Inferencia Estadística II
Reporte Final:
Pruebas de Hipótesis en Gráficas Aleatorias.

Marcos Torres Vivanco

1. Introducción.

Desde su origen, las gráficas se han utilizado para modelar una cantidad muy diversa de problemas, desde mapas, procesos biológicos, interacciones de organismos en un ecosistema, relaciones en neuronas, entre otros. Al agregarle un factor aleatorio a las gráficas es posible realizar los análisis estudiados en estadística, tomando como datos gráficas observadas.

En el contexto de gráficas aleatorias es posible realizar pruebas de hipótesis sobre una gran cantidad de problemas, pruebas de bondad de ajuste sobre la distribución de los grados de los vértices, verificar ciertas propiedades de centralidad, o verificar la presencia de subgráficas con alguna propiedad.

Una comunidad es una subgráfica de nuestra gráfica aleatoria que tiene mayor probabilidad de conectarse. Por ejemplo, en una red social una comunidad puede representar un grupo de amigos, o una comunidad de fans. Encontrar la comunidad de manera explicita se puede plantear como un problema de matrices aleatorias, pero el problema de detectar la presencia de dicha comunidad corresponde a un problema de pruebas de hipótesis y es este es el problema que guiará este trabajo.

En la sección 2 se dará una introducción a la teoría de gráficas y de gráficas aleatorias. Se hablará de su origen y algunas de las definiciones más importantes con las que trabajaremos.

En la sección 3 se presentará el problema de detección de comunidades. Este problema se puede plantear como una prueba de hipótesis. Comenzaremos con el caso particular de la detección de un clique, y avanzaremos hacía el caso más general de la detección de una comunidad usando tres estadísticas diferentes, total de aristas,

escaneo y total de triángulos. Se hablará de las propiedades, ventajes y desventajas de cada una de estas estadísticas.

2. Preliminares.

A lo largo de esta sección se presentarán las definiciones y nociones básicas relacionadas con la teoría de gráficas y posteriormente con la denominada teoría de gráficas aleatorias. La comprensión de dichos conceptos será de vital importancia para el trabajo que realizaremos posteriormente.

2.1. Teoría de Gráficas.

La teoría de gráficas, también llamada teoría de grafos, surgió con un problema planteado al matemático Leonard Euler en el año de 1736. La ciudad de Konigsberg, Prusia Oriental, es atravesada por el río Pregel, el cual tiene una bifurcación que divide el terreno de la ciudad en cuatro regiones distintas, las cuales estaban conectadas por siete puentes. El problema consistía en encontrar un recorrido para cruzar cada uno de los puentes una única vez y regresar al mismo punto de partida.

Para dar la solución a esta cuestión, Euler recurrió a una abstracción del mapa de la ciudad. Para el problema que se le plateaba, no importa la magnitud de las regiones, ni la longitud de los puentes, solo importaba que existía un conjunto de regiones y una relación entre ellos. Con esta abstracción se pudo dar una demostración en términos matemáticos del porque no existe el recorrido solicitado.

Figura 1: Abstracción de los puentes en una gráfica

Con esta idea se puede dar una definición formal del principal objeto matemático con el qe estaremos trabajando. Una **gráfica** G es un par ordenado $G = (V, A)$, donde a V se le denomina el conjunto de vértices y A el conjunto de aristas, el cual está conformado por pares ordenados de elementos de V. En algunos contextos es posible considerar que las aristas aparecen más de una vez en la gráfica (aristas múltiples) o que existen aristas que conectan un vértice con él mismo (loops). Pero en este caso

no consideraremos ninguno de estos casos, es decir trabajaremos con **gráficas simples**.

En este trabajo también agregaremos una condición extra a las gráficas, no consideramos gráficas dirigidas, es decir, únicamente trabajaremos con gráficas G tales que si la arista (u, v) aparece en el conjunto de aristas, entonces (v, u) también.

Trabajar directamente con gráficas puede resultar práctico e intuitivo, pero si se desea realizar trabajos computacionales o para aplicar teoría de otras ramas de las matemáticas, especialmente álgebra lineal, es importante considerar que a toda gráfica se le puede asociar una matriz. A esta matriz se le denomina **matriz de adyaciencia** y si $G = (V, A)$ es una gráfica las entradas de su matriz de adyaciencia están dadas por:

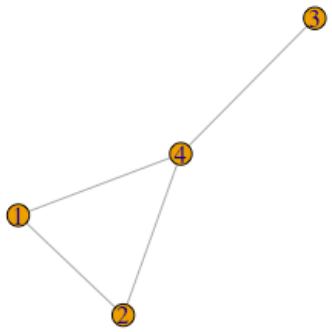
$$(\mathcal{M}_G)_{i,j} = \begin{cases} 1 & \text{si } (i,j) \in A \\ 0 & \text{si } (i,j) \notin A. \end{cases}$$

Notemos que toda matriz de adyaciencia es simétrica, pues G no es dirigida, y tiene ceros en la diagonal, pues no hay loops.

También podemos notar que si M es matriz simétrica con diagonal nula y que solo toma valores 1 y 0, entonces podemos construir una única gráfica G tal que su matriz de adyaciencia es M.

Por ejemplo se muestra la siguiente matriz y su correspondiente gráfica asociada.

$$M_G = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$



Una **subgráfica** $G' = (V', A')$de la gráfica $G = (V, A)$, es una gráfica tal que $V' \subset V$ y $A' \subset A$. Si tomamos un subconjuto de vértices V' de una gráfica $G = (V, A)$, entonces

la subgráfica generada por V' es la mayor subgráfica de G que contiene a los vértices V'.

Una **gráfica completa** es una gráfica $G = (V, A)$ que tiene todas las aristas posibles que pueden unir a sus vértices. Es decir, tal que

$$A = \{(u, v) \ : \ u, v \in V\}.$$

Notemos que una gráfica completa $G = (V, A)$ tiene $\binom{N}{2}$ aristas diferentes.

Un **clique** de una gráfica $G = (V, A)$ es una subgráfica completa de G, es decir es una subgráfica que tiene todas sus aristas posibles.

2.2. Gráficas Aleatorias.

El concepto de **gráfica aleatoria** se refiere a considerar distribuciones de probabilidad sobre conjuntos de gráficas o a un proceso aleatorio para construir una gráfica. La teoría de gráficas aleatorias surgió en 1959, con los estudios de Renyi y Ërdos quienes definieron y usaron dichas estructuras para encontrar propiedades en gráficas. De manera independiente Gilbert también las definió el mismo año, proponiendo un modelo diferente al de Renyi y Ërdos.

- El modelo de Erdos-Renyi, denotado por $G(N, M)$, tiene como parámetros el número de vértices N y el número de aristas M. Por lo tanto la distribución de $G(N, M)$ es uniforme sobre el conjunto de todas las gráficas de N vértices y M aristas.

- El modelo de Gilbert denotado $G(N, p)$, está parametrizado por el número de aristas N y la probabilidad p de que cualesquiera dos de ellas se encuentren unidas por una arista. Por lo tanto la distribución de $G(N, p)$ se obtiene de fijar N vértices y colocar las aristas entre los vértices de manera independiente con probabilidad p. Entonces la probabilidad asociada a una gráfica dada G con N vértices y M aristas es $p^M(1-p)^{\binom{N}{2}-M}$.

A lo largo de este trabajo usaremos el modelo de Gilbert, por resultar más adecuado para los modelos con los que trabajaremos.

Notemos que en una gráfica aleatoria $\mathcal{G} \sim G(N, p)$, la probabilidad de conexión entre dos vértices es una variable aleatoria Bernoulli con parámetro p. Dichas variables corresponden a las aristas de nuestra gráfica, por lo que podemos representar las probabilidades de conexión de una gráfica aleatoria con una matriz de adyaciencia:

$$(\mathcal{M}_{\mathcal{G}})_{i,j} = e_{i,j},$$

donde $e_{i,j}$ es la probabilidad de conectar los vértices i y j. Al igual que con las gráficas no aleatorias, toda matriz de adyaciencia en el contexto aleatorio es simétrica. Notemos que los elementos de la diagonal son todos ceros, pues no hay loops, es decir $e_{i,i} = 0$.

$$\mathcal{G} = \begin{pmatrix} 0 & e_{1,2} & \cdots & e_{1,N-1} & e_{1,N} \\ e_{2,1} & 0 & \cdots & e_{2,N-1} & e_{2,N} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ e_{N-1,1} & e_{N-1,2} & \cdots & 0 & e_{N,N-1} \\ e_{N,1} & e_{N,2} & \cdots & e_{N,N-1} & 0 \end{pmatrix}$$

2.3. Inferencia en gráficas aleatorias.

Supongamos que tenemos una observación de nuestra gráfica aleatoria $\mathcal{G} \sim G(N, p)$, en donde N es conocido, y que deseamos realizar inferencia sobre el parámetro p. Entonces tenemos que la función de veorsimilitud de nuestras $\binom{N}{2}$ aristas observadas es:

$$\begin{aligned} L(p; G) &= \prod_{i,j \in \{1,\ldots,N\}} p^{e_{i,j}} (1-p)^{1-e_{i,j}} \\ &= p^W (1-p)^{\binom{N}{2} - W}. \end{aligned}$$

Donde $W = \sum_{i,j \in \{1,\ldots,N\}} e_{i,j}$ es la cantidad total de aristas.

Por lo tanto la función de log-verosimilitud es:

$$l(p; G) = W \log(p) + \left(\binom{N}{2} - W \right) \log(1-p).$$

Derivando y maximizando, obtenemos que el estimador de máxima verosimilitud es:

$$\hat{p} = \frac{W}{\binom{N}{2}}.$$

Es usual considerar una única observación de la gráfica. Pero en muchas aplicaciones es útil considerar más de una observación de la misma gráfica, pues se considera que la gráfica evoluciona conforme pasa el tiempo. Por ejemplo consideremos una red social donde los vértices son personas y las aristas son una relación de amistad entre dos personas. En este caso conforme pasa el tiempo las amistades pueden terminar y dos personas que antes eran desconocidas son ahora amigos.

Para este trabajo, solo consideraremos el caso en que se tiene una única observación de la gráfica.

3. El problema de detectar una comunidad.

Dada una gráfica aleatoria $\mathcal{G}$ una comunidad es una subgráfica aleatoria de $\mathcal{G}$ con mayor probabilidad de conectarse. Un ejemplo de comunidad en un contexto de redes sociales es un grupo de amigos dentro de un conjunto de personas de la red. El problema principal de este trabajo consiste en detectar una comunidad en un modelo de una gráfica aleatoria.

El problema de detección de comunidades ha resultado de gran importancia en diversos contextos, encontrar grupos de amigos en redes sociales o incluso en biología para encontrar grupos de genes en gráficas de co-ocurrencia de genes [1].

El problema de encontrar explicitamente la comunidad resulta de gran importancia para el área de ciencias de la computación. Pero en nuestro caso estudiaremos el problema de únicamente detectar la existencia o no de dicha comunidad. Este problema surge para reducir la complejidad computacional del problema particular de encontar la comunidad o por el contexto del modelo en el que estamos trabajando. La detección de una comunidad dentro de un modelo de una gráfica aleatoria se puede plantear como un problema de prueba de hipótesis, por lo que resulta de interés estadístico.

Bajo la hipótesis nula tenemos que la distribución de la gráfica sigue un modelo usual de Gilbert con N vértices todo ellos con probabilidad p_0 de unirse entre ellos. En la hipótesis alternativa tenemos un subconjunto de n vértices con probabilidad p_1 de unirse entre ellos, tal que $p_0 < p_1$. Esta última condición es la que nos indica que el subconjunto de n vértices generará una comunidad, en donde sus miembros tienen mayor probabilidad de estar conectados que con el resto de los vértices de la gráfica.

Por lo tanto, las hipótesis con las que trabajaremos son:

$$H_0 : \mathcal{G} \sim G(N, p_0) \text{ v.s. } H_1 : \mathcal{G} \sim G(N, p_0; n, p_1). \tag{1}$$

A lo largo de este trabajo se supondrán conocidos los valores N y n. El valor de p_1 en general será desconocido, pero el hecho de que sea mayor que el valor de p_0 será útil para la definición de algunas pruebas como veremos más adelante. En cuanto al valor de p_0, el caso en que es conocido resulta ilustrativo e interesante matemáticamente pero es el caso cuando p_0 es desconocido el que resulta de interés en la práctica.

3.1. Pruebas de hipótesis.

Una **prueba** para nuestras hipótesis (1) es una función $\phi : \mathcal{G} \to \{0, 1\}$, tal que si G es una observación del modelo $\mathcal{G}$, entonces $\phi(G) = 1$ si detecta la presencia de una comunidad y $\phi(G) = 0$ si no detecta una comunidad. Generalmente una prueba

se obtiene con ayuda de una estadística, es decir, una función de los datos que nos dé información sobre la presencia de una comunidad.

A lo largo de este trabajo presentaremos diversas pruebas, por lo que resulta útil tener una noción de comparación o de optimalidad. Una aproximación para esto es con la función de **riesgo** o de peor caso de la prueba ϕ, la cual definimos como:

$$\gamma_N = P_{H_0}(\phi = 1) + P_{H_1}(\phi = 0).$$

Algunas propiedades asitóticas de las pruebas pueden ser observadas cuando la cantidad de vértices tiende a infinito, $N \to \infty$. Por lo tanto, consideramos una sucesión de hipótesis

$$(H_{0,N} \text{ v.s. } H_{1,N}),$$

donde N denota el número de vértices en las hipótesis (1).

Agregaremos una condición asintótica a nuestras hipótesis, referente al tamaño de la comunidad. Pediremos que el número n, no sea tan pequeño como para las pruebas lo ignoren conforme $N \to \infty$. Esto lo podemos obtener con la condición:

$$n/\log(N) \to \infty.$$

Decimos que una sucesión de pruebas (ϕ_N) para una sucesión de hipótesis $(H_{0,N}$ v.s. $H_{1,N})$ es **asintóticamente potente** si $\gamma_N(\phi_N) \to 0$. De manera intuitiva, una prueba es asintóticamente potente si es mejor que cualquier otra prueba que ignore los datos, es decir, que solo adivine.

3.2. Detección de un clique.

En nuestro primer ejemplo de una prueba para detectar una comunidad, comenzaremos con el caso particular en que la probabilidad de conexión dentro de la comunidad es $p_1 = 1$. Con esta condición, tenemos que bajo la hipótesis nula en nuestra gráfica G se forma una clique de tamaño n.

$$H_0 : \mathcal{G} \sim G(N, p_0) \text{ v.s. } H_1 : \mathcal{G} \sim G(N, p_0; n, 1)$$

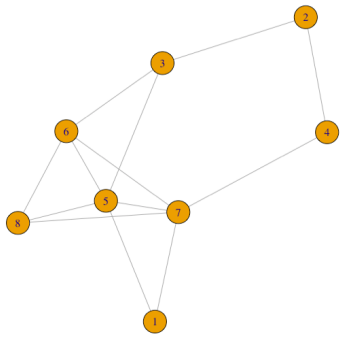


Figura 2: La subgráfica formada por los vértices $\{5, 6, 7, 8\}$ es un clique.

La estadística que usaremos para definir la prueba es el **número de clique**, es decir, el número de vértices del clique más grande contenido en nuestra gráfica G. Para nuestra hipótesis nula, $\mathcal{G} \sim G(N, p_0)$, tenemos que cualesquiera dos vértices tienen probabilidad $p_0 > 0$ de unirse, por lo que la probabilidad de que existan cliques de cualquier tamaño es diferente de cero. Sin embargo, la probabilidad de que aparezcan cliques muy grandes es muy baja. Por ejemplo, la probabilidad de que aparezca un clique de tamaño N, o lo que es lo mismo, que la gráfica sea completa, es de $p_o^{\binom{N}{2}}$.

Denotemos por C a nuestra estadística del número de clique. Dada una observación G, rechazamos para valores muy grandes de $C(G)$, pues esto presenta evidencia a favor de la existencia de un clique anormal en nuestra gráfica.

Esta estadística tiene una doble utilidad. No sólo resuelve el problema de detección de una comunidad, también encuentra la comunidad de manera explicita. Esto puede resultar contraproducente en el aspecto computacional, pero resultará ilustrativa como un primer acercamiento a las pruebas de hipótesis en el contexto de gráficas aleatorias.

Esta prueba resulta ser asintóticamente potente. Esto se sigue de un resultado clásico de gráficas aleatorias:

Teorema 1. *Si se tiene la condición*

$$\binom{N}{n} p_0^{\frac{n(n-1)}{2}} \to 0.$$

Entonces bajo la hipótesis nula el número de clique es a lo más $n-1$ *y bajo la hipótesis alternativa es al menos* n.

Ahora, daremos una descripción más detallada de la prueba para los casos en que p_0 es un valor conocido y cuando p_0 es desconocido.

p_0 conocido:

Para poder definir nuestra prueba y encontrar una región de rechazo para los valores de la estadística C (calibrar la prueba), necesitamos encontrar la distribución que sigue C, esto se puede conseguir con técnicas combinatoriales para calcular probabilidades.Este primer enfoque resulta demasiado complicado, por lo que se aproximará los valores de la distribución usando un método Monte-Carlo:

1. Se toman M muestras de la hipótesis nula $G(N, p_0)$.
2. Se calculan los números de clique de cada una de las muestras.
3. Se obtiene una distribución del número de clique, bajo la hipótesis nula.
4. Se calcula el cuantil q de probabilidad $(1 - \alpha) \times 100\,\%$.
5. La prueba $(C, (q, \infty))$ es de nivel α.

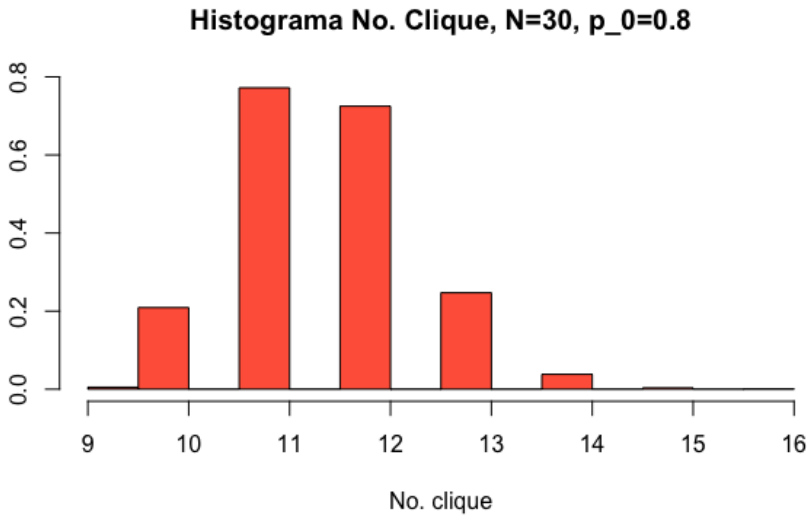


Figura 3: Distribución del número de clique para $G(30, 0.8)$ obtenida con $M = 10,000$ muestras.

Para verificar la potencia de nuestra prueba, simulamos 1000 muestras de la hipótesis nula y 1000 muestras de la hipótesis alternativa para observar las probabilidades de rechazo estimadas, las cuales resultaron ser:

$$P_{H_0}[\phi = 1] = 0.025$$

$$P_{H_1}[\phi = 1] = 0.56$$

Con lo que obtenemos que no rechaza mas del 5 % de las veces en los casos en que no hay un clique y que rechaza en más del 50 % de los casos en que existe la presencia de un clique.

p_0 desconocido:

Para este caso no podemos aplicar un método Monte-Carlo para calibrar nuestra prueba, pues la hipótesis nula incluye todas las distribuciones de la forma $G(N, p_0)$. Entonces recurrimos a un método de Bootstrap paramétrico:

1. Encontramos el estimador de mxima verosimilitud $\hat{p_0}$ del parámetro p_0 de nuestra gráfica observada G.
2. Realizamos M muestras bootstrap de $G(N, \hat{p_0})$.
3. Obtenemos la distribución del número de clique con este parámetro.
4. Calculamos el quantil q de probabilidad $(1 - \alpha) \times 100\,\%$.
5. Realizamos la prueba $(C, (q, \infty))$ de nivel α.

La idea de estimar el parámetro de máxima verosimilitud de p_0 sigue la filosofía de encontrar la distribución en la hipótesis nula que se parezca más a lo que observamos en nuestros datos.

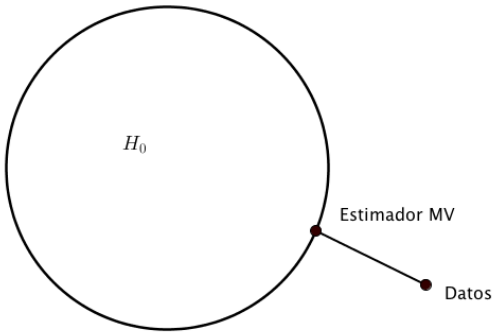


Figura 4: Distribución en H_0 mas parecida a los datos.

En la actualidad no se conoce un algoritmo que calcule el número de clique en tiempo polinomial, por lo que esta prueba resulta bastante costosa en recursos computacionales. Al ser una prueba asintóticamente potente, si existiera un algoritmo que

estime el número de clique en tiempo polinomial, tendríamos una prueba muy buena.

Para fines prácticos, es mejor realizar otra prueba de hipótesis que primero detecte la presencia de un clique, para después encontrar de manera directa dicho clique.

3.3. Estadísticas de total de aristas y escaneo.

En el caso general cuando $p_1 \in (p_0, 1]$ existen diversas estadísticas para detectar una comunidad. Comenzaremos definiendo dos de ellas y discutiremos algunas de sus propiedades.

Si A es la matriz de adyaciencia de una gráfica G, definimos la **estádistica del total de aristas** como:

$$W = \sum_{1 \leq i < j \leq N} A_{i,j},$$

es decir, la cantidad total de aristas en la gráfica.

La idea para considerar el total de aristas está fundamentada en que los vértices que conforman una comunidad tienen una mayor probabilidad de unirse, por lo que la gráfica tiene una mayor cantidad de aristas. Entonces en nuestra prueba rechazaremos cuando la cantidad de aristas es muy grande, por lo que necesitamos encontrar el cuantil que acumula α de probabilidad en la cola derecha de nuestra distribución y la prueba sería de la forma $(W, (q_\alpha, \infty))$.

Cuando p_0 es conocido, podemos determinar el cuantil q_α para formar nuestra región de rechazo de la prueba de dos formas. La primera es usando un método Monte-Carlo similar al que usamos en la estadística del número de clique. La segunda forma es calcular la distribución de W, la cual se puede probar que bajo la hipótesis nula sigue una distribución binomial de parámetros $W \sim \text{Binom}(\binom{N}{2}, p_0)$.

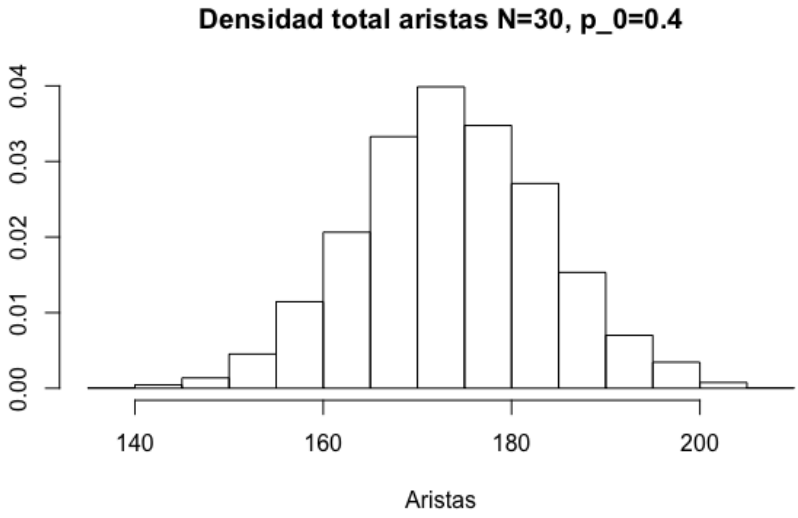


Figura 5: Método Monte-Carlo para total de aristas.

Para un ejemplo en un modelo $G(30, 0.4)$ se calculó el cuantil de probabilidad $\alpha = 0.05$ siguiendo ambos caminos, y en los dos se obtuvo que el valor a partir del cual debemos rechazar es 191.

Cuando p_0 es desconocido ya no conocemos la distribución de W bajo la hipótesis nula, por lo que nuestra única alternativa es usar un método de bootstrap paramétrico. Estimamos el valor de p_0 y calculamos el valor a partir del cual nuestra prueba deberá rechazar, dependiendo del nivel de confianza de la prueba.

Notemos que es posible realizar todo esto en un tiempo polinomial, pues para contar las aristas hace falta verificar las $N(N-1)/2$ parejas de vértices existentes. Es decir, la estadística W se obtiene en un tiempo de orden $O(N^2)$.

Un aspecto importante de esta prueba, es que no depende directamente del valor de p_1 ni de n, que son los valores de los que depende la hipótesis alternativa. La forma en que afecta a la prueba el valor de p_1, es en el hecho de que $p_0 < p_1$, por lo que se rechazará para valores altos de W.

Ahora introduciremos otra estadística. Sea A la matriz de adyaciencia de una gráfica G, entonces definimos la **estadística de escaneo** como:

$$W_n := \max_{|S|=n} A_S, \quad A_S := \sum_{i,j \in S, i<j} A_{i,j},$$

es decir el máximo de aristas contenidas en alguna subgráfica de n vértices.

Esta estadística conserva la idea de la estadística W, pues ante la presencia de una comunidad el valor de W_n será grande, por lo que la prueba rechaza para valores elevados de W_n.

Podemos notar que la estadística W_n, a diferencia de la estadística W, incluye más información presente en la hipótesis nula, pues es necesario conocer el tamaño n de la comunidad, pero sigue sin incluir el valor de la probabilidad p_1.

Para encontrar la región de rechazo asociada a esta estadística y poder construir nuestra prueba, procedemos de manera análoga a como lo hicimos con la estadística de total de aristas, se comporta de manera similiar al rechazar para valores grandes de W_n. Incluso lo podemos dividir en los mismos dos casos, con p_0 conocida y desconocida.

Esta prueba presenta algunas ventajas sobre la de total de aristas, la principal de ellas es que es asintóticamente potente, por lo que trabaja bien para valores grandes de N. Pero su principal desventaja es que no se conoce un algoritmo para calcular la estadística W_n en tiempo polinomial. Por lo tanto esta prueba es muy costosa computacionalmente y puede llegar a tardar mucho si se aplica en gráficas con un valor de N muy grande.

3.4. Estadística total de triángulos

En esta sección presentaremos una nueva estadística, la cual presenta diversas propiedades que resultan de gran interés tanto en la tería, como en la práctica.

Sea G una gráfica. Definimos la **estadística total de triángulos** como:

$$T = |\{\text{Triángulos en G}\}|.$$

La estadística T es de las más populares entre las gráficas que consisten en contar patrones y es de las mas simples y de menor costo computacional. El algoritmo para contar el total de triángulos en una gráfica se basa en el siguiente resultado:

Teorema 2. *Si A es la matriz de adyaciencia de una gráfica G, la cantidad total de caminos para llegar del vértice i al vértice j en k pasos es:*

$$(A^n)_{i,j}.$$

La prueba de este teorema se puede encontrar en [4], junto con otros resultados interesantes de combinatoria en gráficas

.

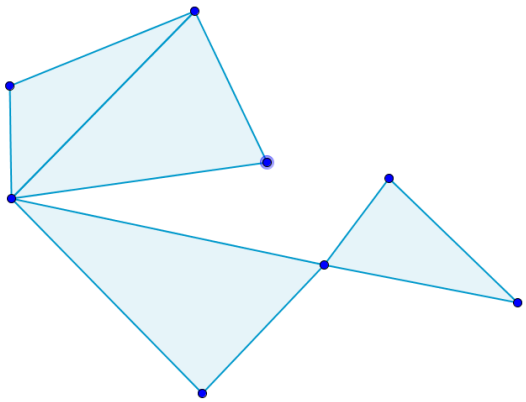


Figura 6: Los 4 triángulos de una gráfica.

Dado un vértice i de la gráfica G, notemos que que los únicos caminos de longitud 3 que parten de i y regresan a i en la gráficas son triángulos. Por lo tanto, la cantidad total de triángulos en una gráfica se puede obtener con la siguiente fórmula:

$$T = \frac{\text{traza}(A^3)}{6}$$

El $6 = 3!$ que aparece dividiendo corresponde al número de veces que repetimos en el conteo a cada triángulo.

Este resultado se puede verificar de manera intuitiva con el siguiente ejemplo, donde A es la matriz de adyaciencia de la gráfica que se muestra enseguida:

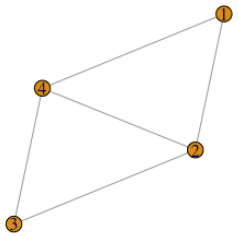


Figura 7: Gráfica con matriz de adyaciencia A.

$$A = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

$$A^2 = \begin{pmatrix} 2 & 1 & 2 & 1 \\ 1 & 3 & 1 & 2 \\ 2 & 1 & 2 & 1 \\ 1 & 2 & 1 & 3 \end{pmatrix} \quad A^3 = \begin{pmatrix} 2 & 5 & 2 & 5 \\ 5 & 4 & 5 & 5 \\ 2 & 5 & 2 & 5 \\ 5 & 5 & 5 & 4 \end{pmatrix}$$

$$T(G) = 12/6 = 2$$

Otro aspecto importante de la estadística de total de triángulos es que nos dá información importante sobre la estructura topológica de la gráfica, que se usa para el estudio de redes que aparecen en la vida real. Mas específicamente, es posible calcular el coeficiente de agrupamiento (cluster coefficient) usando la cantidad de triángulos:

$$C = \frac{3 \times \text{Número de triángulos}}{\text{Número de tripletas}},$$

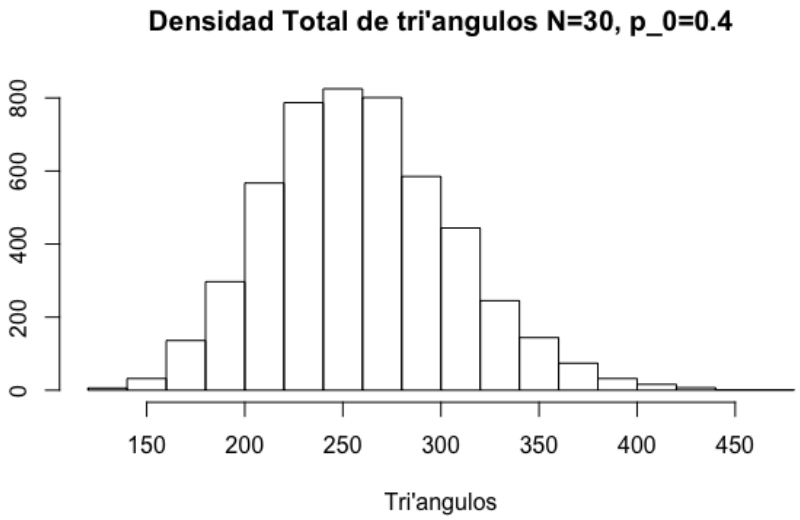
el cual se usa para cuantificar el agrupamiento de redes en el Internet [3].

La estadística de total de triángulos nos resulta particularmente interesante, pues su prueba asociada es asintóticamente potente y, como ya vimos, es posible calcular la estadística en tiempo polinomial. Por lo que combina dos propiedades deseables en las pruebas.

Con el estadístico de total de triángulos rechazaremos para valores grandes de T, dado que bajo la hipótesis alternativa,hay mayor probabilidad de conexión en nuestra comunidad, lo que produce una mayor presencia de triángulos en la gráfica. Otra interpretación equivalente, es que al existir una comunidad se incrementa el coeficiente de agrupamiento.

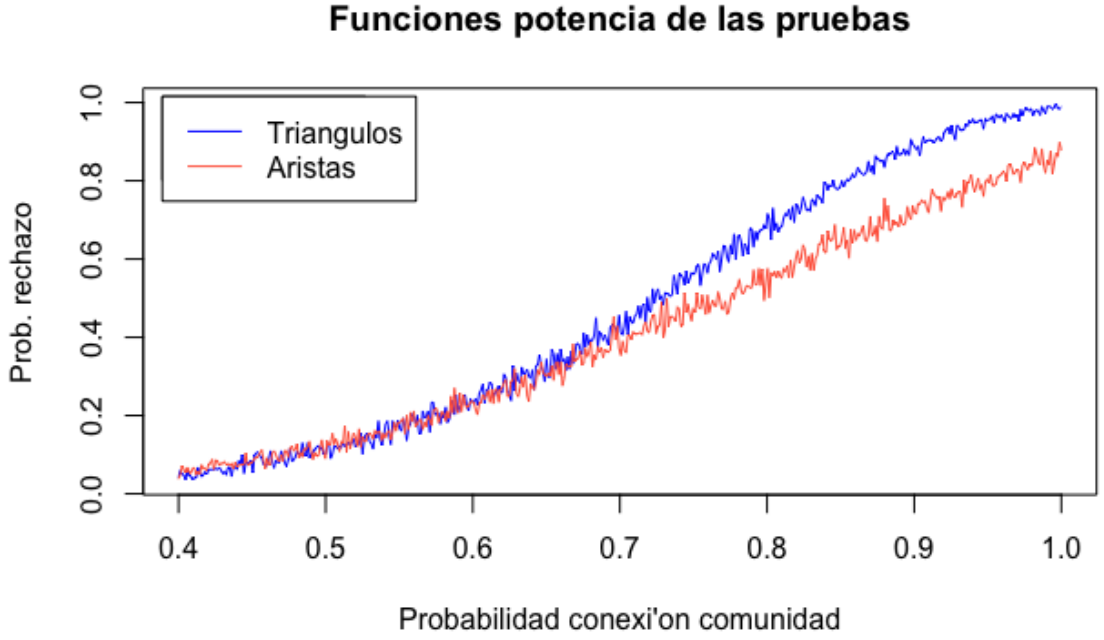
Nuevamente, el proceso para calibrar la prueba, es decir encontrar la región de rechazo, es similar a los procedimientos que ya hemos visto con las otras estadísticas para los casos en que p_0 es conocido o desconocido.

Con un proceso de Monte-Carlo se obtuvo la densidad del número del total de triángulos para un modelo $G(30, 0.4)$, y se obtuvo el valor del cuantil que acumula 5 % en la cola derecha, por lo que rechazaremos para valores mayores a 343



Se realizaron aproximaciones con un método Monte-Carlo para estimar la función potencia de las estadísticas de número de triángulos y de total de aristas para el par de hipótesis:

$$H_0 : \mathcal{G} \sim G(30, 0.4) \text{ v.s. } H_1 : \mathcal{G} \sim G(30, 0.4; 10, p_1), p_1 > 0.4.$$



Podemos notar que dichas las dos pruebas tienen comportamientos similares cuando el valor de p_1 es cercano a $p_0 = 0.4$, pero conforme la probabilidad de conexión de la comunidad aumenta la función potencia de la prueba con la estadística de triángulos es mayor que la prueba con la estadística de aristas.

4. Conclusiones.

La investigación básica en la teoría de gráficas aleatorias resulta ser la base para el análisis y desarrollo de aplicaciones en redes. Por ejemplo, siempre se pueden realizar métodos Monte-Carlo y métodos Bootstrap para encontrar valores relacionados a las distribuciones de las estadísticas vistas, pero se esta realizando investigación actual para encontrar las distribuciones del total de triángulos o de otros patrones presentes en las gráficas.

La investigación y el desarrollo de teoría en la resolución de los problemas computacionales que se presentan en el contexto de gráficas aleatorias es de gran importancia por sus aplicaciones en redes sociales, en donde los datos tienen una dimensionalidad muy elevada.

Referencias

[1] J. Reichardt y S. Bornholdt. "Statistical mechanics of community detection". En: *Phys. Rev* 74 (2006).

[2] E. Arias-Castro y N. Verzelen. "Community detection in dense random networks". En: *The Annals of Statistics* 42 (2014), págs. 940-969.

[3] E. Arias-Castro y N. Verzelen. "Community detection in sparce random networks". En: *Annals of Applied Probability* 25 (2015), págs. 3465-3510.

[4] B. Bollobás. "Random graphs (second ed.)" En: *Cambridge Studies in Advanced Mathematics* 73 (2001).