Detección de comunidades en una red de delfines

Andino C. , Asplanato L. , Murchison, F.

Departamento de Física, UBA

Resumen

Se realizó un análisis de comunidades sobre la red social de delfines de Nueva Zelanda mediante la implementación de cuatro algoritmos de partición -*Infomap, Fast-greedy, Louvain y Edge-betweenness*-. Se estudió el acuerdo de las mismas con la red y la proporción de géneros en cada partición, encontrando una correlación de género en dos de las comunidades.

1. Introducción teórica

El análisis de redes permite, entre otras cosas, inferir interacciones entre nodos a partir de la relación topológica entre ellos. Es decir, mediante algoritmos de análisis de vértices y ejes de una red se puede obtener información aproximada del comportamiento de los nodos en una meso-escala, puesto que se tienen relaciones de pertenencia y similitud entre grupos de ellos, llamados *comunidades* de la red.

Cada comunidad es un subgrafo conexo y localmente denso. Se define de acuerdo a algún parámetro determinado según el algoritmo implementado. La *modularidad* parte de la hipótesis que el número de enlaces entre nodos del mismo tipo (pertenencia a la misma comunidad) será mayor que entre comunidades, al ser las mismas grupos de alta densidad de enlaces. La diferencia entre esta cantidad de enlaces intra-comunidades para la red categorizada y el número para la red recableada es la modularidad $'Q'$, donde $0 < Q < 1$ implica que se presenta un mayor número de enlaces del mismo tipo que los esperados por azar, mientras que $Q < 0$ implica menos enlaces a los esperados por azar. El *silhouette* da un valor de concordancia de cada partición con la naturaleza intrínseca de la red analizando las distancias (camino mínimo entre un nodo y el otro) entre nodos de distintas comunidades en relación al tamaño característico de la comunidad de pertenencia. Con valores análogos a la modularidad, se considera que una partición es adecuada si el valor de silhouette o modularidad es cercano a 1.

El algoritmo *Infomap* [1] determina las particiones a partir de una caminata al azar dentro de red. Se parte de cierto nodo semilla y se camina aleatoriamente a nodos vecinos, etiquetando a cada uno con un dado código de acuerdo a la probabilidad de visita de un caminante

aleatorio en el mismo. Se entiende que, si una comunidad presenta una conectividad local mayor implicaría que el caminante perduraría mayor tiempo en nodos de la misma. La partición óptima se elige a partir de la minimización de la llamada "*Map Equation*"de las particiones posibles para esa red. *Louvain* [2] parte de la red original y agrupa nodos en la misma categoría buscando maximizar la modularidad, la siguiente iteración recibe la red con estas comunidades como 'nodos' y así sucesivamente hasta llegar a máxima modularidad. *Fast-greedy* [3] realiza un proceso similar hasta llegar a una única comunidad, pero optimiza la velocidad del algoritmo aprovechando simetrías y características de las redes para disminuir los cálculos. Mientras estos métodos son aglomerativos, es decir, parten de la red y agrupan en las comunidades, *Edge-betweenness*, o *algoritmo de Girvan Newman*[4] es divisivo, puesto que parte considerando toda la red como una entidad, y parte a la misma eliminando el eje con mayor valor de centralidad *edge-betweenness*, de allí el nombre, hasta tener tantas clasificaciones como nodos. Se calcula la modularidad a cada paso y se selecciona la partición que la maximice.

Se pueden utilizar distintos criterios para comparar dos particiones obtenidas con métodos distintos. Un criterio posible se basa en comparar la información de cada red a partir de calcular la *información mutua* (IM). Si se tienen dos fuentes de un mismo tipo de información, el criterio de información mutua sirve para determinar qué tanto se conoce sobre una fuente dado el conocimiento sobre la otra. La IM entre dos comunidades C_1 y C_2 está dada por:

$$IM = \sum_{C_1} \sum_{C_2} P(C_1, C_2) log \frac{P(C_1, C_2)}{P(C_1)P(C_2)}$$

Al máximo valor de información mutua se lo conoce como entropía de Shannon, y ocurre cuando ambas fuentes son iguales. Para una dada comunidad C_i la entropía de Shannon se define como:

$$S = -\sum_{C_i} P(C_i) log P(C_i)$$

2. Resultados y análisis

La red social de delfines de Nueva Zelanda tiene 62 nodos (delfines) enlazados si interactúan entre sí de alguna manera. Los enlaces no son pesados, por lo que, a menos que se haya tenido en cuenta frecuencia de contacto como umbral para determinar un enlace, en principio todos ellos tienen el mismo peso y son no-dirigidos.

Se realizaron las particiones con cada uno de los algoritmos detallados en la sección 1 para inferir las comunidades presentes en estos animales. Los resultados pueden verse en la figura 1. Los algoritmos determinaron un número de comunidades similares, con Fast-greedy siendo el único método que tuvo 4 en lugar de 5 grupos. Sin embargo, como se puede ver en la figura 2, varía el número de nodos por comunidad a lo largo de las particiones, y los nodos no son necesariamente los mismos. Esto puede apreciarse en la figura 1, donde hay alrededor de 5

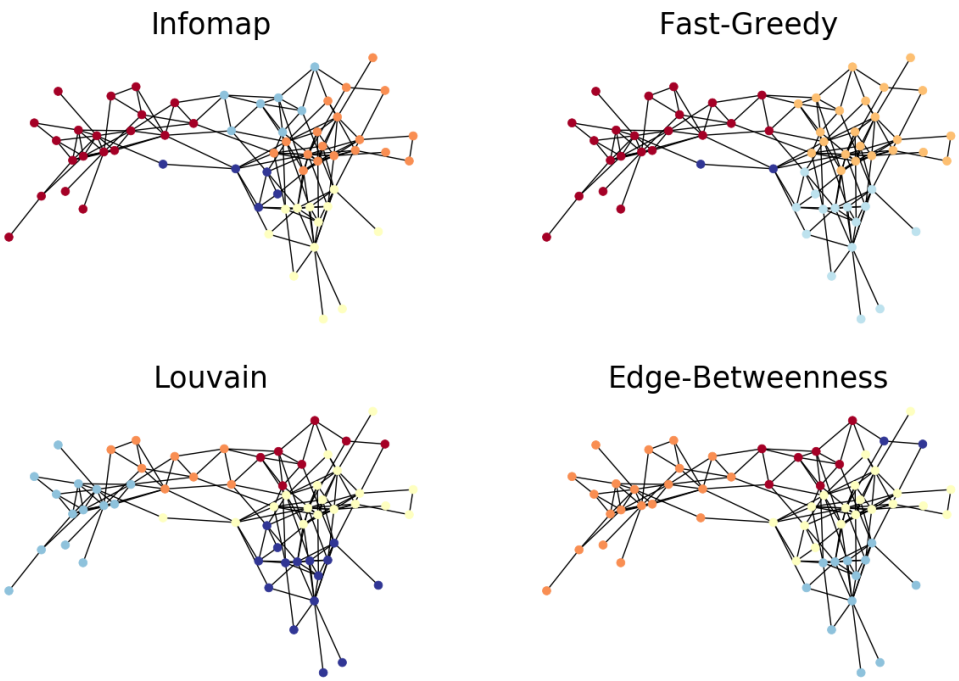

Figura 1: *Partición en comunidades de la red social de delfines según los algoritmos Infomap (5), Fast Greedy (4), Louvain (5) y Edge-betweenness (5)*

regiones de nodos que permanecen aproximadamente constantes a través del análisis con los distintos algoritmos, y es en los límites de las comunidades donde aparece la mayor variación en la clasificación de los algoritmos. Para caracterizar estos cambios de manera cuantitativa se realizó un estudio de superposición de nodos entre comunidades. Tan sólo 3 de las 5 (4 en el caso de Fast-greedy) comunidades de cada método tienen al menos un nodo compartido entre todos los métodos (ver Apéndice A, tabla 3). Además, se buscó la coincidencia promedio de nodos por comunidad de un algoritmo comparado con los métodos restantes (ver Apéndice A, tabla 4).

Dadas las cuatro particiones ilustradas, se caracterizó la bondad de las mismas con los parámetros modularidad y silhouette. Para determinar su relevancia y poder determinar si la red social de delfines es modular por naturaleza se la recableó 2100 veces, manteniendo la distribución de grado en la misma, y se realizó el mismo análisis de comunidades cada vez. Los valores de modularidad y silhouette promedio se muestran en la tabla 1, donde el error reportado es la desviación estándar de los datos. Se realizaron histogramas de los parámetros obtenidos para las redes recableadas (ver Apéndice B).

A partir de los valores reportados, se determinó que la red de delfines es modular ya que cada Q obtenido supera al esperado en una red recableada de manera aleatoria. *Fast-greedy* posee la menor modularidad de los métodos empleados, y se solapa con la modularidad del azar al considerar su error. Este método determinó la existencia de 4 comunidades en lugar de 5, por

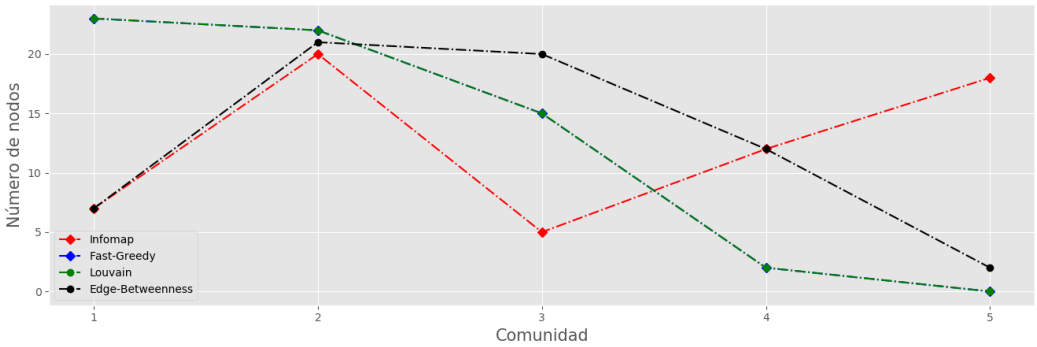

Figura 2: *Número de nodos por comunidad para las cuatro particiones realizadas.*

Modularidad				
	Infomap	Fast - greedy	Louvain	Edge-betweenness
Red	0,53	0,50	0,52	0,52
Azar	$0{,}18 \pm 0{,}15$	$0{,}35 \pm 0{,}2$	$0{,}36 \pm 0{,}01$	$0{,}27 \pm 0{,}03$

Silhouette				
	Infomap	Fast - greedy	Louvain	Edge-betweenness
Red	0,32	0,18	0,29	0,34
Azar	$0{,}23 \pm 0{,}05$	$0{,}22 \pm 0{,}02$	$0{,}23 \pm 0{,}03$	$0{,}33 \pm 0{,}06$

Tabla 1: *Caracterización de las particiones obtenidas en la red de delfines. Comparación con valores obtenidos en* 2000 *recableados al azar de la red.*

lo que nodos que formaban una comunidad por sí mismos pueden haber sido categorizados dentro de otras comunidades, disminuyendo el valor de la modularidad al transformar un número de enlaces de la misma categoría a enlaces entre comunidades. El parámetro silhouette no fue determinante; el valor mínimo, de la partición *Fast-greedy*, fue menor que el esperado por azar y el algoritmo *Edge-betweenness* presenta un valor igual al esperado por el azar.

Con el objetivo de analizar la similaridad entre las comunidades contenidas en dos particiones distintas, se realizó un estudio comparativo con la información como parámetro. La información mutua (IM) se normalizó por la entropía de Shannon, y de esta manera, un 100 % de coincidencia de la IM de dos particiones indica que son iguales.

La entropía de Shannon es una característica de una partición, por lo que existen tantas normalizaciones de la IM como algoritmos de reconocimiento de comunidades utilizados. La tabla 2 muestra los porcentajes de similitud entre cada par de particiones.

En cada una de las cuatro columnas figura la IM entre todas las particiones, normalizando por la ES de un método de referencia dado. La primera columna representa la similitud de cada partición a la obtenida por el método *Infomap* y así respectivamente. El *"porcentaje promedio"* permite caracterizar la partición que posee mayor cantidad de IM con el resto. La partición resultante de aplicar el método *Fast-greedy* es aquella con mayor porcentaje de

Porcentaje de información mutua			
Infomap	89,8 %	72,9 %	89,9 %
71,6 %	Fast-greedy	70,3 %	61,1 %
75,7 %	91,5 %	Louvain	77,0 %
84,8 %	72,3 %	69,9 %	Betweenness
Promedio			
77,4 %	84,5 %	71,0 %	76,0 %

Tabla 2: *Porcentaje de información mutua para cada uno de los cuatro método indica la información mutua entre todas las particiones, utilizando una de las par de uno de los métodos como referencia.*

coincidencia con el resto de las particiones. Esto llama la atención, ya que es en comunidades que se cuestiona al considerar sus parámetros de bondad de

Finalmente, se estudió si había una cierta tendencia de género en la top Se buscaron posibles patrones relacionados al género en la formación de detectadas anteriormente comparando la cantidad de delfines de cada género por comunidad (y para cada método de detección de comunidades) con la cantidad esperada por azar.

La red cuenta con 34 delfines macho, 24 hembra, y 4 identificados con género 'na' (género desconocido). Se distribuyó el género de los delfines al azar, preservando el número total de delfines de cada género, y se contó el número de delfines de cada género en cada partición. El promedio de los mismos se tomó como el valor medio aleatorio de cada género por red. El cociente de los valores originales con estos esperados por azar dio la razón de cada género en cada comunidad (ver Apéndice C, tabla 5 para el detalle). Las proporciones se muestran en la figura 3.

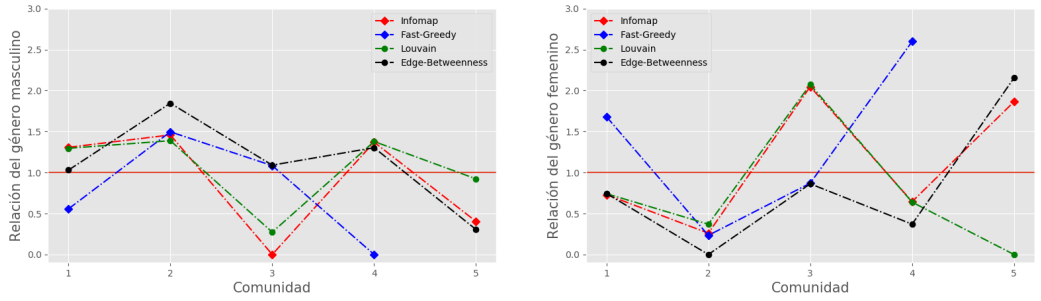

Figura 3: *Proporción de delfines de cada género dentro de cada comunidad, comparados con los valores esperados del promedio de* 2500 *asignaciones de género aleatorias.*

Un cociente cercano a 1 implica un valor comparable con el azar. Es notable la relación de dependencia de los géneros en las comunidades. La comunidad 2 tiene una clara sobrerrepresentación de delfines masculinos, con una consecuente subrrepresentación de delfines femeninos. Lo opuesto es cierto para la comunidad 4 casi en la totalidad de los métodos empleados. El comportamiento de las comunidades determinadas por *Fast-greedy* difieren de la norma. Las comunidades $1, 3$ y 5 presentan un comportamiento menos concluyente, ya que no coinciden

para todos los métodos de partición.

3. Conclusiones

si tuvieran que usar un algoritmo de comunidad cual usarían? y Por q

La red social de 62 delfines de Nueva Zelanda es una red modular mejor definida por 5 comunidades de delfines. El parámetro de modularidad de la red, de acuerdo a la partición, dió mayor al esperado por azar para un 75 % de los algoritmos empleados. La modularidad ofreció un mayor poder de resolución para definir la concordancia de las particiones ya que silhouette mostró poca diferencia respecto del azar.

Se buscó determinar la similitud entre todas las particiones obtenidas a partir de los cuatro métodos de identificación de comunidades analizados utilizando el criterio de Información Mutua. Se pudo determinar que el método que reconoce comunidades más comunes a toda las particiones es el $Fast - greedy$ mientras que el método que parece generar comunidade menos comunes a los demás es el Algoritmo de Louvain.

que significa

Utilizando la información respecto a la distribución del género de los delfines, se observ una tendencia a generar comunidades de modo que haya una de ellas predominantemente femenina y otra predominantemente masculina, independientemente del algoritmo empleado y la cantidad de nodos dentro de cada comunidad.

APÉNDICE

A. Comunidades

El análisis de superposición de nodos por comunidad se basó en el análisis de la intersección de sets de nodos. Viendo la intersección completa, se llega a la tabla 3, donde se pone el número de nodos presentes en cada comunidad para todas las particiones analizadas.

Nodos 'perpetuos'				
Comunidad 1	**Comunidad 2**	**Comunidad 3**	**Comunidad 4**	**Comunidad 5**
5	6	3	0	0

Tabla 3: *Número de nodos compartidos por todas las particiones, en cada comunidad.*

Al realizar un análisis de intersección entre comunidades comparando cada método por separado, tenemos que ciertos métodos de partición coinciden en mayor medida. Mientras que en la figura 1 se puede ver que Louvain y Edge-betweenness comparten en gran medida las particiones, un análisis cuantitativo de los datos se representa en la tabla 4. Para cada método de partición se presenta la intersección de los nodos de cada comunidad respecto a las alternativas, normalizada por el número de nodos de la comunidad del método analizado. Repitiendo para todos los métodos se tiene cada fila de la tabla.

Coincidencia de comunidades					
	Comunidad 1	**Comunidad 2**	**Comunidad 3**	**Comunidad 4**	**Comunidad 5**
Infomap	73,3 %	90,3 %	40,0 %	50,0 %	72,2 %
Fast-greedy	85,7 %	98,8 %	68,8 %	25,0 %	25,0 %
Louvain	68,3 %	48,7 %	68,8 %	25,0 %	47,2 %
Edge-betweenness	73,3 %	91,5 %	55,0 %	50,0 %	27,8 %

Tabla 4: *Porcentaje de semejanza entre cada comunidad determinada por un método, comparado con todo el resto.*

Como se puede ver, las comunidades con mayor variabilidad a través de los métodos son las dos últimas, donde se tiene un menor número de nodos, por lo que las variaciones son más representativas. Además, el método *Fast-greedy* presentó una comunidad menos, por lo que las otras debieron absorber los nodos de la faltante.

B. Caracterización de las particiones

Los histogramas de las figuras 4 y 5 muestran los valores de modularidad y silhouette de los 2000 recableados aleatorios de la red de delfines, con el valor reportado para la red original marcado en negro.

Si bien se presenta un comportamiento anómalo en la distribución de valores de modularidad mediante el algoritmo de partición de *Infomap*, en todos los casos el reportado supera el valor medio de la distribución azarosa. En silhouette, en cambio, no tiene casos favorables respecto al azar la mitad de las veces.

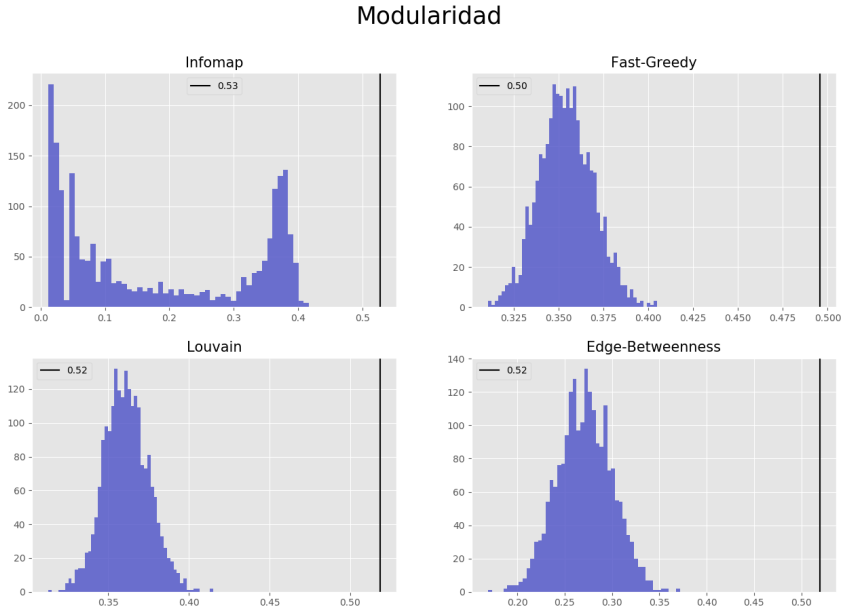

Figura 4: *Histogramas con la distribución de modularidad promedio de las 2000 redes simuladas, con los valores reportados de la red de delfines en negro.*

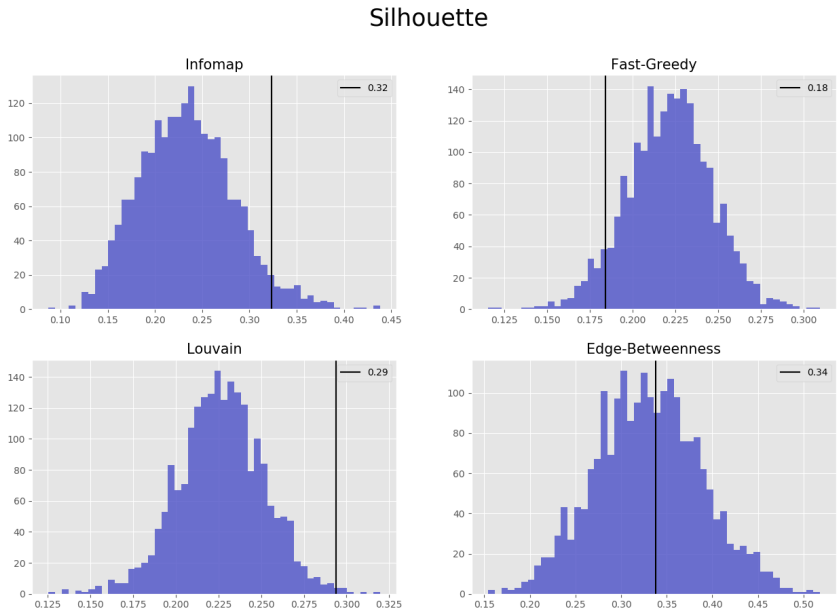

Figura 5: *Histogramas con la distribución de silhouette promedio de las 2000 redes simuladas, con los valores reportados de la red de delfines en negro.*

C.Representación de géneros por comunidad

La tabla 5 muestra el proceso de obtención de la relación de género en cada comunidad, respecto de los delfines presentes en la red, para el caso de las particiones dadas por el método *Infomap*. De manera análoga se obtienen los valores para los algoritmos restantes.

Se parte contabilizando el número de delfines de cada género en cada comunidad para un dado método de partición. Luego se re-asignan 2500 veces los géneros de cada delfín de manera aleatoria y se toma el número promedio de delfines de cada género determinados por el azar. Las columnas de comparación muestran la relación entre el número de delfines encontrado respecto de los esperados por azar. Es decir, un número mayor (o menor) que 1 representa una desviación del azar, dado que se tienen más (o menos) de los esperados.

	Red			Azar			Comparación		
	Machos	Hembras	NA	Machos	Hembras	NA	Machos	Hembras	NA
Comunidad 1	5	2	0	3.84	2.74	0.44	1.30	0.73	0
Comunidad 2	16	2	2	10.97	7.68	1.35	1.46	0.26	1.48
Comunidad 3	0	4	1	2.73	1.95	0.32	0	2.05	3.12
Comunidad 4	9	3	0	6.57	4.65	0.78	1.37	0.65	0
Comunidad 5	4	13	1	9.92	6.98	1.11	0.40	1.86	0.90

Tabla 5: *Cantidad de delfines de cada género dentro de cada comunidad detectada mediante Infomap de la red; se muestra la cantidad en la red real, la cantidad esperada por azar, y la comparación entre ambas.*

Referencias

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

[2] Lambiotte R Lefebvre E Blondel V, Guillaume J.L. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics*, 2008.

[3] Moore C Clauset A, Newman E.J. Finding community structure in very large networks. *Physical Review*, 70.

[4] Newman E.J Girvan M. Community structure in social and biological networks. *PNAS*, 99, 2002.