8-2012

Consistency under Sampling of Exponential Random Graph Models

Cosma R. Shalizi
Carnegie Mellon University, cshalizi@andrew.cmu.edu

Alessandro Rinaldo
Carnegie Mellon University, arinaldo@stat.cmu.edu

Consistency under Sampling of Exponential Random Graph Models

Cosma Rohilla Shalizi*
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA, 15213-3890 USA

Alessandro Rinaldo†
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA, 15213-3890 USA

Abstract

The growing availability of network data and of scientific interest in distributed systems has led to the rapid development of statistical models of network structure. Typically, however, these are models for the entire network, while the data consists only of a sampled sub-network. Parameters for the whole network, which is what is of interest, are estimated by applying the model to the sub-network. This assumes that the model is *consistent under sampling*, or, in terms of the theory of stochastic processes, that it defines a projective family. Focussing on the popular class of exponential random graph models (ERGMs), we show that this apparently trivial condition is in fact violated by many popular and scientifically appealing models, and that satisfying it drastically limits ERGM's expressive power. These results are actually special cases of more general ones about exponential families of dependent random variables, which we also prove. Using such results, we offer easily checked conditions for the consistency of maximum likelihood estimation in ERGMs, and discuss some possible constructive responses.

Keywords: Exponential family, Projective family, Network models, Exponential random graph model, Sufficient statistics, Independent increments, Network sampling

1 Introduction

In recent years, the rapid increase in both the availability of data on networks (of all kinds, but especially social ones) and the demand, from many scientific areas, for analyzing such data has resulted in a surge of generative and descriptive models for network data (Easley and Kleinberg, 2010; Newman, 2010). Within statistics, this trend has led to a renewed interest in developing, analyzing and validating statistical models for networks (Goldenberg *et al.*, 2009; Kolaczyk, 2009). Yet as networks are a non-standard type of data, many basic properties of statistical models for networks are still unknown or have not been properly explored.

In this article we investigate the conditions under which statistical inferences drawn over a sub-network will generalize to the entire network. It is quite rare for the data to ever actually be the *whole* network of relations among a given set of nodes or units[1]; typically, only a sub-network is available. Guided by experience of more conventional problems like regression, analysts have generally fit models to the available sub-network, and then extrapolated them to the larger true network which is of actual scientific interest, presuming that the models are, as it were, consistent under sampling. What we show is that this is only valid

*Email: cshalizi@cmu.edu
†Email: arinaldo@cmu.edu

[1]This sense of the "whole network" should not be confused with the technical term "complete graph", where every vertex has a direct edge to every other vertex.

for very special model specifications, and the specifications where it is *not* valid include some of which are currently among the most popular and scientifically appealing.

In particular, we restrict ourselves to exponential random graph models (ERGMs), undoubtedly one of the most important and popular classes of statistical models of network structure. In addition to the general works already cited, the reader is referred to Frank and Strauss (1986); Wasserman and Pattison (1996); Anderson *et al.* (1999); Snijders *et al.* (2006); Robins *et al.* (2007); Wasserman and Robins (2005); Handcock *et al.* (2008); Park and Newman (2004b), for detailed accounts of these models. There are many reasons ERGMs are so prominent. On the one hand, ERGMs, as the name suggests, are exponential families, and so they inherit all the familiar virtues of exponential families in general: they are analytically and inferentially convenient (Brown, 1986); they naturally arise from considerations of maximum entropy (Mandelbrot, 1962) and minimum description length (Grünwald, 2007), and from physically-motivated large deviations principles (Touchette, 2009); and if a generative model obeys reasonable-seeming regularity conditions while still having a finite-dimensional sufficient statistic, it must be an exponential family (Lauritzen, 1988)[2]. On the other hand, ERGMs have particular virtues as models of networks. The sufficient statistics in these models typically count the number or density of certain "motifs" or small sub-graphs, such as edges themselves, triangles, k-cliques, stars, etc. These in turn are plausibly related to different network-growth mechanisms, giving them a substantive interpretation. (See, e.g., Goodreau *et al.* (2009) as an exemplary application of this idea, or, more briefly, §5 below.) Moreover, the important task of edge prediction is easily handled in this framework, reducing to a conditional logistic regression (Handcock *et al.*, 2008). Since the development of (comparatively) computationally-efficient maximum-likelihood estimators (based on Monte Carlo sampling), ERGMs have emerged as flexible and persuasive tools for modeling network data (Handcock *et al.*, 2008).

Despite all these strengths, however, ERGMs are tools with a serious weakness. As we mentioned, it is very rare to ever observe the whole network of interest. The usual procedure, then, is to fit ERGMs (by maximum likelihood or pseudo-likelihood) to the observed sub-network, and then extrapolate the same model, with the same parameters, to the whole network; often this takes the form of interpreting the parameters as "provid[ing] information about the presence of structural effects observed in the network" (Robins *et al.*, 2007, p. 194), or the strength of different network-formation mechanisms. (Ackland and O'Neil 2011; Daraganova *et al.* 2012; de la Haye *et al.* 2010; Gondal 2011; Gonzalez-Bailon 2009; Schaefer 2012; Vermeij *et al.* 2009 are just a few of the more recent papers doing this.) This obviously raises the question of the statistical (i.e., large sample) consistency of maximum likelihood estimation in this context. Unnoticed, however, is the logically prior question of whether it is *probabilistically* consistent to apply the same ERGM, with the same parameters, both to the whole network and its sub-networks. That is, whether the marginal distribution of a sub-network will be consistent with the distribution of the whole network, for all possible values of the model parameters. The same question arises when parameters are compared across networks of different sizes (as in, e.g., Faust and Skvoretz 2002; Goodreau *et al.* 2009; Lubbers and Snijders 2007). When this form of consistency fails, then the parameter estimates obtained from a sub-network may not provide reliable estimates of, or may not even be relatable to, the parameters of the whole network, rendering the task of statistical inference based on a sub-network ill-posed. We formalize this question using the notion of "projective families" from the theory of stochastic processes. We say that a model is **projective** when the same parameters can be used for both the whole network and any of its sub-networks. In this article, we fully characterize projectibility of discrete exponential families and, as corollary, show that ERGMs are projective only for very special choices of the sufficient statistic.

Outline Our results are not specific just to networks, but pertain more generally with exponential families of stochastic processes. §2 therefore lays out the necessary background about projective families of distributions, projective parameters, and exponential families in a somewhat more abstract setting than that of networks. §3 shows that a necessary and sufficient condition for an exponential family to be projective is that the sufficient statistics obey a kind of additive decomposition. This in turn implies strong independence

[2] Mandelbrot (1962) is still one of the best discussions of the interplay between the formal, statistical and substantive motivations for using exponential families.

properties. We also prove results about the consistency of maximum likelihood parameter estimation under these conditions (§4). In §5, we apply these results to ERGMs, showing that most popular specifications for social networks and other stochastic graphs cannot be projective. We then conclude with some discussion on possible constructive responses. The proofs are contained in §**??**.

Related work An early recognition of the fact that sub-networks may have statistical properties which differ radically from those of the whole network came in the context of studying networks with power-law ("scale-free") degree distributions. On the one hand, Stumpf *et al.* (2005) showed that "subnets of scale-free networks are not scale-free"; on the other, Achlioptas *et al.* (2005) demonstrated that a particular, highly popular sampling scheme creates the appearance of a power-law degree distribution on nearly any network. While the importance of network sampling schemes has been recognized since then (Kolaczyk, 2009, ch. 5), and valuable contributions have come from, e.g., Kossinets (2006); Handcock and Gile (2010); Krivitsky *et al.* (2011); Ahmed *et al.* (2010), we are not aware of any work which has addressed the specific issue of consistency under projection which we tackle here. Perhaps the closest approaches to our perspective are Orbanz (2011) and Xiang and Neville (2011). The former considers conditions under which infinite-dimensional families of distributions on abstract spaces have projective limits. The latter, more concretely, addresses the consistency of maximum likelihood estimators for exponential families of dependent variables, but under assumptions (regarding Markov properties, the "shape" of neighborhoods, and decay of correlations in potential functions) which are basically incomparable in strength to ours.

2 Projective Statistical Models and Exponential Families

Our results about exponential random graph models are actually special cases of more general results about exponential families of dependent random variables, and are just as easy to state and prove in the general context as for graphs. Setting this up, however, requires some preliminary definitions and notation, which make precise the idea of "seeing more data from the same source". In order to dispense ourselves from any measurability issues we will implicitly assume the existence of an underlying probability measure for which the random variables under study are all measurable. Furthermore, for the sake of readability we will not rely on the measure theoretic notion of filtration: though technically appropriate, it will add nothing to our results.

Let $\mathcal{A}$ be a collection of finite subsets of a denumerable set $\mathbb{I}$ partially ordered with respect to subset inclusion. For technical reasons, we will further assume that $\mathcal{A}$ has the property of being an ideal: i.e. if A belongs to $\mathcal{A}$ then all subsets of A are also in $\mathcal{A}$ and if A and B belongs to $\mathcal{A}$, then so does their union. We may think of passing from A to $B \supset A$ as taking increasingly large samples from a population, or recording increasingly long time series, or mapping data from increasing large spatial regions, or over an increasingly dense spatial grid, or looking at larger and larger sub-graphs from a single network. Accordingly, we consider the associated collection of parametric statistical models $\{\mathcal{P}_{A,\Theta}\}_{A\in\mathcal{A}}$ indexed by $\mathcal{A}$, where, for each $A \in \mathcal{A}$, $\mathcal{P}_{A,\Theta} \equiv \{\mathbb{P}_{A,\theta}\}_{\theta\in\Theta}$ is a family of probability distributions indexed by points θ in a fixed open set $\Theta \subseteq \mathbb{R}^d$. The probability distributions in $\mathcal{P}_{A,\Theta}$ are also assumed to be supported over the same $\mathcal{X}_A$, which are countable[3] sets for each A. We assume that the partial order of $\mathcal{A}$ is isomorphic to the partial order over $\{\mathcal{X}_A\}_{A\in\mathcal{A}}$, in the sense that $A \subset B$ if and only if $\mathcal{X}_B = \mathcal{X}_A \times \mathcal{X}_{B\setminus A}$.

For given θ and A, we denote with X_A the random variable distributed as $\mathbb{P}_{A,\theta}$. In particular, for a given $\theta \in \Theta$, we can regard the $\{\mathbb{P}_{A,\theta}\}_{A\in\mathcal{A}}$ as finite dimensional (i.e. marginal) distributions.

For each pair A, B in $\mathcal{A}$ with $A \subset B$, we let $\pi_{B\mapsto A}\colon \mathcal{X}_B \to \mathcal{X}_A$ be the natural index projection given by $\pi_{B\mapsto A}(x_A, x_{B\setminus A}) = x_A$. In the context of networks, we may think of $\mathbb{I}$ as the set of nodes of a possibly infinite random graph, which without loss of generality can be taken to be $\{1, 2, \ldots\}$ and of $\mathcal{A}$ as the collection of all finite subsets of $\mathbb{I}$. Then, for some positive integers n and m, we may, for instance, take $A = \{1, \ldots, n\}$ and $B = \{1, \ldots, n, \ldots, n+m\}$, so that X_A will be the induced sub-graph on the first n nodes and X_B the

[3] Our results extend to continuous observations straightforwardly, but with annoying notational overhead.

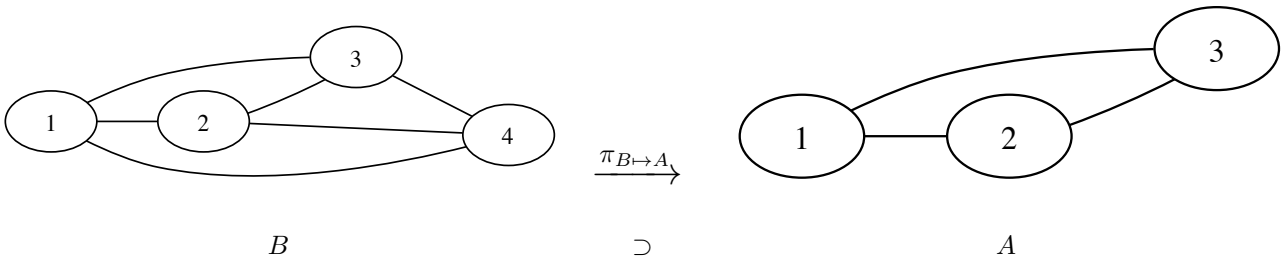

Figure 1: Projective structure for networks: when the set of observables A is contained in the larger set of observables B, X_A (on the right) can be recovered from X_B (on the left) through the projection $\pi_{B\mapsto A}$, which simply drops the extra data.

induced sub-graph on the first $n+m$ nodes. The projection $\pi_{B\mapsto A}$ then just picks out the appropriate sub-graph from the larger graph (see Figure 1 for a schematic example). We will be concerned with a natural form of probabilistic consistency of the collection $\{\mathcal{P}_{A,\Theta}\}_{A\in\mathcal{A}}$ which we call *projectibility*, defined below.

Definition 2.1. The family $\{\mathcal{P}_{A,\Theta}\}_{A\in\mathcal{A}}$ is **projective** if, for any A and B in $\mathcal{A}$ with $A \subset B$,

$$\mathbb{P}_{A,\theta} = \mathbb{P}_{B,\theta} \circ \pi^{-1}_{B\mapsto A}, \quad \forall\theta \in \Theta. \tag{1}$$

See Kallenberg (2002, p. 115) for more general treatment of projectibility. In words, $\{\mathcal{P}_{A,\Theta}\}_{A\in\mathcal{A}}$ is a projective family when $A \subset B$ implies that $\mathbb{P}_{A,\theta}$ can be recovered by marginalization over $\mathbb{P}_{B,\theta}$, for all θ. (Figure 2 illustrates.) Within a projective family, $\mathbb{P}_\theta$ denotes the infinite-dimensional distribution, which thus exists by the Kolmogorov extension theorem (Kallenberg, 2002, Thm. 6.16, p. 115).

Projectibility is automatic when the generative model calls for independent and identically distributed (IID) observations. It is also generally unproblematic when the model is specified in terms of *conditional* distributions: one then just uses the Ionescu Tulcea extension theorem in place of that of Kolmogorov (Kallenberg, 2002, Thm. 6.17, p. 116). However, many models are specified in terms of *joint* distributions for various index sets, and this, as we show in Theorem 3.2, can rule out projectibility.

We restrict ourselves to **exponential family** models by assuming that, for each choice of $\theta \in \Theta$ and $A \in \mathcal{A}$, $\mathbb{P}_{A,\theta}$ has density with respect to the counting measure over $\mathcal{X}_A$ given by

$$p_{A,\theta}(x) = \frac{e^{\langle\theta, t_A(x)\rangle}}{z_A(\theta)}, \quad x \in \mathcal{X}_A, \tag{2}$$

where $t_A \colon \mathcal{X}_A \to \mathbb{R}^d$ is the measurable function of minimal sufficient statistics, and $z_A \colon \Theta \to \mathbb{R}$ is the **partition function** given by

$$z_A(\theta) \equiv \sum_{x\in\mathcal{X}_A} e^{\langle\theta, t_A(x)\rangle}. \tag{3}$$

If $X_A \sim \mathbb{P}_{A,\theta}$, we will write $T_A \equiv t_A(X_A)$ for the random variable corresponding to the sufficient statistic. Equation (2) implies that T_A itself has an exponential family distribution, with the same parameter θ and partition function $z_A(\theta)$ (Brown, 1986, Prop. 1.5). Specifically, the distribution function is

$$\mathbb{P}_{A,\theta}(T_A = t) = \frac{e^{\langle\theta,t\rangle} v_A(t)}{z_A(\theta)} \tag{4}$$

where the term $v_A(t) \equiv |\{x \in \mathcal{X}_A \colon t_A(x) = t\}|$, which we will call the *volume factor*, counts the number of points in $\mathcal{X}_A$ with the same sufficient statistics t. The moment generating function of T_A is

$$M_{\theta,A}(\phi) = \mathbf{E}_\theta\left[e^{\langle\phi,T_A\rangle}\right] = z_A(\theta+\phi)/z_A(\theta). \tag{5}$$

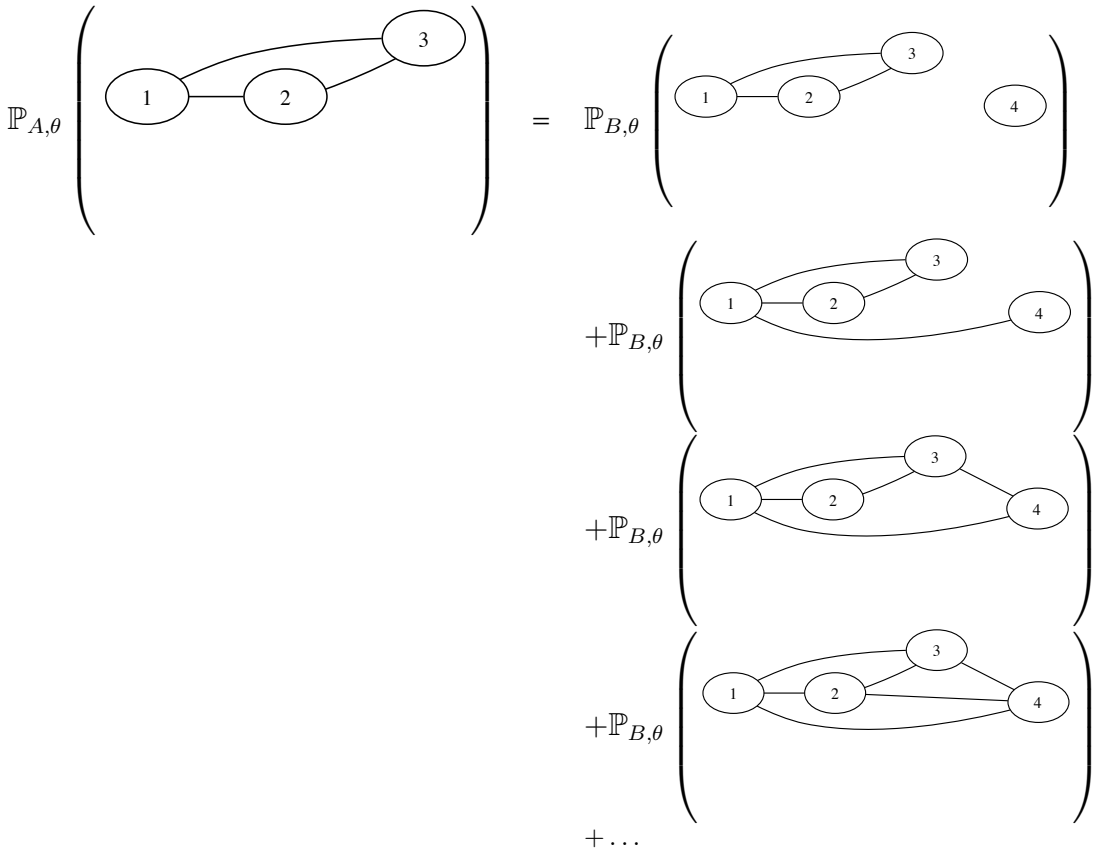

Figure 2: Illustration of projectibility: the probability of a small configuration over A (left), calculated according to $\mathbb{P}_{A,\theta}$, must match the sum of all larger configurations over B containing it (right), calculated according to $\mathbb{P}_{B,\theta}$.

If the sufficient statistic is completely additive, i.e., if $t_A(x_A) = \sum_{i \in A} t_{\{i\}}(x_i)$, then this is a model of independent (if not necessarily IID) data. In general, however, the choice of sufficient statistics may impose, or capture, dependence between observations.

Because we are considering exponential families defined on increasingly large sets of observations, it is convenient to introduce some notation related to multiple statistics. Fix $A, B \in \mathcal{A}$ such that $A \subset B$. Then $t_B : \mathcal{X}_B \mapsto \mathbb{R}^d$, and we will sometimes write this function $t(x, y)$, where the first argument is in $\mathcal{X}_A$ and the second in $\mathcal{X}_{B \setminus A}$. We will have frequent recourse to the increment to the sufficient statistic, $t_{B \setminus A}(x, y) \equiv t_B(x, y) - t_A(x)$. The volume factor $v_B(t_B(x_B))$ is defined as before, but we shall also consider, for each observable value t of the sufficient statistics for A and increment δ of the sufficient statistics from A to B, the *joint volume factor,*

$$v_{A,B \setminus A}(t, \delta) \equiv \left| \{(x, y) \in \mathcal{X}_B \colon t_A(x) = t \text{ and } t_{B \setminus A}(x, y) = \delta\} \right|, \tag{6}$$

and the *conditional volume factor,*

$$v_{B \setminus A | A}(\delta, x) \equiv \left| \{y \in \mathcal{X}_{B \setminus A} \colon t_{B \setminus A}(x, y) = \delta\} \right|. \tag{7}$$

As we will see, these volume factors play a key role in characterizing projectibility.

3 Projective Structure in Exponential Families

In this section we characterize projectibility in terms of the increments of the vector of sufficient statistics. In particular we show that exponential families are projective if, and only if, their sufficient statistics decompose into separate additive contributions from disjoint observations in a particularly nice way which we formalize in the following definition.

Definition 3.1. The sufficient statistics of the family $\{\mathcal{P}_{A,\Theta}\}_{A \in \mathcal{A}}$ have **separable increments** when, for each $A \subset B$, $x \in \mathcal{X}_A$, the range of possible increments δ is the same for all x, and the conditional volume factor is constant in x, i.e. $v_{B \setminus A | A}(\delta, x) = v_{B \setminus A}(\delta)$.

It is worth noting that the property of having separable increments is an intrinsic property of the family $\{\mathcal{P}_{A,\Theta}\}_{A \in \mathcal{A}}$ that depends only on the functional forms of the sufficient statistics $\{t_A\}_{A \in \mathcal{A}}$ and not on the model parameters $\theta \in \Theta$. This follows from the fact that, for any A, the probability distributions $\{\mathbb{P}_{A,\theta}\}_{\theta \in \Theta}$ have identical support $\mathcal{X}_A$. Thus, this property holds for all of θ or none of them.

The main result of this paper is then as follows.

Theorem 3.2. *The exponential family $\{\mathcal{P}_{A,\Theta}\}_{A \in \mathcal{A}}$ is projective if and only if the sufficient statistics $\{T_A\}_{A \in \mathcal{A}}$ have separable increments.*

3.1 Independence Properties

Because projectibility implies separable increments, it also carries statistical-independence inmplications. Specifically, it implies that the increments to the sufficient statistics are statistically independent, and that $X_{B \setminus A}$ and X_A are conditionally independent given increments to the sufficient statistic. Interestingly, independent increments for the statistic are necessary but not quite sufficient for projectibility. These claims are all made more specific in the propositions which follow.

We first show that projectibility implies that the sufficient statistics have independent increments. In fact, a stronger results holds, namely that the increments of the sufficient statistics are independent of the actual sequence. Below we will write $T_{B \setminus A}$ to signify $T_B - T_A$.

Proposition 3.3. *If the exponential family $\{\mathcal{P}_{A,\Theta}\}_{A \in \mathcal{A}}$ is projective, then sufficient statistics $\{T_A\}_{A \in \mathcal{A}}$ have independent increments, i.e. $A \subset B$ implies that $T_B - T_A \perp\!\!\!\perp T_A$ under all θ.*

Proposition 3.4. *In a projective exponential family, $T_{B \setminus A} \perp\!\!\!\perp X_A$.*

We note that independent increments for the sufficient statistics T_A in no way implies independence of the actual observations X_A. As a simple illustration, take the one-dimensional Ising model[4], where $\mathbb{I} = \mathbb{N}$, each $\mathcal{X}_i = \pm 1$, $\mathcal{A}$ consists of all intervals from 1 to n, and the single sufficient statistic $T_{1:n} = \sum_{i=1}^{n-1} X_i X_{i+1}$. Clearly, $T_{1:(n+1)} - T_{1:n} = +1$ when $X_n = X_{n+1}$, otherwise $T_{1:(n+1)} - T_{1:n} = -1$. Since $v_{1:(n+1)|1:n}(+1, x) = v_{1:(n+1)|1:n}(-1, x) = 1$, by Theorem 3.2, the model is projective. By Proposition 3.3, then, increments of T should be independent, and direct calculation shows the probability of increasing the sufficient statistic by 1 is $e^\theta/(1+e^\theta)$, no matter what $X_1, \ldots X_n$ are. While the sufficient statistic has independent increments, the random variables X_i are all dependent on one another.[5]

The previous results provide a way, and often a simple one, for checking whether projectibility fails: if the sufficient statistics do not have independent increments, then the family is not projective. As we will see, this test covers many statistical models for networks.

It is natural to inquire into the converse to these propositions. It is fairly straightforward (if somewhat lengthy) to show that independent increments for the sufficient statistics implies that the joint volume factor separates.

Proposition 3.5. *If an exponential family has independent increments, $T_{B\setminus A} \perp\!\!\!\perp T_A$, then its joint volume factor separates, $v_{A,B\setminus A}(t, \delta) = v_A(t) v_{B\setminus A}(\delta)$, and the distribution of T is projective.*

However, independent increments for the sufficient statistics do *not* imply that separable increments (hence projectibility), as shown by the next counter-example. Hence independent increments are a necessary but not sufficient condition for projectibility.

Suppose that $\mathcal{X}_A = \{a, b, c, d\}$, and $\mathcal{X}_{B\setminus A} = \{i, ii, iii, iv, v\}$. (Thus there are 20 possible values for X_B.) Let

$$\begin{aligned} +1 &= t_A(a) = t_A(b) \\ -1 &= t_A(c) = t_A(d) \end{aligned}$$

so that $v_A(+1) = v_A(-1) = 2$. Further, let

$$\begin{aligned} 2 &= t_B(a, i) = t_B(a, ii) \\ 0 &= t_B(a, iii) = t_B(a, iv) = t_B(a, v) \\ 0 &= t_B(b, i) = t_B(b, ii) \\ 2 &= t_B(b, iii) = t_B(b, iv) = t_B(b, v) \\ t_B(c, y) &= t_B(a, y) - 2 \\ t_B(d, y) &= t_B(b, y) - 2 \end{aligned}$$

It is not hard to verify that $T_{B\setminus A}$ is always either $+1$ or -1. It is also straightforward to check that $v_{A,B\setminus A}(t, \delta) = 5$ for all combinations of t and δ, implying that $v_{B\setminus A}(+1) = v_{B\setminus A}(-1) = 2.5$, and that the joint volume factor separates. On the other hand, the *conditional* volume factors are not constant in x, as $v_{B\setminus A|A}(+1, a) = 2$ while $v_{B\setminus A|A}(+1, b) = 3$. Thus, the sufficient statistic has independent increments, but does not have separable increments. Since projective families have separable increments (Proposition 7.1), this cannot be a projective family. (This can also be checked by a direct and straightforward, if even more tedious, calculation.)

We conclude with section with a final observation. Butler (1986) showed that when observations follow from an IID model with a minimal sufficient statistic, the predictive distribution for the next observation can be written entirely in terms of how different hypothetical values would change the sufficient statistic. (Cf. Lauritzen 1974; Besag 1989.) This predictive sufficiency property carries over to our setting.

[4] Technically, with "free" boundary conditions; see Landau and Lifshitz (1980).

[5] Note that while this is a *graphical* model, it is not a model of a random graph. (The graph is rather the one-dimensional lattice.) Rather, it is used here merely to exemplify the general result about exponential families. We turn to exponential random graph models in §5.

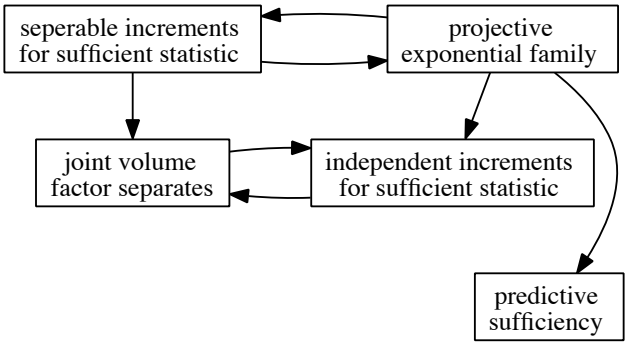

Figure 3: Relations among the main properties of models considered in §3. Probabilistic properties of the models are on the right, algebraic/combinatorial properties of the sufficient statistic are on the left.

Theorem 3.6 (Predictive Sufficiency)**.** *In a projective exponential family, the distribution of $X_{B\setminus A}$ conditional on X_A depends on the data only through $T_{B\setminus A}$.*

The main implications among our results are summarized in Figure 3.

3.2 Remarks, Applications and Extensions

Exponential families of time series As the example of the Ising model in §3.1 (p. 7) makes clear, our theorem applies whenever we need an exponential family to be projective, not just when the data are networks. In particular, they apply to exponential families of time series, where $\mathbb{I}$ is the natural or real number line (or perhaps just its positive part), and the elements of $\mathcal{A}$ are intervals. An exponential family of stochastic processes on such a space has projective parameters if, and only if, its sufficient statistics have separable increments, and so only if they have independent increments.

Transformation of parameters Allowing the dimension of θ to be fixed, but for its components to change along with A, does not really get out of these results. Specifically, if θ is to be re-scaled in a way that is a function of A alone, we can recover the case of a fixed θ by "moving the scaling across the inner product", i.e., by re-defining T_A to incorporate the scaling. With a sample-invariant θ, it is this transformed T which must have separable increments. Other transformations can either be dealt with similarly, or amount to using a non-uniform base measure (see below).

Statistical-mechanical interpretation It is interesting to consider the interpretation of our theorem, and of its proof, in terms of statistical mechanics. As is well-known, the "canonical" distributions in statistical mechanics are exponential families (Boltzmann-Gibbs distributions), where the sufficient statistics are "extensive" physical observables, such as energy, volume, the number of molecules of various species, etc., and the natural parameters are the corresponding conjugate "intensive" variables, such as, respectively, (inverse) temperature, pressure, chemical potential, etc. (Landau and Lifshitz, 1980; Mandelbrot, 1962). Equilibrium between two systems which interact by exchanging the variables tracked by the extensive variables obtains if and only if they have the same values of the intensive parameters (Landau and Lifshitz, 1980). In our terms, of course, this is simply projectibility, the requirement that the same parameters hold for all sub-systems.

What we have shown is that for this to be true, the increments to the extensive variables must be completely unpredictable from their values on the sub-system.

Furthermore, notice the important role played in both halves of the proof by the separation of the joint volume factor, $v_{A,B\setminus A}(t,\delta) = v_A(t)v_{B\setminus A}(\delta)$. In terms of statistical mechanics, a macroscopic state is a collection of microscopic configurations with the same value of one or more macroscopic observables. The Boltzmann entropy of a macroscopic state is (proportional to) the logarithm of the volume of those microscopic states (Landau and Lifshitz, 1980). If we define our macroscopic states through the sufficient statistics, then their Boltzmann entropy is just $\log v$. Thus, the separation of the volume factor is the same as the additivity of the entropy across different parts of the system, i.e., the entropy is "extensive". Our results may thus be relevant to debates in statistical mechanics about the appropriateness of alternative, non-extensive entropies (cf. Nauenberg 2003).

Beyond exponential families It is not clear just how important it is that we have an exponential family, as opposed to a family admitting a finite-dimensional sufficient statistic. As is well-known, the two concepts coincide under some regularity conditions (Barndorff-Nielsen, 1978), but not quite strictly, and it would be interesting to know whether or not the exponential form of equation (2) is strictly required. We have attempted to write the proofs in a way which minimizes the use of this form (in favor of the Neyman factorization, which only uses sufficiency), but have not succeeded in eliminating it completely. We return to this matter in the conclusion.

Prediction We have focused on the implications of projectibility for parametric inference. Exponential families are however often used in statistics and machine learning as generative models in applications where the only goal is prediction (Wainwright and Jordan, 2008), and so (to quote Butler 1986) "all parameters are nuisance parameters". But even in then, it must be possible to consistently extend the generative model's distribution for the training data to a joint distribution for training and testing data, with a single set of parameters shared by both old and new data. While this requirement may seem too trivial to mention, it is, precisely, projectibility.

Growing Number of Parameters In the proof of Theorem 3.2, we used the fact that T_A, and hence θ, has the same dimension for all $A \in \mathcal{A}$. There are, however, important classes of models where the number of parameters is allowed to grow with the size of the sample. Particularly important, for networks, are models where each node is allowed a parameter (or two) of its own, such as its expected degree — see for instance the classic p_1 model of Holland and Leinhardt (1981), or the "degree-corrected block models" of Karrer and Newman (2011). We can formally extend Theorem 3.2 to cover some of these cases — including those two particular specifications — as follows.

Assume that T_A has a dimension which is strictly non-decreasing as A grows, i.e., $d_A \leq d_B$ whenever $A \subset B$. Furthermore, assume that the set of parameters θ_A only grows, and that the meaning of the old parameters is not disturbed. That is, under projectibility we should have

$$\mathbb{P}_{B,\theta_B} \cdot \pi^{-1}_{B \mapsto A} = \mathbb{P}_{A,\pi_{\mathbb{R}^{d_B} \mapsto \mathbb{R}^{d_A}}\theta_B}(\cdot) \tag{8}$$

For any fixed pair $A \subset B$, we can accommodate this within the proof of Theorem 3.2 by re-defining T_A to be a mapping from $\mathcal{X}_A$ to $\mathbb{R}^{d_B}$, where the extra $d_B - d_A$ components of the vector are always zero. The extra parameters in θ_B then have no influence on the distribution of X_A and are unidentified on A, but we have, formally, restored the fixed-parameter case. The "increments" of the extra components of T_B are then simply their values on X_B, and, by the theorem, the range of values for these statistics, and the number of configurations on $\mathcal{X}_{B\setminus A}$ leading to each value, must be equal for all $x \in \mathcal{X}_A$.

Adapting our conditions for the asymptotic convergence of maximum likelihood estimators (§4) to the growing-parameter setting is beyond our scope here.

Non-uniform base measures If the exponential densities in (2) are defined with respect to non-uniform base measures different from the counting measures, the sufficient statistics need not have separable increments. In Appendix A we address this issue and describe the modifications and additional assumptions required for our analysis to remain valid. We thank an anonymous referee and Pavel Krivitsky for independently brining up this subtle point to our attention.

4 Consistency of Maximum Likelihood Estimators

Statistical inference in an exponential family naturally centers on the parameter θ. As is well known, the maximum likelihood estimator $\widehat{\theta}$ takes a particularly simple form, obtainable using the fact (which follows from equation (5)) that $\nabla_\theta z_A(\theta) = z_A(\theta)\mathbf{E}_\theta[T_A]$:

$$\begin{aligned} 0 = \nabla_\theta \frac{e^{\langle\theta, t_A(x)\rangle}}{z_A(\theta)}\Big|_{\theta=\widehat{\theta}} &= \frac{-z_A(\widehat{\theta})t_A(x)e^{\langle\widehat{\theta}, t_A(x_A)\rangle} + e^{\langle\widehat{\theta}, t_A(x)\rangle}z_A(\widehat{\theta})\mathbf{E}_{\widehat{\theta}}[T_A]}{z_A^2(\theta)} \\ t_A(x) &= \mathbf{E}_{\widehat{\theta}}[T_A] \end{aligned} \tag{9}$$

In words, the most likely value of the parameter is the one where the expected value of the sufficient statistic equals the observed value.

Assume the conditions of Theorem 3.2 hold, so that the parameters are projective and the sufficient statistics have (by Lemma 7.3) independent increments. Define the logarithm of the partition function $a_A(\theta) \equiv \log z_A(\theta)$.[6]. Suppose that

$$a_A(\theta) = r_{|A|}a(\theta) \tag{10}$$

where $|A|$ is some positive-valued measure of the size of A, $r_{|A|}$ a positive monotone-increasing function of it, and $a\colon \Theta \mapsto \mathbb{R}$ is differentiable (at least at θ). Then, by equation (5) for the moment generating function, the cumulant generating function of T_A is

$$\kappa_{A,\theta}(\phi) = r_{|A|}(a(\theta+\phi) - a(\theta)) \tag{11}$$

From the basic properties of cumulant generating functions, we have

$$\mathbf{E}_\theta[T_A] = \nabla_\phi \kappa_{A,\theta}(0) = r_{|A|}\nabla a(\theta) \tag{12}$$

Substituting into equation (9),

$$\frac{t_A(x)}{r_{|A|}} = \nabla a(\widehat{\theta}) \tag{13}$$

Thus to control the convergence of $\widehat{\theta}$, we must control the convergence of $T_A/r_{|A|}$.

Consider a growing sequence of sets A such that $r_{|A|} \to \infty$. Since T_A has independent increments, and the cumulant generating functions for different A are all proportional to each other, we may regard T_A as a time-transformation of a Lévy process Y_r (Kallenberg, 2002). That is, there is a continuous-time stochastic process Y with IID increments, such that Y_1 has cumulant generating function $a(\theta+\phi)-a(\theta)$, and $T_A = Y_{r_{|A|}}$. Note that T_A itself does not have to have IID increments, but rather the distribution of the increment $T_B - T_A$ must only depend on $r_{|B|} - r_{|A|}$. Specifically, from lemma 7.6 and equation (10), the cumulant generating function of the increment must be $(r_{|B|} - r_{|A|})[a(\theta+\phi) - a(\theta)]$. The scaling factor homogenizes (so to speak) the increments of T.

Writing the sufficient statistic as a transformed Lévy process yields a simple proof that $\widehat{\theta}$ is strongly (i.e., almost-surely) consistent. Since a Lévy process has IID increments, by the strong law of large numbers $Y_{r_{|A|}}/r_{|A|}$ converges almost surely ($\mathbb{P}_\theta$) to $\mathbf{E}_\theta[Y_1]$ (Kallenberg, 2002). Since $T_A = Y_{r_{|A|}}$, it follows that $T_A/r_{|A|} \to \mathbf{E}_\theta[Y_1]$ a.s. ($\mathbb{P}_\theta$) as well; but this limit is $\nabla a(\theta)$. Thus the MLE converges on θ almost surely. We have thus proved

[6] In statistical mechanics, $-a_A$ would be the Helmholtz free energy.

Theorem 4.1. *Suppose that the model $\mathbb{P}_\theta$ is projective, and that the log partition function obeys equation (10) for each $A \in \mathcal{A}$. Then the maximum likelihood estimator exists and is strongly consistent.*

We may extend this in a number of ways. First, if the scaling relation equation (10) holds for a particular θ (or set of θ), then $T_A/r_{|A|}$ will converge almost surely for that θ. Thus, strong consistency of the MLE may in fact hold over certain parameter regions but not others. Second, when $d > 1$, all components of T_A must be scaled by the *same* factor $r_{|A|}$. Making the expectation value of one component of T be $O(|A|)$ while another was $O(|A|^3)$ (for instance) would violate equation (12) and so equation (10) as well.

Finally, while the exact scaling of equation (10), together with the independence of the increments, leads to strong consistency of the MLE, ordinary consistency (convergence in probability) holds under weaker conditions. Specifically, suppose that log partition function or free energy scales in the limit as the size of the assemblage grows,

$$\lim_{r_{|A|}\to\infty} a_A(\theta)/r_{|A|} = a(\theta) \tag{14}$$

(We give examples towards the end of §5 below.) We may then use the following theorem:

Theorem 4.2. *Suppose that an exponential family shows approximate scaling, i.e., equation (14) holds, for some θ. Then, for any measurable set $K \subseteq \mathbb{R}^d$,*

$$\liminf_{r_{|A|}\to\infty} \frac{1}{r_{|A|}} \log \mathbb{P}_{A,\theta}\left(\frac{T_A}{r_{|A|}} \in K\right) \geq -\inf_{t\in \mathrm{int}K} J(t) \tag{15}$$

$$\limsup_{r_{|A|}\to\infty} \frac{1}{r_{|A|}} \log \mathbb{P}_{A,\theta}\left(\frac{T_A}{r_{|A|}} \in K\right) \leq -\inf_{t\in \mathrm{cl}K} J(t) \tag{16}$$

where

$$J(t) = \sup_{\phi\in\mathbb{R}^d} \langle \phi, t\rangle - [a(\theta+\phi) - a(\theta)] \,, \tag{17}$$

and $\mathrm{int}K$ and $\mathrm{cl}K$ are respectively the interior and the closure of K.

When the limits in equations (15) and (16) coincide, which they will for most nice sets K, we may say that

$$\frac{1}{r_{|A|}} \log \mathbb{P}_{A,\theta}\left(\frac{T_A}{r_{|A|}} \in K\right) \to -\inf_{t\in K} J(t) \tag{18}$$

Since $J(t)$ is minimized at 0 when $t = \nabla a(\theta)$[7], equation (18) holds in particular for any neighborhood of $\nabla a(\theta)$, and for the complement of such neighborhoods, where the infimum of J is strictly positive. Thus $T_A/r_{|A|}$ converges in probability to $\nabla a(\theta)$, and $\widehat{\theta} \xrightarrow{P} \theta$, for all θ where equaiton (14) holds.

Heuristically, when equation (14) holds but equation (10) fails, we may imagine approximating the actual collection of dependent and heterogeneous random variables with an average of IID, homogenized effective variables, altering the behavior of the global sufficient statistic T by no more than $o_P(r_{|A|})$. In statistical-mechanical terms, this means using renormalization (Yeomans, 1992). Probabilistically, the existence of a limiting (scaled) cumulant generating function is a weak dependence condition (den Hollander, 2000, §V.3.2). While under equation (10) we identified the T_A process with a time-transformed Lévy process, now we can only use a central limit theorem to say they are close (den Hollander, 2000, §V.3.1), reducing almost-sure to stochastic convergence. (See Jona-Lasinio (2001) on the relation between central limit theorems and renormalization.) In any event, asymptotic scaling of the log partition function implies $\widehat{\theta}$ is consistent.

[7] For small $\epsilon \in \mathbb{R}^d$, by a second order Taylor expansion, $J(\epsilon + \nabla a(\theta)) \approx \frac{1}{2}\langle \epsilon, I(\theta)\epsilon\rangle$, where $I(\theta)$ acts as the Fisher information rate; cf. Bahadur (1971).

5 Application: Non-projectibility of Exponential Random Graph Models

As mentioned in the introduction, our general results about projective structure in exponential families arose from questions about exponential random graph models of networks. To make the application clear, we must fill in some details regarding ERGMs.

Given a group of n nodes, the network among them is represented by the binary $n \times n$ **adjacency matrix** X, where $X_{ij} = 1$ if there is a tie from i to j and is 0 otherwise. (Undirected graphs impose $X_{ij} = X_{ji}$.) We may also have covariates for each node, say Y_i. Our projective structure will in fact be that of looking at the sub-graphs among larger and larger groups of nodes. That is, A is the sub-network among the first n nodes, and $B \supset A$ is the sub-network among the first $n + m$ nodes. The graph or adjacency matrix itself is the stochastic process which is to have an exponential family distribution, conditional on the covariates:

$$p_\theta(x|y) = \frac{e^{\langle \theta, t(x,y) \rangle}}{z(\theta|y)} \tag{19}$$

(We are only interested in the exponential-family distribution of the graph holding the covariates fixed.) As mentioned above, the components of T typically count the number of occurrences of various sub-graphs or motifs — as edges, triangles, larger cliques, "k-stars" (k nodes connected through a central node), etc. — perhaps interacted with values of the nodal covariates. The definition of T may include normalizing the counts of these "motifs" by data-independent combinatorial factors to yield densities.

A **dyad** consists of an unordered pair of individuals. In a dyadic independence model, each dyad's configuration is independent of every other dyad's (conditional on Y). In an ERGM, dyadic independence is equivalent to the (vector-valued) statistic T adding up over dyads,

$$t(X, Y) = \sum_{i=1}^{n} \sum_{j<i} t_{ij}(X_{ij}, X_{ji}, Y_i, Y_j) \tag{20}$$

That is, the statistic can be written as a sum of terms over the information available for each dyad. In particular, in **block models** (Bickel and Chen, 2009), Y_i is categorical, giving the type of node i, and the vector of sufficient statistics counts dyad configurations among pairs of nodes of given pairs of types. Dyadic independence implies projectibility: since all dyads have independent configurations, each dyad makes a separate additive contribution to T. Going from $n - 1$ to n nodes thus adds n terms, unconstrained by the configuration among the $n - 1$ nodes. T thus has separable increments, implying projectibility by Theorem 3.2. (Adding a new node adds only edges between the old nodes and the new, without disturbing the old counts.)[8] As the distribution factorizes into a product of $n(n-1)$ terms, each of exactly the same form, the log partition function scales exactly with $n(n-1)$, and the conclusions of §4 imply the strong consistency of the maximum likelihood estimator[9]. This result thus applies to the well-studied β-model (Barvinok and Hartigan, 2010; Chatterjee *et al.*, 2011; Rinaldo *et al.*, 2011)

Typically, however, ERGMs are *not* dyadic independence models. In many networks, if nodes i and j are both linked to k, then i and j are unusually likely to be directly linked. This will of course happen if nodes of the same type are especially likely to be friends ("homophily", McPherson *et al.* 2001), since then the posterior probability of i and j being of the same type is elevated. However, it can also be modeled directly. The direct way to do so is to introduce the number (or density) of triangles as a sufficient statistic, but this leads to pathological degeneracy (Rinaldo *et al.*, 2009), and modern specifications involve a large set of

[8] We have assumed the type of each node is available as a covariate. In the stochastic block model, types are latent, and the marginal distribution of graphs sums over type-conditional distributions. Proposition B.1 in Appendix A shows that such summing-over-latents preserves projectibility. For stochastic block models, projectibility also follows from Lovász and Szegedy (2006, Theorem 2.7(ii)).

[9] An important variant of such models are the "degree-corrected block models" of Karrer and Newman (2011), where each node has a unique parameter, which is its expected degree. It is easily seen that the range of possible degrees for each new node is the same, no matter what the configuration of smaller sub-graphs (in which the node does not appear), as is the number of configurations giving rise to each degree. The conditions of §3.2 thus hold, and these models are projective.

triangle-like motifs (Snijders *et al.*, 2006; Wasserman and Robins, 2005; Handcock *et al.*, 2008). Empirically, when using such specifications, one often finds a non-trivial coefficient for such "transitivity" or "clustering", over and above homophily (Goodreau *et al.*, 2009). It is because of such findings that we ask whether the parameters in these models are projective.

Sadly, no statistic which counts triangles, or larger motifs, can have the nice additive form of dyad counts, no matter how we decompose the network. Take, for instance, triangles. Any given edge among the first n nodes *could* be part of a triangle, depending on ties to the next node. Thus to determine the number of triangles among the first $n+1$ nodes, we need much more information about the sub-graph of the first n nodes than just the number of triangles among them. Indeed, we can go further. The range of possible increments to the number of triangles changes with the number of existing triangles. This is quite incompatible with separable increments, so, by (3.2), the parameters cannot be projective. We remark that the non-projectibility of Markov graphs (Frank and Strauss, 1986), a special instance of ERGMs where the sufficient statistics count edges, k-stars and triangles, was noted in Lauritzen (2008).

Parallel arguments apply to the count of any motif of k nodes, $k > 2$. Any given edge (or absence of an edge) among the first n nodes could be part of such a motif, depending on the edges involving the next $k-2$ nodes. Such counts are thus not nicely additive. For the same reasons as with triangles, the range of increments for such statistics is not constant, and non-separable increments imply non-projective family.

While these ERGMs are not projective, some of them may, as a sort of consolation prize, still satisfy equation (14). For instance, in models where T has two elements, the number of edges and the (normalized) number of triangles or of 2-stars, the log partition function is known to scale like $n(n-1)$ as the number of nodes $n \to \infty$, at least in the parameter regimes where the models behave basically like either very full or very empty Erdős-Rényi networks (Park and Newman, 2004b,a, 2006; Chatterjee and Dey, 2010; Chatterjee and Diaconis, 2011; Bhamidi *et al.*, 2011). (We suspect, from Park and Newman 2004b; Xiang and Neville 2011; Chatterjee and Diaconis 2011, that similar results apply to many other ERGMs.) Thus, by equation (18), if we fix a large number n of nodes and generate a graph X from $\mathbb{P}_{\theta,n}$, the probability that the MLE $\widehat{\theta}(X)$ will be more than ϵ away from θ will be exponentially small in $n(n-1)$ and ϵ^2. Since these models are not projective, however, it is impossible to *improve* parameter estimates by getting more data, since parameters for smaller sub-graphs just cannot be extrapolated to larger graphs (or vice versa).

We thus have a near-dichotomy for ERGMs. Dyadic independence models have separable and independent increments to the statistics, and the resulting family is projective. However, specifications where the sufficient statistics count larger motifs cannot have separable increments and projectibility does not hold. Such an ERGM may provide a good description of a given social network on a certain set of nodes, but it cannot be projected to give predictions on any larger or more global graph from which that one was drawn. If an ERGM is postulated for the whole network, then inference for its parameters must explicitly treat the unobserved portions of the network as missing data (perhaps through an expectation-maximization algorithm), though of course there may be considerable uncertainty about just how much data is missing.

6 Conclusion

Specifications for exponential families of dependent variables in terms of joint distributions are surprisingly delicate; the statistics must be chosen extremely carefully, in order to achieve separable increments. (Conditional specifications do not have this problem.) This has, perhaps, been obscured in the past by the emphasis on using exponential families to model multivariate but independent cases, as IID models are always projective.

Network models, one of the outstanding applications of exponential families, suffer from this problem in an acute form. Dyadic independence models are projective models, but are sociologically extremely implausible, and certainly do not manage to reproduce the data well. More interesting specifications, involving clustering terms, never have separable increments. We thus have an impasse which it seems can only be resolved by going to a different family of specifications. One possibility — which however requires more and different data — is to model the evolution of networks over time (Snijders, 2005). In particular, Hanneke

et al. (2010) considers situations where the distribution of the network at time $t+1$ conditional on the network at time t follows an exponential family. Even when the statistics in the conditional specification include (say) changes in the number of triangles, the issues raised above do not apply.

Roughly speaking, the issue with the non-projective ERGM specifications, and with other non-projective exponential families, is that the dependency structure corresponding to the statistics allows interactions between arbitrary collections of random variables. It is not possible, with those statistics, to "screen off" one part of the assemblage from another by conditioning on boundary terms. Suppose our larger information set B consists of two non-overlapping and strictly smaller information sets, $A \subset B$ and $C \subset B$, plus the new observation obtained by looking at both A and C. (For instance, the latter might be the edges between two disjoint sets of nodes.) Then the models which work properly are ones where the sufficient statistic for B partitions into marginal terms from A and C, plus the interactions strictly between them: $t_B(X_B) = t_A(X_A) + T_C(X_C) + T_{B\setminus(A\cup C)}(X_{B\setminus(A\cup C)})$. In physical language (Landau and Lifshitz, 1980), the energy for the whole assemblage needs to be a sum of two "volume" terms for its sub-assemblages, plus a "surface" term for their interface. The network models with non-projective parameters do not admit such a decomposition; every variable, potentially, interacts with every other variable.

One might try to give up the exponential family form, while keeping finite-dimensional sufficient statistics. We suspect that this will not work, however, since Lauritzen (1988) showed that whenever the sufficient statistics form a semi-group, the models must be either ordinary exponential families, or certain generalizations thereof with much the same properties. We believe that there exists a purely algebraic characterization of the sufficient statistics compatible with projectibility, but must leave this for the future.

One reason for the trouble with ERGMs is that every infinite exchangeable graph distribution is actually a mixture over projective dyadic-independence distributions (Diaconis and Janson, 2008; Bickel and Chen, 2009), though not necessarily ones with a finite-dimensional sufficient statistic. Along any one sequence of sub-graphs from such an infinite graph, in fact, the densities of all motifs approach limiting values which pick out a unique projective dyadic-independence distribution (Diaconis and Janson, 2008) (and cf. Lauritzen 1988, 2008). This suggests that an alternative to parametric inference would be non-parametric estimation of the limiting dyadic-independence model, by smoothing the adjacency matrix; this, too, we pursue elsewhere.

Acknowledgments We thank Luis Carvalho, Aaron Clauset, Mark Handcock, Steve Hanneke, Brain Karrer, Sergey Kirshner, Steffen Lauritzen, David Lazer, John Miller, Martina Morris, Jennifer Neville, Mark Newman, Peter Orbanz, Andrew Thomas and Chris Wiggins, for valuable conversations; an anonymous referee of an earlier version for pointing out a gap in a proof; and audiences at the Boston University probability and statistics seminar, and Columbia University's applied math seminar. CRS was supported by grants from the National Institutes of Health (# 2 R01 NS047493) and the Institute for New Economic Thinking. AR was partially supported by grants from the U.S. Air Force Office of Scientific Research (AFOSR #FA9550-12-1-0392) and from the Defense Advanced Research Projects Agency.

7 Proofs

For notation in this section, without loss of generality, fix a generic pair of subsets $A \subset B$ and a value of θ. We will write a representative point $x_B \in \mathcal{X}_B$ as $x_B = (x, y)$, with $x \in \mathcal{X}_A$ and $y \in \mathcal{X}_{B\setminus A}$. Also, we abbreviate $t_B(x,y) - t_A(x)$, for $x \in \mathcal{X}_A$ and $y \in \mathcal{X}_{B\setminus A}$ by $t_{B\setminus A}(x,y)$.

7.1 Proof of Theorem 3.2

For clarity, we prove the two directions separately. First we show that projectability implies separable increments.

Proposition 7.1. *If the exponential family $\{\mathbb{P}_\theta\}_{A\in\mathcal{A}}$ is projective, then the sufficient statistics $\{T_A\}_{A\in\mathcal{A}}$ have separable increments, i.e. $A \subset B$ implies that $v_{B\setminus A|A}(\delta, x) = v_{B\setminus A}(\delta)$.*

Proof. By projectibility, for each θ,

$$p_{A,\theta}(x) = \sum_{y \in \mathcal{X}_{B\setminus A}} p_{B,\theta}(x,y) = \sum_{y \in \mathcal{X}_{B\setminus A}} \frac{e^{\langle \theta, t_B(x,y)\rangle}}{z_B(\theta)} \tag{21}$$

$$= \frac{1}{z_B(\theta)} \sum_{y \in \mathcal{X}_{B\setminus A}} \exp\left\{\langle \theta, t_B(x,y) - t_A(x)\rangle + \langle \theta, t_A(x)\rangle\right\} \tag{22}$$

$$= \frac{e^{\langle \theta, t_A(x)\rangle}}{z_A(\theta)} \frac{z_A(\theta)}{z_B(\theta)} \sum_{y \in \mathcal{X}_{B\setminus A}} \exp\left\{\langle \theta, t_{B\setminus A}(x,y)\rangle\right\} \tag{23}$$

$$= p_{A,\theta}(x) \frac{z_A(\theta)}{z_B(\theta)} \sum_{y \in \mathcal{X}_{B\setminus A}} \exp\left\{\langle \theta, t_{B\setminus A}(x,y)\rangle\right\}, \tag{24}$$

which implies that, for all $x \in \mathcal{X}_A$,

$$\sum_{y \in \mathcal{X}_{B\setminus A}} \exp\left\{\langle \theta, t_{B\setminus A}(x,y)\rangle\right\} = \frac{z_B(\theta)}{z_A(\theta)}. \tag{25}$$

Re-writing the left-hand side of equation (25) as a sum over the set $\Delta(x)$ of values which the increment $t_{B\setminus A}(x,y)$ to the sufficient statistic might take yields

$$\sum_{\delta \in \Delta(x)} v_{B\setminus A|A}(\delta, x) \exp\langle \theta, \delta\rangle = \frac{z_B(\theta)}{z_A(\theta)}. \tag{26}$$

where the joint volume factor is defined in (6). Since the right-hand side of equation (26) is the same for all x, so must the left-hand side.

Observe that this left-hand side is the Laplace transform of the function $v_{B\setminus A|A}(\cdot, x)$. The latter is a non-negative function which defines a measure on $\mathbb{R}^d$, whose support is $\Delta(x)$. Hence,

$$\sum_{\delta \in \Delta(x)} \frac{v_{B\setminus A|A}(\delta, x)}{\sum_{\delta' \in \Delta(x)} v_{B\setminus A|A}(\delta', x)} \exp\langle \theta, \delta\rangle \tag{27}$$

is the Laplace transform of a discrete probability measure in $\mathbb{R}^d$. But the denominator in the inner sum is just $|\mathcal{X}_{B\setminus A}|$, no matter what x might be[10]. So we have that for any $x, x' \in \mathcal{X}_A$, and all $\theta \in \Theta$,

$$\sum_{\delta \in \Delta(x)} \frac{v_{B\setminus A|A}(\delta, x)}{|\mathcal{X}_{B\setminus A}|} \exp\langle \theta, \delta\rangle = \sum_{\delta \in \Delta(x')} \frac{v_{B\setminus A|A}(\delta, x')}{|\mathcal{X}_{B\setminus A}|} \exp\langle \theta, \delta\rangle \tag{28}$$

Since both sides of equation (28) are Laplace transforms of probability measures on a common space, and the equality holds on all of Θ, which contains an open set, we may conclude that the two measures are equal (Barndorff-Nielsen, 1978, Theorem 7.3). This means that they have the same support, $\Delta(x) = \Delta(x') = \Delta$, and that they have the same density with respect to counting measure on Δ. As they also have the same normalizing factor (viz., $|\mathcal{X}_{B\setminus A}|$), we get that $v_{B\setminus A|A}(\delta, x) = v_{B\setminus A|A}(\delta, x') = v_{B\setminus A}(\delta)$. Since the points x and x' are arbitrary, this last property is precisely having separable increments. ■

Next, we prove the reverse direction, namely that separable increments imply projectibility. This is clearer with some preliminary lemmas.

[10] This can be seen either from recalling that exponential families have full support, or from defining T_B as a total and not a partial function on $\mathcal{X}_B$.

Lemma 7.2. *If the sufficient statistics have separable increments, then the joint volume factors factorize, i.e.,*

$$v_{A,B\setminus A}(t,\delta) = v_A(t)v_{B\setminus A}(\delta), \tag{29}$$

for all $A \subset B$, t and δ.

Proof. By definition,

$$v_{A,B\setminus A}(t,\delta) = \sum_{\{x\in\mathcal{X}_A\colon\, t_A(x)=t\}} v_{B\setminus A|A}(\delta,x). \tag{30}$$

When the statistic has separable increments, $v_{B\setminus A|A}(\delta,x) = v_{B\setminus A}(\delta)$, so

$$v_{A,B\setminus A}(t,\delta) = \sum_{\{x\colon\, t_A(x)=t\}} v_{B\setminus A}(\delta) = v_A(t)v_{B\setminus A}(\delta), \tag{31}$$

proving the claim. ■

Lemma 7.3. *If the joint volume factor factorizes, then the sufficient statistics has independent increments, and the distribution of the sufficient static is projective.*

Proof. Without loss of generality, fix a value t for T_A and δ for $T_{B\setminus A}$. By the law of total probability and the definition of the volume factor,

$$\mathbb{P}_{\theta,B}(T_A = t, T_{B\setminus A} = \delta) = v_{A,B\setminus A}(t,\delta)\frac{e^{\langle\theta,t\rangle}e^{\langle\theta,\delta\rangle}}{z_B(\theta)}. \tag{32}$$

If the volume factor factorizes, so that $v_{A,B\setminus A}(t,\delta) = v_A(t)v_{B\setminus A}(\delta)$, then we obtain

$$\mathbb{P}_{\theta,B}(T_A = t, T_{B\setminus A} = \delta) = \left[\frac{1}{z_A(\theta)}v_A(t)e^{\langle\theta,t\rangle}\right]\left[\frac{z_A(\theta)}{z_B(\theta)}v_{B\setminus A}(\delta)e^{\langle\theta,\delta\rangle}\right]. \tag{33}$$

It then follows that

$$\mathbb{P}_{\theta,B}(T_A = t, T_{B\setminus A} = \delta) = \mathbb{P}_{\theta,B}(T_A = t)\mathbb{P}_{\theta,B}(T_{B\setminus A} = \delta), \quad \forall\theta, \tag{34}$$

and thus that T has independent increments. To establish the projectibility of the distribution of T, sum over δ:

$$\begin{aligned}
\mathbb{P}_{\theta,B}(T_A = t) &= \sum_{\delta}\mathbb{P}_{\theta,B}(T_A = t, T_{B\setminus A} = \delta) \\
&= \frac{v_A(t)e^{\langle\theta,t\rangle}}{z_B(\theta)}\sum_{\delta} v_{B\setminus A}(\delta)e^{\langle\theta,\delta\rangle} \\
&= \frac{v_A(t)e^{\langle\theta,t\rangle}}{z_B(\theta)}z_{B\setminus A}(\theta)
\end{aligned}$$

Since $\mathbb{P}_{A,\theta}(T_A = t) = v_A(t)e^{\langle\theta,t\rangle}/z_A(\theta)$, and both distributions must sum to 1 over t, we can conclude that $z_A(\theta) = z_B(\theta)/z_{B\setminus A}(\delta)$, and hence that the distribution of the sufficient statistic is projective. ■

Lemma 7.4. *If the sufficient statistics of an exponential family have separable increments, then*

$$\mathbb{P}_{B,\theta}(X_A = x, T_{B\setminus A} = \delta) = \frac{1}{v_A(t_A(x))}\mathbb{P}_{B,\theta}(T_A = t_A(x), T_{B\setminus A} = \delta) \tag{35}$$

Proof. Abbreviate $t_A(x)$ by t. By the law of total probability,

$$\mathbb{P}_{B,\theta}(T_A = t, T_{B\setminus A} = \delta) = \sum_{(x,y):t_A(x)=t, t_{B\setminus A}(x,y)=\delta} p_{B,\theta}(x,y) \tag{36}$$

Since T_B is sufficient, and $t_B(x,y) = t + \delta$ for all (x,y) in the sum,

$$\mathbb{P}_{B,\theta}(T_A = t, T_{B\setminus A} = \delta) = v_{A,B\setminus A}(t,\delta) e^{\langle \theta, t+\delta \rangle} / z_B(\theta) \tag{37}$$

By parallel reasoning,

$$\mathbb{P}_{B,\theta}(X_A = x, T_{B\setminus A} = \delta) = v_{B\setminus A|A}(\delta, x) e^{\langle \theta, t+\delta \rangle} / z_B(\theta) \tag{38}$$

Therefore

$$\mathbb{P}_{B,\theta}(X_A = x, T_{B\setminus A} = \delta) = v_{B\setminus A|A}(\delta, x) \frac{\mathbb{P}_{B,\theta}(T_A = t, T_{B\setminus A} = \delta)}{v_{A,B\setminus A}(t,\delta)} \tag{39}$$

If the statistic has separable increments, then $v_{A,B\setminus A}(t,\delta) = v_A(t) v_{B\setminus A}(\delta) = v_A(t) v_{B\setminus A|A}(\delta, x)$, and the conclusion follows. ∎

Remark: The lemma does *not* follow merely from the joint volume factor separating, $v_{A,B\setminus A}(t,\delta) = v_A(t) v_{B\setminus A}(\delta)$. The conditional volume factor must also be constant in x.

Proposition 7.5. *If the sufficient statistic of an exponential family has separable increments, then the family is projective.*

Proof. We calculate the marginal probability of X_A in $\mathbb{P}_{\theta,B}$, by integrating out the increment to the sufficient statistic. (The set of possible increments, Δ, is the same for all x, by separability.) Once again, we abbreviate $t_A(x)$ by t.

$$\begin{aligned}
\mathbb{P}_{B,\theta}(X_A = x) &= \sum_{\delta \in \Delta} \mathbb{P}_{B,\theta}(X_A = x, T_{B\setminus A} = \delta) \\
&= \frac{1}{v_A(t)} \sum_{\delta \in \Delta} \mathbb{P}_{B,\theta}(T_A = t, T_{B\setminus A} = \delta) \\
&= \frac{1}{v_A(t)} \sum_{\delta \in \Delta} \mathbb{P}_{B,\theta}(T_A = t) \mathbb{P}_{B,\theta}(T_{B\setminus A} = \delta | T_A = t) \\
&= \frac{\mathbb{P}_{B,\theta}(T_A = t)}{v_A(t)} \\
&= \frac{\mathbb{P}_{A,\theta}(T_A = t)}{v_A(t)} \\
&= \mathbb{P}_{A,\theta}(X_A = x)
\end{aligned}$$

These steps use, in succession: Lemma 7.4; the fact that conditional probabilities sum to 1; the projectibility of the sufficient statistics (via Lemmas 7.2 and 7.3); and the definition of $v_A(t)$. ∎

7.2 Other Proofs

Proof of Proposition 3.3. By Proposition 7.1, a projective family has separable increments, and by Lemma 7.3, separable increments implies independent increments. ∎

Proof of Proposition 3.4. By Proposition 7.1, every projective exponential family has separable increments. By Lemma 7.4, in an exponential family with separable increments,

$$\mathbb{P}_{B,\theta}(X_A = x, T_{B\setminus A} = \delta) = \frac{1}{v_A(t_A(x))} \mathbb{P}_{B,\theta}(T_A = t_A(x), T_{B\setminus A} = \delta) \tag{40}$$

Therefore, using projectibility,

$$\mathbb{P}_{B,\theta}(T_{B\setminus A} = \delta | X_A = x) = \frac{\mathbb{P}_{B,\theta}(T_A = t_A(x), T_{B\setminus A} = \delta)/v_A(t_A(x))}{p_{A,\theta}(x)} \tag{41}$$

By the definition of $v_A(\cdot)$, $p_{A,\theta}(x) = \mathbb{P}_{A,\theta}(T_A = t_A(x))/v_A(t_A(x))$, so

$$\mathbb{P}_{B,\theta}(T_{B\setminus A} = \delta | X_A = x) = \frac{\mathbb{P}_{B,\theta}(T_A = t_A(x), T_{B\setminus A} = \delta)}{\mathbb{P}_{A,\theta}(T_A = t_A(x))} \tag{42}$$

But, by Lemma 7.3, the sufficient statistics have a projective distribution with independent increments, implying

$$\mathbb{P}_{B,\theta}(T_A = t_A(x), T_{B\setminus A} = \delta) = \mathbb{P}_{A,\theta}(T_A = t_A(x))\mathbb{P}_{B,\theta}(T_{B\setminus A} = \delta) \tag{43}$$

Therefore,

$$\mathbb{P}_{B,\theta}(T_{B\setminus A} = \delta | X_A = x) = \mathbb{P}_{B,\theta}(T_{B\setminus A} = \delta) \tag{44}$$

and so $T_{B\setminus A} \perp\!\!\!\perp X_A$. ■

Proof of Proposition 3.5. Below we prove that if the suffiicient statistics of an exponential family have independent increments, then the volume factor separates, and the distribution of the statistic is projective.

Since T_B is a sufficient statistic, by the Neyman factorization theorem (Schervish, 1995, Thm. 2.21,p. 89)

$$\mathbb{P}_{B,\theta}(X_A = x, X_{B\setminus A} = y) = g_B(\theta, t_A(x) + t_{B\setminus A}(x, y))h(x, y) \tag{45}$$

In light of equation (2), we may take $h(x, y) = 1$. Abbreviating $t_A(x)$ by t and $t_{B\setminus A}(x, y)$ by δ, it follows that

$$\mathbb{P}_{B,\theta}(T_A = t, T_{B\setminus A} = \delta) = v_{A,B\setminus A}(t, \delta)g_B(\theta, t + \delta) \tag{46}$$

By independent increments, however,

$$\mathbb{P}_{B,\theta}(T_A = t, T_{B\setminus A} = \delta) = \mathbb{P}_{B,\theta}(T_A = t)\mathbb{P}_{B,\theta}(T_{B\setminus A} = \delta) \tag{47}$$

whence it follows that, for some functions $g_{B\setminus A}, k_A, k_{B\setminus A}$,

$$g_B(\theta, t + \delta) = g_A(\theta, t)g_{B\setminus A}(\theta, \delta) \tag{48}$$

and

$$v_{A,B\setminus A}(t, \delta) = k_A(t)k_{B\setminus A}(\delta) \tag{49}$$

and

$$\mathbb{P}_{B,\theta}(T_A = t, T_{B\setminus A} = \delta) = k_A(t)k_{B\setminus A}(\delta)g_A(\theta, t)g_{B\setminus A}(\theta, \delta) \tag{50}$$

To proceed, we must identify the new g and k functions. To this end, recalling that $v_A(t)$ is the number of x_A configurations such that $t_A(x_A) = t$, we have

$$\sum_\delta v_{A,B\setminus A}(t, \delta) = v_A(t)|\mathcal{X}_{B\setminus A}| \tag{51}$$

and, at the same time,

$$\sum_\delta v_{A,B\setminus A}(t, \delta) = k_A(t)\sum_\delta k_{B\setminus A}(\delta). \tag{52}$$

Clearly, then, $k_A(t) = c_1 v_A(t)$ while $\sum_\delta k_{B\setminus A}(\delta) = c_2|\mathcal{X}_{B\setminus A}|$. Since

$$\sum_t \sum_\delta v_{A,B\setminus A}(t, \delta) = |\mathcal{X}_A||\mathcal{X}_{B\setminus A}| \tag{53}$$

and $\sum_t v_A(t) = |\mathcal{X}_A|$, we need $c_1 c_2 = 1$, and may take $c_1 = c_2 = 1$ for simplicity. This allows us to write

$$v_{A,B\setminus A}(t,\delta) = v_A(t) v_{B\setminus A}(\delta) \tag{54}$$

which is exactly the assertion that the volume factor separates.

Turning to the g functions, we sum over δ again to obtain the marginal distribution of T_A:

$$\begin{aligned}
\mathbb{P}_{B,\theta}(T_A = t) &= \sum_\delta \mathbb{P}_{B,\theta}(T_A = t, T_{B\setminus A} = \delta) \\
&= \sum_\delta v_A(t) g_A(\theta, t) v_{B\setminus A}(\delta) g_{B\setminus A}(\theta, \delta) \\
&= v_A(t) g_A(\theta, t) \sum_\delta v_{B\setminus A}(\delta) g_{B\setminus A}(\theta, \delta).
\end{aligned}$$

Now, we finally we use the exponential-family form. Specifically, we know that

$$g_B(\theta, t+\delta) = \frac{e^{\langle \theta, t\rangle} e^{\langle \theta, \delta\rangle}}{z_B(\theta)}, \tag{55}$$

so that $g_A(\theta, t) \propto e^{\langle \theta, t\rangle}$, $g_{B\setminus A}(\theta, \delta) \propto e^{\langle \theta, \delta\rangle}$. Therefore,

$$\mathbb{P}_{B,\theta}(T_A = t) \propto v_A(t) e^{\langle \theta, t\rangle} \propto \mathbb{P}_{A,\theta}(T_A = t), \tag{56}$$

and normalization now forces

$$\mathbb{P}_{B,\theta}(T_A = t) = \mathbb{P}_{A,\theta}(T_A = t), \tag{57}$$

as desired. ■

Proof of Theorem 3.6. The conditional density of $X_{B\setminus A}$ given X_A is just the ratio of joint to marginal densities (both with the same θ, by projectibility):

$$\begin{aligned}
p_{B|A,\theta}(y|x) &= \frac{p_{B,\theta}(x,y)}{p_{A,\theta}(x)} = \frac{e^{\langle \theta, t_B(x,y)\rangle}/z_B(\theta)}{e^{\langle \theta, t_A(x)\rangle}/z_A(\theta)} && (58) \\
&= \frac{e^{\langle \theta, t_{B\setminus A}(x,y)\rangle}}{z_B(\theta)/z_A(\theta)} && (59)
\end{aligned}$$

which is an exponential family with parameter θ, sufficient statistic $T_{B\setminus A}$, and partition function $z_{B\setminus A|A}(\theta) \equiv z_B(\theta)/z_A(\theta)$. ■

Proof of Theorem 4.2. Under equation (14), the cumulant generating function also scales asymptotically, $\kappa_{A,\theta}(\phi)/r_{|A|} \to a(\theta+\phi) - a(\theta)$. Since a is differentiable, the Gärtner-Ellis theorem of large deviations theory (den Hollander, 2000, ch. V) implies that $T_A/r_{|A|}$ obeys a large deviations principle with rate $r_{|A|}$, and rate function given by equation (17), which is to say, equations (15) and (16). ■

Lemma 7.6. *The moment generating function of $T_{B\setminus A}$ is*

$$\frac{z_B(\theta+\phi) z_A(\theta)}{z_B(\theta) z_A(\theta+\phi)} = \frac{M_{\theta,B}(\phi)}{M_{\theta,A}(\phi)} \tag{60}$$

Proof. From the proof of Theorem 3.6, $X_{B\setminus A}|X_A$ has an exponential family distribution with sufficient statistic $T_{B\setminus A}$. Thus we may use equation (5) to find the moment generating function of $T_{B\setminus A}$ conditional on X_A:

$$M_{\theta,B\setminus A|A}(\phi) = \frac{z_{B\setminus A|A}(\theta+\phi)}{z_{B\setminus A|A}(\theta)} \tag{61}$$

$$= \frac{z_B(\theta+\phi)/z_A(\theta+\phi)}{z_B(\theta)/z_A(\theta)} \tag{62}$$

$$= \frac{z_B(\theta+\phi)z_A(\theta)}{z_B(\theta)z_A(\theta+\phi)} = \frac{M_{\theta,B}(\phi)}{M_{\theta,A}(\phi)} \tag{63}$$

$$\tag{64}$$

Since, however, $T_{B\setminus A} \perp\!\!\!\perp X_A$ (Proposition 3.4), equation (60) must also give the unconditional moment generating function. ■

References

Achlioptas, Dimitris, Aaron Clauset, David Kempe and Cristopher Moore (2005). "On the Bias of Traceroute Sampling (or: Why almost every network looks like it has a power law)." In *Proceedings of the 37th ACM Symposium on Theory of Computing*. URL http://arxiv.org/abs/cond-mat/0503087.

Ackland, Robert and Mathieu O'Neil (2011). "Online collective identity: The case of the environmental movement." *Social Networks*, **33**: 177–190. doi:10.1016/j.socnet.2011.03.001.

Ahmed, Nesreen K., Jennifer Neville and Ramana Kompella (2010). "Reconsidering the Foundations of Network Sampling." In *Proceedings of the 2nd Workshop on Information in Networks [WIN 2010]* (Sinan Aral and Foster Provost and Arun Sundararajan, eds.). URL http://www.cs.purdue.edu/homes/neville/papers/ahmed-win2010.pdf.

Anderson, Carolyn J., Stanley Wasserman and Bradley Crouch (1999). "A p^* Primer: Logit Models for Social Networks." *Social Networks*, **21**: 37–66. doi:10.1016/S0378-8733(98)00012-4.

Bahadur, R. R. (1971). *Some Limit Theorems in Statistics*. Philadelphia: SIAM Press.

Barndorff-Nielsen, O. E. (1978). *Information and Exponential Families in Statistical Theory*. New York: John Wiley and Sons.

Barvinok, A. and J. A. Hartigan (2010). "The number of graphs and a random graph with a given degree sequence." Electronic pre-print, arxiv.org. URL http://arxiv.org/abs/1003.0356.

Besag, Julian (1989). "A Candidate's Formula: A Curious Result in Bayesian Prediction." *Biometrika*, **76**: 183. URL http://www.jstor.org/pss/2336383.

Bhamidi, Shankar, Guy Bresler and Allan Sly (2011). "Mixing Time of Exponential Random Graphs." *Annals of Applied Probability*, **21**: 2146–2170. URL http://arxiv.org/abs/0812.2265. doi:10.1214/10-AAP740.

Bickel, Peter J. and Aiyou Chen (2009). "A Nonparametric View of Network Models and Newman-Girvan and Other Modularities." *Proceedings of the National Academy of Sciences (USA)*, **106**: 21068–21073. doi:10.1073/pnas.0907096106.

Brown, Lawrence D. (1986). *Fundamentals of Statistical Exponential Families: with Applications in Statistical Decision Theory*. Hayward, California: Institute of Mathematical Statistics. URL http://projecteuclid.org/euclid.lnms/1215466757.

Butler, Ronald W. (1986). "Predictive Likelihood Inference with Applications." *Journal of the Royal Statistical Society B*, **48**: 1–38. URL http://www.jstor.org/stable/2345635.

Carrington, Peter J., John Scott and Stanley Wasserman (eds.) (2005). *Models and Methods in Social Network Analysis*. Cambridge, England: Cambridge University Press.

Chatterjee, Sourav and Partha S. Dey (2010). "Applications of Stein's Method for Concentration Inequalities." *Annals of Probability*, **38**: 2443–2485. URL http://arxiv.org/abs/0906.1034. doi:10.1214/10-AOP542.

Chatterjee, Sourav and Persi Diaconis (2011). "Estimating and Understanding Exponential Random Graph Models." E-print, arxiv:1102.2650. URL http://arxiv.org/abs/1102.2650.

Chatterjee, Sourav, Persi Diaconis and Allan Sly (2011). "Random graphs with a given degree sequence." *Annals of Applied Probability*, **21**: 1400–1435. URL http://arxiv.org/abs/1005.1136. doi:10.1214/10-AAP728.

Daraganova, Galina, Pip Pattison, Johan Koskinen, Bill Mitchell, Anthea Bill, Martin Watts and Scott Baum (2012). "Networks and geography: Modelling community network structure as the outcome of both spatial and network processes." *Social Networks*, **34**: 6–17. doi:10.1016/j.socnet.2010.12.001.

de la Haye, Kayla, Garry Robins, Philip Mohr and Carlene Wilson (2010). "Obesity-related behaviors in adolescent friendship networks." *Social Networks*, **32**: 161–167. doi:10.1016/j.socnet.2009.09.001.

den Hollander, Frank (2000). *Large Deviations*. Providence, Rhode Island: American Mathematical Society.

Diaconis, Persi and Svante Janson (2008). "Graph Limits and Exchangeable Random Graphs." *Rendiconti di Matematica e delle sue Applicazioni*, **28**: 33–61. URL http://arxiv.org/abs/0712.2749.

Easley, David and Jon Kleinberg (2010). *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge, England: Cambidge University Press.

Faust, Katherine and John Skvoretz (2002). "Comparing Networks Across Space and Time, Size and Species." *Sociological Methodology*, **32**: 267–299. URL http://www.socsci.uci.edu/~kfaust/faust/research/articles/faust_skvoretz_comparing_sm_2002.pdf. doi:10.1111/1467-9531.00118.

Frank, Ove and David Strauss (1986). "Markov Graphs." *Journal of the American Statistical Associaiton*, **81(395)**: 832–842.

Goldenberg, Anna, Alice X. Zheng, Stephen E. Fienberg and Edoardo M. Airoldi (2009). "A Survey of Statistical Network Models." *Foundations and Trends in Machine Learning*, **2**: 1–117. URL http://arxiv.org/abs/0912.5410.

Gondal, Neha (2011). "The local and global structure of knowledge production in an emergent research field: An exponential random graph analysis." *Social Networks*, **33**: 20–30. doi:10.1016/j.socnet.2010.09.001.

Gonzalez-Bailon, Sandra (2009). "Opening the black box of link formation: Social factors underlying the structure of the web." *Social Networks*, **31**: 271–280. doi:10.1016/j.socnet.2009.07.003.

Goodreau, Steven M., James A. Kitts and Martina Morris (2009). "Birds of a Feather, Or Friend of a Friend?: Using Exponential Random Graph Models to Investigate Adolescent Social Networks." *Demography*, **46**: 103–125. doi:10.1353/dem.0.0045.

Grünwald, Peter D. (2007). *The Minimum Description Length Principle*. Cambridge, Massachusetts: MIT Press.

Handcock, Mark S. and Krista J. Gile (2010). "Modeling Social Networks from Sampled Data." *Annals of Applied Statistics*, **4**: 5–25. URL http://arxiv.org/abs/1010.0891.

Handcock, Mark S., David R. Hunter, Carter T. Butts, Steven M. Goodreau and Martina Morris (2008). "statnet: Software Tools for the Representation, Visualization, Analysis and Simulation of Network Data." *Journal of Statistical Software*, **24**: 1–11. URL http://www.jstatsoft.org/v24/i01. Special issue on statnet.

Hanneke, Steve, Wenjie Fu and Eric P. Xing (2010). "Discrete Temporal Models of Social Networks." *Electronic Journal of Statistics*, **4**: 585–605. URL http://projecteuclid.org/euclid.ejs/1276694116.

Holland, Paul W. and Samuel Leinhardt (1981). "An Exponential Family of Probability Distributions for Directed Graphs." *Journal of the American Statistical Association*, **76**: 33–50. URL http://www.jstor.org/pss/2287037.

Jona-Lasinio, Giovanni (2001). "Renormalization Group and Probability Theory." *Physics Reports*, **352**: 439–458. URL http://arxiv.org/abs/cond-mat/0009219.

Kallenberg, Olav (2002). *Foundations of Modern Probability*. New York: Springer-Verlag, 2nd edn.

Karrer, Brian and Mark E. J. Newman (2011). "Stochastic Blockmodels and Community Structure in Networks." *Physical Review E*, **83**: 016107. URL http://arxiv.org/abs/1008.3926.

Kolaczyk, Eric D. (2009). *Statistical Analysis of Network Data*. New York: Springer-Verlag.

Kossinets, Gueorgi (2006). "Effects of Missing Data in Social Networks." *Social Networks*, **28**: 247–268. URL http://arxiv.org/abs/cond-mat/0306335. doi:10.1016/j.socnet.2005.07.002.

Krivitsky, Pavel N., Mark S. Handcock and Martina Morris (2011). "Adjusting for Network Size and Composition Effects in Exponential-Family Random Graph Models." *Statistical Methodology*, **8**: 319–339. URL http://arxiv.org/abs/1004.5328. doi:10.1016/j.stamet.2011.01.005.

Landau, L. D. and E. M. Lifshitz (1980). *Statistical Physics*. Oxford: Pergamon Press.

Lauritzen, Steffen L. (1974). "Sufficiency, Prediction and Extreme Models." *Scandinavian Journal of Statistics*, **1**: 128–134. URL http://www.jstor.org/pss/4615564.

— (1988). *Extremal Families and Systems of Sufficient Statistics*. Berlin: Springer-Verlag.

— (2008). "Exchangeable Rasch Matrices." *Rendiconti di Matematica e delle sue Applicazioni*, **28**: 83–95. URL http://www.stats.ox.ac.uk/~steffen/papers/rendiconti.pdf.

Lovász, László and Balázs Szegedy (2006). "Limits of Dense Graph Sequences." *Journal of Combinatorial Theory B*, **96**: 933–957. URL http://arxiv.org/abs/math.CO/0408173. doi:10.1016/j.jctb.2006.05.002.

Lubbers, Miranda J. and Tom A. B. Snijders (2007). "A comparison of various approaches to the exponential random graph model: A reanalysis of 102 student networks in school classes." *Social Networks*, **29**: 489–507. doi:10.1016/j.socnet.2007.03.002.

Mandelbrot, Benoit (1962). "The Role of Sufficiency and of Estimation in Thermodynamics." *Annals of Mathematical Statistics*, **33**: 1021–1038. URL http://projecteuclid.org/euclid.aoms/1177704470.

McPherson, Miller, Lynn Smith-Lovin and James M. Cook (2001). "Birds of a Feather: Homophily in Social Networks." *Annual Review of Sociology*, **27**: 415–444. doi:10.1146/annurev.soc.27.1.415.

Nauenberg, Michael (2003). "Critique of q-entropy for Thermal Statistics." *Physical Review E*, **67**: 036114. URL http://arxiv.org/abs/cond-mat/0210561.

Newman, Mark E. J. (2010). *Networks: An Introduction*. Oxford, England: Oxford University Press.

Orbanz, Peter (2011). "Projective Limit Techniques in Bayesian Nonparametrics." Unpublished.

Park, Juyong and Mark E. J. Newman (2004a). "Solution of the 2-star Model of a Network." *Physical Review E*, **70**: 066146. URL http://arxiv.org/abs/cond-mat/0405457.

— (2004b). "The Statistical Mechanics of Networks." *Physical Review E*, **70**: 066117. URL http://arxiv.org/abs/cond-mat/0405566.

— (2006). "Solution for the Properties of a Clustered Network." *Physical Review E*, **72**: 026136. URL http://arxiv.org/abs/cond-mat/0412579.

Rinaldo, Alessandro, Stephen E. Fienberg and Yi Zhou (2009). "On the Geometry of Discrete Exponential Families with Application to Exponential Random Graph Models." *Electronic Journal of Statistics*, **3**: 446–484. URL http://projecteuclid.org/euclid.ejs/1243343761.

Rinaldo, Alessandro, S. Petrović and Stephen E. Fienberg (2011). "Maximum Likelihood Estimation in Network Models." Electronic pre-print, arxiv.org. URL http://arxiv.org/abs/1105.6145.

Robins, Garry, Tom Snijders, Peng Wang, Mark Handcock and Philippa Pattison (2007). "Recent developments in exponential random graph (p^*) models for social networks." *Social Networks*, **29**: 192–215. URL http://www.stats.ox.ac.uk/~snijders/RobinsSnijdersWangHandcockPattison.pdf. doi:10.1016/j.socnet.2006.08.003.

Schaefer, David R. (2012). "Youth co-offending networks: An investigation of social and spatial effects." *Social Networks*, **34**: 141–149. doi:10.1016/j.socnet.2011.02.001.

Schervish, Mark J. (1995). *Theory of Statistics*. Berlin: Springer-Verlag.

Snijders, Tom A. B. (2005). "Models for Longitudinal Network Data." In Carrington *et al.* (2005), pp. 215–247. URL http://stat.gamma.rug.nl/snijders/cup_ch11.pdf.

Snijders, Tom A. B., Philippa E. Pattison, Garry L. Robins and Mark S. Handcock (2006). "New Specifications for Exponential Random Graph Models." *Sociological Methodology*, **36**: 99–153. URL http://www.csss.washington.edu/Papers/wp42.pdf. doi:10.1111/j.1467-9531.2006.00176.x.

Stumpf, Michael P. H., Carsten Wiuf and Robert M. May (2005). "Subnets of Scale-free Networks are not Scale-free: Sampling Properties of Networks." *Proceedings of the National Academy of Sciences (USA)*, **102**: 4221–4224. doi:10.1073/pnas.0501179102.

Touchette, Hugo (2009). "The Large Deviations Approach to Statistical Mechanics." *Physics Reports*, **478**: 1–69. URL http://arxiv.org/abs/0804.0327. doi:10.1016/j.physrep.2009.05.002.

Vermeij, Lotte, Marijtje A. J. van Duijin and Chris Baerveldt (2009). "Ethnic segregation in context: Social discrimination among native Dutch pupils and their ethnic minority classmates." *Social Networks*, **31**: 230–239. doi:10.1016/j.socnet.2009.06.002.

Wainwright, Martin J. and Michael I. Jordan (2008). "Graphical Models, Exponential Families, and Variational Inference." *Foundations and Trends in Machine Learning*, **1**: 1–305. URL http://www.cs.berkeley.edu/~jordan/papers/wainwright-jordan-fnt.pdf. doi:10.1561/2200000001.

Wasserman, Stanley and Philippa Pattison (1996). "Logit Models and Logistic Regressions for Social Networks: I. An Introduction to Markov Graphs and p^*." *Psychometrika*, **61**: 401–425. doi:10.1007/BF02294547.

Wasserman, Stanley and Garry Robins (2005). "An Introduction to Random Graphs, Dependence Graphs, and p^*." In Carrington *et al.* (2005), pp. 148–161.

Xiang, Rongjing and Jennifer Neville (2011). "Relational Learning with One Network: An Asymptotic Analysis." In *Proceedings of the* 14^{th} *International Conference on Artificial Intelligence and Statistics [AISTATS 2011]* (Geoffrey Gordon and David Dunson and Miroslav Dudík, eds.), vol. 15 of *Journal of Machine Learning Research: Workshops and Conference Proceedings*, pp. 779–788. URL http://jmlr.csail.mit.edu/proceedings/papers/v15/xiang11a.html.

Yeomans, Julia M. (1992). *Statistical Mechanics of Phase Transitions*. Oxford: Clarendon Press.

A Non-Uniform Base Measures

For each $A \in \mathcal{A}$, we introduce a finite reference measure μ_A. The exponential family is then given not by Eq. (2) in SR but by

$$p_{A,\theta}(x) = \mu_A(x)\frac{e^{\langle\theta, t_A(x)\rangle}}{z_A(\theta)} \tag{65}$$

with the suitably-modified partition function

$$z_A(\theta) = \sum_{x\in\mathcal{X}_A} \mu_A(x)e^{\langle\theta, t_A(x)\rangle} \tag{66}$$

The family of base measures need not itself be projective. However, we do require that if $\mu_A(x) > 0$, then $\mu_B(x \times \mathcal{X}_{B\setminus A}) > 0$. (If this is not so, then some configurations which are allowed for small samples are forbidden for larger ones, and clearly no exponential family with respect to this μ could possibly be projective.) With this condition, the ratio

$$\mu_{B\setminus A|A}(y|x) = \mu_B(x, y)/\mu_A(x) \tag{67}$$

is well-defined μ_B-almost-everywhere. It defines a finite measure on $\mathcal{X}_{B\setminus A}$.

The use of a non-uniform base measure requires some modifications to the definitions of volume factors. First, marginal volume factors are given in terms of the base measure:

$$v_A(t) = \mu_A(\{x : t_A(x) = t\}) = \sum_{x:t_A(x)} \mu_A(x) \tag{68}$$

The joint volume factor may also be defined directly from the base measure:

$$v_{A,B\setminus A}(t, \delta) = \mu_B(\{x, y : t_A(x), t_{B\setminus A}(x, y) = \delta\}) = \sum_{x,y:t_A(x),t_{B\setminus A}(x,y)=\delta} h_B(x, y) \tag{69}$$

and the conditional volume factor from the ratios of the base measures:

$$v_{B\setminus A|A}(\delta, x) = \sum_{y:t_{B\setminus A}(x,y)=\delta} \mu_{B\setminus A|A}(y|x) \tag{70}$$

Note that all of these definitions reduce to the ones given before when μ is counting measure.

We still say that the sufficient statistics have separable increments when $v_{B\setminus A|A}(\delta, x) = v_{B\setminus A}(\delta)$ for all x. Unfortunately, whether the statistics have separable increments can change with the choice of base measure[11].

[11] We are grateful to an anonymous referee and to Pavel Krivitsky for independently raising this question. To give a trivial example, let $\mathcal{X}_A = \{a, b\}$, $\mathcal{X}_{B\setminus A} = \{\alpha, \beta\}$, with $T_B(a, \alpha) - T_A(a) = T_B(b, \beta) - T_A(b)$, while $T_B(a, \beta) - T_A(a) = T_B(b, \alpha) - T_A(b)$, and these two increments are different from each other. Clearly, T has separable increments under the uniform base measure. If $\mu_{B\setminus A}(\alpha) \neq \mu_{B\setminus A}(\beta)$, then T does not have separable increments.

There needs to be one further restriction on the base measures, which is that, for all x and x' in $\mathcal{X}_A$,

$$\sum_{y \in \mathcal{X}_{B \setminus A}} \mu_{B \setminus A|A}(y|x) = \sum_{y \in \mathcal{X}_{B \setminus A}} \mu_{B \setminus A|A}(y|x') \; . \tag{71}$$

This holds when μ_B is a product measure on $A \times B \setminus A$. (This is the case, for instance, for the model of Krivitsky *et al.* (2011).) It also holds if the family of μ is projective, since then $\mu_{B \setminus A|A}$ is a conditional probability measure and must integrate to 1. We conjecture that Eq. 71 can in fact always be imposed, through a suitable re-scaling of μ, but have not shown this.

Under Eq. 71 and the modified definitions of the volume factors, most proofs go through as given above. The two exceptions are as follows.

In the proof of Proposition 4 in SR, the normalizing factor in Eq. (27) in SR, $\sum_{\delta' \in \Delta(x)} v_{B \setminus A|A}(\delta', x)$ is the same as $\sum_{y \in \mathcal{X}_{B \setminus A}} \mu_{B \setminus A|A}(y|x)$. This is not necessarily the cardinality $|\mathcal{X}_{B \setminus A}|$, but so long as it is constant in x, the rest of the proof holds.

In the proof of Lemma 1 in SR, we modify Eq. (31) in SR to read

$$v_{A,B \setminus A}(t, \delta) = \sum_{x : t_A(x) = t} \mu_A(x) v_{B \setminus A|A}(\delta, x) \tag{72}$$

which is easily verified from the definitions. The proof now goes through as before.

B From Conditional to Unconditional Projectibility

Proposition B.1. *Suppose that, for each $A \in \mathcal{A}$, $\mathcal{Y}_A = \mathcal{X}_A \times \mathcal{C}_A$, and that $\{\mathcal{P}_A\}_{A \in \mathcal{A}}$ is a family of distributions on $\mathcal{Y}_A$, i.e., of joint distributions of X_A and C_A. If (i) the marginal distributions of $\{C_A\}_{A \in \mathcal{A}}$ are projective, (ii) the conditional distributions $X_A|C_A$ are projective (almost always), and (iii) under P_B, X_A is independent of $C_{B \setminus A}$ given C_A, then the marginal distribution of X is projective.*

Proof. Use the law of total probability to expand $P_B(X_A = x)$:

$$\begin{aligned}
P_B(X_A = x) &= \sum_{c,d,y} P_B(X_A = x, X_{B \setminus A} = y, C_A = c, C_{B \setminus A} = d) \\
&= \sum_{c,d} P_B(C_A = c, C_{B \setminus A} = d) \sum_{y} P_B(X_A = x, X_{B \setminus A} = y | C_A = c, C_{B \setminus A} = d) \\
&= \sum_{c,d} P_B(C_A = c, C_{B \setminus A} = d) P_A(X_A = x | C_A = c)
\end{aligned}$$

by conditions (ii) and (iii).

$$\begin{aligned}
P_B(X_A = x) &= \sum_{c} P_B(C_A = c) P_A(X_A = x | C_A = c) \sum_{d} P_B(C_{B \setminus A} = d | C_A = c) \\
&= \sum_{c} P_B(C_A = c) P_A(X_A = x | C_A = c) \\
&= P_A(X_A = x)
\end{aligned}$$

since by condition (i), $P_A(C_A = c) = P_B(C_A = c)$. ∎

In addition to the application to stochastic block models, we note that this (reassuringly) shows that hidden Markov models are projective.