Gráficas aleatorias, redes sociales y el internet

Parte 1

Mariana Olvera-Cravioto

UNC Chapel Hill

molvera@email.unc.edu

23 de septiembre de 2020

Redes sociales y gráficas

- El internet, la red, Facebook, Twitter, LinkedIn, Instagram, WhatsApp, WeChat, Snapchat, Pinterest, Reddit, etc. son todos ejemplos de **redes**.
- En **redes sociales** las conexiones ocurren entre personas.
- Una conexión entre dos personas puede significar muchas cosas dependiendo de la red, e.g., amistad, ligas, relaciones entre seguidores y seguidos, etc.
- También existen muchas redes que no tienen que ver con personas, e.g., el internet, la red de conexiones neuronales en el cerebro, interacciones entre proteínas en biología, citas entre artículos, etc.
- Para analizar redes, muchas veces es conveniente representarlas a través de **gráficas** o **grafos**.

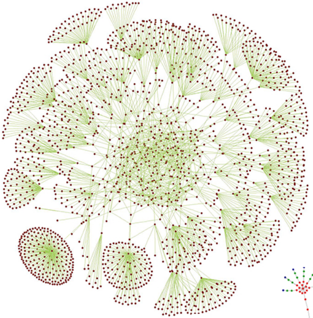
Gráficas

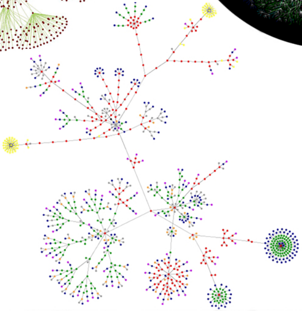
- Una gráfica consiste en un conjunto de **vértices**, V, y un conjunto de **aristas** E.
- Una gráfica puede ser **no dirigida** o **dirigida**.
- En una gráfica no dirigida, la relación entre vértices es simétrica, mientras que en gráficas dirigidas no lo es.
- Llamaremos al conjunto de vértices $V = \{1, 2, \ldots, n\}$, y escribiremos $i \to j$ para indicar que existe una arista (posiblemente no dirigida) del vértice i al vértice j.
- En una gráfica no dirigida el **grado** de un vértice es el número de aristas incidentes a él.
- En una gráfica dirigida, el **grado entrante** es el número de aristas entrantes y el **grado saliente** es el número de aristas salientes.

Ejemplos de gráficas

- **Internet:** los vértices representan computadoras/servidores y las aristas representan conexiones entre ellas/ellos.
- **Web:** los vértices representan páginas web y las aristas son ligas de una página a otra.
- **Facebook:** los vértices representan personas y las aristas son relaciones de amistad.
- **Twitter:** los vértices representan personas y las aristas existen entre seguidores y seguidos.
- **Redes de citas:** los vértices representan artículos scientíficos y las aristas son citas.

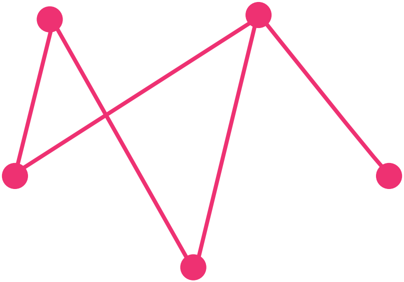
Diferentes tipos de gráficas





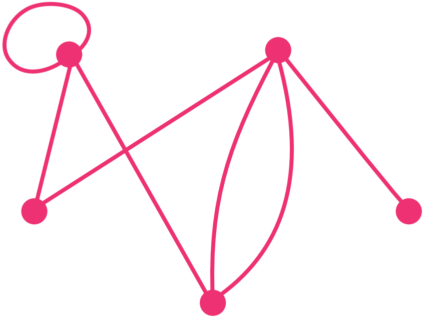
Tipos de gráficas

- **Gráficas simples:** gráficas que no tienen bucles (loops) ni aristas múltiples entre dos vértices.



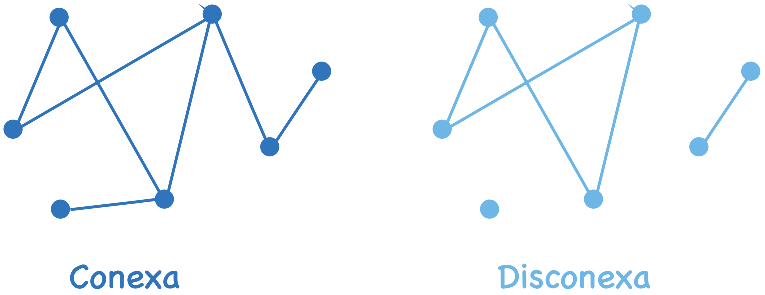
Tipos of gráficas

- **Multigráficas:** gráficas que pueden tener bucles (loops) y aristas múltiples entre dos vértices.



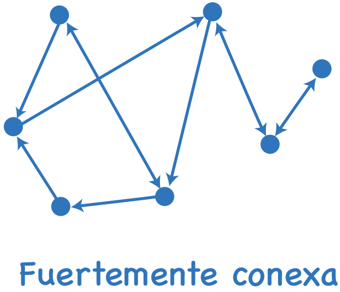
Tipos de gráficas

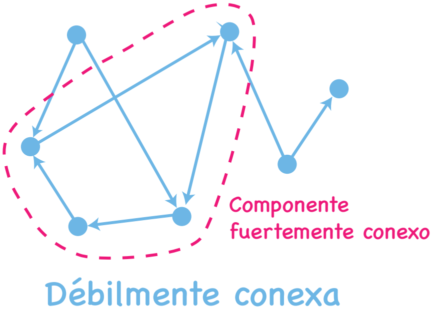
- **Gráficas conexas:** gráficas en las cuales para cada par de vértices existe un camino que los conecta.



Tipos de gráficas

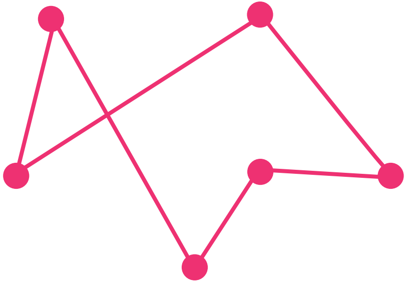
- **Gráficas fuertemente conexas:** gráficas dirigidas en las cuales para cada par de vértices i y j existe un camino dirigido que lleva de i a j y un camino dirigido, posiblemente diferente, que lleva de j a i.





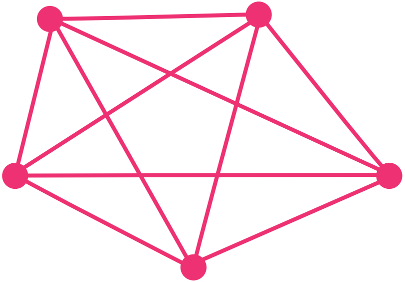
Tipos de gráficas

- **Gráficas regulares:** todos los vértices en la gráfica tienen el mismo grado.



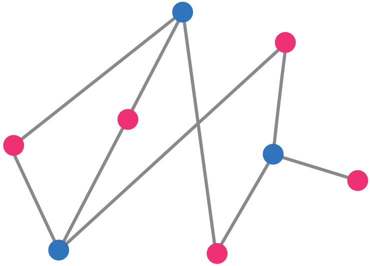
Tipos de gráficas

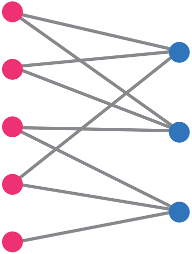
- **Gráficas completas:** existe una arista entre cada par de vértices en la gráfica.



Tipos de gráficas

- **Gráficas bipartitas:** gráficas cuyos vértices pertenecen a dos clases distintas, V_1 y V_2, y en las cuales las aristas ocurren exclusivamente entre un vértice en V_1 y uno en V_2.





Estructuras y propiedades

- Algunas estructuras interesantes cuando estudiamos gráficas:
  - **Ciclos:** caminos que empiezan y terminan con el mismo vértice sin que se repita ningúno.
  - **Cliques:** subgráficas completas.
  - **Distancia etre dos vértices:** la longitud del camino más corto que conecta a dos vértices; en gráficas dirigidas el camino debe ser dirigido.
  - **Componente de un vértice:** el conjunto de vértices que puede ser alcanzado a través de un camino (dirigido) a partir de un vértice dado.

- Algunas propiedades de interés:
  - **Diámetro:** la distancia máxima entre dos vértices en la gráfica.
  - **Componentes:** el tamaño del componente más grande, el segundo más grande, etc.
  - **Longitud de los ciclos:** la longitud típica de los ciclos en la gráfica.
  - **Agrupamiento:** la proporción de triángulos (3-cliques) vs. número de enlaces con 3 vertices y 2 aristas.
  - **Comunidades:** subconjuntos de vértices que tienen más aristas con vertices del subjcnjunto que con otros vértices en la gráfica.

Algunas preguntas interesantes

- ¿Es la gráfica (fuertemente) conexa?
  - Si no lo es, ¿existe un componente (fuertemente) conexo gigante?
  - ¿Cuál es el tamaño de los componentes menores?
- ¿Cuál es el diámetro de la gráfica?
- ¿Cuál es la distancia típica entre los vértices?
- ¿Cuál es la distribución de los grados en la gráfica?
- ¿La gráfica tiene comunidades o agrupamientos?
- ¿Existen vértices que son más "centrales" o "influyentes" en la red?

Redes de un mundo pequeño

- En los 60's, un psicólogo social llamado Stanley Milgram condujo una serie de expermentos para tratar de encontrar la distancia típica entre dos personas escogidas al azar en los Estados Unidos.
- A una persona escogida al azar en el "Midwest" se le daba una carta dirigida a una persona escogida al azar en Boston, sin que las personas se conocieran. La instrucción era tratar de hacer llegar la carta a su destino haciendo uso de una cadena de personas conocidas.
- **Resultado:** en promedio, la cadena de conocidos entre el remitente original y el destinatario de la carta es de longitud 6, un fenómeno conocido como el

  mundo pequeño o los **6 grados de separación**.
- Interesantemente, el fenómeno del mundo pequeño es muy común en redes grandes del mundo real.

Redes libres de escala

- Recordemos que el grado de un vértice $i \in V = \{1, 2, \ldots, n\}$ es una gráfica no dirigida, denotado D_i, es el número de aristas incidentes a él.
- La proporción de vértices que tienen grado $k = 0, 1, 2, \ldots,$ está dada por

$$p(k) = \frac{1}{n} \sum_{i=1}^{n} 1(D_i = k)$$

- Llamamos a $\{p(k) : k \geq 0\}$ la **distribución de los grados**.
- Si la distribución de los grados satisface

$$p(k) \propto k^{-\gamma}$$

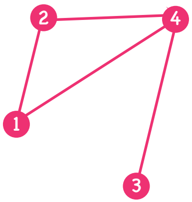
  para alguna $\gamma > 0$ (usualmente $\gamma \in (2, 3)$), decimos que la gráfica es **libre de escala**.
- En una gráfica libre de escala hay vértices que tienen grados muy grandes, aún cuando el grado promedio sea pequeño.

La matriz de adyacencia

- Una manera conveniente de representar a una gráfica es a través de us matriz de adyacencia.
- Para una gráfica $G(V, E)$ con vértices $V = \{1, 2, \ldots, n\}$, su matriz de adyacencia A es la matriz de $n \times n$ que satisface:

$$a_{i,j} = \text{número de aristas del vértice } i \text{ al vértice } j$$

- En una gráfica **simple** tenemos $a_{i,j} \in \{0, 1\}$ para todo par (i, j).
- En una gráfica **no dirigida** la matriz A es simétrica.
- Ejemplo:



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

Contando caminos

- Supongamos que A es la matriz de adyacencia de la gráfica $G(V, E)$.
- Entonces, para toda $k \geq 1$ y todo $i, j \in V$,

$$(A^k)_{i,j}$$

es el **número de caminos de longitud** k del vértice i al vértice j, donde A^k es la matriz A elevada a la potencia k.

Modelos de gráficas aleatorias

- Algunas redes del mundo real son demasiado grandes para ser analizadas de manera exacta.
- Algunas están cambiando constantemente.
- **Idea:** podemos conceptualizar una gráfica específica como un elemento "típico" de una familia más grande.
- Si podemos demostrar que una propiedad es cierta para una familia grande de gráficas, es probable que en particular sea cierta en nuestra gráfica.
- Las **gráficas aleatorias** son modelos matemáticos que nos ayudan a estudiar gráficas grandes del mundo real.
- Ningún modelo puede replicar todas las propiedades de una gráfica específica del mundo real, por lo que nos enfocamos en escoger modelos que comparten propiedades importantes para el problema que queremos analizar.

El límite sobre el tamaño de la gráfica

- Los modelos de gráficas aleatorias consisten en un conjunto de vértices $V_n = \{1, 2, \ldots, n\}$ y un conjunto de reglas basado en eventos aleatorios para decidir si una arista existe o no.
- Su análisis matemático se basa usualmente en tomar el **límite de gráficas crecientes**, $n \to \infty$, en la sucesión de gráficas $\{G(V_n, E_n) : n \geq 1\}$.
- Tomar el límite $n \to \infty$ simplifica los cálculos y nos permite identificar propiedades generales.
- En la práctica, establecer resultados bajo el límite $n \to \infty$ significa que las propiedades encontradas son ciertas para gráficas suficientemente grandes.

Modelos estáticos vs. modelos que evolucionan

- Los modelos de gráficas aleatorias se pueden clasificar en dos categorías: **modelos estáticos** y **modelos que evolucionan**.
- Los modelos estáticos son una "foto instantánea" de una red grande.
- En los modelos estáticos $G(V_n, E_n)$ y $G(V_{n+1}, E_{n+1})$ pueden ser completamente diferentes.
- En los modelos que evolucionan los vértices son añadidos a la gráfica (usualmente uno a la vez) para describir la manera en la que la gráfica va creciendo, por lo tanto, $G(V_n, E_n)$ y $G(V_{n+1}, E_{n+1})$ tienen muchas aristas en común.
- En muchos modelos que evolucionan, las aristas nunca desaparecen, por lo que $G(V_n, E_n)$ es una subgráfica de $G(V_{n+1}, E_{n+1})$.

El modelo Erdős-Rényi

- La gráfica aleatoria más simple es el **modelo Erdős-Rényil**.
- Consideremos una gráfica con vértices $V_n = \{1, 2, \ldots, n\}$.
- Hay $\binom{n}{2}$ posibles aristas en total, y para decidir si cada una pertenece a la gráfica vamos a echar un volado.
- Supongamos que tenemos una moneda que cae 'cara' con probabilidad $p \in (0, 1)$.
- Para cada par de vértices i y j, echamos un volado; si cae cara, dibujamos una arista entre i y j, de lo contrario no hacemos nada.
- De manera equivalente, si A denota la matriz de adyacencia de la gráfica, ponemos

$$a_{i,j} = a_{j,i} = 1(\text{el volado cae cara}), \qquad i \neq j,$$

y escribimos $a_{i,i} = 0$.

Propiedades del modelo Erdős-Rényi

- Este es el modelo de gráfica aleatorias más estudiado que hay.
- Algunas de sus propiedades son:
  - Si $np < 1$ la gráfica consiste sólo de componentes pequeños de tamaño $O(\log n)$.
  - Si $np \to c > 1$ la gráfica contiene un único componente conexo *gigante*, con todos los otros componentes de tamaño $O(\log n)$.
  - Si $np = 1$ el componente más grande es de tamaño $O(n^{2/3})$.
  - Si $p < (1 - \epsilon)n^{-1} \log n$ lo más probable es que la gráfica sea **disconexa**.
  - Si $p > (1 + \epsilon)n^{-1} \log n$ lo más probable es que la gráfica sea **conexa**.
- Cuando la gráfica es conexa, exhibe la propiedad del **mundo pequeño**, con distancias típicas de orden $O(\log n)$.

Distribución de los grados

- La distribución de los grados se calcula usando probabilidades binomiales.
- Fijemos un vértice $i \in V_n$, y notemos que su grado está es:

$$D_i = \sum_{j=1}^{n} \chi_{i,j}, \qquad \chi_{i,j} = 1((i,j) \in E_n)$$

- Observemos que $\chi_{i,j}$ son v.a. Bernoulli con parámetro p.
- Ya que todos los vértices tienen la misma distribución, para todo $i \in V_n$,

$$P(D_i = k) = P(D_1 = k) = P(\text{Binomial}(n,p) = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- Más aún, si $np \to c$ cuando $n \to \infty$, tenemos que

$$\lim_{n\to\infty} P(D_1 = k) = \frac{e^{-c} c^k}{k!}, \qquad k \geq 0,$$

que es una distribución Poisson con media c.... no es **libre de escala**.

Gracias por su atención.