Grafos Small-World

Leonardo Torres

Contents

CHAPTER 1

Introducción

Grafos

- Grafo: nodos y aristas. Grafos simples. Subgrafo.
- Aristas dirigidas y con pesos.
- Matriz de adyacencia.
- Caminos, ciclos y geodésicas.
- Grafos infinitos.
- Formas de estudio: álgebra, algoritmia, combinatoria, probabilidad, redes.

Redes

- Qué es: conceptos y métodos
- Aplicaciones
- Aplicaciones a Neurociencia

Small-Worlds

- Historia, 6 grados de separación
- Ejemplos de la vida real
- Modelo original: Watts y Strogatz

Definiciones formales.

- Small world effect vs small world graph
  - problema: ambigüedad
- Indicadores
  - problema: dependen de otro grafo (no son un número para un grafo, sino una distribución: no se comporta como otros indicadores más claros)
- Definiciones asintóticas
  - problema: dependen del proceso de crecimiento del grafo

Difusión. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam et dignissim sem. Nullam maximus ante ac porttitor blandit. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Pellentesque quis quam tincidunt, suscipit sem quis, eleifend ipsum. Integer eget eleifend nisl. Maecenas et vulputate diam. Curabitur luctus consequat fringilla. Etiam pretium ligula et dui interdum, vitae porta erat maximus. Aliquam erat volutpat. Proin tempus nibh tincidunt quam eleifend venenatis.

En el presente trabajo,...

- Se presenta la defincíon formal asintótica de Cont y Tanimura
- Se discute un modelo small world

CHAPTER 2

Preliminares

Convenciones

Antes de comenzar, indicaremos las convenciones de notación que usaremos en el resto de nuestra discusión.

Si A es cualquier conjunto, escribiremos $|A|$ para denotar su cardinalidad. Cada vez que hablemos de un grafo $G = (\mathscr{N}, \mathscr{E})$, identificaremos los nodos $i \in \mathscr{N}$ con los números naturales $1, 2, ..., |\mathscr{N}|$. Más aún, identificaremos a G con el conjunto de nodos, $G \equiv \mathscr{N}$, es decir $i \in G \iff i \in \mathscr{N}$, $\forall i = 1, 2, ..., |G|$ y $|G| = |\mathscr{N}|$. Por otro lado, escribiremos la arista dirigida que va de i a j como el par ordenado (i, j). En nuestro caso, como trataremos siempre con aristas no dirigidas, tendremos que $(i, j) \in \mathscr{E}$ implica $(j, i) \in \mathscr{E}$. Por ello, escribiremos (i, j) también para la arista no dirigida que une i con j. En este caso, diremos que i y j son *vecinos* y escribiremos $V(i)$ para el conjunto de nodos vecinos de i. Para complicar las cosas, a veces identificaremos también a G con su conjunto de aristas $\mathscr{E}$. Esto no causará ambigüedad, ya que los nodos son números, mientras que las aristas son pares ordenados. De esta manera, escribiremos $i \in G$ y $(i, j) \in G$ sin lugar a confusión. Mientras que escribimos $(i, j) \in G$ para las aristas *existentes* en el grafo G, escribiremos $(i, j) \in G \times G$ para una arista *posible* del grafo. Por ejemplo, si $G = (\{1, 2, 3\}, \{(1, 2), (2, 1)\})$, son verdaderas las afirmaciones $|G| = 3$, $4 \notin G$, $(1, 2) \in G$, $(1, 3) \notin G$, y $(1, 3) \in G \times G$.

A continuación, definiremos los espacios con los que trabajaremos. Llamaremos Γ_n al conjunto de grafos simples, no dirigidos, de aristas binarias y de n nodos,

$$\Gamma_n = \{G = (\mathscr{N}, \mathscr{E}) : \ |G| = n, \ (i, i) \notin G, \ (i, j) \in G \iff (j, i) \in G\}.$$

Usualmente, identificaremos cada grafo con su matriz de adyacencia. De ahí que el conjunto Γ_n esté en biyección con el conjunto de matrices simétricas con entradas 0 ó 1: si $G \in \Gamma_n$ entonces $G \equiv M$, $M \in \{0, 1\}^{n \times n}$, donde M es la matriz de adyacencia correspondiente a G. Por tratarse de grafos simples, si las entradas de la matriz de adyacencia M son (a_{ij}), tendremos $a_{ii} = 0$, $\forall i = 1, 2, ..., |G|$. Además, escribiremos $\Gamma_\infty = \bigcup_{n=1}^{\infty} \Gamma_n$ para el conjunto de todos los grafos simples, no dirigidos y de aristas binarias.

A este último conjunto lo dotaremos de un σ-álgebra, convirtiéndolo en un espacio medible. En realidad, usaremos el σ-álgebra usual de boreleanos del conjunto de matrices $\{0, 1\}^{n \times n}$, y lo trasladaremos a Γ_n mediante la biyección mencionada. Bajo este σ-álgebra, llamémosle $\mathscr{B}$, una función $f : \Gamma_\infty \longrightarrow \mathbb{R}$ es $\mathscr{B}$-medible si cada una de sus proyecciones $f\big|_n : \Gamma_n \longrightarrow \mathbb{R}$ es medible, cuando es vista como función del espacio de matrices.

Definiciones

Pasaremos ahora a definir los elementos sobre los que construiremos los modelos que queremos estudiar.

Indicadores gráficos. La función $f : \Gamma_\infty \longrightarrow \overline{\mathbb{R}}$ es un indicador gráfico si es $\mathscr{B}$-medible. Por lo mencionado anteriormente, para ello es suficiente expresarla como una función medible de las matrices de adyacencia. En nuestra discusión, nos interesarán en particular cuatro indicadores. Pasamos a definir tres de ellos y dejaremos el último para una sección posterior.

Sea el grafo $G \in \Gamma_\infty$, con matriz de adyacencia $M = (a_{ij})$ y consideremos los nodos i y j de G. Entonces definimos los siguientes indicadores.

Grado y grado promedio. Estos indicadores son las cantidades típicas que se manejan en todas las áreas de la Teoría de Grafos. $deg(i)$ es el número de vecinos del nodo i, mientras que $deg(G)$ es la media aritmética de ellos.

$$deg_G(i) = \sum_{j=1}^{|G|} a_{ij} \qquad deg(G) = \frac{1}{|G|} \sum_{i=1}^{|G|} deg_G(i).$$

Coeficiente de agrupación local. Hay varios coeficientes de agrupación usados en Teoría de Redes, todos los cuales miden el grado de transitividad de un grafo de maneras ligeramente distintas. El que nos interesará a nosotros es el *coeficiente de agrupación local*, definido de la siguiente manera,

$$c_i = \frac{\sum_{j \neq i}^{|G|} \sum_{k \neq j, k \neq i} a_{ij} a_{ik} a_{jk}}{deg_G(i)(\frac{deg_G(i)-1}{2})}.$$

Esta expresión puede parecer atemorizante a primer vista. Sin embargo, ella se lee como sigue. *El coeficiente de agrupación local del nodo i es el número de triángulos presentes con un vértice en i, sobre el número total de triángulos que i podría formar.* Como las entradas a_{ij} son solo 0 ó 1, y como los nodos i, j, k forman un triángulo solo si el producto $a_{ij}a_{jk}a_{ik}$ es igual a 1, el numerador de la expresión es igual al número de triángulos presentes en el grafo con un vértice en i. Por otro lado, el denominador es simplemente el número total de aristas que puede haber entre vecinos de i, o, equivalentemente, el número total de posibles triángulos que incluyen a i.

A partir de estos c_i, definimos un indicador más usado, llamado el *coeficiente de agrupación local promedio de G*, denotado $\overline{c}(G)$ ó simplemente $\overline{c}$, cuando no haya lugar a ambigüedad, $\overline{c}(G) = \frac{1}{|G|} \sum_{i=1}^{|G|} c_i$. Abreviaremos $\overline{c}$ como *agrupación local de G.*

Diámetro. Es un indicador global del "tamaño" del grafo.

$$diam(G) = \max_{i \in G,\, j \in G} \{d(i,j)\},$$

donde $d(i,j)$ es el menor número de aristas en todos los caminos que empiezan en i y terminan en j. Si no hay un camino entre los nodos i, j, escribimos $d(i,j) = \infty$ y, en ese caso, tenemos $diam(G) = \infty$.

Claramente, estas funciones tienen el derecho de llamarse indicadores gráficos, ya que las podemos expresar como funciones medibles de las matrices de adyacencia.

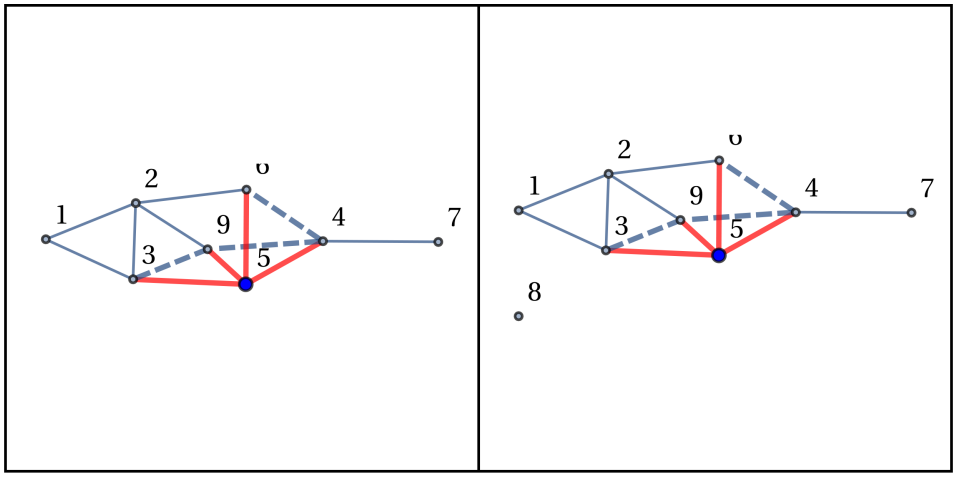

FIGURE 2.0.1. Indicadores gráficos.

Definidos nuestro espacio medible y funciones medibles sobre él, pasaremos a incluir aleatoriedad en nuestro estudio. En lo subsiguiente, supondremos que existe un espacio de probabilidad abstracto $(\Omega, \mathscr{F}, \mathcal{P})$ que contiene el espacio muestral de todos los resultados posibles de nuestros experimentos.

Grafos e Indicadores aleatorios. Un grafo aleatorio es una variable aleatoria $G : \Omega \longrightarrow \Gamma_\infty$. Dados un grafo aleatorio G y un indicador gráfico Q, la variable aleatoria $Q \circ G : \Omega \longrightarrow \overline{\mathbb{R}}$ se llama indicador aleatorio. El cuarto indicador que faltaba mencionar es un indicador que *siempre* es aleatorio, sin importar si el grafo bajo estudio lo es o no.

Distancia típica. Dado un grafo $G \in \Gamma_\infty$, aleatorio o determinístico, definimos $U : \Omega \to G \times G \setminus \{(i,i) : i \in G\}$, un par de nodos distintos $u \neq v$ del grafo G, elegidos de manera uniforme. En este caso, definimos la distancia típica de G, T_G como la distancia entre ellos

$$T_G = d(U) = d(u, v).$$

Modelo de crecimiento de grafos. Un modelo de crecimiento será una sucesión $(G_n)_{n=1}^\infty$ donde $G_n \in \Gamma_n$. Si todos los G_n son grafos aleatorios, la sucesión será un modelo de crecimiento aleatorio.

Comportamiento asintótico de indicadores. Para este trabajo, daremos un modelo específico de crecimiento aleatorio y nos interesará estudiar el comportamiento de ciertos indicadores cuando $|G| \to \infty$. Sean $(G_n)_{n=1}^\infty$ un modelo de crecimiento (determinístico) y Q un indicador. Si $f : \mathbb{N} \longrightarrow \mathbb{R}$ es una sucesión determinística de números, decimos que es una

Cota superior de crecimiento de (G_n). Si se cumple

$$\limsup_{n\to\infty} \frac{Q(G_n)}{f(n)} \leq 1.$$

O una

Cota inferior de crecimiento de (G_n). Si

$$\limsup_{n\to\infty} \frac{Q(G_n)}{f(n)} \geq 1.$$

En el caso en que $f(n)$ sea cota superior y otra sucesión $g(n)$ sea cota inferior del crecimiento de (G_n) y si, además, se cumple

$$\lim_{n\to\infty} \frac{f(n)}{g(n)} = c \in \mathbb{R},$$

diremos que $Q(G_n)$ *crece como* $f(n)$.

Comportamiento asintótico de indicadores aleatorios. Ahora, consideremos un modelo de creimiento aleatorio $(G_n)_{n=1}^{\infty}$. Sean Q un indicador y f una sucesión de números. En este caso, f será una

Cota superior del crecimiento de (G_n) en esperanza. Si

$$\limsup_{n\to\infty} \frac{\mathsf{E}[Q(G_n)]}{f(n)} \leq 1.$$

Cota superior del crecimiento de (G_n) en probabilidad. Cuando

$$\limsup_{n\to\infty} P\big(\frac{Q(G_n)}{f(n)} \leq 1\big) = 1.$$

Cota superior del crecimiento de (G_n) casi ciertamente. En el caso

$$P\big(\limsup \frac{Q(G_n)}{f(n)} \leq 1\big) = 1.$$

Análogamente, se definen las cotas inferiores en esperanza, en probabilidad y casi ciertas. Claramente, estas definiciones van en orden ascendente de rigor.

Small-World. Finalmente, podemos presentar la definición formal de grafo *small-world* usada por Cont y Tanimura.

Un modelo de crecimiento aleatorio $(G_n)_{n=1}^{\infty}$ se llama *small-world* si

(1) $\exists \mathrm{a}_1 \geq 0 :\ f(n) = \mathrm{a}_1 \log(n)$ es cota superior del crecimiento de $deg(G_n)$,
(2) $\exists \mathrm{a}_2 > 0 :\ \overline{c}(G) \geq \mathrm{a}_2,\ \forall n = 1, 2, ..., \infty$,
(3) $\exists \mathrm{a}_3 \geq 0 :\ g(n) = \mathrm{a}_3 \log(n)$ es cota superior del crecimiento de T_{G_n}.

Si el modelo (G_n) es aleatorio, cualquiera de las tres propiedades puede cumplirse en esperanza, en probabilidad o casi ciertamente.

Observación. En la definición, acotamos el grado esperado del modelo aún cuando no mencionamos tal restricción en la discusión previa. Esto se hace por la razón práctica de evitar que ciertos ejemplos triviales cumplan con nuestra definición. Por ejemplo, el grafo completo, como se mencionó antes, cumple con el efecto *small-world,* pero no con esta nueva definición. Más aún, podemos considerar la idea intuitiva de que, si el grado promedio fuera grande, sería muy "fácil" para todos los nodos cumplir con ambas restricciones de agrupación y distancia. Quizás se cumplen hasta por coincidencia. Queremos estudiar aquellos grafos que cumplen con la definición, sin que sea una consecuencia directa de la alta conectividad inherente (grado).

OBSERVACIÓN. La restricción sobre el coeficiente de agrupación es ciertamente bastante leve. Si el modelo es aleatorio, bastaría que, asintóticamente en el número de nodos, el número esperado de triángulos formado por un nodo sea igual a 1 (si el grado de ellos no cambia). Sin embargo, si cada nodo estuviera tan poco conectado como para formar solo un triángulo, sería difícil para la distancia típica ser tan corta como estamos requiriendo.

Ejemplos y contraejemplos. En los siguientes ejemplos, consideremos un modelo de crecimiento de grafos $(G_n)_{n=1}^{\infty}$ arbitrario.

Triviales: Los modelos de grafo completo $G_n = K(n)$, grafo vacío $G_n = \overline{K}(n)$, grafo estrella $G_n = S_{n-1}$ y grafo árbol $G_n = \acute{a}rbol(n)$, no son modelos *small-world.* El grafo completo no cumple con la limitación del grado; los otros tres no cumplen con la cota de agrupación. (Los árboles no pueden tener triángulos).

Erdos-Renyi: Si $G_n = G(n,p)$, entonces se cumplen las restricciones necesarias sobre la distancia típica (de hecho, aún mejor, sobre el diámetro) y sobre el coeficiente de agrupación, pero no la del grado. Si, en cambio, consideramos la versión que mantiene el grado esperado constante, $G_n = \overline{G}(n,p_n)$, donde $\mathsf{E}[deg_{G_n}(i)] = p_n(n-1)$, entonces sí se cumple la definición.

Látices: Un látice regular (por ejemplo, el modelo de Watts y Strogatz con $p = 0$) cumple trivialmente las condiciones de agrupación y grado. Sin embargo, cuando se aumentan nuevos nodos al grafo, la distancia aumenta como $\sqrt[d]{n}$, donde d es la dimensión del látice.

Watts-Strogatz: Si $k > 0$ y $p > 0$ son fijos y $G_n = WS(n,k,p)$, entonces G_n cumple la definición.

Exponencial: POR PROBAR

CHAPTER 3

El modelo *small-world* de Cont y Tanimura

El modelo simplificado

Este modelo viene de la necesidad de construir un modelo de grafo *small-world* que no dependa de una estructura regular subyaciente. La idea principal es empezar con un número de "comunidades" (subgrafos completos) y unirlas de manera aleatoria. En esta versión simplificada, consideraremos el caso en que todas las comunidades tienen el mismo número de nodos. Hacemos esto para esclarecer la prueba y familiarizarnos con el modelo. En la siguiente sección, levantaremos esta restricción.

Construcción. Fijemos los números $n, \delta, M \in \mathbb{N}$ con $n = \delta \times M$ y consideremos $G \in \Gamma_n$ el grafo de n nodos sin aristas. Añadiremos aristas a G en tres pasos.

(1) Particionamos los nodos $1, 2, ..., n$ en M conjuntos disjuntos, $G^1, G^2, ..., G^M$, con δ nodos cada uno. Si $i \in G^k$, decimos que G^k es la *comunidad de origen del nodo i.* Cada i lo unimos con una arista a cada nodo que tenga la misma comunidad de origen: $i, j \in G^k \Longrightarrow (i, j) \in G$.

(2) Para cada $i \in G$, sean las variables aleatorias $X_i : \Omega \longrightarrow \{G^1, G^2, ..., G^M\}$, independientes e indénticamente distribuidas de manera uniforme sobre el conjunto de comunidades, $X_i \sim U(G^1, G^2, ..., G^M)$. En la práctica, identificaremos cada G^k con el número k, excepto en casos ambiguos. De esta manera, si $X_i = k$, diremos que G^k es la *comunidad secundaria o de destino del nodo i.* A veces diremos que el nodo i *escoge* X_k como destino. Para cada i, añadimos a G las aristas (i, j), $\forall j \in X_i$. Es decir, unimos cada nodo con todos los nodos en su comunidad secundaria.

(3) Si los nodos i, j cumplen $X_i = X_j$, añadimos a G la arista (i, j). Esto es, unimos todos los nodos que tengan la misma comunidad secundaria.

Cabe recordar que, como G es un grafo no dirigido, basta añadir la arista (i, j) para que (j, i) también esté presente.

Podemos describir la construcción de G de la siguiente manera. Dado $m = 1, 2, ..., M$, sea

$$A^m = G^m \cup \{i : X_i = G^m\}$$

el conjunto de los nodos que tienen G^m como comunidad de origen o de destino. Entonces,

$$(i, j) \in G \Longleftrightarrow i \neq j \ y \ (i, j) \in \bigcup_{m=1}^{M} A^m \times A^m.$$

De esta manera, podemos ver a G como los M subgrafos completos A^m que están unidos de manera aleatoria.

Hecha la misma construcción para cada n (o, mejor dicho, para cada n múltiplo de δ), queremos demostrar la siguiente afirmación.

AFIRMACIÓN 1. El modelo $(G_n)_{n=1}^{\infty}$ es un modelo de crecimiento de grafos *small-world.*

La demostración la dividimos en tres partes, una para cada característica propia de los grafos *small-world.* Los resultados propios del grado y el coeficiente de agrupación local corresponden a las Proposiciones 1 y 2, respectivamente. El estudio de la distancia típica abarca las Proposiciones 3 y 4.

PROPOSICIÓN 1. *El grado esperado del modelo está acotado superiormente por* $4\delta - 1$.

PROOF. Escribimos $S_m = |\{i : X_i = m\}|$, de donde $|A^m| \leq \delta + S_m$. El número de aristas en G está acotado por

$$|\mathcal{E}| \leq \sum_{m=1}^{M} \frac{|A^m|(|A^m|-1)}{2} \leq \sum_{m=1}^{M} \frac{(\delta + S_m)(\delta + S_m - 1)}{2}.$$

Por ser las variables X_i i.i.d. y uniformes, tenemos

$$P(S_1 = x_1, S_2 = x_2, ..., S_M = x_M) = \begin{cases} \frac{n!}{x_1! x_2! \ldots x_M!}(\frac{1}{M})^n, & \sum_{k=1}^{M} x_k = n \\ 0, & \text{de otro modo} \end{cases},$$

i.e., el vector $(S_1, S_2, ..., S_M)$ tiene distribución multinomial. Entonces, tenemos

$$\mathsf{E}[S_m] = \frac{n}{M} = \delta, \quad \mathsf{Var}[S_m] = \frac{n}{M}\big(1 - \frac{1}{M}\big) = \delta\big(1 - \frac{1}{M}\big), \quad \mathsf{E}[S_m^2] = \delta\big(1 - \frac{1}{M}\big) + \delta^2$$

Y, en consecuencia,

$$\mathsf{E}[|\mathcal{E}|] \leq \frac{1}{2}\sum_{m=1}^{M} \mathsf{E}[(\delta + S_m)(\delta + S_m - 1)] = \frac{\delta}{2}(4\delta M - M - 1).$$

Finalmente,

$$\mathsf{E}[deg(G)] = \frac{2\mathsf{E}[|\mathcal{E}|]}{n} = 4\delta - 1 - \frac{\delta}{n}.$$

□

PROPOSICIÓN 2. *En el modelo* $(G_n)_{n=1}^{\infty}$*, el coeficiente de agrupación local de cada nodo* $i \in G_n$ *cumple*

$$c_i \geq \frac{1}{2} - \frac{1}{\delta - 1}, \forall n = 1, 2, ...$$

PROOF. Fijado un nodo i, sean G^j y G^k sus comunidades de origen y destino, respectivamente. Entonces, podemos particionar $G(i)$ de la siguiente manera:

$$G(i) = (G^j \cup \{u : X_u = G^j\}) \setminus \{i\} \bigcup (G^k \cup \{u : X_u = G^k\}) \setminus \{i\}.$$

Ambos conjuntos tienen al menos $\delta - 1$ nodos. Aplicando el Lema 1 del Apéndice A, con $l = 2$, $h \geq \delta - 1$, tenemos

$$c_i \geq \frac{1}{2} - \frac{1}{\delta - 1}.$$

□

OBSERVACIÓN. Para que el modelo sea *small-world,* necesitamos que cada $c_i > 0$. Claramente, esto ocurrirá en el modelo simplificado cuando $\delta \geq 3$.

PROPOSICIÓN 3. *En el modelo $(G_n)_n^\infty$, cuando $n \to \infty$, es cada vez más probable que podamos alcanzar por lo menos la mitad de los nodos en un número logarítmico de pasos, habiendo empezado desde un nodo arbitrario.*

Concretamente, para n suficientemente grande, sea un nodo $u \in G_n$. Entonces, existe un conjunto de nodos $\overline{S}(u,n) \subseteq G_n$ tal que $|\overline{S}| \geq \frac{n}{2}$ y, si $v \in \overline{S}$, entonces

$$P(d(u,v) \leq 4\log(n) + 4) \geq 1 - \frac{1}{\delta^2 n^2}.$$

PROOF. Fijamos n para el resto de nuestro argumento. Probaremos que, si empezamos desde una comunidad arbitraria, digamos G^1, podemos alcanzar al menos la mitad de las demás comunidades. Como cada comunidad es un subgrafo completo y todas ellas tienen el mismo número δ de nodos, esto probará la Proposición. En lo subsiguiente, identificaremos el conjunto de comunidades $\{G^1, G^2, ..., G^M\}$ con el conjunto de números $\{1, 2, ..., M\}$.

Definimos el proceso de descubrimiento de comunidades C_k, N_k, empezando desde $G^1 \equiv 1$, recursivamente.

$$C_0 = \{1\}, \qquad N_k = \{i \in G_n : \exists d \in C_k,\ i \in d\} = \bigcup_{d \in C_k} G^d$$

$$C_k = \{G \notin \bigcup_{l=0}^{k-1} C_l : \exists j \in N_{k-1},\ X_j = G\},\ k = 0, 1, ...$$

En palabras, C_k es el conjunto de comunidades que se pueden alcanzar en, como mucho, k pasos; mientras que N_k es el conjunto de nodos de estas comunidades. Notamos que los C_k son disjuntos dos a dos.

También definimos los eventos

$$E_k = \{|C_k| > e|C_{k-1}|\},\ k = 1, 2, ...$$

E_k implica que el número de comunidades que alcanzamos en el último paso es al menos e veces mayor que el número anterior. $\bigcap_{m=1}^{l} E_m$ implica que hemos descubierto comunidades a una velocidad exponencial, hasta el l-ésimo paso.

Ahora, estamos interesados en estimar la probabilidad de descubrir nuevas comunidades a una velocidad exponencial hasta descubrir al menos la mitad de ellas. Para ello, definimos los tiempos de parada

$$\tau_u = \min\{t \in \mathbb{N} : \sum_{j=0}^{t} |C_j| \geq u\},\ u = 1, 2, ...$$

Para un T dado, estimamos la probabilidad $P(\bigcap_{j=1}^{T} E_k) = \prod_{j=1}^{T} P(E_k | \bigcap_{m=1}^{j-1} E_m)$.[1] Estudiamos el lado derecho de la igualdad, para cada j, según tres casos distintos: $j = 1$, $2 \leq j \leq \tau_{\sqrt{M}}$ y $\tau_{\sqrt{M}} < j \leq \tau_{\frac{M}{2}}$. Queremos encontrar que estas probabilidades son cada vez más grandes.

[1] Donde tomamos la convención $P(E_1 | \bigcap_{m=1}^{m=0} E_m) = P(E_1)$.

$j = 1$**:** Estimamos la probabilidad $P(E_1^c) = P(|C_1| \leq 2)$. $\{|C_1| \leq 2\}$ ocurre cuando todos los nodos de G^1 eligieron como comunidades secundarias a $\{1, k_1, k_2\}$, para algún par k_1, k_2. Tenemos

$$\{|C_1| \leq 2\} = \bigcup_{k_1,k_2 \in \{1,2,...,M\}, k_1 \neq k_2} \{X_i \in \{1, k_1, k_2\}, \forall i \in G^1\}$$

$$P(|C_1| \leq 2) \leq \sum_{k_1 \neq k_2} P(X_i \in \{1, k_1, k_2\}, \forall i \in G^1)$$

$$P(|C_1| \leq 2) \leq \sum_{k_1 \neq k_2} \prod_{i=1}^{|G^1|} P(X_i \in \{1, k_1, k_2\}) = \binom{M}{2} \left(\frac{3}{M}\right)^{\delta} = \frac{3^\delta}{2} \frac{M-1}{M^{\delta-1}} \leq \frac{3^\delta}{M^{\delta-2}}.$$

Escogemos $\delta \geq 6$ y M suficientemente grande tal que $\frac{3^\delta}{M^{\delta-2}} \leq \frac{1}{M^3}$.

$2 \leq j \leq \tau_{\sqrt{M}}$**:** Escribimos $\mathscr{E}_l = \bigcap_{m=1}^{l} E_m$. Queremos acotar la probabilidad del evento E_j, condicionado a $\mathscr{E}_{j-1}$, cuando $j < \tau_{\sqrt{M}}$. Sea $V = \{1, 2, ..., A\}$ el conjunto de todas comunidades vistas hasta el paso $j-1$ y supongamos que en el paso $j-1$ se alcanzaron exactamente μ nuevas comunidades. Como $j < \tau_{\sqrt{M}}$, tenemos $\mu \leq A < \sqrt{M}$.

Ahora bien, condicionalmente en $\mathscr{E}_{j-1}$, se cumple que $e^{j-1} \leq \mu \leq \delta^{j-1}$. Además, $E_j^C = \{0 \leq |C_j| < e\mu\}$ ocurre cuando todos los nodos en el paso $j-1$ escogieron una comunidad secundaria o bien en V o bien en un conjunto K de comunidades no vistas, $K = \{k_1, k_2, ..., k_{3\mu-1}\} \subset V^C$. Es decir, cuando hay $e\mu - 1$ comunidades nuevas[2], como mucho.

Entonces, tenemos

$$E_j^c \big| \big(\{|C_{j-1}| = \mu\} \cap \mathscr{E}_{j-2} \cap \{j < \tau_{\sqrt{M}}\} \big) \iff \exists K = \{k_1, k_2, ..., k_{e\mu-1}\} \subset V^C : \, X_i \in V \cup K, \, \forall i = 1, 2, ..., \delta\mu.$$

$$\begin{aligned} P\big(E_j^c \big| \{|C_{j-1}| = \mu\} \cap \mathscr{E}_{j-2} \cap \{j < \tau_{\sqrt{M}}\}\big) &\leq \sum_{k_1,k_2,...,k_{e\mu-1} \in V^C} P(\forall X_i, \, X_i \in V \cup K) \\ &= \sum_{k_1,k_2,...,k_{e\mu-1} \in V^C} P(X_i \in V \cup K)^{\delta\mu} = \binom{M-A}{e\mu-1} \left(\frac{A+e\mu-1}{M}\right)^{\delta\mu} \\ &\leq \frac{(M-A)^{e\mu-1}}{(e\mu-1)!} \left(\frac{\sqrt{M}}{M}\right)^{\delta\mu} \leq \frac{1}{M^{\mu(\delta/2-e)+1}} \\ &\leq \frac{1}{M^{e^{j-1}(\delta/2-e)+1}}. \end{aligned}$$

[2]En realidad nos referimos al máximo entero de $e\mu - 1$. Hacemos esta identificación para hacer la exposición más clara y no sobrecargar la notación.

Por último, observamos que $\mathscr{E}_{j-1} = \mathscr{E}_{j-2} \cap \bigcup_{\mu=3^{j-1}}^{\delta^{j-1}} \{|C_{j-1}| = \mu\}$. Entonces,

$$P\big(E_j^c | \mathscr{E}_{j-1} \cap \{j < \tau_{\sqrt{M}}\}\big) = \sum_{\mu=e^{j-1}}^{\delta^{j-1}} P\big(E_j^c | \{|C_{j-1}| = \mu\} \cap \mathscr{E}_{j-2} \cap \{j < \tau_{\sqrt{M}}\}\big) \frac{P(\{|C_{j-1}| = \mu\} \cap \mathscr{E}_{j-2} \cap \{j < \tau_{\sqrt{M}}$$

$$\begin{aligned}
&\leq \sum_{\mu=e^{j-1}}^{\delta^{j-1}} P\big(E_j^c | \{|C_{j-1}| = \mu\} \cap \mathscr{E}_{j-2} \cap \{j < \tau_{\sqrt{M}}\}\big) \\
&\leq \frac{\delta^{j-1} - e^{j-1}}{M^{e^{j-1}(\delta/2-e)+1}} \\
&\leq \frac{M^{j-1}}{M^{e^{j-1}(\delta/2-e)+1}} \\
&\leq \frac{1}{M^3}.
\end{aligned}$$

Donde la última desigualdad se cumple a partir de $j = 3$ para $\delta \geq 7$, y desde $j = 2$ si $\delta \geq 8$.

$\tau_{\sqrt{M}} < j \leq \tau_{\frac{M}{2}}$: POR PROBAR

Conseguida esta cota, pasamos a acotar $P(\tau_{\frac{M}{2}} \geq \log(M))$. Queremos que esta probabilidad sea cada vez más pequeña.

Definamos $L = \min\{t \in \mathbb{N} : e^t \geq \frac{M}{2}\} < \log(M) + 1$ y el tiempo de parada acotado $T = \min\{L, \tau_{\frac{M}{2}}\}$. Tenemos que

- el evento $\bigcap_{m=1}^{L} E_m$ implica $|C_L| > e^L \geq \frac{M}{2}$,
- el evento $\{j \leq \tau_{\frac{M}{2}}\} \cap \bigcap_{m=1}^{j} E_m$ implica $|C_{j-1}| < \frac{M}{2}$,
- el evento $\bigcap_{m=1}^{T} E_m$ implica $L \geq \tau_{\frac{M}{2}}$.

Con todo, tenemos

$$P(L \geq \tau_{\frac{M}{2}}) \geq P(\bigcap_{m=1}^{T} E_m) = \prod_{j=1}^{T} P(E_k | \bigcap_{m=1}^{j-1} E_m) \geq (1 - \frac{1}{M^3})^T \geq (1 - \frac{1}{M^3})^{\log(M)+1}$$

Para concluir, usamos el Lema 2 del Apéndice A, con $f(M) = M^3$, $g(M) = \log(M)$. Tenemos, para M suficientemente grande,

$$P(\tau_{\frac{M}{2}} < \log(M) + 1) \leq P(\tau_{\frac{M}{2}} > L) \leq 1 - (1 - \frac{1}{M^3})^{\log(M)+1} \leq \frac{\log(M) + 1}{M^3}.$$

Finalmente, definimos el conjunto de comunidades $S = \bigcup_{j=0}^{\tau_{\frac{M}{2}}} C_j$, que cumple $|S| \geq \frac{M}{2}$, por definición de $\tau_{\frac{M}{2}}$. Ahora bien, $\tau_{\frac{M}{2}} < \log(M) + 1$ implica

$$\forall G, G' \in S : d(G, G') \leq d(G, S) + d(S, G') \leq 2\log(M) + 2.$$

Por lo tanto,

$$P(d(G, G') \leq 2\log(M) + 2) \geq P(\tau_{\frac{M}{2}} \leq \log(M) + 1) \geq 1 - \frac{\log(M) + 1}{M^3}.$$

Basta tomar los nodos $\overline{S} = \bigcup_{G \in S} G$, un $M = n \times \delta$ grande para el cual $\frac{\log(M)+1}{M^3} \leq \frac{1}{M^2}$ y observar que si $u \in G$, $v \in G'$, entonces $d(u,v) \leq 2d(G,G')$, para obtener

$$P(d(u,v) \leq 4\log(n) + 4) \geq 1 - \frac{1}{\delta^2 n^2}, \ \forall u, v \in \overline{S}$$

□

PROPOSICIÓN 4. *La distancia típica en* $(G_n)_{n=1}^{\infty}$ *cumple*

$$P(\limsup_{n\to\infty} \frac{T(G_n)}{O(\log(n))} \leq 1) = 1$$

donde $O(\log(n))$ *es una función lineal de* $\log(n)$.

PROOF. Sean u, v nodos elegidos de manera uniforme en $G \times G \setminus \{(i,i) : i \in G\}$, G^u y G^v sus comunidades de origen, y sean $S_u = \bigcup_{j=1}^{\tau^u_{\frac{M}{2}}} C^u_j$, $S_v = \bigcup_{j=1}^{\tau^v_{\frac{M}{2}}} C^v_j$, donde C^u_j, C^v_j son los conjuntos de todas las comunidades que se pueden alcanzar en, como mucho, j pasos, a partir desde G^u, G^v, respectivamente. Ahora bien, $S_u \cap S_v \neq \emptyset$ pues $|S_u| + |S_v| \geq M$. Entonces, sea $l \in S_u \cap S_v$. Tenemos

$$d(u,v) \leq d(u,l) + d(l,v).$$

De la proposición anterior, tenemos que, si n es suficientemente grande, entonces $P(d(u,v) > 8\log(n) + 8) \leq \frac{2}{\delta^2 n^2}$. En consecuencia,

$$\sum_{n=1}^{\infty} P(\frac{d(u,v)}{O(\log(n))} > 1) < \infty.$$

Donde $O(\log(n)) = 8\log(n) + 8$ es una función lineal de $\log(n)$.
Por el Lema de Borel-Cantelli, tenemos

$$P(\limsup_{n\to\infty} \frac{T(G_n)}{O(\log(n))} \leq 1) = 1.$$

□

Con esto, hemos probado la Afirmación 1, con lo que el modelo es *small-world*. Sin embargo, Cont y Tanimura mejoran su resultado, con la siguiente proposición.

PROPOSICIÓN 5. *El diámetro cumple*

$$P(\limsup_{n\to\infty} \frac{diam(G_n)}{O(\log(n))} \leq 1) = 1.$$

PROOF. POR PROBAR □

Modelo generalizado

Pasamos a discutir el modelo generalizado de Cont y Tanimura. Este modelo tiene dos cambios importantes con respecto al anterior. Primero, cada comunidad ya no es de un tamaño δ fijado con anterioridad, sino que se agrupan los n nodos en comunidades de un tamaño que cercano a $\log(n)$. Segundo, cada nodo puede estar conectado a un número aleatorio de comunidades secundarias.

Construcción. Fijamos $n \in \mathbb{N}$ y construimos el grafo $G_n \in \Gamma_n$ en tres pasos, empezando desde el grafo vacío.

(1) Separamos los n nodos en M comunidades, cada una de tamaño $\log(n) - 1 \leq |G^k| \leq \log(n) + 1$, $k = 1, 2, ..., M$. Para cada comunidad G^k, unimos cada uno de sus nodos con todos los otros pertenecientes a la misma *comunidad de origen*.
(2) Construimos el grafo auxiliar E_n con M nodos. Identificamos cada comunidad G^kcon el nodo k de E_n. En el grafo E_n, unimos cada par de nodos con una probabilidad $p = \frac{r\log(M)}{M}$, con $r > 1$. De esta manera, el grafo E_n es un grafo de Erdos-Renyi. Si dos nodos i, j de E_n están unidos, escogemos aleatoriamente y de manera uniforme un nodo u en $G^i \cup G^j$. Si $u \in G^i$, lo unimos con una arista a cada nodo de G^j; si $u \in G^j$, lo unimos a cada nodo de G^i. Igual que antes, estas serán las comunidades de destino de cada nodo.
(3) Finalmente, si dos nodos de G_n comparten al menos una comunidad de destino, los unimos con una arista.

De esta manera, la construcción es análoga a la anterior, excepto que ahora cada nodo tiene un número aleatorio de comunidades secundarias, entre 0 y $M - 1$, inclusive.

AFIRMACIÓN 2. El modelo generalizado $(G_n)_{n=1}^{\infty}$ es un modelo *small-world.*

Al igual que antes, probaremos la Afirmación 2 en tres partes.

PROPOSICIÓN 6. *El grado promedio esperado es menor que* $(r^2 + r + 2)\log(n)$.

PROOF. Usamos la misma estrategia que antes para acotar la esperanza del número total de todos en G_n.

Dado que cada comunidad G^m es un subgrafo completo, podemos considerar A^m, el subgrafo completo maximal que contiene a G^m. En este caso, tenemos $G_n = \bigcup_{m=1}^{M} A^m \times A^m$. De esta manera, si $\mathcal{E}$ es el conjunto de aristas de G_n, tenemos

$$|\mathcal{E}| \leq \sum_{m=1}^{M} \frac{|A^m|(|A^m| - 1)}{2}.$$

Ahora bien, cada A^m contiene todos los nodos de G^m, donde $|G^m| \leq \log(n)+1$, así como los nodos i que hayan sido conectados a todos los nodos de G^m por ser esta su comunidad secundaria. Hay como máximo $deg_E(m)$ de estos nodos. Entonces, tenemos $|A^m| \leq deg_E(m) + \log(n) + 1$. Por otro lado, sabemos que $deg_E(m) \sim Bin(M-1, \frac{r\log(M)}{M})$, de donde $\mathsf{E}[deg_E(m)] = r\log(M)\frac{M-1}{M} \leq r\log(M)$ y $\mathsf{E}[deg_E(m)^2] = r\log(M)(\frac{M-1}{M})(\frac{M-r\log(M)}{M}) + (r\log(M)\frac{M-1}{M})^2 \leq r\log(M) +$

$(r\log(M))^2$. Reemplazando, se tiene

$$\begin{aligned}
\mathsf{E}[|\mathcal{E}|] &\leq \frac{1}{2}\sum_{m=1}^{M}\mathsf{E}[(deg_E(m)+\log(n)+1)(deg_E(m)+\log(n))] \\
&\leq \frac{1}{2}\sum_{m=1}^{M}\big(\mathsf{E}[deg_E(m)^2]+(\log(n)+1)\mathsf{E}[deg_E(m)]+(\log(n)+1)\log(n)\big) \\
&\leq \frac{M}{2}\big(r\log(M)+(r\log(M))^2+(\log(n)+1)r\log(M)+(\log(n)+1)\log(n)\big) \\
&\leq \frac{M}{2}\big((r^2+r+1)\log(n)^2+(2r+1)\log(n)\big).
\end{aligned}$$

Finalmente, observamos $\frac{M}{n} \leq \frac{1}{\log(n)-1}$, lo que nos deja, para el grado promedio esperado,

$$\begin{aligned}
\mathsf{E}[deg(G_n)] &= \frac{2\mathsf{E}[|\mathcal{E}|]}{n} \leq \frac{(r^2+r+1)\log(n)^2+(2r+1)\log(n)}{\log(n)-1} \\
\mathsf{E}[deg(G_n)] &\leq (r^2+r+2)\log(n),
\end{aligned}$$

donde la última desigualdad cumple para un n suficientemente grande. □

PROPOSICIÓN 7. *La agrupación local de G_n cumple $\overline{c} \geq \frac{1}{r+2}$ casi ciertamente.*

PROOF. Empezaremos acotando por debajo cada c_i. Para ello, utilizaremos de nuevo el lemma 1 del apéndice A. Escribamos $S_i = |\{m : i \in A^m\}|$, donde A^m es, como antes, el subgrafo completo maximal que contiene a G^m. Tenemos $S_i \geq 1$ ya que $i \in A^i$. Además, cada vecino j debe estar en algún conjunto A^m, de donde $V(i) = \underset{i\in A^m}{\cup}(A^m \setminus \{i\})$. Para aplicar el lemma 1, necesitamos encontrar una partición de $V(i)$ en subgrafos completos. Esto no siempre será posible; sin embargo, si quitamos algunas aristas de $V(i)$ (con cuidado de no quitar muchas) podemos particionar $V(i)$ al mismo tiempo de obtener una cota inferior de c_i (pues, al quitar aristas, estamos reduciendo el número de triángulos presentes). Entonces bien, quitemos de $V(i)$ las aristas que unen dos conjuntos diferentes A^u, A^v, para obtener los conjuntos A'^u, A'^v. Ahora $V(i)$ se puede particionar en los subgrafos completos y disjuntos $A'^m \setminus \{i\}, m = 1, ..., M$. Como no hemos cambiado el número de nodos, tenemos $|A'^m| \geq |G^m \setminus \{i\}| \geq \log(n) - 2$. Entonces, por el lemma 1 del Apéndice A, tenemos $c_i \geq \frac{1}{S_i} - \frac{1}{\log(n)-2}$. Como S_i es todavía un número aleatorio, nos fijaremos más bien en la cota

$$\overline{c} = \frac{1}{n}\sum_{i=1}^{n} c_i \geq \min_{S_1,S_2,...,S_n}\big(\frac{1}{n}\sum_{i=1}^{n}\big(\frac{1}{S_i} - \frac{1}{\log(n)-2}\big)\big).$$

Vemos que el mínimo se alcanza cuando cada S_i toma su valor máximo, lo que resultaría en un valor cercano a cero. Como queremos acotar $\overline{c}$ por debajo, nos interesa, entonces, acotar por arriba los valores de S_i. Para ello, estudiaremos los valores posibles de $\sum_{i=1}^{n} S_i$.

Observamos que por cada arista de E_n, se añade (como máximo) una comunidad secundaria a algún nodo i, lo que a su vez aumenta en 1 el valor de S_i. Como, además, $S_i \geq 1, \forall i$, tenemos que el exceso de la suma de los S_i sobre n no será mayor al número de aristas de E_n. Definamos la variabale aleatoria

$X_j, j = 1, ..., K = \frac{M(M-1)}{2}$, con $X_j = 1$ cuando existe la arista j y 0 en otro caso. Entonces,

$$\sum_{i=1}^{n} S_i \leq n + \sum_{j=1}^{K} X_j.$$

Demostraremos que el número de aristas de E_n está acotado casi ciertamente, de donde la suma de los S_i también lo estará, con lo que podremos hallar el mínimo de la expresión de $\bar{c}$. En efecto, escribamos $p = \frac{r \log(M)}{M}$. Tenemos $X_j \sim Ber(p)$ y, por la desigualdad de Bernstein,

$$P\big(|\sum_{j=1}^{K}(X_j - p)| \geq \sqrt{\epsilon K}\big) \leq 2\exp(-\frac{t^2}{1+\alpha/3}),$$

donde $\sigma^2 \leq p$, $t = \frac{\sqrt{\epsilon}}{\sigma}$, $\alpha = \frac{2\sqrt{\epsilon}}{\sigma^2\sqrt{K}} \leq 1$ y $1 + \alpha/3 \leq 2$ para M suficientemente grande y ϵ pequeño.[3] En este caso, se tiene $\frac{t^2}{1+\alpha/3} \geq \sqrt{M}$ y, entonces,

$$P(\sum_{j=1}^{K} X_j < r\, n) \geq P\big(|\sum_{j=1}^{K}(X_j - p)| < \sqrt{\epsilon K}\big) \geq 1 - 2\exp(-\sqrt{M}).$$

Pero el evento de la mano izquierda implica $\sum_{i=1}^{n} S_i \leq (r+1)n$, con lo que podemos pasar a minimizar la cota inferior de $\bar{c}$. Para ello, consideramos los valores posibles de la variable aleatoria $Y = \min_{S_1, S_2, ..., S_n} \big(\frac{1}{n}\sum_{i=1}^{n}(\frac{1}{S_i} - \frac{1}{\log(n)-2})\big)$. Primero, estudiamos la solución al problema de minimización

$$\min_{\sum_i S_i = H} \frac{1}{n}\sum_{i=1}^{n}\frac{1}{S_i},$$

con un $H > 0$ arbitrario. Supongamos que R_i, $i = 1, ...n$ son solución del problema. Afirmamos que $R_i = R_j$, $\forall i, j$. En efecto, si hubieran dos R_i, R_j distintos, podríamos reemplazarlos por $R_i' = R_j' = \frac{R_i + R_j}{2}$ y la suma $\sum_{i=1}^{n} \frac{1}{R_i}$ sería menor. Concluimos que la solución es $R_1 = R_2 = ... = R_n = \frac{H}{n}$ y el valor mínimo es $\frac{n}{H}$. Luego, tenemos,

$$\min_{\sum_i S_i \leq (r+1)n} \frac{1}{n}\sum_{i=1}^{n}\frac{1}{S_i} = \min\{n, \frac{n}{2}, \frac{n}{3}, ..., \frac{n}{(r+1)n}\} = \frac{1}{r+1}.$$

Hemos probado

$$P(\bar{c} \geq \frac{1}{r+2}) \geq P(Y = \frac{1}{r+1} - \frac{1}{\log(n)-2}) \geq P(\sum_i S_i \leq (r+1)n) \geq 1 - 2\exp(-\sqrt{M}),$$

donde la primera desigualdad se cumple para un n suficientemente grande. Finalmente, por el segundo lemma de Borel-Cantelli,

$$P(\limsup_n \bar{c} \geq \frac{1}{r+2}) = 1.$$

□

[3] Ver Apéndice A.

PROPOSICIÓN 8. *Si $r > 2$, entonces $\frac{2\log(M)}{\log(2\log(M))} + 1$ es cota superior de crecimiento de $diam(G_n)$ casi ciertamente.*

PROOF. Para $r > 2$, el grafo E_n es conexo casi ciertamente y $\frac{\log(M)}{\log(r\log(M))}$ es cota superior del diámetro. (Véase Chung y Lu, 2000, Teorema 3). Como cada comunidad de G_n es un subgrafo completo, el resultado está probado. □

OBSERVACIÓN. El comportamiento del diámetro en el grafo de Erdos-Renyi ha sido estudiado en muchos trabajos dferentes en las últimas seis décadas y queda fuera del alcance de esta monografía. Para un estudio detallado de la situación, véase Chung y Lu, 2000. En este artículo se prueba que, en un grafo de Erdos-Renyi con n nodos y probabilidad $p = \frac{r\log(n)}{n}$, si $r > 2$, el diámetro toma hasta tres valores distintos casi ciertamente, todos ellos cercanos a $\frac{\log(n)}{\log(np)}$.

Una generalización más fuerte: construcción inversa

Evidentemente, para demostrar que el modelo generalizado es small-world, Cont y Tanimura solo utilizan ciertas características del grafo auxiliar E_n. En efecto, r es escogido de tal manera de que E_n cumpla

- El grado promedio esperado en E_n es sub-logarítmico en M.
- El número de aristas de E está acotado por arriba con una probabilidad exponencial en $\sqrt{M}$.
- El diámetro de E_n está acotado por $\log(M)$.

Estas restricciones en la distribución del grado, el número de aristas y el diámetro, logran que, al reemplazar los nodos de E_n por grafos completos de cierto tamaño, el grafo G_n cumpla restricciones parecidas sobre el grado y el diámetro, a la vez que se genera una nueva característica: el coeficiente de agrupación es alto.

Para una nueva generalización del modelo, en vez de empezar a construir G_n y tomar un grafo auxiliar E_n, empezaremos la construcción con un E_n arbitrario que cumpla las características deseadas y reemplazaremos sus nodos por subgrafos completos con un número apropiado de nodos. Estudiaremos cómo las características de E_n se heredan en G_n en los casos en que E_n es determinístico o aleatorio. Cabe resaltar que en ambos de estos casos, el resultado será un grafo aleatorio.

Construcción inversa: caso determinístico. Consideremos un grafo arbitrario E, con $|E| = n$ y fijemos un número natural $\delta > 0$. Construimos un grafo asociado, G, análogamente a las construcciones anteriores, en tres pasos.

(1) Por cada nodo $i \in E$, añadimos un subgrafo completo G^i a G con δ nodos. Si $h \in G^i$, esta es la comunidad original de h.
(2) Por cada arista $(i, j) \in E$, escogemos de manera eleatoria y unirme un nodo $h \in G^i \cup G^j$. Si $h \in G^i$, unimos h con todos los nodos de G^j; Si $h \in G^j$, lo unimos con todos los nodos de G^i. De esta manera, asignamos una comunidad secundaria para h.
(3) Unimos cada par de nodos $h, k \in G$ si ellos comparten al menos una comunidad secundaria.

Identificamos cada nodo $i \in E$ con la comunidad $G^i \subset G$; llamamos a E el *hipergrafo* de G; y a G, el *hipografo* de E. Observamos que $|G| = \delta \times n$ y que, si $h \in G^i$, el número S_h de comunidades de h (originales y secundarias) cumple $S_h \leq 1 + \deg_E(i)$. Observamos también que, aún cuando E_n es determinístico, su hipografo

es aleatorio. Además, si tenemos dos modelos de crecimiento de grafos $(E_n)_{n=1}^{\infty}$ y $(G_n)_{n=1}^{\infty}$ en el que cada G_n es hipografo de E_n, para todo n, decimos simplemente que $(E_n)_{n=1}^{\infty}$ es hipergrafo de $(G_n)_{n=1}^{\infty}$.

En este caso determinístico, las características de E_n se transmiten a su hipografo según la siguiente proposición. Recordemos que denotamos con letras caligráficas $(\mathcal{E})$ el conjunto de aristas de un grafo (E).

PROPOSICIÓN 9. *Sea $(E_n)_{n=1}^{\infty}$ un modelo determinístico arbitrario y $(G_n)_{n=1}^{\infty}$ su hipografo. Pongamos $\bar{c}_n$ para el coeficiente de agrupación local promedio de G_n. Entonces,*

a. $diam(G_n) \leq 2\, diam(E_n) + 1$.
b. $|\mathcal{G}_n| = \frac{\delta(\delta-1)}{2} n + \delta|\mathcal{E}_n| + t_n$, *donde t_n es una variable aleatoria que cumple* $t_n \leq \frac{1}{2}\sum_{i\in E} \deg(i)^2 - |\mathcal{E}_n|$.
c. *Si existe una constante $H_n \in \mathbb{N}$ tal que $H_n \geq \frac{|\mathcal{E}_n|}{\delta n}$, entonces $\bar{c}_n \geq \frac{1}{H_n+1} - \frac{1}{\delta-1}$.*

PROOF. La propiedad a) se cumple de manera trivial desde que cada comunidad G_n^i es un subgrafo completo.

Para la b), consideremos las aristas que se añaden a G en cada uno de los pasos de su construcción. En el primer paso, se añaden $\frac{\delta(\delta-1)}{2}$ aristas por cada nodo en E. En el segundo paso, añadimos δ aristas por cada arista en E. Estas suman en total $n\frac{\delta(\delta-1)}{2} + \delta|\mathcal{E}_n|$. Escribamos t_n para las aristas añadidas a G en el tercer paso; llamémosles "terciarias". Añadimos una arista terciaria entre los nodos $h \in G^{m_1}, k \in G^{m_2}$ solo cuando ocurren las siguientes tres cosas al mismo tiempo: 1) $i \in E$ es adyacente a ambos $m_1, m_2 \in E$; 2) G^i es comunidad secundaria de h; y 3) G^i es comunidad secundaria de k. Ahora bien, 2) y 3) solo se pueden cumplir cuando se cumple 1). De ahí que, para la existencia de una arista terciaria, requerimos primero que exista un camino (m_1, i, m_2) en E y, en ese caso, puede o puede no existir una arista terciaria correspondiente en G. De todo ello, concluimos que podemos acotar t_n por el número de caminos de la forma (m_1, i, m_2) en E, para todo i. Para un i dado, existen exactamente $\binom{\deg_E(i)}{2}$ de estos caminos, de donde se deduce la cota

$$t_n \leq \frac{1}{2}\sum_{i\in E} \deg_E(i)(\deg_E(i) - 1).$$

Para la c), por el razonamiento de las pruebas anteriores y teniendo en cuenta $\sum_{h\in G_n} S_h \leq n\delta + |\mathcal{E}_n| \leq (H_n + 1)n\delta$, tenemos

$$\bar{c}_n \geq \min_{S_1,\ldots,S_{n\delta}} \left(\frac{1}{n\delta}\sum_{h\in G_n} \frac{1}{S_h} - \frac{1}{\delta-1}\right) \geq \frac{1}{H_n+1} - \frac{1}{\delta-1}.$$

□

Construcción inversa: caso aleatorio. Sea el grafo aleatorio E, con $|E| = n$. Construimos su hipografo G en tres pasos. El primer y tercer paso son iguales que antes. Para la asignación de comunidades secundarias, consideremos la probabilidad p_{ij} de que los nodos $i, j \in E$ estén conectados por una arista. Así pues, en el segundo paso de la construcción de G, con una probabilidad p_{ij}, escogemos un

nodo de manera uniforme en $G^i \cup G^j$ y lo conectamos a todos los nodos de la otra comunidad.[4]

En este caso, tenemos la siguiente proposición sobre las características del hipografo de un grafo aleatorio.

PROPOSICIÓN 10. *Sea* $(E_n)_{n=1}^{\infty}$ *un modelo aleatorio arbitrario y* $(G_n)_{n=1}^{\infty}$ *su hipografo. Entonces,*

a. $diam(G_n) \leq 2\, diam(E_n) + 1$.
b. $\mathsf{E}[\deg(G_n)] \leq \delta - 1 + \frac{2\delta-1}{\delta}\mathsf{E}[\deg(E_n)] + \frac{1}{\delta n}\mathsf{E}[\sum_{m\in E_n} \deg_{E_n}(m)^2]$.
c. *Sea una constante* $H \in \mathbb{N}$.
   i. *Si* $P(H \geq \frac{|\mathcal{E}_n|}{\delta n}) \to 1$, *entonces* $P(\bar{c}_n \geq \frac{1}{H+1} - \frac{1}{\delta-1}) \to 1$.
   ii. *Si* $P(\limsup_n H \geq \frac{|\mathcal{E}_n|}{\delta n}) = 1$, *entonces* $P(\limsup_n \bar{c}_n \geq \frac{1}{H+1} - \frac{1}{\delta-1}) = 1$.

PROOF. a) es inmediata. Para b), basta definir A_n^m como el subgrafo maximal que cumple $G_n^m \subset A_n^m$ y observar $G_n = \bigcup_{m\in E_n} A_n^m \times A_n^m$ y $|A_n^m| \leq \delta + \deg_{E_n}(m)$, para luego reemplazar en $\frac{n\delta}{2}\mathsf{E}[\deg(G_n)] = \mathsf{E}[|\mathcal{G}_n|] \leq \frac{1}{2}\sum_{m\in E_n} |A_n^m|(|A_n^m| - 1)$.

Para c), aplicamos el razonamiento de las pruebas anteriores para obtener

$$P(\bar{c}_n \geq \frac{1}{H+1} - \frac{1}{\delta-1}) \geq P(\sum_{h\in G_n} S_h \leq (H+1)n\delta) \geq P(H \geq \frac{|\mathcal{E}_n|}{\delta n}).$$

□

[4]Notamos que solo estamos usando las variables $a_{ij} \sim Ber(p_{ij})$, cuya distribución está inducida por E. Dos grafos aleatorios pueden inducir las mismas a_{ij} al mismo tiempo de ser ellos grafos aleatorios distintos. Concluimos que no podemos recuperar el hipergrafo E conociendo solamente su hipografo.

APPENDIX A

Algunos resultados

LEMA 1. *Sea $G \in \Gamma_n$ un grafo arbitrario y sean $i \in G$ un nodo y $G(i) = G \cap (V(i) \times V(i))$ el subgrafo de G que contiene todos los vecinos de i, pero no i, y todas las aristas entre ellos. Supongamos que $G(i)$ se puede particionar en l subgrafos completos disjuntos. Entonces, se cumple*

$$c_i \geq \frac{1}{l} - \frac{1}{h},$$

donde h es el número de nodos en el subgrafo más pequeño de $G(i)$.

PROOF. Consideremos tal partición y pongamos $q_1, q_2, ..., q_l$ para el número de nodos en cada subgrafo completo de $G(i)$. Tenemos $a = \sum_{j=1}^{l} \frac{q_j(q_j-1)}{2}$ aristas en $G(i)$, mientras que el número total de posibles aristas es $t = \frac{(\sum_{j=1}^{l} q_j)(-1+\sum_{j=1}^{l} q_j)}{2}$.

Luego,

$$c_i \geq \frac{a}{t} \geq \frac{\sum_j q_j(q_j-1)}{(\sum_j q_j)^2} \geq \frac{\overbrace{\sum_j q_j^2}^{\mathcal{A}}}{(\sum_j q_j)^2} - \frac{\overbrace{\sum_j q_j}^{\mathcal{B}}}{\sum_j q_j^2}.$$

Tenemos para $\mathcal{A}$, usando $2q_jq_k \leq q_k^2 + q_j^2$,

$$(\sum_j q_j)^2 \leq \sum_j q_j^2 + \frac{1}{2}\sum_{k\neq j}^{l}(q_k^2 + q_j^2) = \sum_j q_j^2 + \frac{1}{2}\sum_{k\neq j}^{l} q_k^2 + \frac{1}{2}\sum_{k\neq j}^{l} q_j^2$$

$$(\sum_j q_j)^2 \leq \sum_j q_j^2 + (l-1)\sum_j q_j^2 = l\sum_j q_j^2.$$

Es decir, $\mathcal{A} \geq \frac{1}{l}$.

Por otro lado, notamos que $\sum_j hq_j \leq \sum_j q_j^2$, es decir, $\mathcal{B} \leq \frac{1}{h}$.

Finalmente, $c_i \geq \frac{1}{l} - \frac{1}{h}$. □

LEMA 2. *Sean $f(n)$, $g(n)$ dos funciones tal que $\lim_{n\to\infty} f(n) = \lim_{n\to\infty} g(n) = \infty$, donde $g(n)$ es no decreciente. Supongamos que $\lim_{n\to\infty} h(n) = \frac{f(n)}{g(n)} = 0$. Entonces, existe un n_0 tal que*

$$\forall n \geq n_0 : (1 - \frac{1}{f(n)})^{g(n)} \geq 1 - \frac{g(n)}{f(n)}.$$

PROOF. Dado, $i \in \mathbb{N}$, tenemos $g(n)^i \geq g(n)$ y además,

$$\frac{-g(n)}{i\, f(n)^i} \geq \frac{-g(n)^i}{i\, f(n)^i}$$

$$g(n)\sum_{i=1}^{\infty} \frac{(-1)^{i+1}}{i} \big(\frac{-1}{f(n)}\big)^i \geq \sum_{i=1}^{\infty} \frac{(-1)^{i+1}}{i} \big(\frac{-g(n)}{f(n)}\big)^i.$$

Para algún n_0, tendremos $f(n) > g(n) > 1$. Luego,

$$g(n)\log(1 - \frac{1}{f(n)}) \geq \log(1 - \frac{g(n)}{f(n)})$$

$$(1 - \frac{1}{f(n)})^{g(n)} \geq 1 - \frac{g(n)}{f(n)}$$

□

LEMA 3. *Derivación de la media y varianza de la distribución multinomial.*

PROOF. POR PROBAR □

LEMA 4. *Borel-Cantelli.*

PROOF. POR PROBAR □

LEMA 5. *Borel-Cantelli 2*

PROOF. POR PROBAR □

LEMA 6. *Desigualdad de Bernstein*

PROOF. POR PROBAR □

APPENDIX B

Cont & Tanimura: Errata

En el desarrollo de este trabajo, se encontraron algunos errores en el artículo de Cont y Tanimura, algunos más graves que otros. Listamos algunos de ellos a continuación.

- $j \leq \tau_{\sqrt{M}} \implies \overline{F_{j-1}} \leq \sqrt{M}$. Haciendo δ tan grande como queramos, podemos encontrar fácilmente contraejemplos para esta afirmación. En el caso $j \leq \tau_{\sqrt{M}}$ solo podemos afirmar $\overline{F_{j-1}} \leq \delta\sqrt{M}$. Sin embargo, esto no es suficiente para lo que necesitamos y debemos replantear el argumento.
- $L = \min\{l : e^l \geq \frac{M}{2}\} \implies L < \log(M)$. Se tiene, en realidad, $L < \log(M)+1$. Afortunadamente, este error no afecta el trabajo subsecuente y podemos sustituir la cota real en el argumento sin problemas.
- Se usan intercambiablemente los eventos $\{|C_j| \geq e|C_{j-1}|\}$ y $\{|C_j| \geq 3|C_{j-1}|\}$, cuando estos son iguales solo en el caso $j = 1$.
- El conjunto C_j contiene las comunidades que se pueden alcanzar en *como máximo* j pasos, no exactamente en j.
- pg. 16: Se dice que el grado total del nodo i es igual a $2k + X_i$, cuando en realidad se tiene $deg_{sw}(i) \leq 2k + X_i$. En la siguiente línea se usa bien, pero está mal escrito.
- En la última página, se demuestra que $\overline{c} \leq \frac{1}{3}$ casi ciertamente, cuando en realidad se quiere lo contrario: $\overline{c} \geq \frac{1}{3}$.