Exponential-family Random Network Models (ERNM)

Ian Fellows

UCLA

January 9, 2012

The landscape

- Random graphs == Random connections, Fixed nodal attributes
- Gibbs/Markov random fields == Fixed connections, Random nodal attributes
- ERNM == Random connections, Random nodal attributes

ERNM Formulation

Let Y be an n by n matrix who's entries $Y_{i,j}$ indicate whether subject i and j are connected, where n is the size of the population. Further let X be a $n \times q$ matrix of nodal variates. We define the *network* to be the random variable (Y, X). Then a joint exponential family model for the network may be written as:

$$P(X = x, Y = y|\eta) = \frac{1}{c(\eta)} e^{\eta h(x,y)}, \qquad (x, y) \in \mathcal{N} \qquad (1)$$

Uninteresting Example: Separable Models

Suppose that h is composed such that the model can be expressed as

$$P(X = x, Y = y | \eta_1, \eta_2) = \frac{1}{c(\eta_1, \eta_2)} e^{\eta_1 h_1(x) + \eta_2 h_2(y)} \qquad (x, y) \in \mathcal{N}. \tag{2}$$

Then

$$\begin{aligned} P(X = x | \eta_1) &= \frac{1}{c_1(\eta_1)} e^{\eta_1 h_1(x)} \\ P(Y = y | \eta_2) &= \frac{1}{c_2(\eta_2)} e^{\eta_2 h_2(y)}. \end{aligned}$$

Pathological Example: Ising as a Joint Model

Joint model:

$$P(X = x, Y = y | \eta_1, \eta_2) \propto e^{\eta_1 \sum_i \sum_j y_{i,j} + \eta_2 \sum_i \sum_j x_i y_{i,j} x_j}.$$

With conditional distributions being:

$$\begin{aligned} P(Y_{i,j} = y_{i,j} | X = x, \eta_1, \eta_2) &\propto e^{\eta_1 y_{i,j} + \eta_2 x_i y_{i,j} x_j} \\ P(X = x | Y = y, \eta_2) &\propto e^{\eta_2 \sum_i \sum_j x_i y_{i,j} x_j} \end{aligned}$$

Pathological Example: Ising as a Joint Model

Degeneracy: Oh My!!!

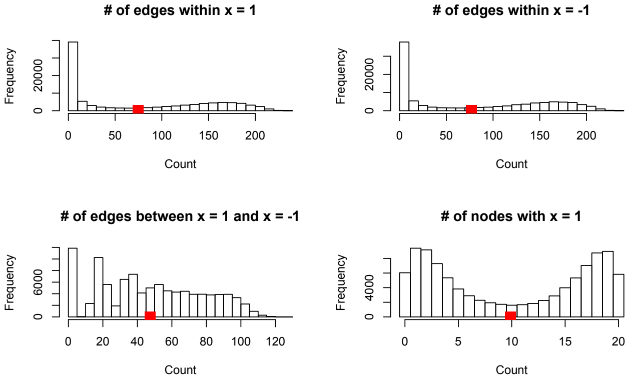

Figure: 100,000 draws from an Ising Joint Model with $\eta_1 = 0$ and $\eta_2 = 0.13$. Mean values are marked in red.

Dispare

Oh well, better give up.

Hope

But wait, is there a better measure of homophily which doesn't display degeneracy?

...

... 6 months pass ....

Regularized Homophily

$$reg\_homophily(k,l) = \sum_{i:x_i=k} \sqrt{d_{i,l}} - E_{binom}(\sqrt{d_{i,l}}),$$

where $d_{i,l}$ is the number of edges connecting node i to nodes in group l, and $E_{binom}(\sqrt{d_{i,l}})$ is the expectation of the square root of a binomial variable, with probability equal to the proportion of nodes in group l and size equal to the out-degree of node i.

Logistic Regression in Network Data

$$P(Z=z,X=x,Y=y|\eta,\beta,\lambda)=\frac{1}{c(\beta,\eta,\lambda)}e^{zx\beta\cdot+\eta h(x,y)+\lambda g(z,y)}. \tag{3}$$

$$P(z_i=1|z_{-i},x_i,Y=y,\beta,\lambda)=\frac{e^{x_i\beta}}{e^{\lambda[g(z^-,y)-g(z^+,y)]}+e^{x_i\beta}}. \tag{4}$$

where z_{-i} represents the set of z not including z_i, z^+ represents z where $z_i=1$, z^- is z where $z_i=0$,and x_i represents the ith row of X.

A Super-population Model for an Add Health High School

	η	Std. Error	Z	p-value
Mean Degree	-167.90	8.51	-19.73	<0.001
Log Variance of Degree	22.18	10.01	2.22	0.027
Degree = 0	3.91	0.47	8.28	<0.001
Degree = 1	2.20	0.38	5.86	<0.001
Degree = 2	0.73	0.35	2.05	0.041
Grade = 9	0.88	0.78	1.13	0.258
Grade = 10	1.74	0.92	1.89	0.058
Grade = 11	2.53	0.79	3.20	0.001
Within Grade Homophily	3.97	0.47	8.44	<0.001
+1 Grade Homophily	0.50	0.33	1.54	0.125
+2 Grade Homophily	-1.07	0.27	-4.03	<0.001
+3 Grade Homophily	-0.59	0.40	-1.47	0.143

Table: ERNM Model with Standard Errors Based on the Fisher Information

A Super-population Model for an Add Health High School

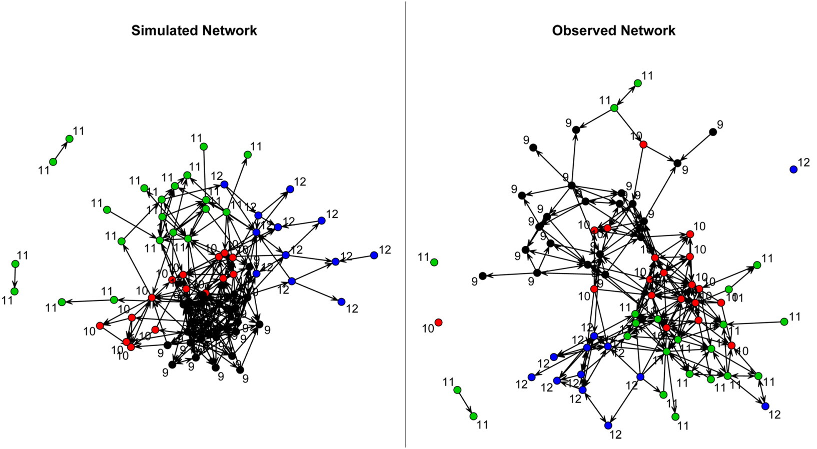

Figure: Model-Based Simulated High School

A Super-population Model for an Add Health High School

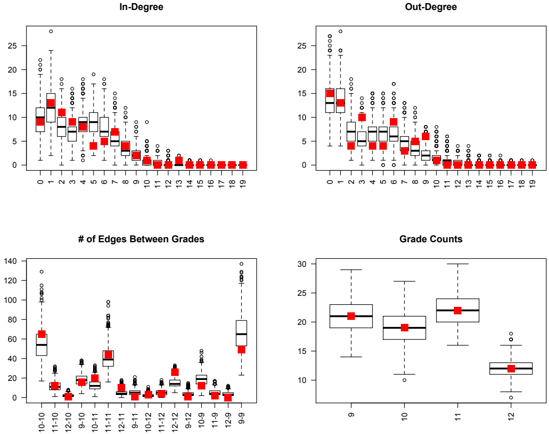

Figure: Model Diagnostics

Logistic Regression on Substance Use: Naive model

	β	Std. Error	Z	p-value
Intercept	-1.70	0.44	-3.84	<0.001
Male	1.18	0.57	2.09	0.037

Table: Simple Logistic Regression Model Ignoring Network Structure

Logistic Regression on Substance Use: ERNM model

		Bootstrap	Asymptotic		
	η	Std. Error	Std. Error	Z	p-value
Mean Degree	-164.18	7.86	8.07	-20.36	<0.001
Log Variance of Degree	20.35	8.85	9.07	2.24	0.025
Degree 0	4.01	0.45	0.44	9.12	<0.001
Degree 1	2.25	0.37	0.35	6.44	<0.001
Degree 2	0.74	0.36	0.35	2.08	0.038
Grade Homophily	3.85	0.46	0.46	8.41	<0.001
+1 Grade Homophily	0.45	0.33	0.33	1.39	0.166
+2 Grade Homophily	-1.14	0.28	0.25	-4.50	<0.001
+3 Grade Homophily	-0.58	0.39	0.38	-1.52	0.129
Sex Homophily	0.98	0.28	0.27	3.56	<0.001
Substance Homophily	0.88	0.25	0.26	3.44	<0.001
Intercept	-1.79	0.49	0.43	-4.11	<0.001
Male	0.94	0.56	0.52	1.81	0.070

Table: ERNM Model Inference

Logistic Regression on Substance Use: homophily diagnostics

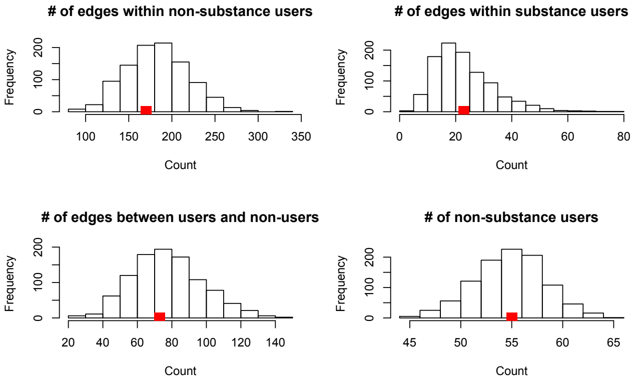

Figure: Substance Use Homophily Diagnostics. The values of the observed statistics are marked in red.

Conclusion

ERNM is a framework for inference about networks, including both the graph and the nodal characteristics.