STAT 538 Lecture 3.1
Exponential Random Graph Models (ERGM)

©Marina Meilă
mmp@stat.washington.edu

1 What is an ERGM?

Let $\mathcal{G} = (V, E)$ be an undirected graph, with $|V| = n$ nodes and edge set $E = \{ij,\ i \neq j\} \subset V \times V$. A **random graph model** is a distribution $P(E|V)$ defined for all finite sets of nodes V. An equivalent way to define a random graph over V is to associate an indicator variable Y_{ij} to each pair of nodes, and then write P as a the distribution of $Y|V$ and some parameters θ. In this context, an **Exponential Random Graph Model (ERGM)** is an exponential family model for $Y_N = [Y_{ij}]_{1 \leq i < j \leq n}$.

$$P_\theta(Y_N) \ = \ \exp\left(\theta^T t_N(Y_N) - \psi_N(\theta)\right) \tag{1}$$

In the above $N = n(n-1)/2$ the dimension of Y and t_N is a vector of sufficient statistics computed from Y. Note that the dependence of V is implicit, through the dependence of N on n. Definition (1) can be generalized by:

1. Allowing for directed graphs, or restricting the possible edges; N will take a value equal to the dimension of Y in each case.
2. Considering that nodes can have features X_i, $i = 1 : n$ which can influence the probabilities of the edges. In this more general case we define

$$P_\theta(Y_N|X_{1:n}) \ = \ \exp\left(\theta^T t_N(Y_N|X_{1:n}) - \psi_N(\theta, X_{1:n})\right) \tag{2}$$

Example 1 (The Erdös-Renyi (ER) model) *For this model, $t_N = \sum y_{ij}$ and there is a single parameter $\theta \in \mathbb{R}$. Thus,*

$$P_\theta(y_N) \ \propto e^{\theta \sum y_{ij}} = \prod_{ij} e^{\theta y_{ij}} \tag{3}$$

In this model, each edge is sampled iid from a Bernoulli with natural parameter θ. The most probable graph is the complete graph if $\theta > 0$ and the empty graph if $\theta < 0$.

Example 2 (The Stochastic Block-Model (SBM)) *The assumption is that the nodes in V are partitioned into K clusters; $X_i \in \{1 : K\}$ denotes the cluster that i belongs to. We have $K(K+1)/2$ sufficient statistics, defined as*

$$t_{kl}(y, x) \ = \ \sum_{x_i = k, x_j = l \ or \ x_i = l, x_j = k} y_{ij} \tag{4}$$

Hence, an edge is sampled independently with a probability that dependins on where its endpoints lie. Note that for known X, the normalization constant for the SBM is tractable.

The ER and the SBM are called **diadic** models, which means that edges are sampled independently conditioned on the features of their endpoints. SBM's have been intensely studied and used in the context of social networks. However, it is generally observed that diadic models, even with richer and more refined node features, do not fit well the real world social-networks. In particular, individuals who have a friend in common tend to be themselves friends with higher probability than a diadic model can predict. In other words, features like triangles and stars have higher frequency in real networks than the frequencies predicted by independent sampling of edges.

This prompted the development of "proper" ERGMs. These are exponential models where the sufficient statistics count other "interesting" features, like triangles, nodes of degree $k = 2, 3, 4 \ldots$, 4 and 5 cliques, in addition to edges.

Example 3 (ERGM with star and triangle features) *Let $t_{1,N}$ count the number of edges, $t_{2,N}$ the number of triangles, $t_{3,N}$ the number of 3-stars (nodes of degree 3), $t_{4,N}$ the number of 4-stars, etc. There is a parameter θ_k for each statistic $t_{k,N}$; when $\theta_k > 0$ the model favors the graphs which contain more of feature k, and when $\theta_k < 0$ then graphs containing fewer of this feature will be more probable.*

$$P_\theta(y_N) \;=\; e^{\theta_1 \# edges + \theta_2 \# triangles + \theta_3 \# 3\text{-}stars + \ldots - \psi_N(\theta_1, \theta_2, \ldots)} \tag{5}$$

Note that these statistics will be dependent on each other, in ways that are fairly complex. Not surprisingly, these models are harder to understand, and this is reflected quantitatively in the fact that the normalization constant Z is generally intractable.

2 Paradigms in modeling with ERGMs

Parameter estimation A first remark on ERGMs is that while the number of parameters p is fixed, even a single sample from P_θ contains N (dependent or independent) random variables y_{ij}. Thus, estimating θ can be done from a single graph, or from a small set of observed graphs. A second feature of ERGMs is that, whenever the Y variables are dependent (for example in a proper ERGM like in Example 3) one would have to estimate the parameters from non-iid data. Sometimes the Y variables are dependent because the node features X, which we assumed are known, are not. This is the case for the SBM. While the model specified by (4) is simple and tractable when the cluster assignments are known, in practice these are not observed, and they have to be estimated simultaneusly with θ, using other, observed, node features and the graph connectivity Y.

Here for simplicity we will only consider ML estimation of ERGMS, but in practice, MAP estimation and full Bayesian estimation of these models (and especially of SBM type models) are widely used, and efficiently implemented for fairly large networks.

Computational issues For most proper ERGMs, ψ is not computable in closed form or tractably. This means that sampling from P_θ and exact inferences under these models are also generally intractable. For example, the simple marginal probability $P_\theta(Y_{ij} = 1|n)$ is intractable in model (5). Thus, Monte Carlo methods for parameter estimation, sampling and other inference tasks (e.g.link prediction) have been developed, and are available in packages like `statnet` [Handcock et al., 2008].

Model interpretation What is the ultimate aim of a model? One possible aim is prediction: if this network doubles in size (i.e $n \to 2n$, what will be its expected properties: number of triangles, diameter, expected degree of a node, number of edges?

Another question is testing: for a new network, we want to know if it was generated from a given model, or in which model class it fits best. Given two networks (usually over different sets of nodes and with different number of nodes), were they generated by the same process?

A third kind of question that a model can answer for us is interpretation of the parameters. Implicit under parameter interpretation is parameter consistency, i.e the assumption that large nets and small nets from the same "source" (e.g. collaboration nets in statistics, needle sharing nets among drug users) have the same parameters, with the parameter estimates concentrating around the "true" parameters $\theta \in \mathbb{R}^p$ as the network sizes grow larger. Unfortunately, research over the last decade has shown that not all ERGMs are consistent.

3 Instability and inconsistency phenomena in ERGMs

3.1 Instability and its consequences

Assume w.l.o.g. that $t_n \in \{0, \dots T_N\}$ *Exercise: Why w.l.o.g ?*. For example the number of edges $t_1 \leq N = n(n-1)/2$, the number of triangles $t_2 \leq n(n-1)(n-2)/6$, the number of 3-stars $t_3 \leq n(n-1)(n-2)(n-3)/24$.

A sufficient statistic t_N is called **stable** iff $\frac{T_N}{N}$ is bounded as $N \to \infty$; otherwise t_n is **unstable**. For example, the number of edges is stable, while the number of triangles is unstable.

Theorem 1 (After [Schweinberger, 2011]) *Assume P_θ is a single parameter model with sufficient statistic t_N unstable.*

1. *We write $y_N \sim y'_N$ if the two random graphs represented by y_N, y'_N differ in the value of a single Y_{ij}. Then $\max_{y_N \sim y'_N} \frac{P_\theta(y_N)}{P_\theta(y'_N)}$ tends to infinity when $N \to \infty$. In other words, P_θ is sensitive to small changes in Y.*
2. *The probability distribution P_θ concentrates on extreme values of the sufficient statistic, i.e. for any θ and any $\epsilon \in (0,1)$, $P_\theta[t_N(Y) \geq (1-\epsilon)T_N] \to 1$, if $\theta > 0$, or $P_\theta[t_N(Y) \leq \epsilon T_N] \to 1$, if $\theta < 0$, when $N \to \infty$.*

In the case of multi-dimensional θ, the current results are not so simple, but they are qualitatively similar.

Instability has negative consequences, known in the field of network modeling as (model) degeneracy. In particular,

- MCMC methods for estimation or sampling will not mix well.
- Since the probability mass of P_θ almost always concentrates on the full or the empty graph, while real networks are usually close to neither (even though they

are sparse), it follows that the estimates of θ will be approximatively 0 for any network that is not very small. (This was demonstrated examples of $n = 32$ and $n = 36$ in [Schweinberger, 2011] and respectively [Fienberg et al., 2009].)

- Unless the estimated parameter $\hat{\theta}$ is very close to 0, sampling for the estimated model $p_{\hat{\theta}}$ will produce graphs that do not ressemble the data (because they will be almost complete or almost empty).
- Hence the model's predictions (of average degree, expected diameter, etc) will not be correct.

Empirical evidence for degeneracy in real world scenarios has accumulated over the last decade. It has been shown that degeneracy in network modeling can be explained/characterized by means of convex geometry and the marginal polytope of the ERGM; [Handcock, 2003] is one of the first papers that studies it theoretically from the point of view of exponential family models. The more recent paper [Fienberg et al., 2009] expands this with a more refined geometric view, complete with vivid illustrations.

3.2 (In)consistency of ERGMs

While the theorem above does not directly prove inconsistency of instable models, it gives very strong hints that this may occur. The characterization of consistency is solved directly in [Shalizi and Rinaldo, 2013][1] which pose the problem from the point of view of projectivity of a model. Essentially, we cannot talk about consistency of a parameter, in the context of dependent data, without assuming that the same parameters can describe both a large amount of data (e.g a large network) and a subset of it (e.g a smaller network). To have consistency, we must have projective exponential family.

The main result of [Shalizi and Rinaldo, 2013] is that projectivity can be characterized by what they call **volume factors**. In brief, denote by $\mathcal{T}_N(y_N) = \{y'_N, t(y'_N) = t(y_N)\}$ the **type** of y_N, i.e. the equivalence class of all other graphs that have the same sufficient statistics as y_N. If $t_N(y_N) = t$ then $\mathcal{T}_N(y_N) = \mathcal{T}_N(t)$. The volume factor $v_N(y_N) = |\mathcal{T}_N(y_N)|$ is the size of this set. It turns out that the way the volume factors grow when y_N is a subset of a larger network y_{M+N} is essential for projectivity, and therefore for consistency. To go with [Shalizi and Rinaldo, 2013] we call the smaller node set A and the larger node set B, and therefore our previous notation becomes $v_A(y_N) = |\mathcal{T}_N(y_N)|$, while for the larger network the same quantity is $v_B(y_{M+N})$. Denote $y_A \equiv y_N$ and $y_B \equiv y_{M+N}$, and $y_{B\setminus A} = y_{M+N} \setminus y_A$ (hope it makes sense).

Now let $t_A(y_N) = t$, $t_B(y_{N+M}) = t + \delta$ (assuming that t counts some features, like triangles, so the count can only grow if the network has more nodes). Define the **conditional volume factor** $v_{B\setminus A|A}(y_A, \delta) = |\{y'_{B\setminus A} \text{ so that } y'{}_A = y_A, t_(y'_B) = t + \delta\}|$; in other words, $v_{B\setminus A|A}$ is the number of all networks equivalent to B whose restriction to A is identical to y_A.

Definition 1 *The sufficient statistic t* **has separable increments** *iff for all set of nodes B, for all $A \subset B$, and for all networks y_A, the range of possible increments $\delta = t_B(y_B) - t_A(y_A)$ is the same, and the conditional volume factor does not depend on y_A, i.e. $v_{B\setminus A|A}(\delta, y_A)$ depends only on δ.*

[1] Note: this is not so long ago!

Theorem 2 ([Shalizi and Rinaldo, 2013]) *The exponential family P_θ is projective iff the sufficient statistics have separable increments.*

For example, when a set of nodes A, with a network y_A on them, is increased with $B \setminus A$, the number of edges in examples 1, and 2, will increase by amounts that depend only on properties of A and B, but not on what edges appear in y_A. However, the number of triangles in $B \setminus A$ will depend on the configuration of edges in y_A, and in particular on the number of triangles in y_A. Hence, diadic models are projective, but ERGMs (that count triangles and stars) are not.

References

[Fienberg et al., 2009] Fienberg, S. E., Rinaldo, A., and Zhou, Y. (2009). On the geometry of discrete exponential families with application to exponential random graph models. Technical report, Carnegie Mellon University.

[Handcock, 2003] Handcock, M. S. (2003). Assessing degeneracy in statistical models of social networks. CSSS Working Paper 39, University of Washington.

[Handcock et al., 2008] Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., and Morris, M. (2008). statnet: Software tools for the representation, visualization, analysis and simulation of network data. *Journal of Statistical Software*.

[Schweinberger, 2011] Schweinberger, M. (2011). Instability, sensitivity, and degeneracy of discrete exponential families. *Journal of the American Statistical Association*, 106(496):1361–1370.

[Shalizi and Rinaldo, 2013] Shalizi, C. and Rinaldo, A. (2013). Consistency under sampling of exponential random graph models. *The Annals of Statistics*.