Cycle Census Statistics for Exponential Random Graph Models*†

Carter T. Butts‡

1/17/06

Abstract

Exponential family models for random graphs (ERGs, also known as $p*$ models) are an increasingly popular tool for the analysis of social networks. ERGs allow for the parameterization of complex dependence among edges within a likelihood-based framework, and are often used to model local influences on global structure. This paper introduces a family of cycle statistics, which allow for the modeling of long-range dependence within ERGs. These statistics are shown to arise from a family of partial conditional dependence assumptions based on an extended form of reciprocity, here called reciprocal path dependence. Algorithms for computing cycle statistic changescores and the cycle census are provided, as are analytical expressions for the first and approximate second moments of the cycle census under a Bernoulli null model. An illustrative application of ERG modeling using cycle statistics is also provided.

Keywords: exponential family models, cycles, random graphs, partial conditional dependence models, social networks

1 Introduction

Exponential family models for random graphs (ERGs) are an increasingly popular tool for the analysis of social networks (see, e.g., Carrington et al.,

*This work was supported by NIH award 5 R01 DA012831-05 and NSF ITR award IIS-0331707.

†The author would like to thank Martina Morris, Mark Handcock, Ulrik Brandes, David Hunter, and Garry Robins for their helpful input.

‡Department of Sociology and Institute for Mathematical Behavioral Sciences; SSPA 2145; University of California, Irvine; Irvine, CA 92697; buttsc@uci.edu

2005, for examples and a review of recent work). Central to the appeal of ERGs is their ability to parameterize complex dependence among edges, while still supporting likelihood based inference. Arguably, this combination allows ERG models to fulfill some of the roles traditionally occupied by approaches such as agent based modeling (e.g., exploring the global consequences of local organizing principles) as well as conventional statistical techniques (e.g., estimation and model selection).

Much – if not most – of the existing work on exponential family models for random graphs is focused on *local structure*, i.e. structural properties which depend only on the first or second order neighborhood of a focal vertex. Prominent examples of such properties include reciprocity, transitivity, clustering, and the degree distribution, as well as covariate-dependent effects such as homophily and propinquity. While such properties are inarguably important, it is not clear that they capture all processes of substantive interest. Indeed, it has long been argued that local structure may itself be affected by large-scale network properties. In negative exchange networks, for instance, changes to distant edges can affect the balance of even/odd paths between actors, thereby changing their payoffs (and, hence, incentives to trade) (Willer, 1999). Long-range cycles have similarly been hypothesized to be of importance in maintaining reputational and generalized exchange systems, serving as redundant conduits for the flow of goods and information (including information about the trustworthiness of one's local trading partners) (Bearman, 1997; Yamagishi and Cook, 1993). The presence of such redundant connections is thought to be an important factor in organizational resilience, particularly within turbulent environments (e.g., disasters) (Kendra and Wachtendorf, 2003); on the other hand, these same features may be selected against within contexts such as sexual contact networks, where they may increase network members' exposure to communicable disease.

Here, we take ERG models in a more global direction by introducing effects for cycles of varying lengths. As we show, these effects can be understood as capturing a form of *extended reciprocity,* in which potential edges are affected by the presence or absence of redundant paths between endpoints. While computation for the associated path counts is potentially expensive, we provide algorithms which allow for the use of cycle effects in typical empirical settings. These algorithms can also be employed to compute a partial (or, for small graphs, total) cycle census; comparison of the observed cycle census versus the expected census under a simple null model of edgewise independence provides a useful exploratory mechanism for identifying structural biases. Finally, we demonstrate the use of cycle statistics

for both exploratory and modeling purposes, using four data sets which span a range of relational types.

1.1 Brief Review of Discrete Exponential Family Models

Before turning to the cycle statistics per se, it is first useful to briefly review the basic definition of the ERG representation.[1] Let G be a random graph with countable support $\mathcal{G}$, and let $\mathbf{t} : \mathcal{G} \mapsto \mathbb{R}^m$ be a vector of *sufficient statistics* for G. We may then represent the probability mass function of G by a discrete exponential family, i.e.

$$\Pr(G = g|\theta) = \frac{\exp\left(\theta^T \mathbf{t}(g)\right)}{\sum_{g' \in \mathcal{G}} \exp\left(\theta^T \mathbf{t}(g)\right)}, \tag{1}$$

where $\theta \in \mathbb{R}^m$ is a vector of parameters. Intuitively, the ERG acts to shift probability mass towards graphs for which $\theta_i t_i(g)$ is large, and away from graphs for which the corresponding statistic is small (or highly negative). $\theta^T \mathbf{t}$ thus acts as a potential function for G, with $\mathbf{t}$ indicating the graph features which are positively/negatively weighted by the model.

Given an ERG model, simulation of draws from G may be accomplished via Markov Chain Monte Carlo (MCMC) methods; see, e.g., Crouch et al. (1998); Snijders (2002). Inference for θ given G is complicated by the presence of the unknown normalizing factor within the likelihood for G (i.e., the sum over $\mathcal{G}$ above) which depends upon θ but not upon G. Since the cardinality of $\mathcal{G}$ (denoted $|\mathcal{G}|$) is typically on the order of 2^{N^2} or greater, direct computation of the normalizing factor is generally infeasible; simple Monte Carlo quadrature is also difficult, due to the high variance of the summand. Current estimation methods thus employ importance sampling for variance reduction, approximate the likelihood by a product of conditional likelihoods (the pseudolikelihood method), or dodge the matter entirely by seeking to solve the likelihood equation,

$$\mathbf{E}_{\hat{\theta}} \mathbf{t}(G) = \mathbf{t}(g_{\text{obs}}) \tag{2}$$

(where g_{obs} is the observed graph, and $\mathbf{E}_{\hat{\theta}} \mathbf{t}(G)$ is the expectation of $\mathbf{t}(G)$ under the MLE). While efficient techniques for the solution of these problems remains an area of active research, we will not treat the matter in detail here. We simply note that workable approaches do exist for many problems

[1] Properly speaking, the ERG/$p*$ framework is a general way of parameterizing probability models on graphs, rather than being a model per se. We will generally avoid this distinction here.

of substantive interest (see Wasserman and Robins, 2005, for a brief review), and that several software implementations of these methods are currently available (e.g., Snijders, 2001; Handcock et al., 2003). Our purpose here, rather, is centered on the definition of $\mathbf{t}$ based on cyclical properties of G, and on the computation for the associated statistics. It is to this problem that we now turn.

2 Cycle Statistics: Parameterization and Computation

To model more global properties within an exponential family context, we here introduce a family of sufficient statistics based on cycle counts. Specifically, let $G = (V, E)$ be a loopless graph on N vertices. We define the ith *cycle statistic*, c_i, of G by $c_i(G) \equiv |\{g \subseteq G : g \cong C_i\}|$, where C_i is the cycle on i vertices (or directed cycle, if G is directed) and $\cong$ is the isomorphism relation. Thus, $c_i(G)$ is the number of i-cycles in G. Trivially, $c_i(G) = 0$ forall $i \notin 2, \ldots, N$ (or $i \notin 3, \ldots, N$, if G is undirected); otherwise, c_i will depend upon the structure of G. The set of all cycle statistics for G is called the *cycle census* of G, in direct analogy with the well-known dyad and triad censuses (Holland and Leinhardt, 1976). Cycle statistics are in general affinely independent of each other, and may be used within an ERG model in the same manner as other statistics such as k-stars, triad counts, etc.

As noted above, an exponential family model acts to place probability mass in log-proportion to the potential $\theta^T \mathbf{t}$. Where cycle statistics are included within $\mathbf{t}$, then, the corresponding θ parameters can be interpreted as promoting or inhibiting cycle formation (depending on the sign of θ_i). Further, we can observe from Equation 2 that the maximum likelihood estimator for an ERG containing cycle statistic c_i corresponds to a model in which the expected number of i-cycles is equal to the number of i-cycles in the observed graph. A complete cycle census model, then, which sets $\mathbf{t} = \mathbf{c}$, can be understood as preserving the expected cycle distribution at all scales. Incomplete cycle census models (for which $\mathbf{t} \subset \mathbf{c}$) likewise preserve cycle distributions at some lengths, but not others; these models may be preferable when structural dependence is believed to be limited to certain scales.

It should be noted that some cycle statistics arise naturally from other models, and are already in wide use. In the directed case, the 2-cycle statistic is identical to the number of mutuals, a statistic which first appears in the

dyad dependence models (including $U|MAN$, p_1, and p_2). 3-cycles (in both the directed and undirected cases) appear with the Markov graphs (Frank and Strauss, 1986), wherein dependence is generalized from dyads to edges sharing at least one endpoint. Pattison and Robins (2002) have further suggested a family of partial conditional dependence models based on 3-paths, which include 4-cycle statistics in their full specification. Thus, local cycle statistics are a familiar component of many ERG models. Longer-range cycles, on the other hand, have not been employed in past work; similarly, little is known about the nature of models which are constructed from cycle statistics *per se*. Before turning to matters of computation, then, we briefly explore some of the properties of cycle census models.

2.1 Models from Reciprocal Path Dependence

While cycle statistics may be motivated by direct, substantive considerations, they may also arise as a result of assumptions regarding conditional dependence among edges. As we have already seen, cycles of length 2 and 3 arise naturally from the Markov graphs (Frank and Strauss, 1986), and cycles of length 4 are implicated in the 3-path models of Pattison and Robins (2002). It is thus natural to ask whether there is a family of models which is parameterized through more general cycle statistics, and (if so) whether this may give us some insight into the properties of the cycle census. As it happens, it is possible to generate such a family as a partial conditional dependence model, using the notion of *reciprocal path dependence*.

The core idea behind reciprocal path dependence is that two possible edges are conditionally dependent only if their respective endpoints are joined by (appropriately directed) paths. While this assumption can be motivated in a number of ways, possibly the most obvious is via an extended notion of reciprocity. Under dyadic reciprocity, the probability of an edge from vertex i to vertex j is dependent upon the existence of an edge from vertex j to vertex i; intuitively, an edge which establishes a mutual relationship is not the same as an edge which establishes an asymmetric relationship. A similar argument may be made in the triadic case, wherein an (i, j) edge may be made more or less likely by the existence of an indirect relationship from j to i mediated by some third party, k. If we continue to extend the potential for indirect reciprocity, allowing dependence at ever greater distances, we eventually arrive at the case in which the potential for an (i, j) edge to reciprocate a j, i path of any length could (potentially) prove consequential. This limiting case is the reciprocal path dependence

model.[2]

To express this intuition more formally, we define a series of *reciprocal path conditions* on the edges of G.

Definition 1. Let $G = (V, E)$ be a directed graph, with $(i, j), (k, l)$ being vertex pairs such that $i \neq j, k \neq l$, $i, j, k, l \in V$. Let P_{ab} denote a directed a, b path in G. Then (i, j) and (k, l) are said to satisfy the *strong reciprocal path condition* if

1. $\{i, j\} \cap \{k, l\} = 2$
2. $i = l$ and $\exists\ P_{jk} \in G$;
3. $j = k$ and $\exists\ P_{li} \in G$; or
4. $\exists\ P_{jk}, P_{li} \in G : P_{jk} \cap P_{li} = \varnothing$.

We say that $(i, j), (k, l)$ satisfy the *weak reciprocal path condition* if (1), (2), or (3) above is true, or if $\exists\ P_{jk}, P_{li} \in G$. Now, consider the case in which G is an undirected graph, with vertex pairs $\{i, j\}, \{k, l\} : i \neq j, k \neq l$, $i, j, k, l \in V$. Let P_{ab} denote an undirected a, b path in G. Then $\{i, j\}$ and $\{k, l\}$ are said to satisfy the strong reciprocal path condition if

1. $\{i, j\} \cap \{k, l\} = 2$;
2. $\{i, j\} \cap \{k, l\} = 1$ and $\exists\ P_{ab} \in G : a \cup b = (\{i, j\} \cup \{k, l\}) \setminus (\{i, j\} \cap \{k, l\}), (\{i, j\} \cap \{k, l\}) \notin P_{ab}$; or
3. $\exists\ P_{ab}, P_{cd} \in G : \{a, c\} = \{i, j\}, \{b, d\} = \{k, l\}, P_{ab} \cap P_{cd} = \varnothing$.

If (1) or (2) above is true, or if $\exists\ P_{ab}, P_{cd} \in G : \{a, c\} = \{i, j\}, \{b, d\} = \{k, l\}, \{\{i, j\}, \{k, l\}\} \cap (P_{ab} \cup P_{cd}) = \varnothing$, then $\{i, j\}, \{k, l\}$ are said to satisfy the weak reciprocal path condition. Given a choice of condition, we use the expression aRb to denote the statement "a and b satisfy the reciprocal path condition." The negation of this statement is written as $a\overline{R}b$. R is a symmetric binary relation, and thus $aRb \Leftrightarrow bRa$, $a\overline{R}b \Leftrightarrow b\overline{R}a$. ■

The above definitions can be restated as follows. A pair of directed dyads satisfy the strong reciprocal path condition if they are identical, or if there is a directed path from the receiver of the first dyad to the sender of the first dyad which includes the second directed dyad (treating the second directed

[2] In practice, it is possible (and generally desirable) to limit the range at which dependence can occur. This is demonstrated in Section 3, below.

edge as being present for purposes of the condition). Such a pair satisfies the weak reciprocal path condition if it satisfies the strong condition, or if there are paths connecting the receiver of each directed dyad to the sender of the other. Thus, the strong and weak conditions differ only in that the former requires the paths connecting the respective dyads to be independent. The definitions for the undirected case are exactly analogous to those of the directed case, save that the orientation of edges (and hence paths) is ignored.

Now, what can be said of an ERG model which treats edges as dependent only if they satisfy a reciprocal path condition? In particular, what sufficient statistics are needed to implement such a model? Although the Hammersley-Clifford Theorem (Besag, 1974) provides a means of identifying the sufficient statistics associated with a given conditional dependence structure, that is not adequate here. Let $\mathbf{X}$ be the adjacency matrix associated with random graph G, such that $X_{ij} = 1$ if $(i,j) \in E$ and $X_{ij} = 0$ otherwise. (In a slight abuse of notation, we will treat (i,j) as equivalent to $\{i,j\}$ in the remainder of this subsection, where G is undirected.) Denote the cells of $\mathbf{X}$ not corresponding to pairs $(a,b),(c,d),\ldots$ by $\mathbf{X}^c_{ab,cd,\ldots}$. Now let $\mathcal{E} = \{(i,j) : i \neq j, i,j \in V\}$, and let $D = (\mathcal{E}, E')$ be a simple graph such that $\{(i,j),(k,l)\} \in E'$ iff $X_{ij} \not\perp X_{kl}|\mathbf{X}^c_{ij,kl}$. D is then the *conditional dependence graph* of G. By the Hammersley-Clifford Theorem, a discrete exponential family model whose sufficient statistics are the products of cells in $\mathbf{X}$ whose associated edges form cliques in D is sufficient to implement the conditional dependence structure of G. Thus, it would appear, we must merely identify the D associated with each reciprocal path condition to find the sufficient statistics of G.

As it happens, the reality of the situation is a bit more complex: since, in general, any edge pair can satisfy the reciprocal path conditions, it follows that D is a complete graph for any G implementing a reciprocal path dependence model. If D is a complete graph, then all subsets of D are cliques; thus all possible subsets of edges can generate sufficient statistics for G. To resolve the situation, we note that the statistics generated by the Hammersley-Clifford construction are always sufficient to implement the associated model, but are not always *necessary*. As Pattison and Robins (2002) have demonstrated, it is often possible to use additional constraints on the dependence structure to pare down the set of sufficient statistics. Specifically, they show that assumptions of edgewise dependence conditional on particular realizations of the rest of the graph can serve to generate more restrictive models. In our case, the reciprocal path dependence assumption is precisely of this type: we posit conditional dependence of edges *given* that the appropriate path condition is satisfied, assuming edges to be otherwise

independent. Pattison and Robins describe this as "partial conditional dependence," and provide a theorem for the construction of sufficient statistics given certain assumptions of this type. Although we cannot use their result directly (since the reciprocal path dependence conditions do not satisfy the assumptions of their theorem), we can derive a similar theorem which does apply in our case.

Theorem 1. *Let* $\mathbf{X}$ *be a random adjacency matrix whose pmf is a discrete exponential family satisfying a reciprocal path dependence assumption under condition* R*. Then the sufficient statistics for* $\mathbf{X}$ *are functions of edge sets* S *such that* $(i,j)R(k,l)\ \forall\ \{(i,j),(k,l)\} \subseteq S$*.*

Proof. Our argument initially shadows that of Pattison and Robins (2002: 332–334), departing from it in the nature of the partial conditional dependence assumed. We begin (as do they) with Besag's (1974) definition of $\mathcal{Q}(\mathbf{x}) = \ln\left[\Pr\left(\mathbf{X}=\mathbf{x}\right)/\Pr\left(\mathbf{X}=\mathbf{0}\right)\right]$, where $\mathbf{X}$ is a random adjacency matrix (with realization $\mathbf{x}$), and $\mathbf{0}$ is a zero-matrix with the same dimensions as $\mathbf{x}$. Let $\mathbf{X}_{ij}^{-}$ represent the matrix $\mathbf{X}$ with X_{ij} set equal to 0. As per Besag (1974),

$$\mathcal{Q}(\mathbf{x}) - \mathcal{Q}(\mathbf{x}_{ij}^{-}) = \ln\left[\frac{\Pr\left(\mathbf{X}=\mathbf{x}\right)}{\Pr\left(\mathbf{X}=\mathbf{x}_{ij}^{-}\right)}\right] \tag{3}$$

$$= \ln\left[\frac{\Pr\left(X_{ij}=x_{ij}\,\middle|\,\mathbf{X}_{ij}^{c}=\mathbf{x}_{ij}^{c}\right)}{\Pr\left(X_{ij}=0\,\middle|\,\mathbf{X}_{ij}^{c}=\mathbf{x}_{ij}^{c}\right)}\right], \tag{4}$$

and from the Hammersley-Clifford Theorem,

$$= x_{ij}\left[\sum_{A\subseteq M\setminus\{(i,j)\}} \lambda_{A\cup\{(i,j)\}} \prod_{(k,l)\in A} x_{kl}\right] \tag{5}$$

where M is the set of directed dyads on V, $\lambda_{A\cup\{(i,j)\}} \in \mathbb{R}$, and $\lambda_{A\cup\{(i,j)\}} = 0$ if $A\cup\{(i,j)\}$ is not a clique of D.

As noted above, D is a complete graph under the reciprocal path conditions – thus, the Hammersley-Clifford Theorem alone cannot place useful restrictions on the number of sufficient statistics required to parameterize such models. Like Pattison and Robins (2002), we thus invoke a form of partial conditional dependence to force certain λ coefficients to 0. We begin by noting that, under a reciprocal path dependence model, $x_{ij} \perp x_{kl}$ unless

the appropriate reciprocal path condition is met. For some fixed (k, l), let $\mathbf{x}^c_{ij}, \mathbf{x}'^c_{ij}$ be such that $(i, j)\overline{R}(k, l)$, with $\left(\mathbf{x}^c_{ij}\right)_{kl} = 1$ and $\left(\mathbf{x}'^c_{ij}\right)_{kl} = 0$. By assumption, $(i, j)\overline{R}(k, l)$ implies $X_{ij} \perp X_{kl}|\mathbf{X}^c_{ij,kl}$. It follows, then, that

$$\ln\left[\frac{\Pr\left(X_{ij} = x_{ij}\left|\mathbf{X}^c_{ij} = \mathbf{x}^c_{ij}\right.\right)}{\Pr\left(X_{ij} = 0\left|\mathbf{X}^c_{ij} = \mathbf{x}^c_{ij}\right.\right)}\right] = \ln\left[\frac{\Pr\left(X_{ij} = x_{ij}\left|\mathbf{X}^c_{ij} = \mathbf{x}'^c_{ij}\right.\right)}{\Pr\left(X_{ij} = 0\left|\mathbf{X}^c_{ij} = \mathbf{x}'^c_{ij}\right.\right)}\right], \quad (6)$$

and hence,

$$0 = x_{ij}\left[\left(\sum_{A\subseteq M\setminus\{(i,j)\}} \lambda_{A\cup\{(i,j)\}} \prod_{(m,n)\in A} x_{mn}\right) - \left(\sum_{A\subseteq M\setminus\{(i,j)\}} \lambda_{A\cup\{(i,j)\}} \prod_{(m,n)\in A} x'_{mn}\right)\right] \quad (7)$$

$$= x_{ij}\left[\sum_{A\subseteq M\setminus\{(i,j)\}} \lambda_{A\cup\{(i,j)\}}\left(\prod_{(m,n)\in A} x_{mn} - \prod_{(m,n)\in A} x'_{mn}\right)\right] \quad (8)$$

This condition is satisfied (in general) iff $\lambda_S = 0$ forall S such that $(k, l) \in S$, $(i, j)\overline{R}(k, l)$. Note that we do *not* have to assume that $\lambda_S = 0$ for S such that $(i, j)R(k, l)$, because (by construction) the sufficient statistics associated with any such terms in Equation 8 must be 0.

The above deals with the dependency of a single pair of edges; to restate, we have shown that a reciprocal path dependency assumption implies that any sufficient statistic which is a function of edge pair $(i, j), (k, l)$ must be associated with a zero coefficient unless $(i, j)R(k, l)$. Since this condition applies to all edge pairs simultaneously, it follows that the only sufficient statistics with nonzero λ values can be those for which all pairs of included edges satisfy R. Stated more formally, $\lambda_S \neq 0$ only if $(i, j)R(k, l)$ $\forall$ $\{(i, j), (k, l)\} \subseteq S$. Note that this result also extends naturally to the undirected case, e.g., by restricting attention to the lower triangle of $\mathbf{X}$. $\square$

The practical importance of Theorem 1 lies in the fact that only a small number of configurations satisfy the various reciprocal path conditions. In the strong, directed case, S consists entirely of edge variables (which trivially satisfy R) and directed cycles. Applying the standard assumption of homogeneity under isomorphism, we arrive at a model whose sufficient statistics are the edge count, together with the number of directed cycles of each length (i.e., the cycle census). The weak directed case is less restrictive: in addition to edges, various directed closed walks (cycles with repeated edges

and/or vertices) are also admissible. These statistics are more difficult to fully characterize, and we will not attempt to do so here. In the undirected case, the strong condition implies that S consists of all subgraphs of G having a spanning cycle; chords are thus permitted here, unlike the directed case. Statistics for the weak condition in the undirected case are analogous to the directed case, save that the cycles involved are undirected.

The above suggests a particularly important role for the cycle census: together with the edge count, the cycle census is sufficient for the (homogeneous) strong reciprocal path dependence model in the directed case. Unlike dyadic, Markov, or three-path models, the RPD model is truly global in character. Edges within a RPD model can depend on other edges at geodesic distances of $\mathcal{O}(N)$, as compared with distances of 1 to 3 for the local structure models. At the same time, this long-range dependence is managed within a small number of sufficient statistics (in the strong case, at least), greatly facilitating estimation and interpretation.

Of course, these attractive properties do not make the RPD model a panacea. The specific form of long-range dependence implied by the reciprocal path constraint may be reasonable for certain processes (e.g., generalized exchange), but is likely insufficient to capture other phenomena (e.g., heterogeneity in number of partners). Furthermore, actually computing the cycle census is itself a substantial challenge. It is to this problem that we now turn.

2.2 Computation

Practical inference for cycle parameters would seem to pose formidable computational challenges. Since computation of the cycle census is at least NP-complete (note that it includes the Hamiltonian graph problem as a special case[3]), it is not immediately obvious that it is possible to employ cycle statistics for graphs of more than minimal size. While the complexity of cycle computation does indeed place some limits on what models can be fit, these limits are less severe than one might expect. With that in mind, we here provide some simple algorithms for use in computing cycle census statistics for ERG models; examples of the use of these algorithms may be found in Section 3.

Although there is no known means of fully circumventing the problem of NP-completeness, there are several aspects of the ERG context which allow us to employ cycle census statistics in practice. First, most networks

[3]A graph is Hamiltonian if it contains a spanning cycle; determining whether a graph is Hamiltonian is a well-known NP-complete problem (West, 1996).

of empirical interest are exceedingly sparse. Social actors typically have upper bounds on the number of relationships they can sustain (Mayhew and Levinger, 1976), leading to densities which fall approximately as $1/N$ in graph size. This limits the number of cycles which are realized in practice, a fact which can be exploited when designing cycle counting algorithms. A second useful observation is the fact that we do not have to calculate the entire cycle census in order to fit or simulate draws from an ERG model – following Snijders (2002), we need only calculate the *changes* in the cycle statistics associated with the addition or deletion of a given edge. Finally, we note that the phenomena of interest to most researchers imply dependence over a relatively restricted range of distances. Interdependencies induced by inbreeding taboos or incentives from generalized exchange are anticipated to weaken as one moves from relatively short cycles (e.g., length 4-5) to long cycles (e.g., length 6 or greater). As a result, we may often restrict our attention to cycles of fixed maximum length. Computation for these bounded cycles is often feasible even for large graphs, despite the fact that a complete cycle census may be unavailable. Our approach is thus to focus on computation for all cycles of length $2, \ldots, \ell_{\max}$, where $\ell_{\max}$ is some user-specified upper bound. We provide algorithms for computing changescores for such statistics, and for computation of the associated (partial) cycle census.

2.2.1 Changescore Algorithms

Let $G = (V, E)$ be a (possibly directed) graph, and let (v, v') be a pair of distinct vertices. We seek to compute the changes in the numbers of cycles of lengths $2, \ldots, \ell_{\max}$ (respectively) which result from setting $E := E \cup (v, v')$ (if $(v, v') \notin E$), or setting $E := E \setminus (v, v')$ (if $(v, v') \in E$). These changescores are computed by Algorithm 1, such that c_i is the change in the number of cycles of length $i + 1$ produced by a change in the state of the (v, v') edge. (Note that $c_1 = 0$ in the undirected case.) Algorithm 1 works by counting the numbers of paths of length $1, \ldots, \ell_{\max} - 1$ from v' to v in G. If $(v, v') \in E$, then each of these paths join with the (v, v') edge to form cycles; removal of the (v, v') edge thus removes exactly these cycles from G. By turns, if $(v, v') \notin E$, each path from v' to v will contribute one cycle to $G \cup (v, v')$. Since no other cycles can be added/removed by a change to the (v, v') edge, it follows that the changescore computation reduces to the problem of counting (v', v) paths.

Counting the (v', v) paths at each length is fairly straightforward. One approach is provided in Algorithms 2 and 3. Algorithm 2 begins in lines 3–5

Algorithm 1 Routine to Compute Changescores for Cycle Terms

1: **procedure** `CycleChangescore`$(G, v, v', \ell_{\max})$
2: $\mathbf{p}$:=`PathCount`$(G, v', v, \ell_{\max} - 1)$
3: **if** $(v, v') \in E$ **then**
4: $\quad \mathbf{c} := -\mathbf{p}$
5: **else**
6: $\quad \mathbf{c} := \mathbf{p}$
7: **end if**
8: **return c**

by checking for adjacency (a 1-path) in the directed case; this is handled here (rather than in the subsequent recursive step) due to the fact that directedness of G affects only the initial adjacency. Algorithm 2 next creates A (the set of potentially available internal vertices) in line 6. Each member of A adjacent to the initial vertex is then used as a seed for a depth-first recursion (lines 7–11). Upon termination, the vector of path counts ($\mathbf{p}$) is returned.

Algorithm 2 Path Counting Routine

1: **procedure** `PathCount`$(G, v, v', \ell_{\max})$
2: $\mathbf{p} := \{0\}_{\max}^{\ell}$
3: **if** (G is directed) $\wedge$ $((v, v') \in E)$ **then**
4: $\quad p_1 := 1$
5: **end if**
6: $A := V \setminus \{v, v'\}$
7: **for** $v_c \in A$ **do**
8: $\quad$ **if** $(v, v_c) \in E$ **then**
9: $\quad\quad$ `PathCountRecurse`$(G, v_c, 2, v', A, \mathbf{p}, \ell_{\max})$
10: $\quad$ **end if**
11: **end for**
12: **return p**

The recursive step (Algorithm 3) proceeds as follows. Given that we have arrived at vertex v_c, we check for adjacency to the destination vertex (v_d), incrementing the path count if such an adjacency is found (lines 2-4). Assuming that we have neither exhausted the available vertices nor reached the maximum path length (line 5), we update A by removing the current vertex (line 6). For each neighbor of v_c remaining in A, we further recurse using the updated available set and incremented path length (lines 7-11).

Algorithm 3 Recursive Subroutine for Path Counting Algorithm

procedure `PathCountRecurse`$(G, v_c, \ell, v_d, A, \mathbf{p}, \ell_{\max})$
if $(v_c, v_d) \in E$) **then**
  $p_\ell := p_\ell + 1$
end if
if $(|A| > 0) \wedge (\ell < \ell_{\max})$ **then**
  $A := A \setminus v_c$
  for $v \in A$ **do**
    if $(v_c, v) \in E$ **then**
      `PathCountRecurse`$(G, v, \ell + 1, v_d, A, \mathbf{p}, \ell_{\max})$
    end if
  end for
end if

Recursive subtraction from the initial set of available vertices in these two algorithms guarantees that no vertices are visited twice in any sequence of recursions, and recursion on each element of the set at each iteration guarantees that all permissible sequences of vertices are employed. The maximum depth of recursion is bounded by the maximum path length (two less than the initial $\ell_{\max}$), and the depth-first structure further guarantees that at most $\ell_{\max} - 2$ instances of the subroutine will be active at any given time. This is important, since the total number of paths to be traced may be quite large. Worst case performance for Algorithms 2 and 3 clearly occurs on the complete graph, where all vertices are adjacent. In that case, $N - 1 - i$ branches must be explored at each depth-i branch, for a total number of $\prod_{i=1}^{\ell_{\max}-2}(N-1-i)$ calls to `PathCountRecurse`. This is obviously bounded above by $N^{\ell_{\max}-2}$, which provides a conservative sense of the time complexity involved. A somewhat more reasonable approximation to the average time is given by $\bar{d}^{\ell_{\max}-2}$, where $\bar{d}$ is the mean degree of G. As this suggests, it may be possible to consider reasonably long cycles where $\bar{d}$ is small, so long as G does not contain too many large cliques.

2.2.2 The Cycle Census

While the above algorithms are obviously designed for changescore computation, they can also be used to obtain exact cycle census statistics. A simple method proceeds as follows. Given target graph $G = (V, E)$, let $G' = (V, \varnothing)$ and let $\mathbf{c} = \{0\}^{\ell_{\max}-1}$. Then, for each $(v, v') \in E$, add `CycleChangescore`$(G', v, v', \ell_{\max})$ to $\mathbf{c}$ and let $G' = G' \cup (v, v')$. This method

is somewhat slow (it requires $|E|$ calls to the changescore routine), but is often adequate for sparse G.

2.3 Cycle Statistics in the Homogeneous Bernoulli Case

While our focus here is on the use of cycle terms in more general exponential family models, it is also worth noting some simple results regarding the distribution of cycle census statistics under a common null model – the homogeneous Bernoulli graph (Erdös and Rényi, 1960). For random graph G, the homogeneous Bernoulli pmf is given by

$$\Pr(G = g|\theta) = \begin{cases} \theta^{|E(g)|}(1-\theta)^{2\binom{N}{2}-|E(g)|} & \text{if } G \text{ directed} \\ \theta^{|E(g)|}(1-\theta)^{\binom{N}{2}-|E(g)|} & \text{if } G \text{ undirected} \end{cases} \tag{9}$$

where $N = |V|$. Since all edges of G are (by definition) iid Bernoulli, it follows that the expected number of k-cycles is equal to the number of k-subsets of V, multiplied by θ^k (the probability of observing k edges) times the number of distinct cycles per k-set. If $t_i(G)$ is the number of length-i cycles in G, it thus follows that

$$\mathbf{E}_\theta\big(t_i\,(G)\big) = \begin{cases} \frac{\prod_{j=0}^{i-1}(N-j)}{i}\theta^i & \text{if } G \text{ directed} \\ \frac{\prod_{j=0}^{i-1}(N-j)}{2i}\theta^i & \text{if } G \text{ undirected} \end{cases} \tag{10}$$

(see Bollobás (1998)). While the exact variance of t_i under the Bernoulli graph model is nontrivial to compute, a rough approximation may be obtained by treating all potential cycles as independent; this implies that t_i is approximately binomial, and hence that

$$\mathbf{V}_\theta\big(t_i\,(G)\big) \approx \begin{cases} \frac{\prod_{j=0}^{i-1}(N-j)}{i}\theta^i(1-\theta^i) & \text{if } G \text{ directed} \\ \frac{\prod_{j=0}^{i-1}(N-j)}{2i}\theta^i(1-\theta^i) & \text{if } G \text{ undirected} \end{cases}. \tag{11}$$

Although crude, this approximation provides a quickly calculated heuristic which may be used to flag census statistics (e.g., using z-scores) which deviate greatly from a random baseline. More accurate estimation/testing of cycle effects can then be performed using the appropriate ERG model. It should be noted, in any event, that cycle census statistics become extremely right-skewed for longer cycles; thus, z-scores will only serve as reasonable approximations for cycles of fairly short length.

3 Application

To illustrate the use of cycle census statistics in practice, we here apply these measures to sample networks taken from a variety of settings. Specifically, we employ data on the gift-exchange of taro root among households in a Papuan village (Schwimmer, 1973); reciprocal communication among organizations conducting search and rescue operations following the flooding in the Texas hill country (Drabek et al., 1981); friendship nominations among adolescents (Coleman, 1964); and militarized interstate disputes among nations during the year 2000 (Ghosen et al., 2004). Sociograms for the four sample networks are provided in Figure 1. As the figure shows, these networks vary substantially in size and structure. (The friendship and MID relations are also directed, unlike the taro exchange and Texas organizational network.)

3.1 Cycle Census Comparison

We begin with a comparison of observed cycle census statistics with those expected under the Bernoulli null model. Using the algorithms of Section 2.2, we may calculate the cycle census for each network. These are shown (up to length 12) in Table 1. As the table illustrates, very different cycle distributions are observed across networks. In the taro exchange and the Texas emergent multiorganizational network (EMON), we see an increasing number of cycles at longer lengths. This is consistent with the observation that (per Figure 1) these two networks are heavily triangulated, and only weakly clustered. By contrast, the friendship and MID relations show a preponderance of cycles at short lengths, with counts dying out (quite abruptly, for the MIDs) as one approaches the 6-7 range. As this would suggest, both networks are poorly connected, and their larger components are either tree-like or weakly triangulated. Such structures generate few long cycles, suggesting that dependence may remain relatively localized in these networks.

A somewhat different view of these statistics is obtained when we consider the expected cycle counts. Figure 2 shows the log ratio of observed versus expected counts (under the Bernoulli model) for each graph. Contrary to what might be expected, each network shows the same general pattern of above-average local clustering, accompanied by a tendency towards underrepresentation of long cycles. As this illustrates, it is somewhat hazardous to evaluate the cycle census without controlling for the effects of size and density; even where cycle counts seem to be growing rapidly (as in the Texas EMON), this growth may be below that expected by chance.

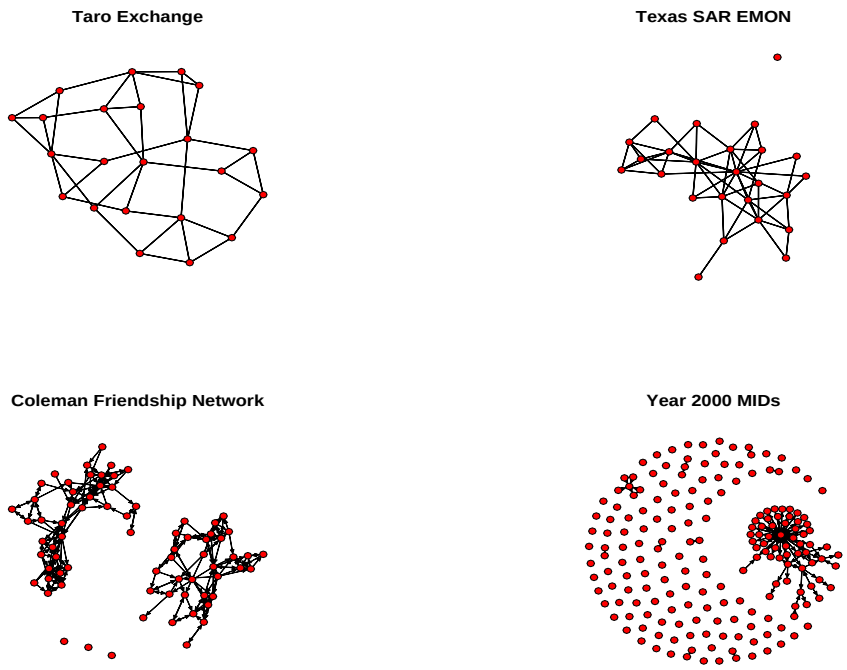

Figure 1: Sociograms for Sample Networks

	Taro Exchange	Texas EMON	Friendship	MIDs
2-Cycles	0	0	62	24
3-Cycles	10	40	88	4
4-Cycles	4	89	136	2
5-Cycles	7	226	202	1
6-Cycles	20	592	240	0
7-Cycles	48	1411	164	0
8-Cycles	94	3068	19	0
9-Cycles	152	6078	20	0
10-Cycles	247	11059	15	0
11-Cycles	430	18889	8	0
12-Cycles	697	30403	2	0

Table 1: Cycle Counts for Four Sample Networks

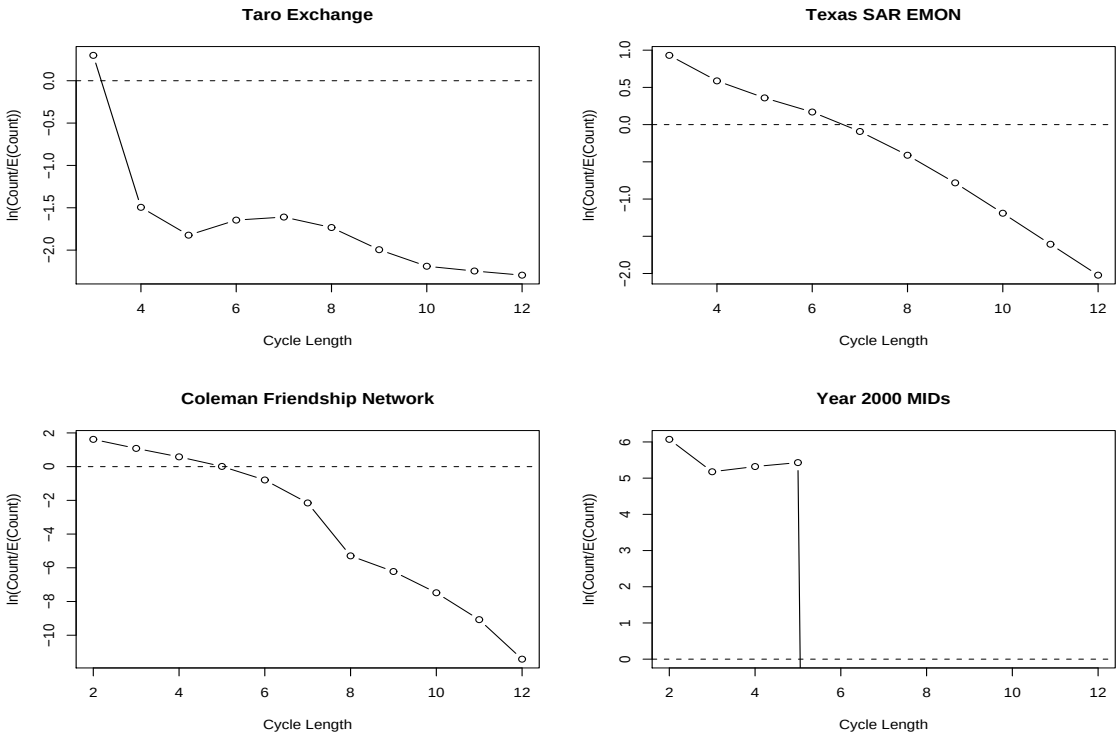

Figure 2: Logged Cycle Census Ratios (Observed versus Expected), by Cycle Length

Of course, even this yardstick does not take into account the dependence among cycles of varying order. For this, we must turn to the ERG models.

3.2 Cycle Census Models

As a major objective of this paper has been to elucidate the use of cycle statistics within exponential family models, it is useful to demonstrate the use of such models in action. Table 2 shows the results for MCMC-MLE fits of an incomplete cycle census model to each of the four data sets, using all cycles up to length 6.[4] In addition to the MLEs themselves, asymptotic

[4] 2-cycle terms are only meaningful for the directed graphs, and were thus not used for the undirected graphs. Further, the 6-cycle term was dropped for the MIDs data, due to the fact that the realized count was equal to 0 (and hence the MLE for this parameter

standard errors and associated p-values are also shown; note that these last should be used heuristically, since their usual normal-theory justifications have not been proven for ERGs. Null and residual deviance statistics are given for each model, which provide some sense of overall fit. In general, the cycle census models seem to fit quite well: they account for 76%-98% of the deviance in these networks, despite the small number of parameters involved. While these models were chosen for illustrative simplicity rather than fit, it is nonetheless useful to know that they provide a reasonable characterization of the data.

Table 2 suggests rather different underlying influences across the four networks. In the case of the taro exchange network, there appears to be a general tendency to suppress long cycles, particularly those of even length. This would be compatible with a mechanism such as in-group bias in trading (as opposed to an out-group bias, which would tend to suppress cycles of odd length), as well as a "nearest neighbor" phenomenon (which would suppress long cycles more generally). In the case of the Texas EMON, the strongest effect by far (other than density suppression) is the tendency towards the formation of 3-cycles. This is consistent with the high degree of observed triangulation in this graph, and may reflect a tendency for emergency response organizations to seek redundancy amongst their local contacts. For the friendship network, we see a strong tendency towards creation of short cycles, which steadily diminishes with cycle length. One suspects that (given the Bernoulli comparison plots) cycle effects of even longer length would begin to show significant negative effects, although the model fits well without these terms. Finally, the network of militarized international disputes reveals strong suppression of odd cycles, together with a positive effect for even cycles of short length. Given the negative valence of MID ties, one is tempted to interpret this in balance theoretic terms; to be sure, such a result is what would be expected from a signed balance model (see, e.g. Harary, 1959). While the model fits well, some skepticism is still in order: as per Table 1, many of the estimated effects here are based on very small numbers of realized cycles in an extremely sparse graph.

4 Conclusion

In this paper, we have explored cycle statistics as a device for capturing long-range dependence within social networks. As we demonstrated, cycle statistics arise naturally from the assumption that one edge depends

would not exist).

	Taro Exchange			Texas EMON			Friendship			MIDs		
	$\hat{\theta}$	s.e.	$\Pr(>\|Z\|)$	$\hat{\theta}$	s.e.	$\Pr(>\|Z\|)$	$\hat{\theta}$	s.e.	$\Pr(>\|Z\|)$	$\hat{\theta}$	s.e.	$\Pr(>\|Z\|)$
Edges	2.0526	1.4914	0.1687	−2.5933	0.4064	0.0000	−4.1778	0.0957	0.0000	−6.9336	0.3406	0.0000
Cycle2							1.5615	0.2082	0.0000	7.8360	2.4368	0.0013
Cycle3	1.1489	1.0175	0.2588	2.6117	0.9033	0.0038	0.7222	0.2092	0.0006	−3.0203	0.7638	0.0001
Cycle4	−2.1619	0.8713	0.0131	−0.7302	0.5911	0.2167	0.6866	0.1819	0.0002	43.3479	0.0188	0.0000
Cycle5	−0.0789	0.6297	0.9003	0.1765	0.2081	0.3964	0.1663	0.1062	0.1173	−1.9328	0.0029	0.0000
Cycle6	−0.4999	0.2772	0.0714	−0.0300	0.0316	0.3423	−0.0063	0.0334	0.8508			
Null Deviance			320.234			415.89			7286.4			50308.62
Residual Deviance			56.112			97.14			1384.4			988.48
Model df			5			5			6			5
Residual df			226			295			5256			36285

Table 2: Cycle Census Models for Four Sample Networks

upon another if that edge has the potential to create a secondary "conduit" between its endpoints. Where this conduit must be a directed path, the resulting model is fully parameterized by the cycle census, together with the edge count; weaker conditions imply a larger set of sufficient statistics, but all involve subgraphs with spanning cycles (and/or unions of non-disjoint cycles). Computation for cycle statistics can be performed using a recursive path-counting algorithm, an algorithm which can also be used (albeit inefficiently) to obtain the cycle census. Although this calculation is exponential time in the worst case, it is still possible to count reasonably long cycles (e.g., 10-12) for many data sets of substantive interest. This was illustrated with the application of cycle counts and associated ERG models to four sample data sets, taken from a range of substantive domains. While much remains to be learned about the behavior of cycle census models – and exponential random graph models more generally – it is hoped that the present work serves to advance our understanding in this area.

5 References

Bearman, P. (1997). Generalized Exchange. *American Journal of Sociology*, 102:1383–1415.

Besag, J. (1974). Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society, Series B*, 36(2):192–236.

Bollobás, B. (1998). *Modern Graph Theory*. Springer, New York.

Carrington, P. J., Scott, J., and Wasserman, S., editors (2005). *Models and Methods in Social Network Analysis*. Cambridge University Press, Cambridge.

Coleman, J. S. (1964). *Introduction to Mathematical Sociology*. Free Press, New York.

Crouch, B., Wasserman, S., and Trachtenburg, F. (1998). Markov Chain Monte Carlo Maximum Likelihood Estimation for $p*$ Social Network Models. Paper presented at the XVIII International Sunbelt Social Network Conference, Sitges, Spain.

Drabek, T. E., Tamminga, H. L., Kilijanek, T. S., and Adams, C. R. (1981). *Managing Multiorganizational Emergency Responses: Emergent Search and Rescue Networks in Natural Disaster and Remote Area Settings*. Number Monograph 33 in Program on Technology, Environment,

and Man. Institute of Behavioral Sciences, University of Colorado, Boulder, CO.

Erdös, P. and Rényi, A. (1960). On the evolution of random graphs. *Public Mathematical Institute of Hungary Academy of Sciences*, 5:17–61.

Frank, O. and Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association*, 81:832–842.

Ghosen, F., Palmer, G., and Bremer, S. (2004). The MID3 Data Set, 1993–2001: Procedures, Coding Rules, and Description. *Conflict Management and Peace Science*, 21:133–154.

Handcock, M. S., Hunter, D. R., Butts, C. T., and Morris, M. (2003). The statnet Library for R. Software library.

Harary, F. (1959). On the measurement of structural balance. *Behavioral Science*, 4:316–323.

Holland, P. W. and Leinhardt, S. (1976). Local Structure in Social Networks. In Heise, D., editor, *Sociological Methodology*, pages 1–45. Jossey-Bass, San Francisco.

Kendra, J. M. and Wachtendorf, T. (2003). Elements of Resilience After the World Trade Center Disaster: Reconstituting New York City's Emergency Operations Center. *Disasters*, 27(1):37–53.

Mayhew, B. H. and Levinger, R. L. (1976). Size and density of interaction in human aggregates. *American Journal of Sociology*, 82:86–110.

Pattison, P. and Robins, G. (2002). Neighborhood-Based Models for Social Networks. *Sociological Methodology*, 32:301–337.

Schwimmer, E. (1973). *Exchange in the Social Structure of the Orokaiva*. St. Martins, New York.

Snijders, T. A. B. (2001). SIENA: Simulation Investigation for Empirical Network Analysis. Software package.

Snijders, T. A. B. (2002). Markov Chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure*, 3(2).

Wasserman, S. and Robins, G. (2005). An introduction to random graphs, dependence graphs, and $p*$. In Carrington, P. J., Scott, J., and Wasserman, S., editors, *Models and Methods in Social Network Analysis*, chapter 10, pages 192–214. Cambridge University Press, Cambridge.

West, D. B. (1996). *Introduction to Graph Theory.* Prentice Hall, Upper Saddle River, NJ.

Willer, D., editor (1999). *Network Exchange Theory.* Praeger, Westport, CN.

Yamagishi, T. and Cook, K. (1993). Generalized Exchange and Social Dilemmas. *Social Psychology Quarterly*, 56:235–248.