Detección de comunidades en redes de enfermedades y caracterización según sus sistemas fisiológicos

Andino C. , Asplanato L. , Murchison, F.
Departamento de Física, UBA

Resumen

Se recreó el trabajo realizado por Goh. et. al [1] sobre la base de datos de interacción gen-enfermedad de DisGeNet [2] proyectando redes pesadas desde la red bipartita a cada espacio. Se emplearon dos métodos, una proyección pesada publicada por Zhou et. al [3] y una proyección directa. Se realizó un estudio de comunidades en cada espacio y se les asignó, en caso de una presencia considerable de una dada categoría fisiológica de MeSH [4], una fisiología a la comunidad.

1. Introducción

Una *red bipartita* tiene nodos de dos categorías y los enlaces ocurren únicamente entre nodos de distintos grupos. En este trabajo, se estudió la red bipartita formada por la vinculación entre mutaciones genéticas y la manifestación de enfermedades a partir de las mismas en el ser humano desde la base de datos de DisGeNet.

Dado el volumen de datos extraído se empleó el *score* propio, basado en el número de fuentes curadas, la literatura y el número de modelos en animales para cada asociación, para un filtrado a volumen manejable y que representara asociaciones de confianza. Para esto se unificaron todas las apariciones de vinculación dando como puntuación la suma de los *scores* individuales, filtrando para aquellos con valores mayores a 0,23, donde 0,2 es el mínimo posible. Se obtiene una red bipartita con 6149 enlaces, 3629 enfermedades y 2946 genes.

Las proyecciones de una red bipartita constan de asociar dos elementos de una categoría de la red si comparten vínculos con al menos un mismo nodo de la categoría opuesta (si comparten primeros vecinos). En este caso, la proyección sobre el espacio de enfermedades genera enlaces entre enfermedades que comparten al menos una mutación genética como origen. Permite ver la cercanía topográfica entre enfermedades dados sus vínculos con genes compartidos. La proyección al espacio de genes es análoga.

Las proyecciones a los espacios de nodos de una red bipartita deben realizarse tomando precaución de preservar la mayor cantidad de información de la red de la cual se parte. Si se establecen vínculos entre dos nodos de la misma categoría sin distinguir el número de vecinos que comparte se pierde información ya que es equivalente compartir 3 primeros vecinos o 100. Como primera corrección, se puede hacer una *proyección pesada directa*, donde se enlazan dos nodos si comparten al menos un primer vecino pero se genera un enlace pesado por el número de primeros vecinos en común. Este segundo tipo de proyección ignora la distribución de grado a lo largo de cada categoría, donde un nuevo vínculo a un *hub* no genera el mismo cambio de información como el vincular dos nodos previamente disconexos. Zhou et. al [3] discute un método de proyección de *pasaje de información* para generar redes con enlaces pesados y dirigidos, donde el peso toma en cuenta el número de vecinos compartidos entre dos nodos de la misma categoría y el grado del nodo "*objetivo*", como se ve en la siguiente ecuación.

$$w_{ij} = \frac{1}{k(j)} \sum_{l} \frac{1}{k(l)} \tag{1}$$

En la ecuación, w_{ij} es el peso del enlace entre el nodo i y j, $k(j)$ el grado del nodo j, y la sumatoria es sobre todos los vecinos l compartidos. La matriz de adyacencia de las redes creadas mediante esta proyección no será simétrica.

En este trabajo se consideraron el segundo y tercer método de proyección, llamados *proyección directa* y *proyección pesada* respectivamente.

Se utilizó una categorización de las enfermedades obtenidas del *Medical Subject Headings* (MeSH), una organización jerárquica de términos biomédicos construída por el *U.S. National Library of Medicine*.

2. Caracterización de las proyecciones

Ambos métodos de proyección resultaron en una red en el espacio de enfermedades de 1634 nodos, 9778 enlaces pesados y grado medio de ~ 12, y en el de genes, 1288 nodos, 20333 enlaces y ~ 32 de grado medio. Los coeficientes de clustering local (0,74 y 0,78) y global (0,39 y 0,55) fueron iguales para ambos métodos de proyección, por lo que la red de genes tiene en todos los casos mayor clausura.

2.1. Análisis de distribución de grado

Se analizaron las distribuciones de grado de las redes en cada proyección mediante el ajuste lineal del la distribución de grado en escala logarítmica. En el caso de la proyección directa, los valores fueron de $\gamma_e = 1{,}0 \pm 0{,}9$ y $\gamma_g = 0{,}7 \pm 0{,}2$ para las redes de enfermedad y genes respectivamente, por lo que se descarta que tenga una distribución de grado que cumpla ley de potencia libre de escala (ver Apéndice, sección A, figura 7). Como se puede ver en la figura 1, las redes de la proyección pesada tienen $\gamma_e = 3{,}4 \pm 0{,}4$ y $\gamma_g = 4{,}3 \pm 0{,}4$. Dado que los valores de la proyección pesada son $\gamma_i > 2$ para ambas redes, éstas son libre de escala. Esto puede implicar la existencia de estructuras mesoscópicas bien definidas, como comunidades, en las redes. Debido a que se observa que en las redes libre de escala la distribución de coeficientes de clustering disminuye con el grado de los nodos, eso puede implicar que estas redes poseen nodos de grado bajo pertenecientes a subgrafos muy densamente poblados, los cuales se conectan entre sí por hubs.

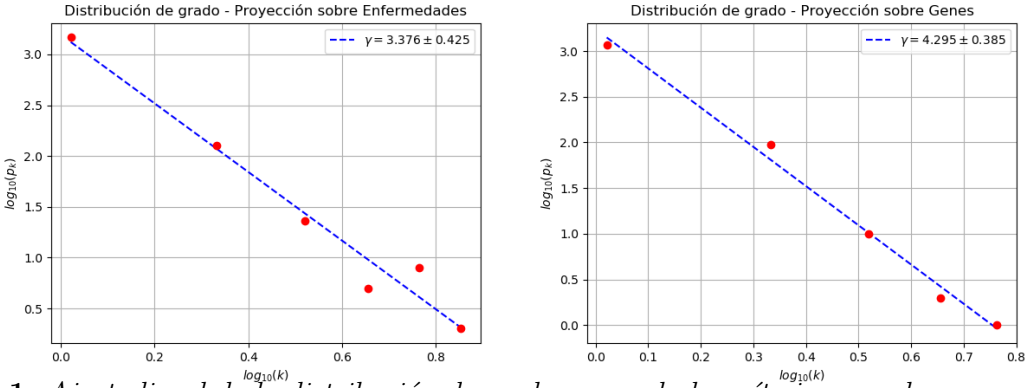

Figura 1: *Ajuste lineal de la distribución de grado en escala logarítmica para la proyección pesada en el espacio de enfermedaes (izquierda) y genes (derecha).*

Como análisis posterior, se analizó la asortatividad de las redes en el caso de las proyecciones directas. Como se puede ver en la figura 2, ambas proyecciones son asortativas, con

coeficientes $\mu_e = 0{,}10 \pm 0{,}03$ y $\mu_g = 0{,}19 \pm 0{,}02$. Esto indicaría que los nodos de mayor grado tienden a asociarse con nodos grado similar. Esto es de esperarase por la manera de construir las redes, ya que los nodos se asocian si comparten un cierto número de primeros vecinos. Las enfermedades asociadas a genes cuya mutación genera múltiples problemas fenomenológicos tendrá asociados a todos ellos. Y viceversa en el caso de la red de genes.

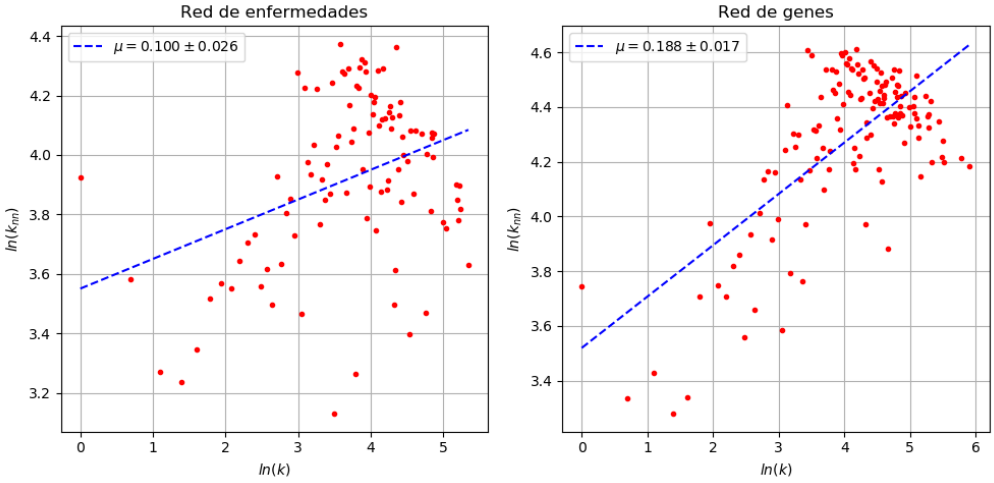

Figura 2: *Ajuste del coeficiente de asortatividad para las redes proyectadas de manera directa.*

2.2. Análisis de comunidades

Se emplearon los algoritmos *Infomap* y *Louvain* para el cómputo de las comunidades en las componentes gigantes de las redes de enfermedad y genes para cada método de proyección. Los datos del número de comunidades detectadas y la modularidad en cada caso pueden verse en la tabla 1.

Proyección directa

	Enfermedades		**Genes**	
	Comunidades	**Modularidad**	**Comunidades**	**Modularidad**
Infomap	140	0,56	76	0,57
Louvain	25	0,60	19	0,60

Proyección pesada

	Enfermedades		**Genes**	
	Comunidades	**Modularidad**	**Comunidades**	**Modularidad**
Infomap	171	0,74	108	0,68
Louvain	35	0,82	25	0,72

Tabla 1: *Número de comunidades en cada red proyectada, de acuerdo al mecanismo de proyección, con la modularidad de la partición.*

Se aprecia que el método de proyección pesada logra obtener una mayor modularidad en ambos espacios de proyección, independientemente del método empleado para la partición de comunidades. Esto puede deberse a que el peso asignado a cada enlace logra captar una mayor cantidad de información de la red bipartita que la proyección directa. De esta forma, cualquier algoritmo de determinación de comunidades asigna una vinculación apropiada relativa a este peso.

Ante la diferencia en el número de comunidades detectadas por cada método se realizó un análisis de la topología de las mismas con histogramas de densidad, coeficientes de clustering local y global y la fracción de nodos de cada comunidad respecto de la proyección. Los histogramas para la proyección directa en el espacio de enfermedades y genes puede verse en la figura 3. Se aprecia que en todos los casos el método de Infomap, que presentó menor modularidad, resulta en promedio en comunidades más intra-conectadas y de menor dimensión.

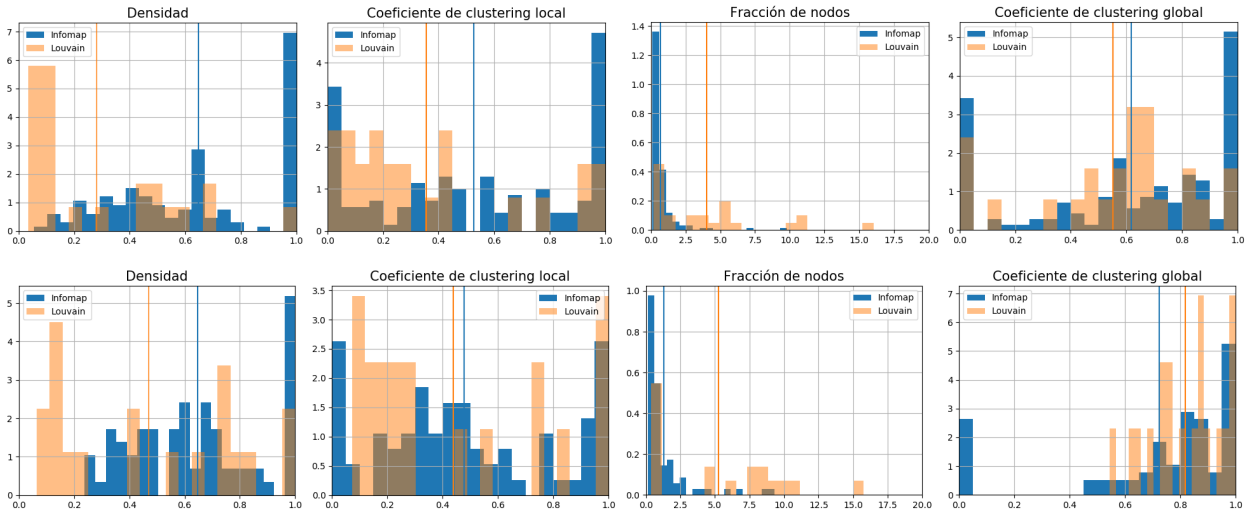

Figura 3: *Histogramas de los parámetros de densidad, coeficiente de clustering local, fracción de nodos y coeficiente de clustering global para las comunidades generadas con los métodos Infomap en azul y Louvain en naranja. Red de enfermedades (arriba) y genes (abajo). Las rectas verticales muestran la media de cada distribución.*

La figura 4 muestra las comunidades de mayor tamaño como sub-grafo de la componente gigante de la red de enfermedades con ambos métodos. En proyección directa se seleccionaron 6 comunidades a analizar ya que la quinta y sexta presentaron el mismo número de nodos. Las componentes gigantes totales de cada red pueden verse en el Apéndice, sección B, figura 8.

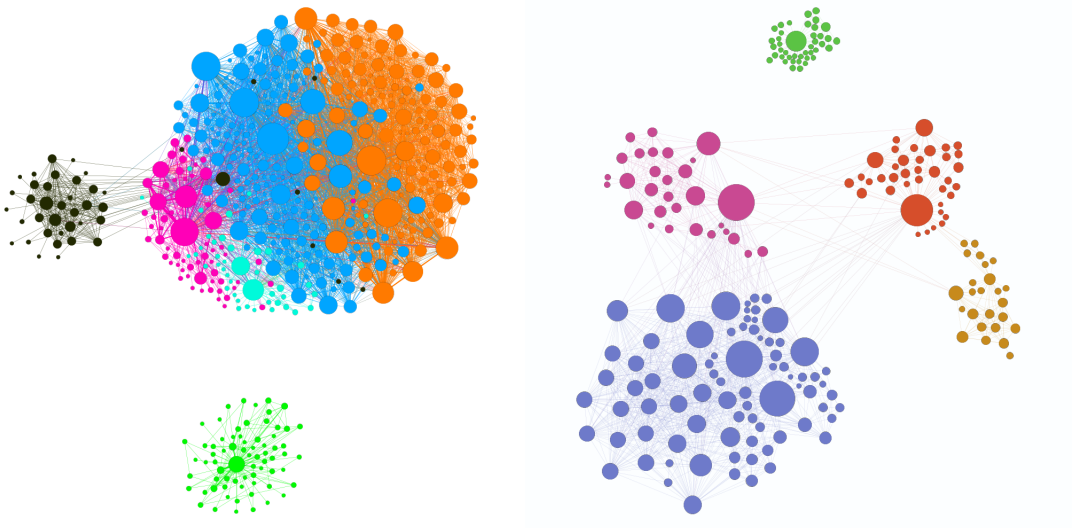

Figura 4: *Subgrafo de las comunidades más importantes de la proyección de enfermedades para la proyección directa (izquierda) y pesada (derecha) con el algoritmo de clustering Infomap.*

La distribución generada por el método pesado genera comunidades claramente delimitadas y con pocos enlaces inter-comunidades mientras que el método directo presenta un gran

solapamiento entre comunidades. Esto se condice con los valores de modularidad previamente mencionados. Si bien la modularidad de la proyección directa indica un valor mayor por azar para toda la componente gigante, el método no parece clusterizar de manera rotunda las comunidades de mayor tamaño.

2.2.1. Clasificación de enfermedades

Para evaluar si las comunidades inherentes a la topología de la red coinciden con alguna propiedad propia del sistema de enfermedades en el ser humano, se asignaron clases a los nodos utilizando la base de datos MeSH (*Medical Subject Headings*) [4] para categorizar a cada una de acuerdo a su origen o cualidad fisiológica. Se restringió el estudio a las comunidades preseleccionadas en la figura 4.

En primera instancia, dado que se busca conectar la organización en el espacio de enfermedades con el de genes, se ignoraron clasificaciones que no implicaran una condición genética como causante, por ejemplo, *'Wounds and Injuries'*. Cada nodo recibió un atributo con la lista de clasificaciones posibles, detalladas en la tabla 2, ya que las enfermedades pueden estar asociadas a más de un sistema fisiológico.

Código	**Nombre de Categoría**	**Código**	**Nombre de Categoría**
C04	Neoplasms	C14	Cardiovascular Diseases
C05	Musculoskeletal Diseases	C15	Hemic & Lymphatic Diseases
C06	Digestive System Diseases	C16	Congenital, Hereditary, & Neonatal Diseases & Abnormalities
C07	Stomatognathic Diseases	C17	Skin & Connective Tissue Diseases
C08	Respiratory Tract Diseases	C18	Nutritional & Metabolic Diseases
C09	Otorhinolaryngologic Diseases	C19	Endocrine System Diseases
C10	Nervous System Diseases	C20	Immune System Diseases
C11	Eye Diseases	C23	Pathological Conditions, Signs & Symptoms
C12	Male Urogenital Diseases	F01	Behavior & Behavior Mechanisms
C13	Female Urogenital Diseases & Pregnancy Complications	F03	Mental Disorders

Tabla 2: *Relación entre los códigos y categorías utilizadas de la base de datos MeSH.*

Para determinar la clasificación de cada comunidad se calculó el porcentaje de nodos de cada tipo, considerando que la base de datos disponible no lograba calificar a todos los nodos de la red. De esta manera, se consideraron los gráficos de barra de la figura 5 que contabilizan el porcentaje de nodos de cada clasificación dentro de las comunidades. Se ignoraron las comunidades donde más de la mitad de los nodos estuviera sin clasificar. De las seleccionadas, se muestran únicamente aquellos casos donde las fracciones fueran lo suficientemente representativas, es decir, donde alrededor de la mitad de los nodos etiquetados pertenecieran a una misma categoría.

El orden de la numeración de las comunidades es decreciente en el número de nodos. El método de proyección directo brindó dos comunidades clasificadas, la *Comunidad 2*, con una fuerte predominancia de enfermedades del tipo *'C4'*, *Neoplasmas*, y la *Comunidad 4*, que presentó un 44 % de nodos sin clasificar, por lo que predominaron las enfermedades de la categoría *'F03'*, de *Desórdenes Mentales*. En la proyección pesada se encontraron dos comunidades de interés, las comunidades *1* y *4* donde predominan las categorías *'C04'* y *'C14'*, correspondientes a *Neoplasmas* y *Enfermedades Cardiovasculares* respectivamente.

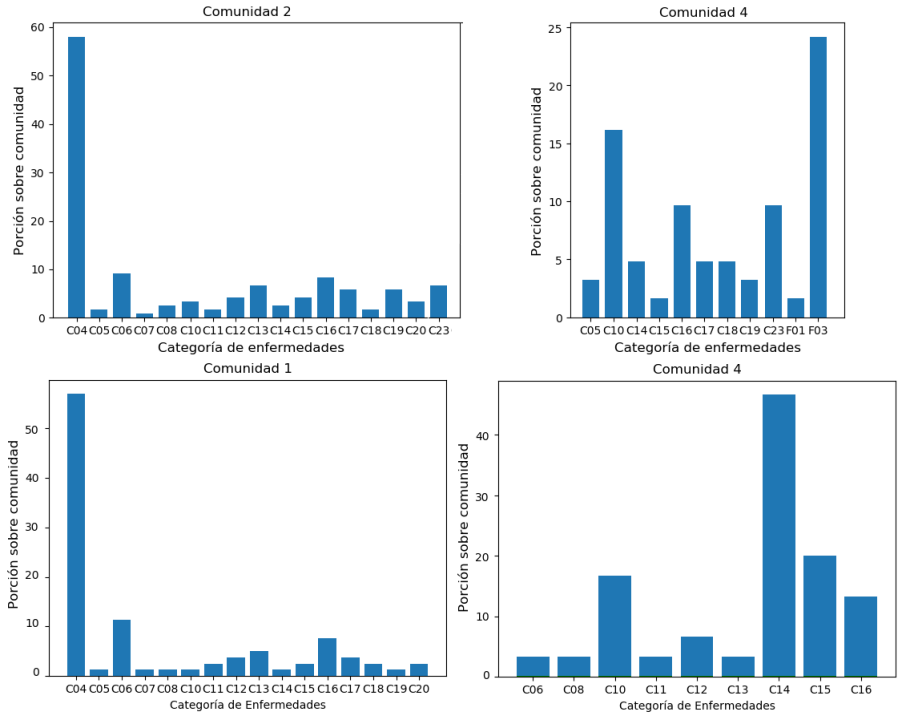

Figura 5: *Composición de las comunidades con alguna fracción fisiológica representativa, dentro de las 5 comunidades de enfermedades con mayor cantidad de nodos. Comunidades con proyección directa (arriba) y pesada (abajo).*

En ambas proyecciones surge una comunidad topológica con una clara composición predominante de enfermedades del tipo de Neoplasmas (tumores benignos o cancerígenos). Esto puede deberse a que los tumores aparecen en todo el cuerpo y pueden originarse por múltiples mutaciones. Así, era de esperarse que se tuviera mucha información sobre los mismos a partir de la red bipartita original y, además, que coincidiera con las comunidades de mayor número de nodos.

Cabe destacar que en ambos casos, la fracción de nodos puede sumar más del 100 %, ya que se tienen múltiples asignaciones por cada nodo.

Asignando el caracter fenomenológico de las enfermedades a los genes cuyas mutaciones lo causan (asociaciones dadas por la base de datos de DisGeNet), se pudo hacer el mismo estudio a las comunidades encontradas por el método *Infomap* en el espacio de genes a partir de la proyección directa. El análisis de fracción de carácter fisiológico por comunidad puede verse en la figura 6.

Se encontraron tres comunidades con al menos la mitad de los nodos clasificados en alguna categoría y donde una de ellas fuera preponderante. La comunidad 1 presenta nuevamente un máximo en *'C04'*, *Neoplasmas*, la comunidad 2 tiene un gran número de nodos de genes cuyas mutaciones implican una enfermedad del tipo *'C14'*, o *Cardiovascular* y la comunidad 3 tiene su máximo en *'F03'*, con *Desórdenes Mentales*. Como puede verse, dos de estas tres comunidades coinciden en la clasificación con las comunidades de mayor tamaño detectadas en la red proyectada sobre el espacio de enfermedades (Neoplasmas y Desórdenes Mentales).

Para ver la correlación entre los espacios, se analizó la intersección entre estas comunidades de igual clasificación en ambos espacios de las proyecciones directas. Se obtuvieron los primeros vecinos de las enfermedades en la red bipartita y se los comparó con la lista de nodos de la comunidad genética de la misma clasificación.

Las enfermedades de la comunidad de neoplasmas tenían 178 nodos génicos vecinos en la

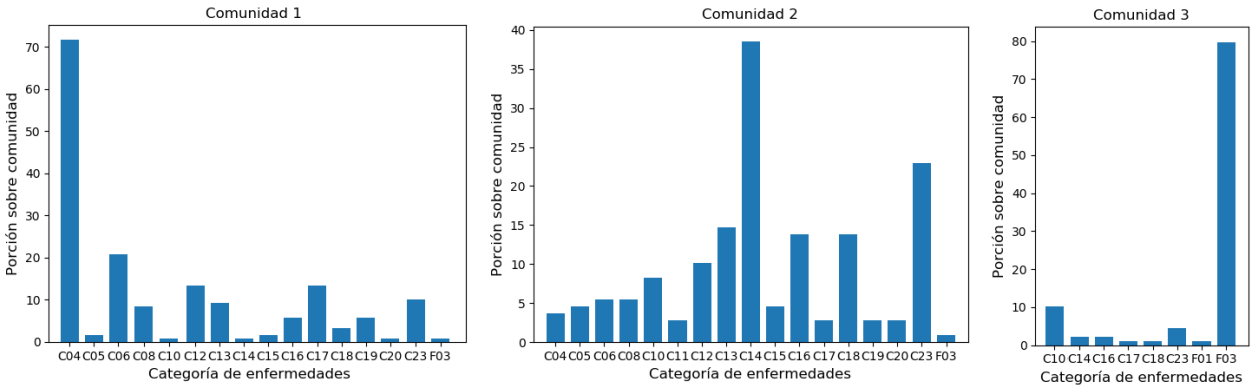

Figura 6: *Composición de las comunidades con alguna fracción fisiológica representativa, dentro de las 5 comunidades de genes con mayor cantidad de nodos. Comunidades con proyección directa (izquierda) y pesada (derecha).*

red bipartita, y 120 genes en la comunidad de neoplasmas de ese espacio. La intersección entre los mismos fue de 111. Esto quiere decir que el 92,5 % de nodos que quedaron dentro de la comunidad genética de neoplasmas representan un $\sim$ 62,4 de todos los genes que generaron esas enfermedades 'neoplasmáticas'. Los nodos de desórdenes mentales tenían 199 genes vecinos en la red bipartita, y una comunidad de genes de 89 con una intersección de 88 nodos, por lo que un $\sim$ 98,8 % de los nodos genéticos definidos como precursores de desórdenes mentales representan el 44,2 % de todos los genes asociados a enfermedades calificadas como 'desórdenes mentales'. Mientras que la comunidad de neoplasmas parece ser coincidente, la de desórdenes mentales tendría menos información. Esto puede deberse a que la calificación de las comunidades de neoplasmas se realizó con mayor certeza que la de desórdenes mentales al comparar número de nodos sin asignació y razón entre las fracciones preponderantes.

Restó verificar estas proporciones para las comunidades provenientes de la proyección pesada, así como contrastarlo con lo obtenido de un análisis avanzando por los vecinos en la red bipartita de manera inversa, viendo los vecinos de las comunidades genéticas y compararlo con las comunidades de enfermedades.

3. Conclusiones y consideraciones futuras

En este trabajo se lograron proyecciones de la base de datos de *DisGeNet*, filtrando los datos para mantener aquellos con un *score* $> 0{,}23$, mediante los métodos de proyección directa y pesada. Éstos resultaron en redes de topología similar, con mismo número de nodos, enlaces, grado medio y coeficientes de clustering local y global. Mientras las proyecciones pesadas brindaron redes con distribución de grado de cola pesada, la proyección directa brindó una distribución de red de potencias con un coeficiente menor al necesario para ser considerado de este tipo. La asortatividad de estas últimas proyecciones muestra que, como es de esperar debido a cómo se vinculan los nodos en cada espacio, se tiene una red con cierta homofilia de grado.

En ambas proyecciones se identificaron comunidades mediante los métodos de *Infomap* y *Louvain*, con un número de comunidades encontradas siempre mayor por parte de Infomap. La proyección pesada mostró ser, tanto para enfermedades como para genes, de mayor modularidad. Esto quiere decir que el número de enlaces entre nodos de la misma comunidad es mayor que los esperados por azar. Se corroboró, mediante la visualización de las 5 o 6 comunidades de mayor tamaño de la componente gigante de la red de enfermedades, que la partición lograda por el mismo algoritmo (*Infomap*) sobre la red de proyección pesada tenía comunidades mejor

diferenciadas.

Empleando la base de datos MeSH pudieron asignarse clasificaciones a los nodos de ambas proyecciones en el espacio de enfermedades y, mediante su relación en la red bipartita a los genes que las causan, al espacio de genes en el caso de la proyección directa. Esto permitió analizar la composición de fisiologías involucradas en cada comunidad, dentro de las 5 mayores para restringir el volumen de análisis. Se descartaron también aquellas comunidades con más de la mitad de nodos de asignación desconocida (por no estar incluida en la base de datos de MeSH) y con proporciones similares de dos fisiologías como para definir unívocamente su tipo. Ambos métodos de proyección brindaron una comunidad con clasificación, de acuerdo a la proporción de sistemas fisiológicos presentes, Neoplasmática, y una segunda que no compartieron de Desórdenes Mentales y Enfermedades Cardiovasculares en el caso de la proyección directa y pesada respectivamente.

En el espacio de genes de proyección directa se estudiaron de la misma manera las comunidades, encontrando tres donde preponderaba alguna clase de enfermedaes. Las mismas fueron de Neoplasmas, Cardiovasculares y Desórdenes Mentales. Al encontrar comunidades de igual clasificación a las del espacio de enfermedades para el caso de la proyección directa, se analizó la correlación entre las mismas evaluando la coincidencia del grupo de nodos clasificados como pertenecientes a los Neoplasmas (o Desórdenes Mentales) en el espacio de genes con el grupo de nodos de genes asociados a las enfermedades de la misma clasificación en la red bipartita. Se encontró que las comunidades de Neoplasmas presentaron mayor coincidencia ($\sim 62{,}4\,\%$) que la de Desórdenes Mentales ($\sim 44{,}2\,\%$).

Para que la aplicación de éste método sea más rotundo, debería continuarse el trabajo empleando algún algoritmo de propagación o "*Functional Flow*"[5] para inferir las clasificaciones del gran número de nodos sin asignar y repetir el análisis de fracción de clasificación de nodos. Además, debería trasladarse este análisis a las redes creadas mediante la proyección pesada que parecen presentar comunidades mas definidas pero no se cuenta con la información cuantitativa del solapamiento entre espacios, por ejemplo, para compararlos uno a uno.

Referencias

[1] David Valle Barton Childs Marc Vidal Kwan-II Goh, Michael E. Cusick and Albert-László Barabási. The human disease network. *PNAS*, 2007.

[2] Disease gene net. http://www.disgenet.org/web/DisGeNET/menu/dbinfo. Accessed: 2018-11/12.

[3] Matúš Medo Tao Zhou, Jie Ren and Yi-Cheng Zhang. Bipartite network projection and personal recommendation. *Phys. Rev*, 2007.

[4] Medical subject headings. https://www.nlm.nih.gov/mesh/. Accessed: 2018-11/12.

[5] Agarwal A Chazelle B Singh M. Nabieva E, Jim K. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics*, 2005.

Apéndice

A. Distribución de grado

Se muestran los ajustes para las proyecciones directas de la red bipartita. En el caso de la proyección sobre genes se tiene no sólo un menor coeficiente, si no que además el ajuste presenta mayores diferencias respecto de los datos obtenidos. En principio estos datos cumplirían una ley de potencias pero no serían de cola pesada.

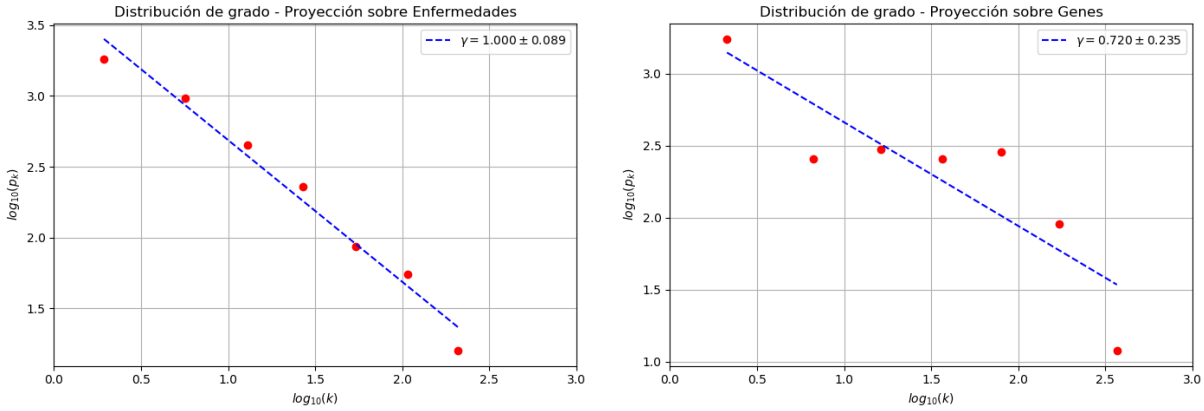

Figura 7: *Ajuste lineal de la distribución de grado en escala logarítmica para la proyección directa en el espacio de enfermedades (izq.) y genes (der).*

B. Redes - Componente Gigante

Se muestran las componentes gigantes de las proyecciones a cada espacio (genes o enfermedades). En el caso del método de proyección directo, se muestran las comunidades encontradas por los métodos Infomap y Louvain. Los colores representan cada comunidad encontrada, el tamaño de los nodos es proporcional al grado de cada nodo, y llevan nombre los primeros nodos en el ranking de betweenness centrality que, a partir del análisis de vulnerabilidad de las redes, se encontró que era un parámetro indicativo de centralidad de los nodos. Las redes azules son las proyecciones a cada espacio mediante el método de proyección pesada.

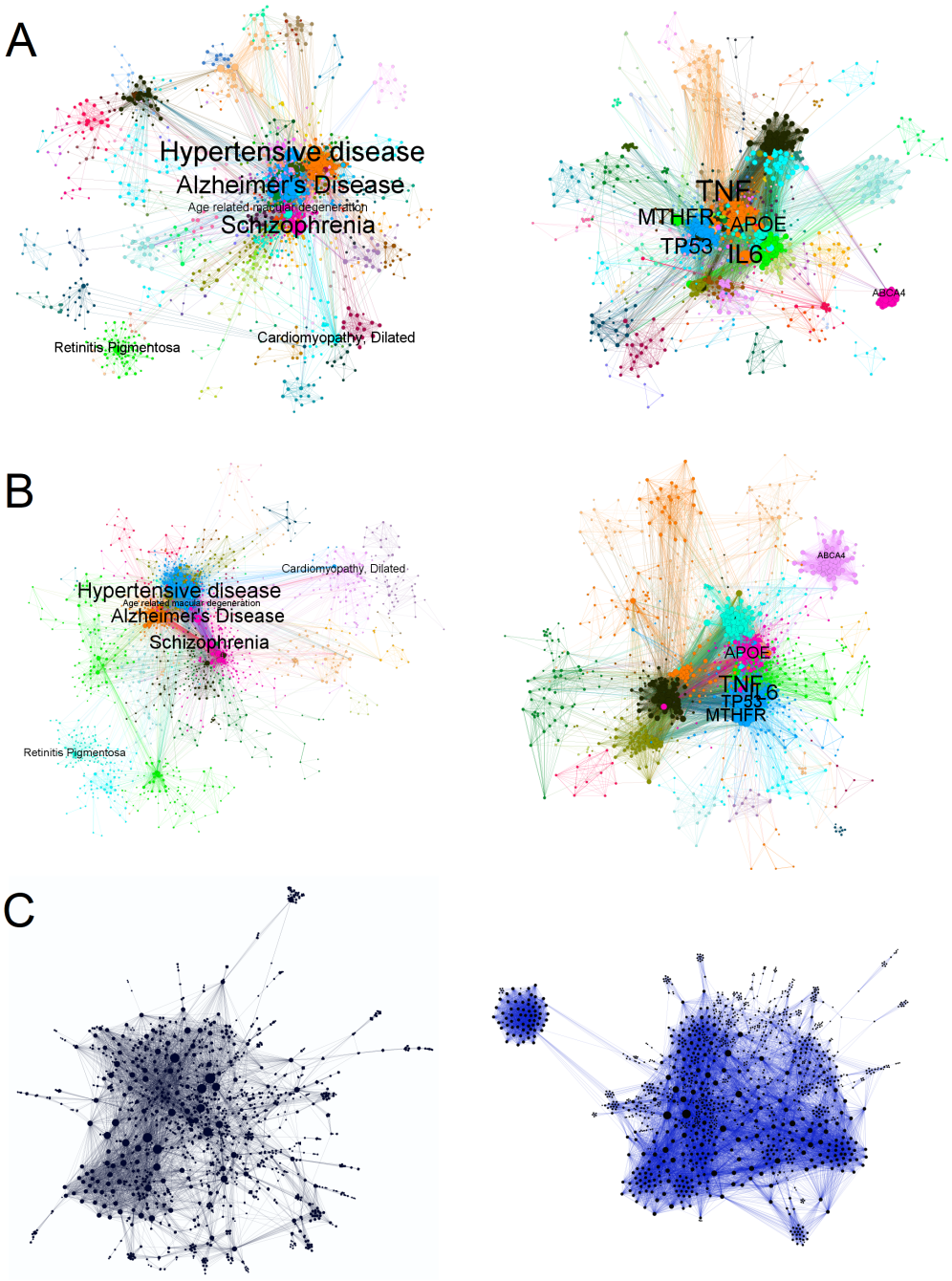

Figura 8: *Componente gigante de la red proyectada en el espacio de enfermedades (izquierda) y genes (derecha) mediante método de proyección directa con clasificación de comunidades por método de Infomap (A), Louvain (B) y proyección pesada, sin clasificación de comunidades (C).*