“Exponential Random Graph Models” para el análisis de Redes Sociales

Diego Gómez-Zará

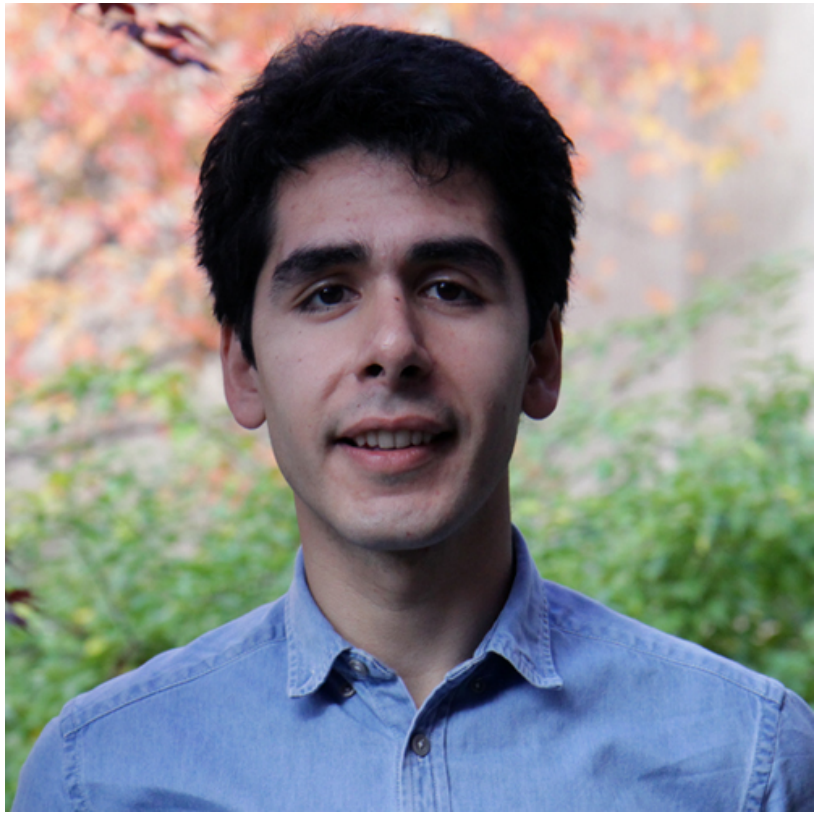
Presentación SocVis 11 de agosto 2017

Patrocinada por el proyecto Fondecyt de Iniciación 11150783,
“Dealing with information overload using intelligent recommender system interfaces” y
Army Research Office W911NF-14-10686

Northwestern University

SONIC
advancing the science of networks in communities

¿Quién soy yo?



- Magíster en Ciencias de la Computación PUC.
- Hoy: 2nd year Ph.D. student en Northwestern University
- Mi programa es “Technology and Social Behavior” (Computer Science + Communications)
- Advisor: Prof. Noshir Contractor

Introducción

- Esta presentación explica los fundamentos y pasos básicos para crear ERGMs
- Especificamente:
  - Escribir el modelo
  - Calcular los valores a partir del método de máxima varianza (MLEs)
  - Chequear la convergencia de las simulaciones (MCMC)
  - Simular redes aleatorias a partir del modelo
  - Chequear los resultados mediante Goodness-of-fit (GOF)

¿Qué es un ERGM?

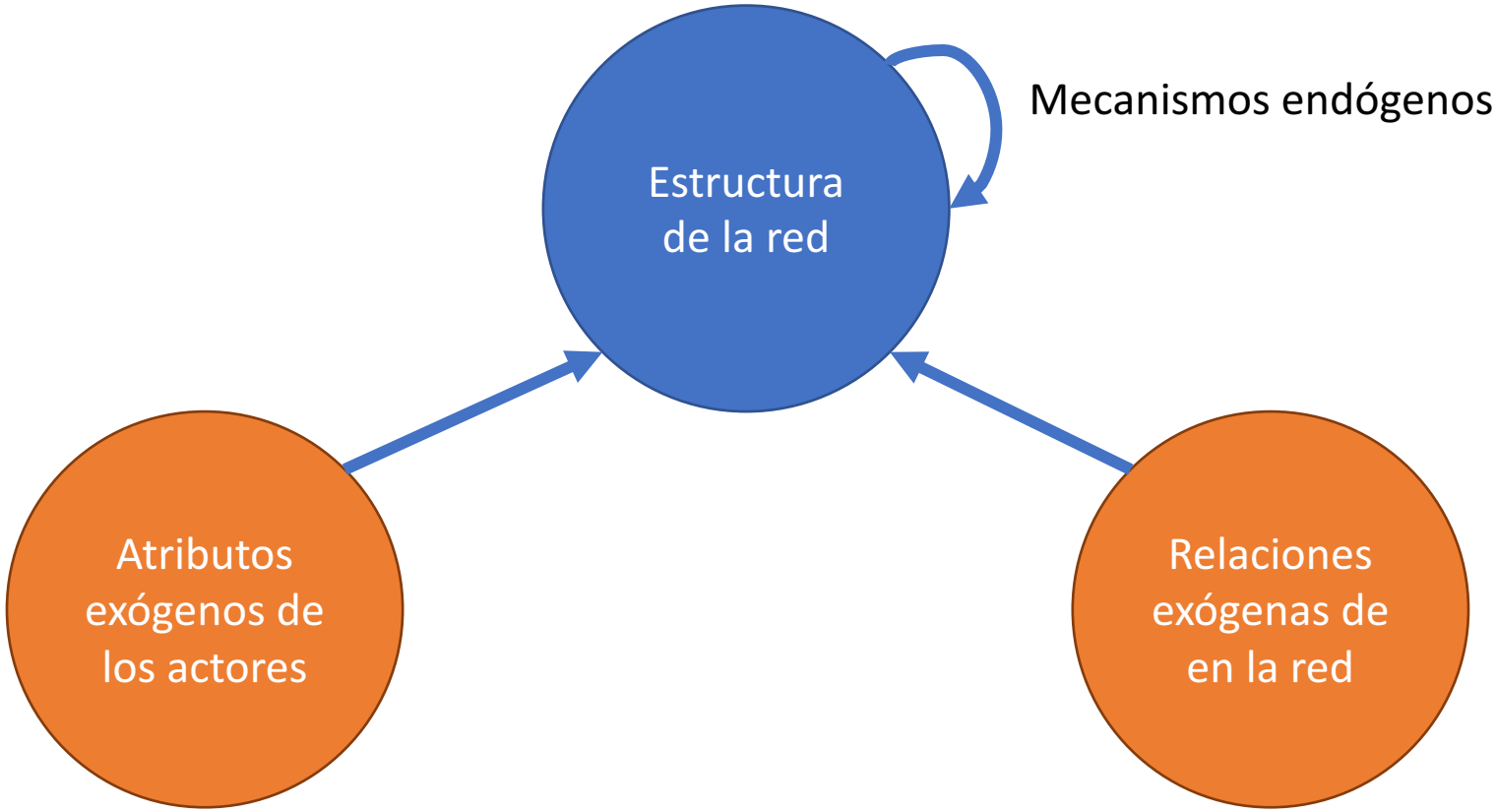
- Los Modelos de Grafos Aleatorios Exponenciales (Exponential Random Graph Models) son una familia de modelos estadísticos que analizan información de redes.
- Su propósito es describir las configuraciones que posee la estructura de una red.
- Los resultados pueden ser usados para entender un fenómeno particular o generar nuevas simulaciones (Hunter et al. 2008).

Motivación

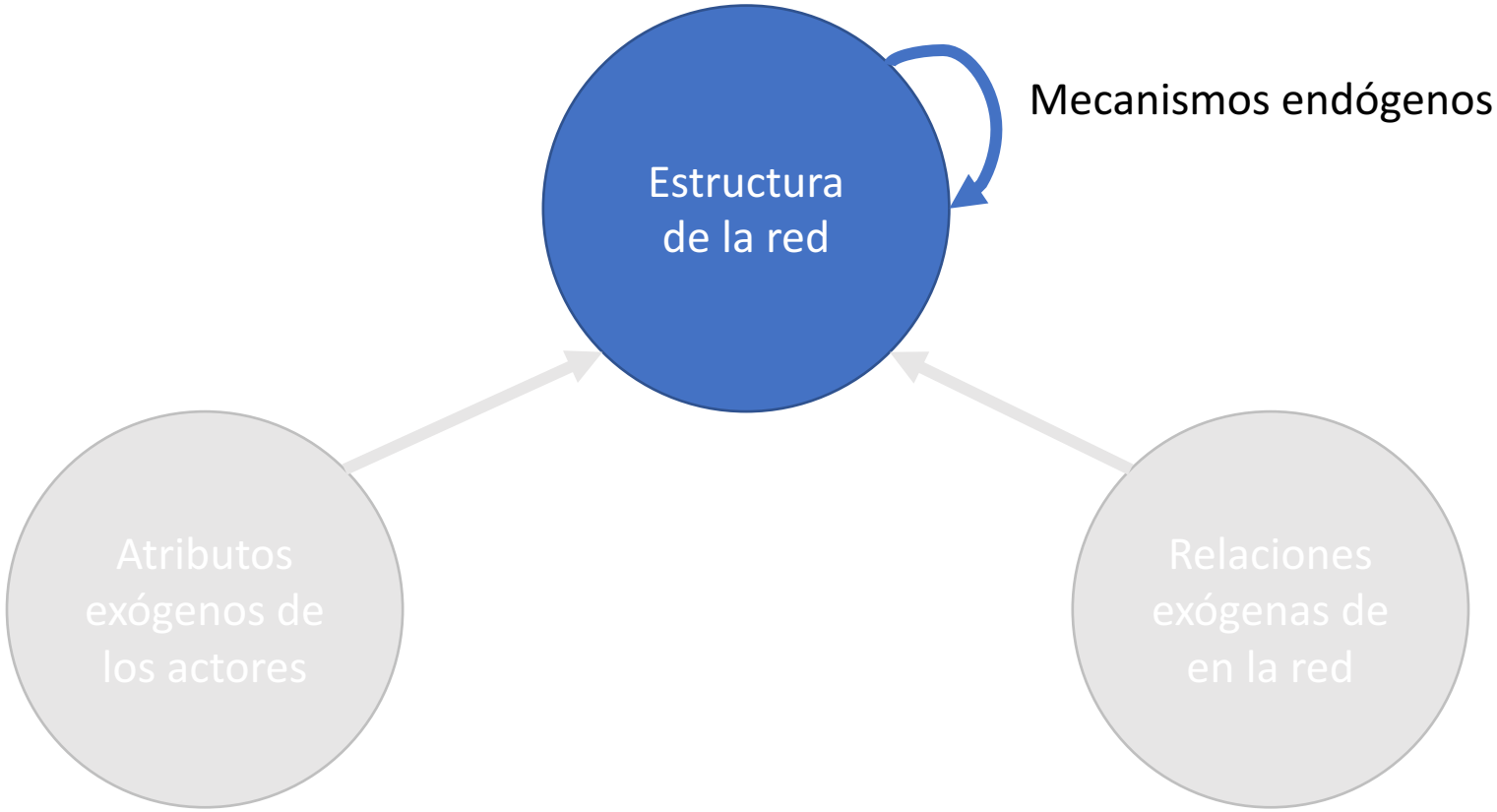
¿Por qué las personas creamos, mantenemos, disolvemos, y reconstituimos conexiones?

- Teorías de interés propio
- Teorías de interés mutuo y acción colectiva
- Teorías de intercambio de recursos/capital social
- Teorías de contagio
- Teorías de balance
- Teorías de homofilia
- Teorías de proximidad
- Teorías de co-evolución

Motivación


Estructura de la red
Mecanismos endógenos
Atributos exógenos de los actores
Relaciones exógenas de en la red

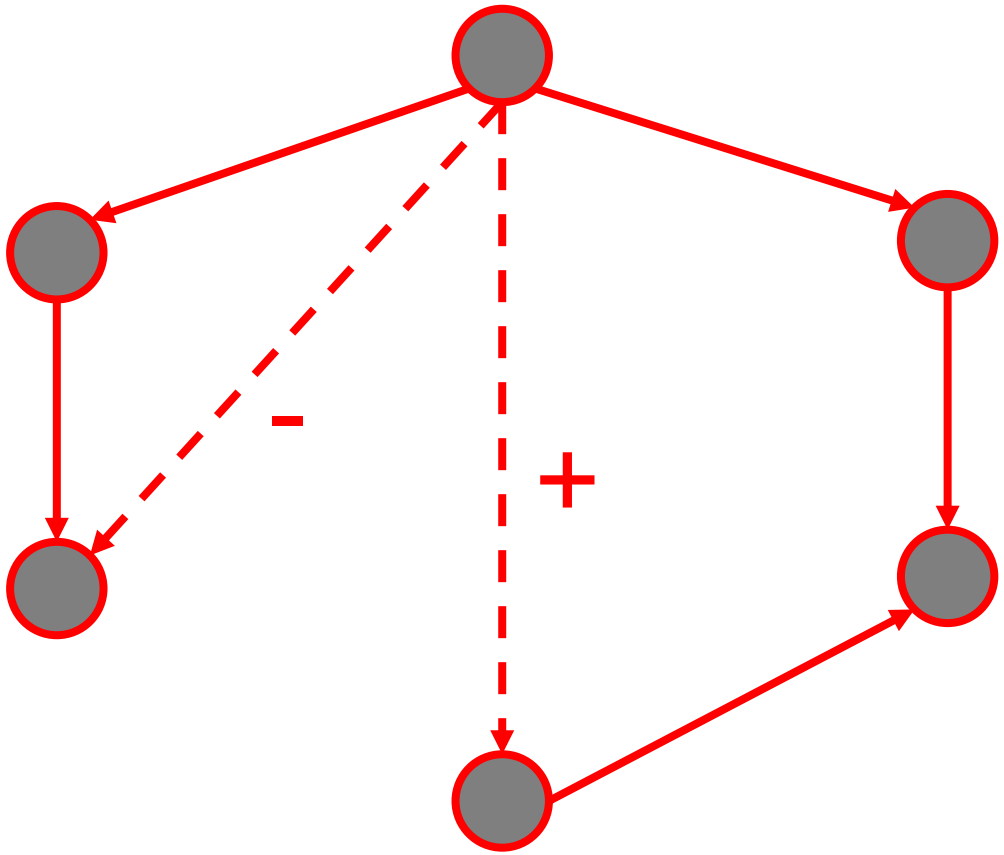
Motivación


Estructura de la red
Mecanismos endógenos
Atributos exógenos de los actores
Relaciones exógenas de en la red

Motivación

Mecanismos endógenos

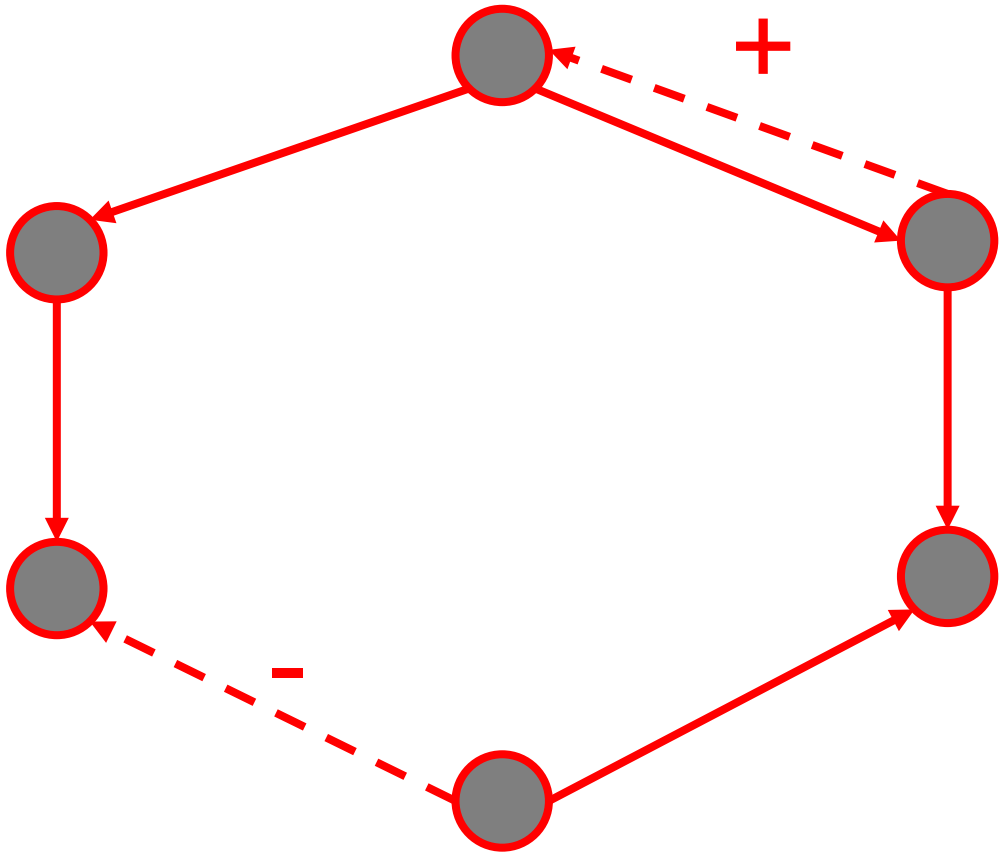
Teoría de vacíos estructurales



Motivación

Mecanismos endógenos

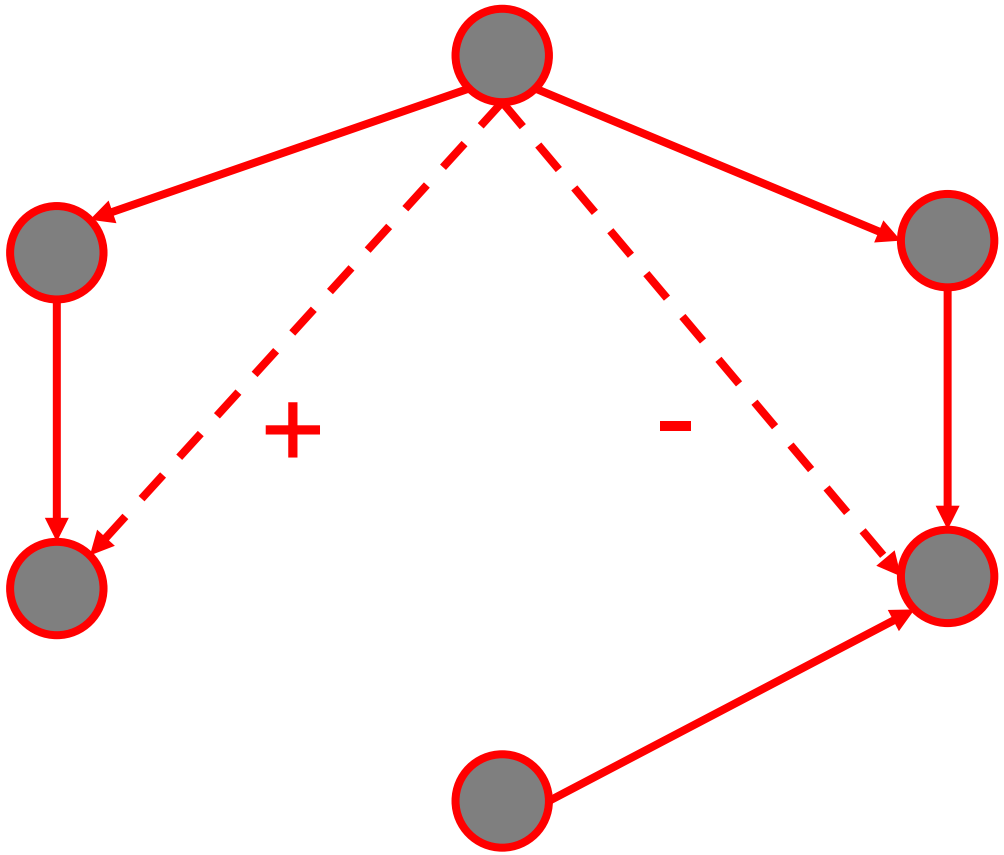
Teorías de intercambios



Motivación

Mecanismos endógenos

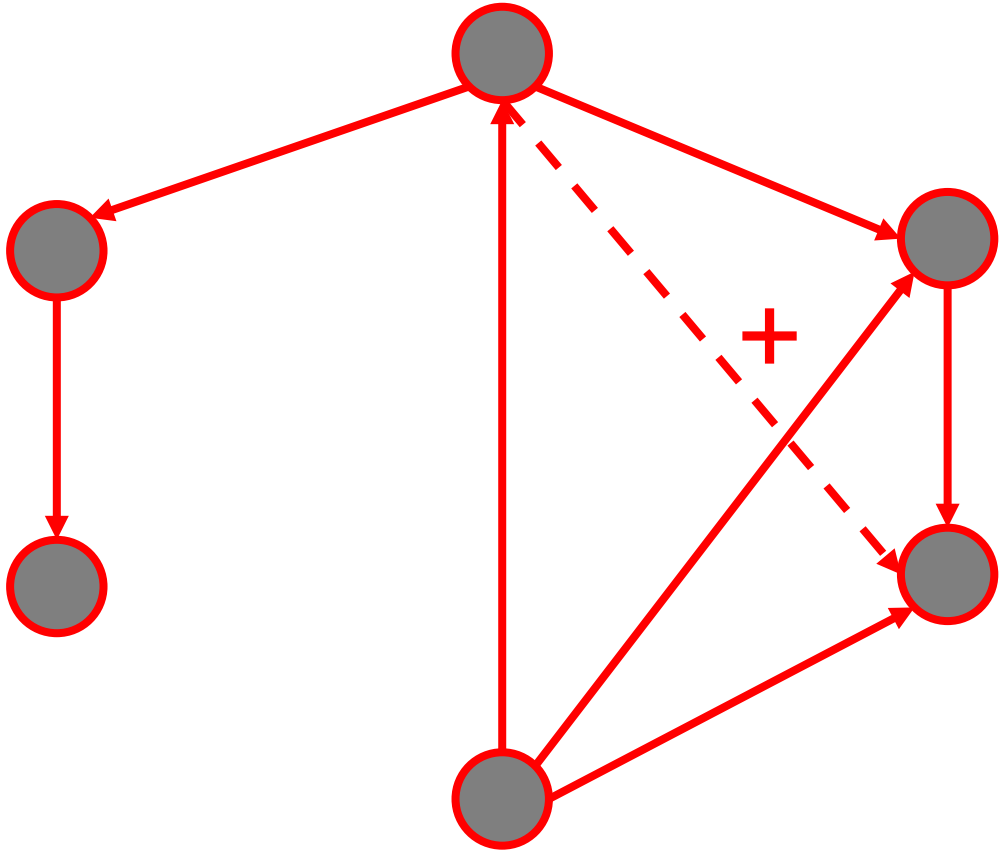
Teorías de balance



Motivación

Mecanismos endógenos

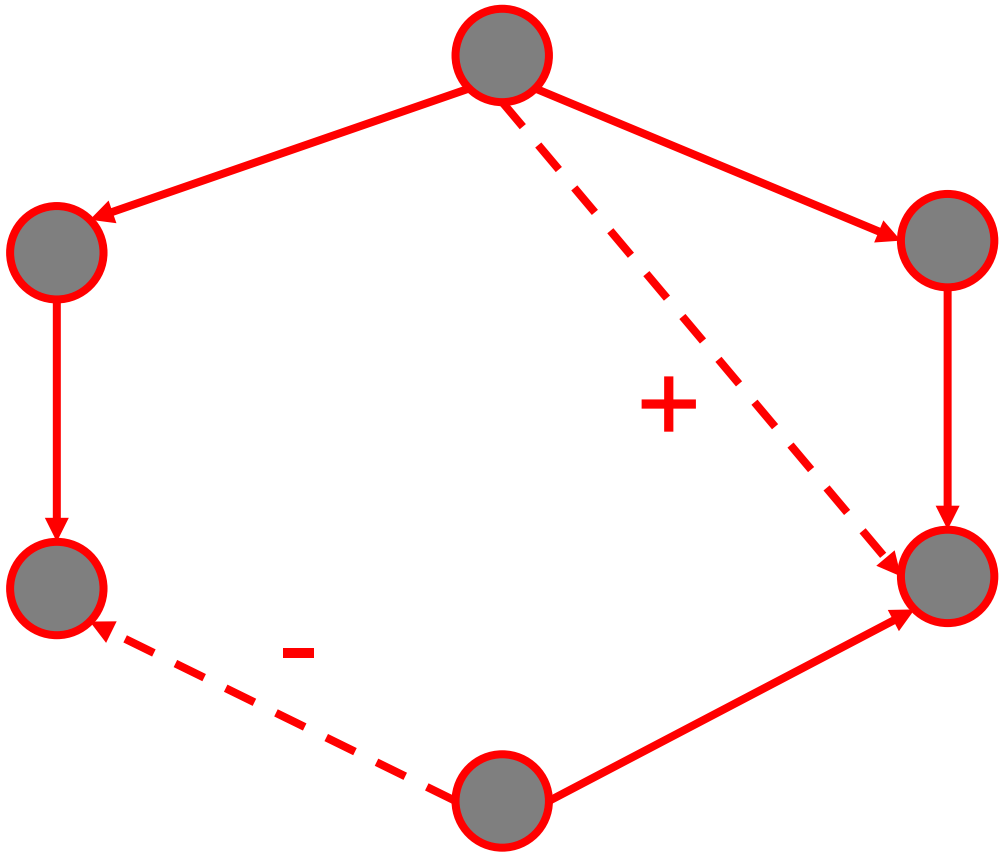
Teorías de cohesión



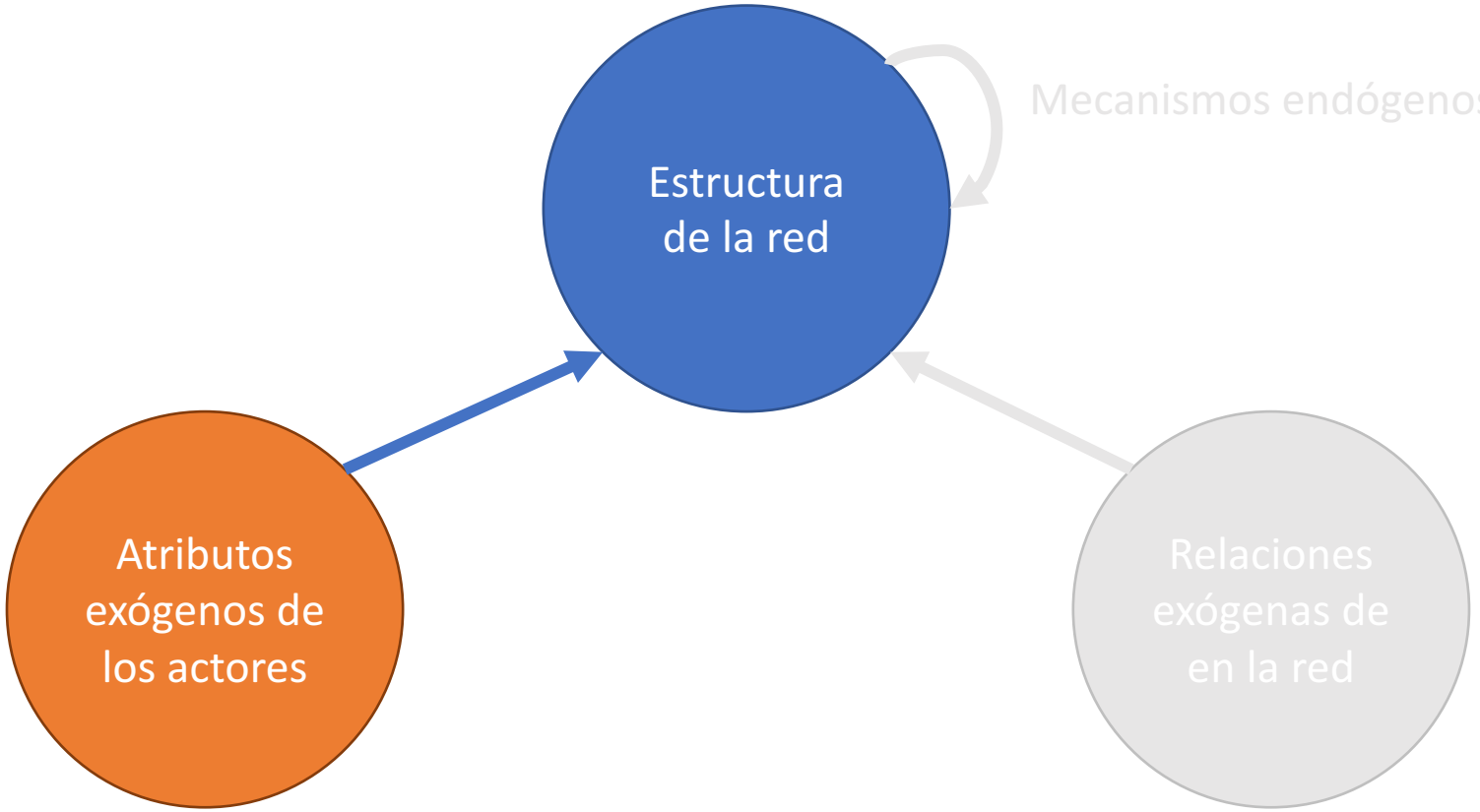
Motivación

Mecanismos endógenos

Acción colectiva



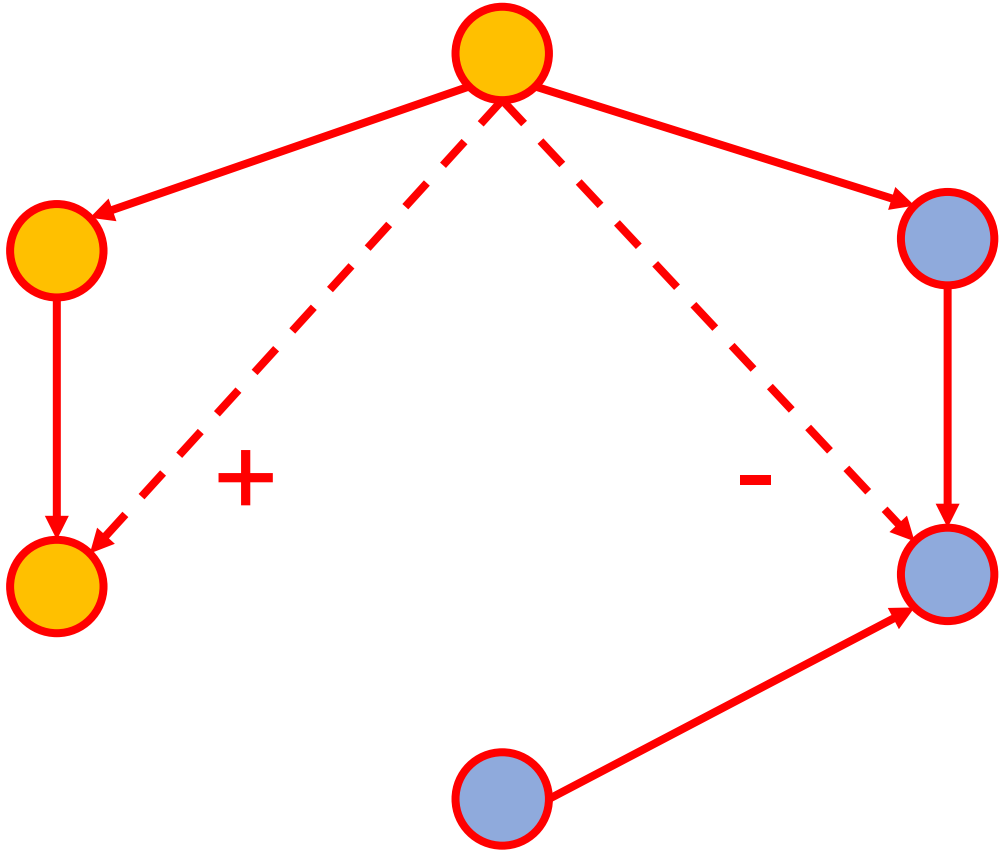
Motivación


Mecanismos endógenos
Estructura de la red
Atributos exógenos de los actores
Relaciones exógenas de en la red

Motivación

Atributos de los actores

Teorías de homofolia



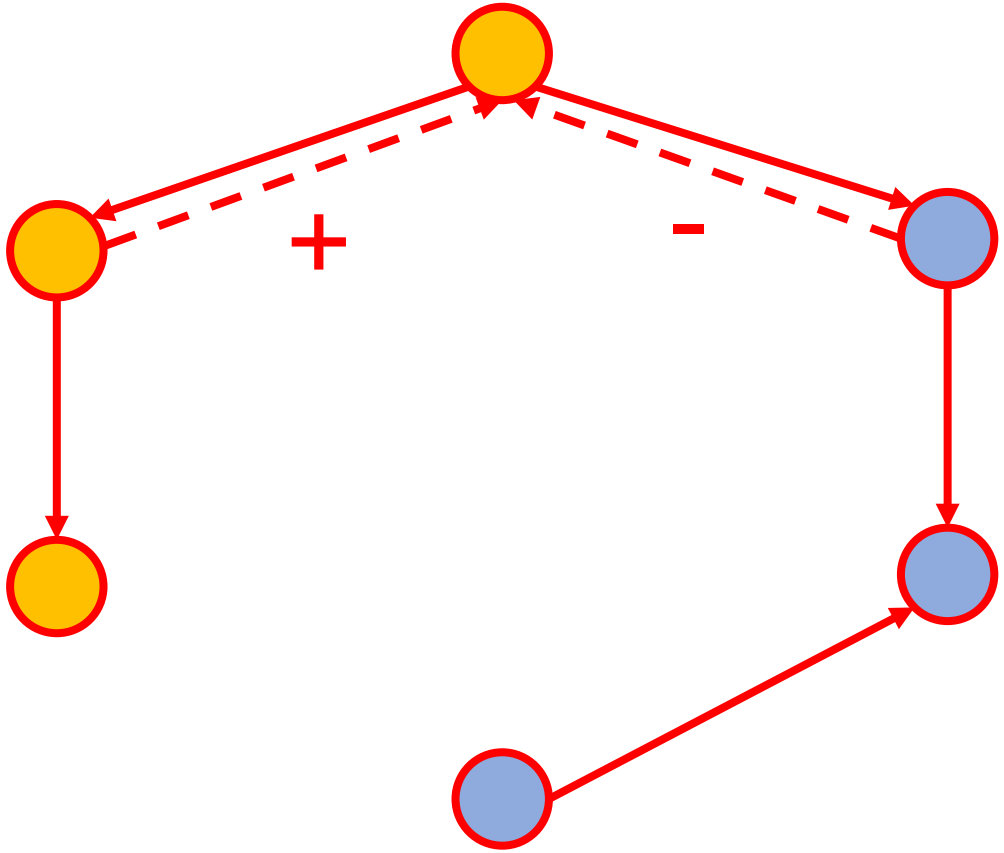
Universidad de Chile

Universidad Católica

Motivación

Atributos de los actores

Teorías de dependencia



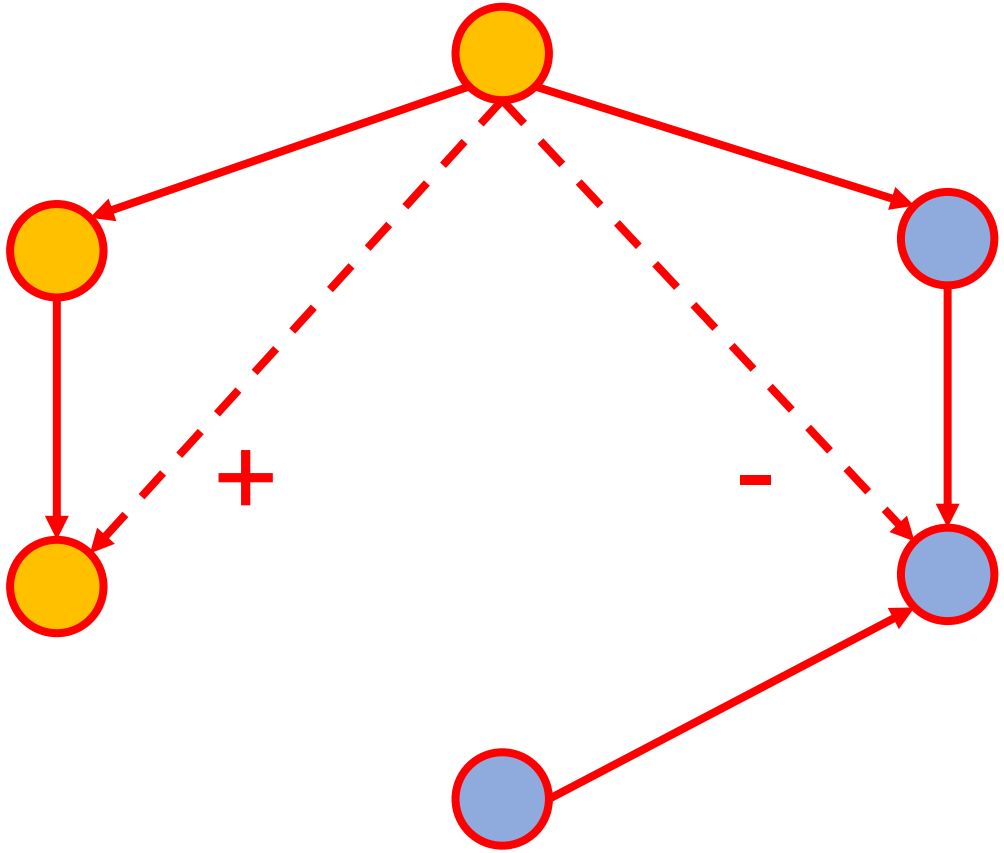
Universidad de Chile

Universidad Católica

Motivación

Atributos de los actores

Teorías de balance



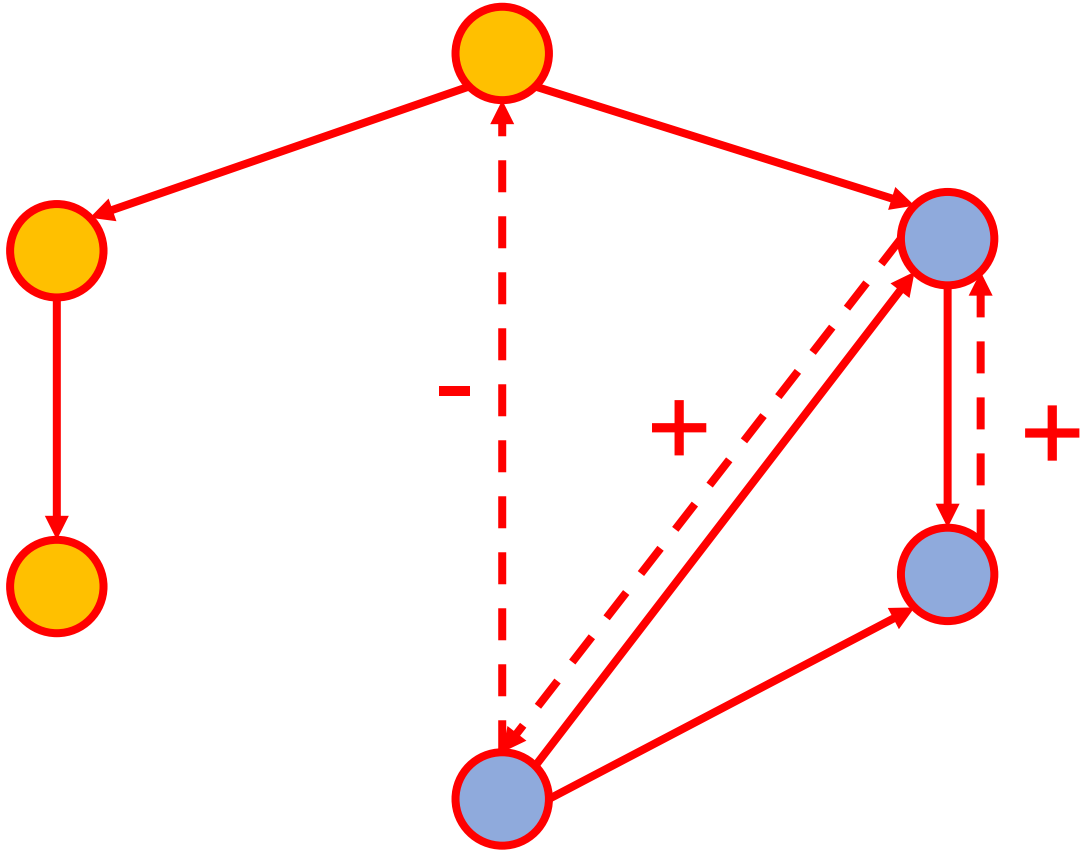
Universidad de Chile

Universidad Católica

Motivación

Atributos de los actores

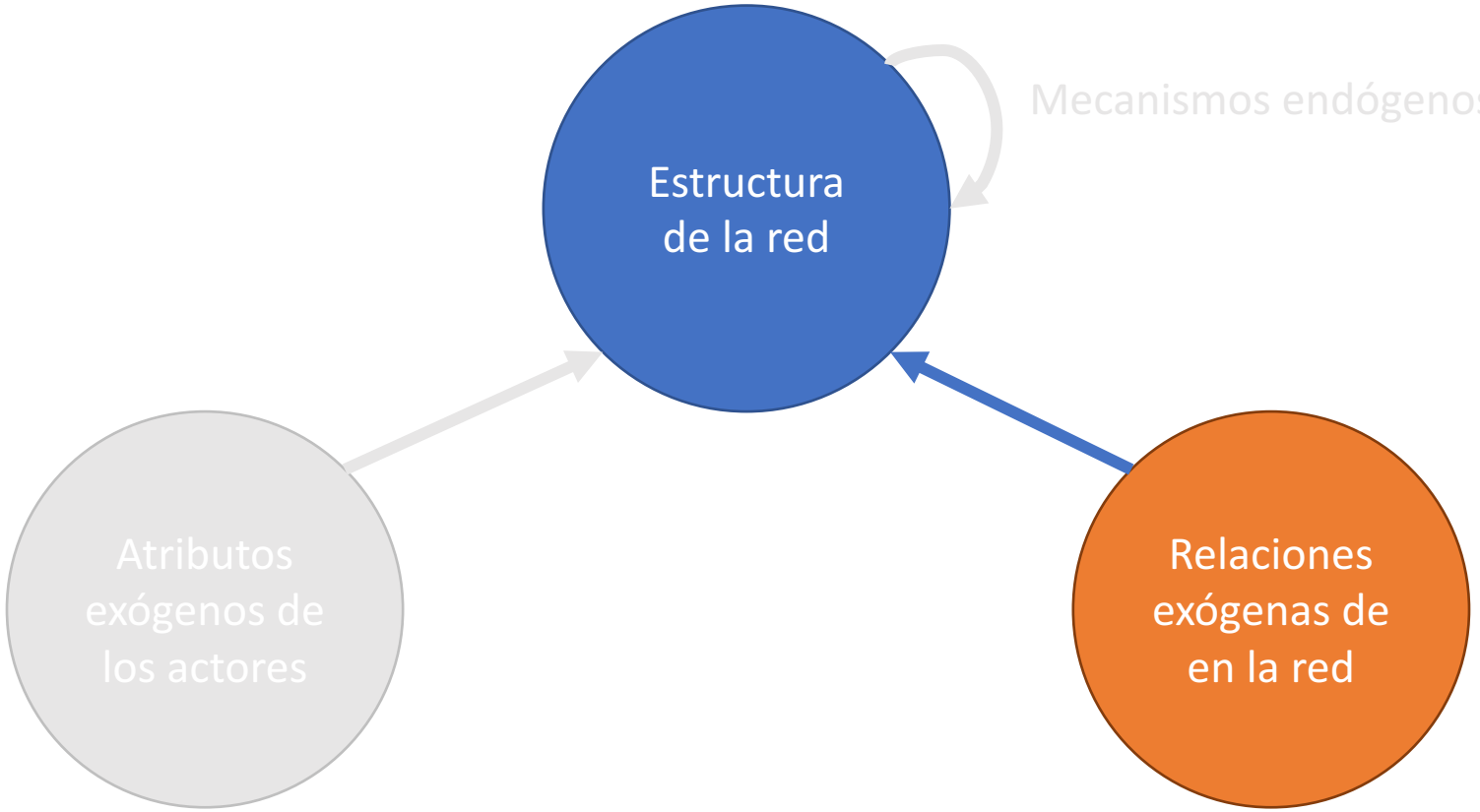
Teorías de cohesión



Universidad de Chile

Universidad Católica

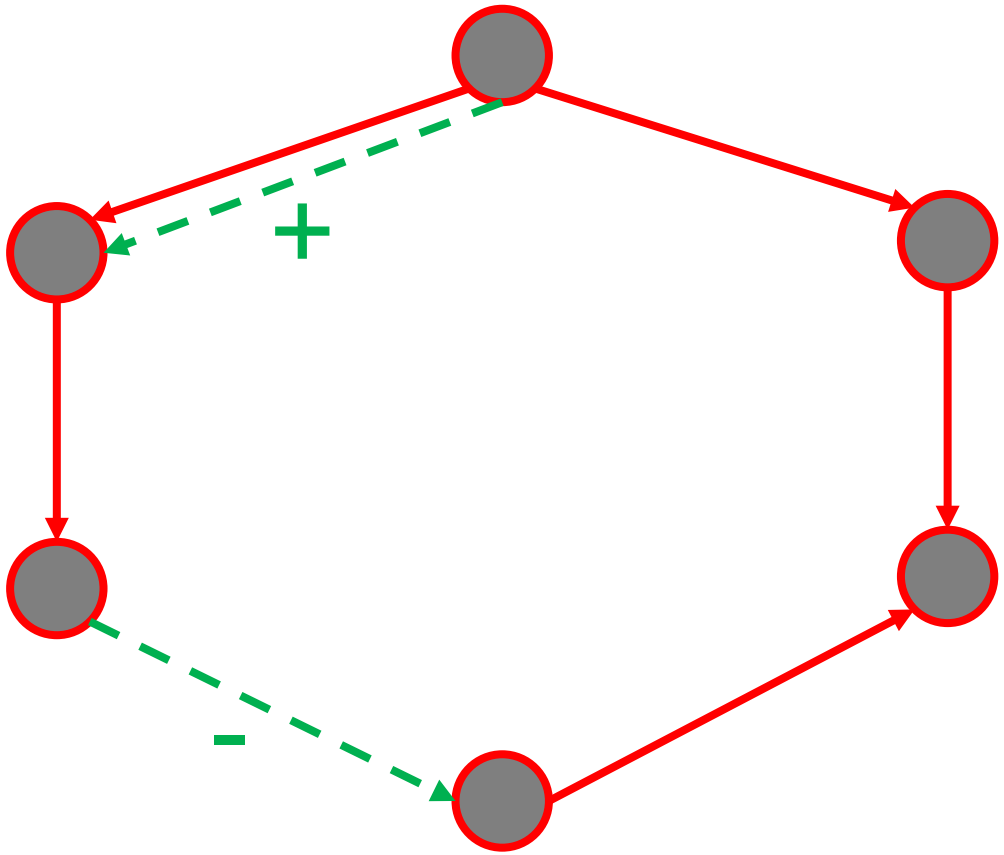
Motivación


Mecanismos endógenos
Estructura de la red
Atributos exógenos de los actores
Relaciones exógenas de en la red

Motivación

Relaciones endógenas

Teorías cognitivas



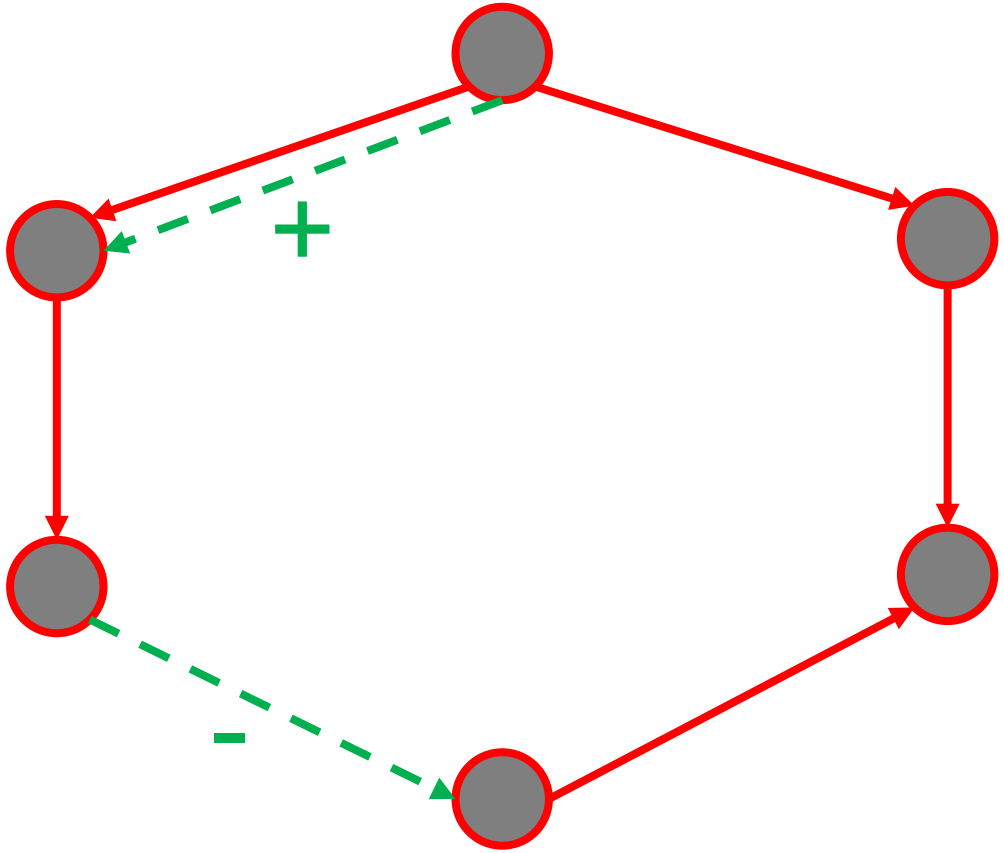
Comunicación

Amistad

Motivación

Relaciones endógenas

Teorías de evolución



Comunicación

Comunicación previa

Motivación

- El análisis de redes sociales posee muchos niveles: nodos, atributos, diadas, triadas, etc.
  - Si analizamos solo un nivel, cubrimos solo ciertos procesos.
  - Los atributos afectan la estructura, y la estructura afecta a los atributos
  - Necesitamos controlar los efectos exógeneos y endógeneos bajo un mismo marco.
- Buscamos un modelo estadístico que permita testear estos múltiples niveles de análisis.

Motivación

- Un modelo estadístico es una representación de procesos estocásticos especificados a ciertos niveles y que explican niveles superiores.
- ¿Por qué son útiles?
  - Logran estimar y parametrizar los procesos
  - Generan inferencias de una observación a otros casos
  - Buscamos representaciones adecuadas de las observaciones (Goodness of fit)

¿Por qué usar modelos estadísticos?

Métodos descriptivos

- Corresponden a un resumen de métricas.
  - Nivel de nodos: centralidad, distribución geodésica.
  - Nivel de configuración: censo de triadas.
  - Nivel de red: centralización, clustering, small-worlds, core/periphery.

Métodos generadores

- Sucesos micro que explican patrones macro
- Recuperan los procesos surgidos a partir de ciertas observaciones
- Testear hipótesis alternativas
- Extrapolar y generar simulaciones a partir del modelo.

¿Por qué usar modelos estadísticos?

- Para una red de n nodos, es computacionalmente complejo realizar todas las combinaciones en una red (directa o indirecta).
  - Red directa con 100
  - Posibles estados: $2^{100*99} = 2^{9900} >> 10^{80}$
- Además, la mayoría de las redes observadas en la realidad no se parecen a redes aleatorias.
  - Comparar redes observadas con aleatorias puede llevar a conclusiones equivocadas.
- Evaluar una teoría/hipótesis en base a un enfoque probabilístico
  - Generar distribuciones e intervalos de confianza

¿Por qué usar modelos estadísticos?

Modelos previos:

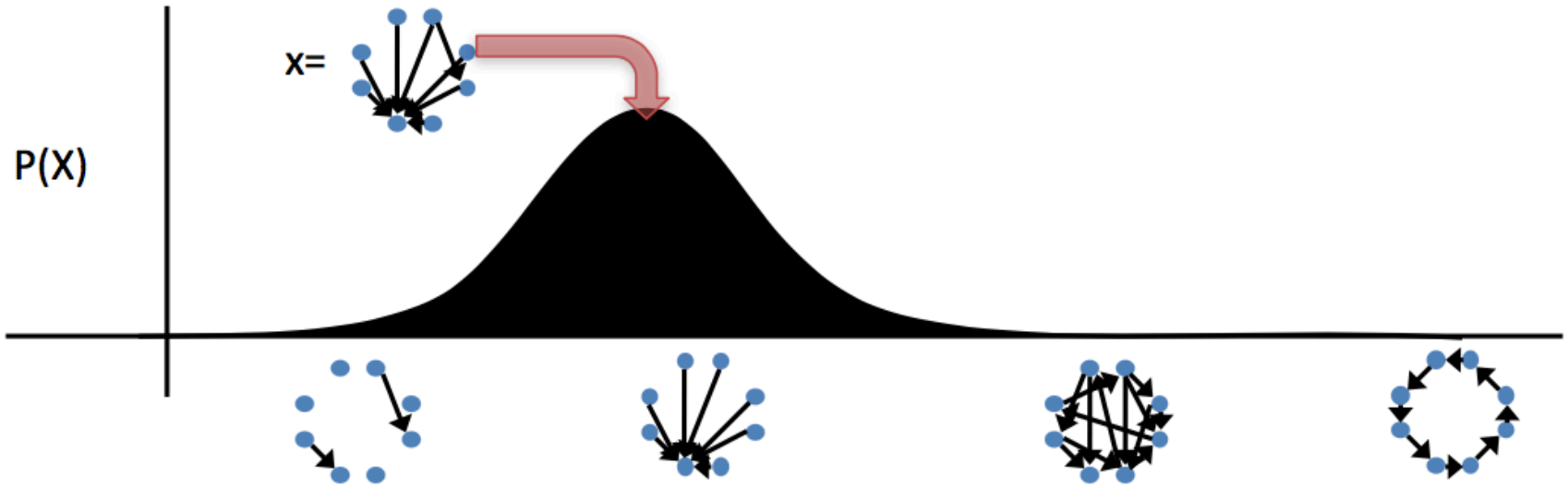
- Diadas independientes
  - Cada enlace entre una diada es independiente de otras
  - No considera dependencias
- Distribución de posibles redes
  - La red observada es una de tantas probables.
  - Problemas para generar más observaciones
- Distribución uniforme/Bernoulli
  - Cada enlace tiene la misma probabilidad de ocurrir.
  - Es un supuesto poco realista.

¿Por qué usar modelos estadísticos?

- La realidad es….
  - No hay independencia entre los distintos enlaces
  - La presencia/ausencia de un enlace está influida por la presencia/ausencia de otros enlaces
  - Diferentes *micro* procesos pueden llevar a *macro* procesos
- Se requieren métodos que puedan observar estos procesos sociales a nivel local, interactivo, y estocástico.

¿Por qué usar modelos estadísticos?

Las redes que exhiben aquellas características de interés deben tener probabilidades más altas que las redes que no exhiben aquellas características.



Descripción

$$P(Y = y) = \frac{1}{\kappa(\theta)} e^{\theta^T g(Y)}$$

- *Y*: la realización de una red, similar a una variable aleatoria.
- *y*: la red observada.
- *g(y)*: un vector de estadísticos de red.
- θ: un vector de coeficientes correspondientes a *g(y)*.
- $\kappa(\theta)$: escalar de normalización.

ERGMs son regresiones lineales

- El término es similar a los coeficientes y términos de un regresión lineal múltiple.

$$P(Y = y) = e^{\theta_1 x_1 + \cdots + \theta_n x_n}$$

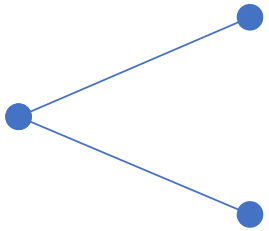
- Variable dependiente: La red observada
- Variables independinetes: Los parámetros y los coeficientes
- El intercepto corresponde al parámetro de los arcos

Parámetros

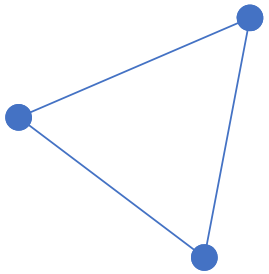
Caso indirecto

Arco

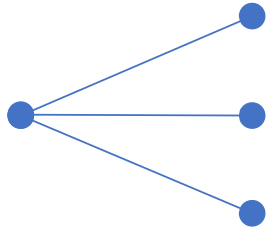
2-estrella



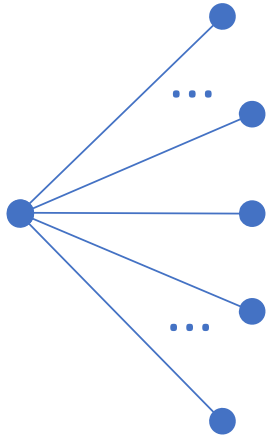
Triángulo



3-estrella



K-star

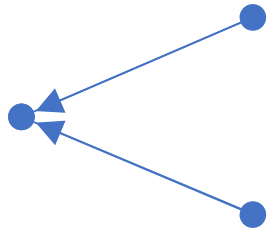


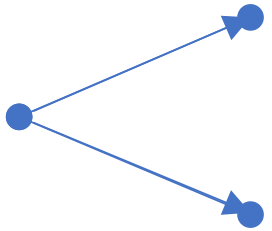
Parámetros

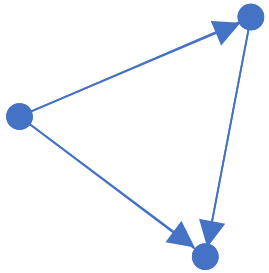
Caso directo

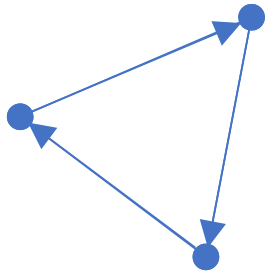
Arco 

Reciprocidad 

2-in-star 

2-out-star 

Transitividad 

Ciclo 

...

Ejemplo

$$P(Y = y) = \frac{1}{\kappa(\theta)} e^{\theta_1 * arco + \theta_2 * reciprocidad}$$

$$Red \sim \theta_1 * arco + \theta_2 * reciprocidad$$

Estimación

- En las regresiones lineales, típicamente se utiliza Mínimos Cuadrados Ordinarios para obtener los valores de los coeficientes.
- En ERGMs, no conocemos que valores de parámetros usar para asignar probabilidades
- Utilizando la red observada como una guía, el modelo estima los mejores valores utilizando el método de máxima verosimilutd (MLE)
  - Este método es frecuentemente utilizado en regresiones logísticas y de multinivel.

Interpretación de los valores

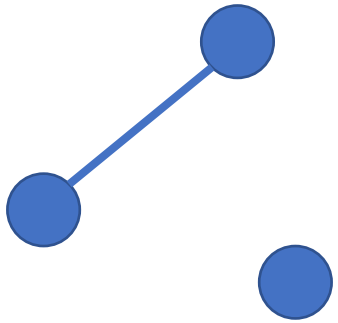
$\theta < 0$	$\theta = 0$	$\theta > 0$
El parámetro tiene menores probabilidades de ocurrir que lo esperado.	El parámetro tiene la misma probabilidad de ocurrir que lo esperado	El parámetro tiene mayores probabilidades de ocurrir que lo esperado

Ejemplo simple

- Modelo ERGM:

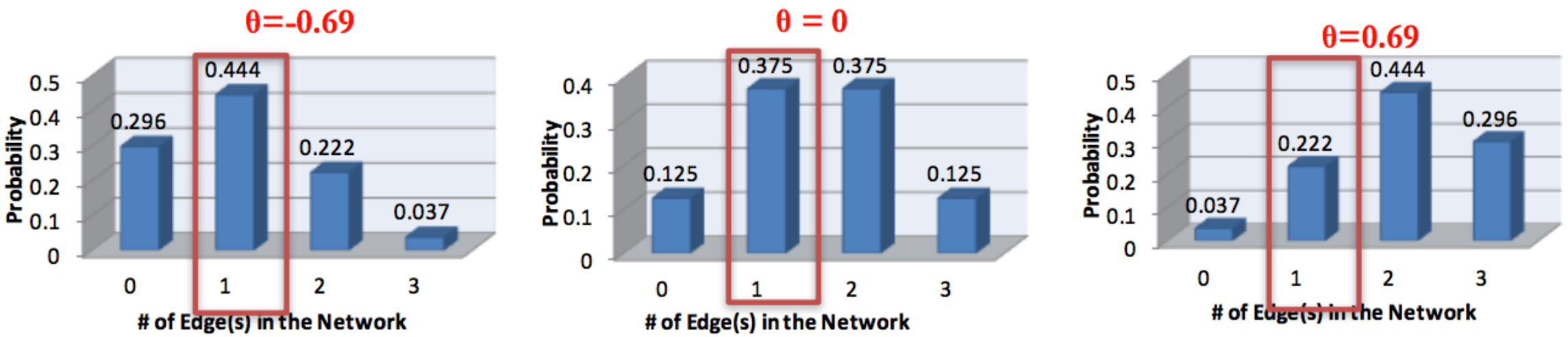
$$Red \sim \theta * arco$$

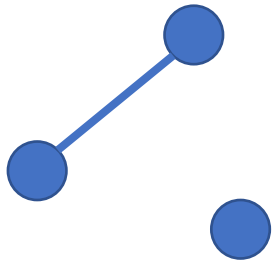
- Red observada:



- Dada la siguiente red, ¿Cuál es el mejor θ que explica el número probable de enlaces en la red observada?

Ejemplo simple

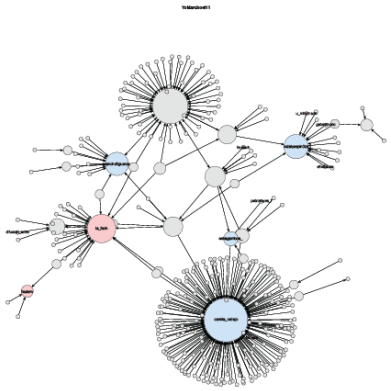


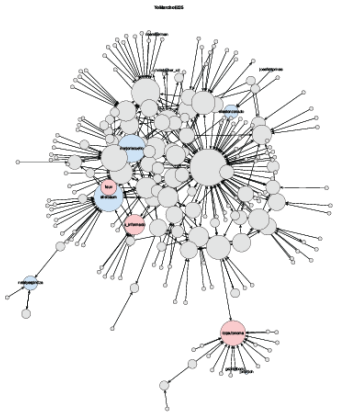


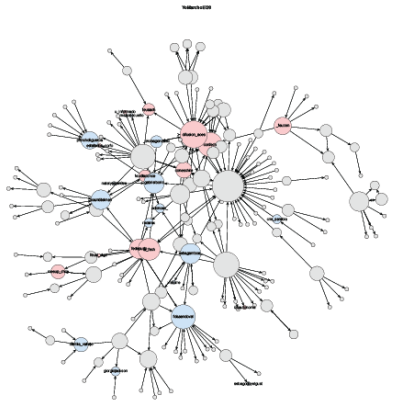
De los tres casos, θ= -0.69 explica mejor esta red.
Entrega la mayor posibilidad para la red observada

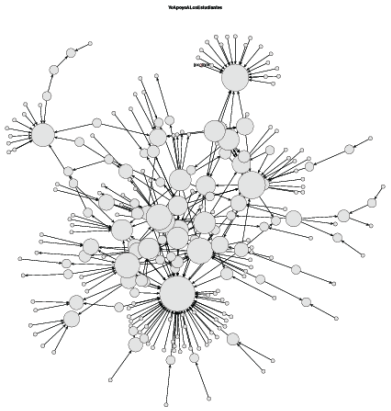
Ejemplo

Informational Networks

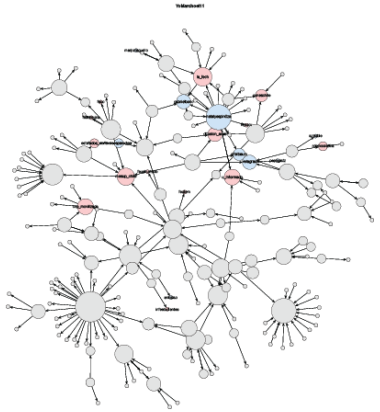


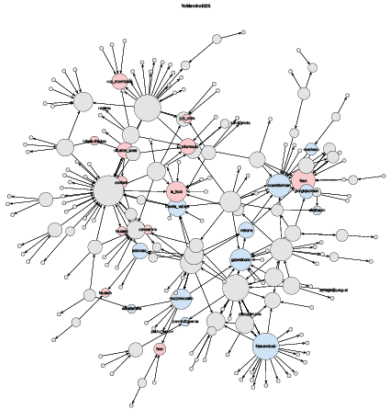


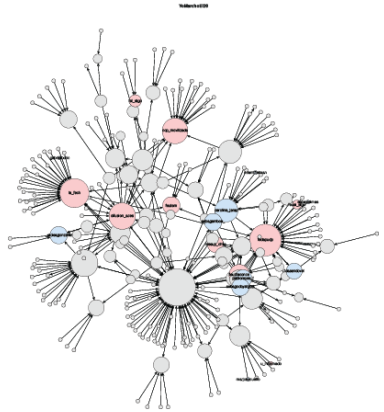


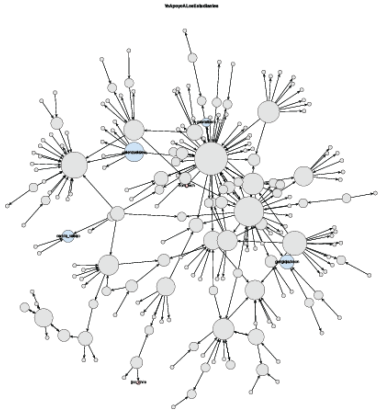


Conversational Networks









#YoMarchoel11 *#YoMarchoel25* *#YoMarchoel28* *#YoApoyoAlosEstudiantes*

Gómez-Zará et al. 2017, "The role of social movement organizations in Twitter: Evidence from the Chilean Student Movement"

Ejemplo: Red informacional

Gómez-Zará et al. 2017, "The role of social movement organizations in Twitter: Evidence from the Chilean Student Movement"

N = 4 hashtags, 1,077 user accounts, 1,382 edges; MLE Log likelihood: 85,924.56 (d.f.=14)

Controls	**Estimate (SD)**	**Odds-ratio**
Structure Effects		
Edges	-4.97 (0.07) ***	0.01
Reciprocity	0.33 (0.41) †	1.39
Popularity	-6.71 (0.18) ***	< 0.01
Activity	2.7 (0.13) ***	14.89
Hierarchical structure	1 (0.09) ***	2.71
Same targets	0.04 (0.01) ***	1.04
Actor Attributes		
Leader to Leader	-1.82 (0.54) ***	0.16
Organization to Leader	-0.68 (0.61) †	0.51
People to Leader	0.63 (0.05) ***	1.88
Leader to Organization	-0.32 (0.43) †	0.72
Organization to Organization	0.22 (0.52) †	1.24
People to Organization	0.41 (0.06) ***	1.51
Leader to People	-1.91 (0.37) ***	0.15
Organization to People	0.27 (0.27) †	1.31

*† $p < 0.1$, * $p < 0.05$, ** < p, 0.01, *** $p < 0.001$*

Ejemplo: Red conversacional

Gómez-Zará et al. 2017, "The role of social movement organizations in Twitter: Evidence from the Chilean Student Movement"

N = 4 hashtags, 831 user accounts, 1,031 edges; MLE Log likelihood: 64,685.67 (d.f.=14)

Controls	**Estimate (SD)**	**Odds-ratio**
Structure Effects		
Edges	-4.77 (0.06) ***	0.01
Reciprocity	2.07 (0.17) ***	7.93
Popularity	-0.44 (0.11) ***	0.65
Activity	-1.58 (0.11) ***	0.21
Hierarchical structure	0.91 (0.07) ***	2.47
Same targets	0.02 (0.01) **	1.02
Actor Attributes		
Leader to Leader	0.73 (0.14) ***	2.07
Organization to Leader	0.94 (0.25) ***	2.55
People to Leader	0.58 (0.09) ***	1.78
Leader to Organization	0.51 (0.28) †	1.67
Organization to Organization	0.38 (0.34)	1.46
People to Organization	1.05 (0.08) ***	2.86
Leader to People	0.02 (0.11)	1.02
Organization to People	-0.84 (0.2) ***	0.43

*† $p < 0.1$, * $p < 0.05$, ** < p, 0.01, *** $p < 0.001$*

Lista de pasos

1. Especificar el modelo con parámetros $g(Y)$ a evaluar al evolución del gráfico.
2. Utilizando la red observada, usar el método de máxima verosimilitud (MLE en inglés) para estimar los coeficientes θ de los parámetros $g(Y)$ utilizados en el modelo.
3. Simular redes aleatorias basadas en el modelo
    1. Generar una muestra de redes aleatorias basadas en esas reglas.
    2. Las redes simuladas "deberían verse" como la observada.
    3. Utilizamos técnica de Markov-Chain Monte Carlo
4. Realizar *Goodness of fit* de las redes simuladas respecto a las redes observadas

Pasos a seguir

1. Examinar la red y sus propiedades
2. Crear un modelo nulo
3. Crear un modelo de base
4. Crear un modelo endógeno
5. Crear un modelo exógeno

Pasos a seguir

1. Examinar la red y sus propiedades

- Antes de comenzar a modelar, es bueno analizar la estructura de la red. Podemos realizar esto a partir de visualizaciones.
- Explorar:
  - Número de enlaces (densidad)
  - Número de vértices
- Chequear la distribución de los grados (in-out)
  - Como el modelo tomará el mismo número de vértices y simulará diferentes combinaciones de enlaces, las distribuciones de grado deben ser similares a la red observada.

Pasos a seguir

2. Crear un modelo nulo

- Comenzamos creando el modelo más simple de interés: un solo parámetro que todos los enlaces tienen la misma probabilidad de existir.
- También conocido como modelo Bernoulli o grafo Erdös-Rènyi.
- Este modelo nos permite que parámetros seleccionar posteriormente.

Pasos a seguir

2. Crear un modelo nulo

- El objetivo de la simulación es ver si el modelo estimado captura las características de la red observada
- Generamos 100 simulaciones a partir del modelo creado.
- Comparamos la distribución de los parámetros de redes (grado, estrellas, triángulos, etc) en las simulaciones con respecto a la red observada.

Pasos a seguir

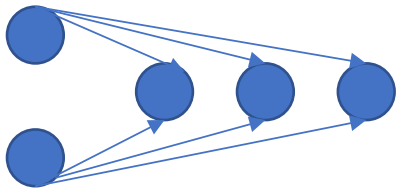
3. Crear un modelo base

- Generalmente, debemos corregir las distribuciones de los grados. Para eso añadimos el término para el grado (in/out).
- Usualmente, añadimos un término de decaimiento a estos términos.
  - Le otorga mayor relevancia a los nodos con menor grado.
  - Asegura mayores posibilidades de convergencia del modelo

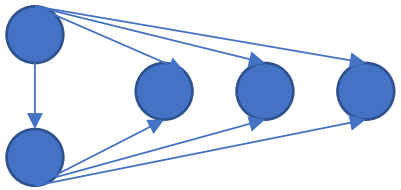
Pasos a seguir

4. Crear un modelo endógeno

- Añadimos los términos sugereidos por el modelo de Robins et al.'s (2007) para redes directas:
  - Densidad (edges)
  - Reciprocidad (mutual)
  - Diadas: Compañeros compartidos (gwdsp)



  - Triadas: Nodos unidos, compañeros compartidos (gwesp)



- Para los factores de decaimiento, se recomienda partir con valores pequeño y estándares: 0.25, 0.5, 1.0, 2.0, etc.

Pasos a seguir

5. Crear un modelo exógeno

- Finalmente, añadimos los términos exógeneos que puedan explicar la red observada, y sean parte de nuestra hipótesis planteadas.
- Términos frecuentes
  - Nodematch: homofilia
  - Nodemix: homofilia cruzada
  - Nodecov: atributo del actor
  - Edgecov: atributo de los enlaces
  - Dyadcov: atributo de las diadas

Demo

Utilizar archivo “ergm_tutorial.R”

Limitaciones

- Redes gigantes (> 2,000 nodos)
- Convergencia
- Inclusión de nuevos actores
- Dimensión temporal
  - Otros modelos posibles: t-ergms, Siena Models, Relational Event Models.

Referencias

- Monge, P. R., & Contractor, N. S. (2003). *Theories of communication networks*. Oxford University Press, USA.
- Lusher, Dean, Johan Koskinen, and Garry Robins. *Exponential random graph models for social networks: Theory, methods, and applications*. Cambridge University Press, 2012.
- Goodreau, Steven M., et al. "A statnet Tutorial." *Journal of statistical software* 24.9 (2008): 1.
- Hunter, David R., et al. "ergm: A package to fit, simulate and diagnose exponential-family models for networks." *Journal of statistical software* 24.3 (2008).
- Robins, Garry, et al. "Recent developments in exponential random graph (p*) models for social networks." *Social networks* 29.2 (2007): 192-215.
- EUSN 2014 Workshop using statnet.

¡Muchas gracias!

Agradecimientos especiales al Profesor Noshir Contractor, Yun Huang, y al equipo de SONIC Lab.

Email: dgomezara@u.northwestern.edu | Twiter: @dgzara