Universidad de Buenos Aires

Informe 3:

# Comunidades

*Raúl Barriga*
*Mariela Celis*
*Jimmy Masías*
*Sebastían Pinto*

9 de noviembre de 2016

# Índice

## 1. Introducción

La red de este informe trata de una comunidad de delfines de Doubtful Sound, Nueva Zelanda. La comunidad, que se constituye de 62 ejemplares identificados por una marca en la aleta dorsal, fue fotografiada entre 1995 y 2001. A partir de esos datos se construyó la red que contiene 159 links, donde se establece que existe un link entre aquellos individuos que fueron vistos juntos de forma más frecuente que la esperada aleatoriamente, es decir, por un criterio de "compañía preferida". [D. Lusseau, The emergent properties of a dolphin social network, Proc. R. Soc. London B (suppl.) 270, S186-S188 (2003).]

## 2. Partición en clusters

Implementamos diferentes algoritmos para la detección de comunidades en la red de delfines. La asignación de cada algoritmo se puede observar en la figura **??**, en la cual se puede observar que todos los algoritmos detectan entre 4 y 6 comunidades presentes en la red, aunque en algunos casos aparecen comunidades compuestas por solo dos delfines, es decir, de un tamaño considerablemente menor que el las restantes, lo cual se podría considerar la absorción de esta pequeña comunidad por parte de otra de mayor tamaño.

En la tabla **??** calculamos la modularidad y el silhouette dada por cada algoritmo. Para cuantificar qué tan pertinente es la partición en clusters de esta red, estudiamos estos mismos observables recablenado en forma aleatoria la red original, manteniendo la membresía a las particiones inalterable. Eso nos resultó en una distribución de valores de modularidad y silhouette de la figura **??**. Si se compara esta distribución con los valores de la tabla **??**, se puede concluir que tanto la modularidad como el silhouette es significativamente mayor en la red actual que la esperada por azar (hipótesis nula), con lo cual decimos que esta red se describir correctamente como una red compuesta por comunas.

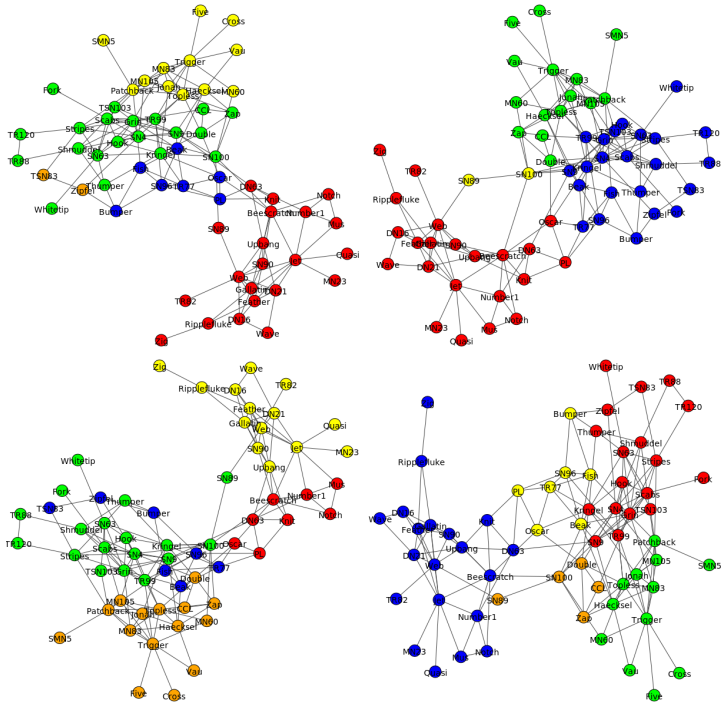

**Figura 1:** Layouts indicando el cluster asignado por cada algoritmo. De izquierda a derecha, y arriba hacia abajo: Edge betweenness, Fast greedy, Louvain e Infomap.

| Algoritmo | Modularidad | Silhouette |
|---|---|---|
| Edge-betweenness | 0.519 | 0.338 |
| Fast greddy | 0.495 | 0.184 |
| Louvain | 0.519 | 0.294 |
| Infomap | 0.529 | 0.328 |

**Tabla I:** Modularidad y silhouette de las particiones dadas por diferentes algoritmos.

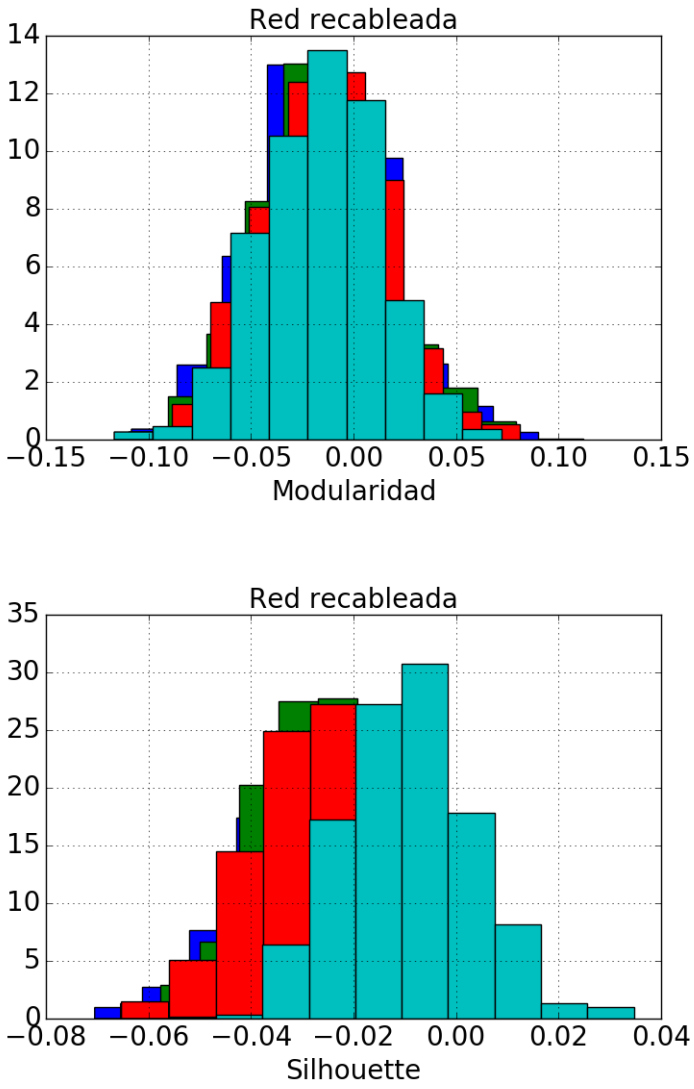


**Figura 2:** Modularidad y silhouette recableando en forma aleatoria, manteniendo la pertenencia a cada cluster dada por los algoritmos de detección de particiones.Se puede observar que la probabilidad de obtener los valores de la tabla **??** dada una reconexión aleatoria es prácticamente nula.

## 3. Relación entre comunidades

A través del índice de *Información Mutua* podemos cuantificar la similitud entre particiones, de comunidades de la red, definidas por dos conjuntos etiquetas $\{C_1\}$ y $\{C_2\}$. Este está dado por

$$I(\{C_1\},\{C_2\}) = \sum_{C_1,C_2} p(C_1,C_2)\log\frac{p(C_1,C_2)}{p(C_1)p(C_2)}, \tag{1}$$

o su versión normalizada

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

donde

$$H(C) = -\sum_{c_i\in C} p(c_i)\log(p(c_i)) \tag{3}$$

es la información total de la partición $C \equiv \{c_i\}$.

Las comunas construidas en el presente grafo fueron deducidas usando los algoritmos *greedy, betweenness, infomap* y *louvain* (RB & MC: Esto me parece que que es redundante... y el siguiente parrafo no lo entendemos). La definición **??** cuantifica en cuánto coinciden las particiones obtenidas por dos algoritmos diferentes. En el caso particular en que los conjuntos $\{C_1\}$ y $\{C_2\}$ estén descorrelacionados, entonces se dice que el conjunto $\{C_1\}$ no brinda ninguna información sobre el conjunto $\{C_2\}$, y de acuerdo a la ec. **??** obtenemos $I_n = 0$. Yi en el caso particular en que $\{C_1\}$ y $\{C_2\}$ son el mismo conjunto, obtenemos la información mutua normalizada $I_n = 1$, es decir, dos algoritmos diferentes encuentran exactamente la misma comuna.

### 3.1. Comparación entre algoritmos de reconocimiento de comunidades

La cuantificación de información dada por la Ec. **??** consta tanto de: la medición de la probabilidad de que un nodo pertenezca a una comunidad $C_i$ ($p(C_i)$), como de la probabilidad conjunta de que un nodo pertenezca a una comunidad $C_i$ en la partición $\{C_i\}$ y pertenezca a la comunidad $C_j$ en la partición $\{C_j\}$ ($p(C_i,C_j)$). La primera distribucón de pertenencia a etiquetas/comunidades se puede ver en la figura **??**. En ella se puede observar que el etiquetado muestra distintas distribuciones en cada caso, además es importante notar que etiquetados iguales no representan las mismas comunidades entre cada algoritmo, por lo tanto, no existe una única distribución

que represente cada caso; Por otro lado, el caso de la probabilidad conjunta es mostrado en las matrices de la figura **??**.

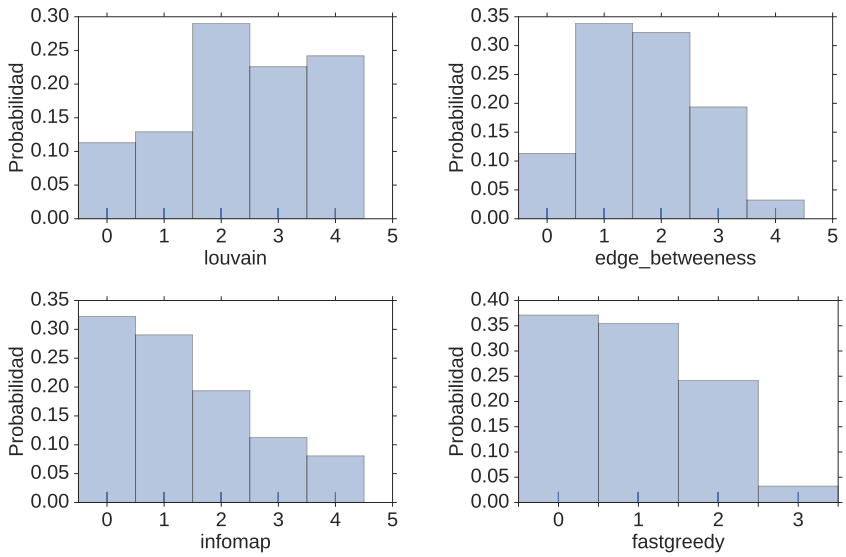


**Figura 3:** Distribución de probabilidad de pertenencia de un nodo a una comunidad para los diferentes algoritmos utilizados.

La *Información Mutua* total, normalizada, es mostrada en la Tabla **??** de la cual se puede observar que los algoritmos *infomap* y *louvain* son los más similares con una semejanza del 86,2 %, las rutinas *Fast Greedy* y *Edge Betweeness* muestran la menor correlación con una similitud del 66,2 % mientras que en general el resto coinciden en un rango de $70\,\% - 80\,\%$

**Tabla II:** Información mutua entre las particiones encontradas por cada algoritmo.

| | Fast Greedy | Edge betweeness | Infomap | Louvain |
|---|---|---|---|---|
| Fast Greedy | 1.000 | 0.662 | 0.767 | 0.794 |
| Edge Betweeness | | 1.000 | 0.771 | 0.732 |
| Infomap | | | 1.000 | 0.862 |
| Louvain | | | | 1.000 |

## 3.2. Relación de las comunas con género

Para cuantificar la relación entre las comunas deducidas por los diferentes algoritmos (e.g. *greedy*) y el género, usamos la ec. **??** identificando a las comunas con $\{C_1\}$ y a las etiquetas de género con $\{C_2\}$. En la figura **??**

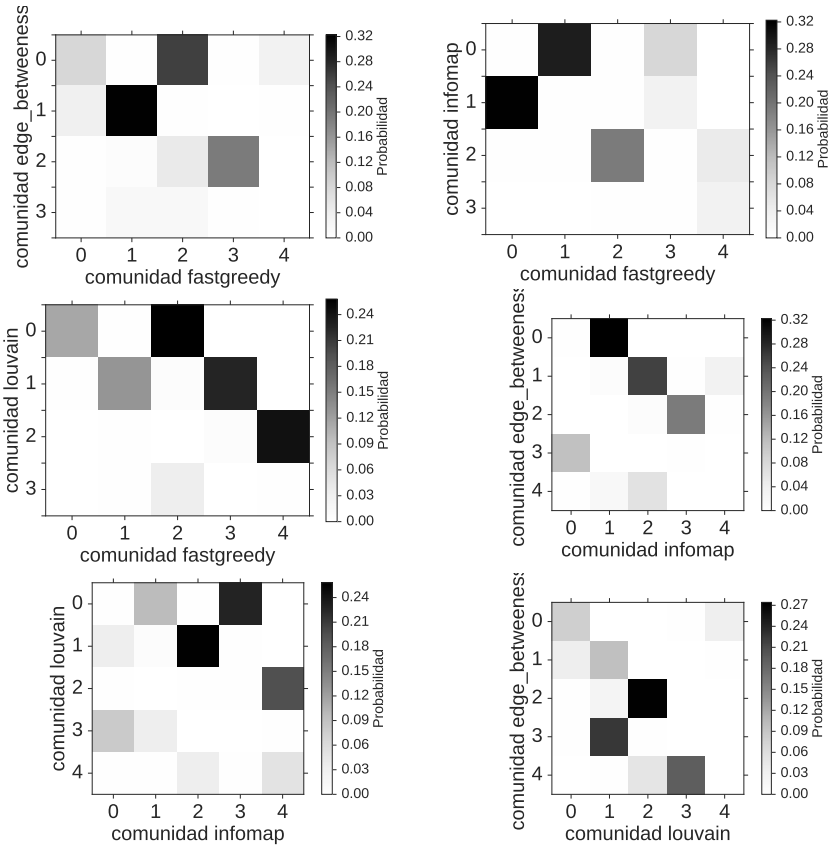


**Figura 4:** Distribución de probabilidad conjunta de pertenencia de un nodo a una comunidad de cada par de algotirmos.

mostramos, en el encabezado de cada panel, los valores de la información mutua $I_n$, los cuales caen en el intervalo $(0{,}10 - 0{,}21)$, es decir que $I_n \ll 1$ en todos los casos; esto nos dice que el conjunto de comunas ($\{C_1\}$) deducido por cierto algoritmo (e.g. *greedy*) no nos da mucha información sobre el género ($\{C_2\}$). Como test de consistencia para esto último, hicimos sorteos del género de cada nodo(manteniendo constante el número total de masculinos y femeninos por separado), y contabilizamos el número de enlaces entre pares de géneros distintos $n_i g$. En la figura **??** mostramos un histograma de $n_i g$, y en línea negra el valor asociado para la red real (original). De aqui vemos que el valor de la red real esta apartado $\sim 1\sigma$ del valor medio del histograma; lo cual significa que hay una ligera tendencia a que las comunas tengan muchos ejemplares de un sexo en particular. Esto último es consistente con el bajo

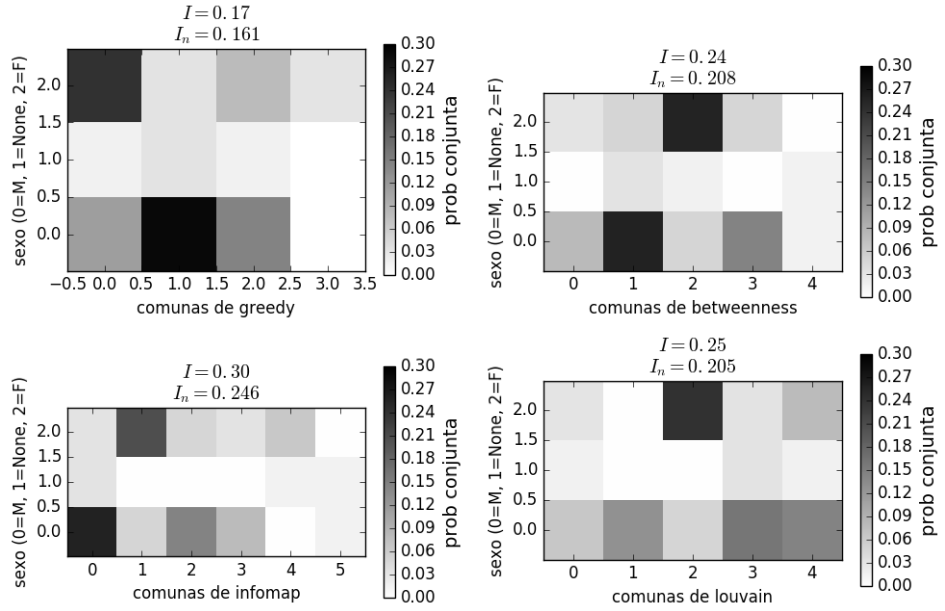


**Figura 5:** Valores de las matrices de probabilidad conjunta para los algoritmos *greedy* (izquierda, arriba) *betweenness* (derecha, arriba), *infomap* (izquierda, abajo) y *louvain* (derecha, abajo).

valor de $I_n$ discutido mas arriba. Sin embargo, la diferencia no parece ser significativa: la probabilidad de obtener el valor actual de número de links entre delfines de distinto género y misma comuna, dada una distribución de sexos al azar, es de aproximadamente 5 %, por lo tanto no es tan improbable obtener este valor en una asignación aleatoria de géneros.

# 4. Conclusiones

Estudiamos distintos algoritmos de detección de comunidades en la red delfines, obteniendo que esta red puede ser bien descripta como una red compleja conformada por comunas. Si bien se observó una ligera tendencia de las comunas a estar conformadas por ejemplares de un mismo sexo, uno de los test estadísticos realizados no nos dió una diferencia significativa de lo esperado al azar, por lo que no podemos descartar la hipótesis nula.

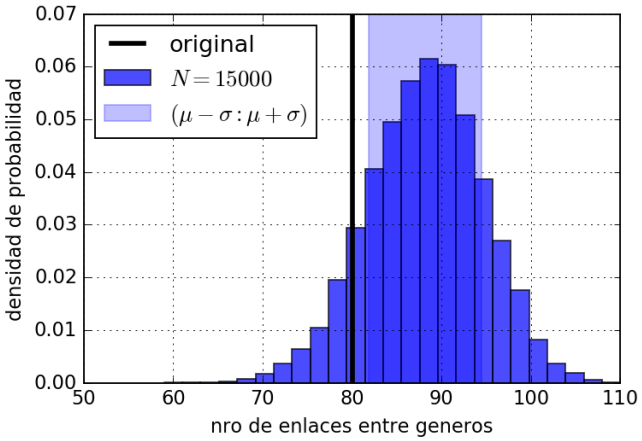


**Figura 6:** Distribución del número de enlaces entre géneros diferentes, para diferentes realizaciones de sorteo del sexo de los nodos de la red (manteniendo constante el número de masculinos y femeninos por separado). La línea negra muestra el valor que corresponde a la red original que caracterizamos en este trabajo. La zona sombreada en celeste representa la región que cubre la desviación estándar respecto de la media. El valor de la red original (o real) se aparta $\sim 1\sigma$ respecto del centro de la distribución, lo cual muestra una ligera tendencia a la existencia de comunas que tienen muchos ejemplares de un sexo en particular. Esto es consistente con el bajo valor ($\ll 1$) de la información mutua $I_n$ (ver ec. **??** y Secc. **??**).