Understanding Exponential Random Graph Models (ERGM)

Shouyang Wang

Background

Much work in classical Social Network Analysis (SNA) has been concerned with measuring and visualizing networks. Many of the standard SNA measures center around the descriptive properties of observed relational data - this naturally leads to the measurement of the structural properties and the importance of an individual's position in the structure.

Wasserman and Faust spent a significant portion of their book (1994) addressing questions related to prestige (i.e., centrality and structural imbalance), cohesive subgroups (i.e., k-cores and cliques), and role/positional analysis (i.e.blockmodels and relational algebras). In the last chapter of the book, they note that

> *"(Wasserman and Faust) believe that statistical models will be a major focus for the continued development and expansion of network methods. Clearly, scientific understanding is advanced when we can test propositions about network properties rather than simply relying on descriptive statements. Great steps have been made in statistical models for dyads, including p1 and its relatives. . . ."*

P*, the successor of the p1 Markov random graph models (Holland & Leinhardt, 1981; Frank & Strauss, 1986), is also known in the SNA community as the Exponential Random Graph Models - ERGMs (Krivitsky, Morris, Handcock, Butts, Hunter, Goodreau, Klumb de-Moll, 2021; Butts, 2008; Snijders, Pattison, Robins & Handcock, 2005; Contractor, Wasserman, & Faust, 2006; Robins, Elliott, & Pattison, 2001).

Why should social scientists study ERGM?

Several reasons motivate researchers to apply ERGM in social science studies. In numerous cases, social data are relational (i.e., friendships, kinship ties, love, conflicts, social media following). Some random network models (i.e., Erdős–Rényi) lack features observed in human social networks, such as clustering. Also, many classical models and regression techniques strongly assume that the observations are independent, leading to the issue that we model individuals without considering the social relationships and context.

Karl Marx said, "it is not the consciousness of men that determines their being, but, on the contrary, their social being that determines their consciousness." In other words, the social environment (that an individual is embedded in) directly impacts how the individual thinks or behaves. From a social and behavioral science perspective, it is unrealistic to model social actors as isolated entities. For instance, if A is friends with B, and B is friends with C. It's reasonable to model the probability of A being friends with C NOT being the same as A being friends with a random person.

Furthermore, in social science studies, researchers often have more than one compelling theory that could explain an observed phenomenon. It is insightful to understand the significant local social processes (i.e., reciprocity, transitivity, preferential attachment) that result in the formation of a network. An ERG model, just like many other regression models, can be used to fit many covariates simultaneously and test one theory against another for hypothesis testing purposes. The restrictions on the covariates are loose: the characteristics of social actors (i.e., nodal attributes like ethnicity (discrete data) or height (continuous data) of an individual) and network statistics (i.e., graph theory measurements that have implicit social meanings: the number of reciprocated ties, triangles, stars, or user-defined components/subgraphs) can all be used to introduce local features and dependencies to model the formation of an observed network.

Lastly, in economic and political science settings, rather than concerning the entire network at the global level, it is often valuable to understand the probability of forming a tie between two social actors. For instance, knowing the likelihood of nation A trading with nation B helps policymakers adjust and react in international politics. The generic form of an ERGM can be re-expressed as a function of change statistics (i.e., how a given network statistics change by adding or removing a social tie) at the dyadic level. The flexibility and simplicity of ERGM make it a promising and more realistic network modeling approach in social science research.

Theoretical Foundations

At first sight, an ERGM model can be daunting, not necessarily because the logic is challenging to comprehend but because the notations are intimidating.

In reality, ERGMs are relatively effortless to apply with the help of the R ergm package (Statnet Development Team, 2021). The output of calling the ergm function and interpreting results is similar to that of conducting other classical regression models and hypothesis testing in R. Understanding basic statistical techniques, such as coding and analyzing a simple logistic regression, is often sufficient to apply an ERGM and get concrete results.

Intuitively, the observed network (i.e., the relational data collected) is perceived as one potential outcome derived from a sample space of networks. For instance, given a friendship network in a classroom. One possible outcome is that people are all strangers. An alternative outcome is that everyone is friends with each other. Between those two extreme outcomes, there are many different potential network configurations, and one of those theoretical networks has the exact configuration as the observed network. We ask, what is the probability of observing this network?

One trivial approach is that the probability of observing any network configuration is the same for all theoretical networks in the sample space. Clearly, this does not help gain any insights. It is sensible to argue that not all configurations are equally probable.

$$P(Y = y) = \frac{\exp\left(\theta' g(y)\right)}{k(\theta)}$$

where

- Y is the random variable for the state of the network (with realization y),
- $g(y)$ is a vector of model statistics for network y
- θ is the vector of coefficients for those statistics, and
- $k(\theta)$ represents the quantity in the numerator summed over all possible networks (typically constrained to be all networks with the same node set as y).

The left-hand side denotes the probability of observing a network. This probability is a function of network statistics and coefficients/parameters. The denominator term on the right-hand side is the sum of all possible networks. Some refer to the denominator term as the normalizing factor (i.e., we reduce the probability function to the probability density function with a total probability of 1). In statistical modeling, the exponential family is heavily studied and includes many of the most common distributions (i.e., Exponential, Dirichlet, Poisson, Bernoulli). The exponential term also ensures that we always get a positive probability (i.e., e^x is always positive).

Instead of assigning the same probability to all theoretical networks in the sample space, we want to pick the coefficients that maximize the probability of observing the observed network. In statistics, this is called Maximum Likelihood Estimation. Based on the generic ERGM equation, the chance of observing a graph depends on how many configurations of interests are present (i.e., it could be the number of reciprocated ties, stars, triangles, etc.). The parameters tell us the significance of each configuration.

Further, often in a more practical sense, we are interested in probabilities in terms of a single tie at the dyadic level. We re-express the general form in terms of the conditional log-odds of a tie. On the left-hand side is the log of the ratio of a probability. The probability of forming a tie can be obtained by applying the inverse logit function (e.g., a logit of 0 corresponds to a probability of 50%).

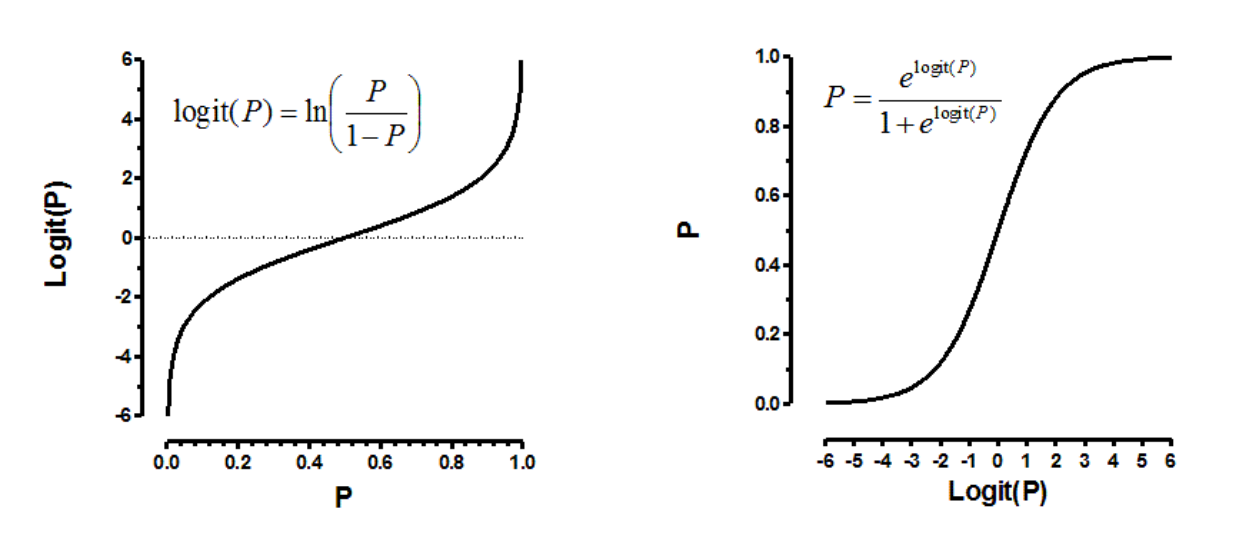


Figure 1: logit function (left) and sigmoidal logistic function (right)

We express the conditional log-odds of a tie in terms of change statistics, and θ is the term's contribution to the log-odds of an individual tie, conditional on all other dyads.

$$\text{logit}\left(Y_{ij} = 1 \mid y_{ij}^{c}\right) = \theta'\delta\left(y_{ij}\right)$$

where

- Y_{ij} is the random variable for the state of the actor pair i, j (with realization y_{ij}), and
- y_{ij}^{c} signifies the complement of y_{ij}, i.e. all dyads in the network other than y_{ij}.
- $\delta\left(y_{ij}\right)$ is a vector of the "change statistics" for each model term. The change statistic records how the $g(y)$ term changes if the y_{ij} tie is toggled on or off. So:

$$\delta\left(y_{ij}\right) = g\left(y_{ij}^{+}\right) - g\left(y_{ij}^{-}\right)$$

where

- $y_{ij}^{+}$is defined as y_{ij}^{c} along with y_{ij} set to 1 , and
- $y_{ij}^{-}$is defined as y_{ij}^{c} along with y_{ij} set to 0.

Application and Implementation using R Programming

Network Data Source: Visiting ties among families in San Juan Sur, Costa Rica, 1948

Network Description:

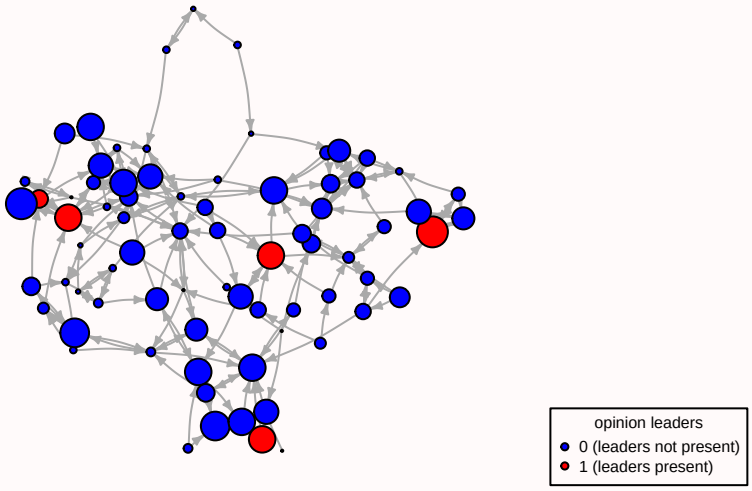
> *"In 1948, American sociologists executed a large field study in the Turrialba region, which is a rural area in Costa Rica (Latin America). They were interested in the impact of formal and informal social systems on social change. Among other things, they investigated visiting relations*

between families living in haciendas (farms) in a neighborhood called Attiro. This is the network for a village in the area called San Juan Sur. The network of ties is a simple directed graph: each arc represents"frequent visits" from one family to another. The exact number of visits was not recorded. The investigators proposed an ethnographic classification of the families into six family-friendship groupings on substantive criteria. In rural areas where there is little opportunity to move up and down the social ladder social groups are usually based on family relations. This is the vertex attribute grouping."

```
pacman::p_load(sna, ergm, igraph, intergraph, dplyr)
# visualize the network
network_viz <- function(network) {
  G = asIgraph(network)
  gen_colors <- c("blue", 'red')
  V(G)$leader.type = ifelse(V(G)$leaders == 0, 1, 2)
  V(G)$color <- gen_colors[V(G)$leader.type]
  plot(G, edge.arrow.size=0.3, edge.curved=0.1,
  vertex.size = V(G)$status, vertex.label=NA)
  # add a legend
  legend("bottomright", inset=c(0,0),
  legend = c("0 (leaders not present)", "1 (leaders present)"),
  pt.bg = gen_colors,
  title = "opinion leaders", pch = 21, cex = 0.5)

}

(SanJuanSurVisiting) %>% network_viz
```



Our research questions are

- Do families in the same group tend to visit one another more frequently than those in different groups?
- Do visiting behaviors tend to be reciprocal?

```
rand_graph <- ergm(SanJuanSurVisiting ~ edges # density
                            + nodematch("grouping", diff = TRUE) # group homophily
                            + nodefactor('grouping') # control skewed distributions
                            + mutual, # reciprocity
                            control = control.ergm(seed = 0))
predict(rand_graph) %>% head()
```

```
##   tail head           p
## 1   33   64 0.012233819
## 2   70   22 0.006436296
## 3   49   30 0.004349164
## 4   26   41 0.011954609
## 5   67   46 0.004349164
## 6   38   72 0.007274199
```

```
rand_graph %>% summary()
```

```
## Call:
## ergm(formula = SanJuanSurVisiting ~ edges + nodematch("grouping",
##     diff = TRUE) + nodefactor("grouping") + mutual, control = control.ergm(seed = 0))
##
## Monte Carlo Maximum Likelihood Results:
##
##                       Estimate Std. Error MCMC % z value Pr(>|z|)
## edges                 -5.24200    0.64105      0  -8.177  < 1e-04 ***
## nodematch.grouping.1   3.34698    0.68741      0   4.869  < 1e-04 ***
## nodematch.grouping.2   2.42492    0.71597      0   3.387 0.000707 ***
## nodematch.grouping.3   2.44701    0.63807      0   3.835 0.000126 ***
## nodematch.grouping.4   1.78949    0.48102      0   3.720 0.000199 ***
## nodematch.grouping.5   2.71704    0.84210      0   3.227 0.001253 **
## nodematch.grouping.6   3.24227    0.79702      0   4.068  < 1e-04 ***
## nodematch.grouping.7   3.47463    0.75876      0   4.579  < 1e-04 ***
## nodematch.grouping.8   3.68859    0.63810      0   5.781  < 1e-04 ***
## nodematch.grouping.9   3.72482    0.79267      0   4.699  < 1e-04 ***
## nodefactor.grouping.2  0.48241    0.45550      0   1.059 0.289563
## nodefactor.grouping.3  0.60563    0.44207      0   1.370 0.170692
## nodefactor.grouping.4  0.73905    0.41506      0   1.781 0.074979 .
## nodefactor.grouping.5  0.62714    0.47140      0   1.330 0.183397
## nodefactor.grouping.6  0.24513    0.49643      0   0.494 0.621464
## nodefactor.grouping.7  0.08834    0.48214      0   0.183 0.854617
## nodefactor.grouping.8 -0.27975    0.45223      0  -0.619 0.536171
## nodefactor.grouping.9 -0.03580    0.50177      0  -0.071 0.943126
## mutual                 2.08027    0.28132      0   7.395  < 1e-04 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##      Null Deviance: 7694  on 5550  degrees of freedom
```

```
##  Residual Deviance: 1174  on 5531  degrees of freedom
##
## AIC: 1212  BIC: 1338  (Smaller is better. MC Std. Err. = 0.4483)
```

The nodematch term for grouping is positive and significant, indicating strong tendencies for interactions in the same groups - families within each group are more likely to interact with each other than those in different groups. Further, reciprocity term is also significant, indicating visiting behaviors tend to be reciprocal.

ERGM is a powerful tool for modeling social networks. We provide an introduction to the theoretical foundations of ERGM. When applying ERGM, researchers need to specify the vector of model statistics or covariates. Social science theories should drive and guide the specification of those network statistics. In short, we assume that the global network structure is an aggregation of local processes. When employing ERGM, the researcher's job is to propose the socially-meaning stochastic process(es) that could result in the observed network.

References

Butts, Carter. (2008). Social Network Analysis: A Methodological Introduction. Asian Journal of Social Psychology. 11. 13 - 41. 10.1111/j.1467-839X.2007.00241.x.

Charles P. Loomis, Julio O. Morales, Roy A. Clifford & Olen E. Leonard, Turrialba. Social Systems and the Introduction of Change (Glencoe (Ill.): The Free Press, 1953): p. 43 (Attiro), 45 and 78 (San Juan Sur).

Harris, J. K. (2014). An introduction to exponential random graph modeling. SAGE Publications, Inc. https://dx.doi.org/10.4135/9781452270135

Handcock M, Hunter D, Butts C, Goodreau S, Krivitsky P, Morris M (2018). ergm: Fit, Simulate and Diagnose Exponential-Family Models for Networks. The Statnet Project (http://www.statnet.org). R package version 3.9.4, https://CRAN.R-project.org/package=ergm.

Hunter D, Handcock M, Butts C, Goodreau S, Morris M (2008). "ergm: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks." Journal of Statistical Software, 24(3), 1–29.

Lusher, D., Koskinen, J., & Robins, G. (Eds.). (2012). Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications (Structural Analysis in the Social Sciences). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511894701

Wasserman, S., & Faust, K. (1994). Social Network Analysis: Methods and Applications (Vol. 8). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511815478