Aplicación de métodos de detección de comunidades a una red de delfines

Favio Di Ciocco, Daniel Pinto, Diego Espejo, Alejandro Barton

7 de noviembre de 2018

1. Datos del TP

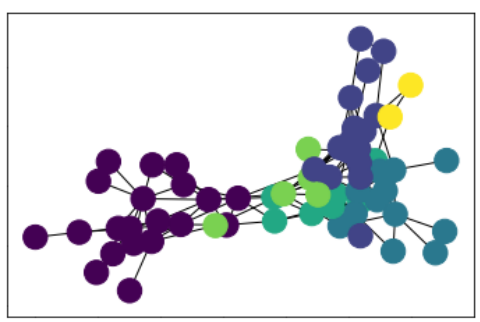
Para este trabajo se analizó nuevamente la red social de 62 delfines extraída de Lusseau et al.(2003). Sobre esta red se aplicarán los siguientes métodos para hallar comunidades, (Clústers), entre los nodos: Infomap, Fast-Greedy, Algoritmo de Louvain y Edge-Betwenness. Cada conjunto de comunidades halladas es lo que se llamará **partición**, y sobre estas particiones se estudiarán ciertas propiedades como modularidad y Silhoutte para caracterizarlas.

Dada una red con su partición, la modularidad Q nos dice en qué medida los nodos de una misma comunidad se interconectan entre sí, con respecto a lo que cabría esperar de repartir nodos al azar, conservando la distribución de grados. Q es positiva si hay más enlaces entre nodos de la misma comunidad que los que esperaríamos por azar.

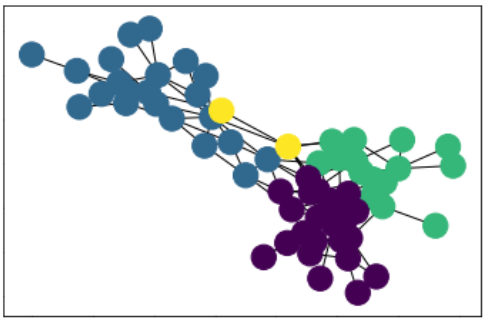
Si tenemos una red con cierta partición en comunidades, los algoritmos descriptos anteriormente tratan de maximizar la modularidad, cambiando las particiones en cada iteración.

La Silhouette de un nodo mide cuánto está incuido dentro de su cluster, comparándolo con los otros clusters. Una Silhouette nula o negativa se puede interpretar como que el nodo está en la frontera del cluster, o del lado de "afuera".

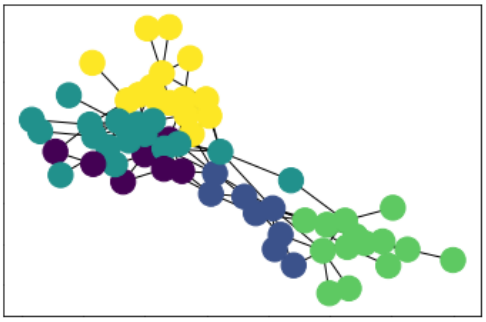
2. Visualización de los clústers dados por diferentes métodos



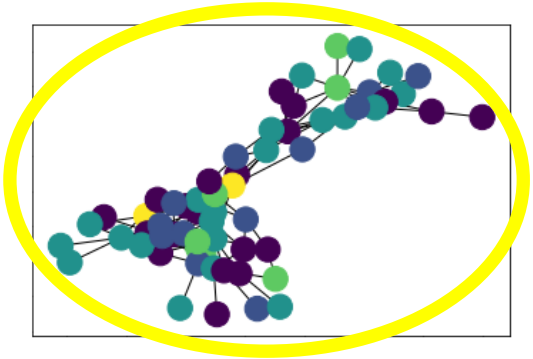
(a) Clústers dado el método Infomap



(b) Clústers dado el método Fast-Greedy



(c) Clústers dado el método Louvain



(d) Clústers dado el método Edge-Betweenness

Figura 1: Clústers para distintos métodos

Analizando la figura 1 se puede observar que el método de Edge-Betweenness para esta red resulta ineficaz, ya que los colores están muy mezclados y las particiones no resultan naturales a la vista.

En cambio en el método Fast-Greedy y en el algoritmo de Louvain se pueden notar bien diferenciadas las comunidades, con comunidades que apenas se superponen. Aunque claramente los dos dan particiones diferentes, ya que el método Fast-Greedy sólo presenta cuatro comunidades, mientras que el de Louvain separa los nodos en cinco comunidades.

No es casualidad que estos dos métodos coincidan en ser óptimos, ya que nodos en cada iteración, aunque de maneras distintas.

Aunque el grafico de una idea int

En cambio el método Infomap opera a través de realizar caminos aleator hallar las comunidades y por último el método de Edge Betweenness opera eliminando enlaces a través de un criterio de esencialidad para estos, y formando componentes separadas a través de la remoción de estos enlaces.

3. Caracterización de particiones según Modularidad y Silhouette

Los valores de Modularidad y Silhouette media para cada partición se muestran en la tabla1, donde para el valor de Silhouette se calculó el promedio sobre el valor de todos los nodos.

Se puede ver que el valor de modularidad del método Edge Betwenness da negativo, confirmando la observación antes hecha respecto de las comunidades formadas al visualizar la partición de este método.

Por otro lado, la modularidad del método Infomap es la máxima, seguida por el de Louvain y el de Fast-Greedy, lo cual nos marca que utilizando este criterio, el método Infomap es el que produce la mejor

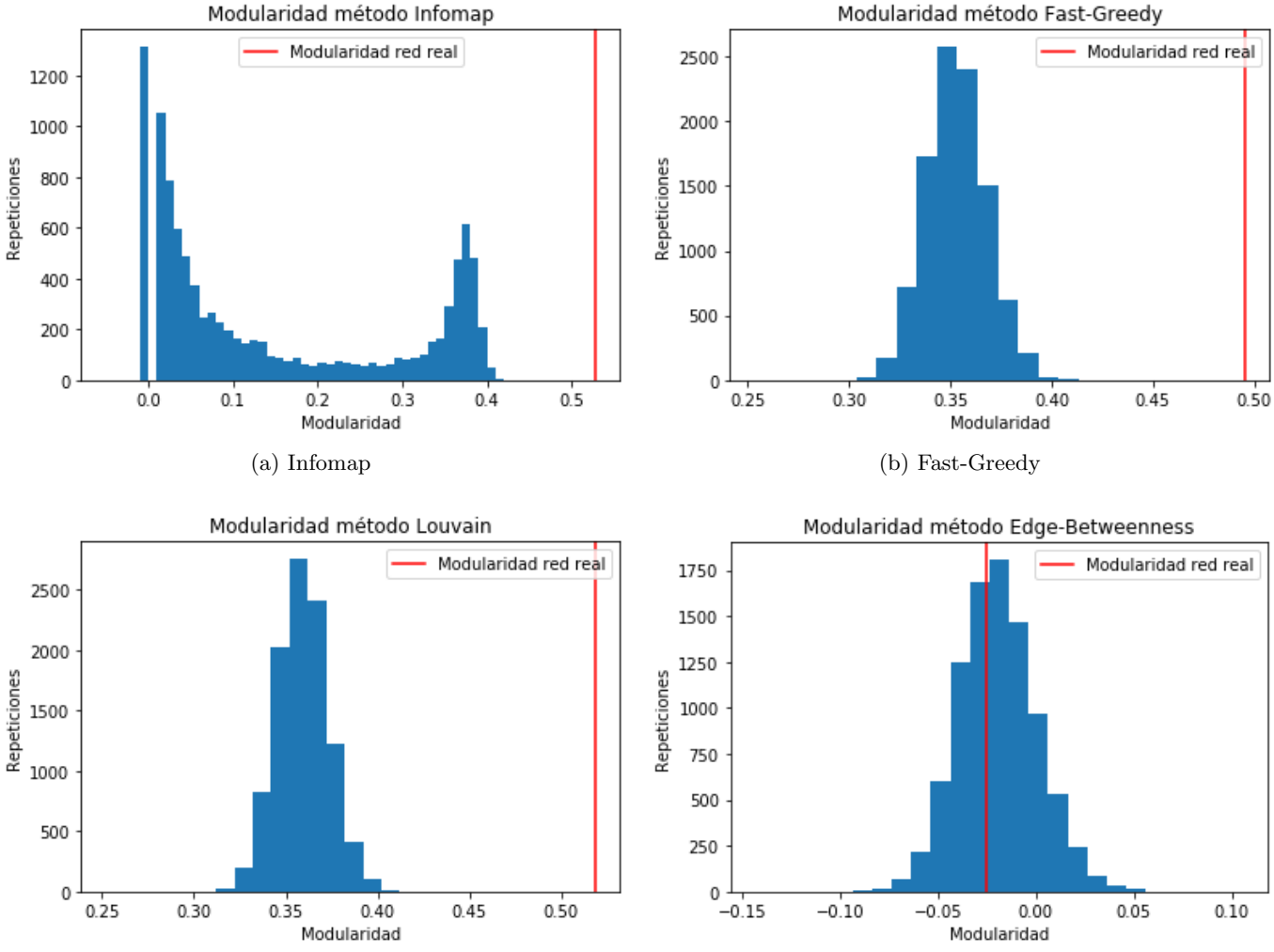
	Infomap	Fast-Greedy	Louvain	Edge-Betweenness
Modularidad	0.5189	0.4954	0.5185	-0.0255
Silhouette media	0.2632	0.3458	0.2769	0.2663

Tabla 1: Valores de Modularidad y Silhouette para diferentes metodos

partición.

Se define una red como "modular"si dado el conjunto de nodos y enlaces que tiene, al aplicarle algún metodo que identifique comunidades, permite obtener una partición con un valor alto de modularidad en comparación con el resto de redes posibles que se podrían armar. Para analizar si nuestra red es modular para los métodos utilizados, se recableo la red muchas veces de manera aleatoria y se graficó en un histograma los valores de modularidad hallados en comparación con el valor de modularidad de la red real.

A continuación se ven los histogramas de Modularidad para los cuatro métodos para hallar comunidades aplicados a la red recableada y su comparación con el valor obtenido en la red real.



(a) Infomap

(b) Fast-Greedy

(c) Louvain

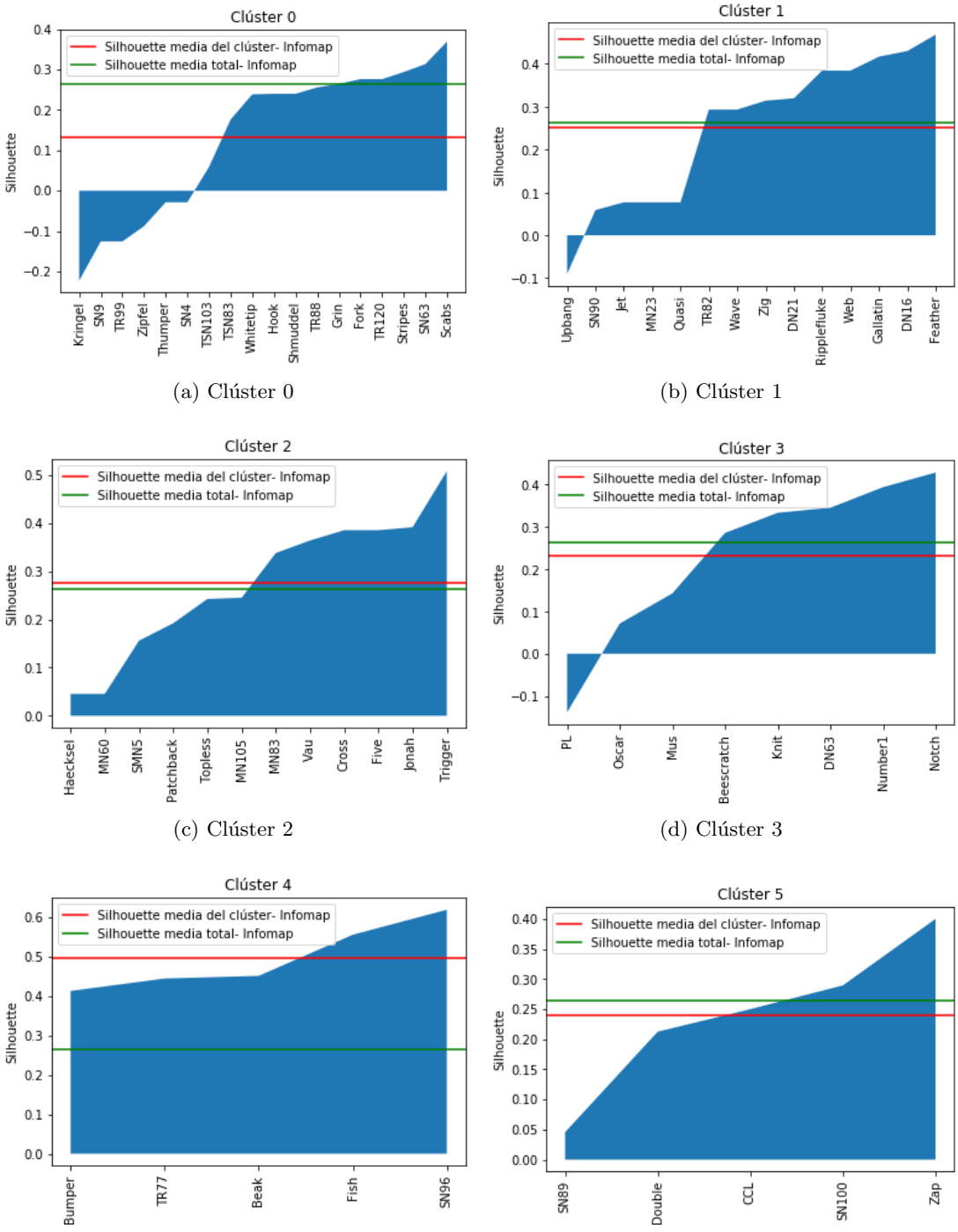
(d) Edge-Betweenness

Figura 2: Modularidad para diferentes Clústers del método Fast-Greedy

En la figura 2a se pueden observar valores de modularidad cercanos a cero y repetidos muchas veces, ésto es debido a que el algoritmo muchas veces resulta en comunidades muy pequeñas con valores muy bajos de modularidad. De todas formas se puede ver en las figuras 2a, 2b, 2c que la red se comporta de forma modular al mismo tiempo que su valor dista mucho del valor dado por la hipótesis nula, con lo cual

descartamos dicha hipótesis para explicar las comunidades. También se puede observar en la figura 2d que para éste método la modularidad da muy cercana al valor de la hipótesis nula, lo cual indica que para el método de Edge-Betwenness, la red no es modular.

A continuación se muestran los valores de Silhouette para diferentes clústers y según diferentes métodos y una comparacion con el valor obtenido dado el método para la red real:



(a) Clúster 0

(b) Clúster 1

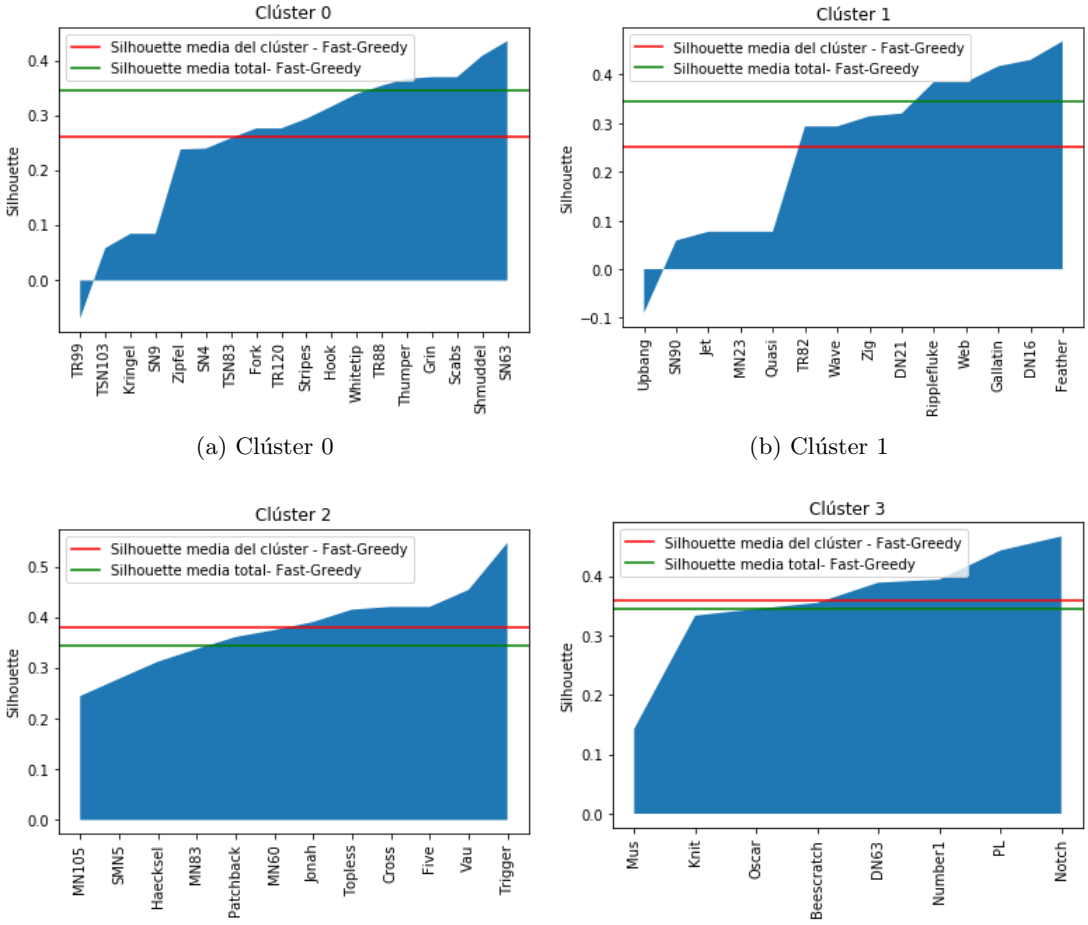
(c) Clúster 2

(d) Clúster 3

(e) Clúster 4

(f) Clúster 5

Figura 3: Silhouette para los Clústers del método Infomap



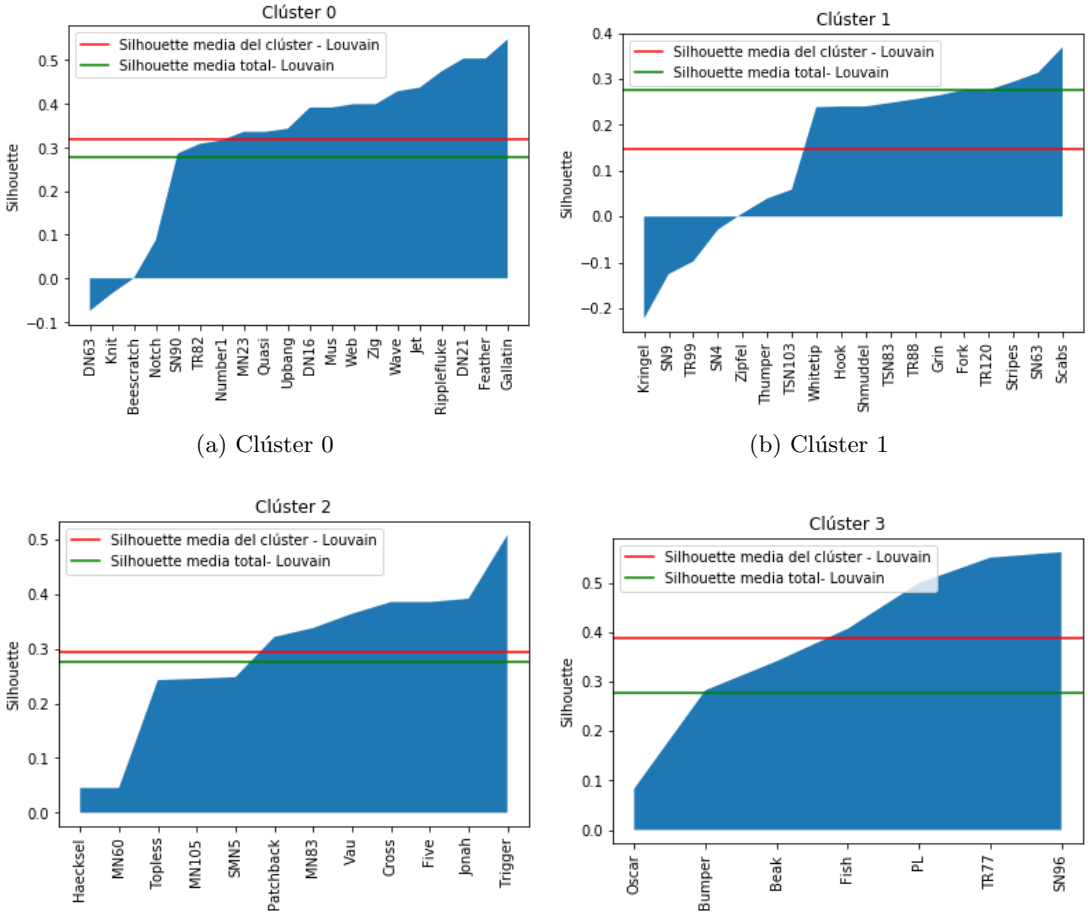
(a) Clúster 0

(b) Clúster 1

(c) Clúster 2

(d) Clúster 3

Figura 4: Silhouette para los Clústers del método Fast-Greedy



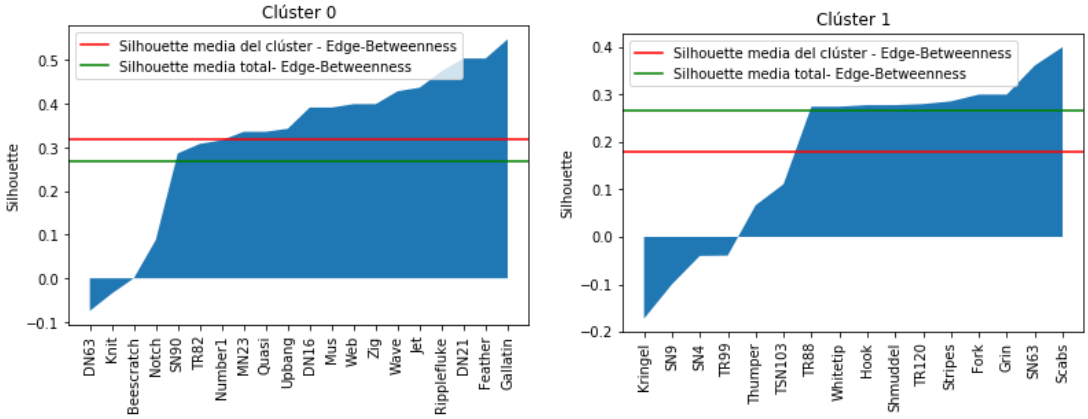
(a) Clúster 0

(b) Clúster 1

(c) Clúster 2

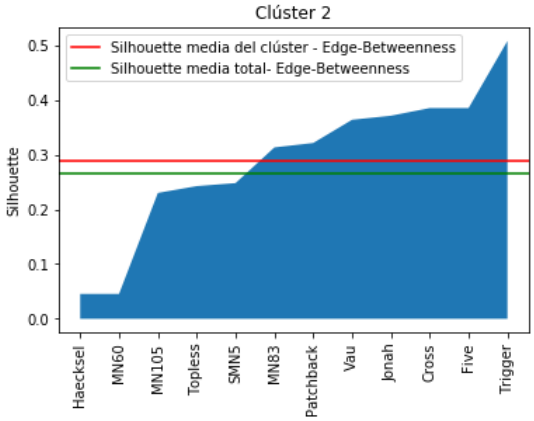
(d) Clúster 3

Figura 5: Silhouette para los Clústers del método Louvain

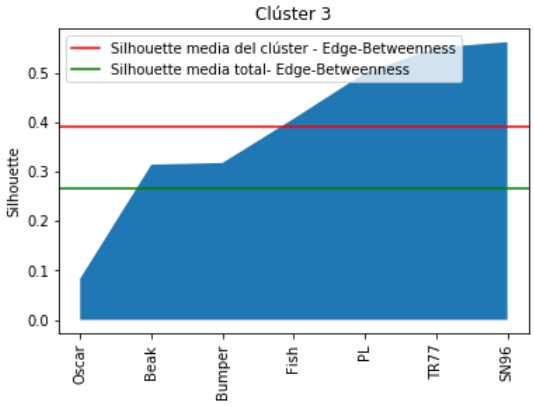


(a) Clúster 0

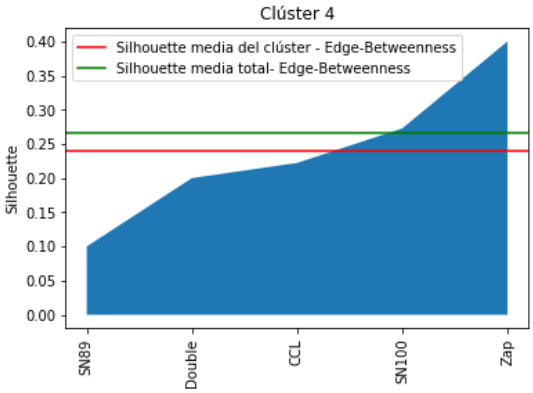
(b) Clúster 1



(c) Clúster 2



(d) Clúster 3



(e) Clúster 4

Figura 6: Silhouette para los Clústers del método Edge-Betweennes

4. Acuerdo entre particiones

Cada uno de los métodos utilizados para hallar comunidades devuelve particiones diferentes. Sería de interés poder cuantificar que tan similares son estas particiones, es decir cuánto acuerdo hay entre ellas, y cuanto acuerdo hay entre las comunidades formadas. Es por esto que se utilizarán dos métodos para calcular este acuerdo entre particiones: El método de *Información Mutua* y el de *Precisión*.

Información mutua \ Precision	Infomap	Fast-Greedy	Louvain	Edge-Betweenness
Infomap		0.8709	0.9386	0.8804
Fast-Greedy	0.7903		0.8646	0.8434
Louvain	0.8621	0.7948		0.8730
Edge-Betweenness	0.7718	0.6621	0.7329	

Tabla 2: Acuerdo entre particiones según los métodos de *Información Mutua* y *Precisión*. Los valores por encima de la diagonal corresponden a la *Precisión* entre dos particiones mientras que los que se encuentran por debajo se corresponden con la *Información Mutua*.

Al observar la tabla 2 se puede ver que excepto para el valor de acuerdo entre el método Edge-Betwenneess y el de Infomap según Información mutua, todos los valores son mayores a 0,73, indicando un buen acuerdo entre las particiones.

Además se puede observar que los va ok, si, efectivamente, son metricas diferentes, solo tiene sentido ha que los de *Información Mutua* para cualesquiera dos métodos. Esto está relacionado con que estas dos métricas para medir acuerdo se calculan de manera diferente. Aún así se puede notar que ambos muestran la misma tendencia, ya que el valor de mínimo acuerdo para *Información Mutua* se da para los métodos Edge-Betweenness y Fast-Greedy, mientras que para la métrica de *Precisión* el mínimo valor se corresponde a los mismos dos métodos.

De la misma manera, el máximo para el método de *Información Mutua* se da para los mismo dos métodos de detección de comunidades que para el de *Precisión*.

5. Distribuciones de género en las comunidades

Finalmente se analizaron las distribuciones de género en las distintas comunidades. Para esto se realizó el siguiente test sobre las cuatro particiones:

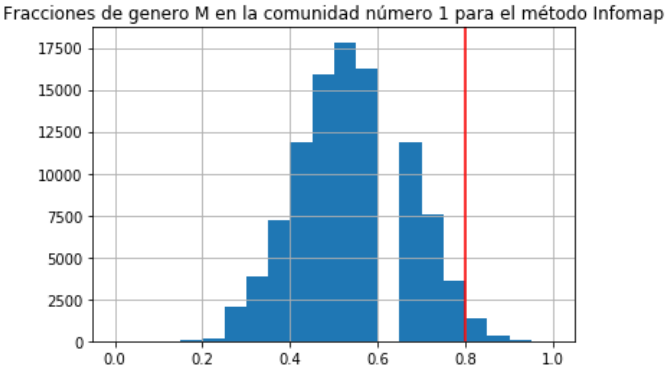
Lo primero que se hizo fue calcular la probabilidad de que al sacar un nodo al azar de la red, este sea de género Macho, Hembra o NA. Estas probablidades resultaron ser:

$$P_M = 0,548 P_H = 0,387 P_{NA} = 0,064$$

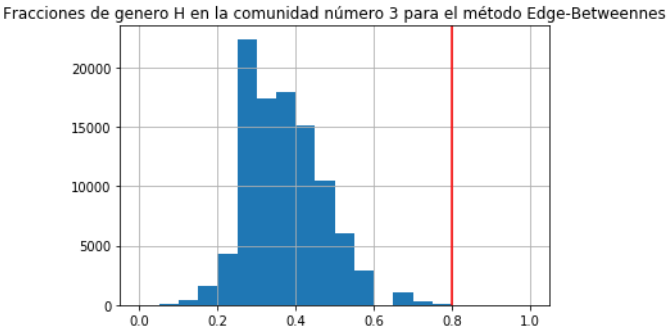
A continuación, se construyeron comunidades de igual tamaño a las originales, asignando género a los miembros de las comunidades de manera aleatoria en función de la probabilidad correspondiente a cada género. Este proceso se realizó 100000 veces para cada comunidad de cada uno de los métodos.

A cada una de estas comunidades creadas aleatoriamente se le calculó la fracción de delfines machos, fracción de delfines hembras y fracción de delfines NA para luego poder armar un histograma.

Finalmente, se comparó la distribución de fracciones obtenidas en los histogramas contra la fracción de delfines de cada género en la red real y se pudo observar que en muchas comunidades la fracción real de delfines de alguno de los géneros estaba claramente alejada de la distribución aleatoria. Como ejemplo se muestran las siguientes figuras:



(a) Histograma de fracción de género Macho



(b) Histograma de fracción de género Hembra

Figura 7: Ejemplos representativos de los histogramas obtenidos de las fracciones de género de las comunidades construidas aleatoriamente. La línea roja en vertical representa el valor de la fracción de género para la comunidad real

Se puede observar claramente en la figura cómo la línea roja se encuentra en la zona más extrema de la distribución de fracciones de género.

A continuación se presenta una tabla con tres comunidades por método, para todos los géneros, en la cual se muestran los valores promedio y de desviación estándar de la fracción de género de cada uno de los histogramas, esto es lo que va en la columna de *Aleatorio*, comparado con el valor real de cada comunidad.

En la tabla se encuentran coloreadas las celdas en las cuales el valor real se encuentra alejado del valor promedio por una distancia mayor a la de la desviación estándar. Se utilizó esto como criterio para determinar que en esa comunidad había una distribución de género que se salía de lo esperable para una simple distribución aleatoria, y que por tanto la comunidad presentaba algún grado de homofilia para alguno de los géneros. Vale aclarar que están coloreados los casos en los que el valor real es muy superior al promedio como también aquellos en los que el valor real es mucho menor al promedio.

Método	Género	Comunidad 1		Comunidad 2		Comunidad 3	
		Aleatorio	Real	Aleatorio	Real	Aleatorio	Real
Infomap	Macho	(0.55±0.11)	0.80	(0.55±0.12)	0.18	(0.55±0.14)	0.75
	Hembra	(0.37±0.11)	0.1	(0.39±0.11)	0.81	(0.39±0.14)	0.25
	NA	(0.065±0.055)	0.1	(0.065±0.058)	0	(0.065±0.071)	0
Fast-Greedy	Macho	(0.55±0.10)	0.3	(0.55±0.11)	0.82	(0.55±0.13)	0.6
	Hembra	(0.39±0.10)	0.65	(0.39±0.10)	0.09	(0.39±0.13)	0.34
	NA	(0.064±0.051)	0.04	(0.065±0.052)	0.09	(0.064±0.063)	0.06
Louvain	Macho	(0.55±0.19)	0.57	(0.55±0.18)	1	(0.55±0.12)	0.16
	Hembra	(0.39±0.19)	0.29	(0.39±0.17)	0	(0.39±0.12)	0.83
	NA	(0.064±0.092)	0.14	(0.064±0.087)	0	(0.065±0.058)	0
Edge-Betweenness	Macho	(0.55±0.19)	0.71	(0.55±0.11)	0.76	(0.55±0.11)	0.15
	Hembra	(0.39±0.18)	0.29	(0.39±0.11)	0.14	(0.39±0.11)	0.8
	NA	(0.064±0.092)	0	(0.065±0.053)	0.095	(0.065±0.055)	0.05

Tabla 3: Tabla de valores de fracción de géneros para cada comunidad obtenidos de manera aleatoria en comparación con valores obtenidos de la red real

Observando la tabla 3 se puede notar que los valores promedios en cada género para todas las comunidades dan muy parecidos a la probabilidad original de obtener un nodo de un cierto género. Esto es esperable, ya que fueron construidas respetando estas probabilidades.

Si bien cada método tenía más de tres comunidades, se reportan sólo 3 para no hacer el cuadro muy extenso y ya que con sólo tres se puede representar el comportamiento de la red para cada partición.

Se puede observar que en la mayoría de las comunidades uno de los géneros predomina, teniendo un valor de fracción que supera ampliamente al promedio, mientras que los otros dos quedan relegados a valores menores al promedio.

En los casos del género NA en los cuales la desviación estándar resulta mayor al promedio, se atribuye esto al hecho de que la cantidad de delfines de género NA en la red real es muy chica, pero al construir las comunidades aleatorias podría haber sucedido que en muchas haya habido una cantidad de delfines NA de manera que la campana de distribución haya quedado muy ancha.

Con este test se puede concluir que las comunidades formadas por los delfines responden en su mayoría a algún grado de homofilia en la variable género, donde miembros del mismo género interactuaban más entre ellos. Un caso extremo de esto se ve en la comunidad 2 del Método del Algoritmo de Louvain, donde la fracción real de Machos para la comunidad es 1.