ANÁLISIS ESTADÍSTICOS CON R. APLICACIONES

UNIVERSIDAD DE GRANADA

Proyecto Fin de Máster

Autor:

Salvador Nebro Ramos

Tutora:

Yolanda Román Montoya

Junio 2022

Resumen

En el documento que se presenta a continuación se ha tratado el análisis de redes a partir de una base de datos.

En primer lugar se encuentra un capítulo en el que se encuentra recogido los principales apuntes teóricos partiendo desde conceptos básicos, pasando por modelos matemáticos que dan razón de ser al modelo estadístico con el que se concluye, el Modelo de Gráfico Aleatorio Exponencial.

El segundo capítulo está compuesto por los principales paquetes que se han empleado para la construcción del modelo a través del Software R.

Por último, se realiza el análisis de un conjunto de datos formado por un conjunto de bandas callejeras y las relaciones existentes entre ellas para la distribución de dogras en Montreal, con el que se ha querido observar si se cumplen ciertas hipótesis partiendo del Modelo de Gráfico Aleatorio Exponencial.

Índice general

Capítulo 1

Marco teórico

1.1. Introducción.

Según la (RAE, 2020) una **red** "*es un conjunto de elementos organizados para determinado fin o un conjunto de personas relacionadas para una determinada actividad.*" En el análisis de redes, dependiendo de la red que estemos estudiando, usaremos una definición u otra, ya que se puede analizar una red de transporte, la organización para la aprobación de leyes, las relaciones sociales existentes entre individuos, entre otras muchas. Sin embargo, las redes más estudiadas a lo largo de la historia son las **redes sociales**.

El estudio de las relaciones que existen entre los individuos de una comunidad particular y su comportamiento dentro de la misma es algo que ha creado gran interés en el ámbito de las Ciencias Sociales y que se ha ido caracterizando con el tiempo.

El **Análisis de Redes Sociales o Social Network Analysis** (ARS) no surge como aislado, nuevo y disconexo, sino que se conforma a través de la aportación de los estudios realizados por diferentes ciencias. Esto se aprecia en trabajos como el realizado por Jacob Moreno, psicólogo que explicó la influencia mutua entre sus pacientes mediante gráficos de redes. Para el sociólogo Harrison White tanto la sociedad como el mercado tienen que ser observados como una red de conexiones. Clave fue la aportación de Mark Granovetter, discípulo de Harrison White, quien formuló una teoría en la que se apreciaba la influencia que tienen las redes en la vida social, como la fortaleza de los vínculos débiles y la inserción de la acción económica. El principal aporte de las matemáticas a todo esto es la teoría de grafos que determina diversos conceptos y procesos para explicar las relaciones entre los datos (Cardenas, 2020).

Por lo que podemos observar, el Análisis de Redes Sociales no se trata de algo aislado creado por alguien, sino que se ha ido enriqueciendo a partir de las aportaciones de investigadores pertenecientes a diferentes ámbitos. Por ello, podemos decir que el análisis de redes es un aparato metodológico y técnico que da lugar a una teoría interdisciplinar que se encuentra consolidada y expandida dentro de las Ciencias Sociales. La idiosincracia de dicha teoría es la fundamentación de modelos teóricos basados en elementos matemáticos y el análisis de datos

empíricos. Pero el desarrollo del Análisis de Redes Sociales no se entiende sin tener en cuenta los avances experimentados en la tecnología informática, el álgebra, la topología y la teoría de grafos. La particularidad que posee dicho instrumento es que mediante modelos teóricos formales que cobran forma gracias a las matemáticas y el estudio de datos de carácter empírico, se ha logrado llevar a la práctica lo que se conoce como estructura social. Este hecho cobra gran importancia, pues en ciertas escuelas y tradiciones pertenecientes al campo de las Ciencias, por ejemplo la Sociología o la Antropología, era un concepto principal (Garrido, 2001).

En base a ello se puede afirmar que el desarrollo del Análisis de Redes tiene un papel importante en las Ciencias Sociales, ya que aporta un instrumento nuevo en base a un concepto muy extendido, de manera que se puede aplicar de manera plausible a otros ámbitos como la organización política, los acuerdos a los que se llega en el Senado, el contagio de enfermedades, las relaciones entre personas u organizaciones, la difusión de bulos entre otras muchas aplicaciones como el deporte.

Como ya se ha comentado, las matemáticas son claves para el Análisis de Redes. Gracias al desarrollo de las ramas nombradas se sentaron las bases para poder llevarlo a cabo, pero hoy en día, cobra gran importancia la rama de Estadística. Su importancia reside en que permite obtener información previa acerca de los datos que estamos analizando, crear modelos adecuados que se ajusten de manera adecuada a dichos datos y obtener tanto conclusiones como predicciones sobre qué va a ocurrir con dicho modelo en función de las variables utilizadas. Llevándolo al campo del Análisis de Redes, la Estadística permite estudiar las relaciones existentes sobre un conjunto de individuos en función de ciertas características que van a permitir hallar relaciones posibles entre esos individuos sin olvidar las ya existentes desde un primer momento.

En el trabajo realizado en este documento se parte de una serie de conceptos teóricos y modelos entre los que vamos a usar el modelo estadístico que se describe para analizar la organización de las operaciones de distribución de drogas en Montreal Norte. Ello se llevará a cabo a través del software estadístico R, del cual también se describen los paquetes y funciones utilizadas.

1.2. Conceptos básicos.

Según se recoge en Dunn (1983), una red se puede entender como un conjunto de actores, que en el lenguaje de teoría de grafos serían llamados nodos, conectados por una relación específica, dicha relación las identificamos con las aristas. El análisis de redes es una herramienta que nos va a permitir interpretar cómo es la estructura de una red, dicha red puede ser social, política, de transporte, entre otros. Desde el punto de vista matemático, si es posible llevar a acabo una representación gráfica de la red, ésta se corresponde con un grafo.

Debido a que actualmente la ciencia pretende estudiar objetos o sistemas en los que nos encontramos con gran cantidad de variables o hay que llevar a cabo muchas interacciones para concretar la relación existente entre ellas, lo más adecuado es simplificar la realidad y así poder sacar conclusiones acerca de lo que estamos estudiando. De esta manera aparecen modelos matemáticos a partir de los cuales es posible llevar a cabo una abstracción y posterior modelización de la realidad, con la complejidad inherente en el proceso. .

En general, el proceso se inicia a partir de la observación de una base de datos. Es la estructura de la base y la medición de variables adecuadas la que permitirá establecer y analizar el conjunto de relaciones derivadas del análisis. Es preciso por tanto iniciar el estudio con la descripción de las componentes de la base, que darán lugar al establecimiento de las componentes de la red. dicho análisis descriptivo conlleva varias etapas resumen, gráficos y medidas de resumen. Lo cual podemos extrapolar a las redes:

- Identificación de elementos que conforman la red.
- Número de relaciones entre ellos.
- Representación gráfica, si es posible.
- Medidas que resumen la red.

Para entender todo lo que se recoge en este documento, es conveniente dejar claro qué estructura poseen las redes. Para ello vamos a seguir a Garrido, (2001).

Existen ciertas características formales que se tienen más en cuenta que otras a la hora de construir una red. Una de ellas es la **intensidad relacional**, la cual aporta información sobre la posición que posee cierto nodo en la red, sabiendo ésto, se obtiene la posibilidad y capacidad de acción de dicho nodo. Dicha característica hace referencia a las relaciones que se establecen con ese nodo y viene dada en función del tamaño de la red. Por lo que depende de la **densidad** de la red, que es la proporción de relaciones que presenta respecto a todas las posibles, y el **grado** de la red, que son las relaciones que existen por nodo en término medio. Ambos conceptos se pueden particularizar a los nodos directamente. La importancia de esta característica reside en que se usa para la estimación de la **centralidad** de las posiciones de cada nodo. Ya que el nodo que se considera central, va a ser el más influyente a la hora de establecer relaciones dentro de la red. Dicho término admite dos medidas diferentes:

- Diremos que un nodo es **central** según aparezca en las relaciones. Nótese que dicho término no debe confundirse con **jerarquía**, que hace referencia al prestigio de una posición.

- El segundo tipo tiene en cuenta las conexiones indirectas, es decir, se tiene en cuenta aquellos nodos entre los que no existe una relación directa pero sí que se puede llegar de uno a otro pasando por una serie de nodos. Al camino más corto para establecer esa relación entre los nodos se le llama **geodésica**. En este caso, la centralidad viene dada en función al número de geodésicas que pasan por dicho nodo.

Los nodos que actúan como intermediarios ganan peso en la red, ya que sin ellos a la hora de establecer relaciones entre nodos que no están relacionados directamente, el coste sería mayor.

Hemos visto el papel de los nodos de manera individual, pero en las redes es habitual que se den subgrupos en los que conviene hallar la posición de los nodos que los forman. Para ello existen dos métodos, la detección de camarillas y el enfoque de la equivalencia estructural.

El primero se basa en que los nodos se organizan según las relaciones que existen entre ellos, esto se denomina **cohesión social**, y las **camarillas** se producen cuando todos los nodos del subgrupo están relacionados directamente entre sí. Pero ésto no es lo habitual, sino que suele haber una subred con alta densidad, que se denomina **círculo social**. Su principal inconveniente es que no considera la red al completo. Por ello, se suele utilizar la **equivalencia estructural**. Dicho método se basa en que dos nodos son estructuralmente equivalentes si poseen relaciones similares con los demás nodos de la red sin tener en cuenta si existe una relación directa entre ambos.

Una vez que se ha establecido la estructura que posee una red, se van a presentar una serie de conceptos que se deben de tener en cuenta, ya que de haber representación gráfica de la red, ésta se identifica con un grafo (Kolaczyk y Csárdi, 2014; Barber, 2018) :

- En primer lugar, se entiende por **grafo** a la estructura matemática dada por el par $G = (V, E)$, donde V es el conjunto de los vértices (nodos) y E es el conjunto de las aristas (relaciones).

- El **orden** de un grafo es el número de vértices que contiene y se denota por n_v.

- El **grado** de un vértice se refiere al número de aristas que llegan a él.

- Un grafo es **regular** si todos los vértices tienen el mismo grado.

- Llamamos grafo **dirigido** a aquel en el que las aristas tienen un sentido definido.

- Diremos que un grafo es **simétrico** si es dirigido y si existe una arista de u a v, entonces existe otra de v a u.

- Un grafo **reflexivo** es aquel en el que $\forall\, v \in V$, existe la relación $(u, u) \in E$.

- Se dice que un grafo **transitivo** si dadas (u, v), $(v, z) \in E \Rightarrow (u, z) \in E$.

- Un grafo **completo** es aquel que para cualquier par de nodos $u, v \in V$, existe la relación $(u, v) \in E$.

- Un **camino** de v_0 a v_i es una sucesión $\{v_0, e_1, v_1, e_2, \ldots, v_{i-1}, e_i, v_i\}$, donde los extremos de e_k son los vértices $\{v_{k-1}, v_k\}$. Diremos que la **longitud** de dicho camino es i. Este concepto se puede refinar tomando un camino en el que no se repite ninguna arista y recibe el nombre de **recorrido**. Si el camino empieza y acaba en el mismo vértice diremos que tenemos un **circuito**. En caso de que los vértices del camino sean distintos dos a dos diremos que se trata de un camino **simple**.

- Si nos encontramos ante un grafo no dirigido en el que para cualquier par de nodos existe un camino que los une, diremos que tenemos un grafo **conexo**. Si el grafo es dirigido, entonces tendremos un grafo **fuertemente conexo**.

- El concepto de **distancia** entre vértices en un grafo se define como la longitud del camino o caminos más cortos entre los vértices. Esta distancia se suele denominar **distancia geodésica**, siendo "geodésica" otro nombre para los caminos más cortos.

- Un grafo conexo sin ciclos se llama **árbol**. La unión disjunta de tales se denomina **bosque**. Los árboles tienen una importancia fundamental en el análisis de redes. Ya que se usan, por ejemplo, como estructura de datos clave en el diseño eficiente de muchos algoritmos computacionales. Cuando en un árbol existe un único vértice desde el que hay un camino dirigido a todos los demás vértices del grafo se denomina **raíz**.

Para una red en la que nos encontramos con relaciones directas, lo usual es representarla mediante un grafo dirigido y su **matriz de adyacencia** $X = X_{ij}$, donde el elemento X_{ij} representa el vínculo existente entre el nodo i y el nodo j. Por lo que los elementos de dicha matriz son de la forma:

$$X_{ij} = \begin{cases} 1 & \text{si existe una arista de i a j} \\ \\ 0 & \text{si no existe una arista de i a j} \end{cases}$$

Puesto que otros nodos pueden influir en las relaciones existentes entre dos de ellos, los modelos no pueden construirse a partir de supuestos de independencia amplios. Algunos efectos que se producen sobre la red son (Huisman y Snijders, 2003):

- **Efecto de densidad**: es la tendencia que tienen los elementos de una fila a ser 1 en lugar de 0.

- **Efecto de reciprocidad**: la dependencia dentro del par (X_{ij}, X_{ji}) dada por el número de relaciones recíprocas.

- **Efecto de transitividad**: se refiere a la dependencia dentro de las ternas (nodos y sus relaciones entre ellos), dada por el número de relaciones transitivas.

- **Equilibrio**: la dependencia entre las aristas que salen de i y las de otros nodos con los que está relacionado, muestra la preferencia por otros que toman las mismas decisiones que i.

- **Efecto de relaciones indirectas**: es la dependencia entre i y otro nodo a través de uno intermedio y viene dado por el número de nodos con los que i está indirectamente relacionado (a distancia dos).

- **Popularidad**: se refiere a la dependencia de un elemento X_{ij} con los demás elementos de la fila X_j, dada por la suma de los grados de las aristas de entrada de los otros j con los que i está relacionado.

- **Actividad**: se refiere a la dependencia de un elemento X_{ij} con los demás elementos de la fila X_j, dada por la suma de los grados de las aristas de salida de los otros j con los que i está relacionado.

Debido a que en las redes encontramos una cantidad de información que resulta imposible de estudiar nodo a nodo, habría que hallar una serie de características comunes para poder compararlas entre sí y elegir un modelo eficiente a la hora de modelar. El problema es que, a día de hoy, no existe dicho conjunto de características que identifiquen por completo a cada red, aunque sí que existen una serie de propiedades que son importantes estudiar (Sancho, 2016):

- **Longitud promedia de caminos.** Las definiciones respecto a dicho concepto que tenemos que saber son:

  - **Distancia.** Recordemos que es la longitud del camino entre los vértices.
  - **Diámetro.** Es la distancia máxima existente entre dos nodos.
  - **Longitud promedia.** Es la media de las distancias entre todos los pares de nodos.

    Se observó que la longitud promedia de los caminos que se formaban en la mayoría de redes reales complejas era menor respecto a lo esperado, aún cuando el número de enlaces era más reducido. Dicha característica se conoce como **efecto de mundo pequeño**.

- **Coeficiente de clustering.** Siguiendo a Sancho (2016) *"es la proporción media de pares de vecinos de un nodo que también son vecinos entre sí."*

  Sea el nodo i y supongamos que tiene k_i vecinos en la red, por lo que como máximo existen $k_i(k_i-1)/2$ conexiones entre ellos. El coeficiente de clustering, C_i, viene dado por el cociente entre las conexiones que existen, E_i, y todas las posibles, es decir,

  $$C_i = \frac{E_i}{k_i(k_i-1)/2}$$

  En cuanto al **coeficiente de clustering de la red**, C, lo definiremos como la media de los coeficientes de los nodos que la forman, por lo que $C \leq 1$ y únicamente alcanza el valor 1 si la red es completa. Se ha comprobado que las redes reales tienden a la agrupación,

ya que C es bastante más grande que $O(1/N)$, siendo N el número de nodos. De ambos conceptos se obtiene que la mayoría de redes reales no son ni aleatorias ni completas.

- La **Distribución de grados** de los nodos viene dada en términos de una distribución de probabilidad, $P(k)$, con la que se establece la probabilidad de que, seleccionado un nodo al azar, este tenga k conexiones. Si la red es regular, todos los nodos tienen las mismas conexiones y se dice que la distribución es de tipo delta. Si existe aleatoriedad, dicha distribución será de Poisson, que decrece de manera exponencial conforme nos alejamos del grado medio de la red.

Más allá de preguntarnos cómo es una red observada y de caracterizar su estructura a un nivel más fundamental, podemos estar interesados en comprender cómo puede haber surgido. Por lo general, se distingue entre dos formas de modelizar las redes, mediante modelos matemáticos o estadísticos. Con **modelos matemáticos** nos referimos principalmente a aquellos dados mediante reglas probabilísticas simples que generan una red para intentar concluir que se cumple cierto mecanismo o principio. Mientras que los **modelos estadísticos** pretenden ajustarse a los datos que se tienen, con lo que se puede evaluar la influencia de ciertas variables a la hora de proporcionar las relaciones dentro de la red. Además, el ajuste se realiza y evalúa usando los principios formales de la inferencia estadística. Aunque es cierto que existen similitudes entre ambos modelos (Kolaczyk y Csárdi, 2014).

Sin embargo, desde el punto de vista estadístico los modelos matemáticos son muy simples como para poder ajustarse a los datos reales de una red. Aunque, además de para tener una visión formal de cómo se establecen los mecanismos para la formación de las relaciones y su influencia sobre la red, también se usan para definir clases nulas en las redes que permiten evaluar la "importancia" de las características estructurales en una red. Sin más, procedamos a explicar los distintos modelos matemáticos y estadísticos existentes.

1.2.1. Modelos matemáticos

Consideremos el conjunto $\{\mathbf{P}_\theta(G),\, G \in \mathcal{G} : \theta \in \Theta\}$, donde $\mathcal{G}$ es el conjunto de posibles grafos, $\mathbf{P}_\theta$ es una distribución de probabilidad sobre G y θ es un vector de parámetros que toma valores en Θ.

Este tipo de modelos tiene diversos usos, como la comprobación de la significación de algunas características predefinidas en el grafo de red o la evaluación de posibles valores de predicción para las relaciones. El buen uso de dichos modelos depende en gran medida de la elección de $\mathbf{P}_\theta$.

Este tipo de modelos utilizados de manera matemática suelen ser más simples y modificables para realizar el análisis. Pero también es posible que no se puedan aplicar técnicas estadísticas formales de ajuste y evaluación del modelo, (Kolaczyk y Csárdi, 2014; Sancho, 2016; Cordón, 2013).

Modelo de red aleatoria clásico

Los modelos de red aleatoria clásicos hacen referencia a un conjunto de grafos G que vienen dados junto con una probabilidad $\mathbf{P}$ uniforme sobre G. Dicho modelo se basa en la teoría de Erdös-Rényi. De modo que hacemos referencia al conjunto $\mathcal{G}_{N_v,N_e} = \{G = (V, E) : \mid V \mid = N_v, \mid E \mid = N_e\}$, con probabilidad $\mathbf{P}(G) = \binom{N}{N_e}^{-1}$, donde $N = \binom{N_v}{2}$ es el total de parejas de vértices distintas que podrían existir.

Una variante que suele usarse más en la práctica es la aportada por Gilbert, en la que los grafos de orden N_v pueden obtenerse asignando una arista independiente a cada par de vértices distintos con probabilidad $p \in (0, 1)$. Este modelo se denomina modelo de red aleatoria de Bernoulli si p está bien definida en N_v y $N_e \sim pN_v^2$, siempre que N_v sea grande.

Los grafos que forman las redes no tienen porqué ser conexos, pero por lo general, cuando $p = \dfrac{c}{N_v}$, con $c > 1$, diremos que es muy probable que exista una **componente gigante**, (que son grafos aleatorios en las que cualquier arista puede unir un par de vértices de manera independiente a las otras y con probabilidad p), que contenga a la mayoría de los nodos. Si tenemos que $c > 0$ y N_v es grande, la distribución de los grados se aproximará mediante una distribución de Poisson con media c. Lo que es evidente, ya que el grado de cualquier vértice se distribuye según una binomial con parámetros $N_v - 1$ y p.

Modelo de red aleatoria generalizado

El modelo de Erdös-Rényi se puede generalizar considerando aquel conjunto de grafos en los que se tiene un orden fijo, N_v, alguna característica determinada y cada grafo tiene asignada la misma probabilidad. La característica que se suele elegir es tomar una secuencia de grados fija $\{d_1, \ldots, d_{N_v}\}$. Además, como N_v es fijo, todos los grafos tienen el mismo número de aristas, pues se tiene que cumplir $\bar{d} = 2N_e/N_v$, donde $\bar{d}$ es el grado medio de $\{d_1, \ldots, d_{N_v}\}$.

Modelo de red de mundo pequeño

Los impulsores del modelo de red de mundo pequeño fueron Watts y Strogatz, los cuales estaban intrigados debido a que muchas redes del mundo real presentaban altos niveles de clustering pero pequeñas distancias entre los nodos. Lo cual no podía ocurrir con el modelo de red aleatoria clásica, ya que el diámetro de escala era de O(log N_v), por lo que indicaba nodos muy

juntos, y el coeficiente de clustering se comportaba como N_v^{-1}, lo que induce un bajo clustering.

El modelo presentado por Watts y Strogatz parte de un conjunto de nodos N_v, dispuestos de forma periódica, y se une cada vértice a r de sus vecinos a cada lado. Entonces, para cada arista tomada al azar y con probabilidad p, se hace incidir uno de sus extremos con cualquier otro nodo de manera uniforme, evitando aristas múltiples y bucles.

El efecto de reconducir un número relativamente pequeño de aristas de forma aleatoria hace que se reduzca notablemente la distancia entre los vértices, manteniendo un nivel similar de clustering. Es decir, tanto el coeficiente de clustering como la longitud del camino promedio podemos tomarlos como funciones de la probabilidad de reconexión p.

Modelos libres de escala

Una de las motivaciones de la introducción de este mecanismo fue el deseo de reproducir los tipos de distribuciones de grado amplias que se observan en muchas redes grandes del mundo real. Sin embargo, son Barabási y Albert quienes asientan el interés por dichos modelos. Barabási y Albert se vieron motivados por el crecimiento de la World Wide Web, observando que las páginas web que poseen gran cantidad de conexiones, tienen mayor probabilidad de tener aún más conexiones. Por ejemplo, al redactar un artículo, se tiende a citar otros artículos que han sido reconocidos, por lo que dichos artículos serán bastante más citados que otros menos reconocidos.

El modelo de Barabási y Albert para grafos no dirigidos se construye como sigue. Partimos de un grafo $G^{(0)}$ que posee $N_v^{(0)}$ nodos y $N_e^{(0)}$ aristas. En cada etapa el grafo $G^{(t-1)}$ se modifica añadiendo un vértice de grado $m \geq 1$ y dando lugar a $G^{(t)}$, donde las m nuevas aristas están unidas a m vértices diferentes procedentes de $G^{(t-1)}$, y la probabilidad de que ese nuevo vértice se una al vértice v viene dada por

$$\frac{d_v}{\sum\limits_{v' \in V} d_{v'}}$$

Después de t iteraciones, el algoritmo da lugar a una red con $N_v^{(t)} = N_v^{(0)} + t$ nodos y $N_e^{(t)} = N_e^{(0)} + tm$ conexiones, dando lugar a una red libre de escala en la que la distribución de los grados no cambia con el tiempo. La distribución de grados, cuando t tiende a infinito, sigue una ley de potencias d^{-3}, en otras palabras, la probabilidad de hallar un nodo con d conexiones es proporcional a d^{-3}.

1.2.2. Modelos estadísticos

Los modelos vistos hasta ahora resultan realmente útiles para ciertas aplicaciones, pero no están diseñados explícitamente para ser modelos estadísticos. En este caso, nos encontramos con tres modelos principalmente de los cuales nos vamos a centrar y explicar con detalle uno de ellos (Kolaczyk y Csárdi, 2014; Van der Pol, 2019; Rodríguez-Bocca, 2019).

Modelos de grafos aleatorios exponenciales

Los modelos de grafos aleatorios exponenciales surgen por la necesidad de realizar un análisis de redes en base a medidas estructurales para obtener conclusiones acerca de la formación de dichas redes. Para ello, se incluye en el modelo cierta variable que explica la red y el ERGM se encarga de aportar significación estadística sobre la variable de manera parecida a los modelos de regresión lineal, aunque se expresará la diferencia existente entre ambos modelos más adelante. Las variables a las que se hace referencia se van a denominar estructurales y exógenas.

Se llamarán variables estructurales a las que estudian la presencia de relaciones dando lugar a un modelo cuya configuración se basa en la estructura de la red estudiada. Estas variables pueden ser la reciprocidad, los triángulos o los subgrafos.

Por su parte, las variables exógenas son aquellas que no son determinantes a la hora de construir la red pero sí que pueden tener gran influencia a la hora de dar lugar a relaciones. Por ejemplo, el género o la edad.

Algo que hay que tener muy claro a la hora de saber cómo funcionan los ERGM y sus configuraciones, es que los supuestos que se tienen que cumplir para la dependencia entre vínculos son más exigentes que los de los modelos lineales. Ya que hay que saber definir qué configuraciones locales se van a poder observar y estimar mediante el modelo.

En concreto, los ERGM poseen gran similitud de la manera en la que se construyen con los modelos de regresión logística modificada. Los cuales hacen que las relaciones que se dan en la red, venga determinadas en función de las ya existentes. Por su parte, en los ERGM podemos considerar las interacciones, tanto ponderadas como dirigidas que se producen entre los nodos. Con los ERGM también se puede hallar la distribución de probabilidad de una red, por lo que se puede tomar una muestra de redes de gran tamaño con las que se podrá hacer inferencia sobre las probabilidades de cierta relación en la red (Palacios y Villalobos, 2016).

- **Modelización**
  Sea el grafo $G = (V, E)$ y consideremos la variable aleatoria binaria $Y_{ij} = Y_{ji}$ que indique la pertenencia de cierta arista $e \in E$ entre el par de nodos i, j. Llamemos $Y = [Y_{ij}]$ al conjunto de todas las matrices de adyacencia de G que contienen la conexión entre los nodos i y j . Sea $y = y_{ij}$ una posible red perteneciente al conjunto Y. Un modelo de grafo aleatorio exponencial es un modelo específico de la familia exponencial para la distribución conjunta de los elementos de Y. La ecuación de la forma canónica de un ERGM viene dada por

$$\mathbf{P}_\theta(Y = y) = \left(\frac{1}{\kappa}\right) exp\left\{\sum_H \theta_H g_H(y)\right\},$$

  donde:

i) H es una *configuración* dada por un conjunto de posibles aristas entre un subconjunto de los vértices de G.

ii) $g_H(y) = \prod_{y_{ij} \in H} y_{ij}$, por lo que es 1 si todos los vértices y aristas de H están en y o 0 en otro caso.

iii) un valor no nulo de θ_H significa que los Y_{ij} son dependientes en H para todos los pares de vértices $\{i, j\}$ condicionados al resto del grafo.

iv) $\kappa = \kappa(\theta)$ es una constante normalizada

$$\kappa(\theta) = \sum_{y} exp\left\{\sum_{H} \theta_H g_H(y)\right\},$$

en el que el sumatorio se realiza sobre todas las configuraciones posibles. Nótese la importancia de esto ya que predetermina la elección de las g_H y los coeficientes θ_H, por lo que se tiene una estructura con cierta (in)dependencia de los elementos de Y, lo que resulta interesante dado el carácter intrínsecamente relacional de una red. De manera general, dichas estructuras se caracterizan porque las variables aleatorias $\{Y_{ij}\}_{(i,j)\in A}$ son independientes de $\{Y_{i'j'}\}_{(i',j')\in B}$, condicionadas a los valores $\{Y_{i''j''}\}_{(i'',j'')\in C}$, para A, B, y C conjuntos de índices. Inversamente también se puede partir de una serie de conexiones (in)dependientes entre subconjuntos de Y y tratar de hallar la forma que se induce para g_H y θ_H.

Los EGRM permiten incluir variaciones o extensiones. Cuando tratamos con grafos dirigidos o no dirigidos podemos incluir información sobre los nodos más allá de su conectividad, ello se consigue incluyendo unas variables llamadas **variables de nodo o atributos de nodo**, que se basan en la idea de que la probabilidad de que dos nodos estén conectados dependa del atributo del nodo. Dada una realización x de un vector aleatorio X en los vértices de G, se especifica una forma exponencial para la distribución condicional $\mathbf{P}_\theta(Y = y|X = x)$, que implica características adicionales sobre $g(-)$, que son funciones tanto de y como de x.

Los ERGM nos permiten formular modelos muchos más específicos que el que acabamos de dar, pues es desde el que partimos. El problema surge a la hora de modelizar, por ejemplo, si asumimimos que la red es homogénea, es decir, todos los θ_{ij} valen θ, y que la presencia o ausencia de una arista entre un par de nodos es independiente del posible estado de esa arista, tendríamos que

$$\mathbf{P}_\theta(Y = y) = \left(\frac{1}{\kappa}\right) exp\left\{\theta L(y)\right\},$$

donde $L(y) = \sum_{i,j} y_{ij} = N_e$. Este modelo puede tomarse como una especificación de que la matriz de adyacencia de una red determinada G, es proporcional al número de aristas

de la red. Usando el modelo de grafo aleatorio de Bernoulli se tendría que $p = \dfrac{\exp\theta}{1+\exp\theta}$.

A lo largo del tiempo, lo usual ha sido añadir también estadísticos análogas de la estructura global de la red de orden superior. Por ejemplo, el número de k-stars $S_k(y)$ y el de triángulos $T(y)$. Usando dichas características y teniendo en cuenta que si dos aristas comparten un vértice, vamos a decir que son dependientes y que la condición de la arista Y_{ij} viene determinada por cualquier otra arista que involucre a i o a j, tenemos el concepto de **dependencia de Markov** y diremos que se trata de un *grafo aleatorio de Markov* sí y sólo sí

$$\mathbf{P}_\theta(Y=y) = \left(\frac{1}{\kappa}\right) exp\left\{\sum_{k=1}^{N_v-1}\theta_k S_k(y) + \theta_T T(y)\right\},$$

El inconveniente de estos modelos es que no suelen incluir efectos de estrella para k mayor o igual que cuatro ya que a la hora de realizar estimaciones, éstas resultan muy complejas. Por ello, son modelos que no se ajustan a la realidad de los datos que se suelen tener. Para paliar esta carencia, (Snijders, Pattison, Robins, y Handcock, 2006) proponen introducir una restricción paramétrica dada por $\theta_k \propto (-1)^k\lambda^{2-k}$ sobre los parámetros estrella $\forall\ k \geq 2$, con $\lambda > 1$, en una única estadística k-estrella alternativa:

$$AKS_\lambda(y) = \sum_{k=2}^{N_v-1}(-1)^k\frac{S_k(y)}{\lambda^{k-2}}$$

Alternativamente, si en dicho modelo se considera el número de aristas, el modelo que acabamos de dar es equivalente a un modelo de grado ponderado geométricamente dado por

$$GWD_\gamma(y) = \sum_{d=0}^{n_v-1} e^{-\gamma d}N_d(y),$$

con $N_d(y)$ el número de vértices de grado d y $\gamma = log\left[\dfrac{\lambda}{\lambda-1}\right] > 0$.

Si en lugar de considerar las k-estrellas como la estructura triádricas, tomamos los k-triángulos, nos quedaría un estadístico de la forma

$$AKT_\lambda(y) = 3T_1 + \sum_{k=2}^{N_v-2}(-1)^{k+1}\frac{T_k(y)}{\lambda^{k-1}},$$

donde T_k es el número de k-triángulos, siendo un k-triángulo un conjunto de k triángulos que comparten una base.

Nótese que el desarrollo y las especificaciones que se están proporcionando acerca del modelo únicamente contempla la relación existente entre dos nodos, es decir, si existe o no una arista entre ellos para establecer la red. Ahora bien, lo más usual es que la probabi-

lidad de establecer una arista entre dos vértices no solo dependa del estado, sino también de los atributos propios de los vértices. Para poder incluir dichos atributos al modelo, debemos tenerlos medidos, pues lo que se va a producir es una modificación sobre la constante normalizada κ.

Lo cual se va a realizar de la manera que se presenta a continuación

$$g(y, x) = \sum_{1 \leq i < j \leq N_v} y_{ij} h(x_i, x_j),$$

con h función simétrica de x_i y x_j, donde x_i es el vector de atributos observados del vértice i-ésimo. Por tanto, se puede entender que h mide la similitud total entre los vecinos de la red.

Podemos definir ficha función h d dos maneras distintas, haciendo que se produzcan efectos principales y efectos de segundo orden de manera análoga sobre los atributos.

- **Efectos principales:** Se define para un único X y viene dada por $h(x, x) = x + x$.
- **Efectos de segundo orden:** Viene dada a través de un indicador de equivalencia del atributo perteneciente a dos vértices, $h(x_i, x_j) = I\{x_i = x_j\}$.
  Con dichos efectos se intenta averiguar la tendencia que tienen los nodos a establecer relaciones con otros cuyos atributos toman valores parecidos.

Los ejemplos dados son algunos de los muchos efectos que se producen por los estadísticos introducidos al modelar los gráficos de red utilizando ERGMs. En particular, se va a especificar el modelo

$$\mathbb{P}_{\theta,\beta}(Y = y | X = x) = \left(\frac{1}{\kappa(\theta, \beta)} \right) exp\left\{ \theta_1 S_1(y) + \theta_2 AKT_\lambda(y) + \beta^T g(y, x) \right\},$$

donde g es el vector de atributos y β el de parámetros.

Mediante esta especificación, podemos controlar la densidad de la red, algunos efectos de la transitividad y se puede evaluar las consecuencias que tiene el establecer de ciertas relaciones.

Ajustes del modelo

Los modelos exponenciales traen consigo un problema asociado, el cual es obtener el valor que toman los parámetros usando las observaciones. Este fenómeno en Estadística toma el nombre de Inferencia Paramétrica, y el método más usado para la obtención de parámetros es

el que se llama método de máxima verosimilitud (MLE). Cuyos estimadores son de la forma

$$L(\theta, x_1, ..., x_m) = \prod_{i=1}^{m} P_\theta(x_i),$$

donde $x_1, ..., x_m$ son independientes que siguen un modelo aleatorio cuya distribución es P_θ. Con dicho método lo que se pretende es obtener un θ que maximice la función L.

Si suponemos que partimos de un modelo exponencial en forma canónica, dicho modelo puede ajustarse mediante el método de máxima verosimilitud (MLE). Lo que da lugar a un estimador del vector de parámetros $\theta = \theta_H$ dado por

$$\hat{\theta} = argmax_\theta \left\{\theta^T g(y) - \psi(\theta)\right\},$$

donde $\theta^T g(y) - \psi(\theta)$ es la razón de verosimilitud logarítmica, $\psi(\theta) = log\ \kappa(\theta)$. y también se tiene que $g(y) = \nabla\psi(\theta)|_{\theta=\hat{\theta}}$.

Si tomamos derivadas en ambos lados y usamos que $\mathbb{E}_{\hat{\theta}}\left[g(Y)\right] = \nabla\psi(\theta)$, obtenemos que el estimador de máxima verosimulitud no es más que la solución del sistema de ecuaciones dado por

$$\mathbb{E}_{\hat{\theta}}\left[g(Y)\right] = g(y).$$

El problema que surge con dicho estimador es que involucra a la función $\psi(\theta)$, la cual no puede evaluarse salvo para los casos más triviales, pues implica la aparición de $2^{\binom{N_v}{2}}$ posibles opciones para y, para cada θ. Luego, debido a ello, nos vemos obligados a tener que recurrir a la utilización de métodos numéricos para hallar valores aproximados para $\hat{\theta}$.

Surge la idea de utilizar un vector de parámetros de partida, θ_0, de manera que la razón logarítmica de verosimilitud puede escribirse como

$$r(\theta, \theta_0) = (\theta - \theta_0)^T g(y) - [\psi(\theta) - \psi(\theta_0)].$$

El MLE será el resultante de maximizar dicho estimador, por lo que se van a generar n muestras, $Y_1, \ldots, Y_n$ del EGRM usando θ_0. Se va a aproximar $exp(\psi(\theta) - \psi(\theta_0)]) = \mathbb{E}_{\theta_0}\left[exp\{(\theta - \theta_0)^T g(Y))\}\right]$ usando el promedio de las muestras. Por último, solo nos quedaría evaluar una aproximación de la razón logarítmica de verosimilitud $r(\theta, \theta_0)$.

Bondad de ajuste

Hay ocasiones en las que utilizamos un modelo perteneciente a una clase de modelos el cual es el mejor ajuste de toda esa clase, pero no ocurre lo mismo con los datos, es decir, el modelo no se ajusta bien a los datos. Lo cual puede ocurrir porque la clase de modelo que hemos seleccionado no cuenta con una variedad razonable de modelos entre los que elegir el que mejor se

adapta a nuestros datos.

Debido a lo que acabamos de describir, el concepto de bondad de ajuste tiene una gran importancia en el empleo de un modelo determinado. Centrándonos en los ERGM, se evalúa la bondad del ajuste simulando gran cantidad de grafos aleatorios del modelo que se ha ajustado. A continuación, se realiza una comparación entre las características observadas mediante estos grafos con las del grafo observado, (por ejemplo centralidad o distribuciones de los grados), de manera que si dichas características no coinciden entre ambos, se saca como conclusión que existen diferencias sistemáticas entre lo que arroja el modelo y los datos que teníamos, por lo que existe una falta de bondad de ajuste.

En el caso de los ERGM las características que se suelen emplear para la evaluación de la bondad del ajuste suelen ser las aportadas por los estadísticos suficientes.

Capítulo 2

Aproximación al análisis de redes con R

Lo primero que se va a hacer en este capítulo es justificar la utilización del entorno computacional de R.

Algunas de las facilidades que nos permite R es trabajar con multitud de paquete y funciones para poder trabajar con prácticamente cualquier conjunto de datos, sobre los que se puede ejecutar gran cantidad de órdenes. Dicho programa incluye paquetes específicos para el análisis de redes que ofrecer la ventaja de trabajar con un entorno de programación estadística que no se limitan únicamente al análisis de redes, por lo que ofrece muchas opciones más que otros programas. Aunque uno de os rasgos característicos de dicho programa es que es abierto y libre.

Una ventaja de este programa respecto a los demás es que alberga ciertos paquetes diseñados exclusivamente para el análisis de redes que se unen a las demás funciones proporcionadas por dicho programa. Ello permite que el programa traiga integradas ciertas bases de datos que permiten la realización de este tipo de trabajos. Dicho entorno proporciona herramientas que permiten centrar la atención en combinar los datos, así como en el código y los resultados.

Los paquetes que se encuentran en R exlusivamente para el análisis de datos permiten la manipulación de la misma junto con la visualización, descripción y modelaje de las mismas. Para ello, se cuenta con los siguientes paquetes:

- tidyverse: Para tratar la base de datos de la que se parte.
- igraph: Para guardar y transformar los datos pertenecientes a la red.
- ergm: Para realizar modelos sobre dichos datos.

También cabe destacar que R permite tratar con distintos tipos de redes como son las estocásticas, que son con las que se va a trabajar, modelos con actores dinámicos que se pueden estudiar a lo largo del tiempo, entre otros.

2.1. tidyverse

La ciencia de Datos surge ante la necesidad de interpretar y predecir resultados a partir de una base de datos, la cual se tiene que importar, filtrar en función de lo que se quiera estudiar, modelar, representaciones gráficas (si se puede), y extracción de conclusiones.
Para todo ello, gracias a Hadley Wickham principalmente, R cuenta con un paquete específico llamado `tidyverse`.

Cuando analizamos datos, nuestro fin es obtener información útil a partir de un conjunto de datos que cumplen una serie de características dadas, las variables. Dicha información se obtiene a partir de los datos existentes y se pueden crear nuevas variables, extraer alguno de los datos o variable que nos resulte de especial interés o cambiar las unidades de alguna de ellas. Todo ello podemos conseguirlo utilizando el entorno `tidyverse`, ya que es un conjunto de paquetes de R que ha sido creado para cargar, cambiar, observar y reproducir los el conjunto de datos con el que estamos trabajando. De entre todo el conjunto de paquetes que integran el *universo* tidyverse, nos centraremos en dos fundamentalmente: (Araneda, 2021):

- `readr`: Se utiliza a la hora de importar el conjunto de datos en el programa. En el caso que se presentará a continuación, se trata de un fichero .csv, es decir, separado por comas. Con la función `read_csv()` se realiza esto y especificamos el archivo al que estamos llamando, si tenemos en cuenta el nombre de las columnas, entre otros aspectos. El conjunto de datos cargado recibe el nombre de **data frame**.

- `dplyr`: Dicho paquete aporta un conjunto de funciones que nos resultan muy útiles a la hora de llevar a cabo acciones sobre los data frames. Cabe destacar las pipeline o tuberías, que nos permiten enlazar funcione una tras otra a un data frame a través del comando `%>%`. El uso de las tuberías permite no tener que estar llamando a las funciones, sino que directamente se va aplicando una tras otra al data frame.

- `tidyr`: Permite someter a transformaciones un conjunto de datos con el objetivo de que sea más eficiente.

2.2. igraph

Gracias a la librería `igraph` podemos representar de varias formas distintas una red, ya que cuenta con múltiples opciones para ello. A la hora de dicha representación, podemos tener en cuenta ciertas características sobre los parámetros que influyen en la determinación de una red. Las principales funciones a evaluar por dicho paquete son:

- Aportación de gráficos.

- Descripción de los mismos.

- Evaluación del entorno.

- Hallar subconjuntos.

- Análisis estadístico.

Entre esos parámetros citados anteriormente se encuentran los nodos, cuyas características vienen dadas por `vertex.` y las aristas, dadas por `edge.`, describamos las principales características asociadas a ambos conceptos que se pueden reflejar en el gráfico de la red. La explicación que se refleja a continuación es de los parámetros existentes aplicando la función `plot()` (Csardi y Nepusz, 2006).

- Los parámetros que se van a ver para los **nodos** son:
  - `vertex.size`: Se refiere al tamaño de los nodos. También puede venir dado por un vector si cada nodo es de un tamaño diferente.
  - `vertex.shape`: Es el comando utilizado para la forma del nodo (rectangular, circular).
  - `vertex.color`: Sirve para cambiar el color de los nodos.
  - `vertex.frame.color`: Es el color del borde del nodo.
  - `vertex.frame.width`: Es la anchura del borde del nodo.
  - `vertex.label`: Son las etiquetas de los nodos.
  - `vertex.label.font`: Para cambiar la fuente en las etiquetas de los nodos.
  - `vertex.label.cex`: Es el tamaño de la fuente.
  - `vertex.label.dist`: Es la distancia de la etiqueta desde el centro del nodo. Si es 0, la etiqueta está centrada en el no. Si es 1 entonces la etiqueta se muestra al lado del nodo.
  - `vertex.label.degree`: Define la posición de las etiquetas de los nodos, en relación con el centro de los mismos.
- Se presentan a continuación aquellos en relación a las aristas:
  - `edge.color`: Sirve para cambiar el color de las aristas.
  - `edge.width`: Es la anchura de la arista.
  - `edge.arrow.size`: Es el tamaño de la flecha. Por defecto, es el mismo para todas. Si se toma el atributo de una arista, sólo se utiliza el atributo de la primera arista para todas las flechas.
  - `edge.arrow.width`: Es la anchura de la flecha. Ocurre lo mismo que con el tamaño.
  - `edge.lty`: Es el tipo de línea para las aristas.
  - `edge.label`: Son las etiquetas de las aritas.
  - `edge.label.font`: Para cambiar la fuente en las etiquetas de los nodos.
  - `edge.label.cex`: Es el tamaño de la fuente.

  - `edge.label.color`: Es el color de la etiqueta.
  - `edge.label.x`: Es la manera de dar explícitamente las coordenadas horizontales de las etiquetas.
  - `edge.label.y`: Es la manera de dar explícitamente las coordenadas verticales de las etiquetas.
  - `edge.curved`: Especifica si se dibujan o no los bordes curvos. Puede ser un vector lógico o numérico o un escalar. Primero se replica el vector para que tenga la misma longitud que el número de aristas en el gráfico. Luego se interpreta para cada arista por separado.
  - `edge.arrow.mode`: Este parámetro puede utilizarse para especificar para qué aristas deben dibujarse flechas. Si se da, el usuario tiene que especificar qué aristas tendrán flechas hacia adelante, hacia atrás, ambas, o ninguna flecha.

- Otros parámetros que podemos tener en cuenta son:

  - `layout`: Es una función o una matriz numérica. Especifica cómo se colocarán los nodos en el gráfico.
  - `margin`: La cantidad de espacio vacío debajo, encima, a la izquierda y a la derecha del gráfico, se trata de vector numérico de longitud cuatro. Para los valores que están entre 0 y 0.5, se dice que son significativos, aunque pueden ser negativos, lo que hará que el gráfico se acerque a una parte del mismo.
  - `palette`: Es la paleta de colores que se usa para los nodos.
  - `rescale`: Para reescalar las coordenadas a [-1,1]x[-1,1].
  - `asp`: Es una constante que da la relación de aspecto en un gráfico.
  - `frame`: Es un Booleano que indica si se traza un marco alrededor del gráfico.
  - `main`: Es el título del gráfico.
  - `sub`: Es el subtítulo del gráfico.
  - `xlab`: Es el título para los valores del eje x.
  - `ylab`: Es el título para los valores del eje y.

2.3. ergm

Se tiene que reconocer al grupo de Ciencias de Redes de la Universidad de California en Davis por su aportación a la realización del paquete `ergm`. Que aunque ya se ha descrito anteriormente de manera teórica, se recuerda que sirve para predecir vínculos, tratar la dependencia entre nodos, la influencia de variables exógenas y endógenas y simular redes.. Se va a proceder a explicar las principales funciones de dicho paquete (Handcock y cols., 2022).

- `network( x, directed = TRUE, hyper = FALSE, loops = FALSE, multiple = FALSE, bipartite = FALSE, ignore.eval = TRUE, names.eval = NULL, edge.check = FALSE, density = NULL, init = NULL, numedges = NULL, ... )`: No es una función propiamente dicho del paquete `ergm`, pero al estar cargado, esta función actúa como `as.network.numeric()` en la que introducimos como argumento una matriz, (al usar `network()`), en lugar de un entero que se corresponda con el número de nodos existente. Como resultado se obtiene una red aleatoria de Bernoulli. Por ello, los parámetros que aparecen son los de `as.network.numeric()`, que se van a explicar a continuación:

  - `directed`: Si las aristas tienen que interpretarse como dirigidas.
  - `hiper`: Si puede haber hiper aritas.
  - `loops`: Si puede haber bucles.
  - `multiple`: Si puede haber aristas múltiples.
  - `bipartite`: Si se puede interpretar la red como bipartita.
  - `ignore.eval`: Si no hay que tener en cuenta el valor de las aristas.
  - `names.eval`: El nombre del atributo en el que deben almacenarse los valores de las aristas.
  - `edge.check`: Si se quiere comprobar la consistencia de las nuevas aristas.
  - `density`: Es la probabilidad de que haya una arista en una red de Bernoulli. Si no se da ni densidad ni init, el valor por defecto es el número de nodos dividido por el número de díadas.
  - `init`: Es la probabilidad logarítmica de que haya una arista en una red de Bernoulli.
  - `numedge`: El número de aristas al que se quiere condicionar la red, si es que se indica.
  - `...`: Argumentos adicionales.

  Como dicha red no contiene atributos para los nodos, ni para las aristas, ni para la red, éstos se pueden añadir usando *%v %, %n %, %e %*.

- `ergm(formula, response = NULL, reference = ~Bernoulli, constraints = ~., obs.constraints = ~. - observed, offset.coef = NULL, target.stats = NULL, eval.loglik = getOption("ergm.eval.loglik"), estimate = c("MLE", "MPLE", "CD"), control = control.ergm(), verbose = FALSE, ..., basis = ergm.getnetwork(formula))`: Dicha función se utiliza para ajustar modelos de grafos aleatorios de familia exponencial en la que la probabilidad de una red y está dada por la fórmula del capítulo anterior

$$\mathbf{P}_\theta(Y = y) = \left(\frac{h(y)}{\kappa}\right) exp\left\{\sum_H \theta_H g_H(y)\right\},$$

con la única diferencia de que en nuestro caso consideramos la medida de referencia, $h(y)$, igual a 1, pues es lo habitual. Dicha función toma como argumentos los que se han presentado anteriormente, aunque únicamente se van a describir los relativos a *formula*.

- `formula`: Es un objeto de la forma $y \sim$ "términos del modelo", donde y es una red o una matriz que puede ser transformada en una red. Existe gran diversidad de "términos del modelo" dependiendo del modelo especificado. Por ello, se presentan únicamente los que se van a emplear en la práctica.

    1. `nodecov`: Este término añade una única estadística de red para cada atributo cuantitativo o columna de la matriz igual a la suma de attr(i) y attr(j) para todas las aristas (i, j) del grafo.

    2. `nodematch(attr,diff)`: Cuando diff=FALSE , este término añade un estadístico de red al modelo, que cuenta el número de aristas (i, j) para las que attr(i)==attr(j). Lo cual se llama "homofilia uniforme", pues se asume que cada grupo tiene la misma propensión a que haya aristas entre los nodos pertenecientes al mismo grupo. Cuando nos encontramos con atributos que tienen varios nombres, el estadístico sólo cuenta las aristas en los que coinciden todos los atributos. Cuando diff=TRUE, se añaden p estadísticos de red al modelo, donde p es el número de valores únicos del atributo attr. El k_{th} estadístico cuenta el número de aristas (i, j) para las que attr(i) == attr(j)== valor(k) , donde valor(k) es el k_{th} valor único más pequeño del atributo attr. Esto se denomina también "homofilia diferencial", pues se considera que cada grupo puede tener una propensión única a crear aristas dentro del grupo. Nótese que en una prueba estadística de homofilia uniforme frente a la diferencial debe completase con la función ANOVA. Por defecto, se cuentan las coincidencias entre todos los niveles. Esto funciona tanto para diff=TRUE como para diff=FALSE .

    3. `triangle`: Por defecto, este término añade un estadístico al modelo igual al número de triángulos de la red. Para una red no dirigida, un triángulo se define como cualquier conjunto (i, j),(j, k),(k, i) de tres aristas.

    4. `gwesp(decay, fixed=FALSE, cutoff=30)`: Este término añade un estadístico igual a la distribución de pareja compartida geométricamente ponderadas por las aristas con el parámetro de decaimiento, `decay`, que debe ser no negativo y se utiliza para contar los nodos compartidos o dirigidos por dos caminos distintos. Por lo que es obligatorio si `fixed=TRUE`.

    5. `edges`: Este término añade un estadístico de red igual al número de aristas (es decir, valores distintos de cero) en la red. Para las redes no dirigidas, `edges` es igual a `kstar(1)`.

- `summary.formula`: Se presentan por pantalla los estadísticos del grafo, el término de la izquierda, que estamos explicando en función de los términos del modelo, que se encuentran a la derecha.

- `summary.ergm`: Proporciona una tabla resumen una vez evaluada la función `ergm` formada por las siguientes columnas:

  - Estimaciones de los coeficientes y sus errores estándar.
  - El porcentaje de error estándar atribuible, el cual se indica si `total.variation = TRUE`. Para que ello ocurra, el error estándar tiene que estar formado a partir de la suma de la varianza de las verosimilitudes y la varianza aportada por el MCMC (método de Monte Carlo usando cadenas de Markov).
  - Los z-valores para ver la significatividad de los elementos en el modelos.

  Además, debajo de esta tabla también se encuentran las varianzas nula y residual y los coeficientes AIC, criterio de información de Akaike, y BIC, criterio de información bayesiano, los cuales indican la calidad del modelo en función del ajuste y los términos que intervienen.

- `anova.ergm`: Muestra por pantalla un análisis de varianza mediante una tabla en la que se encuentra la suma de los cuadrados de la variabilidad residual y la explicada por el modelo. Además, proporciona un p-valor que compara ambas variabilidades.

- `gof`: Proporciona los p-valores de los resúmenes de distancia geodésica, grado y alcanzabilidad para obtener la bondad de ajuste del ERGM que se está analizando. Para ello, se extrae de manera aleatoria una muestra de gráficos de dicho modelo. Para obtener un buen ajuste, los estadísticos observados tienen que estar cerca de la mediana de la muestra (a distancia 5 como máximo).

- `plot.gof`: Se trata de la representación gráfica de lo descrito anteriormente.

- `simulate.ergm`: Se emplea para extraer de manera aleatoria un número determinado de redes basadas en el modelo que se ha introducido en función de las restricciones, los atributos y los parámetros del ajuste.

Capítulo 3

Aplicación práctica

En este capítulo se va a llevar a cabo un análisis de redes mediante el modelo del grafo aleatorio exponencial y sus variaciones. En concreto, la base de datos sobre la que se va a trabajar ha sido obtenida del Sistema Automatizado de Información Criminal (ACIIS) de la policía de Montreal, reconstruídos a partir de (Descormiers y Morselli, 2011) que se encuentran en repositorio de datos UCINET (Borgatti, Everett, y Freeman, 2002).

La base de datos que se va a analizar trata de reconstruir la organización de la distribución de drogas en Montreal Norte. Para ello, se ha tenido en cuenta la información obtenida en tres investigaciones distintas en las que se creía que la banda que controlaba dicha distribución era los Bo-Gars.

En el transcurso de las operaciones se vigiló a 45 personas en la primera, 30 en la segunda y 26 en la tercera, un total de 101 personas de las que se parte para reconstruir la red final, que finalmente estaría compuesta por 70 participantes y se forma a partir de tres fuentes de información.

Partimos de una matriz 35×35 en la que se definen las relaciones existentes entre las bandas de manera que si en el lugar (i,j) hay un 1, existe relación y si hay un 0, no existe. Es decir, se nos presentan los datos en forma de matriz de adyacencia en la que las bandas son los nodos y las aristas muestran aquellas bandas con las que se relaciona cada una de ellas. Como partimos de datos que se encuentran en un fichero `.cvs`, lo primero que tenemos que hacer es cargarlos y transformarlos. No se puede olvidar cargar los paquetes que se han explicado en el capítulo 2. El fichero a priori parece estar bien pero se encuentran ciertas carencias, pues hay relaciones existentes entre bandas que aparecen en el lugar (i,j) pero no en el lugar (j,i), por lo que no se tendría una matriz de adyacencia y no se podría hacer nada de lo que se muestra a continuación, de ahí las transformaciones que se realizan en el código. Además, existen bandas cuyo nombre es un número y el programa añade una X delante de dichos nombres para detectarlos como tipo carácter, cosa que también se modifica.

A continuación, se procede a cargar las características cuyo efecto se va a estudiar sobre la

red. Dichas características son atributos de nodos, entre los que se va a diferenciar:

- Afiliación: Por regla general, las bandas callejeras norteamericanas se han dividido en dos grandes coaliciones, los Bloods y los Crips. Por lo que este atributa va a ser clasificatorio, siendo 1) Bloods, 2) Crips y 3) Otros.
- Etnia: La composición étnica de las bandas. Se diferencian: 1) Hispano, 2) Afrocanadienses, 3) Caucásicos, 4) Asiáticos y 5) Sin asociación principal o mixto.
- Territorio: Zona en la que opera la banda 1) Centro, 2) Este y 3) Oeste.

Sin embargo, existen cierto valores de estos atributos que no se han conseguido recuperar. En el fichero de datos, esos valores toman el valor 99, por lo que para no entrar en especulaciones e incongruencias, se eliminan tanto las relaciones que atañan a las bandas de dichos entrevistados, como los atributos que tienen asociados.

```
library(tidyverse)
library(igraph)
library(ergm)

relacionesBandas <- MONTREALGANG
colnames(relacionesBandas) <- c("Bo-Gars","BMF","Green_Land",
"Lg Side","Plan riel", "18th", "187","AYB.2","AYB","Uptown_Posse",
"Dangerous_Street","DPC","J.O.K.E.R.S","Pie-IX","Px-80",
"50 Niggaz", "13th", "67", "99", "146","Black_Dragons","Outlaws",
"South_Side","White_Tigers","Motard","Bronx","LPE","PSC","St.Henri",
"V_Block", "47","Blue_Devil.1","Blue_Devil.2","RTC","Ve_Crew")
relacionesBandas[6,23] = 1
relacionesBandas[19,17] = 1
relacionesBandas[34,19] = 1
relacionesBandas <- relacionesBandas[-c(8,26,27,28,29,30,31,32,33),
                    -c(8,26,27,28,29,30,31,32,33)]

attr <- MONTREALGANG_ATTR
attr <- attr[-c(5,10,13,14,16,21,25,31,33),]

A <- data.matrix(relacionesBandas)
```

Para comprobar que efectivamente dicha matriz es de adyacencia, como ya se sabe que está formada por 0 y 1, se comprueba si es simétrica.

```
isSymmetric(A)
[1] TRUE
```

Una vez que se comprueba que la matriz de adyacencia es correcta, se construye la red especificando que no es una red directa y se añaden los atributos, obteniendo una red con 26 nodos y 57 aristas.

```
bandas <- network(A, directed = FALSE)

bandas %v% "afiliacion" <- attr[,1]
bandas %v% "etnia" <- attr[,2]
bandas %v% "territorio" <- attr[,3]

bandas
Network attributes:
  vertices = 26
  directed = FALSE
  hyper = FALSE
  loops = FALSE
  multiple = FALSE
  bipartite = FALSE
  total edges= 57
    missing edges= 0
    non-missing edges= 57

 Vertex attribute names:
    afiliacion etnia territorio vertex.names

No edge attributes
```

Una vez que se puede empezar a trabajar, el lector puede plantearse el sentido que tiene estudiar el problema que se planteaba al comienzo del capítulo, ver si realmente los Go-Bars eran los principales distribuidores de drogas en Montreal del Norte. Como se han visto mermados los datos de los que se partía a priori y con los ERGM se puede obtener la probabilidad a establecer una relación en función de los atributos. Se van a estudiar los siguientes casos:

- Las bandas formadas por personas de la misma etnia tienen la misma tendencia a establecer relaciones con bandas de su etnia que con etnias distintas.
- Las bandas que se encuentran en el mismo territorio tienen la misma tendencia a establecer relaciones con bandas de su territorio que con bandas que se encuentran más alejadas.
- Las bandas que están afiliadas a la misma coalición tienen la misma tendencia a establecer relaciones con bandas de su coalición que con bandas que pertenecen a otra distinta.
- Los atributos que presentan las bandas no influyen en la interconexión de las mismas.

Ahora que está claro los problemas que se van a abordar, un primer paso puede ser hacerse una idea intuitiva de la red mediante su representación gráfica.

```
plot(bandas, vertex.cex=(bandas %v% "afiliacion"+
bandas %v% "etnia"+bandas %v% "territorio")/3,
label = network.vertex.names(bandas))
```

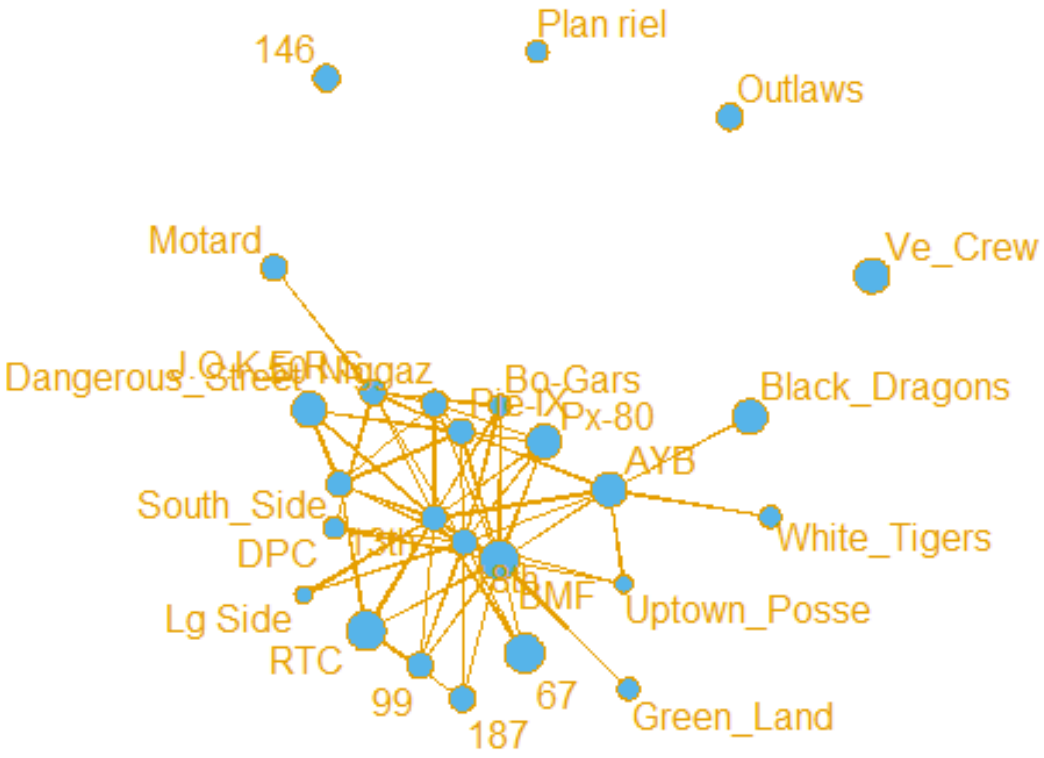

Figura 3.1: Gráfico de la red

En el gráfico que se muestra, se puede apreciar que hay cuatro bandas que no establecen relaciones con las demás. En cambio, las demás sí.

El primer modelo que se va a realizar es un modelo simple, que únicamente contiene un término que representa el número total de aristas en la red. El nombre de este término ergm es `edges`, y que cuando se introduce en un ERGM, dicho coeficiente sirve para controlar la densidad global de la red. A continuación, se pide el resumen del modelo con la función `summary` que devuelve los valores numéricos de los estadísticos de la red en el modelo.

```
summary(bandas ~ edges)
edges
   78

bandasmodel.01 <- ergm(bandas ~ edges)
```

```
Starting maximum pseudolikelihood estimation (MPLE):
Evaluating the predictor and response matrix.
Maximizing the pseudolikelihood.
Finished MPLE.
Stopping at the initial estimate.
Evaluating log-likelihood at the estimate.

summary(bandasmodel.01)

Call:
ergm(formula = bandas ~ edges)

Maximum Likelihood Results:

      Estimate Std. Error MCMC % z value Pr(>|z|)
edges  -1.5479     0.1459      0  -10.61   <1e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Null Deviance: 450.5  on 325  degrees of freedom
 Residual Deviance: 301.8  on 324  degrees of freedom

AIC: 303.8  BIC: 307.6  (Smaller is better. MC Std. Err. = 0)
```

Con este modelo lo que se obtiene es la probabilidad homogénea de que haya una arista entre dos nodos. Para su interpretación, lo usual es partir de la probabilidad logarítmica, quedando de la siguiente forma:

$$logit(p(y)) = -1.5479 \times 1,$$

para cada arista, ya que añadir una arista suponer incrementar en una unidad el total de aristas de la red. La probabilidad total se obtiene tomando la inversa, es decir, haciendo

$$= \frac{\exp -1.5479}{1 + \exp -1.5479} = 0.1754$$

luego, la probabilidad de añadir una nueva arista al modelo es 0.1754. Esta probabilidad se corresponde con la densidad de la red.

Ahora se tiene en cuenta el efecto de los atributos teniendo en cuenta los problemas que se han planteado. Como tenemos atributos cualitativos entre los que se quiere ver si existen diferencias entre las distintas categorías, se tendrá que usar `nodefactor` para cada uno de ellos. La influencia de compartir un atributo no es más que ver los efectos homofílicos de los mismos sobre el modelo, por lo que se empleará `nodematch`.

```
bandamodel.02 <- ergm(bandas~edges + nodefactor('etnia')
+ nodefactor('territorio') + nodefactor('afiliacion') +
nodematch('etnia') + nodematch('territorio') +
nodematch('afiliacion'))

summary(bandamodel.02)

Call:
ergm(formula = bandas ~ edges + nodefactor("etnia") +
nodefactor("territorio") + nodefactor("afiliacion") +
nodematch("etnia") + nodematch("territorio") +
    nodematch("afiliacion"))

Maximum Likelihood Results:

                       Estimate Std. Error MCMC % z value Pr(>|z|)
edges                  -5.56900    1.55820      0  -3.574 0.000352 ***
nodefactor.etnia.2     -0.11241    0.47333      0  -0.237 0.812288
nodefactor.etnia.3      2.68118    0.61729      0   4.343  < 1e-04 ***
nodefactor.etnia.4      1.64172    0.65785      0   2.496 0.012575 *
nodefactor.etnia.5      1.25919    0.48323      0   2.606 0.009166 **
nodefactor.territorio.2  1.08932    0.85726      0   1.271 0.203837
nodefactor.territorio.3  0.37463    0.56422      0   0.664 0.506706
nodefactor.afiliacion.2  1.42409    0.34715      0   4.102  < 1e-04 ***
nodefactor.afiliacion.3 -0.41225    0.57632      0  -0.715 0.474411
nodematch.etnia          0.30629    0.46980      0   0.652 0.514425
nodematch.territorio    -0.37867    0.73692      0  -0.514 0.607357
nodematch.afiliacion    -0.08704    0.38781      0  -0.224 0.822411
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Null Deviance: 450.5  on 325  degrees of freedom
 Residual Deviance: 248.8  on 313  degrees of freedom

AIC: 272.8  BIC: 318.2  (Smaller is better. MC Std. Err. = 0)
```

Nótese que los cálculos que se reflejan se basan en la comparación con la clase 1) de cada atributo. Por lo que se puede afirmar que:

- No se observan diferencias significativas entre las bandas según la etnia para establercer relaciones.
- El territorio en el que actúa cada banda no resulta significativo para establecer relaciones.

- Tampoco se obervan diferencias significativas a la hora de establecer relaciones según la afiliación.
- En cuanto a la interconexión de las bandas sí que se observan diferencias. Lo que se puede observar en la comparación entre Hispanos y Caucásicos, donde las diferencias son altamente significativas, pues p-valor < 0.001, y al comparar los Bloods con los Crips, que ocurre lo mismo.

Seguidamente se evalúa la convergencia del modelo y se interpretan los resultados.

```
gof.bandas.ergm <- gof(bandamodel.02)

plot(gof.bandas.ergm)
```

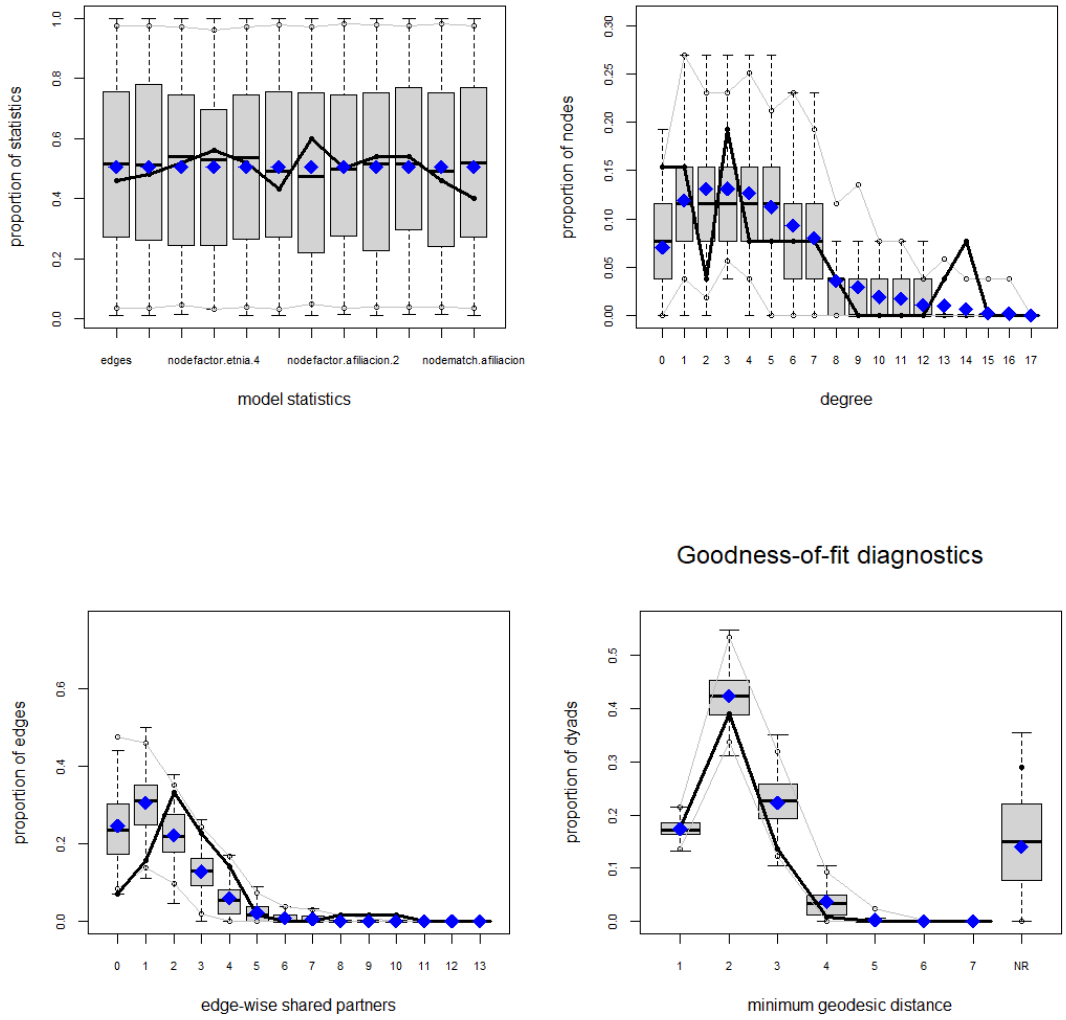

Figura 3.3: Gráficos bondad del ajuste

Se procede a analizar los gráficos obtenidos para evaluar la bondad del ajuste del modelo. En cuanto a los estadísticos del modelo, no hay errores de estimación, pues los observados

se encuentran dentro del rango de los esperados. Para los valores del grado de centralidad de cada nodo se observa que hay grados subestimados y sobreestimados. Mientras que la distancia geodésica no es del todo acertada para valores ”pequeños”. Por lo tanto, el modelo tiene margen de mejora pero es bastante útil.

3.1. Conclusiones y futuras líneas

Con los datos analizados se puede concluir que la homofilia para establecer relaciones entre las bandas no ha resultado significativa para ninguno de los tres atributos sobre los que se ha trabajado. Por otro lado, sí que se han encontrado diferencias significativas para el establecimiento de relaciones de las bandas independientemente de los atributos que éstas poseen.

Para mejorar el modelo cabe la posibilidad de introducir ciertos atributos nuevos como el tipo de droga que distribuye cada banda o la rivalidad existente entre ellas. Además, se podría tener un modelo que sea más representativo si se consiguiera hallar los valores de los atributos de las bandas que se han tenido que eliminar.

Finalmente, el análisis realizado se podría llevar a cabo con el modelo modelo de bloques de red y con el modelo de bloques latentes para ver si mejorarían los resultados obtenidos o no serían útiles para lo que se ha realizado en el trabajo descrito en este documento.

Referencias

Araneda, P. (2021). *Tidyverse para data análisis.* Descargado de https://rpubs.com/paraneda/tidyverse

Barber, F. (2018). Apuntes tema 15. *Universidad de Valencia*.

Borgatti, S. P., Everett, M. G., y Freeman, L. C. (2002). Ucinet for windows: Software for social network analysis. *Harvard, MA: analytic technologies*.

Cardenas, J. (2020). *Curso analisis de redes.* Descargado de http://networksprovidehappiness.com/curso-online-analisis-de-redes/

Cordón, O. (2013). *Tema 5: modelos de redes.* Descargado de https://sci2s.ugr.es/sites/default/files/files/Teaching/GraduatesCourses/RedesSistemasCompejos/Tema05-1-ModelosdeRedesAleatorias-13-14.pdf

Csardi, G., y Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, *Complex Systems*, 1695. Descargado de https://igraph.org

Descormiers, K., y Morselli, C. (2011). Alliances, conflicts, and contradictions in montreal's street gang landscape. *International Criminal Justice Review*, *21*(3), 297–314.

Dunn, W. N. (1983). Social network theory. *Knowledge*, *4*(3).

Garrido, F. J. (2001). El análisis de redes en el desarrollo local. *Facultad de Ciencias Política y Sociología. Universidad Complutense de Madrid*.

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., Krivitsky, P. N., y Morris, M. (2022). ergm: Fit, simulate and diagnose exponential-family models for networks [Manual de software informático]. Descargado de https://CRAN.R-project.org/package=ergm (R package version 4.2.2)

Huisman, M., y Snijders, T. A. (2003). Statistical analysis of longitudinal network data with changing composition. *Sociological methods & research*, *32*(2), 253–287.

Kolaczyk, E. D., y Csárdi, G. (2014). *Statistical analysis of network data with r* (Vol. 65). Springer.

Palacios, D., y Villalobos, C. (2016, 06). Redes académicas al interior de las escuelas chilenas: Un estudio exploratorio utilizando modelos exponenciales de grafos aleatorios (ergm). *Redes. Revista hispana para el análisis de redes sociales*, *27*, 33. doi: 10.5565/rev/redes.631

RAE. (2020). *Diccionario de la lengua española.* 23.ª ed., [versión 23.4 en línea]. Descargado de https://www.rae.es/

Rodríguez-Bocca, P. (2019). *Modelos de grafos.* Descargado de https://eva.fing.edu.uy/pluginfile.php/132941/course/section/14369/block_4_networks_models_part_b_models.pdf

Sancho, F. (2016). *Introducción a las redes complejas.* Descargado de http://www.cs.us.es/~fsancho/?e=80

Snijders, T. A., Pattison, P. E., Robins, G. L., y Handcock, M. S. (2006). New specifications for exponential random graph models. *Sociological methodology*, *36*(1), 99–153.

Van der Pol, J. (2019). Introduction to network modeling using exponential random graph models (ergm): theory and an application using r-project. *Computational Economics*, *54*(3), 845–875.