Tarea 2+3

Observaciones:

- Muchas de las preguntas requieren programar, y calcular cosas. Algunas pueden calcularse en Gephi, especialmente si uno baja plugins apropiados, pero lo más probable es que les resulte más cómodo hacer casi todo desde código, usando una buena librería. Recomiendo ampliamente **iGraph**, que tiene para C, R y Python, y calcula todo o casi todo lo que se pide aquí, si saben buscarlo en la (excelente) documentación. Nótese que algunos nombres pueden no ser obvios: coeficientes de clustering se llaman "transitivity", la aleatorización de grafos es "rewire", etcétera. Una vez que se familiaricen con eso verán que las partes que efectivamente tienen que programar no son muchas.
- Algunas preguntas requieren archivos; están en
`https://www.dropbox.com/sh/lxscsvy6opa9pxh/AADp4GTNmdwn6YZ0i9aVGCEaa?dl=0`
- Incluyan *todo* el código generado para las distintas preguntas, indicando además qué librerías usaron. Además incluyan las redes aleatorias generadas (o, en algunos casos en que hay que generar hartas, incluya una de muestra).
- PREGUNTAS PARA MAGISTER. Con frecuencia alguien convalida este curso para el programa de magíster, y lo que hago ahí es encargarles algún trabajo extra. Una alternativa es hacer desde ya el trabajo extra, y de esa forma después el curso estará pre-convalidado (bastará con pedirlo). Por eso hay una serie de preguntas aquí que están marcadas como **[MAG]** (y sus puntajes indicados en *itálicas*). Si no pretenden hacer magíster, ignórenlas con toda confianza.

1. **[MAG]** *[10 pt]* Sean A la matriz de adyacencia de una red no dirigida y **1** el vector columna formado por puros 1. En términos de esas dos cosas, y usando sólo operaciones matriciales (producto, transposición, traza, etc., pero *no* sumatorias) obtenga expresiones para
   - El vector cuyos elementos son los grados de los distintos vértices.
   - La cantidad de aristas en la red.
   - La matriz $N_{i,j}$ que da la cantidad de vecinos que los nodos i y j tienen en común.
   - La cantidad total de triángulos en la red.
   - El vector k_{nn} que contiene, para cada nodo, la suma de los grados de sus vecinos.
2. [12 pt] Sea G un grafo completo no orientado (todos conectados con todos) de 400 nodos. A partir de las 00:00 horas comenzamos a borrarle aristas, escogidas al azar, a razón de 1 por segundo. ¿A qué hora, en promedio, debiera dejar de existir una componente conexa gigante?

   No lo simule: calcúlelo (y explique su cálculo y cualquier supuesto extra que requiera).
3. [14 pt] Sea G el grafo con matriz de adyacencia

   0 1 1 0 0 0
   1 0 1 0 0 0
   1 1 0 1 0 0
   0 0 1 0 1 1

0 0 0 1 0 1
0 0 0 1 1 0

Escriba la matriz laplaciana y determine el valor de Fiedler, junto al vector propio asociado. Grafique la red incluyendo los valores del vector propio, y úselo para determinar la (bastante evidente) partición de la red en dos comunidades.

4. [14 pt] ¿Cuál es la diferencia entre una 2-componente y un 2-core? Explique, y construya un ejemplo de red pequeña que tenga un 2-core pero dos 2-componentes.

5. **[MAG]***[15 pt]* Considere un grafo aleatorio con una gran cantidad de nodos (piénselo como el límite con $n \to \infty$) y con una distribución de grados dada por $P(k) = C \times \alpha^k$, para $k \geq 0$, donde $0 < \alpha < 1$ y C es una constante de normalización.

   (a) Dé una expresión para C en función de α (para que P efectivamente sea una distribución de probabilidad).

   (b) Dé una expresión cerrada para la función generadora de la distribución de grados. ("Cerrada" se refiere a que no sea una sumatoria, sino una expresión directa).

   (c) Determine la condición sobre α que hará que el grafo tenga o no una componente gigante.

6. [15 pt] Baje la red "redchica.gdf".

   (a) Grafíquela, indicando junto a cada nodo su grado de entrada, su betweenness, y su valor de PageRank.

   (b) Haga un ranking de los nodos en función de cada uno de esos tres índices.

   (c) Comente sobre las posibles correlaciones entre esos valores, y sus divergencias. P. ej., ¿hay nodos con mejor PageRank que el esperable por su grado de entrada? ¿O con peor PageRank? ¿Y qué hay de su betweenness? Interprete, a la luz de la red, el por qué de esas discrepancias.

7. [30 pt] Para esta pregunta consideraremos cuatro redes: las de los archivos "gnutella.gdf" y "delfines.gml", y además un par de redes Erdös-Renyi que usted debe generar con la misma cantidad de nodos y la misma densidad de aristas.

   Ambas son redes no orientadas; la primera es un fragmento de la red p2p Gnutella, años ha, mientras que la segunda corresponde a la red de interacciones sociales de una comunidad de delfines en Nueva Zelandia.

   Ataque a piratas y delfines de las siguientes tres maneras:

   - Eliminando nodos al azar
   - Eliminando nodos en orden de grado decreciente
   - Eliminando nodos en orden de betweenness decreciente

   De ser posible, recalcule grados y betweenness en la medida que vaya eliminando nodos (salvo si la eliminación por betweenness se le vuelve demasiado lenta en Gnutella).

   Para cada red y para cada modo de ataque, determine el porcentaje de nodos que hace falta eliminar para que la componente gigante caiga a 1/2 de su tamaño inicial.

   Comente sus resultados.

8. [30 pt *+10 pt*] En esta pregunta trabajará con cuatro redes: la que estudió en la tarea 1, una Erdös- Renyi y una Barabási-Albert que mantengan (aprox.) la cantidad de nodos y la densidad, y una versión aleatorizada de su red (obtenida aleatorizando conexiones, de modo que sólo se preserve la distribución de grados; si su red tiene m aristas, conviene hacer al menos $2m$ "rewires"). Si su red era dirigida, pásela a no dirigida (ignore la orientación).

(a) Determine la estructura de k-cores de la red que estudió en la tarea 1, indicando los tamaños de cada capa obtenida. Hágalo también para las otras tres redes. Comente sobre las similitudes y diferencias.

(b) Evalúe la modularidad[1] de la red que estudió en la tarea 1, y evalúela también para las 3 otras redes. Compare con el valor de la red real y comente sobre la probable presencia (o no) de comunidades en ella.

(c) Determine la "asortatividad" de su red. ¿Sus nodos son selectivos, antiselectivos, o no hay tendencia? Use el coeficiente de Newman y también el gráfico "knn".

(d) **[MAG]** Haga el gráfico de k vs $C(k)$ para cada red; esto es, el gráfico que relaciona grado k con coeficiente de clustering de los nodos de grado k. Compare los gráficos y comente (en particular, acaso hay indicios de modularidad jerárquica en su red).

9. **[MAG]***[30 pt]* Baje "correos.gdf" y repita los análisis de la pregunta anterior, para esta red y para una versión aleatorizada que usted debe generar (no haremos ER y BA en este caso). Por si acaso, se trata de una red no dirigida indicando intercambios de correo electrónico al interior de una universidad catalana.

10. [30 pt] Genere un grafo ER de 80 nodos, con probabilidad de conexión 0.2. Si llamamos a los nodos $\{a_1, a_2, \ldots, a_{80}\}$, entonces definamos ahora una partición de los nodos en dos grupos como $B_1 = \{a_1, \ldots, a_{40}\}$ y $B_2 = \{a_{41}, \ldots, a_{80}\}$, y otra partición, también en dos grupos, como $C_1 = \{a_1, \ldots, a_{20}, a_{41}, \ldots, a_{60}\}$ y $C_2 = \{a_{21}, \ldots, a_{40}, a_{61}, \ldots, a_{80}\}$.

    Convierta ahora las aristas en arcos orientados, escogiendo al azar cuál punta es cuál, excepto en el caso de las aristas entre B_1 y B_2 : esas oriéntelas con probabilidad p desde B_1 hacia B_2 (y con probabilidad $1-p$ en dirección contraria).

    Repita esto para p desde 0 hasta 1, con pasos de 0.1 (o sea, p=0, 0.1, 0.2, 0.3, ..., 1). En cada ocasión evalúe la medida de modularidad Q^d, esto es, la medida modificada por Newman y Leich para aplicarla a redes dirigidas. Evalúela sobre las dos posibles particiones: (B_1, B_2), y (C_1, C_2). Grafique sus resultados (deberían ser dos curvas, cada una con 11 puntos), e interprételos.

    Nota: para evitar mucho ruido en los resultados, puede repetir el experimento entero varias veces —generando diversas redes— y promediar los $Q^d(p)$ resultantes.

11. [30 pt *+5 pt*] Baje "pescado.gdf". Es una red dirigida y con pesos, que representa las exportaciones de pescado de un país a otro en 1998, medidas en millones de US$. O sea: un arco de A a B con peso w indica que el país A le vendió al país B w millones de dólares de pescado.

    (a) Obtenga la matriz de adyacencia de la red y úsela para evaluar la reciprocidad corregida de la red. Si corregimos un typo del ppt (donde olvidé un exponente obvio) y hacemos más explícitos los índices, la reciprocidad corregida es

$$\rho = \frac{\sum_{i \neq j} (a_{i,j} - \overline{a})(a_{j,i} - \overline{a})}{\sum_{i \neq j} (a_{i,j} - \overline{a})^2}$$

    que se puede reescribir como

$$\rho = \frac{\rho_1 - \overline{a}}{1 - \overline{a}}$$

    donde

$$\rho_1 = \frac{1}{m} \sum_{i \neq j} a_{i,j} a_{j,i} \qquad \text{y} \qquad \overline{a} = \frac{1}{n(n-1)} \sum_{i \neq j} = \frac{m}{n(n-1)}$$

[1] Explicite qué usará como "modularidad". Lo más estándar es usar el Q obtenido tras aplicar el algoritmo glotón de Newman; en iGraph puede aplicar el algoritmo glotón (fastgreedy) y luego la función que calcula modularidad. Alternativamente, puede calcular modularidad en Gephi, pero ahí el valor que le darán será el Q obtenido por el algoritmo de Lovaina. Lo importante es que use el mismo método para todas las redes.

para un grafo con m arcos y n nodos. Comente su resultado: ¿es recíproca, antirrecíproca o ninguna de las anteriores?

(b) **[MAG]** Extienda la noción de reciprocidad al caso con peso. Es decir, invente un ρ^w. Debiera ser más o menos directo a partir de la primera expresión que recordamos arriba; no necesariamente va a tener una fórmula chica como la segunda expresión que salió en ese caso. Evalúe su ρ^w y compárelo con la versión sin peso; comente.

(c) Grafique las distribuciones $P(s^{in})$ y $P(s^{out})$, donde s^{in} (resp. s^{out}) de un nodo se define como la suma de los pesos de los arcos que entran a él (resp., salen de él), ¿Corresponden a algún tipo de distribución conocida? De ser así, estime sus parámetros.

(d) Plotee (s^{in}, s^{out}), para el conjunto de nodos. ¿Se ve algún tipo de relación entre esas 2 variables?

Ahora convierta su red en una red no dirigida. Para eso, pondremos una arista entre par de nodos que estaban conectados por uno o dos arcos, y le asociaremos como peso la suma de los pesos de esos arcos. En iGraph de R se puede hacer usando as.undirected(g, mode="collapse", edge.attr.comb=list(weight="sum")).

(e) Evalúe $P(k)$, $P(s)$ y $P(w)$ para esta nueva versión de la red. Nuevamente, si se parece a alguna distribución conocida, estímele los parámetros (e idealmente superponga en el gráfico la versión estimada).

(f) Plotee el grado vs la fuerza de los nodos. ¿Es lineal la relación? Comente.

(g) Evalúe el coeficiente de clustering de la red, con y sin pesos, y grafique también su dependencia respecto al grado (es decir, estudie $C(k)$ y $C^w(k)$). Comente sus resultados.

12. [25 pt] Baje "epinion.net". Se trata de relaciones de confianza o desconfianza (indicado por el atributo de las aristas, que puede ser +1 o -1) entre usuarios de epinion.com, un sitio en que la gente reseña productos. Si bien la red es dirigida, la tomaremos como no dirigida.

Nos interesa estudiar los triángulos que pueden presentarse, y que podemos clasificar según la cantidad de signos + presente, que puede ser 0, 1, 2 ó 3 (correspondientes a un triángulo en que cada uno desconfía de los otros 2 partícipes, uno en que existe confianza entre dos personas pero ninguno confía en un tercero, uno en que alguien tiene confianza en —y de— dos personas, pero éstas no confían entre sí, y uno en que cada uno confía en los otros 2).

(a) Determine la cantidad total de triángulos de cada tipo presentes en la red: t_0, t_1, t_2, t_3 .

(b) Determine la cantidad total de aristas con signo + (llamémosle m_+) y la cantidad total de aristas (m). Por lo tanto, la probabilidad de que una arista tomada al azar tenga signo + es $p = m_+/m$, y la probabilidad de que tenga signo - es $1 - p$.

(c) ¿Coincide m_+ con $t_1 + 2t_2 + 3t_3$? Explique por qué debiera o no coincidir.

(d) Calcule la probabilidad de que al asignarle pesos al azar (con probabilidades p y $1 - p$ de + y -, resp.) a las aristas de un triángulo, el resultado sea un triángulo de tipo 0, 1, 2 ó 3.

(e) ¿Cómo se compara el resultado en (d) con las proporciones observadas en (a)? Comente (puede especular libremente; después de todo, el curso no es de sociología).

13. **[MAG]***[30 pt]*. Considere el siguiente modelo de contagio tipo SIR:

- En el instante 0 habrá un nodo infectado (estado I), escogido al azar, y los demás estarán susceptibles.
- En cada vuelta, cada nodo susceptible que tenga un vecino (nodo conectado) infectado tendrá una probabilidad 0.2 de contagiarse y pasar a infectado.
- Un nodo que ha estado infectado durante 10 iteraciones pasa al estado R, que ya no es contagioso ni susceptible (y se queda R para siempre).

Sean G_1, G_2 y G_3 tres grafos de 10.000 nodos cada uno. G_1 serán grafos aleatorios de tipo Erdös-Renyi y Barabási-Albert respectivamente, con un grado promedio igual a 4, mientras que G_3 será una malla de 100×100, donde cada nodo (i, j) estará conectado a sus vecinos $(i+1, j)$, $(i-1, j)$, $(i, j+1)$, $(i, j-1)$, con condición de borde periodica (es decir, el vecino superior de $(i, 100)$ es $(i, 1)$, etc.).

Disponemos de 100 vacunas, que inmunizan a los nodos por completo (para efectos prácticos, corresponde a poner a esos nodos inmediatamente en estado R). Consideramos cuatro estrategias:

- E_0: no hacer nada.
- E_1: escoger 1000 nodos al azar y vacunarlos.
- E_2: escoger a los 1000 nodos de mayor grado y vacunarlos.
- E_3: escoger un nodo, vacunarlo a él y a todos sus vecinos, y luego repetir eso hasta que se acaben las vacunas.

Simule la evolución de la enfermedad en cada uno de los grafos, para cada una de las posibles estrategias. Muestre con un gráfico la evolución de la enfermedad en el tiempo (en términos de la cantidad de nodos infectados, hasta un tiempo dado). Determine la duración y cantidad total de infectados de la (posible) epidemia, promediando sobre un conjunto de simulaciones (suficientes para obtener valor promedio representativo). A la luz de los resultados, evalúe la conveniencia de las distintas estrategias de acuerdo al tipo de grafo.