Small team statistics

George G. Vega Yon
Kayla de la Haye

October 1, 2018

Exponential Random Graph Models

Formals

The distribution of $\mathbf{Y}$ can be parameterized in the form

$$\Pr(\mathbf{Y}=\mathbf{y}|\theta,\mathcal{Y}) = \frac{\exp\theta^{\mathsf{T}}\mathbf{g}(\mathbf{y})}{\kappa(\theta,\mathcal{Y})},\quad \mathbf{y}\in\mathcal{Y} \tag{1}$$

Where $\theta\in\Omega\subset\mathbb{R}^q$ is the vector of model coefficients and $\mathbf{g}(\mathbf{y})$ is a q-vector of statistics based on the adjacency matrix $\mathbf{y}$.

- Model (1) may be expanded by replacing $\mathbf{g}(\mathbf{y})$ with $\mathbf{g}(\mathbf{y},\mathbf{X})$ to allow for additional covariate information $\mathbf{X}$ about the network. The denominator,

$$\kappa(\theta,\mathcal{Y}) = \sum_{\mathbf{z}\in\mathcal{Y}}\exp\theta^{\mathsf{T}}\mathbf{g}(\mathbf{z})$$

- Is the normalizing factor that ensures that equation (1) is a legitimate probability distribution.
- Even after fixing $\mathcal{Y}$ to be all the networks that have size n, the size of $\mathcal{Y}$ makes this type of models hard to estimate as there are $N=2^{n(n-1)}$ possible networks! (Hunter et al. 2008)

How does ERGMs look like (in R at least)

```r
network ~ edges + nodematch("hispanic") + nodematch("female") +
  mutual +  esp(0:3) +  idegree(0:10)
```

Here we are controlling for:

- `edges`: Edge count,
- `nodematch(hispanic)`: number of homophilic edges on race,
- `nodematch(female)`: number of homophilic edges on gender,
- `mutual`: number of reciprocal edges,
- `esp(0:3)`: number of shared parterns (0 to 3), and
- `indegree(0:10)`: indegree distribution (fixed effects for values 0 to 10)

(See Hunter et al. 2008).

Example with accuracy

For each team T, we defined the following statistic:

$$A_T \equiv 1 - \frac{1}{n(n-1)} \sum_{i \in N} H(G_i, G_T)$$

Where H is the hamming distance, G_i is i's Cognitive Social Structure, and G_T is the true network.

The statistic is normalized so that it lies wihin 0 and 1, with 0 been complete missmatch, and 1 perfect match.

Simulation process

For each set of experiments, generate N teams by doing:

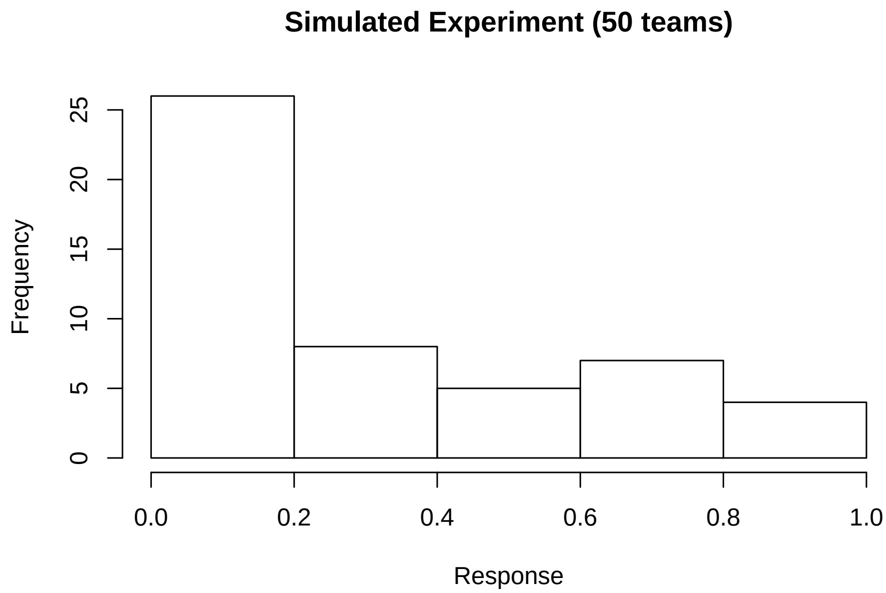
1. Draw a random graph of size n_i from a bernoulli distribution with parameter p_i, call it G_i.
2. Generate n_i other graphs by permuting G_i with different levels of accuracy a_{ij}
3. Generate $Y_i \sim \text{Beta}\left(\exp\left(\theta^{\text{t}} X_i\right), 1.5\right)$, where X_i is a vector of team level statistics, including $\hat{a}_i = n_i^{-1} \sum_j \hat{a}_{ij}$, the average level of accuracy. The resulting value Y_i will be between 0 and 1.

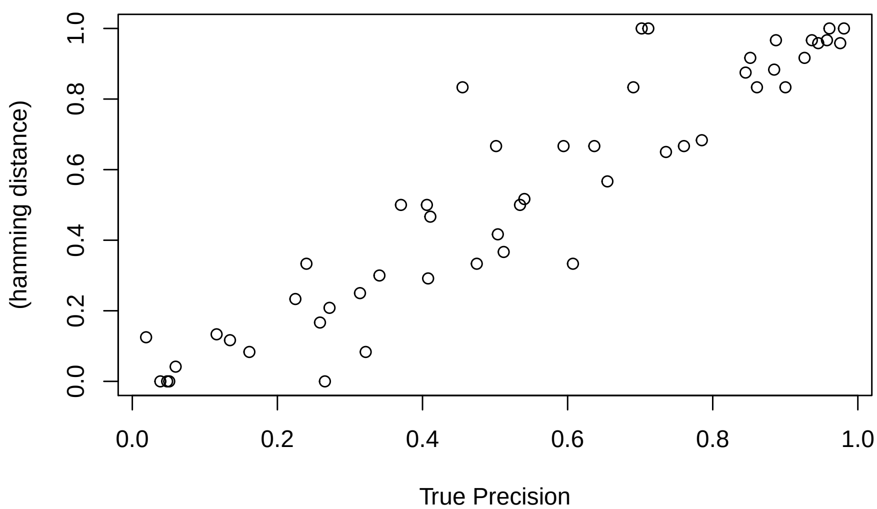
Once all N teams have been simulated, estimate the model using MLE

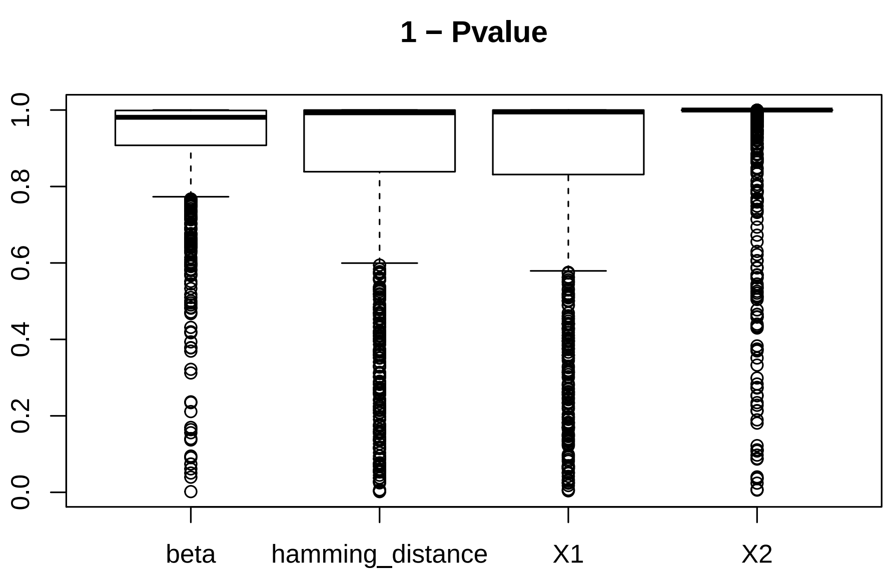
Monte Carlo Experiments

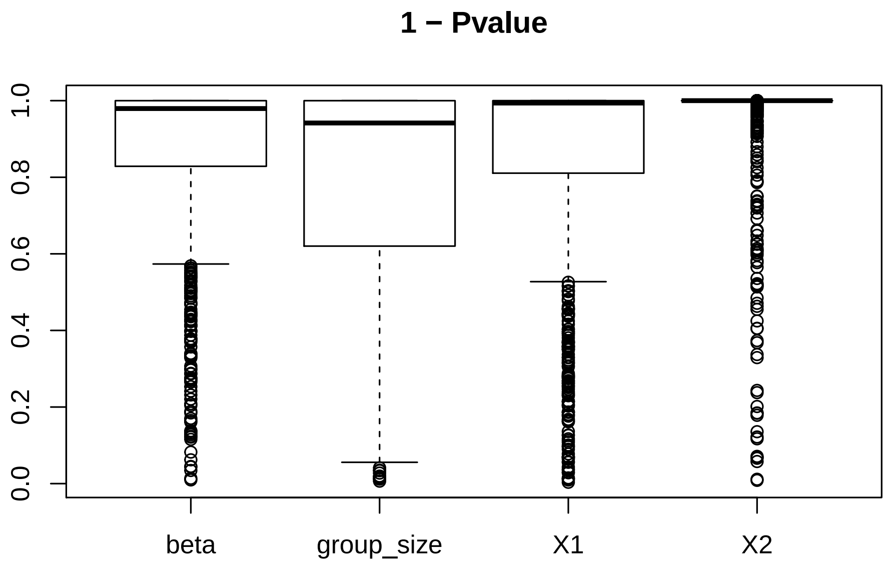
Table 1: Estimates from simulated data

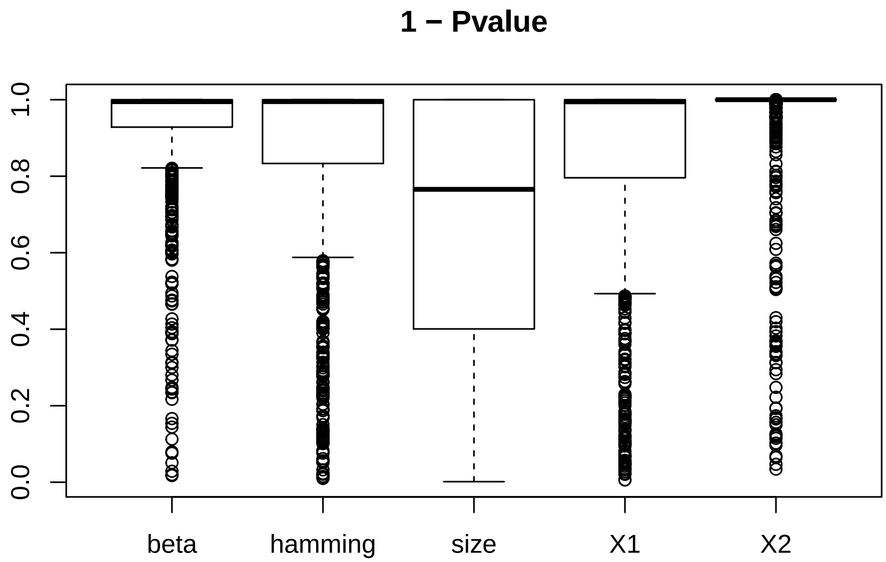
	Estimate	Std. Error
beta	-0.2431621	0.2106463
h	-0.0614361	0.3242206
X1	-2.4709608	0.4129149
X2	0.6836943	0.1096756


Simulated Experiment (50 teams)
Frequency
0
5
10
15
20
25
0.0
0.2
0.4
0.6
0.8
1.0
Response


(hamming distance)
0.0
0.2
0.4
0.6
0.8
1.0
True Precision


1 − Pvalue
0.0
0.2
0.4
0.6
0.8
1.0
beta
hamming_distance
X1
X2





References

Hunter, David R., Mark S. Handcock, Carter T. Butts, Steven M. Goodreau, and Martina Morris. 2008. "ergm : A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks." *Journal of Statistical Software* 24 (3). https://doi.org/10.18637/jss.v024.i03.