What drives social networks?
A gentle introduction to exponential random graph models (with a focus on small networks)

George G Vega Yon

Department of Preventive Medicine

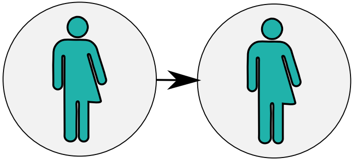
LAERUG
June 10, 2019

Social networks

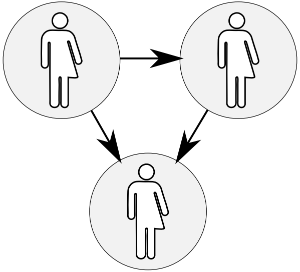
Figure 1: Friendship network of a UK university faculty. Source: **igraphdata** R package (Csardi, 2015). Figure drawn using the R package **netplot** (yours truly, https://github.com/usccana/netplot)

Exponential Family Random Graph Models (ERGMs)

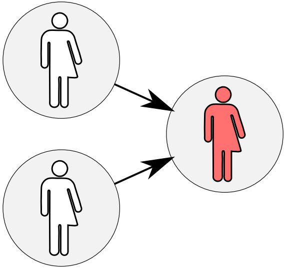
Why are you and I are *[blank]* ? (friends, collaborators, etc.)



Homophily



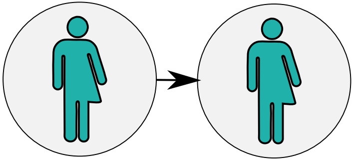
Transitive Triad



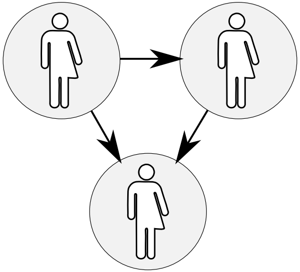
Popularity

Exponential Family Random Graph Models (ERGMs)

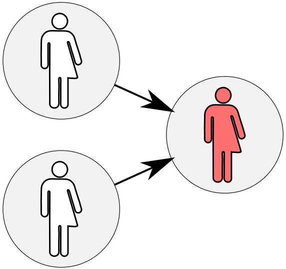
Why are you and I are *[blank]* ? (friends, collaborators, etc.)



Homophily



Transitive Triad



Popularity

Let's build a model for this!

ERGMs from scratch

We need to build a probability function…

ERGMs from scratch

We need to build a probability function...

- First, we will focus on counts: “How many edges?”, “How many homophilic ties?”. We will call them “sufficient statistics”

  $\#edges, \#homophilic\ ties, \dots$

ERGMs from scratch

We need to build a probability function...

- First, we will focus on counts: "How many edges?", "How many homophilic ties?". We will call them "sufficient statistics"

  $\#\textit{edges}, \#\textit{homophilic ties}, \dots$

- As we do in ~~life~~ statistics, let's assume it is an additive model (we add stuff up), in a weighted fashion (i.e. we have model parameters!)

  $\theta_1 \times \#\textit{edges} + \theta_2 \times \#\textit{homophilic ties} + \dots$

ERGMs from scratch

We need to build a probability function...

- First, we will focus on counts: "How many edges?", "How many homophilic ties?". We will call them "sufficient statistics"

  $\#edges, \#homophilic\ ties, \dots$

- As we do in ~~life~~ statistics, let's assume it is an additive model (we add stuff up), in a weighted fashion (i.e. we have model parameters!)

  $\theta_1 \times \#edges + \theta_2 \times \#homophilic\ ties + \dots$

- And since we like things to be positive... we just exponentiate it!

  $\exp\left\{\theta_1 \times \#edges + \theta_2 \times \#homophilic\ ties + \dots\right\}$

ERGMs from scratch

We need to build a probability function...

- First, we will focus on counts: "How many edges?", "How many homophilic ties?". We will call them "sufficient statistics"

  $\#edges, \#homophilic\ ties, \dots$

- As we do in ~~life~~ statistics, let's assume it is an additive model (we add stuff up), in a weighted fashion (i.e. we have model parameters!)

  $\theta_1 \times \#edges + \theta_2 \times \#homophilic\ ties + \dots$

- And since we like things to be positive... we just exponentiate it!

  $\exp\left\{\theta_1 \times \#edges + \theta_2 \times \#homophilic\ ties + \dots\right\}$

- Finally, as probabilities should add up to 1, we will divide the thing by the sum of all possible cases (the "normalizing constant")

  $\frac{\exp\{\theta_1 \times \#edges + \theta_2 \times \#homophilic\ ties + \dots\}}{\sum \exp\{\dots\}}$

ERGMs from scratch

We need to build a probability function...

- First, we will focus on counts: "How many edges?", "How many homophilic ties?". We will call them "sufficient statistics"

  $$\#\textit{edges}, \#\textit{homophilic ties}, \dots$$

- As we do in ~~life~~ statistics, let's assume it is an additive model (we add stuff up), in a weighted fashion (i.e. we have model parameters!)

  $$\theta_1 \times \#\textit{edges} + \theta_2 \times \#\textit{homophilic ties} + \dots$$

- And since we like things to be positive... we just exponentiate it!

  $$\exp\left\{\theta_1 \times \#\textit{edges} + \theta_2 \times \#\textit{homophilic ties} + \dots\right\}$$

- Finally, as probabilities should add up to 1, we will divide the thing by the sum of all possible cases (the "normalizing constant")

  $$\frac{\exp\{\theta_1 \times \#\textit{edges} + \theta_2 \times \#\textit{homophilic ties} + \dots\}}{\sum \exp\{\dots\}}$$

You got yourself an ERGM!

ERGMs... the *lingua franca* of SNA

- Seeks to answer the question: What local social structures gave origin to a given observed graph?

ERGMs... the *lingua franca* of SNA

- Seeks to answer the question: What local social structures gave origin to a given observed graph?
- The model is centered around a vector of **sufficient statistics** $s()$, and is operationalized as:

$$\Pr\left(\mathbf{G}=\mathbf{g} \mid \theta, \mathbf{X}\right) = \frac{\exp\left\{\theta^{\mathbf{t}} s\left(\mathbf{g}, \mathbf{X}\right)\right\}}{\kappa\left(\theta, \mathbf{X}\right)}, \quad \forall \mathbf{g} \in \mathcal{G} \tag{1}$$

Where $\kappa\left(\theta, \mathbf{X}\right)$ is the normalizing constant and equals $\sum_{\mathbf{g}' \in \mathcal{G}} \exp\left\{\theta^{\mathbf{t}} s\left(\mathbf{g}', \mathbf{X}\right)\right\}$.

ERGMs. . . the *lingua franca* of SNA

- Seeks to answer the question: What local social structures gave origin to a given observed graph?
- The model is centered around a vector of **sufficient statistics** $s\left(\right)$, and is operationalized as:

$$\Pr\left(\mathbf{G}=\mathbf{g}\mid\theta,\mathbf{X}\right)=\frac{\exp\left\{\theta^{\mathbf{t}}s\left(\mathbf{g},\mathbf{X}\right)\right\}}{\kappa\left(\theta,\mathbf{X}\right)},\quad\forall\mathbf{g}\in\mathcal{G}\tag{1}$$

  Where $\kappa\left(\theta,\mathbf{X}\right)$ is the normalizing constant and equals $\sum_{\mathbf{g}'\in\mathcal{G}}\exp\left\{\theta^{\mathbf{t}}s\left(\mathbf{g}',\mathbf{X}\right)\right\}$.
- The set of sufficient statistics reflects social and psychological mechanisms that are hypothesized to drive the network structure. Figure 2 shows some examples of values in $s\left(\right)$.

ERGMs... the *lingua franca* of SNA

- Seeks to answer the question: What local social structures gave origin to a given observed graph?
- The model is centered around a vector of **sufficient statistics** $s()$, and is operationalized as:

$$\Pr\left(\mathbf{G}=\mathbf{g} \mid \theta, \mathbf{X}\right) = \frac{\exp\left\{\theta^{\mathbf{t}} s\left(\mathbf{g},\mathbf{X}\right)\right\}}{\kappa\left(\theta,\mathbf{X}\right)},\quad \forall \mathbf{g}\in\mathcal{G} \tag{1}$$

  Where $\kappa\left(\theta,\mathbf{X}\right)$ is the normalizing constant and equals $\sum_{\mathbf{g}'\in\mathcal{G}}\exp\left\{\theta^{\mathbf{t}} s\left(\mathbf{g}',\mathbf{X}\right)\right\}$.

- The set of sufficient statistics reflects social and psychological mechanisms that are hypothesized to drive the network structure. Figure 2 shows some examples of values in $s()$.
- In the case of directed networks, $\mathcal{G}$ has $2^{n(n-1)}$ terms.

ERGMs... the *lingua franca* of SNA

- Seeks to answer the question: What local social structures gave origin to a given observed graph?
- The model is centered around a vector of **sufficient statistics** $s\left(\right)$, and is operationalized as:

$$\Pr\left(\mathbf{G} = \mathbf{g} \mid \theta, \mathbf{X}\right) = \frac{\exp\left\{\theta^{\mathbf{t}} s\left(\mathbf{g}, \mathbf{X}\right)\right\}}{\kappa\left(\theta, \mathbf{X}\right)}, \quad \forall \mathbf{g} \in \mathcal{G} \tag{1}$$

  Where $\kappa\left(\theta, \mathbf{X}\right)$ is the normalizing constant and equals $\sum_{\mathbf{g}' \in \mathcal{G}} \exp\left\{\theta^{\mathbf{t}} s\left(\mathbf{g}', \mathbf{X}\right)\right\}$.

- The set of sufficient statistics reflects social and psychological mechanisms that are hypothesized to drive the network structure. Figure 2 shows some examples of values in $s\left(\right)$.
- In the case of directed networks, $\mathcal{G}$ has $2^{n(n-1)}$ terms.
- See Wasserman, Pattison, Robins, Snijders, Handcock, Butts, and others.

Structures

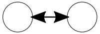
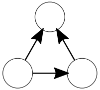
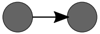
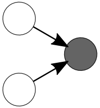
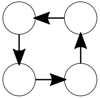
Representation	Description
	Mutual Ties (Reciprocity) $\sum_{i \neq j} y_{ij} y_{ji}$
	Transitive Triad (Balance) $\sum_{i \neq j \neq k} y_{ij} y_{jk} y_{ik}$
	Homophily $\sum_{i \neq j} y_{ij} \mathbf{1}(x_i = x_j)$
	Covariate Effect for Incoming Ties $\sum_{i \neq j} y_{ij} x_j$
	Four Cycle $\sum_{i \neq j \neq k \neq l} y_{ij} y_{jk} y_{kl} y_{li}$

Figure 2: Besides of the common edge count statistic (number of ties in a graph), ERGMs allow measuring other more complex structures that can be captured as sufficient statistics.

References

Allaire, JJ, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. 2018. Rmarkdown: Dynamic Documents for R. https://rmarkdown.rstudio.com.

Csardi, Gabor. 2015. Igraphdata: A Collection of Network Data Sets for the 'Igraph' Package. https://CRAN.R-project.org/package=igraphdata.

Handcock, Mark, Peng Wang, Garry Robins, Tom Snijders, and Philippa Pattison. 2006. "Recent developments in exponential random graph (p*) models for social networks." Social Networks 29 (2): 192–215. https://doi.org/10.1016/j.socnet.2006.08.003.

R Core Team. 2018. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wasserman, Stanley, and Philippa Pattison. 1996. "Logit models and logistic regressions for social networks: I. An introduction to Markov graphs andp." Psychometrika 61 (3): 401–25. https://doi.org/10.1007/BF02294547.

Xie, Yihui. 2018. Knitr: A General-Purpose Package for Dynamic Report Generation in R. https://yihui.name/knitr/.