nonlinear Models of Politics*. University of Michigan Press, pp. 239-268.