Gráficas aleatorias, redes sociales y el internet

Parte 4

Mariana Olvera-Cravioto

UNC Chapel Hill

molvera@email.unc.edu

14 de octubre de 2021

Google

- ¿Qué imaginas cuando piensas en el nombre Google?

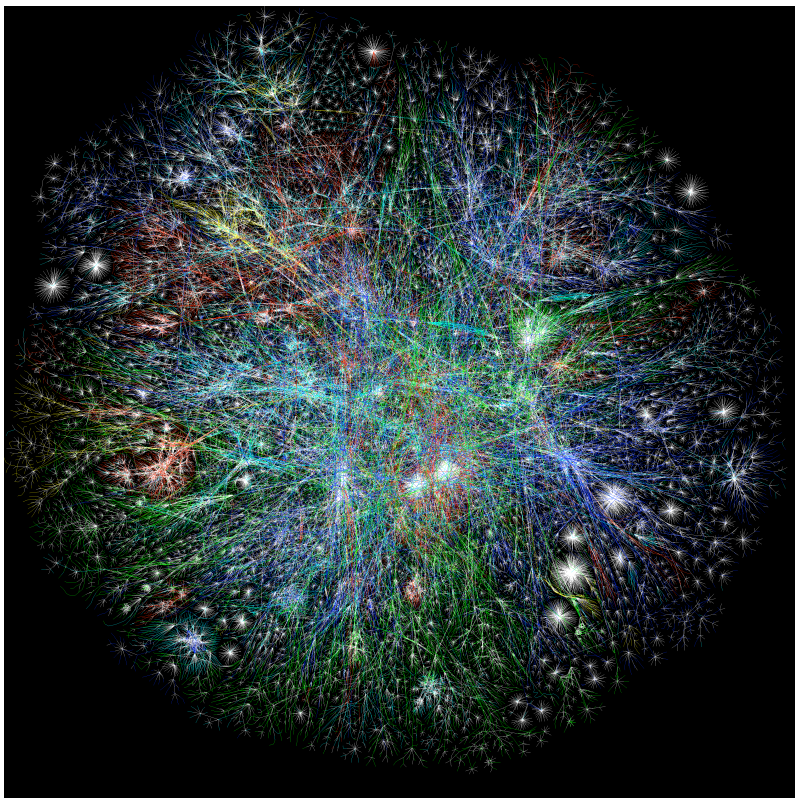
Google

- ¿Qué imaginas cuando piensas en el nombre Google?
- Google, la compañía que conocemos hoy, empezó con un buscador, i.e., un programa de computadora diseñado para ayudarnos a encontrar páginas específicas en la red WWW.

El Internet

- El internet es una red gigantesca de computadoras alrededor del mundo conectadas a través de conexiones físicas o inalámbricas.
- Pensamos en el internet como una gráfica con vértices y aristas:

vértices = computadoras/servidores
aristas = conexiones físicas o inalámbricas entre ellas/ellos



La red WWW (World Wide Web)

- La red WWW es una red "virtual" que conecta páginas a través de ligas.
- Define una gráfica **dirigida** donde:

vértices = páginas web

aristas = ligas dirigidas de una página hacia otra

- La red WWW nació oficialmente en 1991 con la creación del primer *explorador*, un programa con una interface que le permite a los usuarios accesar archivos de diferentes tipos guardados en muchas computadoras diferentes.

La red WWW... cont.

- Con la creación del primer explorador *gratuito*, Mosaic, la red empezó a crecer rápidamente.
- Muchos otros exploradores lo siguieron: Internet Explorer, Netscape Navigator, Mozilla, Firefox, Safari, Google Chrome, etc.
- Conforme la red WWW creció, se volvió cada día más importante poder encontrar infomación rápida y eficientemente.

Buscadores

- Los buscadores empezaron a aparecer desde el principio de la red WWW.
- Su objetivo es encontrar y organizar las grandes cantidades de información contenidas en el Internet.
- Durante los inicios de la red WWW había muchos buscadores que competían por nuestra atención.

¿Qué pasó?

- Muchos de los primeros buscadores estaban basados en *búsquedas simples* dentro de grandes bases de datos:

  palabras clave $\longrightarrow$ lista de páginas que contienen las palabras clave

- Conforme creció la red WWW, las búsquedas simples empezaron a regresar demasiados resultados... se volvió importante ordenarlos.
- Google fue el primer buscador en intentar asignar un **rango universal** a todas las páginas en la WWW para determinar el orden en el cual los resultados serían mostrados.
- Page and Brin, los creadores de Google, propusieron un algoritmo, llamado **PageRank**, capaz de capturar la **relevancia** de cada página con una sola calificación, la cual podía ser usada para ordenar los resultados de cada búsqueda.

¿Qué pasó?

- Muchos de los primeros buscadores estaban basados en *búsquedas simples* dentro de grandes bases de datos:

  palabras clave $\longrightarrow$ lista de páginas que contienen las palabras clave

- Conforme creció la red WWW, las búsquedas simples empezaron a regresar demasiados resultados... se volvió importante ordenarlos.
- Google fue el primer buscador en intentar asignar un **rango universal** a todas las páginas en la WWW para determinar el orden en el cual los resultados serían mostrados.
- Page and Brin, los creadores de Google, propusieron un algoritmo, llamado **PageRank**, capaz de capturar la **relevancia** de cada página con una sola calificación, la cual podía ser usada para ordenar los resultados de cada búsqueda.

¡Su idea funcionó!

Recordatorio sobre gráficas

- Una gráfica consiste en un conjunto de vértices y aristas que los conectan.
- Como hemos visto, las gráficas pueden ser dirigidas o no dirigidas.
- La estructura de una gráfica es importante cuando consideramos cuestiones como:
  - ¿Existe un camino que conecta a cualquier par de vértices?
  - ¿Cuál es la longitud del camino más largo que conecta a dos vértices?
  - ¿Cuál es la distancia típica entre vértices?
  - ¿Cuál es el número promedio de vecinos (entrantes/salientes) que tiene un vértices?
- ¡No todas las gráficas tienen la misma estructura!
- Muchas de las *grandes* gráficas famosas que conoces son bastante especiales...

Modelando redes complejas

- Muchas de las gráficas del mundo real son extraordinariamente grandes, e.g., tienen millones o billones de vértices.
- La gran mayoría son poco densas, i.e.,la razón

$$\frac{\#\text{ aristas}}{\#\text{ vértices}}$$

  no es muy grande.
- Muchas tienen dos propiedades en particular:
  - **Mundo pequeño:** la distancia típica entre dos vértices es relativamente corta comparada con el número de vértices.
  - **Libres de escala:** la proporción de vértices con k vecinos (entrantes/salientes) decae como una potencia de k, e.g.,

$$\frac{\#\text{ vértices con } k \text{ vecinos}}{\#\text{ total de vértices}} \approx Ck^{-\alpha}$$

Otras propiedades de las redes complejas

- Muchas gráficas interesantes son disconexas, pero suelen tener un subjonjunto grande de vértices que es conexo.
- Algunas gráficas tienen muchos agrupamientos (subconjuntos de vértices que tienen más aristas entre ellos que con el resto de la gráfica).
- Algunas gráficas dirigidas exhiben correlación alta entre el número de vecinos entrantes y el número de vecinos salientes de un mismo vértice.

Otras propiedades de las redes complejas

- Muchas gráficas interesantes son disconexas, pero suelen tener un subjonjunto grande de vértices que es conexo.
- Algunas gráficas tienen muchos agrupamientos (subconjuntos de vértices que tienen más aristas entre ellos que con el resto de la gráfica).
- Algunas gráficas dirigidas exhiben correlación alta entre el número de vecinos entrantes y el número de vecinos salientes de un mismo vértice.
- Estas propiedades influencían qué tan rápido se puede esparcer un mensaje a través de la red y/o cuántos vértices lo pueden recibir.
- **También influencían las características de los vértices más "centrales" en la red.**

Relevancia y centralidad

- Regresemos al tema de los buscadores....

Relevancia y centralidad

- Regresemos al tema de los buscadores....
- Intuitivamente, un vértice en una gráfica es **central** si muchos caminos pasan por él.
- La idea detrás del algoritmo PageRank de Google es que las páginas **relevantes** deben ser aquellas que son **centrales** en la red.
- **¿Por qué?**

Relevancia y centralidad

- Regresemos al tema de los buscadores....
- Intuitivamente, un vértice en una gráfica es **central** si muchos caminos pasan por él.
- La idea detrás del algoritmo PageRank de Google es que las páginas **relevantes** deben ser aquellas que son **centrales** en la red.
- **¿Por qué?** Las ligas son creadas por personas, y las personas tienden a crear ligas hacia páginas que contienen información relevante/interesante.
- ¿Cómo encuentra PageRank los vértices centrales?

El algoritmo PageRank

- Sea n el número de vértices en la red WWW.
- El PageRank de la página i, denotado r_i, es un número en $[0, 1]$ que mide su "centralidad" en la red.
- r_i es un rango "universal", i.e., no cambia de una búsqueda a otra, y no tiene nada que ver con el contenido de la página i.
- r_i depende sólo de la topología de la gráfica, i.e., la estructura determinada por las aristas que conectan a los vértices.
- **La relevancia es contagiosa:** Si una página relevante tiene una liga hacia otra página, la hace relevante también, pero si apunta a demasiadas páginas este efecto se reduce.

Calculando el vector PageRank

- Para calcular el vector PageRank $\mathbf{r} = (r_1, \ldots, r_n)$ debemos resolver el sistema de ecuaciones lineales:

$$r_i = \frac{1-\alpha}{n} + \alpha \sum_{j \to i} \frac{r_j}{d_j^+},$$

donde la suma se toma sobre todos los vecinos entrantes de la página i, d_j^+ es el número de vecinos salientes de la página j, y $\alpha \in (0, 1)$ es una constante conocida como el *factor de amortiguación*, usualmente $\alpha = 0.85$.

Calculando el vector PageRank

- Para calcular el vector PageRank $\mathbf{r} = (r_1, \ldots, r_n)$ debemos resolver el sistema de ecuaciones lineales:

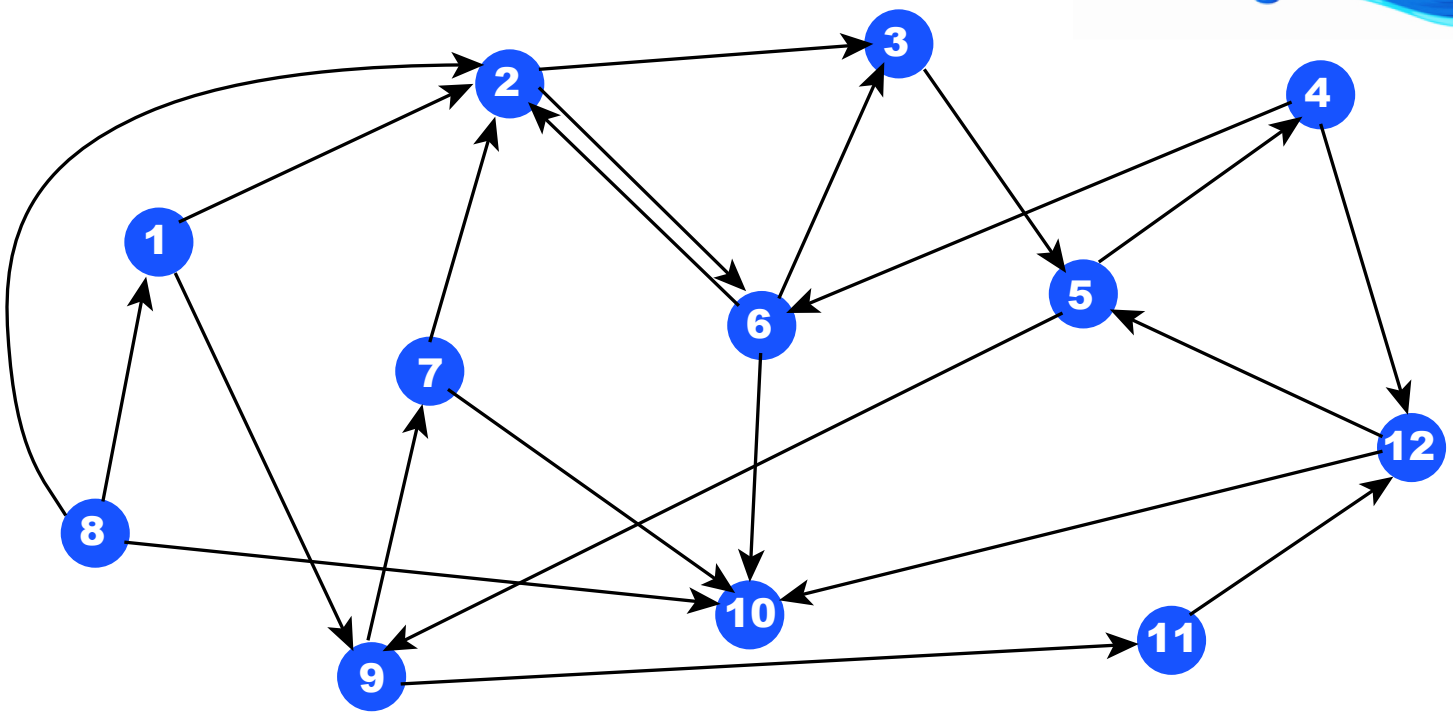
$$r_i = \frac{1-\alpha}{n} + \alpha \sum_{j \to i} \frac{r_j}{d_j^+},$$

  donde la suma se toma sobre todos los vecinos entrantes de la página i, d_j^+ es el número de vecinos salientes de la página j, y $\alpha \in (0, 1)$ es una constante conocida como el *factor de amortiguación*, usualmente $\alpha = 0.85$.

- **¿Por qué funciona ésto?**

La interpretación basada en el surfista aleatorio

- Recordemos que la meta es asignar un rango a todos los vértices de acuerdo a su "centralidad" en la red.
- Imaginemos a un surfista de la red que navega la WWW escogiendo al azar qué ligas seguir.
- Específicamente, cuando nuestro surfista visita la página i, escoge con la misma probabilidad una de las ligas salientes que tiene la página i.
- En el campo de la probabilidad llamamos a este proceso una **caminata aleatoria**.


1
2
3
4
5
6
7
8
9
10
11
12

Caminatas aleatorias en gráficas conexas

- Denotemos $\{X_k : k \geq 0\}$ el proceso estocástico que nos dice qué vértice visita nuestro surfista en el paso k.
- $\{X_k : k \geq 0\}$ es lo que llamamos una cadena de Markov.
- Si la gráfica donde la caminata ocurre es conexa, y dejamos que $k \to \infty$, la proporción de visitas al vértice i converge, i.e.,

$$\lim_{k\to\infty} \frac{\text{Número de visitas al vértice } i \text{ en los primeros } k \text{ pasos}}{k} = \pi_i$$

  existe, y se conoce como la **probabilidad estacionaria** del vértice i.
- La **probabilidad estacionaria** del vértice i tiene la interpretación de ser la proporción de tiempo a la larga que nuestro surfista pasa en el vértice i.
- Si el factor de amortiguación es $\alpha = 0$, tenemos que $r_i = \pi_i$!

Caminatas aleatorias en gráficas disconexas

- El problema con la red WWW es que es una gráfica disconexa.
- En una gráfica disconexa nuestro surfista se puede "atorar".
- Para arreglar este problema, nuestro sufista tiene una moneda que cae *cara* con probabilidad α y *cruz* con probabilidad $1-\alpha$.
- En cada paso, antes de escoger a dónde ir, la/el surfista echa un volado con esa moneda:
  - Si cae *cara* escoge con la misma probabilidad cualquiera de las ligas salientes, si hay alguna, o escoge alguna de las n páginas en la red si no hay ligas.
  - Si cae *cruz* escogemos con la misma probabilidad cualquiera de las n páginas en la WWW.
- La **probabilidad estacionaria** del vértice i es igual a su PageRank, i.e.,

$$\pi_i = r_i!$$

Otras medidas de centralidad en redes

- **Centralidad por grado:** para el vértice i,

$$C_D(i) = D_i = \sum_{j \neq i} a_{ij}$$

  En gráficas dirigidas definimos el grado entrante y saliente por separado.
- **Centralidad por proximidad:** denotemos como $d(i,j)$ la distancia entre los vértices i y j, y definamos

$$C_C(i) = \frac{n-1}{\sum_j d(i,j)}$$

  donde n es el número de vértices en la gráfica.
- **Centralidad de enmedio:** denotemos como g_{jk} el número de caminos que conectan a los vértices j y k, y denotemos como $g_{jk}(i)$ el número de esos caminos que pasan por el vértice i,

$$C_B(i) = \sum_j \sum_{k \neq j} \frac{g_{jk}(i)}{g_{jk}}$$

PageRank hoy en día

- El algoritmo que usa Google hoy en día ha evolucionado mucho desde la versión original de PageRank.
- Cada vértice en la red WWW todavía tiene un rango "universal", sin embargo, la forma en la que es calculado se ha vuelto bastante sofisticada.
- El orden en el que se muestran los resultados de una búsqueda también depende de la computadora del usuario, i.e., los resultados son **personalizados**.
- PageRank Personalizado:

$$r_i = (1-\alpha)q_i + \alpha \sum_{j \to i} \frac{r_j}{d_j^+},$$

donde $\mathbf{q} = (q_1, \ldots, q_n)$ es un vector de probabilidad que determina a dónde ir después de una *cruz*.

PageRank hoy en día... cont.

- La gente ha encontrado maneras de "engañar" a Google para alcanzar rangos altos de manera artificial.
- Google continúa encontrando maneras de cachar a los tramposos.
- Los PageRanks de todas las páginas en la red WWW están siendo constantemente actualizados.
- El PageRank evolucionado que usa Google continúa siendo sumamente eficiente en casi todos los casos, aún tomando en cuenta el impresionante tamaño de la red WWW actual.

Gracias por su atención.