Trabajo computacional 3

Di Filippo Juan, Catoni Josefina, Yalovetzky Romina

Octubre 2018

Resumen

En el presente trabajo se estudió una red social, compuesta por las relaciones entre 62 delfines de Nueva Zelanda, con el objetivo de analizar en profundidad la partición en clusters de la misma. Para efectuar dicha partición se trabajó con una base de criterios estándar: Infomap, Fastgreedy, Louvain y Edge-Betweenness. Además, se realizó un modelo nulo a partir de hacer estadística de la red recableada, manteniendo la distribución de grado constante, y se contrastó la red original con dicho modelo en base a dos parámetros: la modularidad y el coeficiente de Silhouette. En segundo lugar se caracterizó el acuerdo entre las particiones obtenidas por los distintos criterios, calculando y comparando la cantidad de nodos que hay en cada comunidad por un lado y utilizando la matriz de probabilidad conjunta por otro. Finalmente, se estudió la relación entre el género de los delfines y la estructura de comunidades del grupo. Para ello, se determinó una hipotesis nula con la cual comparar las comunidades y la fracción de delfines del mismo género en ellas.

1. Criterios de partición

La red de delfines que se desea caracterizar se muestra en la figura 1. Esta red consta de 62 delfines (24 hembras, 34 machos y 4 sin género asignado) que interactúan entre sí.

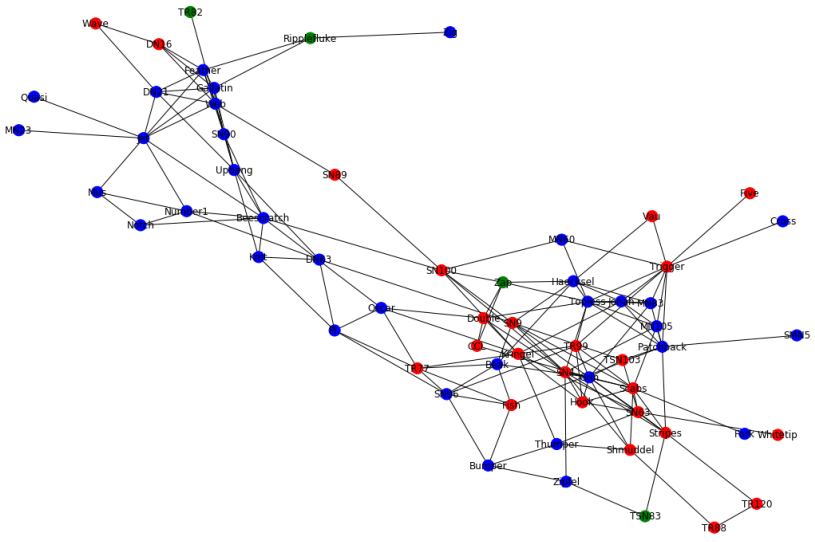

Figura 1: Red de delfines de Nueva Zelanda. El color azul corresponde a los de género masculino, el rojo al femenino, y el verde a los delfines sin género asignado.

En primer lugar, se encontraron las particiones en clusters de la red, utilizando distintas metodologías: *Infomap* [1], *Fast-Greedy* [2], *Louvain* [3] y *Edge Betweenness*.

Para el caso particular de edge betweenness, el algoritmo consiste en ir removiendo los enlaces progresivamente. Lo hicimos hasta que la red queda con un único enlace aplicando el algoritmo de

Girvan Newman[4]. En cada iteración de quitado de enlaces calculamos la modularidad. Entonces, podemos ver en Fig. 2 para qué cantidad de comunidades (y cómo están conformadas) ésta magnitud se maximiza. Nos quedamos con ésta distribución en comunidades según éste criterio.

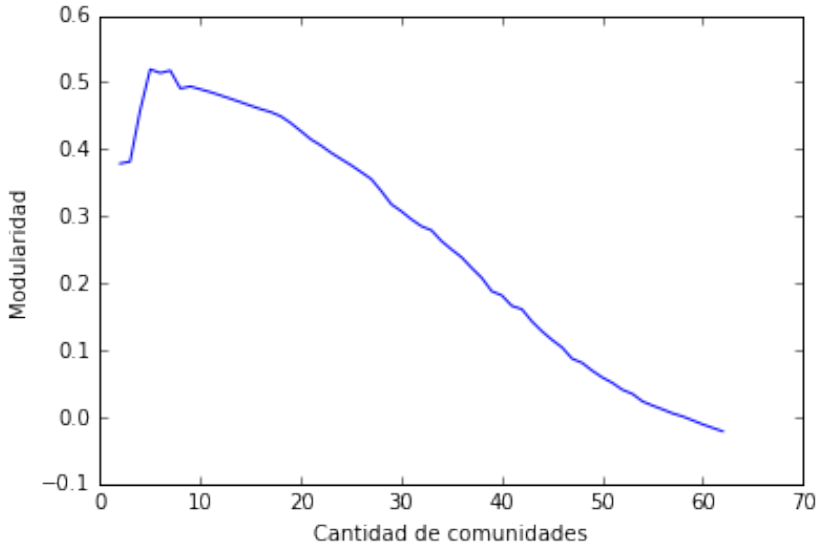

Figura 2: Modularidad de la red en función de la cantidad de comunidades en las que se separe al grupo de delfines por método de Edge Betweenness

En la figura 3 se muestran las particiones obtenidas con los criterios mencionados y se puede apreciar a simple vista que la separación en comunidades obtenidas de la misma red varía con cada uno de ellos. Uno de los objetivos de los próximos apartados será caracterizar cuantitativamente esta variación. Cabe destacar, entre las diferencias observables a partir de esta figura, que todas los criterios parten la red en 5 comunidades, salvo fast-greedy, que lo hace en 4.

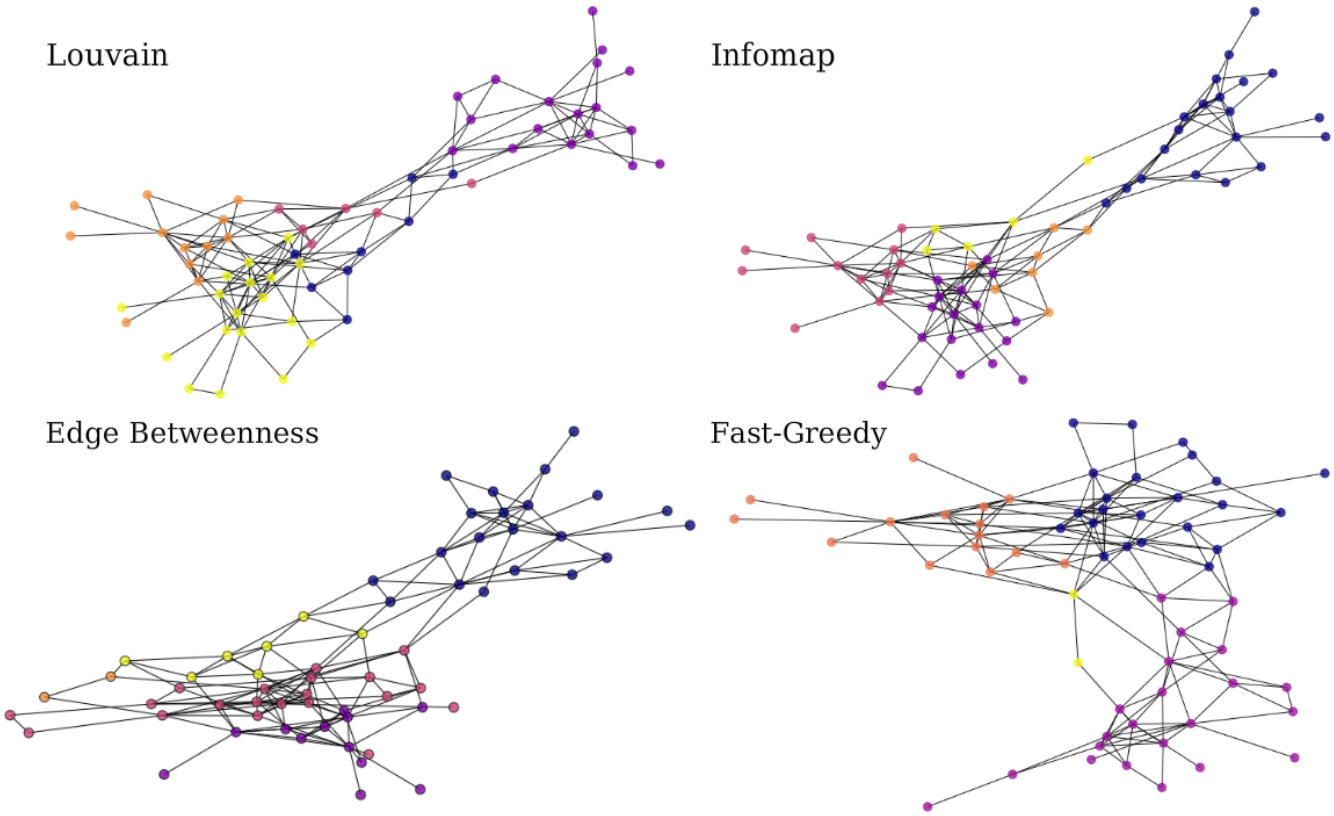

Figura 3: Comunidades obtenidas usando distintas metodologías de partición para la red de delfines de Nueva Zelanda.

1.1. Comparación de los criterios

Para caracterizar la distribución en comunidades de la red adoptamos dos parámetros: la modularidad y el coeficiente de Silhouette medio. Por un lado, cuánto más grande sea la modularidad, mejor resultará la distribución en comunidades. Por otro lado, el coeficiente de Silhouette indica la relación existente entre la distancia promedio entre un nodo y los nodos de su comunidad y la distancia promedio mínima entre un nodo y nodos de comunidades distintas. Este coeficiente se calcula para cada nodo de la red, pero en valor medio da una noción de cuan buena es la separación en comunidades. En el cuadro 1 se pueden observar los valores de modularidad y coeficiente de Silhouette medio obtenidos para los criterios estudiados.

Criterio	Modularidad	Silhouette
Infomap	0.528	0.265
Fast Greedy	0.495	0.130
Louvain	0.519	0.261
Edge Betweenness	0.519	0.285

Cuadro 1: Valores de modularidad para distintos criterios de clustering

Puestos que estos valores son absolutos se procedió a generar un modelo nulo con el cual se puedan comparar dichos parámetros. Para ello se generaron 1000 redes, producto de recablear la red original manteniendo la distribución de grado, y para cada una de ellas se recalculó la distribución de comunidades y los coeficientes pertinentes. En las figuras 4 y 5 se muestran los histogramas obtenidos contrastados con los parámetros calculados en la red real.

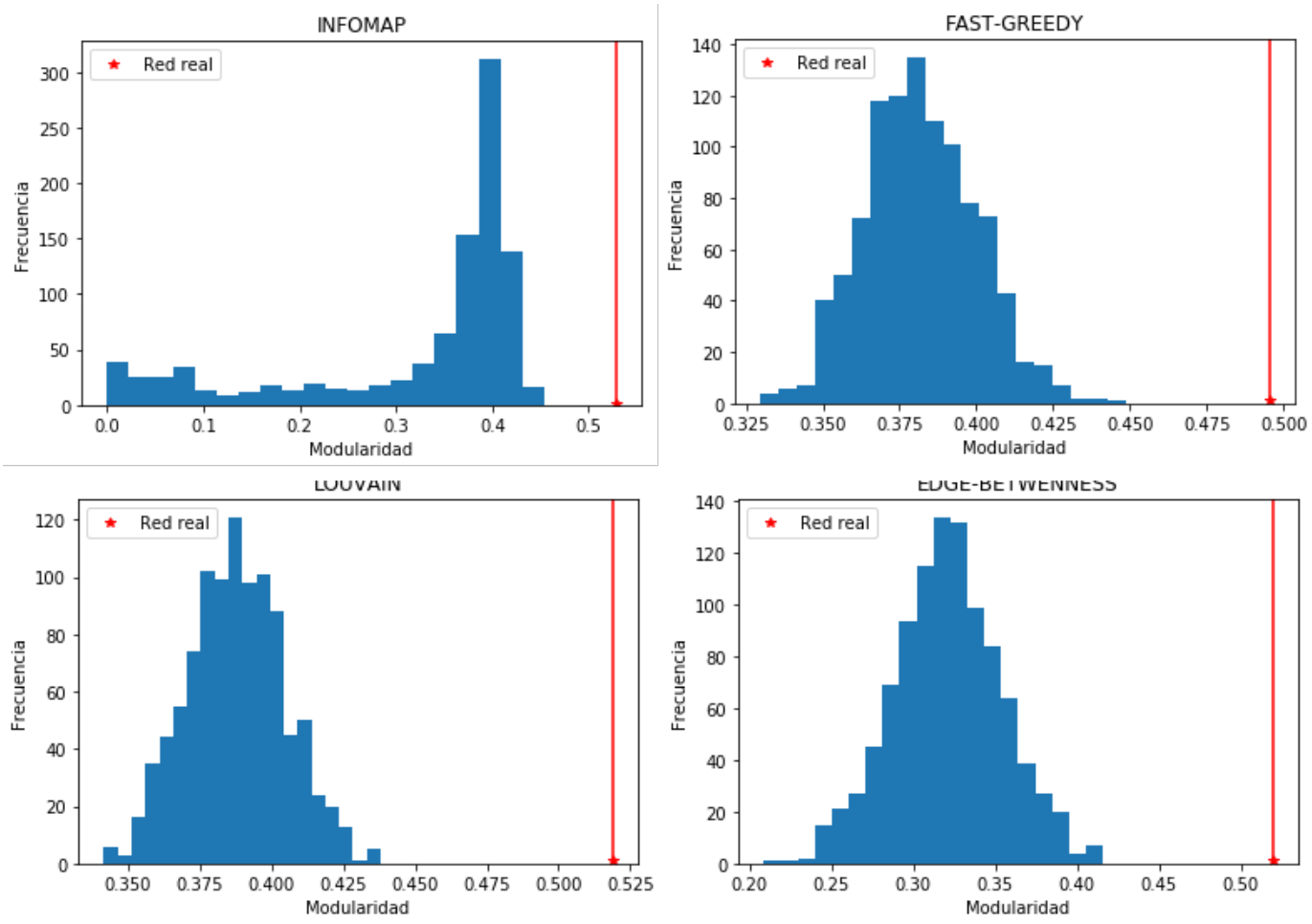

Figura 4: Histograma de modularidad con los distintos criterios de clustering

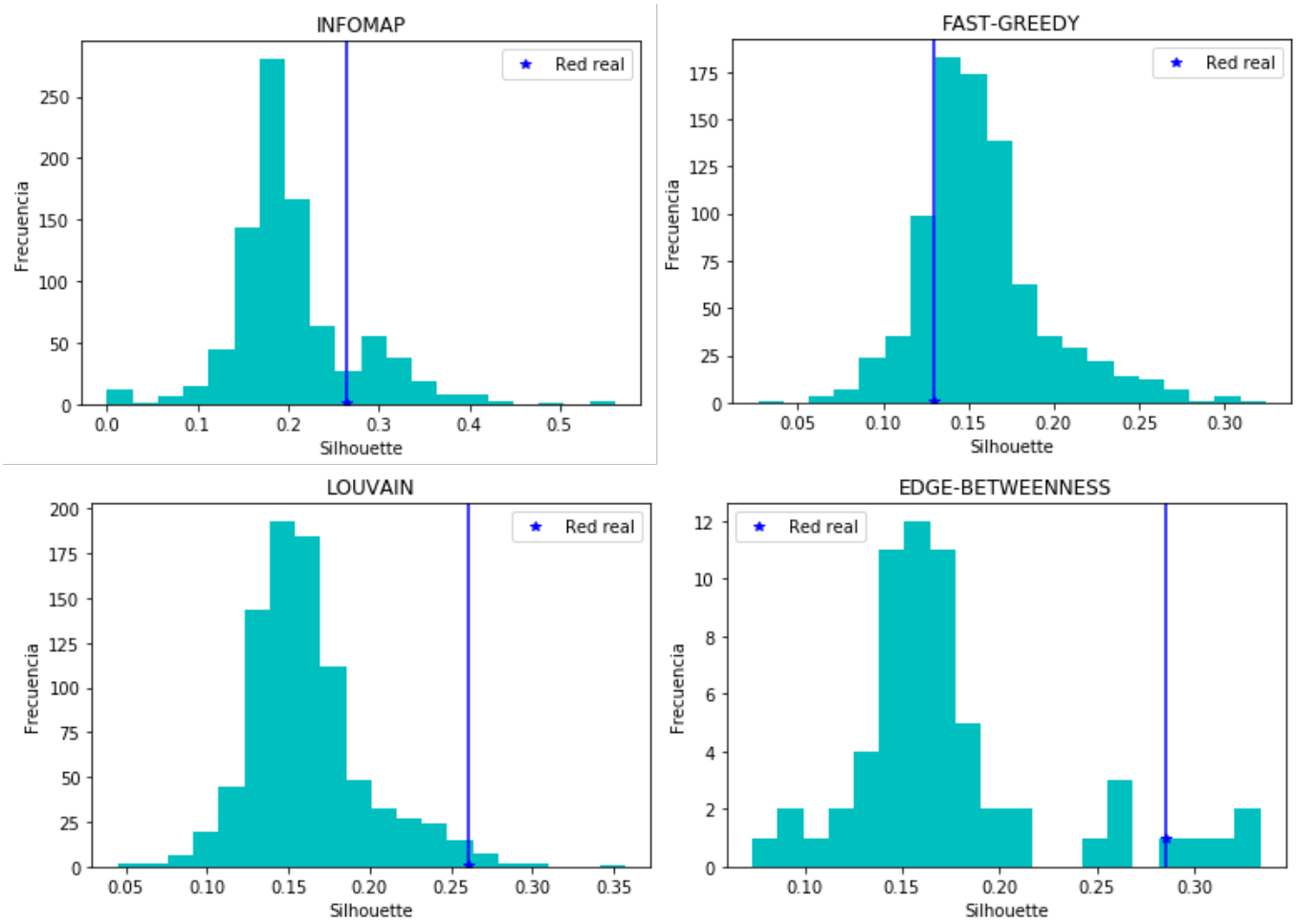

Figura 5: Histograma del coeficiente de Silhouette con los distintos criterios de clustering

A partir de estos análisis se puede concluir que la red de delfines es modular, puesto que la modularidad real tiene una diferencia significativa con la modularidad media obtenida del modelo nulo. También se observa que en la mayoría de los casos el coeficiente de Silhouette medio real tiene una desviación con respecto a la media considerable. El caso que resulta llamativo es en el cual se utilizó el criterio fast-greedy:

Es menester tener en cuenta que el coeficiente de Silhouette medio se ve afectado fuertemente por la cantidad de clusters, y esto mismo varía al recablear la red, con lo cual no resulta una medida muy afortunada. Prueba de esta variación la da el histograma mostrado en la figura 6, donde se puede observar la cantidad de comunidades determinadas por el método fast-greedy al realizar los recableados de la red original.

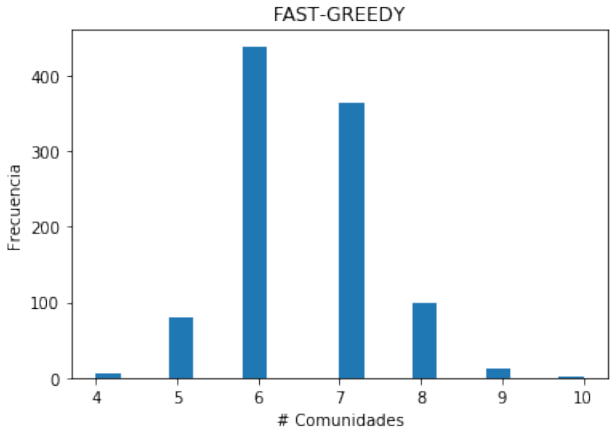

Figura 6: Histograma de la cantidad de clusters hayados en los recableados de la red original para el criterio fast greedy.

2. Acuerdo entre las particiones

Uno de los observables que nos permite hacer una comparación gráfica de los métodos es calcular la cantidad de nodos que hay en cada comunidad según las distribuciones como vemos en Fig. 7. Cabe aclarar que la denominación en comunidades C_i viene de menor a mayor cantidad de nodos para poder comparar las comunidades entre distintas particiones. Lo que uno esperaría es que la comunidad C_i de una dada partición se asemeje a la C_i de otra.

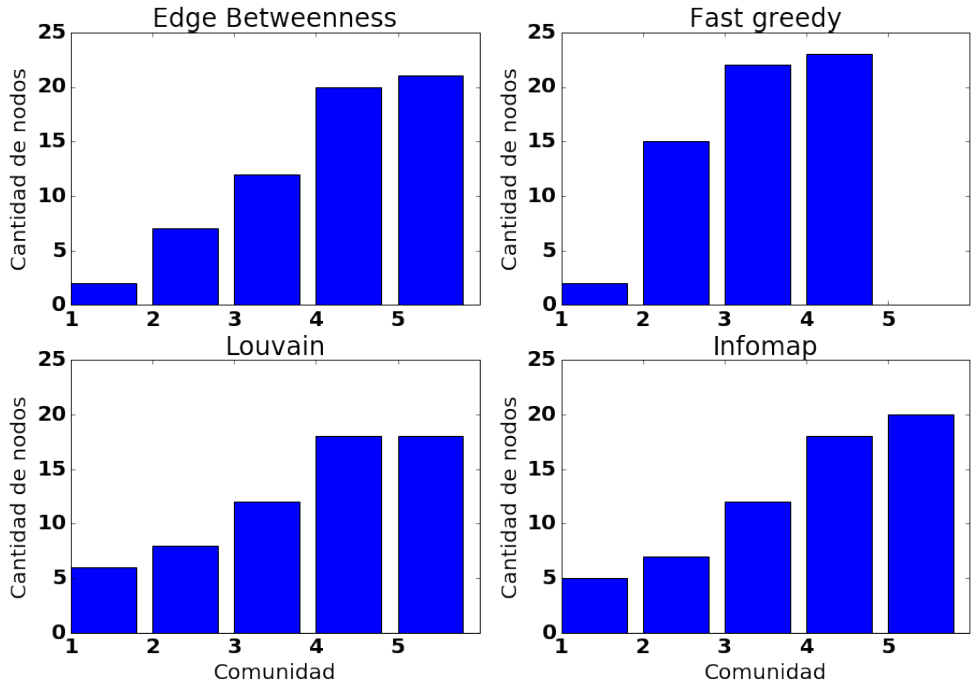

Figura 7: Histograma de la probabilidad no normalizada de cada comunidad para cada partición.

Otro observable es la matriz de la probabilidad conjunta. Esta expresa para cada par de distribución en comunidades la probabilidad conjunta de que un nodo elegido al azar pertenezca a la comunidad C1 de la primera partición y a la C2 de la segunda.

$$p(c_1, c_2) = \frac{N_{c_1,c_2}}{\sum_{c_1,c_2} N_{c_1,c_2}} \tag{1}$$

A continuación vemos la probabilidad conjunta entre partición edge betweenness y las otras estudiadas. En cada una de las filas se muestra a las 5 comunidades de la partición edge betweenness (EB) y en las columnas las respectivas comunidades de las demás particiones.

Se puede ver que éstas 3 matrices son idénticas ilustrando un acuerdo entre las particiones contra la que se compara el de Edge Betweenness. También se ve que los mayores valores se encuentran en la diagonal expresando la similitud entre las comunidades detectadas en cada partición.

		Fast Greedy					Louvain					Info map				
		C1	C2	C3	C4	C5	C1	C2	C3	C4	C5	C1	C2	C3	C4	C5
EB	C1	0	0	0	0	3.2	0	0	0	0	3.2	0	0	0	0	3.2
	C2	1.6	9.7	0	0	0	1.6	9.7	0	0	0	1.6	9.7	0	0	0
	C3	0	0	19.4	0	0	0	0	6.5	0	25	0	0	6.5	0	25
	C4	6.5	0	0	0	25.8	6.5	0	0	0	25.8	6.5		0	0	25.8
	C5	1.6	3.2	0	0	0	1.6	3.2	0	0	0	1.6	3.2	0	0	0

Cuadro 2: Matrices de probabilidad conjunta para la partición Edge Betweenness (sus comunidades en filas) y Fast Greedy, Louvain e Info Map (respectivas comunidades en columnas). Notése que para la partición de Fast Greedy hay 4 comunidades en vez de 5 como en el resto.

Otra forma de cuantificar cuánto se parecen dos particiones es a partir del cálculo de la información mutua ya que considera a todas las comunidades dentro de las particiones. Tiene en cuenta la probabilidad conjunta entre todo par de comunidades dadas dos particiones. Físicamente representa la información que tiene una partición con respecto a la otra. Se define en ec. 2. Cabe señalar que nosotros trabajamos con la fórmula normalizada por la mitad de la multiplicación de las entropías de Shannon de cada partición.

$$I(P_1, P_2) = \sum_{C_1} \sum_{C_2} p(C_1, C_2) log(\frac{p(C_1, C_2)}{p(C_1)p(C_2)}) \tag{2}$$

Un valor alto (cercano a 1) de ésta magnitud nos está diciendo que comparten bastante información por lo que entonces ambas particiones se podría decir que son buenas representaciones de las comunidades del grafo que estudiamos. Se puede ver en Cuadro 3 que los valores que de la información mutua que involucra a Fast greedy son bajos (los valores de su fila y de su columna) por lo que entonces se podría decir que no es una buena partición bajo éste criterio. Esto se condice con lo que se afirmó a partir del criterio de la modularidad.

	Edge Betweenness	Fast Greedy	Louvain	Infomap
Edge Betweenness		0.53	0.78	0.80
Fast Greedy	0.53		0.54	0.53
Louvain	0.78	0.54		0.79
Infomap	0.80	0.79	0.79	

Cuadro 3: Información mutua entre las particiones considerando todas las comunidades de cada respectiva partición. El elemento que se encuentra en la fila, por ejemplo, Louvain y en la columna Edge Betweenness es el valor de la información mutua entre dichas particiones

3. Género de los delfines vs estructura de comunidades del grupo

Finalmente, se analizó cuantitativamente la relación entre el género de los delfines y la estructura en comunidades del grupo.

Para ello, se eligió como método el de comparar el comportamiento del grupo con una hipótesis nula. Para cada metodología de partición en comunidades, se calculó la fracción de delfines de un mismo género que había por comunidad. Luego, se realizó una reasignación aleatoria de géneros a los delfines, y se calculó nuevamente dicha fracción. Se repitió 10.000 veces este último paso de reasignar los géneros y calcular la fracción de delfines del mismo género para cada comunidad.

Es importante destacar, que en la reasignación de géneros se mantuvo constante la cantidad de delfines con cada género.

Para todas las metodologías, con los resultados obtenidos se realizaron para cada comunidad, tres histogramas, cada uno correspondiente a cada género (se consideró el caso de los delfines sin género asignado). En la figura 8 se muestran a modo de ejemplo los histogramas para la metodología de *Fast-Greedy*.

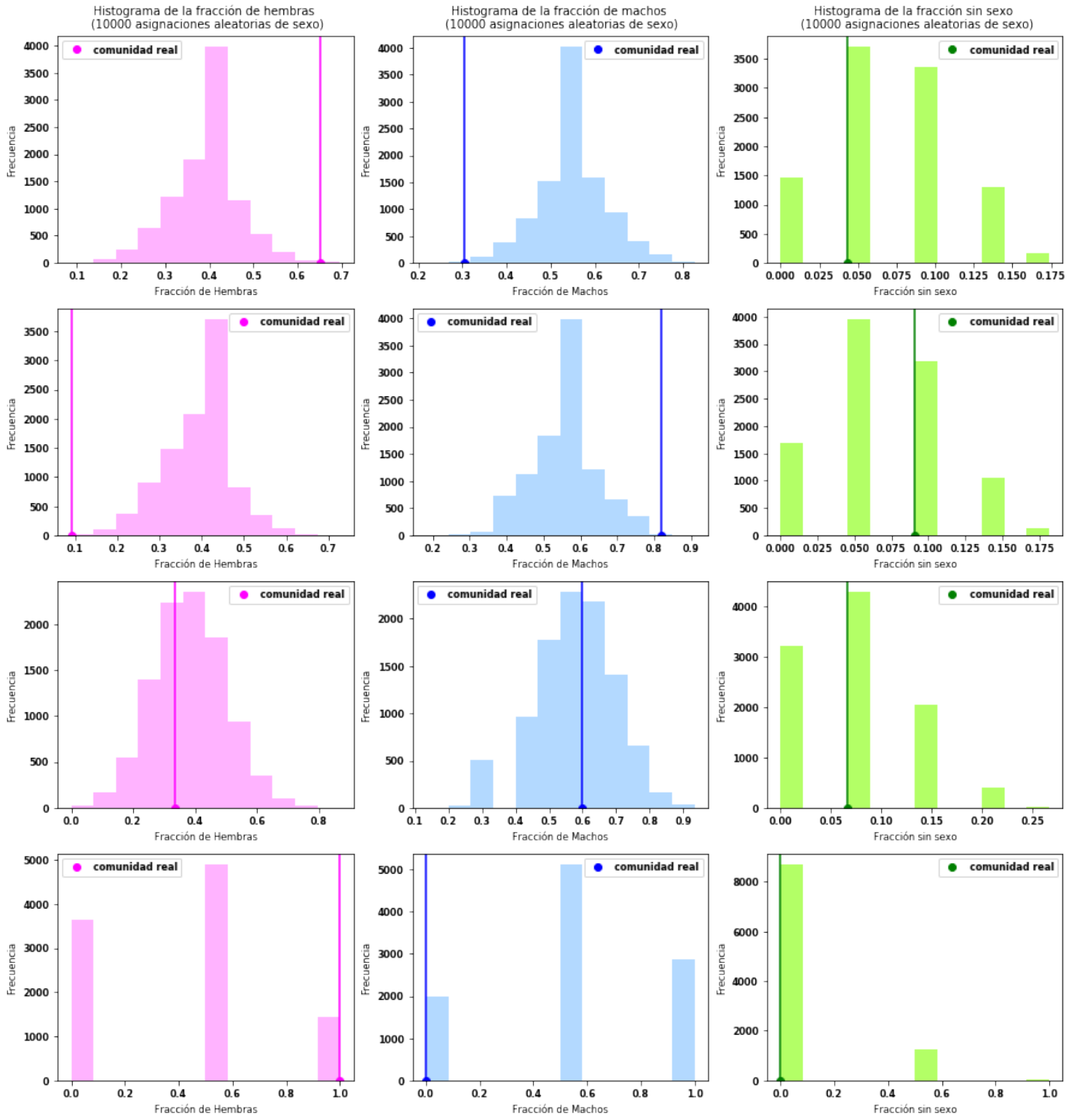

Figura 8: Histogramas de la fracción de delfines de un mismo género por comunidades a partir del método de Fast Greedy. Cada fila corresponde a una comunidad de la red, y se grafica la fracción de hembras, machos y delfines sin género a partir de asignación de género al azar. Con una línea recta se marca el valor obtenido para la estructura en comunidades del grupo real.

Bajo la hipótesis de que las comunidades tiene un carácter marcado de homofilia, se esperaría tener valores lejanos a la la media en los histogramas. En la figura 8 se puede apreciar que las primeras dos comunidades muestran una naturaleza de este tipo. Para la cuarta comunidad, uno podría aventurarse a decir que también presenta carácter de homofilia, ya que el 100 % de los delfines son hembras. Sin embargo, esta comunidad está conformada por solo dos delfines, valor que no resulta lo suficientemente grande para afirmar que la razón de que ambos delfines tengan el mismo género sea la tendencia a estar entre pares del mismo género. También se obtuvo una comunidad dentro de la media esperada al azar en cuanto a fracción de delfines del mismo género.

Para todas las metodologías se obtuvo un comportamiento parecido. En la tabla 4 se muestra, para

cada metodología, que porcentaje de comunidades tienen naturaleza homofílca (como las comunidades 1 y 2 de la figura 8), qué porcentaje se encuentra dentro del valor esperado al azar, y qué porcentaje no tiene suficientes delfines como para determinar un comportamiento conclusivo.

	Evidencia de homofilia	Sin evidencia de homofilia	Resultado no conclusivo
Infomap	60 %	20 %	20 %
Fast-Greedy	50 %	25 %	25 %
Louvain	40 %	40 %	20 %
Edge-Betweenness	40 %	40 %	20 %

Cuadro 4: Porcentaje de comunidades en el grupo de delfines que presentan cada comportamiento, para las distintas metodologías de partición en comunidades.

Referencias

[1] Rosvall Bergstrom, PNAS 2008, Maps of random walks on complex networks reveal community structure

[2] A Clauset, MEJ Newman, C Moore. Finding community structure in very large networks

[3] Blondel, J.Stat.Mech 2008.

[4] M. Girvan M.E.J. Newman, PNAS 99 (2002).