Introducción a las Redes Complejas con Aplicaciones a la Biología

Práctico computacional 3

Algoritmos de detección y medidas de evaluación de comunidades

Integrantes
Yuditsabet Burgos
Alan Givré
Lucía Pedraza

Introducción

La partición en clusters o comunidades en una red está dada por cantidades de nodos que comparten similaridades topológicas entre sí dentro de un mismo grupo que respecto a nodos de otros grupos.

Varios métodos han sido desarrollados basados en heurísticas de agrupamiento jerárquico. En estos casos primero se construye la matriz de similaridad de vértices a partir de la matriz de adyacencia y luego, iterativamente se van identificando grupos de nodos con alta similaridad por medio de estrategias aglomerativas o divisivas. De este ordenamiento resultan estructuras similares a un árbol jerárquico o dendograma desde las cuales es posible definir las particiones en comunidades buscada.

Para extrapolar dicha metodología a grafos, es necesario suponer dos hipótesis. Primero, que exista una estructura en comunidades embebida en la red y segundo, establecer como criterio que una comunidad es un subgrafo conexo y localmente denso; donde un nodo puede alcanzar cualquier otro nodo dentro de la misma comunidad y entre ellos existe alta probabilidad de conexión.

En el presente trabajo serán aplicados los algoritmos Edge betweenness, Fast greedy, Louvain e Infomap para encontrar y caracterizar comunidades dentro de una red de delfines mixta que presenta un total de 62 nodos y 152 enlaces. A continuación se explican brevemente los algoritmos antes mencionados.

Edge betweenness: Parte de un cluster gigante y va particionando en grupos a partir de la remoción de enlaces de más alto betweenness. Se repite iterativamente recalculando la centralidad de enlaces.

Fast greedy: Asigna nodos a comunidades; estima el incremento de modularidad que resulta de la combinación de dos comunidades que estén conectadas al menos por un enlace y las une si ese incremento es máximo. Luego guarda esa comunidad y su modularidad iterativamente hasta obtener una única comunidad. Finalmente elige la partición de máxima modularidad.

Louvain: Busca optimizar la modularidad en pasadas de dos pasos. En el paso uno, trata de unir un nodo con sus vecinos y recalcula el incremento de modularidad eligiendo el máximo siempre que sea positiva. Hace esto para cada nodo. Luego arma una nueva red donde cada nodo es una comunidad

encontrada en el paso anterior y se generan autoenlaces dentro de las comunidades. Se repite el paso 1 hasta no obtener más incrementos en la modularidad.

Infomap: Se basa en caminatas aleatorias como indicador de estructura de la red. Map Equation: Dada una partición de la red en comunidades, la map equation es la manera de cuantificar que tan eficiente es la descripción de una caminata aleatoria sobre la red. Esta descripción de cada nodo consta en un término que señala la comunidad a la que pertenece un nodo, y otro término que señala la identidad del nodo al interior de la comunidad. Se minimiza la Map Equation frente al conjunto de todas las posibles particiones de la red en comunidades

Resultados

Encuentre la partición en clusters de esta red utilizando la metodología Louvain, infomap, Fast_greedy y Edge_betweenness. Visualice los resultados gráficamente.

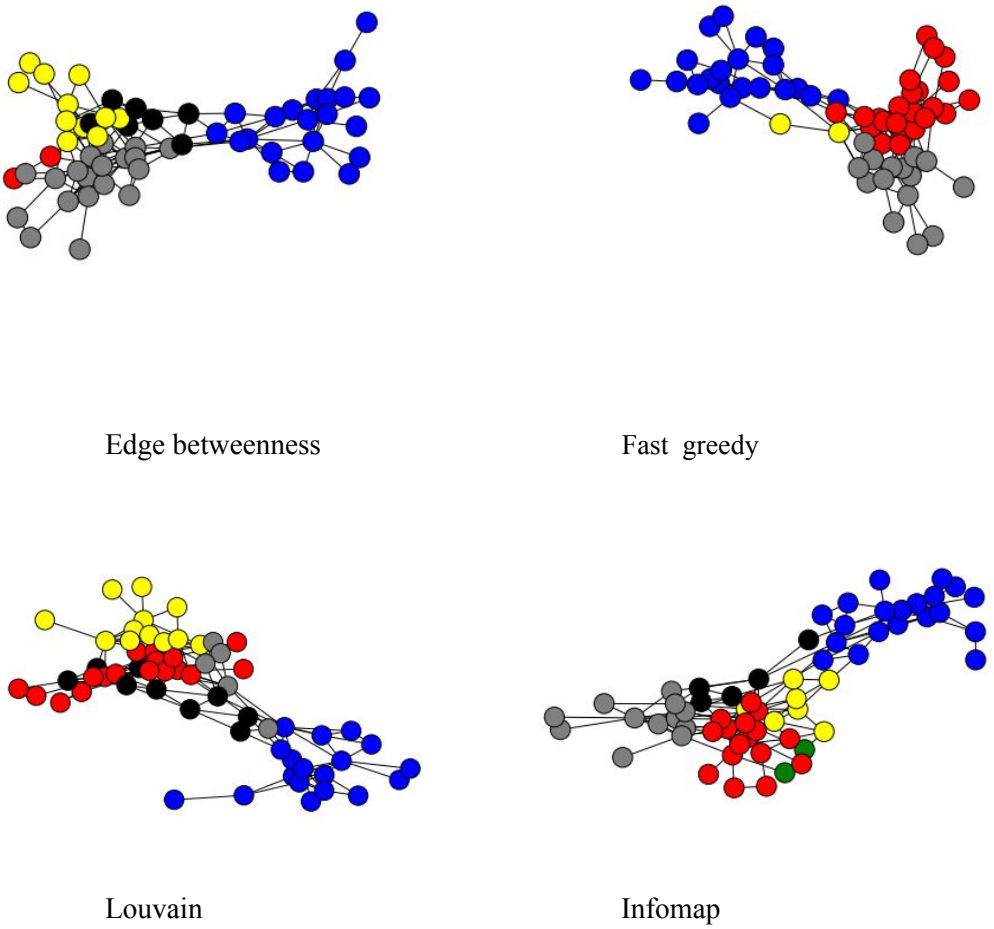

Figura 1. Partición en clusters

El gráfico muestra los clusters encontrados por cada uno de los algoritmos utilizados. Edge betweennes y Louvain detectaron 5 clusters, mientras que Fast greedy solo encontró 4 clusters e Infomap 6 clusters.

Caracterice las particiones obtenidas en términos de modularidad y Silhouettes de cada partición. Compare con valores esperados en redes recableadas y establezca si tiene derecho a llamar modular a esta red

Como medidas internas para evaluar la separación en comunidades de cada partición se utilizaron las nociones de modularidad y Silhouette. En el primer caso se evalúa si la cantidad de enlaces entre vértices del mismo tipo son más o menos que los esperados por azar. En el segundo se evalúa la distancia media a los nodos de un mismo cluster menos la distancia mínima a nodos de otro cluster dividido el máximo entre las dos distancias (y se calcula el valor promedio entre los nodos). La Silhouette total del grafo se calcula como el promedio de todas la Silhouette de cada nodo.

En la tabla 1 se muestra el resultado de las evaluaciones. Los valores encontrados para la modularidad son mayores que cero para todas las particiones, esto indica que la cantidad de enlaces entre vértices dentro de un mismo cluster, es mayor que lo esperado por azar. Además, como es un valor que oscila entre 0 y 1, cuanto más cercano a 1 más modular es la red. Para la categorización considerando el género de los delfines se obtiene el valor de modularidad más alto de la red. Finalmente se puede concluir que podemos considerar la red de delfines una red modular.

Los valores de Silhouettes que encontramos son considerablemente dudosos para esta red; dado que la distancia media entre los nodos dentro del mismo cluster es muy similar a la distancia mínima de un nodo dentro del cluster a un nodo de otro cluster. Se comprobó que el programa media la Silhouette correctamente en una red más simple pero aun así, no nos convenció el resultado, dado que se sospecha que puede llegar a haber un problema de programación. En el código ajunto puede chequerase esta información.

Tabla 1. Medidas internas de evaluación por particiones. Modularidad y Silhouettes

	Modularidad	**Silhouettes**
Edge betweenness	0.540	-0.005
Fast greedy	0.419	0.014
Louvain	0.389	0.029
Infomap	0.385	0.009
Géneros	0.727	0.040

No realizaronparticiones

El tema es que hubo dos

Caracterice cuantitativamente el acuerdo entre las particiones obtenidas utilizando de los observables vistos en clase.

Se consideró el cálculo de la información mutua entre cada par de particiones. Para esto s la posición de cada nodo en su correspondiente cluster para cada una de las particiones. Es

resultado una cardinalidad de la intersección de cada par de clusters, y esto se corresponde con la probabilidad de encontrar un nodo en esa dada intersección de clusters.

Si se llama H_{AB} a la entropía de la intersección, y H_A y H_B a la entropía de cada partición por separado, la información mutua normalizada para el par de particiones es $2(H_A + H_B - H_{AB}) / H_A + H_B$, también conocida como entropía de Shanon.

Tabla 2. Medida externa de evaluación por particiones. Información mutua

Edge betwenness	0.678	0.769	0.782	0.057
0.678	**Fast greedy**	0.812	0.911	0.052
0.769	0.812	**Louvain**	0.909	0.073
0.782	0.911	0.909	**Infomap**	0.097
0.057	0.052	0.073	0.097	**Géneros**

Se puede observar en la tabla 2 que las particiones que cada medida usa son muy similares, con la partición dada por la medida edge beteweenness menos similar al resto. También se compara la partición dada por el label de género, y se puede observar que es una partición mucho menos similar al resto, que las otras entre sí.

Analice cuantitativamente la relación entre el género de los delfines y la estructura de comunidades del grupo. Puede utilizar para ello, por ejemplo, tests de sobre-representación y/o sub-representacion. Qué hipótesis puede aventurar sobre propiedades comportamentales de este grupo de delfines a partir de lo encontrado?

Se busca conocer cómo la clasificación en sexos condiciona la interacción entre nodos, y por lo tanto la formación de clusters. De esta manera queremos calcular la cantidad de individuos de cada género en los clusters y compararla con la cantidad que habría si se hiciera una asignación azarosa. Para el test de sobre-representación y sub-representación, se consideraron los atributos de género, obtenidos por información externa. Se cuantificó la cantidad de delfines con cada género para cada cluster de una de las particiones (infomap), asumiendo (teniendo en cuenta el ejercicio anterior) que el resultado sería similar más allá del método elegido. Luego, se comparó con la cantidad que se obtendría a partir de una redistribución azarosa de los atributos. Se realizaron 1000 distribuciones azarosas que fueron representadas a través de un histograma para cada atributo y para cada cluster (Fig. 2).

En los gráficos podemos ver que los delfines de género masculino están sobre-representados en los clusters 1 (azul) y 3 (gris) Fig. 2a y Fig. 2c respectivamente, mientras que los de género femenino está sobre-representado en el cluster 2 (rojo) Fig. 2e.

A partir de este resultado podemos decir que la agrupación de delfines en cluster no está influenciada por el género; debido a que el número de delfines de ambos sexos por cluster es muy similar entre sí, si no considerasemos la sobre-representación.

Que significa "si no consideramos la sobre-representación"?

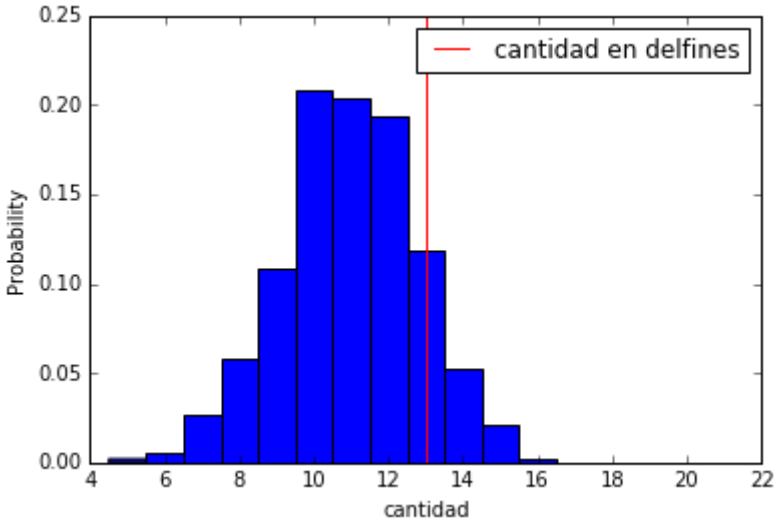

Figura 2.a Cantidad de machos en el cluster 1 (azul) en comparación con el azar.

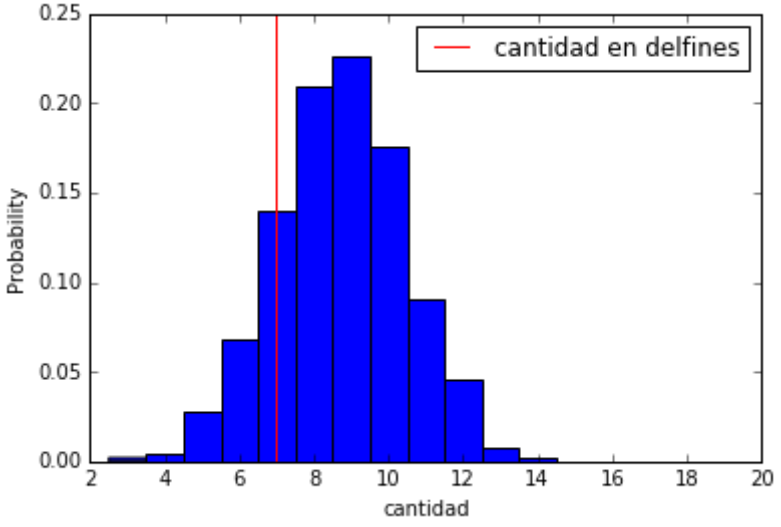

Figura 2.b Cantidad de machos en el cluster 2 (rojo) en comparación con el azar.

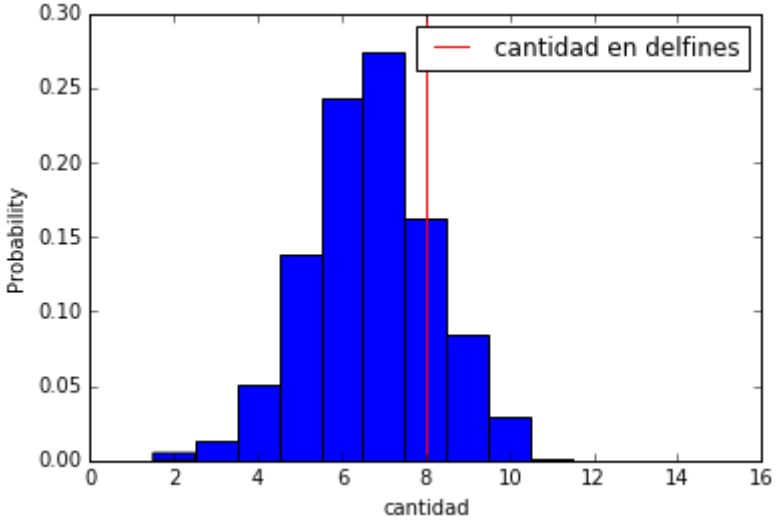

Figura 2.c Cantidad de machos en el cluster 3 (gris) en comparación con el azar.

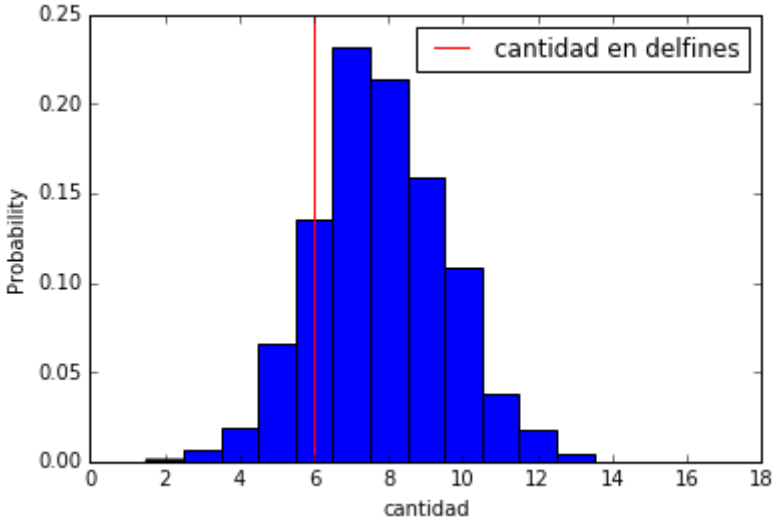

Figura 2.d Cantidad de hembras en el cluster 1 (azul) en comparación con el azar.

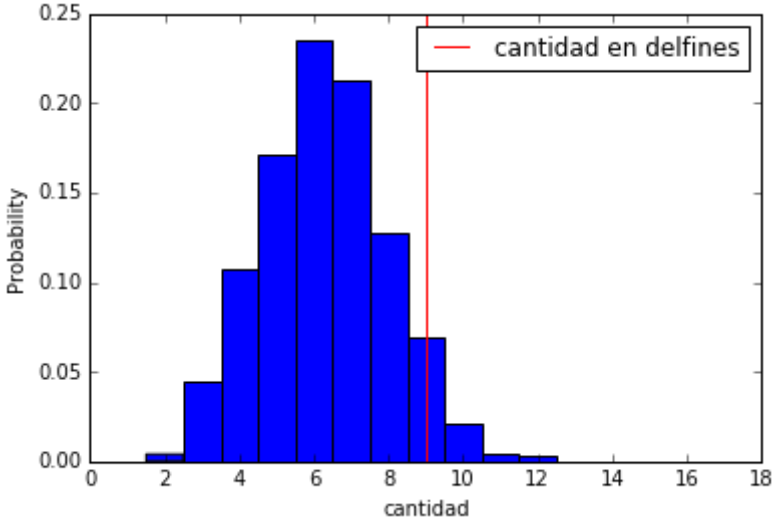

Figura 2.e Cantidad de hembras en el cluster 2 (rojo) en comparación con el azar.

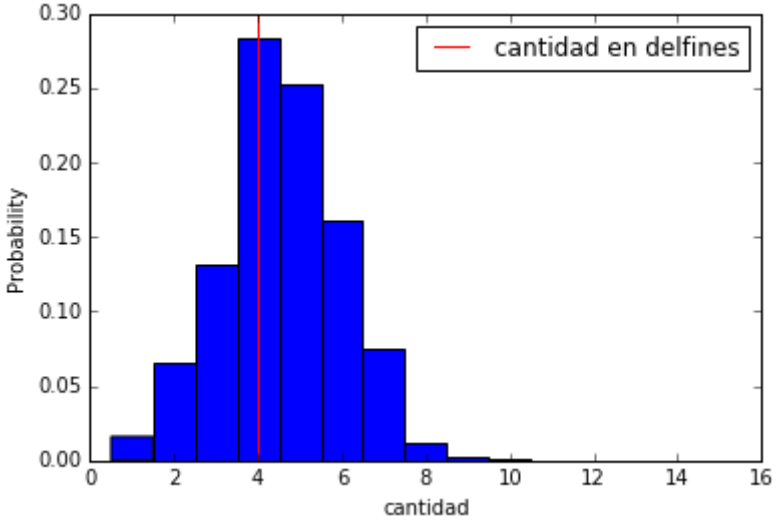

Figura 2.c Cantidad de hembras en el cluster 3 (gris) en comparación con el azar.