Trabajo Computacional 3

Estudio comparativo de comunidades

Heli Magalí García Álvarez, Juan Ignacio Gossn, Santiago Scheiner

Resumen

En este trabajo se estudió la estructura de comunidades de una red social de 62 delfines de Nueva Zelanda mediante cuatro métodos de partición diferentes. Los cuatro métodos se compararon entre sí y se analizó si la red podía ser considerada modular, a partir de la comparación con redes recableadas de manera aleatoria. Por último, se produjo un método para el reconocimiento de comunidades basado en el procedimiento de percolación de cliques.

Los resultados mostraron consistencia entre los algoritmos implementados con leves diferencias en las comunidades halladas, que pudieron adjudicarse a la naturaleza de los algoritmos utilizados. Pudo determinarse también que la red analizada presenta características compatibles con las de una red modular. Para finalizar, pudieron determinarse con éxito las comunidades correspondientes al método de percolación de cliques y discutir un criterio para determinar los individuos más sociables de la red.

1. Introducción

En las redes reales resulta usual encontrar una estructura de comunidades en las que pueden ser agrupados sus nodos, dentro de las que los elementos se hallan densamente conectados entre sí. Determinar la estructura interna de una red mediante una división en comunidades permite crear un mapa a escala de la red original, dado que las comunidades tienden a comportarse como nodos en sí mismas, lo que simplifica notablemente su estudio [1].

En este trabajo se estudió la estructura de comunidades de una red social de 62 delfines de Nueva Zelanda [2] mediante cuatro métodos de partición diferentes: *Louvain*, *Fast-greedy*, *Infomap* y *Newman-Girvan*. La información que brinda la red consiste en los 62 nodos, que representan a cada uno de los individuos de la comunidad de delfines, los enlaces que se dan entre ellos, y el género de cada individuo.

El método Louvain es un algoritmo iterativo que tiene como objetivo central optimizar en cada iteración la modularidad, de modo de obtener el mejor forma de agrupamiento de los nodos [3]. El método consiste en la generación de comunidades provisorias a partir de la maximización de la modularidad de forma local hasta agotar las posibilidades de optimización. Una vez alcanzado este estado, cada comunidad local es colapsada en un único nodo, lo que resulta en una red nueva, a la que vuelve a aplicársele el algoritmo. El proceso global finaliza cuando no es posible comenzar un nuevo ciclo, lo que significa que ya se han encontrado las comunidades definitivas.

De igual manera, el método Fast-Greedy también tiene como objetivo la maximización de la modularidad [4]. En el caso de este algoritmo, al igual que todos los algoritmos de tipo *greedy*, la estrategia de maximización es otorgar un valor siempre creciente de la función de rendimiento en cada paso local (es decir, la modularidad en el paso i-ésimo es mayor o igual que en el paso *i-1*-ésimo) con la esperanza de llegar a una solución general óptima. Este tipo de algoritmos se caracteriza por tener un tiempo de ejecución menor, pero con una notable pérdida de precisión.

Por otro lado, el algoritmo Infomap propone una manera completamente diferente de atacar el problema de la delimitación de las comunidades de una red [5]. El método consiste en describir de la manera más eficiente posible la trayectoria de un caminante aleatorio sobre la red. El concepto está basado en el etiquetado de cada nodo de la trayectoria mediante el código de compresión de Huffman [6] según la probabilidad de visita. Utilizando un etiquetado de dos niveles (uno para la entrada y salida de una sección y uno interno de cada una), es posible reutilizar las etiquetas de los nodos y reducir las descripciones de las trayectorias. De este modo, las secciones que minimizan la descripción de la trayectoria son las comunidades de la red.

Por último, el método de Girvan–Newman consiste en un algoritmo de detección de comunidades a partir de la remoción progresiva de enlaces entre nodos cuyo valor de *edge-betweeness* (número de trayectorias de distancia mínima que atraviesan el enlace) resulte máximo [7]. De esta manera, las comunidades son los conjuntos de nodos que quedan desconectados luego de un número dado de enlaces eliminados. En un primer paso, se mide la centralidad de los enlaces y se elimina el de máxima centralidad. Luego se calcula nuevamente la centralidad y se repite el método hasta agotar los enlaces de la red y obtener únicamente nodos aislados. Con este método es posible determinar diferentes estados intermedios y por lo tanto, construir un dendrograma completo de la partición de la red en comunidades.

Este trabajo se centró en la caracterización de las particiones obtenidas con los cuatro métodos detallados arriba utilizando como parámetros la modularidad y *silhouette* media[8] de cada método. Los resultados se compararon también con los valores esperados para redes recableadas al azar. Para esto se utilizó el método de Maslov-Milo, que mantiene la distribución de grado de los nodos pero rompe la estructura local de las redes mediante la implementación de intercambios de manera iterativa de las conexiones entre pares de nodos, tomados al azar [9] [10]. La comparación con las redes reorganizadas al azar permitió analizar si la red estudiada podía ser considerada como modular.

El acuerdo entre las diferentes particiones obtenidas fue caracterizado cuantitativamente, así como también la relación entre el género de los delfines y la estructura de comunidades del grupo. Para esto último se utilizó el test exacto de Fisher.

Finalmente, se implementó un algoritmo de reconocimiento de comunidades basado en la metodología de percolación de *k-cliqués* [11]. El método consiste en: i) identificar todos los cliqués de grado k embebidos en la red; ii) asociar cada k-cliqué hallado a un nodo de una meta-red; iii) establecer la existencia de enlace entre cada nodo de la meta-red si los k-cliqués representados son adyacentes ente sí, es decir, comparten k-1 nodos, y por último, iv) establecer cada comunidad en la red original como la asociada a cada uno de los subgrafos de la meta-red.

A partir de este último resultado, se estudió qué individuos de la red resultaron los más sociables de la comunidad.

2. Resultados y discusión

2.1. Análisis mediante los algoritmos de clustering de Louvain, Fast-Greedy, Infomap y Newman-Girvan

Como se explicó en la introducción de este trabajo, para comenzar el análisis de las comunidades se aplicaron los algoritmos de Louvain, Fast-Greedy, Info-Map y Newman-Girvan a la red de delfines. En la figura 1 se muestran los resultados de cada uno de ellos.

Es importante observar que el algoritmo de Newman-Girvan brinda un dendrograma con todas las cantidades de particiones compatibles con la cantidad de nodos de la red (es decir, desde una única comunidad, hasta una comunidad por nodo cuando la red está totalmente desarmada). A fin de posibilitar la comparación de las comunidades detectadas, para este

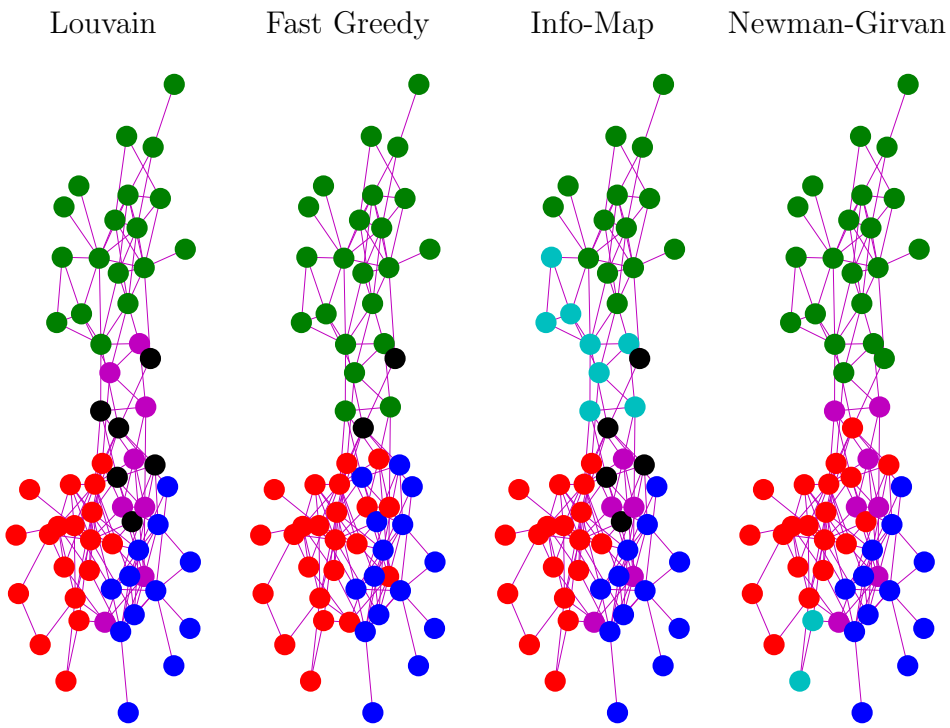

Figura 1: Comparación de las comunidades obtenidas para la red de 62 delfines mediante los cuatro algoritmos empleados. Tanto en esta figura como en las restantes, la distribución espacial de los nodos en los grafos corresponde al método de Kamada-Kawai.

algoritmo se escogió la fase de cinco comunidades, que es la misma cantidad obtenida mediante el método de Louvain, y se corresponde con un máximo de modularidad y *silhouette* media (local) para la red (ver figura 2)

Lo primero que puede observarse es la presencia de tres comunidades que conservan su estructura de manera considerable para los cuatro métodos (graficadas en verde, rojo y azul). Puede verse además que el método Fast-Greedy arroja una comunidad menos que los métodos de Louvain e Infomap. Este fenómeno puede adjudicarse a que, como ya se mencionó, el método Fast-Greedy maximiza la modularidad de manera local, lo que no siempre lleva a una optimización global de la modularidad.

En la figura 2 se muestra la modularidad y la *silhouette* media obtenidas para la red mediante los cuatro algoritmos utilizados.

Como se ya indicó, existen múltiples valores para el método de Newman-Girvan, que corresponden a cada uno de los estados intermedios de partición de la red en comunidades, mientras que para los otros tres, existe un único valor. Lo que puede apreciarse entonces es que para los métodos de Louvain, Fast-Greedy e Infomap la modularidad de la red alcanza su máximo para un valor alrededor de 5 comunidades.

Por otro lado, el método de Newman-Girvan alcanza un máximo de modularidad en las particiones en 5, 6 y 7 comunidades, y un máximo local para la *silhouette* media en la partición en 5 comunidades. Este comportamiento refuerza la presunción (mencionada ya para la figura 1) de la existencia de 5 comunidades en la red. Resulta interesante notar que en ambos gráficos el método de Fast-Greedy es, de los cuatro algoritmos, el de peor rendimiento en cuanto a la

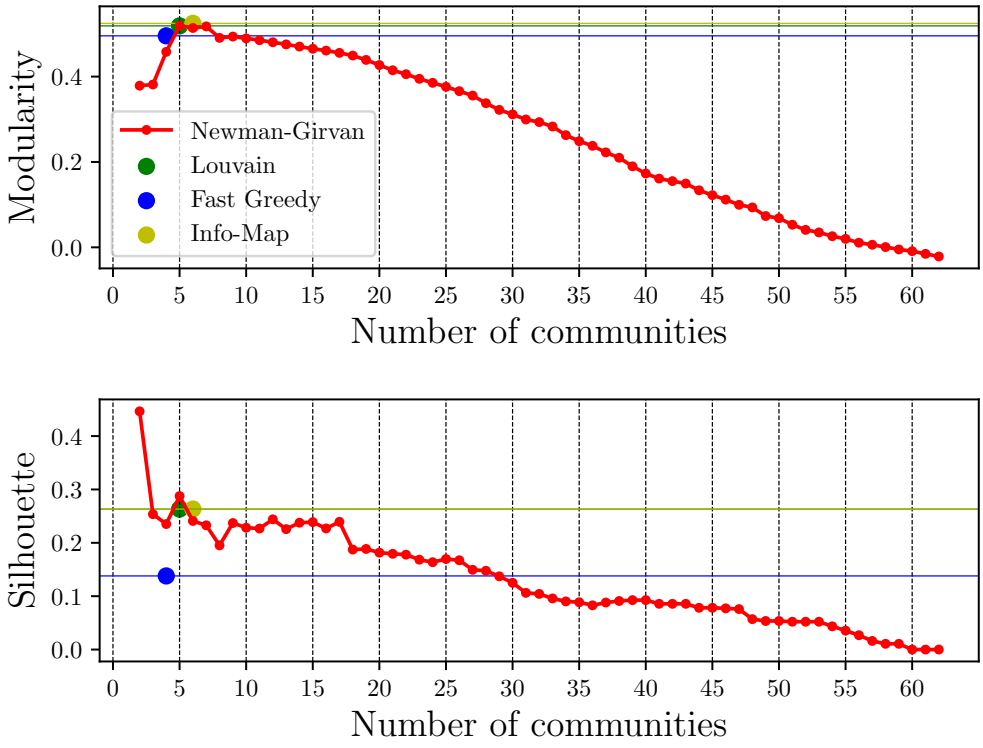

Figura 2: Modularidad (arriba) y *silhouette* media (abajo) en función del número de comunidades halladas por cada método. Se observa que los métodos de Info-Map y Louvain y Newman-Girvan (para 5 comunidades) son los más eficientes en términos de dichos indicadores.

precisión del resultado o partición final obtenida. En otras palabras, si bien el número de comunidades que se detecta por ese método es similar a los otros, la precisión en la manera en que realiza las particiones en comunidades no es el óptimo, lo que puede verse fácilmente teniendo en cuenta que la modularidad y la *silhouette* media, más particularmente, de este método son inferiores que las correspondientes a los otros tres métodos. En cuanto a los métodos de Louvain e Infomap podemos ver que arrojan valores muy similares de modularidad y *silhouette* media para la red agrupada en 5 y 6 comunidades respectivamente. El método de Newman-Girvan para la red partida en 5 comunidades también arroja valores de los parámetros estudiados muy cercanos a los obtenidos por los métodos de Louvain e Infomap.

En relación con estos conceptos, la figura 3 muestra los valores obtenidos para 2000 redes recableadas de manera aleatoria y la comparación de la red original de delfines estudiada en este trabajo. Todos los gráficos muestran que la red posee una modularidad mayor a la esperada por azar, lo que significa que se está en presencia de una red modular.

Una forma útil de intercomparar la efectividad de cada uno de los métodos de partición es mediante la comparación de la *sillhouette* nodo a nodo (Fig. 4). A partir de dicha figura se puede ver no sólo la silueta media, si no que también se pueden detectar nodos donde la silueta es negativa. Dado que la silueta de cada nodo es equivalente a la diferencia entre la distancia media a la propia partición y a la partición ajena más próxima, un valor negativo indica que dicho nodo se halla más cercano a la comunidad vecina que a la propia comunidad. Si bien en términos de silueta media, el método de Newman-Girvan para 5 comunidades resulta el más óptimo, el mismo presenta 9 nodos con silueta negativa; es decir, 50 % más que en el caso de Info-Map. Es decir, si bien la silueta media de Info-Map es menor que la silueta media de Newman-Girvan (5 com), aquella partición presenta mejor silueta nodo a nodo que esta última.

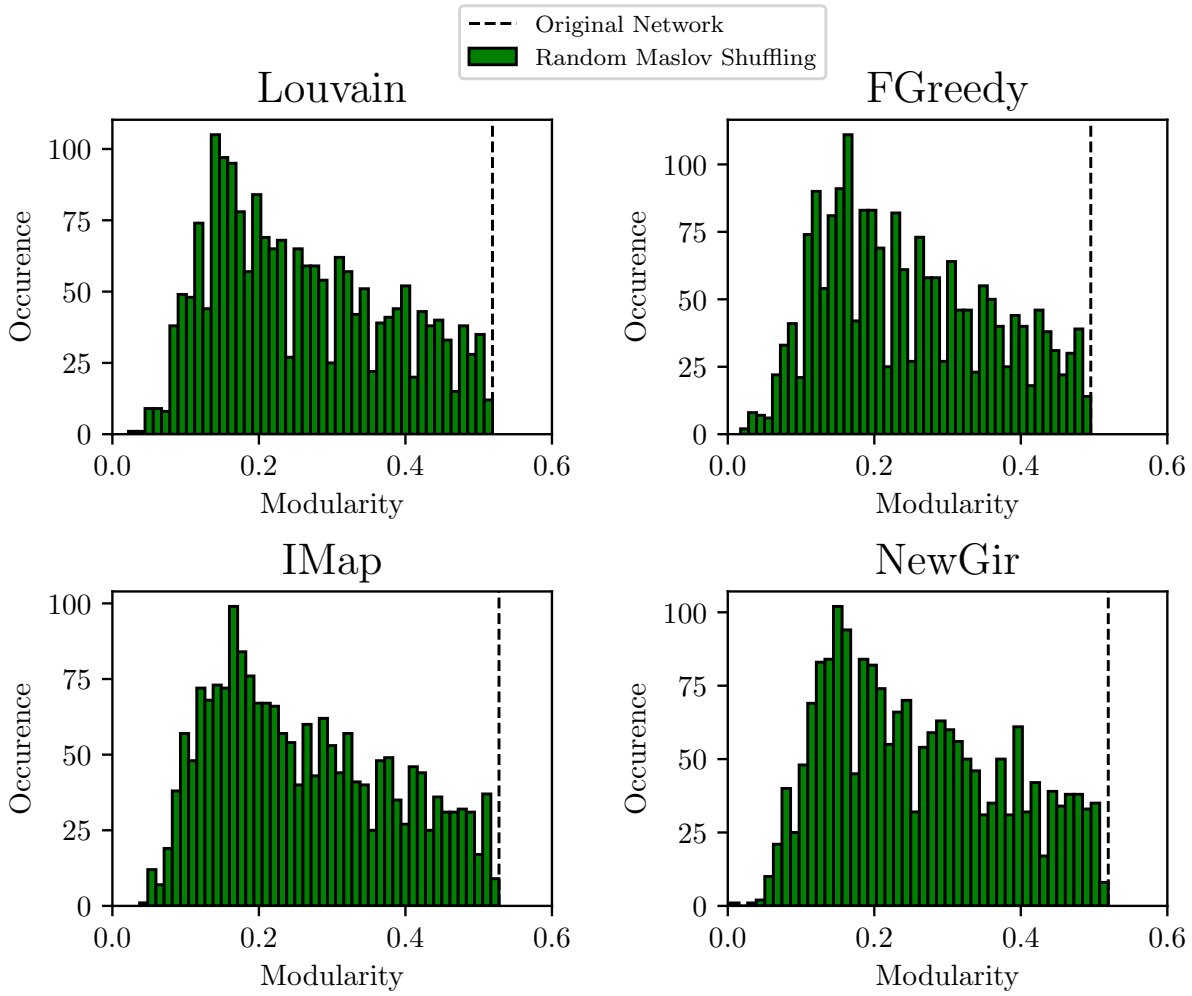

Figura 3: Comparación de modularidades entre redes recableadas (N = 2000) a la Maslov-Milo (histogramas verdes) y red original (líneas punteadas) para los cuatro métodos estudiados. Se observa que la red original es más modular que las redes recableadas en todos los casos.

Por otra parte, como se explicó en la introducción, al tratarse de una red de la que se tiene información de los sexos de los individuos (atributo de los nodos), resulta interesante analizar si las comunidades que se forman en la red responden a sesgos dados por esta característica. La figura 5 muestra el estudio de este fenómeno.

Lo primero que puede observarse es que, en las comunidades más conservadas (graficadas en verde, rojo y azul), se observan los mismos sesgos de género para todos los métodos de partición de redes mostrados. La comunidad verde tiene una clara sobre-representación de machos mientras que la comunidad roja presenta una sobre-representación de hembras en todos los métodos mostrados. En cuanto a la comunidad azul, la misma presenta una sobre-representación de machos para todos los métodos de clustering salvo para Fast Greedy. En principio, debido a que este es el algoritmo que tiene peor performance en cuanto al valor de la modularidad y *silhouette* media de la red final definida, podríamos inclinarnos a pensar que efectivamente la partición azul está mal definida e incluye nodos que deberían pertenecer a otras comunidades, razón por la cual el test de Fisher para la misma dió que esta partición no tiene un sesgo de género cuando en realidad lo debería tener si estuviese "mejor"delimitada (como en los otros métodos). En cuanto a las particiones restantes no es posible realizar una comparación entre los distintos métodos ya que las mismas no están conservadas. Estas poseen menos nodos y para algunas de ellas también se observan sesgos de género, por ejemplo para la comunidad magenta en Louvain se observa sobre-representación de hombres mientras que para la comunidad tam-

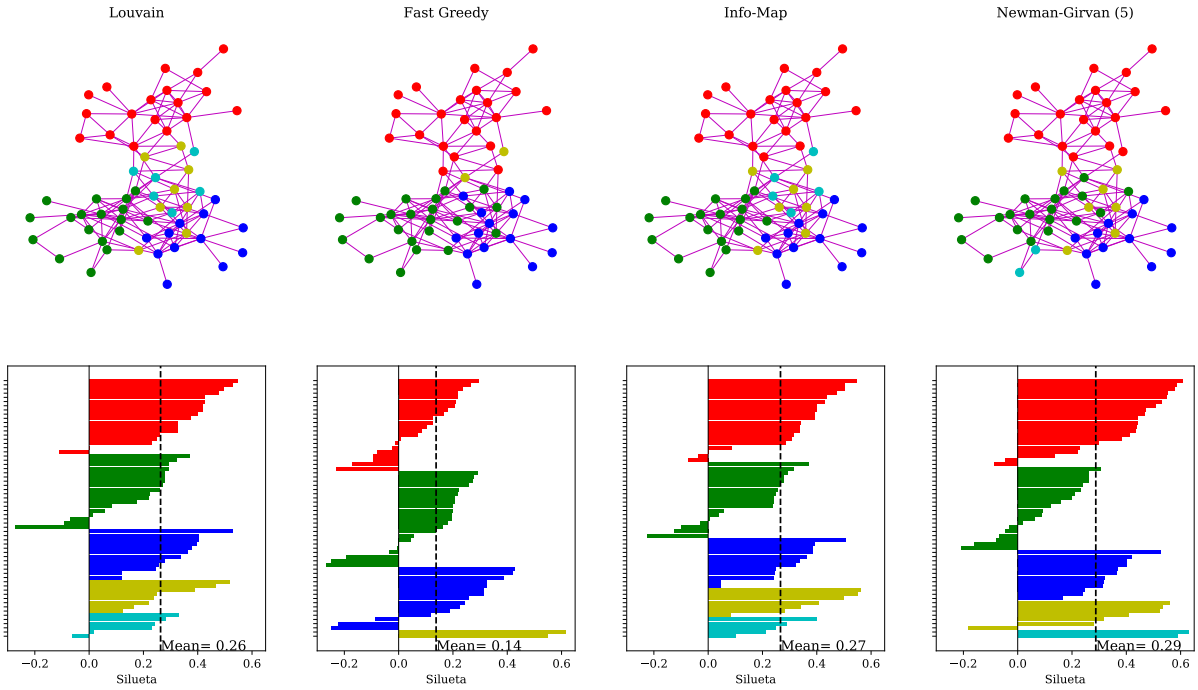

Figura 4: Silueta nodo a nodo para las particiones otorgadas por cada uno de los métodos contemplados. Arr.: grafos de la red delfines distribuido espacialmente mediante el método de Kamada-Kawai, cuyos nodos fueron coloreados según las particiones halladas por cada método. Abj. Silueta nodo a nodo, ordenada en forma creciente por comunidades.

bién coloreada de magenta en el método Info-Map se observa sobre-representación de mujeres. En suma, que para las comunidades definidas en forma más consistente según los diferentes algoritmos de clustering empleados haya sesgos de género, ya sea sobre-representación de un género o lo que es lo mismo sub-representación del otro género, concuerda con los resultados del trabajo computacional 1 en el cual se halló que la red era homofílica.

2.2. Análisis mediante el método de percolación de cliques

Para finalizar el trabajo, se generó un algoritmo de reconocimiento de comunidades basado en el método de percolación de cliques.

La figura 6 corresponde al resultado de la implementación de este método.

Resulta apreciable que la cantidad de nodos de la red que forman parte de 3-cliques es mayor que la que forman parte de 4-cliques, y estos a su vez aparecen en mayor número que los que forman 5-cliques. Esto puede entenderse fácilmente si se tiene en cuenta que un k-clique contiene k (k-1)-cliques en su interior.

A partir de este método puede estudiarse qué individuos son los más sociables de la red.

Una de las maneras de responder esta pregunta es detectando los individuos que se encuentren en comunidades más grandes. De esta manera, se entiende la sociabilidad de un individuo por la cantidad de individuos que están en su grupo más cercano y no por sus enlaces directos.

Por otro lado, no es absurdo argumentar que los individuos más sociables de la red son simplemente los que formen más enlaces con otros individuos, lo que en términos de la red significa que sus nodos tienen los grados más altos de la red.

El cuadro 1 muestra una comparación de estas dos maneras de responder a la pregunta.

Para la confección de esta tabla, se compararon todas las comunidades obtenidas y se las ordenó en orden decreciente de número de nodos. De esta manera, la comunidad 1 es la más grande (con 25 nodos) y la comunidad 2 es la segunda más grande (con 13 nodos). En la tabla

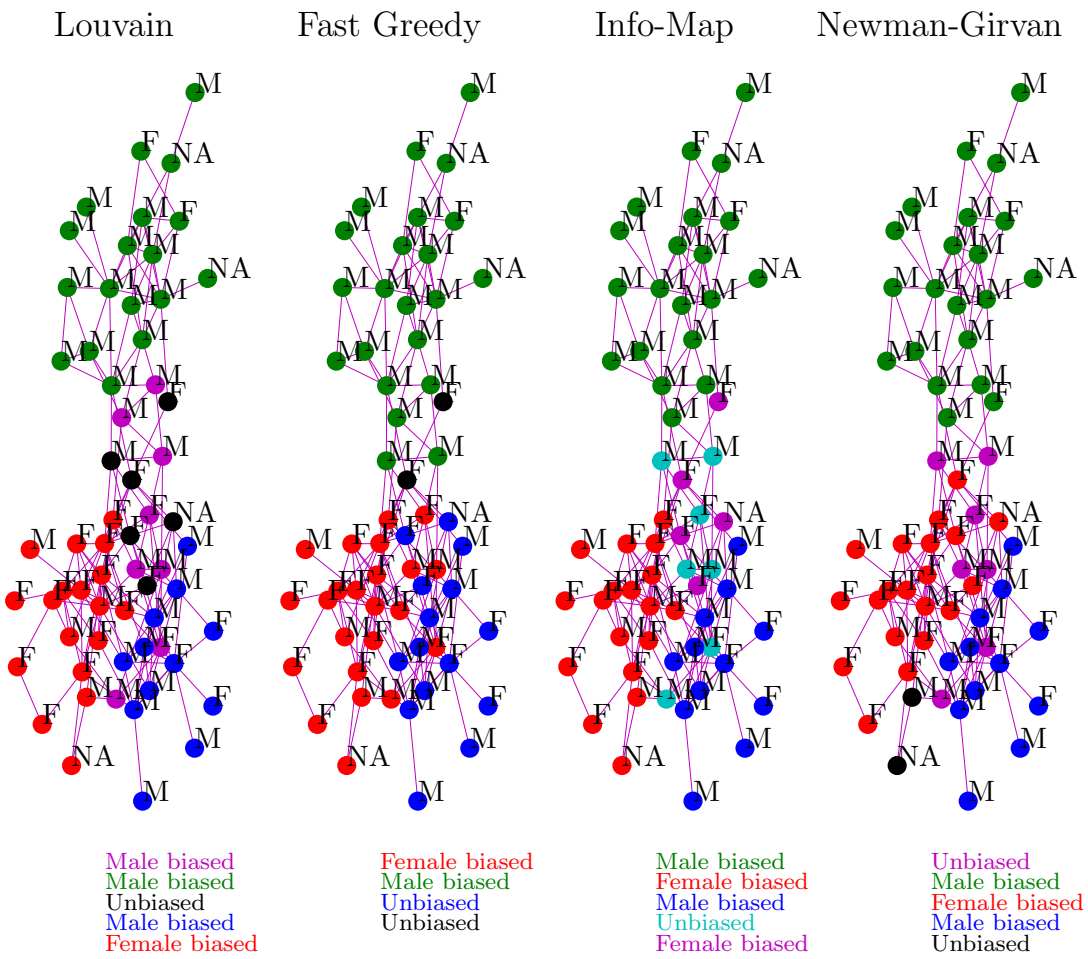

Figura 5: Determinación de la existencia de sesgos en el género de los miembros de cada comunidad de cada método de partición a partir del test exacto de Fisher. El p-valor considerado para rechazar la hipótesis nula (no sesgo) es de 5 %.

se muestran los diez nodos con mayor grado de la red y la comunidad a la que pertenecen.

Lo que se ve es que los ocho nodos con mayor grado de la red pertenecen a la comunidad 1, que es la mayor comunidad presente en la red, y los dos siguientes pertenecen a la segunda comunidad.

Este resultado sugiere entonces que existe una correlación entre las dos maneras sugeridas de analizar la sociabilidad de los individuos, ya que los que tienen mayor grado, son también los que pertenecen a las comunidades más grandes.

3. Conclusiones

Se estudió la estructura de comunidades de una red social de 62 delfines de Nueva Zelanda mediante cuatro métodos de partición diferentes.

Los resultados de los cuatro métodos fueron comparados entre sí utilizando la maximización de modularidad y *silhouette* media y se concluyó que todos los algoritmos resultaron consistentes para la identificación de cuatro a cinco comunidades. En esta misma línea, se comparó la estructura de la red contra 2000 redes aleatorias conformadas según el método de Maslov-Milo, y se determinó sólidamente que la red analizada es modular.

Se analizó el comportamiento de la *silhouette* nodo a nodo para cada una de las particiones

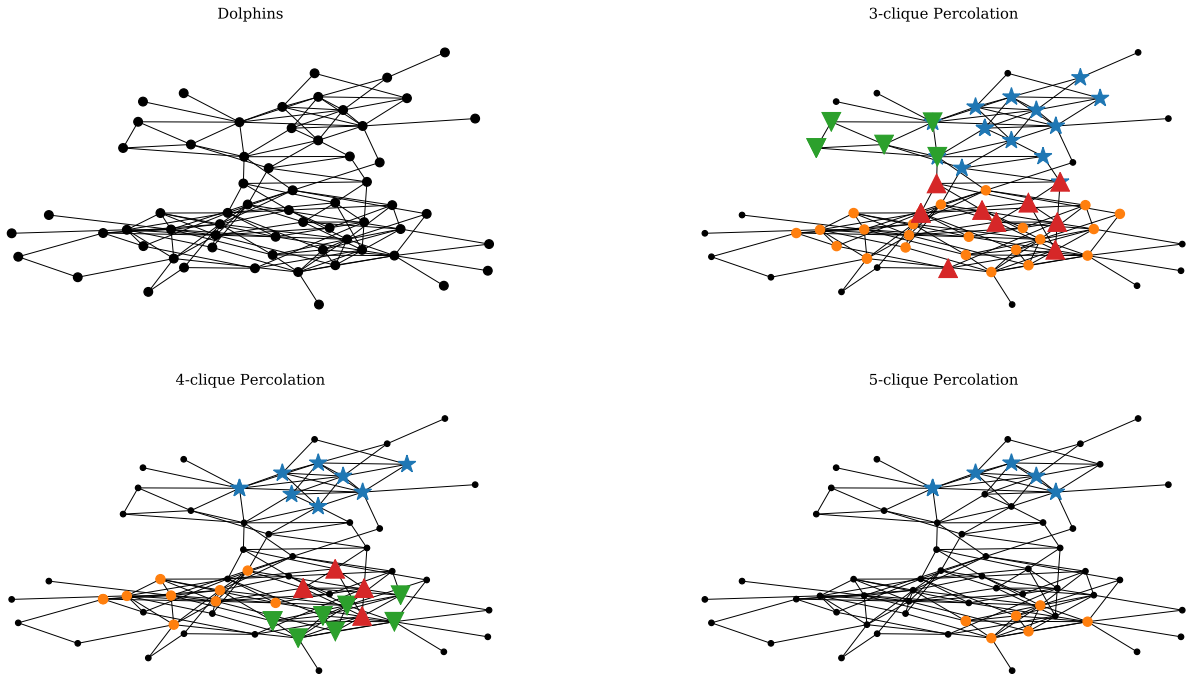

Figura 6: Red de delfines (arr., izq.) junto con la determinación de comunidades a partir del método de percolación de k-cliqués para k = 3 (arr., der.), 4 (abj., izq) y 5 (abj., der.).

Individuo	Grado	Comunidad		Individuo	Grado	Comunidad
Grin	12	1		Jet	9	1
SN4	11	1		Kringel	9	1
Topless	11	1		Patchback	9	1
Scabs	10	1		Web	9	2
Trigger	10	1		Beescratch	8	2

Cuadro 1: Comparación del grado de los diez nodos con más enlaces y la comunidad a la que pertenecen. Las comunidades fueron previamente ordenadas en orden decreciente de cantidad de nodos, lo que significa que la comunidad 1 es la de mayor tamaño.

estudiadias, y se concluyó que, a pesar de que el método Newman-Grivan(5) posee una *silhouette* media levemente mayor que los métodos de Info-Map y Louvain (siendo Fast Greedy el de peor *performance*), estos poseen mejor *sillhouette* local, dado que poseen al menos 50 % menos de nodos con valores negativos.

En todos los métodos estudiados, para las comunidades más numerosas, existe un fuerte sesgo asociado al género. Es decir, las comunidades más numerosas están constituidas mayoritariamente por individuos de igual género. Esto es consistente con la fuerte homofilia que prevalece sobre la red, estudiada en el trabajo computacional TC1.

Para finalizar, pudieron determinarse con éxito las comunidades correspondientes al método de percolación de cliques y discutir un criterio para determinar los individuos más sociables de la red, así como también mostrar la equivalencia entre dos modos de analizar la sociabilidad.

Referencias

[1] NEWMAN, Mark. Finding community structure in networks using the eigenvectors of matrices. Physical review E, 2006, vol. 74, no 3, p. 036104.

[2] LUSSEAU, David, et al. The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. Behavioral Ecology and Sociobiology, 2003, vol. 54, no 4, p. 396-405.

[3] BLONDEL, Vincent D., et al. Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment, 2008, vol. 2008, no 10, p. P10008.

[4] NEWMAN, Mark EJ. Fast algorithm for detecting community structure in networks. Physical review E, 2004, vol. 69, no 6, p. 066133.

[5] ROSVALL, Martin; BERGSTROM, Carl T. Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences, 2008, vol. 105, no 4, p. 1118-1123.

[6] HUFFMAN, David A. A method for the construction of minimum-redundancy codes. Proceedings of the IRE, 1952, vol. 40, no 9, p. 1098-1101.

[7] GIRVAN, Michelle; NEWMAN, Mark EJ. Community structure in social and biological networks. Proceedings of the national academy of sciences, 2002, vol. 99, no 12, p. 7821-7826.

[8] ROUSSEEUW, Peter J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 1987, vol. 20, p. 53-65.

[9] MASLOV, Sergei; SNEPPEN, Kim. Specificity and stability in topology of protein networks. Science, 2002, vol. 296, no 5569, p. 910-913.

[10] MILO, Ron, et al. Network motifs: simple building blocks of complex networks. Science, 2002, vol. 298, no 5594, p. 824-827.

[11] PALLA, Gergely, et al. Uncovering the overlapping community structure of complex networks in nature and society. Nature, 2005, vol. 435, no 7043, p. 814.

¡Muy buen trabajo!Muy completo, bien explicado y con análisis extras además de bue