INTRODUCCIÓN A LOS MODELOS EXPONENCIALES DE GRAFOS ALEATORIOS

Alejandro Espinosa-Rada

Twitter: @aespinosarada

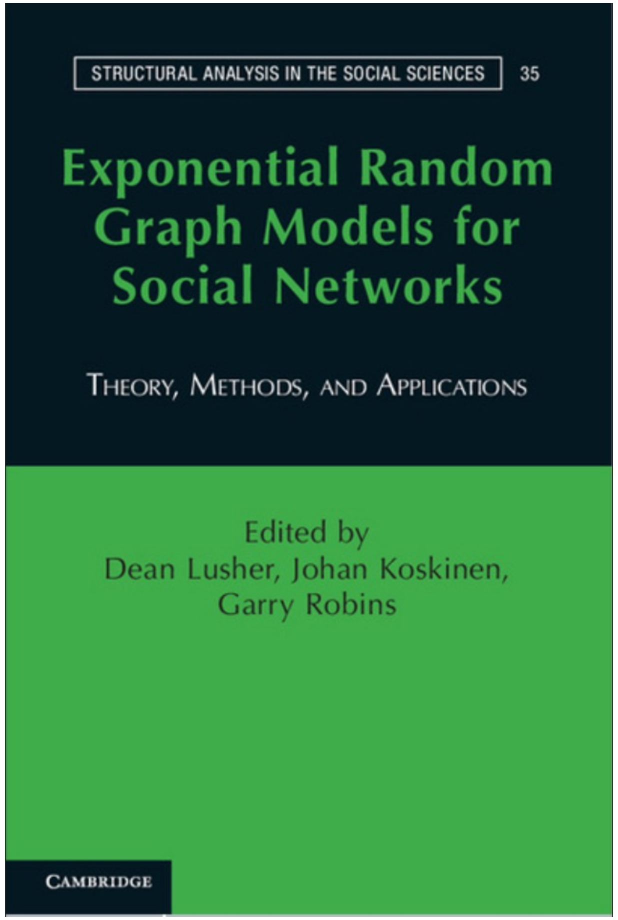
ETH zürich

Social Networks Lab

¿Qué son los ERGMs?

- **ERGMs:** *Exponential Random Graph Models* (Modelos Exponenciales de Grafos Aleatorios).
- Son modelos estadísticos que consideran las estructuras de las redes
- Permiten hacer inferencias sobre cómo la red genera patrones
- Aproximación basada en vínculos (i.e., *tie-based*) que permite entender cómo y por qué los vínculos sociales emergen.

Libro de ERGMs



American Sociological Association

Mathematical Sociology Award Recipient History

The Section on Mathematical Sociology's Harrison White Outstanding Book Award

2016: Dean Lusher, Johan Koskinen, and Garry Robins, *Exponential Random Graph Models for Social Networks*. Cambridge University Press. 2013.

Softwares para estimar ERGMs

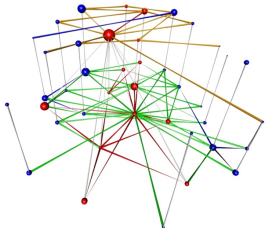
PNet is a software for the statistical analysis of social network data using Exponential Random Graph Models (ERGM) and Auto-logistic Actor Attribute Models (ALAAM).

PNet, XPNet and MPNet run on Windows and are free for noncommercial use.

Required Environments
In order to run the PNet software you will need to have one of the following operating environments on your computer:

Java Runtime Environment (JRE) 6.0 or above

Microsoft .NET Framework Version 4.0 or above

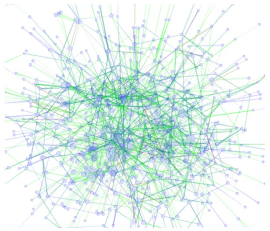


MPNET FOR MULTILEVEL NETWORKS

In addition to most of functions implemented under PNet, MPNet is also designed for:

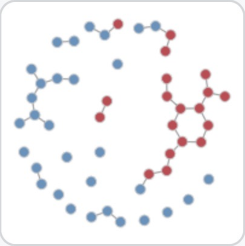
PNET FOR ONE-MODE NETWORKS

PNet is for the simulation and estimation of ERGMs for one-mode networks.



XPNET FOR BIVARIATE ANALYSIS

PNet is for the simulation and estimation of ERGMs for two one-mode networks.



Statnet

Statistical software for the analysis, simulation and visualization of network data.

http://statnet.org/ contact@statnet.org

2019 Award

Martina Morris, Mark Handcock, Carter Butts, Pavel Krivitsky, David Hunter, Steve Goodreau and Skye Bender-deMoll

The winners of the INSNA's 2019 William D. Richards Software award are Martina Morris, Mark Handcock, Carter Butts, Pavel Krivitsky, David Hunter, Steve Goodreau and Skye Bender-deMoll for STATNET package for R.

`ergm`: Fit, Simulate and Diagnose Exponential-Family Models for Networks

downloads 8142/month | CRAN 4.1.2 | codecov 69% | R-CMD-check passing

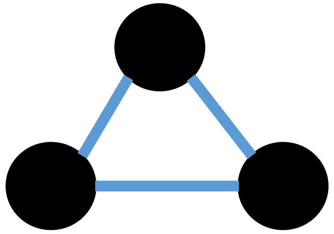
An integrated set of tools to analyze and simulate networks based on exponential-family random graph models (ERGMs). 'ergm' is a part of the Statnet suite of packages for network analysis. See Hunter, Handcock, Butts, Goodreau, and Morris (2008) doi:10.18637/jss.v024.i03 and Krivitsky, Hunter, Morris, and Klumb (2021) arXiv:2106.04997.

¿Cuáles son los procesos subyacentes que crean y mantienen el sistema social (basada en redes), dado que solo observamos una red?

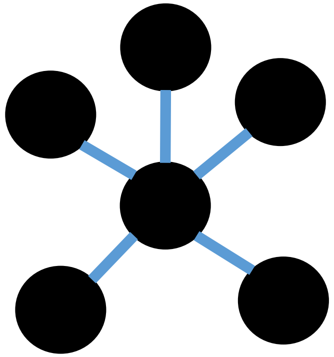
Teorías y mecanismos de redes sociales

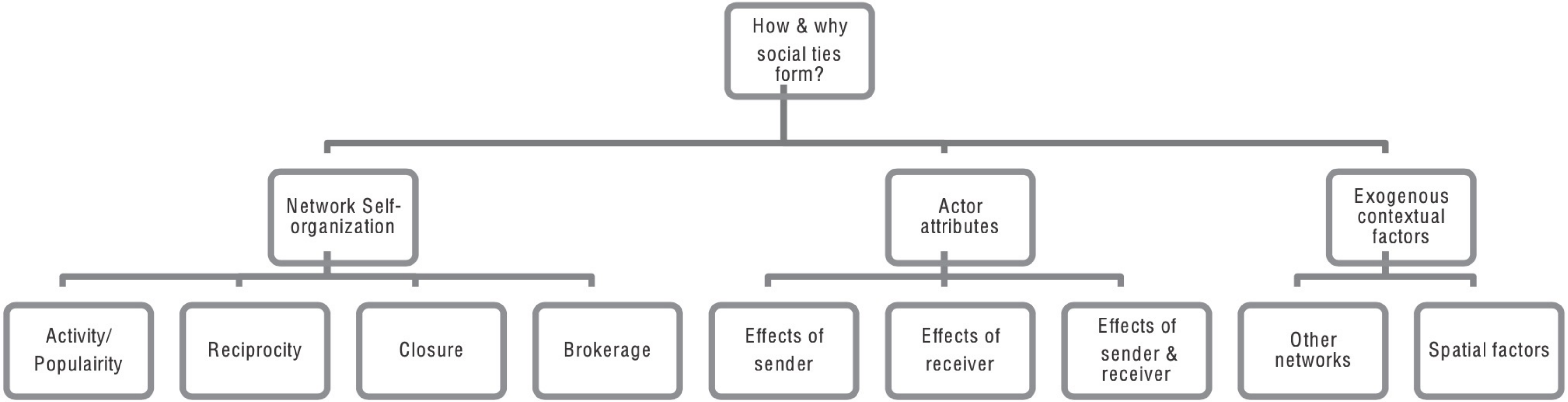
- Díadas:
  - Reciprocidad (Gouldner, 1960; Blau, 1964; Emerson, 1976)
  - Homofilia (Lazarsfeld y Merton, 1954)

- Tríadas:
  - Balance estructural (Heider, 1958)
  - Clustering (Davis, 1969)
  - Agujeros estructurales y *brokerage* (Burt, 1992; Krackhardt, 1999)



- Basados en grados nodales/posiciones
  - Popularidad Efecto Mateo (Merton, 1968)





Lusher, Koskinen y Robins (2013: 24)

Supuestos del modelo

- Las redes sociales emergen de forma local
- Las redes no solo se auto-organizan (dependencia de los vínculos), sino que ellas también se encuentran influenciadas por los atributos de los actores y otros factores exógenos.
- Los patrones de las redes pueden ser considerados como evidencia de procesos estructurales que se encuentran acualmente operando (!)
- Múltiples procesos pueden operar de forma simultánea
- Las redes sociales son estructuras, pero estocásticos

Configuraciones

- En general, cuando analizarmos redes sociales, usualmente consideramos medidas de resúmen. Por ejemplo, identificando el número de vínculos que ésta posee, el número de vínculos mutuos, medidades de centralidades, censo de tríadas, entre otros.
- En el marco de los *ERGMs* estas medidas de resúmen son usualmente denominadas 'estadísticos de redes' ($z(\theta)$) que representan lo que en esta perspectiva se denomina *configuraciones* o *sugrafos locales*.

Ejemplo (1): “La Corporación”

Contexto: Ejecutivos en una industria de entretenimiento, en donde se observa la “red de comunicación” en una red que es directa y binaria. Esta “corporación” buscaba organizarse para dominar el mercado en el participa.

Vínculo: identificar qué comunicación con otro ejecutivo fue importate para que el trabajo se realizara de forma exitosa.

¿Hay alguna configuración que destaca por sobre las otras?

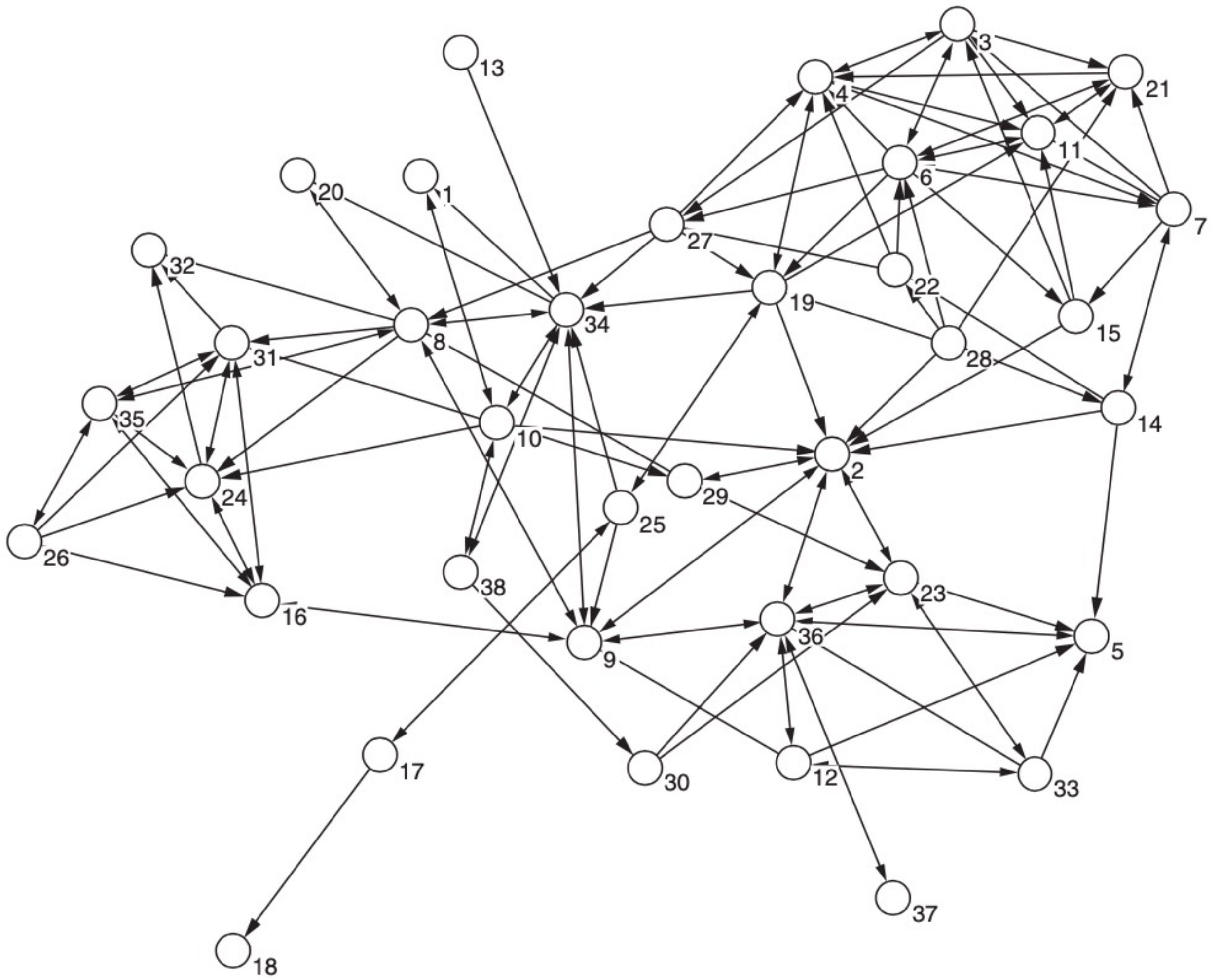


Figure 5.1. Communication network of The Corporation (n = 38 actors).

Lusher, Koskinen & Robins, 2013: 38

Ejemplo (2)

¿Lazos mutuos entre actores?

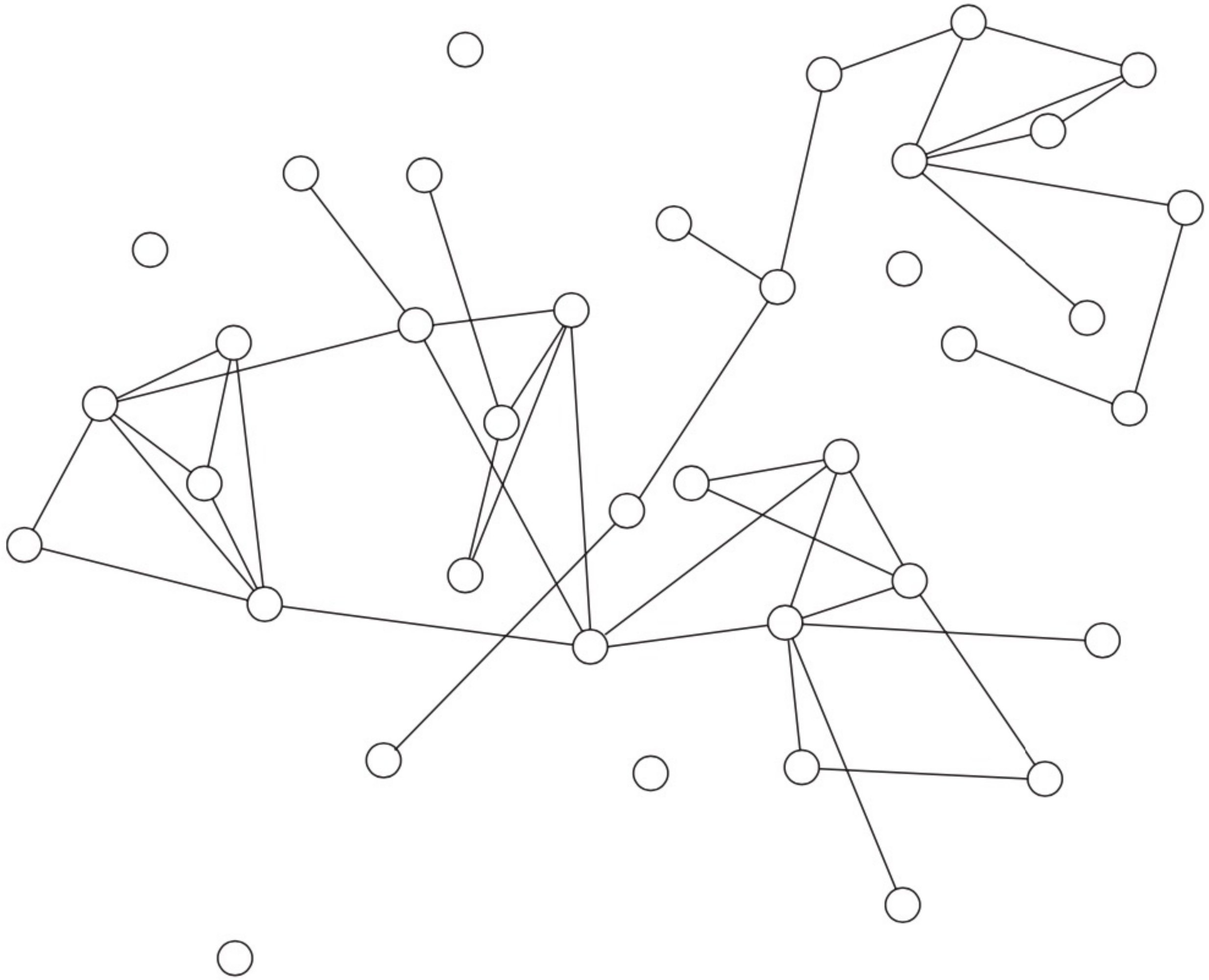


Figure 5.2. Mutual ties only (asymmetric ties removed) in communication network.

Lusher, Koskinen & Robins, 2013: 39

Ejemplo (3)

¿Atributos de los actores?
Ejemplo: cantidad de projectos realizados por ejecutivo (nodos más grandes corresponden a más proyectos)

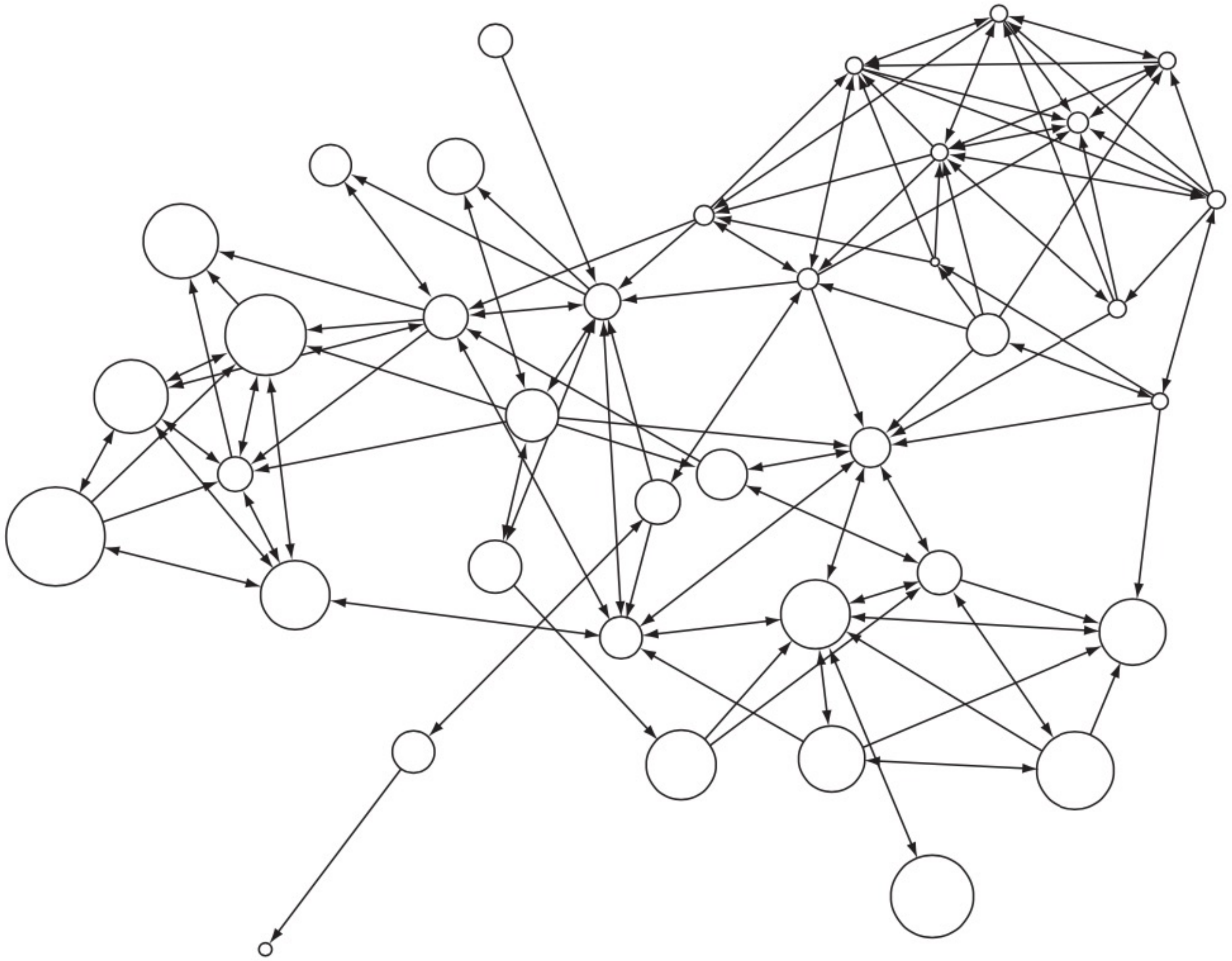


Figure 5.3. Communication network with employee experience represented by size (larger = more experience).

Lusher, Koskinen & Robins, 2013: 40

Ejemplo (4)

¿Atributos de los actores?
Ejemplo: Estatus de los ejecutivos
(“senior” / “otros”)

¿Se observa alguna tendencia de
homofilia en los datos?

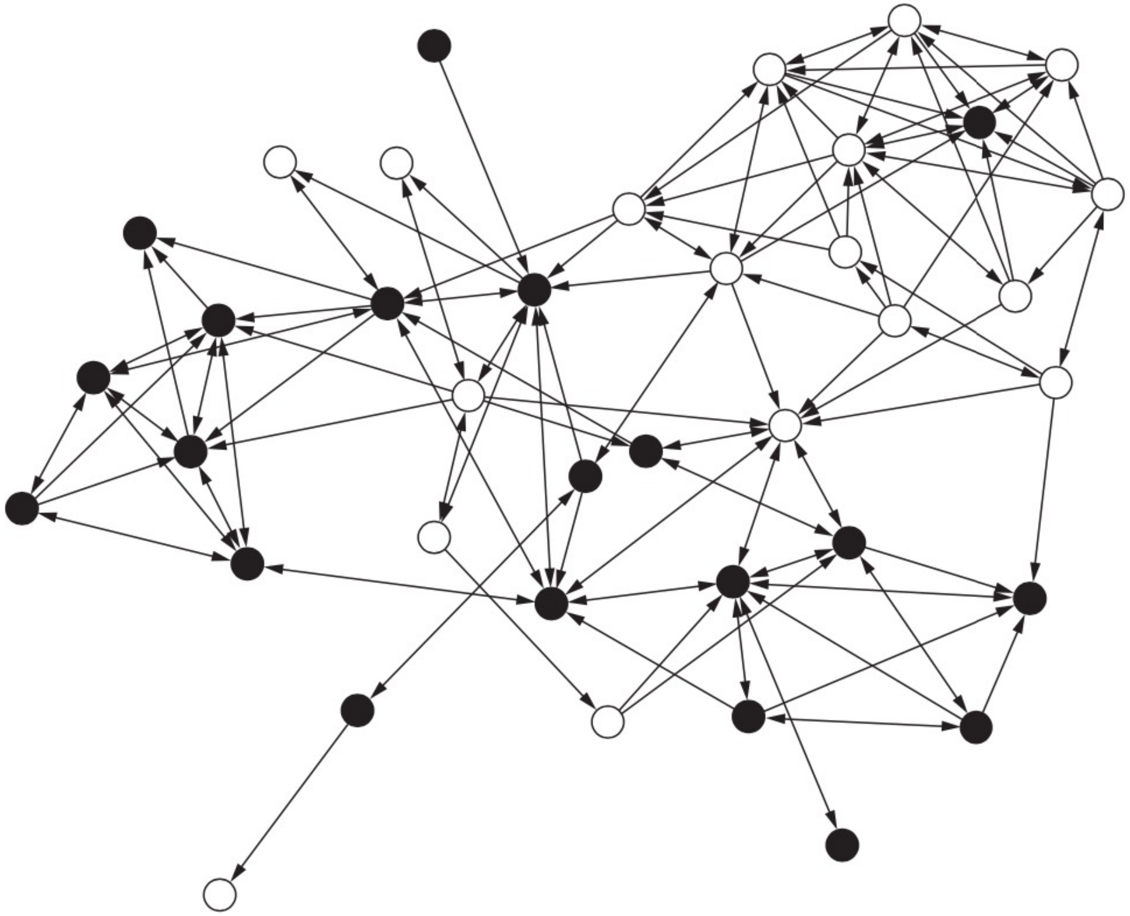


Figure 5.4. Communication network with seniority (black = senior, white = other).

Lusher, Koskinen & Robins, 2013: 40

Ejemplo (5)

¿Atributos de los actores?
Ejemplo: departamento en que trabajan los ejecutivos.

¿Se observa alguna tendencia de homofilia en los datos?

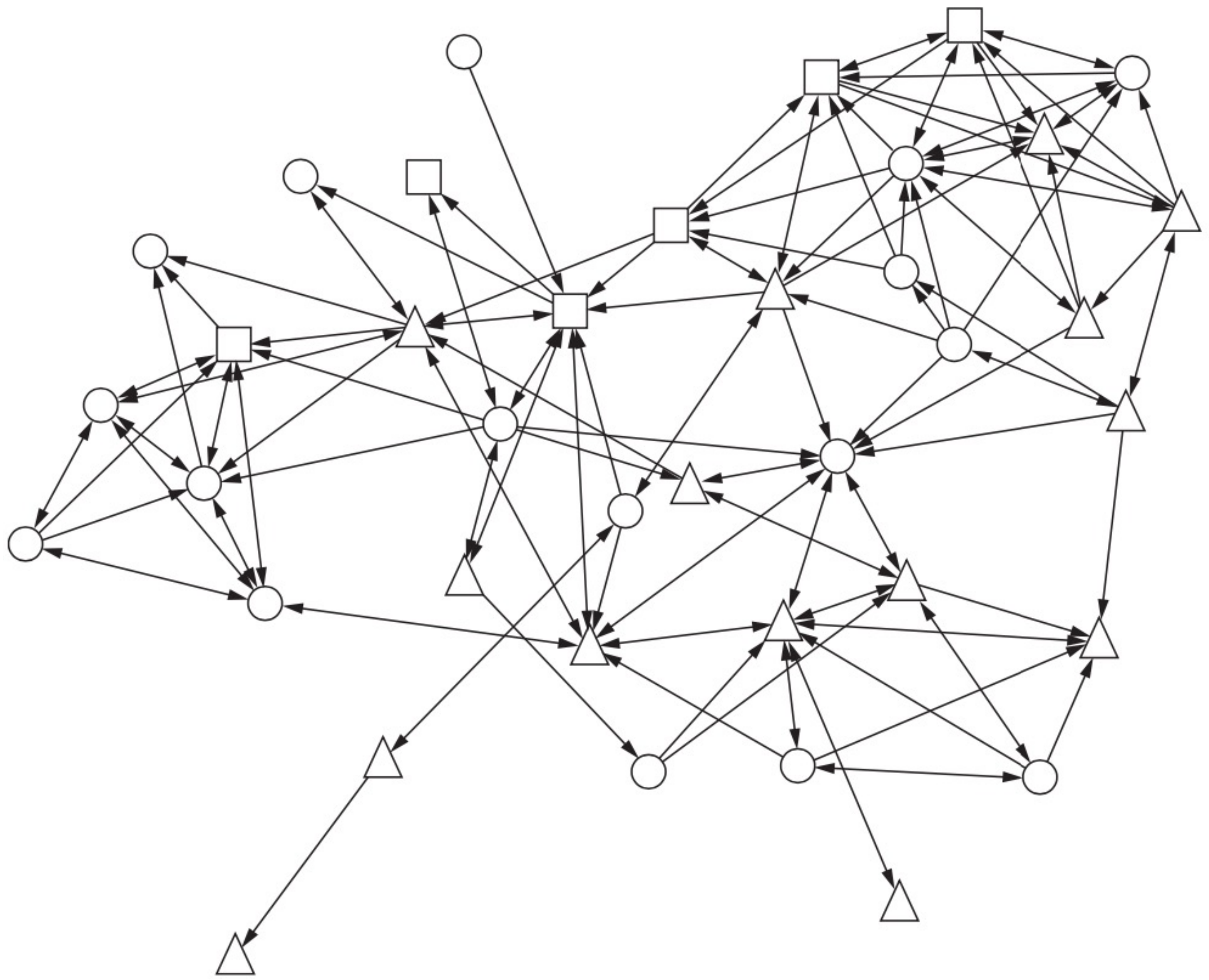


Figure 5.5. Communication network with office membership represented by shape.

Lusher, Koskinen & Robins, 2013: 41

Ejemplo (6)

Comunicaciones entre ejecutivos pueden estar afectadas por la presencia de relaciones de consejos entre ejecutivos.

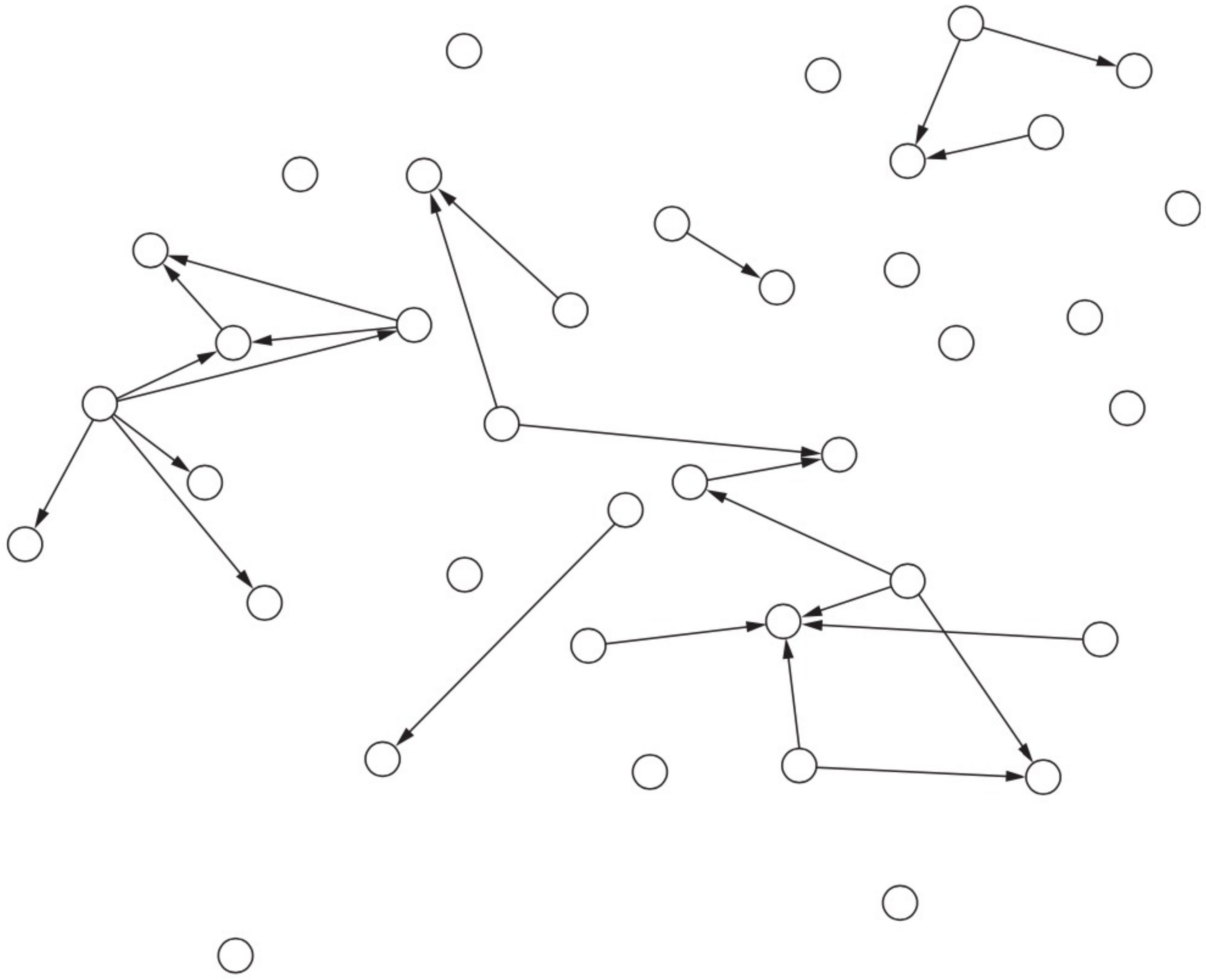
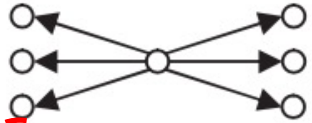
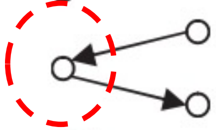
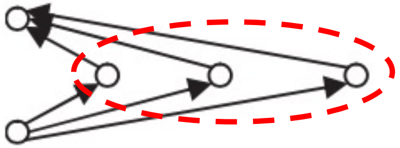
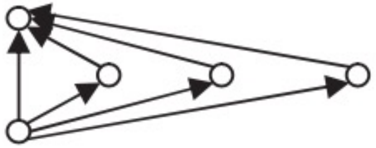
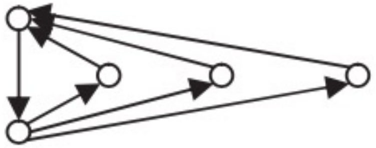


Figure 5.6. Multiplex communication and advice ties (all other ties removed).

Lusher, Koskinen & Robins, 2013: 42

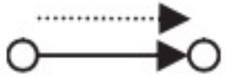
Ejemplo (7)

¿Hay más o menos configuraciones en la red en comparación a lo esperado (y dado los otros efectos en el modelo)?

Network effect		Estimate (SE)
Purely structural effects (endogenous)		
Arc		−1.96 (0.73)*
Reciprocity		2.88 (0.46)*
Popularity (in-degree)		−0.27 (0.32)
Activity (out-degree)		−0.34 (0.34)
Simple 2-path[3]		−0.06 (0.08)
Multiple 2-paths		−0.06 (0.09)
Transitivity (transitive path closure of multiple 2-paths)		1.22 (0.19)*
Cyclic closure (cyclic closure of multiple 2-paths)		−0.37 (0.17)*

Lusher, Koskinen & Robins, 2013: 43

Ejemplo (8)

Actor relation effects (exogenous) (black nodes indicates actor with attribute)		
Sender (seniority)		–0.56 (0.29)
Sender (projects)		0.01 (0.02)
Receiver (seniority)		0.08 (0.23)
Receiver (projects)		–0.02 (0.02)
[4] Homophily (seniority)		0.64 (0.26)*
Heterophily (projects)		–0.08 (0.02)*
Homophily (office)		–0.01 (0.17)
Covariate network (exogenous)		
Advice entrainment (covariate arc)		1.76 (0.30)*

Lusher, Koskinen & Robins, 2013: 43

Probabilidad de los ERGMs (1)

$$P(x;\theta) = \frac{\exp(\sum_k \theta_k z_k(x))}{\kappa}$$

En donde,

x= es un grafo

z = la suma de estadísticos de redes (ejemplo: número de vínculos, relaciones recíprocas, tripletas transitivas, distribuciones de grados, homofilia)

θ= parámetro ponderado

κ = Constante para normalizar la ecuación

Probabilidad de los ERGMs (2)

$$P(x;\theta) = \frac{\exp(\sum_k \theta_k z_k(x))}{k}$$

$$\kappa = \sum_{x\prime \in X} \exp\left(\sum_k \theta_k z_k(x')\right)$$

En donde,

κ = Constante para normalizar la ecuación de todos los grafos posibles del mismo tamaño de $x' \in X$

- Cuando se busca calcular la probabilidad exacta nos encontramos con el problema de que, a excepción de grafos muy pequeños (ejemplo: Vega et al., 2021), no es possible calcular la verosimulitud $L(\theta, x)$.
- Una red directa, binaria, con n nodos tiene $|X| = 2^{n(n-1)}$ estados!

Probabilidad de los ERGMs (3)

- Buscamos estimar el parámetro θ_i:
  - Parámetros positivos indican que existen más configuraciones en la red observada que aquellos esperados aleatoriamente.
  - Parámetros negativos indican que existen menos configuraciones en la red observada que aquellos esperados aleatoriamente.
- Los parámetros permiten estimar y hacer inferencia sobre cómo la estructura de la red global podría haberse formado producto de pequeñas estructuras locales
- Al igual que en regresiones logísticas, para interpreter los coeficientes se suelen utilizar *odds ratio* (condicionados). Pero hay debates sobre si es pertinente hacerlo…

Cuatro generaciones de supuestos de dependencia

- **Dependencia de Bernoulli (Modelo aleatorio simple o Erdös-Rényi)**

  Variables-relacionales son consideradas como independientes entre sí (ejemplo: como si tirásemos una moneda para ver si crear o no vínculos)

Cuatro generaciones de supuestos de dependencia

- **Dependencia de Bernoulli (Modelo aleatorio simple o Erdös-Rényi)**

  Variables-relacionales son consideradas como independientes entre sí (ejemplo: como si tirásemos una moneda para ver si crear o no vínculos)

- **Dependencia Diádica (ejemplo: modelos p1 y p*)**

  Para grafos directos: dependencia entre díadas (el vínculo i y j depende del vínculo j e i).

Cuatro generaciones de supuestos de dependencia

- **Dependencia de Bernoulli (Modelo aleatorio simple o Erdös-Rényi)**

  Variables-relacionales son consideradas como independientes entre sí (ejemplo: como si tirásemos una moneda para ver si crear o no vínculos)

- **Dependencia Diádica (ejemplo: modelos p1 y p*)**

  Para grafos directos: dependencia entre díadas (el vínculo i y j depende del vínculo j e i).

- **Dependencia de Markov (Frank & Strauss, 1986)**

  Variables de redes son condicionalmente independientes a menos que compartan un nodo (un nodo i puede conectar dos vínculos posibles (i, j) y (i, h), entonces se señala que hay una *dependencia* condicionada al resto del grafo)

Cuatro generaciones de supuestos de dependencia

- **Dependencia de Bernoulli (modelo aleatorio simple o Erdös-Rényi)**

  Variables-relacionales son consideradas como independientes entre sí (ejemplo: como si tirásemos una moneda para ver si crear o no vínculos)

- **Dependencia Diádica (ejemplo: modelos p1 y p*)**

  Para grafos directos: dependencia entre díadas (el vínculo i y j depende del vínculo j e i).

- **Dependencia de Markov (Frank & Strauss, 1986)**

  Variables de redes son condicionalmente independientes a menos que compartan un nodo (un nodo i puede conectar dos vínculos posibles (i, j) y (i, h), entonces se señala que hay una *dependencia* condicionada al resto del grafo)

- **Dependencia de Circuito Social (modelos de órdenes superiores)**

  Variables de redes son condicionales y dependientes si pueden crear ciclos de cuatro. En ciertas circunstancias, dos variables-relacionales, X_{ij} y X_{hm} pueden ser condicionalmente dependientes entre sí, inclusive si no comparten un nodo entre ellos.

Modelo (Homogéneo) de Bernoulli

- El modelo asume que los vínculos posibles pueden ser pensados como si tirásemos una moneda. Al respecto, si la moneda es "cara" tendrá una probabilidad p. Por ello, la probabilidad condicional puede ser expresada como:

$$\Pr\left(X_{ij} = 1 \middle| X_{-ij} = x_{-ij}, \theta\right) = \Pr\left(X_{ij} = 1 \middle| \theta\right) = \frac{e^{\theta L(x)}}{1 + e^{\theta L(x)}}$$

- El *logit* de la probabilidad de un vínculo se encuentra dado simplemente por θ, que corresponde al parámetro de "vínculos" (*edge*). Expresado en una función de masa de ERGMs para la matriz de adjacencia:

$$P_\theta(x) = \frac{1}{\kappa(\theta)} \exp\{\theta_L L(x)\}$$

En donde,

θ_L= es el "parámetro de vínculos" *(edge parameter)*

$L(x) = \sum_{i<j} x_{ij}$ corresponde a un solo vínculo

Modelo de Independencia Diádica (1)

$(p_1, p_2 \text{ o } p^*)$

- Este modelo incorpora arcos y reciprocidades. A diferencia del modelo anterior, esta version consideraba un principio simple de dependencia estadística en donde es posible utilizar modelos de regresión logística multinomial.

$$P_\theta(x) = \frac{1}{\kappa(\theta)} \exp\{\theta_L L(x) + \theta_M M(x)\}$$

En donde,

θ_L= es el parámetro de los “número de arcos” *(arcs)*

$L(x) = \sum_{i<j} x_{ij}$ corresponde a un solo vínculo

θ_M= es el parámetro de los números de vínculos recíprocos o mutuos

$M(x) = \sum_{i<j} x_{ij} x_{ji}$ es el nímero de arcos recíprocos (o mutuos) en un grafo

Modelo de Independencia Diádica (2)

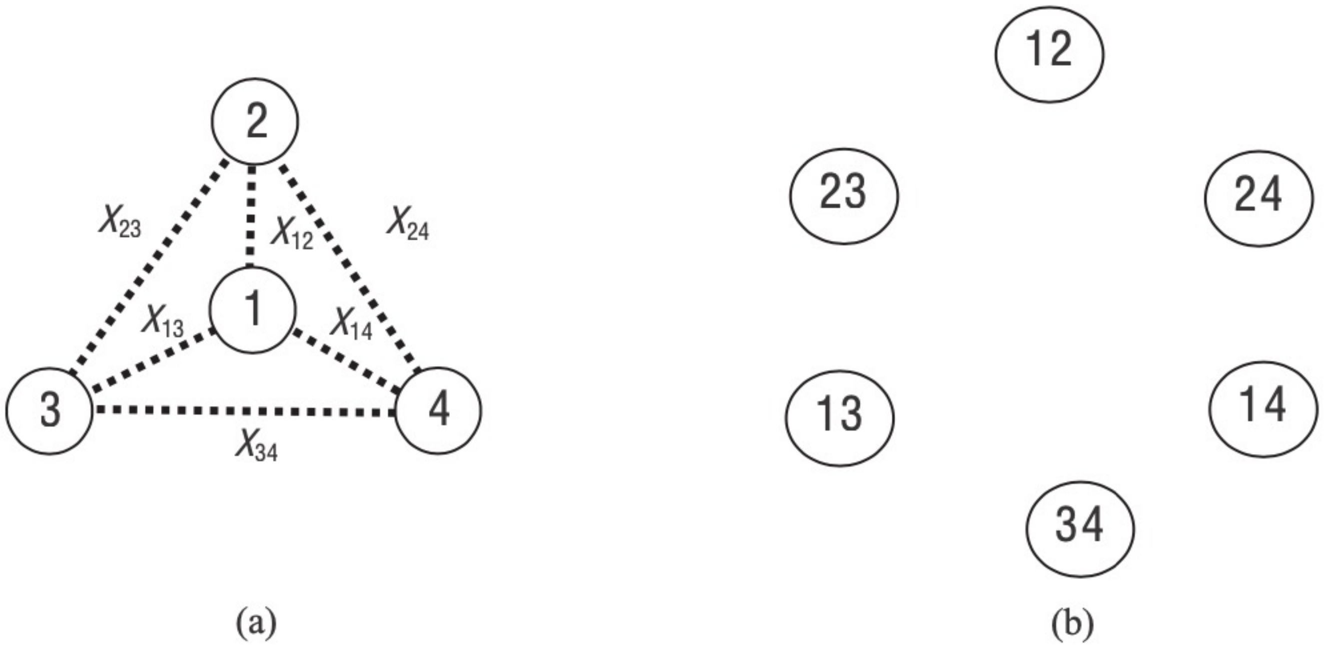


Figure 7.1. Tie-variables of (a) four-node graph and (b) associated Bernoulli dependence graph.

Lusher, Koskinen & Robins, 2013: 78

Modelo de Markov (1)

- Este modelo assume que dos variables son dependientes si “comparten un nodo”. En la configuración considerada todos los nodos del subrago tendrán un vínculo con al menos un nodo.

$$P_\theta(x) = \frac{1}{\kappa(\theta)} \exp\{\theta_L L(x) + \theta_{S_2} S_2\ (x) + \theta_{S_3} S_3\ (x) + \cdots + \theta_{S_{n-1}} S_{n-1}\ (x) + \theta_T T(x) + \}$$

En donde,

θ_{s_k}= es el parámetro de los “número de estrellas” *(k-stars)*

$S_k(x) = \sum_{i \in N} \binom{x_{i+}}{k}$ número de k-estrellas de diferentes tamaños $(2 \leq k \leq (n-1))$ y

$x_{i+} = \sum_j x_{ij}$

θ_T= es el parámetro de los números de triángulos

$T(x) = \sum_{i<j<k} x_{ij} x_{jk} x_{ki}$ es el nímero de triángulos

Modelo de Markov (2)

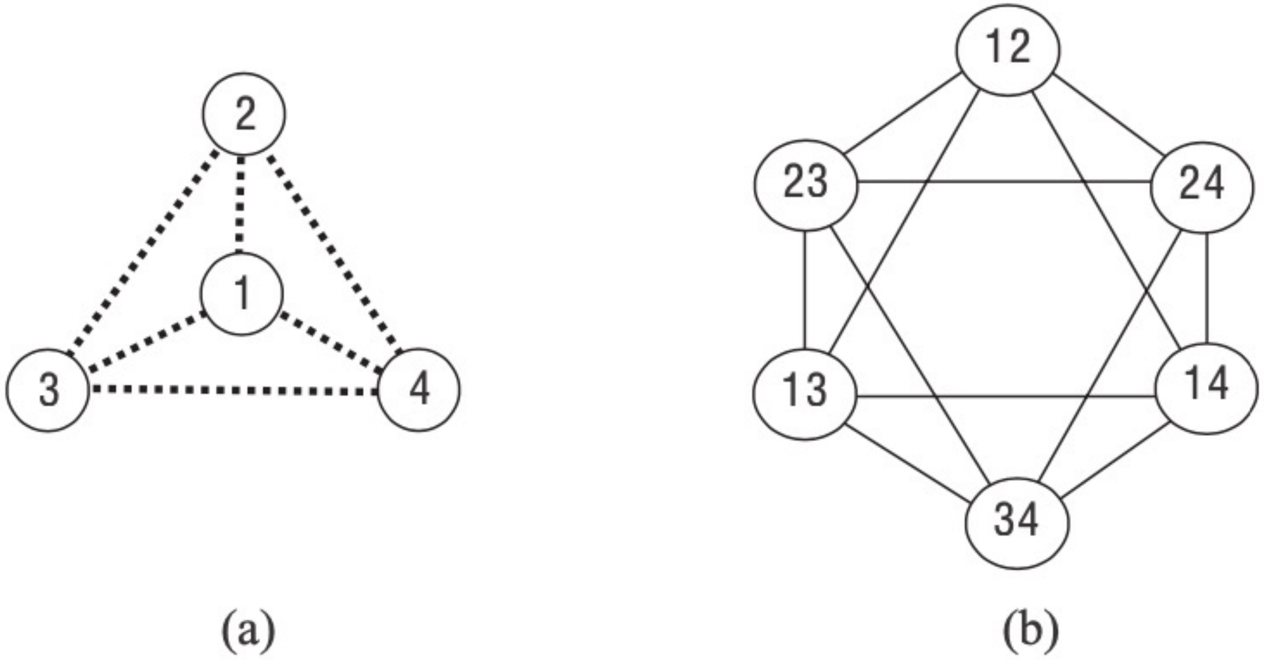
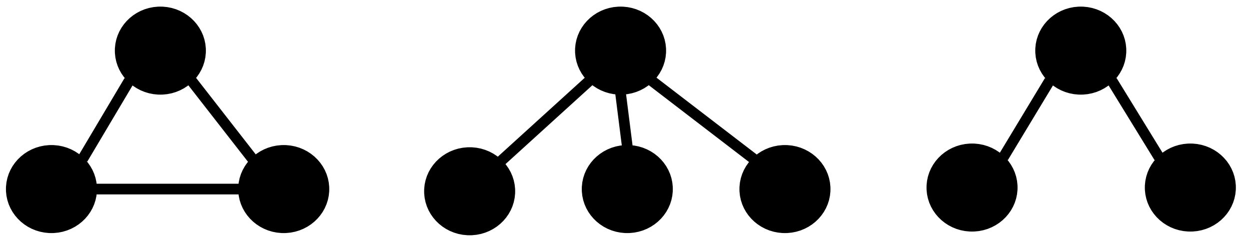


Figure 7.2. Tie-variables of (a) four-node graph and (b) associated Markov dependence graph.

Lusher, Koskinen & Robins, 2013: 79

Modelos que producen *degeneración* (degeneracy)

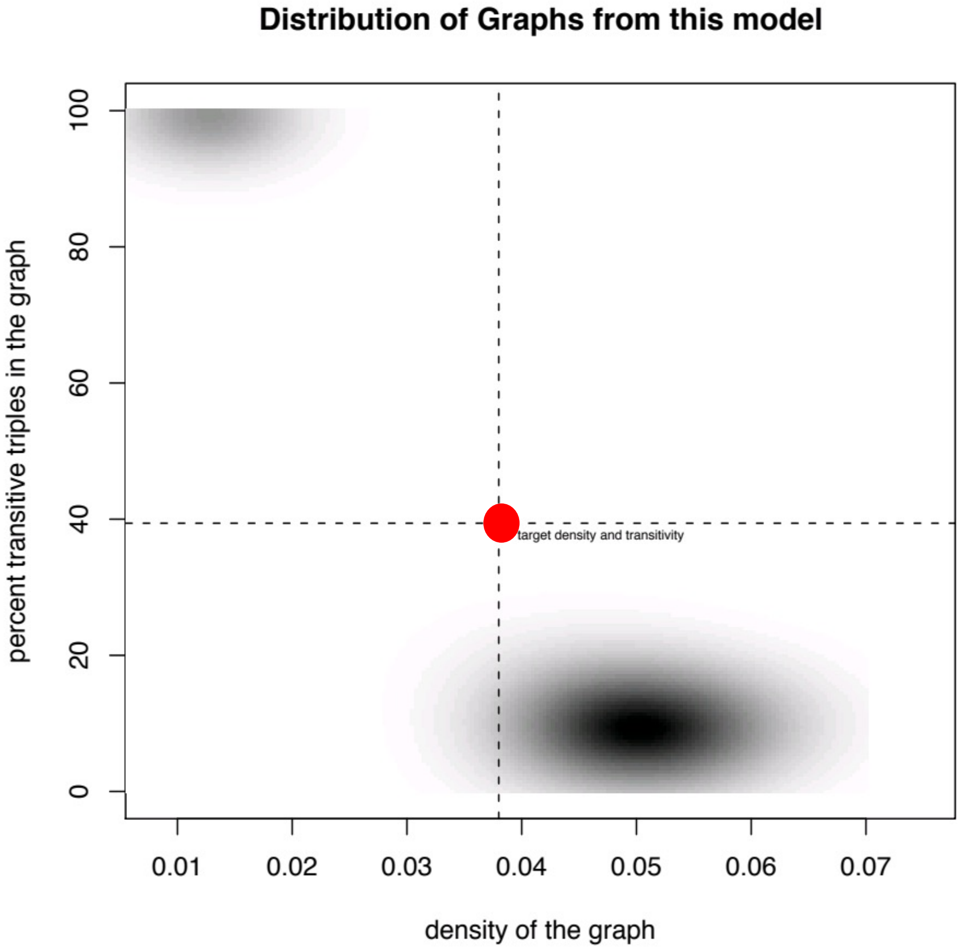
- En un principio los modelos de redes utilizaban configuraciones simples para estimar modelos tales como k-estrellas o triángulos (i.e., dependencias de Markov):



- Estos modelos suelen generar un problema común en **ERGMs** denominado "*degeneracy*". El cual consiste en que pocas configuraciones de redes – usualmente en redes muy dispersas o muy densas – producen una alta masa de probabilidad en donde al realizar la estimación de los "cambios estadísticos" (*change statistics*) los grafos resultantes se situarán ya sea en gráficos altamente disperso o completo.

Degeneración (degeneracy)

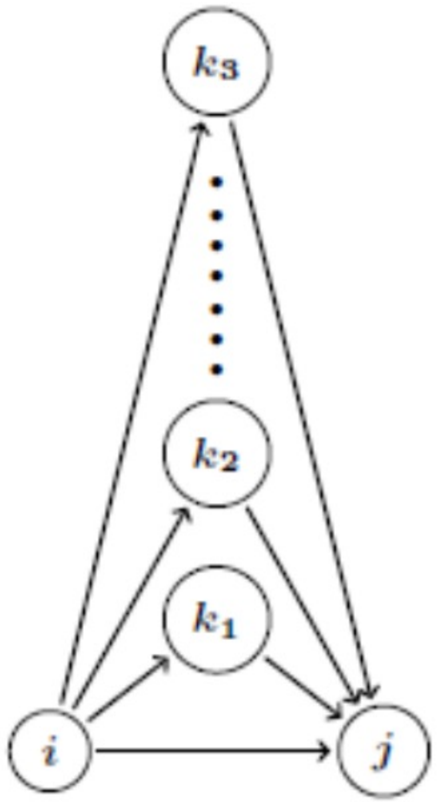
- Cuando dos parámetros son estimados *(edges* y *transitivity)* es esperable que la transitividad sea positiva. Sin embargo, al utilizar los parámetros del modelo estimado a través de simulaciones fue descubierto que los grafos resultantes entregaban comportamientos extraños.
- Los grafos simulados o se concentraban en grafos muy densos o muy dispersos, y difícilmente en el promedio (el valor real), produciéndose una "multi-modalidad" (i.e., dos modas en la distribución)
- En otras palabras, la existencia de un "vínculo" puede generar un efecto en expansion que hace que se produzcan más y más "triángulos".



Handcock et al., 2008: 6

Modelos Considerando Circuitos Sociales (1)

(modelos parcialmente condicionales e independientes)



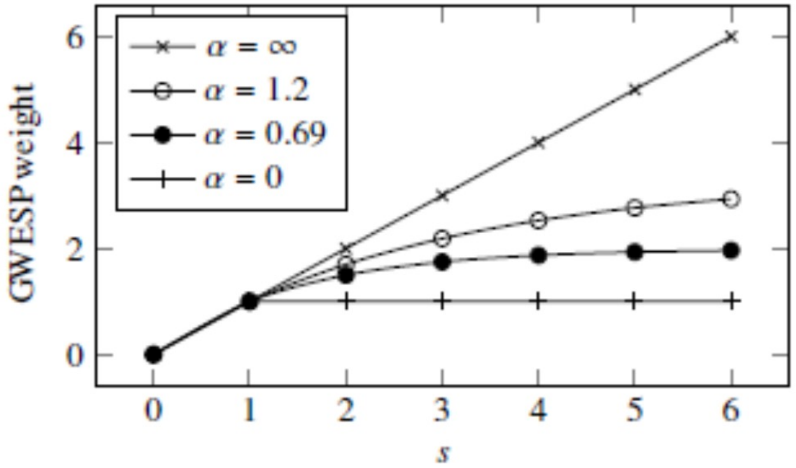


Figure 2: GWESP weight for a tie closing s two-paths for $\alpha = \infty, 1.2, 0.69, 0$.

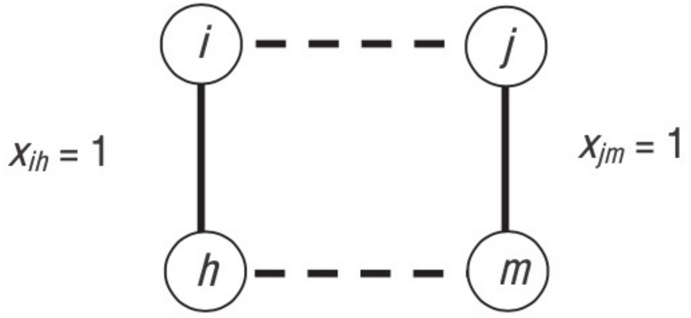


Figure 6.2. Social circuit dependence.

Lusher, Koskinen & Robins, 2013: 58

Figura en Ripley et al. (2021), referencia en Pattison y Robins (2002) y/o Snijders (2006)

Modelos Considerando Circuitos Sociales (1)

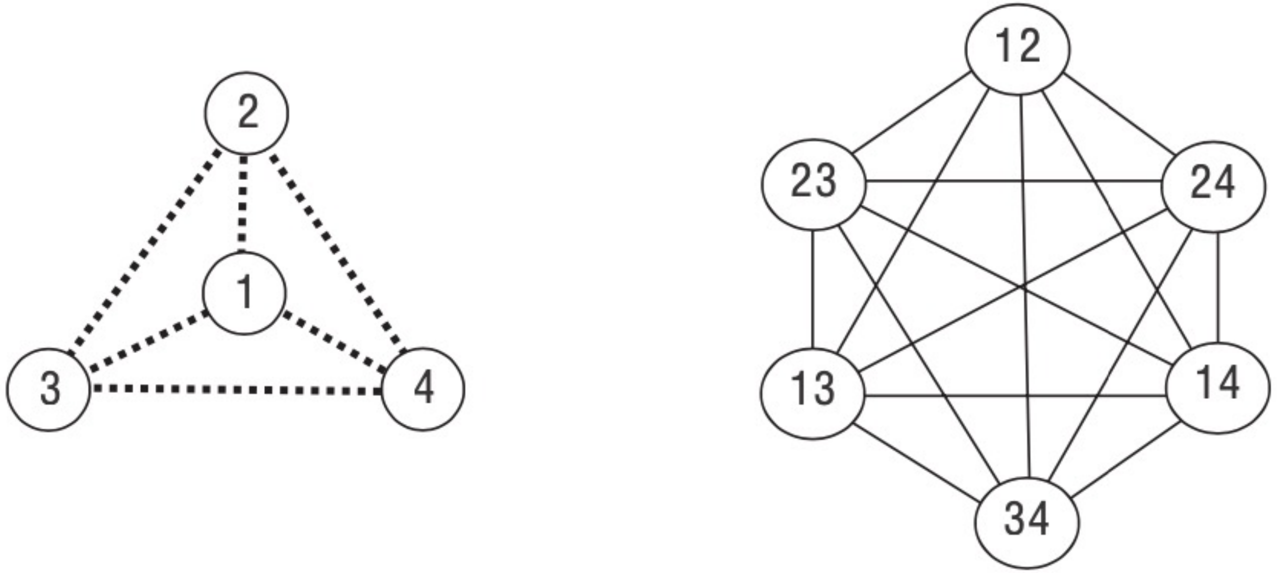
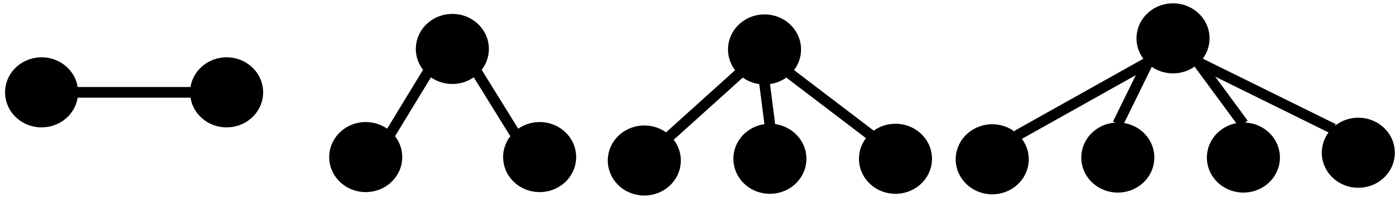


Figure 7.3. Tie-variables of social circuit graph and its dependence graph (top) and dependence graph conditional on some tie-variables being zero (greyed out vertices not in partial dependence graph).

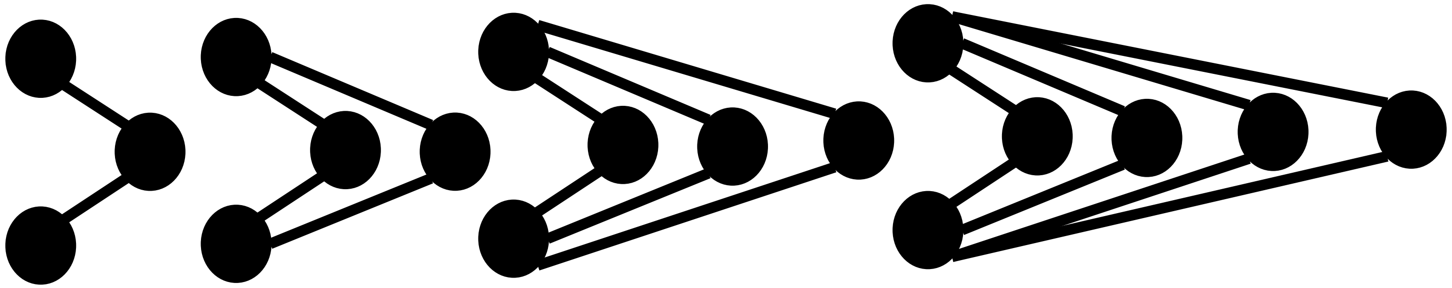
Lusher, Koskinen & Robins, 2013: 80

Estadísticos de órdenes-superiores

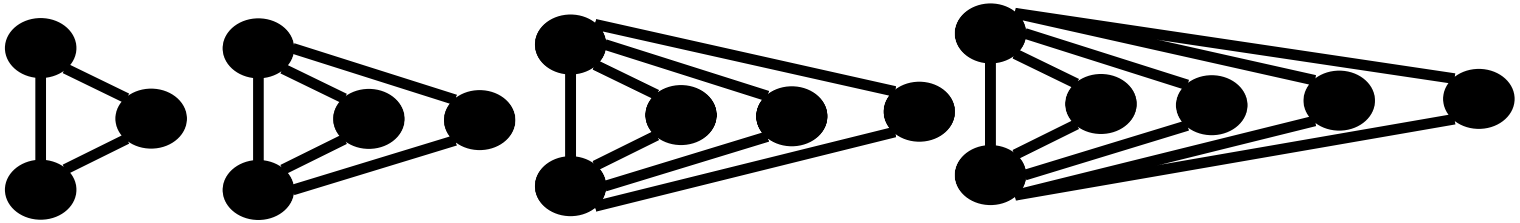
Relaciones y estrellas
(k-estrellas)
Weighted degree (gwdegree)/ Alternating star (AS)



Múltiples caminos
(k-caminos)
Dyad-wise shared partners (gwdesp)/alternating 2-path shared partners (A2P)



Múltiples triángulos
(k-triángulos)
Edgewise shared partners (gwesp)/alternating triangle closure (AT)



Bondades de Ajuste (1)

¿El modelo logra representar adecuadamente estructuras observadas en el grafo?

Para ello se utiliza una contrastación entre la red observada y la distribución de modelos simulados a partir de los parámetros estimados, considerando *efectos que no fueron explícitamente incorporados en el modelo*. Por ejemplo, podemos evaluar la bondad de ajuste a través del siguiente estadístico:

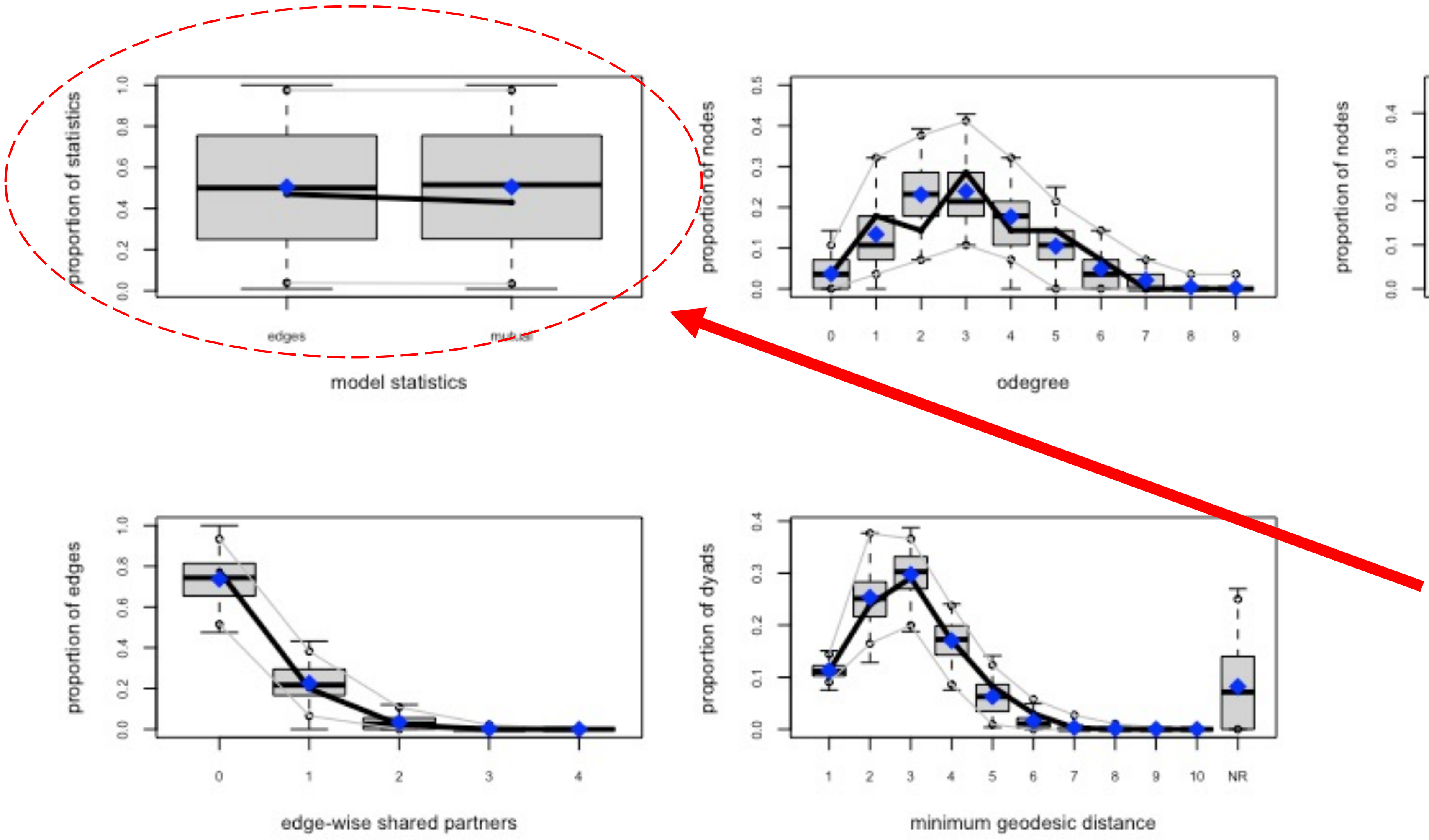
$$[S_k(x_{obs}) - \bar{S}_k]/SD(S_k(x))$$

En donde,

$\bar{S}_k$= Promedio

$SD(S_k(x))$= Desviación estándar

Bondades de Ajuste (2)



¡Cuidado!: Este gráfico es un test de convergencia

Referencias

- Pattison, P., & Robins, G. (2002). Neighborhood-based models for social networks. *Sociological Methodology*, *32*(1), 301-337.
- Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., & Morris, M. (2008). statnet: Software tools for the representation, visualization, analysis and simulation of network data. *Journal of statistical software*, *24*(1), 1548.
- Lusher, D., Koskinen, J., & Robins, G. (Eds.). (2013). *Exponential random graph models for social networks: Theory, methods, and applications*. Cambridge University Press.
- Ripley, R. M., Snijders, T. A. B., Boda, Z., Vörös, A., & Preciado, P. (2021). Manual for SIENA (Version April 6, 2021). *University of Oxford*.
- Snijders, T. A., Pattison, P. E., Robins, G. L., & Handcock, M. S. (2006). New specifications for exponential random graph models. *Sociological methodology*, *36*(1), 99-153.
- Vega Yon, G. G., Slaughter, A., & de la Haye, K. (2021). Exponential random graph models for little networks. *Social Networks*, *64*, 225-238.