Trabajo Práctico Computacional 03: Comunidades

Emanuel Ferreyra, Bruno Kaufman, Ariel Salgado

31 de octubre de 2018

1. Introducción

En este trabajo consideraremos nuevamente la red social con 62 delfines de Nueva Zelanda en conjunto con la información de su género. En esta oportunidad estudiaremos las comunidades presentes en la red, utilizando diferentes métodos para particionarla en clusters: `infomap`, `fast_greedy`, `louvain` y `edge_betweenness`.

Luego caracterizaremos las particiones vía modularidad y silouhette y finalmente observaremos la relación entre el género de los delfines y la estructura de las comunidades obtenidas con las diferentes metodologías.

Para esto utilizaremos el paquete `igraph` de `R`, aprovechando las funciones que nos permiten trabajar con los resultados de detección de comunidades en redes, a través de `membership` y la clase `communities`.

Comenzamos visualizando gráficamente la red de los delfines de cuatro formas distintas, que nos muestran la estructura de comunidades obtenida con los algoritmos `infomap`, `fast_greedy`, `louvain`, y `edge_betweenness`. Se conserva el mismo layout para todos los gráficos, para que se pueda comparar sólamente las comunidades.

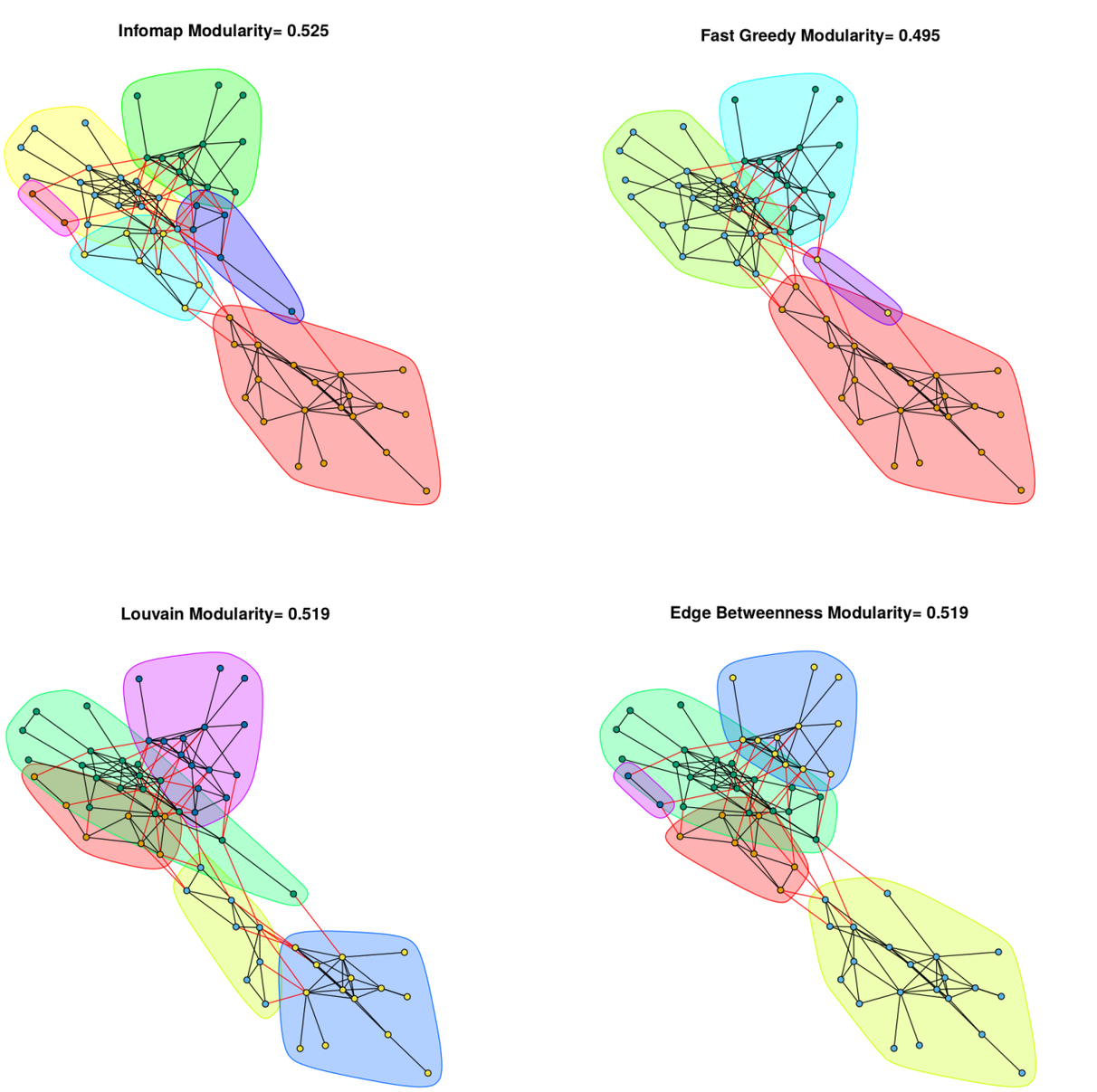

Figura 1: Particiones en clusters obtenidas con las diferentes metodologías

.

Vemos que todas las comunidades tienen valores similares de modularidad, barriendo un rango pequeño alrededor de 0.5. La máxima modularidad se obtiene con algoritmo `infomap`, mientras que tiene su valor más bajo cuando se usa `fast_greedy`. En la tabla 1 se pueden ver resumidas las modularidades de la red frente a las comunidades obtenidas con los distintos algoritmos, asi como el valor de la silhouette media para cada uno. Ambas medidas asignan el valor más bajo a fastgreedy, pero silohuoette prefiere el método de `edge_betweenness` por sobre el resto, lo que muestra que esa es la estructura de comunidades para la cual la cercanía entre miembros de una comunidad se ve más enfatizada.

	infomap	fast_greedy	louvain	edge_betweenness
Modularity	0.525	0.495	0.519	0.519
Silhouette media	0.263	0.138	0.234	0.288

Tabla 1: Modularidad y Silhoutte promedio obtenida para las distintas particiones encontradas por cada método.

Además utilizamos los datos visualizados en la figura 1 para comparar al valor obtenido para la modularidad de cada partición con una distribución de modularidades que serían obtenidas con recableados aleatorios de la misma red. Por la diferencia amplia entre las distribuciones y los valores obtenidos en la red real, podemos concluir que la red es efectivamente modular bajo todas las particiones.

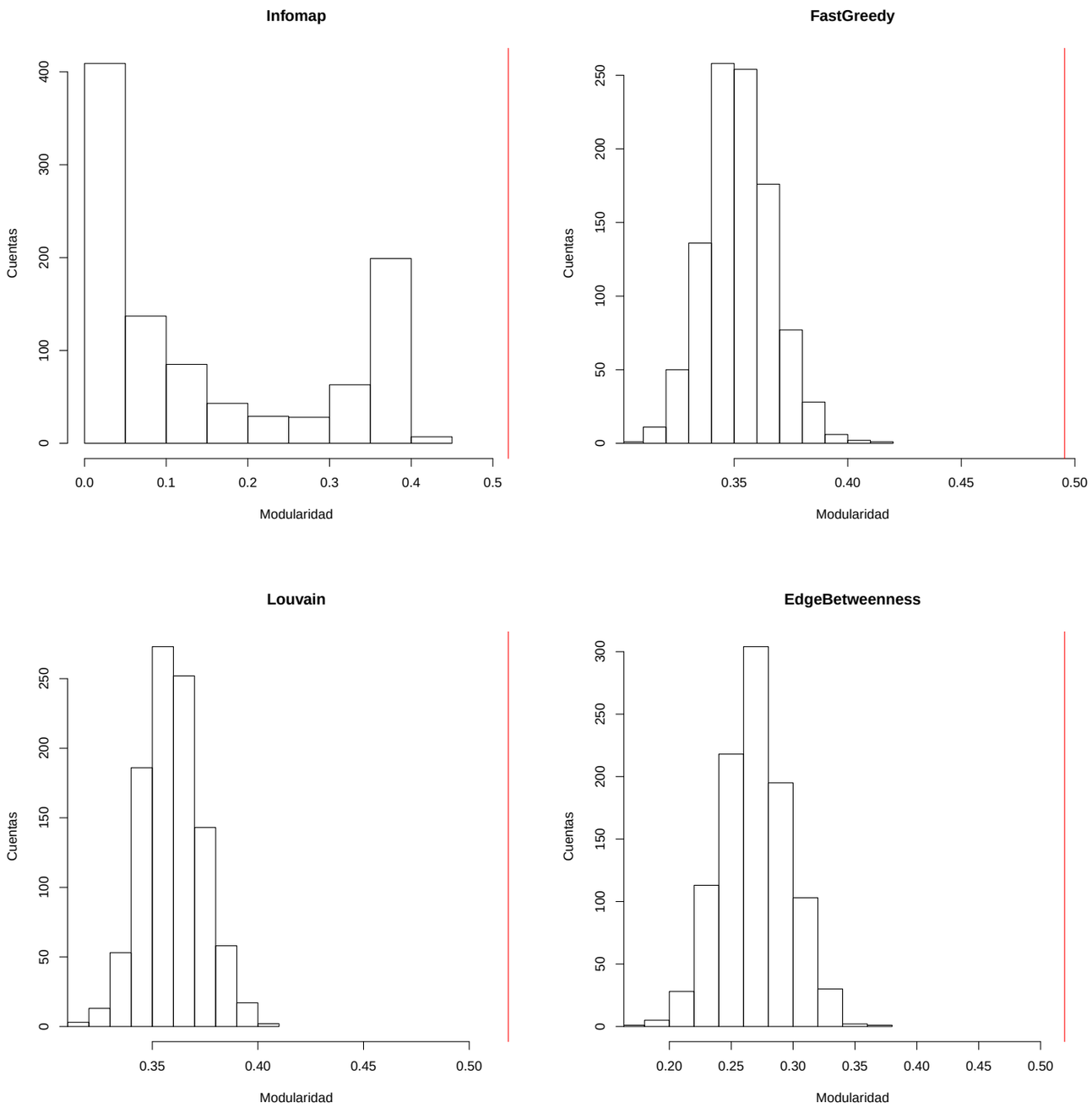

Figura 2: Comparación del valor obtenido para la modularidad de la red bajo cada algoritmo (en rojo) con la distribución de modularidades obtenidas en distintas realizaciones de recableado de red.
.

Algo interesante para notar es que la distribución de modularidades para las comunidades obtenidas con infomap es bimodal. Esto ocurre porque infomap es volátil en decidir cuantas comunidades plantear. Al recablear la red, a veces quedan clusters partido. Infomap asigna toma estos como las comunidades elegidas. Cuando plantea pocas comunidades, la modularidad es menor generando el pico en valores bajos. De acuerdo al recableado de los nodos, el algoritmo tiende a estructuras de comunidad bastante distintas. Esto no sucede en los otros algoritmos al punto de que se note en la modularidad obtenida.

La tabla 1 sólo muestra la silhouette promedio, sin embargo. En la figura 1 se puede ver la distribución de las silhouettes individuales para cada nodo, donde el nombre de cada nodo se ve en el eje x. Para comparar, se dibuja una línea en la silhouette media.

En esta figura, quienes tienen silhouettes altos están muy cerca de los miembros de su comunidad, comparado con los que existen fuera de su comunidad. En cambio, quienes tienen silhouette baja se encuentran más alejados, acentuandose en valores negativos, donde el nodo termina estando más cerca de comunidades ajenas que de la propia; estos nodos podrían entenderse como puentes entre comunidades. Posiblemente por esto haya acuerdo entre el método basado en la betweenness y silhouette.

Es de distinguir que bajo la agupación lograda con el algoritmo `louvain`, hay muy pocos nodos con silhouette negativa. En esta agrupación, las comunidades estan definidas de forma que sus miembros son casi siempre más cercanos entre ellos

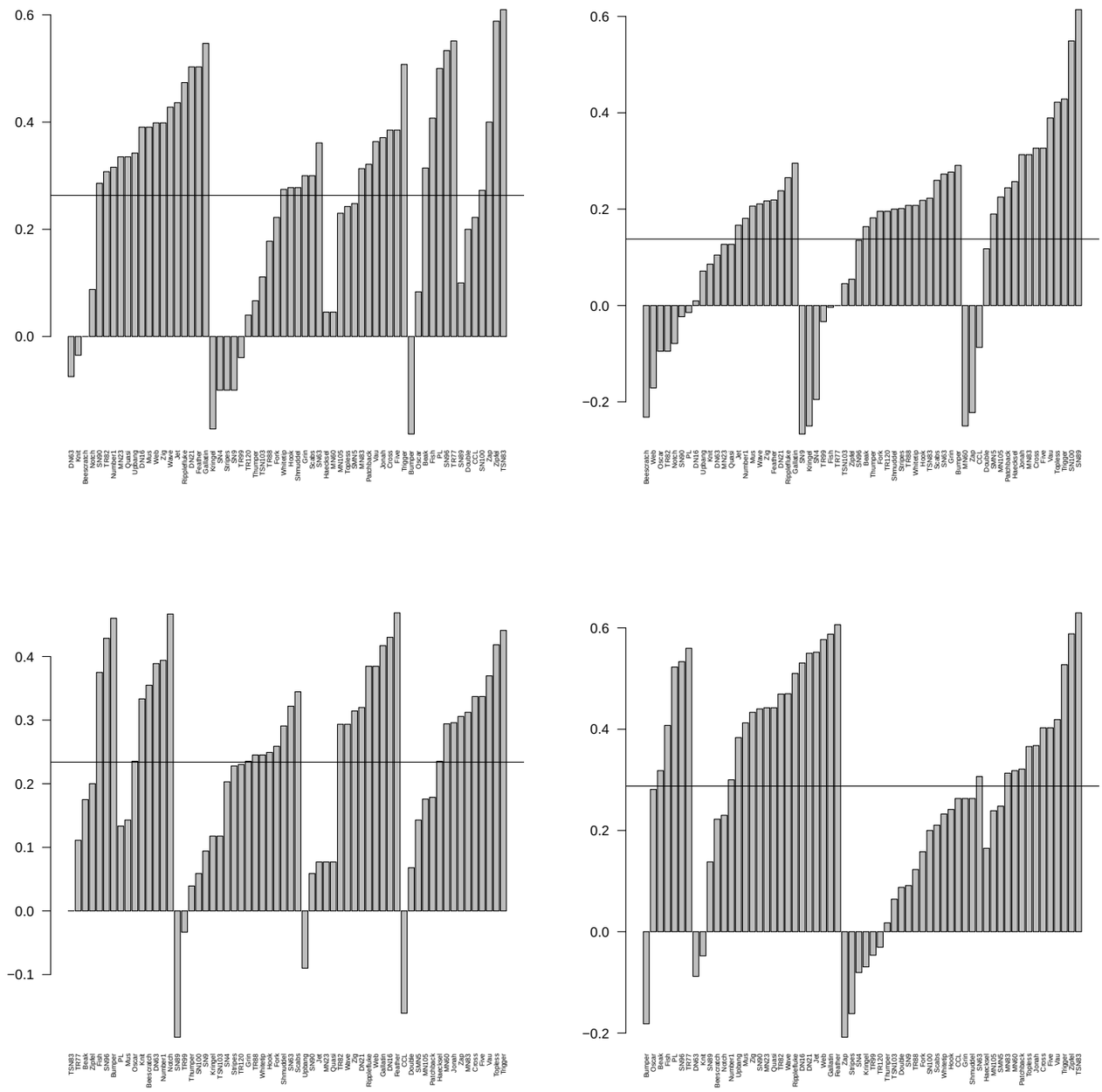

Figura 3: Distribuciones de silhouette para las distintas particiones. De izquierda a derecha y de arriba a abajo: `infomap`, `fast_greedy`, `louvain` y `edge_betweenness`.

.

que de otras comunidades.

2. Relación entre sexo y comunidad

Para analizar la relación entre el género de los delfines y las comunidades reasignamos el sexo a los delfines de manera aleatoria pero conservando las proporciones de cada tipo. En la red hay 24 delfines hembra (f), 34 machos (m) y 4 sin sexo específico (NA). Para cada método, comunidad y sexo calculamos un z-score, restando la cantidad real observada en esa comunidad, con el número medio encontrado en esa comunidad en las reasignaciones, y diviendo por la desviación estándar del número hallado en los distintos sampleos.

A continuación tabulamos, para cada método y cada comunidad, la cantidad de machos, hembras y NA's presentes, y los correspondientes z-scores.

En todos los casos, analizando las cantidades de cada sexo en la comunidad, podemos ver una representación mayoritaria de un sexo en particular. La información de las columnas con los z-scores indican que lo que observamos dista considerablemente de lo esperado aleatoriamente. Esto nos lleva a pensar que los delfines se agrupan más preferentemente con delfines del mismo género. Este comportamiento es consistente con la homofilia observada en la red en el TC01.

Al mismo tiempo, como encontramos 4 o 5 comunidades con todos los métodos, podemos concluir que esta agrupación homofílica igualmente se particiona, pues salvo en la segunda comunidad encontrada por `fastgreedy`, la proporción de delfines de otro sexo al mayoritario es considerablemente pequeña.

3. Comunidades que se solapan

Para explorar un método de detección de comunidades que permitiese considerar comunidades que se solapasen, realizamos una implementación en R del método de percolación de cliques, basandonos en la función `cliques` de `igraph`. Esta función nos permite obtener todos los cliques del grafo con tamaño k. Partiendo de eso, el método de percolación de cliques propone

Comunidad	N_m	N_f	N_{NA}	z_m	z_f	z_{NA}
1	16	2	2	2.73	-3.11	0.83
2	4	13	1	-3.25	3.41	-0.24
3	9	3	0	1.50	-1.07	-1.01
4	5	2	0	0.90	-0.58	-0.76
5	0	4	1	-2.56	1.96	1.21

Tabla 2: Población de cada sexo por comunidad, según las comunidades provistas por Infomap. Se agrega el z-score de cada cantidad.

Comunidad	N_m	N_f	N_{NA}	z_m	z_f	z_{NA}
1	4	2	1	0.11	-0.57	0.90
2	8	0	0	2.74	-2.44	-0.80
3	3	15	0	-3.90	4.57	-1.34
4	10	2	2	1.50	-2.05	1.33
5	9	5	1	0.49	-0.54	0.03

Tabla 3: Población de cada sexo por comunidad, según las comunidades provistas por Louvain. Se agrega el z-score de cada cantidad.

construir una red donde los nodos son los cliques, y se conectan si comparten $k-1$ nodos de la red original. Las comunidades se obtienen asignando una misma comunidad a todos los nodos que se encuentran en los cliques que comparten un mismo cluster. En la figura 4 podemos observar el resultado de la clusterización para 3-cliques. En la figura 5 tenemos el número de comunidades a las que pertenecen en función del grado del delfín. Si bien no es válido para todos los casos, vemos que en general hay una correlación positiva entre el número de comunidades a las que pertenece un delfín y su grado.

En las figuras 6 y 7 se encuentra el resultado de percolar 4-cliques. En este caso el número de comunidades es menor, aunque la conclusión en términos del grado es la misma.

4. Conclusiones

A lo largo del trabajo pudimos explorar distintos métodos de detección de comunidades. Mientras que según la modularidad el método preferido es Infomap, según silhouette el mejor es el basado en la betweenness.

El análisis de comunidades permite observar la homofilia de esta red, al asociar en la misma comunidad a un número mayor al esperado por azar de delfines del mismo sexo. Esto está en acuerdo con lo observado en el trabajo práctico anteriormente realizado sobre la misma red.

Pudimos implementar un método de detección basado en percolación de cliques, y observamos que los delfines con más conexiones tienden a encontrarse en un mayor número de comunidades, indicando que tienden a funcionar como puente entre las comunidades.

Comunidad	N_m	N_f	N_{NA}	z_m	z_f	z_{NA}
1	18	2	2	3.00	-3.65	0.60
2	7	15	1	-2.95	3.22	-0.47
3	9	5	1	0.49	-0.49	0.03
4	0	2	0	-1.5	1.79	-0.40

Tabla 4: Población de cada sexo por comunidad, según las comunidades provistas por Fastgreedy. Se agrega el z-score de cada cantidad.

Comunidad	N_m	N_f	N_{NA}	z_m	z_f	z_{NA}
1	5	2	0	0.94	-0.55	-0.76
2	16	3	2	2.53	-2.88	0.73
3	3	16	1	-4.32	4.67	-0.33
4	9	3	0	1.54	-1.13	-0.99
5	1	0	1	-0.11	-1.20	2.33

Tabla 5: Población de cada sexo por comunidad, según las comunidades provistas por EdgeBetweenness. Se agrega el z-score de cada cantidad.

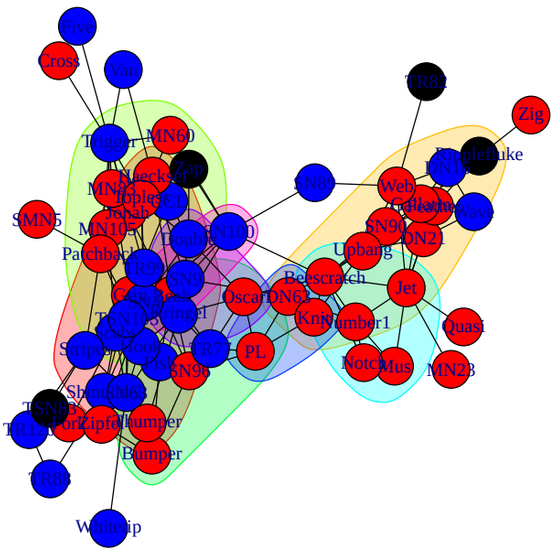

Figura 4: Estructura de comunidades encontrada realizando una percolación de 3-cliques.

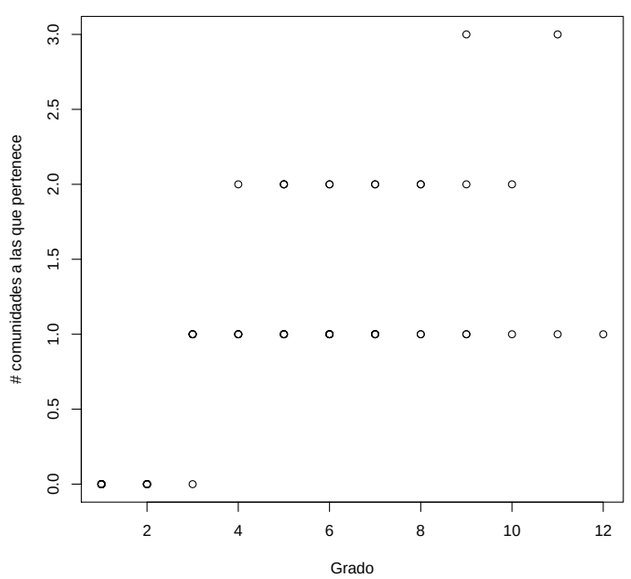

Figura 5: Número de comunidades obtenidas via percolación de 3-cliques a las que un delfín pertenece en función de su grado.

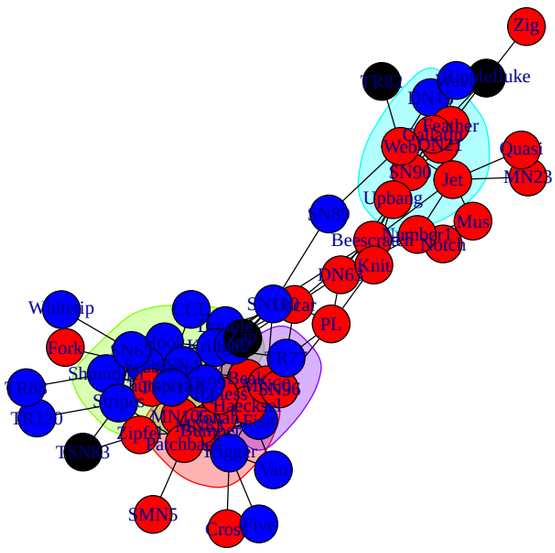

Figura 6: Estructura de comunidades encontrada realizando una percolación de 4-cliques.

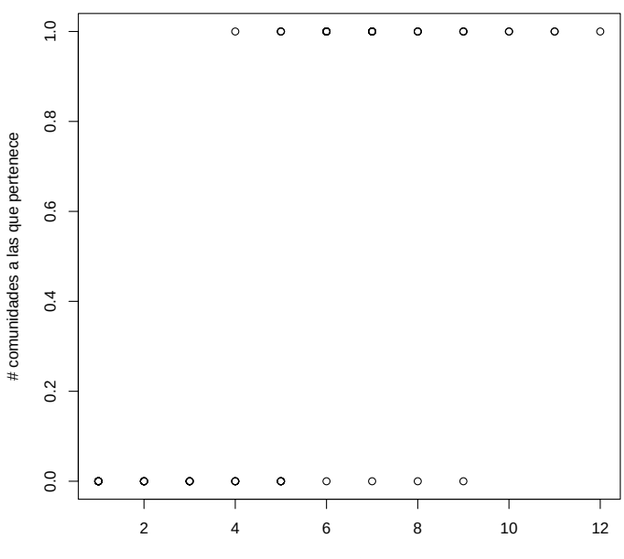

Figura 7: Número de comunidades obtenidas via percolación de 4-cliques a las que un delfín pertenece en función de su grado.