(Duxbury 2023)

Duxbury, Scott W. 2023. “The Problem of Scaling in Exponential Random Graph Models.” *Sociological Methods and Research* 52 (2): 764–802. https://doi.org/10.1177/0049124120986178.

\citet{Duxbury2021} replicates \citet{Goodreau2009}, correcting for scaling. The goal of this article is to introduce the problem of scaling in ERGM, to outline its sources, and to propose methods for overcoming the issue in ERGM.

Residual variation is an unexplained variation in tie probabilities. (p.2)

Coefficients and exponentiated coefficients cannot be interpreted as effect sizes. Scaling can produce differences in coefficient size. Scaling can bias interaction coefficients yielding incorrect assessments of direction, interaction effect size, and significance. (p.2)

Residual variation can impact ERGM coefficients, which is common in both sparse and incredibly dense networks and is related to the known problem of degeneracy. (p.2)

``It is common for researchers to interpret coefficients as effect sized, to compare coefficients between models, and rely on interaction coefficients to interpret interaction'' \citep[p.~2-3]{Duxbury2021}. However, there is a problem with scaling in ERGM, and it affects the evaluation of the parameter interpretation \citep[p.~3]{Duxbury2021}.

Given a network \( Y \) with ties \( y\_{ij} \) connecting actors \( i \) and \( j \), the Exponential Random Graph Model (ERGM) estimates the probability of observing \( Y \) as a function of exogenous actor-level characteristics and sufficient graph statistics. The probability mass function for ERGM is given by:

\begin{equation}

Pr(Y = y | z(y, x)) = \frac{\exp(\theta^T z(y, x))}{\kappa(\theta)},

\end{equation}

where \( \theta \) is the parameter vector, \( z(y, x) \) is a vector of exogenous characteristics \( x \) and endogenous graph statistics computed on \( Y \), and \( \kappa(\theta) = \sum \exp(\theta^T z(y, x)) \) is a normalizing constant ensuring that the sum of probabilities over all possible networks is 1. Due to the intractability of \( \kappa(\theta) \) in most networks of interest, this denominator is typically approximated using Markov chain Monte Carlo (MCMC) sampling \citep{Duxbury2021}.

\citet{Duxbury2021} identifies a key limitation in ERGMs pertaining to the random error assumed in the formation of ties between actors within a network, which is not explicitly accounted for in standard ERGM equations. To rectify this, \citet{Duxbury2021} recommends conceptualizing network ties as manifestations of an underlying, continuous latent variable influenced by both observed and unobserved factors, along with an error term. The typical assumption in stochastic models, including ERGMs, is that this error has zero mean and a fixed variance, conventionally taken to be approximately 3.29. This fixed variance can lead to biases, particularly when the actual error variance diverges from this assumption, affecting the accuracy of the model's coefficients. To address the discrepancy, \citet{Duxbury2021} introduces a scaling factor *τ* to align the assumed error variance with the actual variance. While such scaling issues do not influence the predicted probabilities or the directionality of coefficients, they can significantly distort the magnitude of coefficients and thus impede accurate comparisons across different models or groups within a model.

Scaling in ERGMs is impacted by residual variation, which obscures the true scale of coefficients. This variation stems from factors like unaccounted nodal characteristics, hierarchical nesting structures, and data collection inaccuracies. Strategies like introducing nodal random effects, modeling nesting patterns, and applying pseudolikelihood estimation are employed to counteract these issues, yet residual variations can persist, subtly distorting scaling.

Decay parameters, indicating how tie probability declines with node distance or difference, are sharpened by properly modeling these nesting structures. However, two notable concerns persist: the logistic model's fit and omitted variables. The logistic model may misrepresent tie distributions in networks that are exceptionally dense or sparse, challenging its applicability. Additionally, any variables left out of the model, regardless of their level or interaction within the network, can introduce scaling errors. Given the practical impossibility of including every potential variable, ERGMs inherently risk scaling inaccuracies due to these latent factors.

Top of Form

The scaling issue in ERGMs, tied to untested residual variation, implies coefficients are scaled and not absolute effect sizes. Omitted variables during data collection can lead to substantial rescaling, especially if these variables have strong effects or high variance, distorting coefficient magnitudes.

Key consequences include:

1. **Effect Sizes**: Coefficients and their exponentiated forms should not be construed as effect sizes due to potential rescaling effects, which impact the magnitude but not the direction or significance of noninteraction coefficients.
2. **Comparability**: Coefficients between models can't be reliably compared without assuming constant scaling factors (τ), an assumption that's usually invalid, particularly when variables affecting tie probabilities are added or removed.
3. **Interaction Coefficients**: The significance and interpretation of interaction effects, such as homophily or heterophily, are unreliable unless scaling factors are uniform across groups, an unlikely scenario given expected heterogeneity.

The presence of omitted variables can cause scaling, a common yet unverifiable condition in ERGM applications. This makes scaling a prevalent and unresolved problem in statistical network analysis, challenging the interpretation of ERGMs and necessitating further evaluation through simulation studies.