Trabajo Práctico Computacional 03

Redes Complejas 2018

Noelia Parzajuk, Sofía Nicoletti
noelparzajuk@gmail.com - nicolettisofia1@gmail.com

7 de Noviembre 2018

1. Introducción

Al momento de caracterizar una red, parte de dicha tarea consiste en hallar grupos de nodos que se destacan del resto de la red debido a similitudes en sus propiedades topológicas, dando a lugar a una *comunidad*. En el presente trabajo exploraremos cuatro metodologías para obtener distintas particiones en comunidades, aplicadas a una red de delfines. Caracterizaremos la estructura de comunidad en cada partición y la correspondencia en la composición de clústers entre distintas particiones. Además, analizaremos la relación entre el género de los individuos y la estructura de comunidades del grupo.

2. Algoritmos de partición

Girvan-Newman algorithm (edge betweenness): Remueve iterativamente enlaces de alta intermediatez (betweenness), exponiendo así la estructura de comunidad[1].

Greedy algorithm: Elige, de forma iterativa, pares de nodos enlazados de modo tal que su pertenencia a una misma comunidad provoque el mayor incremento en la modularidad de la red hasta llegar al máximo.

Louvain: Agrupa nodos con sus vecinos de modo tal que obtenga el mayor incremento en la modularidad de forma local y arma una nueva red en la cual los nodos son las comunidades obtenidas previamente. Repite el procedimiento con la nueva red, hasta que la modularidad no vuelva a modificar su valor.

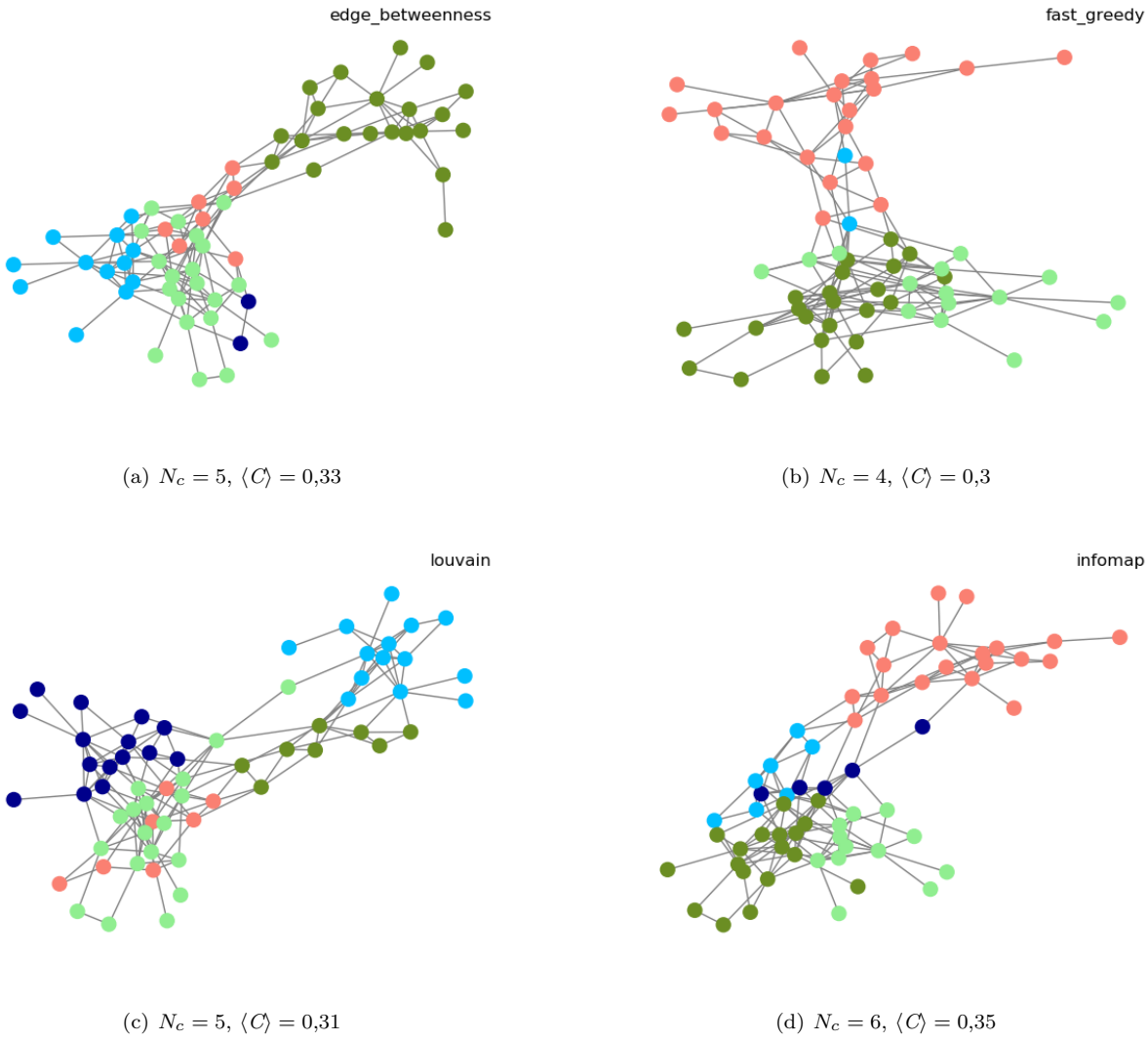
Infomap: Busca describir la trayectoria de un caminante al azar con la menor cantidad de símbolos, explotando la hipótesis de que éste tiende a quedarse atrapado en comunidades. De esta forma, el código que finalmente describe la trayectoria, corresponderá a las comunidades que haya atravezado.

Aplicamos estos algoritmos en la red de delfines usando el módulo de Python Networkx. En la Figura 1 se encuentran las particiones obtenidas con cada método. Se calculó el coeficiente de clústering de las tres comunidades más grandes de cada partición. Esto es

$$C = \frac{1}{n} \sum_{v \in com} c_v,$$

donde n es el número de nodos de la comunidad y c_i el coeficiente de clústering del nodo i. Luego se promediaron estos tres valores, obteniendo $\langle C \rangle$. Su valor oscila entre los 0,3 y 0,35 en los cuatro casos. Esto indica un mayor aglutinamiento dentro de cada comunidad si consideramos que el coef. de clústering total de la red es 0,26. De todas maneras, en las Secciones que siguen haremos un análisis exhaustivo para caracterizar qué tan "buenas"son estas particiones.

[1] Edge betweennes: centralidad asociada al número de caminos más cortos entre pares de nodos que pasan por él.



(a) $N_c = 5$, $\langle C \rangle = 0{,}33$

(b) $N_c = 4$, $\langle C \rangle = 0{,}3$

(c) $N_c = 5$, $\langle C \rangle = 0{,}31$

(d) $N_c = 6$, $\langle C \rangle = 0{,}35$

Figura 1: **Algoritmos de partición.** Red de delfines particionada según los diferentes métodos: 1(a) Edge-Betweenness, 1(b) Fast-Greedy, 1(c) Louvain y 1(d) Infomap. N_c indica la cantidad de comunidades y $\langle C \rangle$ es el coeficiente de clústering de las tres comunidades más grandes promediado. El coeficiente de clústering total de la red es 0,26.

3. Estructura de comunidad

Modularidad

La modularidad nos da una idea de la asortatividad de la red respecto a un cierto atributo en el siguiente sentido: una red es asortativa si hay una fracción significativa de enlaces dentro de un mismo grupo, y la modularidad nos permitirá cuantificar esto. Pues Q es la diferencia entre la fracción de enlaces que ocurren dentro de un mismo grupo y la fracción que esperaríamos tener si los todos los enlaces se posicionaran de manera aleatoria conservando la distribución de grado. Es decir, la modularidad compara la cantidad de enlaces intra-grupo contra una hipótesis nula teórica: recableado de la red preservando el grado de los nodos. Este coeficiente es estrictamente menor que 1, toma valores positivos si hay más enlaces intra-grupo que los que esperaríamos tener, y es negativa si tenemos menos que lo esperado por azar. La misma es descrita de la siguiente forma:

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j),$$

donde m es el número de enlaces, A la matriz de adyacencia, k_q el grado del nodo q y c_q su atributo de interés asociado.

La modularidad de la red para cada una de las particiones da cercana a 0,5. Es positiva, lo cual indica asortatividad, pero nos gustaría un ir poco más allá. Para ello consideramos comparar el coeficiente obtenido con el de un modelo nulo. El mismo consiste en un recableado real de la red, conservando el grado y sexo de cada nodo. Una vez aleatorizados los enlaces de la red, se particiona la nueva red según cada algoritmo, obteniendo nuevas particiones y se mide Q. En la Figura 2 se encuentra la distribución de dicho coeficiente para 1000 redes aleatorias con su respectivo valor medio (linea punteada). En línea sólida azul se encuentra la modularidad de la red real. Para Edge-Betweenness, Louvain y Fast-Greedy, la partición de la red presenta una mayor modularidad, es decir mayor conectividad entre nodos de la misma comunidad, de la que resultaría de recablear aleatoriamente la red. Por lo tanto, puede decirse que las particiones obtenidas bajo estos algoritmos representan la estructura de comunidad del conjunto de delfines. El histograma correspondiente al algoritmo Infomap presenta dos picos de modularidades destacadas. No obstante, el valor medido sigue indicando que la partición de la red posee una mayor modularidad que la que resulta de recablearla aleatoriamente.

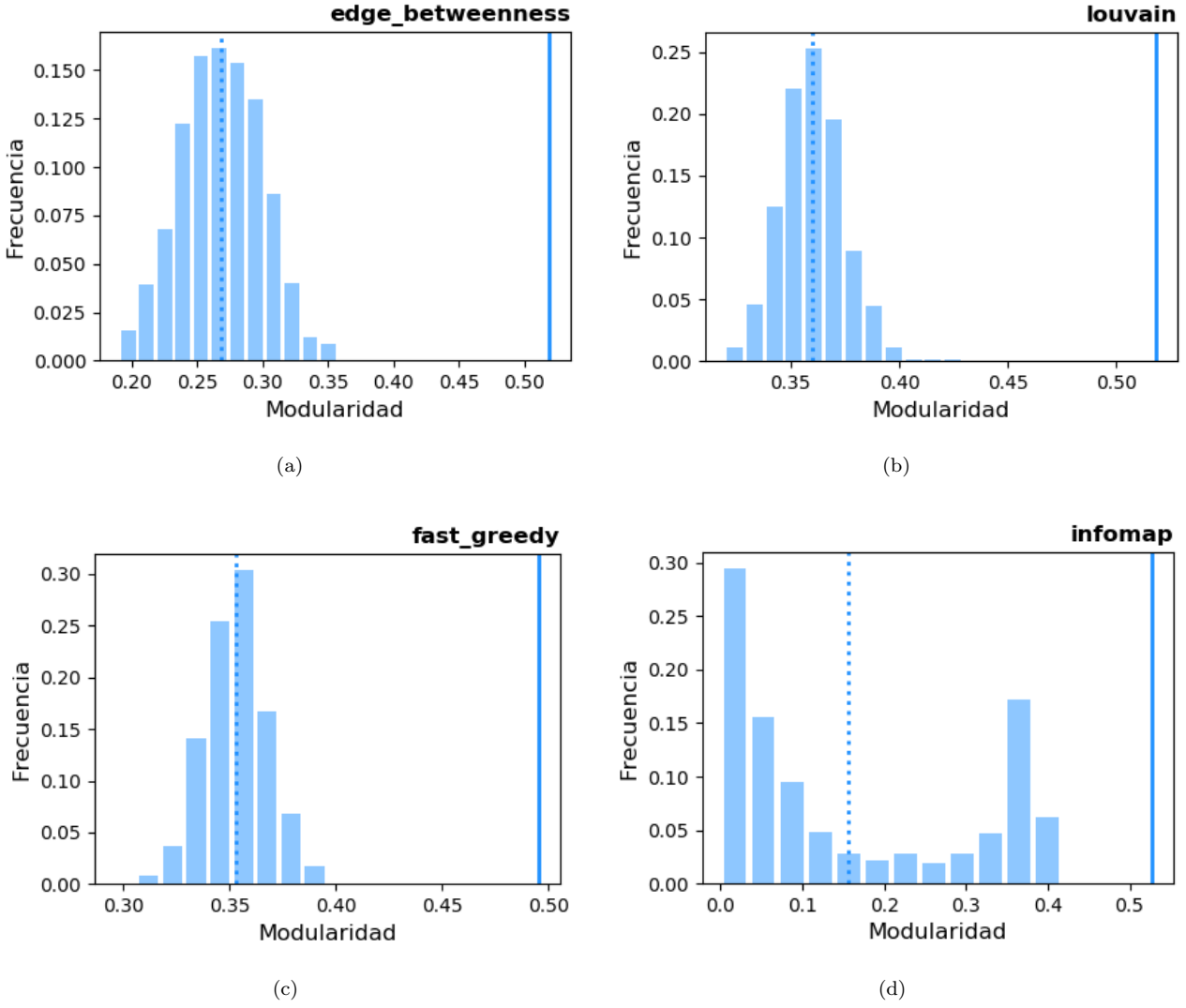


Figura 2: **Estructura de comunidad en red real vs. red recableada: coeficiente de modularidad.** La distribución representa el coeficiente de modularidad medido para 1000 redes recableadas aleatoriamente conservando el grado y sexo de cada nodo. Cada red aleatoria se volvió a particionar según los algoritmos vistos y a partir de allí calculado el coeficiente. En línea punteada la media de la dsitribución, en línea sólida la modularidad de la red real.

Luego obtener las distribuciones, podemos decir que $Q \approx 0{,}5$ es realmente significativo, pues se encuentra por

fuera de la distribución esperada por azar, en todos los casos. De todas maneas, la media de las distribuciones sigue dando positiva. Esto tiene sentido: pues una vez recableada la red, se vuelven a calcular particiones y el algoritmo capta cierta estructura de comunidades haciendo que el coeficiente de modularidad siga dando mayor a cero. Lo interesante, es que en la red real la estructura de comunidad pareciera ser mucho más fuerte: una red más *modular* que lo esperado por azar.

Silhouette

Otra forma de cuantificar la eficiencia de una partición, es ver qué tan similar es un nodo con su comunidad. Esta similaridad se cuantifica con el coeficiente de *silhouette*. Este valor toma valores de -1 a +1, donde un valor cercano a +1 indica que el nodo esá más relacionado con su propia comunidad que con las otras. Si la mayoría de los nodos toman valores altos, luego la particion es adecuada. En la Figura 3 se puede ver el promedio del coeficiente para todos los nodos de la red (linea azul), y la distribución que resulta de recablear la red 1000 veces aleatoriamente. El valor medio de dicha distribución es la línea punteada.

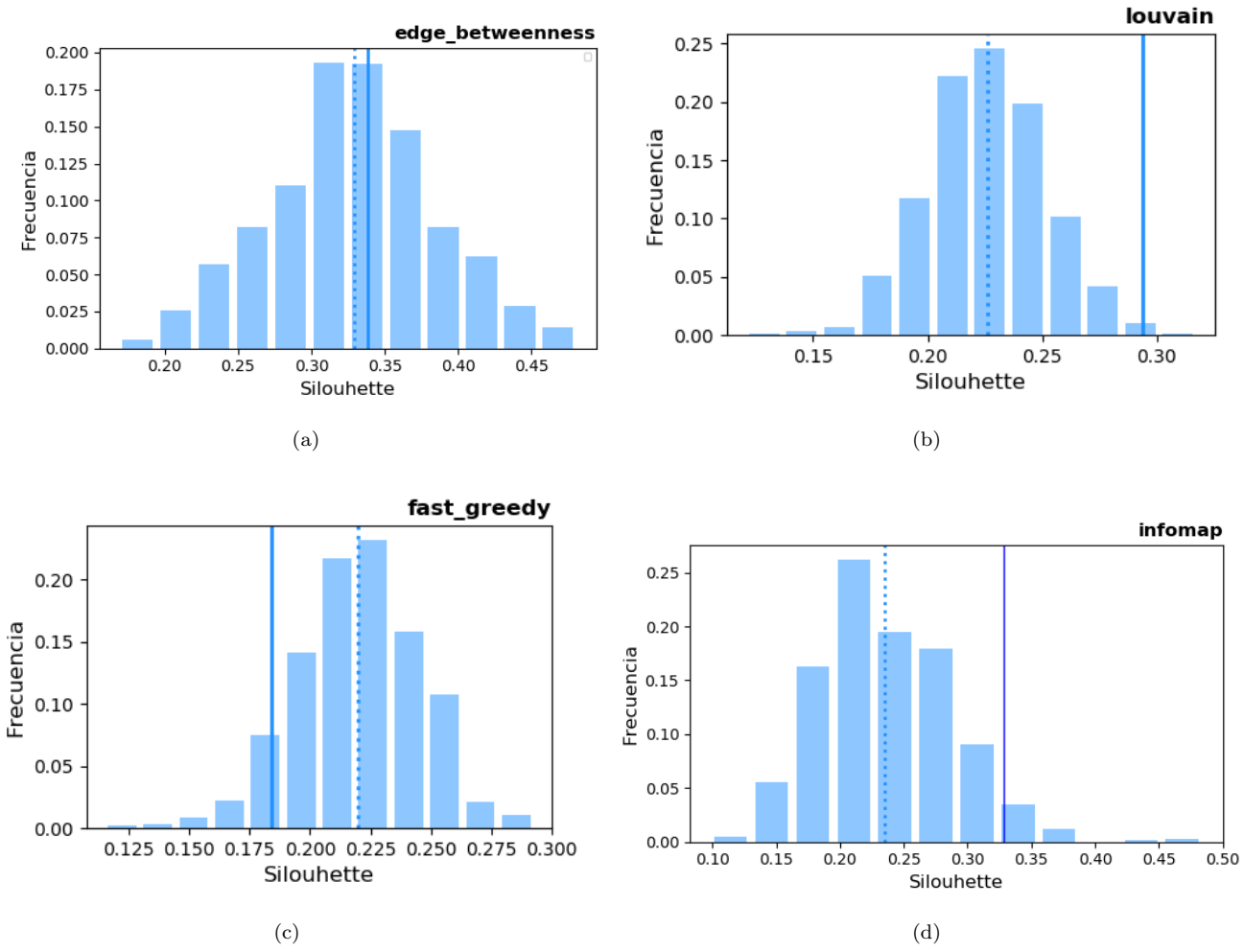


Figura 3: **Estructura de comunidad en red real vs. red recableada: coeficiente de silhouette.** Línea sólida: promedio de los coeficientes de silhouette para todos los nodos de la red real. Línea punteada: media de la disitribución generada a partir de recablear la red conservando el sexo y grado de cada nodo. Se realizaron 1000 iteraciones y para cada una de ellas se volvió a particionar la red según cada algortimo.

En todas las particiones obtenidas, el coeficiente detecta estructura de comunidad, dado que los valores medidos son positivos. Aún así, es necesario poner en contexto al coeficiente obtenido, por lo tanto generamos su distribución para 1000 redes obtenidas mediante un recableado aleatorio que preservaba la distribución de género. A partir de

esto, podemos observar que los algoritmos infomap y louvain arrojan particiones, de la red de delfines, con un coeficiente mayor al esperado tras particionar redes aleatoriamente recableadas. Con éste criterio, son éstos dos algoritmos los que generan particiones en los que los nodos son más similares con su comunidad. Por otro lado, fast-greedy generó una particion con comunidades mas debilmente relacionadas, mientras que edge-betweenness no se destaca respecto al recableado aleatorio.

4. Caracterización

Información mutua

Para evaluar la habilidad de un algoritmo de hallar comunidades, es necesario medir la capacidad de que éste reproduzca las comunidades de redes cuya partición conocemos.

En el presente trabajo no tenemos una partición de referencia, pero lo que podemos hacer es medir qué tan parecidas son las particiones que se obtienen a partir de los distintos algoritmos. Una forma de caracterizar esta similitud es el coeficiente de *información mutua*, definido de la siguiente manera:

$$I(C_1, C_2) = \sum_{C_2} p(C_1, C_2) \ \log \frac{p(C_1, C_2)}{p(C_1)\ p(C_2)}$$

donde $p(C_1, C_2)$ es la probabilidad de que un nodo elegido al azar pertenezca a la comunidad C_1 de la primera particion y la comunidad C_2 de la segunda y $p(C_i$ es la probabilidad de que un nodo pertenezca a la comunidad i de una dada partición. Presentaremos dicha medida de forma normalizada con la *entropía de Shannon* de ambas particiones (i.e. con la cantidad de información obtenida al conocer una variable aleatoria), que tiene la siguiente forma:

$$I_n(C_1, C_2) = \frac{1}{2} \ \frac{I(C_1, C_2)}{H(C_1) + H(C_2)}$$

En la Tabla 1 se puede ver ésta medida aplicada en los algoritmos desarrollados en el trabajo. Infomap y Edge-Betweenness son los algoritmos que generan las particiones más parecidas (apesar incluso de que su proceso iterativo sea completamente distinto), seguidos inmediatamente por Fast-Greedy y Louvain (en este caso, el segundo es una optimización del primero). Por otro lado, Fast-Greedy y Edge-Betweenness son los que poseen la partición menos parecida (en comparación con los valores obtenidos entre los demas algoritmos)

Louvain	0.79	0.73	0.78
0.79	F-G	0.66	0.76
0.73	0.66	E-B	0.91
0.78	0.76	0.91	Infomap

Tabla 1: Información mutua entre las particiones obtenidas a partir de cada algoritmo

Precisión

Otra forma de cuantificar la similitud entre particiones es utilizando la *matriz de confusión*. La misma contiene en sus elementos la cantidad de pares de nodos que poseen las siguientes características:

a. Pertenecen a la misma comunidad en ambas particiones

b. Pertenecen a distintas comunidades en la partición 1 pero la misma en la partición 2

c. Pertenecen a la misma comunidad en la partición 1 pero a distintas en la partición 2

d. Pertenecen a distintas comunidades en ambas particiones

El coeficiente de precisión cuantifica cuantos nodos siguen perteneciendo a la misma comunidad, y cuantos siguen sin pertenecer a la misma comunidad, normalizado de la siguiente manera:

$$\text{precision} = \frac{a+d}{a+b+c+d} = 2\,\frac{a+d}{\mathrm{N(N-1)}}$$

donde a, b, c y d representan la cantidad de pares de nodos que cumplen alguna de las características descritas previamente, y N es la cantidad de nodos del grafo.

La Tabla 2 muestra los resultados obtenidos para los algoritmos usados. Nuevamente, Infomap y Edge-Betweenness proveen de la estructura de comunidades más parecidas entre sí, mientras que Fast-Greedy y Edge-Betweenness son los que poseen la estructura menos parecida (en comparación con los valores obtenidos entre los demas algoritmos)

Louvain	0.86	0.87	0.88
0.86	F-G	0.84	0.90
0.87	0.84	E-B	0.93
0.88	0.90	0.93	Infomap

Tabla 2: Coeficiente de precisión entre las particiones obtenidas para cada algoritmo

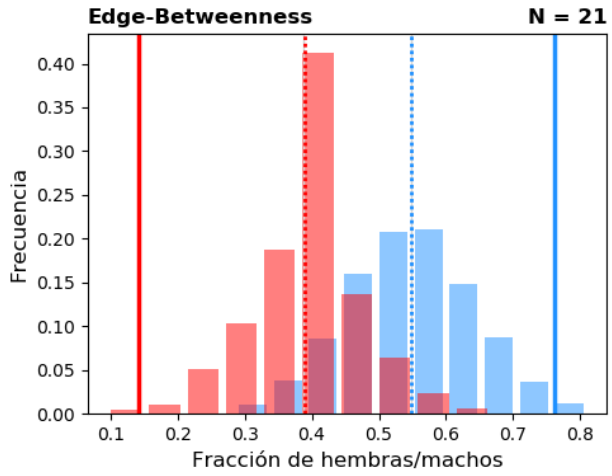
5. Género y estructura de comunidad

Como última instancia de análisis, resulta de interés estudiar si existe alguna relación entre el género de los individuos y la estructura de comunidad subyacente. La red consiste en un grupo de 62 delfines, de los cuales 34 son machos (54.8 %) y 24 hembras (38.7 %), los 4 restantes tienen género no identificado. Dada una partición, arrancaremos viendo cuál es la fracción de machos y hembras en cada comunidad y cómo difiere ésta de la esperada por azar. Es decir, el modelo nulo consistirá en asignar los géneros de manera aleatoria en toda la red, conservando la proporción de machos y hembras.

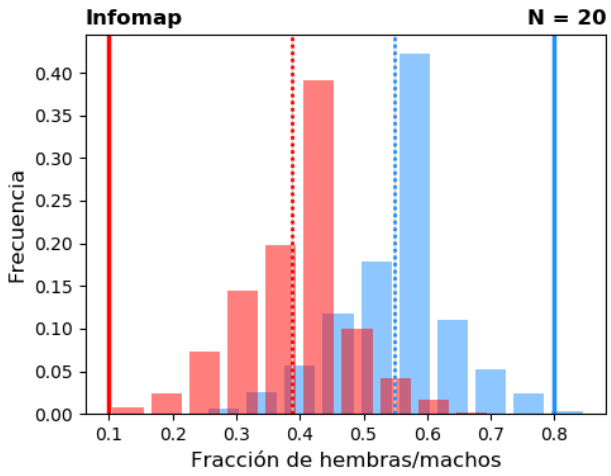
Consideramos cuatro particiones, cada una correspondiente a un algoritmo de partición. Calculamos la distribución nula asignando géneros al azar 3000 veces y midiendo la fracción de machos y hembras cada vez. En la Figura 4 se encuentran los histogramas para alguna de las comunidades más grandes de cada partición. La línea punteada indica la media de la distribución, y línea gris, apenas distinguible por debajo de la punteada, indica la fracción real de toda la red. Es decir: la fracción esperada por azar en cada comunidad es la misma que la proporción real de toda la red, como era de esperarse. Es alevoso el sesgo de género que se observa en estas cuatro comunidades de la Fig. 4: en todos los casos la fracción real del género sobre-representado se encuentra al menos dos desviaciones estándar desplazada respecto de la media.

Posterior a la visualización de las distribuciones, pasemos a un análisis un tanto más cuantitativo y general. En la Tabla 3 se encuentran para cada método las comunidades más representativas de cada partición ($> 10\,\%$ de la red). Para cada comunidad, se encuentra allí el número de individuos, el género sobre-representado y qué porción de la comunidad ocupa (*fracción real*). La sobre-representación viene necesariamente acompañada de la media de la distribución nula (*fracción media*), pues diremos que un género está sobre-resprensetado si la fracción real es mayor la que fracción media, pero ¿qué tanto mayor? Para ello se utilizaron dos métricas: por un lado el p-value (qué tan probable es medir por azar una fracción mayor a la observada), y por el otro, la distancia del valor observado respecto de la media en unidades de la desviación estándar de la distribución (*desplazamiento*). Es decir: a cuántas desviaciones estándar se encuentra el valor real de la media.

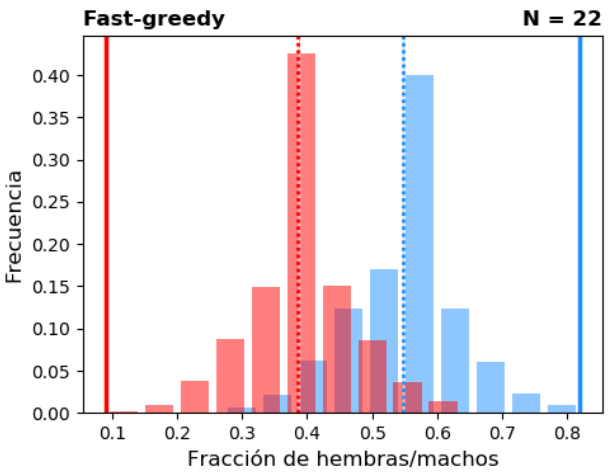
La relación género-comunidad es notable. En las cuatro particiones vemos una sola comunidad con sobre representación femenina. Lo interesante es que en todos los métodos resulta ésta la comunidad más sobre-representada: el p-value es 0.00 y el desplazamiento respecto de la media es el mayor de cada partición, siendo siempre $> 3{,}3$. Por otro lado, el p-value vemos que es en la mayoría de los casos del orden de 0.05 o menor, lo cual es una razón robusta para afirmar que la estructura de comunidad presenta en la red no puede disociarse del género de los individuos.



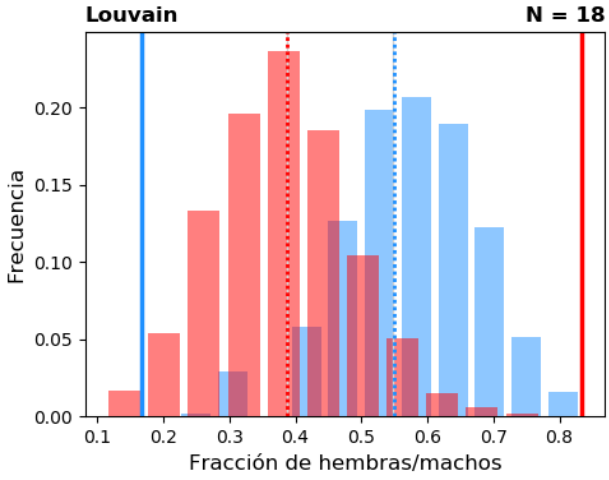
(a) Machos - fracción real: 0,76, media: $0{,}55 \pm 0{,}09$



(b) Machos - fracción real: 0,80, media: $0{,}55 \pm 0{,}09$



(c) Machos - fracción real: 0,81, media: $0{,}55 \pm 0{,}09$



(d) Hembras - fracción real: 0,83, media: $0{,}39 \pm 0{,}10$

Figura 4: **Relación entre género y estructura de comunidad.** En línea sólida azul y roja, fracción de indivudos machos y hembras de algunas de las comunidades de cada partición. La distribución nula corresponde a asignar de manera aleatoria los géneros, preservando la proporción total de la red. En línea punteada la media de la distribución y cercana ella en gris claro la fracción real de toda la red. Hay evidente sobre-representación de género en al menos una comunidad de cada método.

Louvain

N	sobre-representación	fracción media	fracción real	desplazamiento	p-value
18	female	$0{,}39 \pm 0{,}10$	0.83	4.50	0.00
15	male	$0{,}55 \pm 0{,}11$	0.60	0.45	0.44
14	male	$0{,}55 \pm 0{,}12$	0.71	1.34	0.04
8	male	$0{,}55 \pm 0{,}17$	1.00	2.69	0.00

Fast-greedy

N	sobre-representación	fracción media	fracción real	desplazamiento	p-value
23	female	$0{,}39 \pm 0{,}08$	0.65	3.27	0.00
22	male	$0{,}55 \pm 0{,}09$	0.82	3.15	0.00
15	male	$0{,}55 \pm 0{,}11$	0.60	0.44	0.23

Edge-Betweenness

N	sobre-representación	fracción media	fracción real	desplazamiento	p-value
21	male	$0{,}55 \pm 0{,}09$	0.76	2.43	0.01
20	female	$0{,}39 \pm 0{,}09$	0.80	4.46	0.00
12	male	$0{,}55 \pm 0{,}13$	0.75	1.57	0.02
7	male	$0{,}55 \pm 0{,}18$	0.71	0.92	0.08

Infomap

N	sobre-representación	fracción media	fracción real	desplazamiento	p-value
20	male	$0{,}55 \pm 0{,}09$	0.80	2.72	0.00
18	female	$0{,}39 \pm 0{,}10$	0.72	3.35	0.00
12	male	$0{,}55 \pm 0{,}13$	0.75	1.52	0.03
7	male	$0{,}55 \pm 0{,}18$	0.71	0.95	0.08

Tabla 3: **Sobre-representación de género en las comunidades.** Para cada método se encuentran las comunidades que representan una porción mayor al 10 % de la red total. Para cada comunidad se especifica el número de individuos N, el género sobre-representado, la *fracción media* de la distribución nula y la *fracción real* o observada. Por último están las medidas de cuantitativas de sobre-representación: el p-value y el desplazamiento respecto de la media (cant. de desviaciones estándar del valor real a la media).