Priorización de genes-enfermedades mediante procesos de difusión no lineales en redes complejas

Director: Ariel Chernomoretz

Autor: Bautista Buyatti

Licenciatura en Ciencias Físicas

Departamento de Física

Facultad de Ciencias Exactas y Naturales

Universidad de Buenos Aires

Resumen

El presente trabajo se centra en la aplicación de la teoría de redes complejas como herramienta fundamental para organizar y extraer conocimiento acerca de las asociaciones entre genes y enfermedades. Uno de los desafíos abordados en este estudio es el problema de la priorización en redes, donde se emplean métodos de difusión, destacando especialmente la exploración de modelos de difusión no lineales. La priorización en redes implica la identificación y clasificación de nodos relevantes en una red, en este caso, genes relacionados con enfermedades, mediante la propagación de información a lo largo de la red.

Para comprender y analizar el comportamiento de los algoritmos de difusión, se generan redes sintéticas mediante un modelo basado en redes bipartitas. Estos experimentos proporcionan un ambiente de prueba controlado donde se pueden evaluar las características de diferentes tipos de difusión.

Posteriormente, la investigación se traslada de las redes sintéticas a las redes reales. Se emplea una red de interacciones de proteínas (productos génicos) como base, a la cual se incorpora información adicional mediante una red que representa las asociaciones entre genes y enfermedades. Esta información adicional sirve como punto de partida para los procesos de difusión y también para la posterior evaluación de los modelos.

En resumen, esta tesis combina la construcción de modelos sintéticos para comprender mejor los algoritmos de difusión con la aplicación de estos modelos en redes biológicas reales, con el objetivo de entender los procesos de difusión no lineales y de mejorar la priorización de genes asociados a enfermedades.

Agradecimientos

Quisiera agradecer a la UBA, en particular al departamento de Física de la FCEN y los departamentos involucrados a lo largo de la carrera. A los profesores y ayudantes con los que cursé y que me ayudaron y acompañaron a lo largo de las materias. Quiero agradecer al Instituto Leloir por permitirme realizar mi tesis de licenciatura en sus establecimientos. Especialmente a mi director Ariel Chernomoretz quien me permitió trabajar con él, estando presente aún en los momentos en los que se encontraba más ocupado, guiándome a lo largo de este trabajo y enseñándome una incontable cantidad de cosas sobre el área especifica pero también de cosas por fuera del trabajo en sí. A los chicos del laboratorio y a las tesistas que siempre tuvieron la buena voluntad para ayudar y aportar.

Por sobre todo quiero agradecer a mi familia, a mis padres por haberme dado la libertad y el apoyo de seguir el camino que elegí para estudiar esta carrera por todos estos años. A mis hermanos con quienes nos ayudamos, apoyamos y aconsejamos. Al resto de la familia y amigos por siempre mostrar interés en mis estudios y alentar en los momentos necesarios.

Este largo camino no hubiera sido el mismo sin ustedes.

Índice general

Capítulo 1

Introducción

Uno de los grandes desafíos de la investigación biomédica actual es tratar de zanjar la brecha que existe entre las diferentes escalas de organización que conviven en un organismo resolviendo lo que se conoce como la relación genotipo-fenotipo. Esto se traduce, en última instancia, a tratar de encontrar cuales son las bases moleculares de diferentes funciones biológicas o de patologías y enfermedades.

En este contexto el enfoque de redes complejas aparece como una aproximación sumamente apropiada para sacar provecho de la información relevante contenida en la enorme cantidad de evidencia experimental acumulada a diferentes niveles. Ya desde hace algunos años, aplicaciones de la teoría de redes complejas han sido consideradas para, entre otras cosas: estudiar patrones de coexpresión génica, identificar biomarcadores de diversas patologías, y asignar funcionalidad a productos génicos no anotados. Este enfoque también ha comenzado a ser utilizado de manera creciente en el ámbito del análisis de enfermedades humanas con aplicaciones en temas tales como la identificación y caracterización de bases moleculares y etiología de enfermedades, el estudio de propiedades topológicas de los nodos asociados a genes vinculados con enfermedades específicas, la identificación de sub-redes enriquecidas en nodos asociados a enfermedades y el diseño de fármacos, entre otros.

En este trabajo se pretende estudiar el uso de campos escalares desplegados sobre redes complejas a través de operadores de difusión para poder establecer predicciones sobre nuevas asociaciones relevantes entre genes y enfermedades. Utilizaremos para ello un corpus de conocimiento existente relevado a diferentes niveles que involucra: asociaciones ya establecidas entre genes y enfermedades de base hereditaria y redes de interacciones físicas entre proteínas.

Mas específicamente, el objetivo es caracterizar el uso de operadores de difusión no-lineales sobre redes complejas en la tarea de priorizar nuevos vínculos con el fenotipo asociado al conjunto de nodos semilla utilizado para propagar el campo escalar de 'recomendación'. En particular, interesará aplicar lo aprendido al problema de predecir nuevas asociaciones entre genes y fenotipo patológicos (i.e.enfermedades). La hipótesis de trabajo es que el co-

nocimiento embebido en el patrón de conexiado de las redes utilizadas permitirá establecer relaciones de similaridad topológica que resulten relevantes para proponer potenciales nuevas asociaciones con el fenotipo de interés. Adicionalmente, la hipótesis es que el uso de operadores de difusión no lineales permitirán incorporar la idea de efecto sinérgico entre el conjunto completo de semillas de partida.

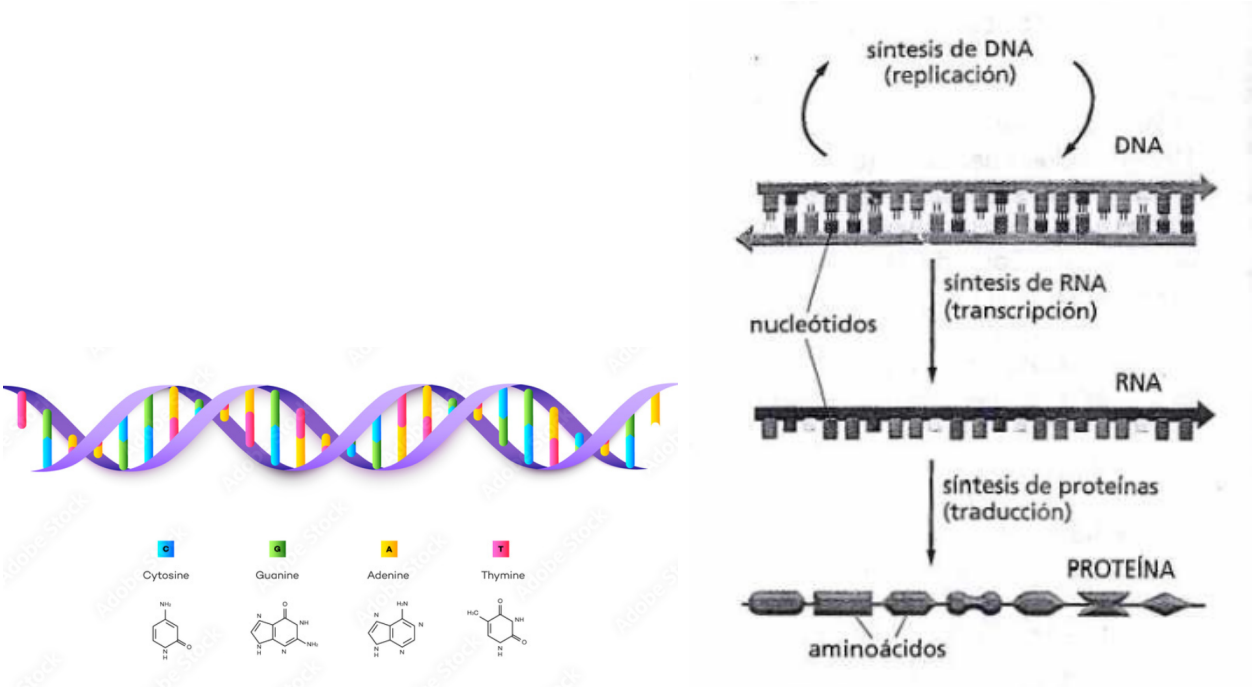
1.1. Conceptos de biología celular

Todos los organismos vivos están formados por células, los organismos unicelulares son la forma de vida mas simple que existe. En cambio los organismos mas complejos como el ser humando están formados por comunidades de células que cumplen funciones especializadas coordinadas por sistemas complejos de comunicación.

Las células que forman a los humanos contienen en su núcleo al genoma, que es el conjunto completo de moléculas de ADN (ácido desoxirribonucleico). Este funciona como una biblioteca completa de información genética que almacena las instrucciones para informar a la célula como generar a las proteínas que son necesarias para su funcionamiento y supervivencia. Estas instrucciones son leídas o transcritas en un conjunto de polímeros llamado ARN (ácido ribonucleico), cuya clase principal actúa como ARN mensajero para transportar la información fuera del núcleo de la célula donde son a su vez traducidos a otro tipo de polímeros llamados proteínas. En la figura (1.1) se muestra por un lado un esquema del ADN y por otro el proceso que se lleva a cabo en la célula por el cual se producen proteínas.

La información para producir una proteína se encuentra codificada en los genes, que son secuencias de ADN que codifican un producto génico (figura 1.2) y no todas las células producen todas las proteínas sino que cada célula expresa algunos genes y otros no. Son las proteínas las que dominan el comportamiento de la célula actuando como soporte estructural, catalizadores químicos, motores moleculares, entre otras muchas funciones. Todas estas funciones son posibles gracias a que varias de ellas se juntan formando complejos y entramados de interacción de los que emergen las funcionalidades que permiten realizar tareas. Estos complejos pueden representarse mediante redes de interacción donde se codifican mediante enlaces las diferentes interacciones que tienen lugar resultando en una estructura de grafo que es de gran utilidad para comprender los diferentes caminos y proteínas involucradas en el funcionamiento celular.

Por otro lado, una definición de enfermedad es aquella condición sobre el organismo o una de sus partes que perjudica o evita su normal funcionamiento y típicamente se manifiesta a través de distinguidos signos y síntomas. Esta definición también describe el mal funcionamiento de una célula o un grupo de ellas y de hecho muchas enfermedades pueden



(a) Esquema del ADN donde se muestran las moléculas que lo componen llamadas nucleótidos.

(b) Proceso de producción de proteínas [ver 18, cap.1].

Figura 1.1

ser definidas a una escala celular. Alteraciones en cualquiera de los procesos que normalmente ocurren en las interacciones moleculares o en las rutas moleculares pueden por lo tanto conducir a enfermedades. De igual forma modificaciones a nivel genético cambiarían los productos resultantes y también llevarían a alteraciones en el funcionamiento normal dando lugar a enfermedades de base genética, que son en las que se tiene interés en esta tesis [ver 18, cap.1].

1.2. Redes

Una red o grafo en lenguaje matemático es una colección de nodos unidos entre sí por enlaces. Esta construcción permite pensar y modelar muchos objetos de interés en el mundo de la física, biología, ciencias sociales, etc y hacerlo ofrece a menudo un conocimiento nuevo y útil sobre el comportamiento del sistema en su conjunto y los patrones de conexión que subyacen en ellos. Algunos ejemplos de redes del mundo real pueden ser ciudades (nodos) conectadas por autopistas (enlaces), personas (nodos) en redes sociales que se siguen o establecieron una relación de amistad (enlaces), el cerebro, redes eléctricas, las páginas en internet, redes de interacciones de proteínas, etc. En la figura (1.3) se muestra un ejemplo de una red de direcciones de internet [10].

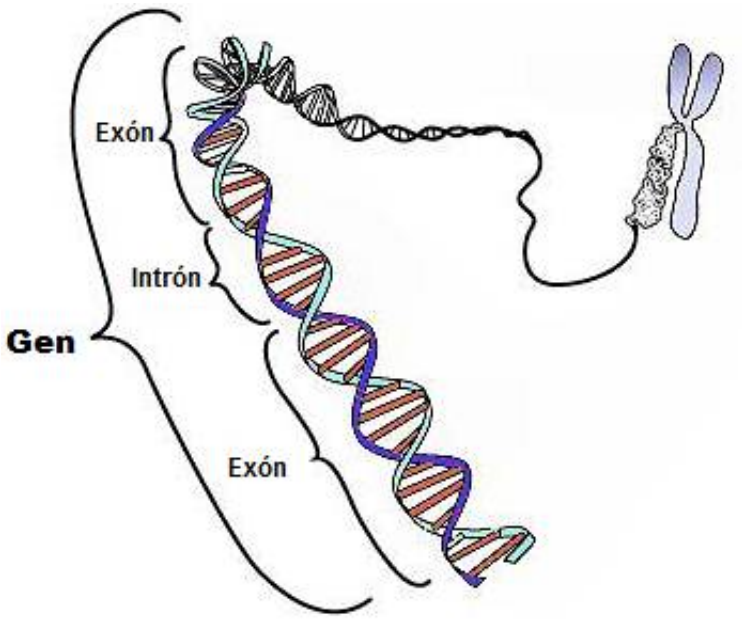


Figura 1.2: Esquema de un gen en el ADN, compuesto por una parte llamada exón y una llamada intrón.

Una red es una representación simplificada que reduce un sistema complejo a una estructura abstracta que captura los aspectos mas básicos de los patrones de conexión. Los nodos y enlaces de la red pueden llevar información adicional pero en general mucha información del sistema original se pierde al reducirlo a una representación de redes. Esto tiene desventajas pero también tiene ventajas que veremos.

Muchas veces es posible hallar comportamientos emergentes gracias al conjunto de interacciones entre las partes que componen a la red pero que no serian posibles o esperables si solo se estudiasen por separado. A lo largo de los años se han desarrollado herramientas para analizar, modelar y entender las redes. A través de cálculos, operaciones y algoritmos se pueden encontrar características de la red como cuál es el nodo más conectado o la longitud de un camino entre dos nodos. También se han desarrollado técnicas para producir modelos y predicciones sobre procesos que ocurren en la red, como por ejemplo la forma en la que se va a desarrollar el tráfico de internet o la forma en la que se propaga una enfermedad en la sociedad.

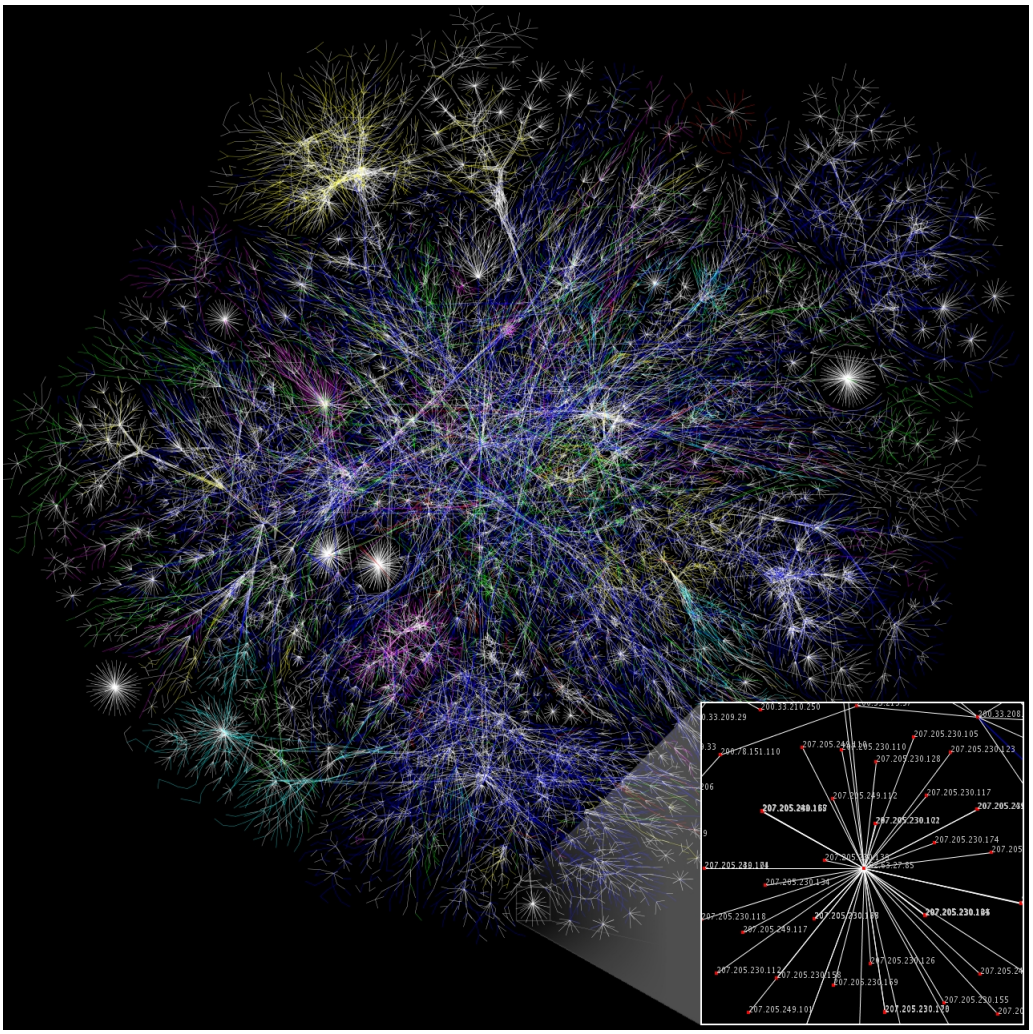


Figura 1.3: Red de internet parcial hasta el año 2005. Los nodos son direcciones de ip.

Como se mencionó anteriormente una red es un conjunto de nodos unidos por enlaces. Estos pueden tener dirección, lo que significa que un enlace que va desde el nodo A hacia el nodo B no es igual a un enlace que va desde el nodo B hacia el A, lo que indica que hay una relación no simétrica entre los nodos. Las redes que tienen enlaces con dirección se llaman redes dirigidas y las que no simplemente redes no dirigidas. Los enlaces también pueden tener asignados números reales indicando la fuerza del enlace entre dos nodos en lugar de solo indicar la existencia o no de una conexión, estos se llaman pesos de los enlaces y las redes que los tienen se llaman redes pesadas.

En general hay a lo sumo un solo enlace entre dos nodos de la red pero puede haber redes donde esto no sucede y hay mas de un enlace entre el mismo par de nodos, estos se llaman enlaces múltiples. Las redes que no tienen enlaces múltiples ni auto enlaces (que conectan un nodo consigo mismo) se consideran redes simples o grafos simples.

Existen varias formas de representar un grafo en el lenguaje matemático. Consideremos por ejemplo una red simple, no dirigida y no pesada de n nodos a los que etiquetamos como

$1, ..., n$. Así es posible etiquetar a los enlaces con tuplas indicando el nodo i desde el que sale y el nodo j hacia el que llega el enlace como (i, j). De esta forma la red completa puede ser descripta dando el número n de nodos y una lista con todos los enlaces de la red. Esta representación es comúnmente usada para el almacenamiento de las redes en computadoras pero para desarrollos matemáticos no es tan conveniente. En su lugar para estos casos se usa otra representación llamada matriz de adyacencia. Para una red simple, no dirigida y no pesada, la matriz de adyacencia A es una matriz binaria y simétrica de elementos A_{ij} tales que $A_{ij} = 1$ si existe un enlace entre los nodos i y j y es cero en otro caso. En el caso de que la red sea pesada la matriz usada es la de pesos W que es igual a la de adyacencia pero en lugar de $1's$ en los elementos distintos a cero tiene los pesos de los enlaces $W_{ij} = \omega_{ij}$, donde ω_{ij} es el peso del enlace entre los nodos i y j. Por ultimo, si la red es dirigida, la matriz de adyacencia ya no es simétrica y por lo tanto $A_{ij} \neq A_{ji}$. En general la convención que se establece sin importar si la red es dirigida o no es que el elemento A_{ij} indica un enlace que va *desde* el nodo j *hacia* el i.

Ahora veamos algunas propiedades estructurales de interés de las redes que son útiles para su caracterización y comparación.

Empecemos por el grado de un nodo, esto es el número de enlaces que se conectan a el. En redes no dirigidas se nota simplemente como k_i y se calcula como $k_i = \sum_j A_{ij}$ pero en redes dirigidas hay que tener en cuenta si los enlaces son entrantes (k_i^{in}) o salientes (k_i^{out}). En una red pesada además se puede calcular la fuerza del nodo que es similar al grado pero suma los pesos de los enlaces conectados a un nodo, $S_i = \sum_j W_{ij}$.

Una de las propiedades más fundamentales de una red es su distribución de grado y ofrece una idea global de las conexiones de la red. Se define p_k como la fracción de nodos de grado k de la red. Al estudiar redes que representan sistemas reales se encontró que por lo general estas tienen una distribución de grado que siguen una distribución de cola pesada, que muchas veces es compatible con una ley de potencias. Esto significa que hay muchos nodos de bajo grado y unos pocos de alto grado llamados hubs que conectan a una gran cantidad de nodos acortando distancias en la red. Una ley de potencias de exponente α sigue la ecuación $P(k) \propto k^{-\alpha}$.

Siguiendo con otras características y definiciones también se encuentra el concepto de componentes de una red. Una componente es un subconjunto de nodos de la red que forman un subgrafo para el cual existe al menos un camino dentro de el entre cualquier par de nodos pero no es posible agregar otro nodo de la red manteniendo esa propiedad. Una red de una sola componente es una red conexa mientras que una dividida en componentes es disconexa. La componente mas grande de la red se llama componente gigante y si la red consta de una sola entonces la componente gigante es igual a la red completa.

En una red también puede haber comunidades definidas por la pertenencia de los nodos a distintos grupos o por compartir características similares. Muchas veces estas en realidad no están definidas y es de interés encontrarlas por lo que se han desarrollados varios algoritmos de detección de comunidades como Louvain, infomap, Leiden, entre otros. Estos intentan agrupar nodos de la red en conjuntos disjuntos buscando características topológicas que sean compartidas dentro de esos conjuntos o se basan en alguna medida de similaridad topológica para agruparlos. Por ejemplo, los algoritmos de Louvain y Leiden se basan en el uso de la modularidad. Esta es una medida global de la estructura de la red y se basa en una división previa de la red en comunidades. Lo que hace es calcular qué tan conectados están los nodos dentro de cada comunidad comparado con una hipótesis nula que corresponde al modelo configuracional (un modelo que intercambia aleatoriamente los enlaces de la red manteniendo los grados de los enlaces) midiendo así la intensidad con la que se divide la red en comunidades. La forma matemática de calcular la modularidad es

$$Q = \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta_{(c_i c_j)},$$

donde m es el número de enlaces de la red y c_i son las comunidades. Esta suma sobre los nodos de la red tendrá términos distintos de cero cuando ambos nodos pertenezcan a la misma comunidad, por lo que se está midiendo el número de enlaces dentro de la comunidad y se la compara con el número de enlaces esperado en el mencionado modelo configuracional.

Por ultimo, se introduce el concepto de superposición topológica (topologial overlap) que será utilizado más adelante. Este es una medida de similaridad entre dos nodos de la red que mide la cantidad de vecinos comunes que tienen dos nodos. La fórmula que define esta cantidad está dada por

$$TO_{ij} = \frac{n_{ij} + A_{ij}}{min(k_i, k_j) + 1 - A_{ij}},$$

donde n_{ij} es el número de primeros vecinos en común entre los nodos i y j. El mínimo valor es $TO_{ij} = 0$ y lo toma cuando $A_{ij} = 0$ (no hay enlace entre i y j) y además $n_{ij} = 0$ (no comparten vecinos). El máximo valor es $TO_{ij} = 1 + \frac{1}{min(k_i,k_j)}$ y lo toma cuando $A_{ij} = 1$ y además $n_{ij} = max(n_{ij}) = min(k_i, k_j)$.

Existe otro tipo de redes llamadas bipartitas que comúnmente se utilizan para representar la pertenencia de nodos en grupos. Estas redes cuentan con dos tipos de nodos que pueden ser divididos en conjuntos disjuntos y los enlaces van solamente entre nodos de diferente tipo. Un ejemplo podría ser una red bipartita donde un tipo de nodo son actores y el otro tipo de nodos son películas con enlaces entre un actor y una película si el actor formo parte de esa película. Un ejemplo de red bipartita se muestra en la figura (1.4a). El equivalente a una matriz de adyacencia para redes bipartitas es la matriz de incidencia I donde las filas se

asocian a un tipo de nodo y las columnas al otro y los elementos I_{ij} son tales que $I_{ij} = 1$ si existe un enlace entre el nodo i y el j. Estas matrices pueden no ser cuadradas y en general no lo son ya que el número de nodos de un tipo no tiene que ser igual al número de nodos del otro tipo.

A partir de una matriz bipartita se puede generar una red que solo contiene a uno de los dos tipos de nodo cuyos enlaces se infieren a partir de la red bipartita, esto se llama proyección sobre uno de los dos tipos de nodos. La forma más sencilla de hacer una proyección es estableciendo un enlace entre dos nodos del mismo tipo si esos dos nodos tienen un enlace con un mismo nodo del tipo opuesto en la red bipartita. Con el ejemplo de actores y películas se podría querer proyectar sobre la red de actores y un enlace entre dos actores se establece si esos dos formaron parte de una misma película. En el esquema de la figura (1.4b) se muestra un ejemplo de proyección.

(a) Ejemplo de una red bipartita con dos tipos de nodos, naranja y azules.

(b) Proyección de la red bipartita sobre los nodos azules.

Figura 1.4

Haciendo una proyección de la forma mencionada se pierde bastante información, una forma de mejorar esto podría ser generar una red pesada donde el peso de los enlaces es el número de nodos que comparten en la red bipartita. Volviendo al ejemplo de los actores y películas, el peso de un enlace entre actores seria el numero de películas en las que aparecieron juntos. Otra forma de proyectar una red bipartita que intenta perder la menor cantidad de información posible al proyectar es conocida como ProbS [16] [17] (probabilistic spreading) y se basa en simular una propagación de información en la red normalizada por el grado del nodo que esta repartiendo la información.

Para comprender mejor esto se presenta el esquema de la figura (1.5), supongamos que la red bipartita tiene dos tipos de nodos A y B y que se quiere proyectar sobre los nodos de tipo A que son los que aparecen arriba en el esquema (figura 1.5a). Se comienza por asignar una cantidad x, y y z a cada nodo como en el esquema, luego se reparten las cantidades asignadas hacia los nodos de tipo B pero normalizadas por los grados de cada nodo de partida como se ve en la figura (1.5b). Finalmente se vuelven a repartir las cantidades que recibieron los

nodos de tipo B hacia los nodos A normalizando ahora por el grado del nodo de partida que en este caso son los de tipo B (figura 1.5c).

Entonces se obtiene una relación entre las cantidades iniciales y finales que en el caso general se puede expresar de forma matricial como

$$x'_i = \sum_j w_{ij} x_j$$

donde en el ejemplo los x'_i son los valores finales y los x_i son los valores iniciales x, y y z, y la matriz w_{ij} está dada por

$$w_{ij} = \frac{1}{k_j} \sum_l \frac{I_{il} I_{jl}}{k_l}$$

con I_{ij} la matriz de incidencia de la red bipartita y k_i el grado del nodo i. La matriz resultante w_{ij} es la matriz de adyacencia que resulta de la proyección y la red proyectada se basa en ella. Esta es una matriz pesada ya que sus elementos son números reales y además es dirigida ya que no es simétrica ($w_{ij} \neq w_{ji}$).

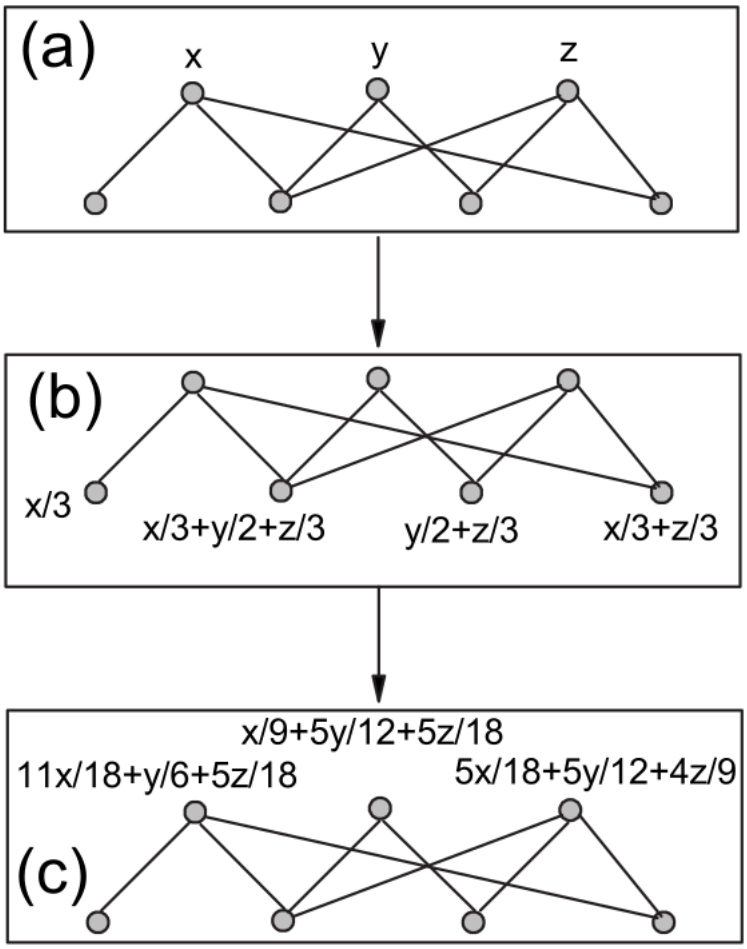


Figura 1.5: Esquema del proceso de proyección de redes bipartitas probabilistic spreading.

1.3. Priorización de genes y enfermedades en redes

La intrincada relación e interacción entre los genes, sus productos y los procesos que llevan a las enfermedades hace que identificar a todos los genes relevantes asociados a una enfermedad sea difícil. Por eso se han desarrollado varias técnicas experimentales para descubrir nuevas asociaciones, como por ejemplo GWA (genome-wide association). Este es un método de alto rendimiento que, analizando la distribución de una patología dentro de un conjunto de individuos filialmente relacionados (es decir, que comparten una parte más o menos sustancial de su bagaje genético), intenta vincular un dado fenotipo con un intervalo del genoma y este puede incluir cientos de genes candidatos. La validación experimental de la relevancia para la enfermedad de cada uno de estos genes es una tarea costosa y que requiere de mucho tiempo y es por ello que se desarrollaron técnicas computacionales de priorización de genes candidatos cuyo objetivo es acotar aún mas la cantidad de genes para proporcionar los mejores candidatos [3] [14] [8] [2].

El problema específico de priorización se puede formular de la siguiente manera: dada una enfermedad (o mas generalmente un fenotipo) de interés y una lista de genes candidatos, el objetivo es identificar potenciales asociaciones entre genes y la enfermedad mediante un ordenamiento de los genes candidatos en orden decreciente según la relevancia que tengan a la enfermedad o fenotipo específico, es decir, producir un ranking de los genes candidatos. La mayoría de los métodos existentes se basan en la información previa existente sobre la enfermedad de interés y los genes sobre los que se conoce y han sido verificados que están asociados a ella.

De esos métodos muchos además usan redes de interacción de proteínas como fuentes de información para encontrar relaciones entre productos de genes candidatos y de genes previamente asociados a la enfermedad de interés.

Estas redes están formadas por nodos que representan proteínas y enlaces que representan interacciones entre ellas. Estas interacciones pueden ser reacciones químicas o físicas pero también relaciones funcionales. Todas estas interacciones son esenciales para llevar a cabo funciones biológicas necesarias para la vida de la célula. Por lo tanto, este tipo de redes resultan de utilidad para la identificación de proteínas involucradas en ciertos procesos o mecanismos que pueden derivar en enfermedades partiendo de algunas proteínas para las que se sabe de esta asociación. Esto se debe a que la cercanía de proteínas en la red indican algún tipo de interacción entre ellas y por lo tanto una potencial participación del proceso en el que se involucran las proteínas de las que se parte.

Los métodos de priorización que utilizan redes de interacción de proteínas se dividen en métodos locales y globales. Los locales se enfocan en la vecindad de los nodos que representan

los productos de genes asociados a la enfermedad de interés mientras que los globales tienen en cuenta la totalidad de la red y su topología. Estos algoritmos funcionan asociando un grado de confianza (un valor numérico) a cada candidato partiendo de la información que ya se tiene sobre la enfermedad y con estos puntajes asociados a los genes se puede elaborar un ranking de los genes más relevantes. Estos métodos suelen estar bajo la categoría de aprendizaje semi-supervisado ya que inicialmente se cuenta con información parcial del problema.

Para este trabajo los métodos de interés son los globales, en particular los que se basan en difusión de información. Consisten en simular un proceso difusivo sobre la red que inicialmente asigna un recurso a aquellos genes de los que se tiene información previa que los asocia con una patología dada y propaga este recurso hacia el resto de los nodos de la red a través de los enlaces siguiendo la topología y la estructura de conexiones. El resultado de la difusión es la cantidad de recurso con la que finalizó el proceso cada gen, esto es lo que se usa para establecer la relevancia de cada uno y poder hacer un ranking. Esta es una manera de relevar distancia en el grafo, con la esperanza de que eso hable de la biología.

En particular, en este trabajo se utilizará la teoría de redes complejas en conjunto con un modelo particular de difusión no lineal en redes para abordar el problema de priorización de genes y enfermedades. El objetivo es poder caracterizar este modelo para emplearlo en la tarea de priorización.

Específicamente, se utiliza una red de interacción de proteínas (PPI) complementada con una red de asociaciones gen-enfermedad (DGI). Utilizando el modelo de difusión seleccionado se propaga un campo escalar de recomendación sobre la PPI a partir del conocimiento previo que provee la DGI. Esto resulta en un puntaje para cada proteína de la red, lo cual permite elaborar un ranking para la recomendación de nuevas asociaciones entre genes y enfermedades.

El trabajo se estructura en 4 capítulos. En el capítulo 2 se introducen los conceptos teóricos de la difusión en redes. Se comienza por conceptos generales para luego abordar otros más específicos de la difusión no lineal. Además, también se explican los métodos de evaluación del modelo de recomendación. En el capítulo 3 se presentan los experimentos realizados para comprender y caracterizar los procesos difusivos. Para ello primero se introduce un modelo para generar redes sintéticas con ciertas características deseadas que sirven como un entorno de prueba controlado. Con estas redes de control se ajustan ciertos parámetros del modelo de difusión y se hacen experimentos para comprender el comportamiento de los procesos difusivos. Finalmente, en el capitulo 4 se introducen las redes utilizadas para tratar el problema de priorización, los detalles específicos de la evaluación y los resultados obtenidos al aplicar los algoritmos de difusión. Las redes que se utilizan son la de interacción de proteínas y la de asociaciones entre genes y enfermedades. Para evaluar se utiliza la métrica de AUC de la curva ROC calculada a partir de los rankings generados con los resultados de

la difusión.

Capítulo 2

Métodos

2.1. Difusión

La difusión es, entre otras cosas, el proceso por el cual un gas se mueve de regiones de alta densidad a regiones de baja densidad, llevado a cabo por la presión relativa de las diferentes regiones.

La difusión es un proceso físico que consiste en el flujo neto de algo (partículas, gas o alguna otra especie) dentro de un material o medio y es caracterizado por el desplazamiento desde regiones de alta concentración a regiones de baja concentración de lo que se difunde y es inducido por el gradiente de concentración. Describe el comportamiento macroscópico que resulta de los procesos microscópicos en el material o medio en el que se da la difusión.

La ecuación que gobierna la dinámica de la difusión está dada por:

$$\frac{\partial \phi(\mathbf{r}, t)}{\partial t} = \nabla[D(\phi, \mathbf{r})\nabla\phi(\mathbf{r}, t)],$$

donde $\phi(\mathbf{r}, t)$ es la densidad de lo que se está difundiendo en la posición $\mathbf{r}$ y tiempo t y $D(\phi, \mathbf{r})$ es el coeficiente de difusión que en el caso general puede depender tanto de la densidad como de la posición (del medio). Cuando el coeficiente de difusión es constante la ecuación se simplifica a:

$$\frac{\partial \phi(\mathbf{r}, t)}{\partial t} = D\nabla^2\phi(\mathbf{r}, t).$$

Es posible considerar procesos difusivos en redes [ver 10, cap. 6.13] y estos pueden usarse como modelos de propagación en la red, como por ejemplo la propagación de una idea en redes sociales o el esparcimiento de un virus en la sociedad. Supongamos que se tiene una cierta cantidad o sustancia y se puede cuantificar como una función o campo escalar sobre los nodos de la red con una cantidad ψ_i sobre el nodo i. Además supongamos que esta sustancia puede esparcirse de un nodo hacia sus vecinos a través de los enlaces a un ritmo dado por $C(\psi_j - \psi_i)$ donde C se conoce como constante de difusión y j es un vecino de i. Esto se interpreta como la cantidad que entra o sale del nodo i hacia el j, si $\psi_i > \psi_j$ la cantidad

es negativa y se va de i mientras que en el caso contrario es positiva y llega hacia i. Por lo tanto en un intervalo de tiempo dt, la variación de ψ_i debido a todos sus vecinos es

$$d\psi_i = C \sum_{j \in N_i} (\psi_j - \psi_i) dt,$$

donde N_i es el conjunto de vecinos del nodo i. Reformulando la ecuación se obtiene

$$\frac{d\psi_i}{dt} = C \sum_j A_{ij}(\psi_j - \psi_i),$$

donde se usa la matriz de adyacencia A_{ij} para sumar sobre los vecinos de i. Esta ecuación se puede reescribir para encontrar una nueva expresión

$$\frac{d\psi_i}{dt} = C \sum_j A_{ij}\psi_j - C\psi_i k_i = C \sum_j (A_{ij} - k_i\delta_{ij})\psi_j,$$

donde se usó que $k_i = \sum_j A_{ij}$. Finalmente se puede escribir en forma vectorial

$$\frac{d\psi}{dt} = C(A - D)\psi,$$

donde se introdujo la matriz diagonal de grados $D_{ij} = k_i\delta_{ij}$. La matriz $L = D - A$ se define como el laplaciano combinatorio o simplemente laplaciano de la red. Finalmente la ecuación toma la forma

$$\frac{d\psi}{dt} = -CL\psi,$$

donde se puede ver la similaridad con la ecuación de difusión en el caso continuo identificando a L con el operador ∇^2.

Consideremos un vector u en donde cada posición i representa un nodo de la red y el valor u_i representa la cantidad de un cierto recurso, entonces cada nodo de la red tiene asignada una cantidad de este. Para ver la interpretación del laplaciano es conveniente ver el resultado para el nodo i de aplicar L sobre el vector u:

$$(Lu)_i = \sum_j L_{ij}u_j = \sum_j (D_{ij} - A_{ij})u_j = \sum_j (k_i\delta_{ij} - A_{ij})u_j = k_i u_i - \sum_{j \in N_i} u_j = \sum_{j \in N_i} (u_i - u_j),$$

donde N_i son los vecinos del nodo i. Ahora, escribamos la compoonente i de la ecuación de difusión discretizando el tiempo con el uso de diferencias finitas donde $\frac{du}{dt} = \frac{u(t+h)-u(t)}{h}$ donde h es un intervalo pequeño de tiempo:

$$u_i(t+h) = u_i(t) - h(Lu(t))_i = u_i(t) - h \sum_{j \in N_i} (u_i - u_j),$$

Esto nos dice cual es la cantidad de recurso que va a recibir el nodo i en un momento de tiempo posterior a t y depende por un lado de la cantidad que tenia el nodo en tiempo t y tambien del resultado de aplicar el laplaciano a $u(t)$.

De la ecuación anterior se ve que aplicar el laplaciano sobre el vector de recurso inicial le asigna a cada nodo una cantidad que es la diferencia de recurso de ese nodo con cada uno de sus vecinos. Entonces finalmente la cantidad que recibe el nodo i en tiempo $t + h$ será la cantidad que tenia a tiempo t y se suma o resta un valor proporcional a las diferencias de recurso con sus vecinos a tiempo t, recibiendo algo positivo en caso de tener menos cantidad y negativo en el caso contrario. Por lo tanto se recupera esa noción de que el recurso fluye desde nodos donde está mas concentrado hacia nodos en donde lo esta en menor medida.

En la figura (2.1) se muestra un esquema con una red de cuatro nodos A, B, C y D en donde se indica con barras verticales el valor de una función f definida sobre ellos. Se ve que por ejemplo el nodo B tiene mas cantidad de recurso que sus vecinos A y C por lo que parte de sus recursos irán hacia ellos. En definitiva, el laplaciano nos da la cantidad en la que debería cambiar el recurso de un nodo y la ecuación indica como evoluciona en el tiempo el proceso.

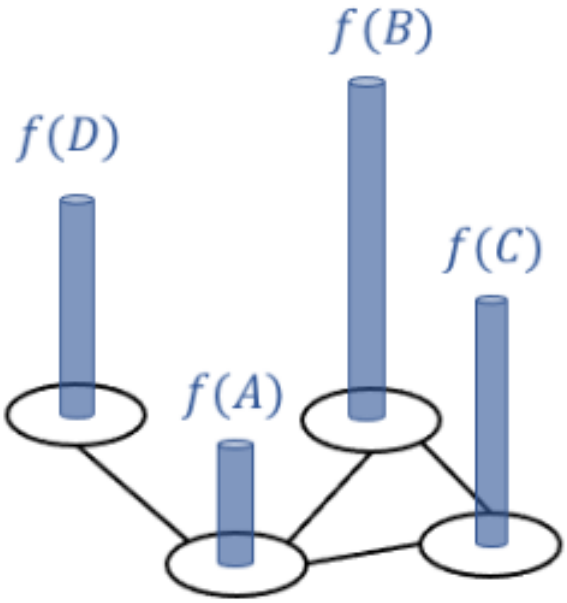


Figura 2.1: Ilustración de una función f definida sobre el grafo de nodos A, B, C, D.

2.2. Operadores diferenciales en redes y difusión no lineal

En la sección anterior se presentó brevemente la intuición o interpretación del proceso de difusión. En esta sección se introduce un proceso similar pero con algunas modificaciones y se pretende hacerlo más formalmente a través de operadores diferenciales definidos en grafos [ver 11, cap. 13]. Las modificaciones que se hacen respecto al proceso introducido

anteriormente están relacionadas a la normalización por grado de los operadores y también a la introducción del caso no lineal.

Antes de definir los operadores son necesarios algunos conceptos como los de espacios de funciones definidas sobre enlaces y nodos del grafo, $\mathcal{H}(E)$ y $\mathcal{H}(V)$ respectivamente. Además también la definición de producto interno entre funciones definidas sobre los nodos $\mathcal{H}(V)$ (análogo para funciones definidas sobre los enlaces $\mathcal{H}(E)$) como

$$\langle f, g\rangle_{\mathcal{H}(V)} = \sum_{v\in V} f(v)g(v),$$

Se define el gradiente sobre el grafo como el operador $\nabla : \mathcal{H}(V) \to \mathcal{H}(E)$. Este opera sobre funciones definidas en los nodos y devuelve una función definida sobre enlaces:

$$(\nabla f)([u, v]) = \sqrt{\frac{a[u, v]}{d(v)}} f(v) - \sqrt{\frac{a[u, v]}{d(u)}} f(u),$$

donde $[u, v]$ es el enlace que conecta a los nodos u y v, la función es tal que $f \in \mathcal{H}(V)$, $a[u, v]$ es el elemento de la matriz de adyacencia (puede ser un grafo pesado) que corresponde a los nodos u, v y $d(v)$ es el grado del nodo v. La interpretación del gradiente es similar a la de su análogo continuo, este operador aplicado sobre la función f y evaluado en el enlace $[u, v]$ es una medida del cambio de la función a lo largo del enlace o entre dos nodos conectados por el. El valor de la función en cada nodo se pesa con el inverso del grado del nodo para que nodos de distinta conectividad sean comparables.

Continuando con las definiciones, la divergencia es un operador que se define como el negativo del adjunto del gradiente del grafo $div : \mathcal{H}(E) \to \mathcal{H}(V)$. Toma funciones definidas sobre los enlaces y devuelve una función definida sobre los nodos:

$$\langle \nabla f, h\rangle_{\mathcal{H}(E)} = \langle f, -div(h)\rangle_{\mathcal{H}(V)},$$

para cualquier $f \in \mathcal{H}(V), h \in \mathcal{H}(E)$. El lado izquierdo de la ecuación indica un producto interno entre funciones definidas en el espacio $\mathcal{H}(E)$ mientras el lado derecho lo hace sobre el espacio $\mathcal{H}(V)$.

Usando la definición de producto interno en cada espacio junto con la definición de gradiente se puede encontrar la expresión para la divergencia:

$$(div\ h)(v) = \sum_{u\sim v} \sqrt{\frac{a[u, v]}{d(v)}} \left(h[v, u] - h[u, v]\right),$$

donde $u \sim v$ corresponde a los nodos u adyacentes a v (primeros vecinos de u) y $h \in \mathcal{H}(E)$. Al igual que sucede con el gradiente, la interpretación de la divergencia en grafos tiene un

parecido con su interpretación continua. Este operador es una medida de cuánto de la cantidad que representa h 'entra' o 'sale' del nodo en el que se evalúa y como en el gradiente también se hace una ponderación según el grado del nodo.

Finalmente, teniendo los operadores de gradiente y divergencia es posible definir el operador laplaciano $\Delta : \mathcal{H}(V) \to \mathcal{H}(V)$ como $\Delta f = -\frac{1}{2} div(\nabla f)$. Este toma funciones definidas sobre nodos y devuelve una función también definida sobre ellos. Usando las definiciones previas de gradiente y divergencia se puede obtener la expresión para el laplaciano:

$$(\Delta f)(v) = f(v) - \sum_{u \sim v} \frac{a[u,v]}{\sqrt{d(v)d(u)}} f(u).$$

Esta ecuación vectorial se puede expresar de forma matricial como $(\Delta f)(v) = (\mathcal{L}u)(v)$ donde u es un vector en el que cada posición (asociada a un nodo de la red) contiene el valor de la función f en el nodo, $u(v) = f(v)$. La matriz $\mathcal{L}$ es lo que se conoce como laplaciano normalizado y su definición es $\mathcal{L} = D^{-1/2} L D^{-1/2}$ donde D es la matriz diagonal de grados $D(u,v) = d(v)\delta(u,v)$ y L es el laplaciano del grafo ya introducido.

Entonces la interpretación de aplicar el laplaciano a la función f en el nodo v es que este es una medida de cuánto varía $f(v)$ con respecto a sus vecinos como ya se expresó en la sección anterior, pero en este caso los valores están pesados por los grados de ambos nodos. Lo que esto significa es que un nodo de alto grado será penalizado tanto al repartir recurso a otros nodos como al recibir recurso de otros nodos. Esto intenta evitar que nodos de alto grado reciban rápidamente mucho mas recurso debido a tener muchas conexiones como se sabe que sucede con el laplaciano sin normalizar [15].

Ya con la definición de este operador podemos expresar la ecuación de difusión vista antes haciendo uso de los operadores definidos acá

$$\frac{du}{dt} \propto \Delta u,$$

donde u es una función definida sobre los nodos del grafo. Esta ecuación representa cómo se reparte una cantidad que inicialmente se encuentra concentrada en algunos nodos (condición inicial de la difusión) y al pasar el tiempo se distribuye sobre el resto de los nodos de la red.

Para resolverla numéricamente de forma iterativa se puede utilizar el método de diferencias finitas para discretizar el tiempo

$$u_{t+h} = u_t + h\mathcal{L}u_t,$$

donde u_t es el vector que contiene los valores de la función en los nodos a tiempo t y h es un intervalo temporal que se da a cada paso de la iteración.

En este trabajo es de interés usar algoritmos de difusión no lineal. El modelo mas simple de difusión no lineal es el de medios porosos que para el caso continuo toma la forma de la siguiente ecuación

$$\frac{\partial \phi}{\partial t} = \Delta(\phi^p),$$

donde ϕ es la densidad que se difunde y $p > 1$. El caso análogo para redes [12] toma la forma

$$\frac{du}{dt} \propto \Delta u^p,$$

y en este caso p es un número real positivo y u una función definida sobre los nodos como ya se viene explicando. Entonces haciendo diferencias finitas como se mostró antes se obtiene

$$u(t+h) = u(t) + h\Delta u(t)^p = u(t) + h\mathcal{L}u(t)^p,$$

donde se usó el laplaciano normalizado $\mathcal{L}$.

Este modelo modifica los valores de la cantidad que contiene cada nodo previamente a ser difundido y es el modelo no lineal que se utiliza en el trabajo. Las difusiones que toman lugar con valores de $p < 1$ se consideran difusiones rápidas ya que como el vector de recurso es una densidad está normalizado a la unidad y cada elemento pertenece al intervalo [0, 1] por lo que el efecto de elevarlos a una potencia menor a 1 es el de aumentar los valores. Esto hace que se reparta más recurso en cada paso temporal. Por otro lado, cuando $p > 1$ son difusiones lentas ya que el efecto es el contrario al caso previamente mencionado. Por último, cuando $p = 1$ corresponde al caso lineal.

La implementación de estos algoritmos para el problema que se quiere resolver es encontrando la solución a la ecuación de forma iterativa partiendo de una condición inicial dada por el vector u_0 ($u(t = 0)$). En este se codifica la información que se tiene previamente del problema y del cual se desprende la difusión.

El problema particular que se aborda en esta tesis es el de priorización de genes asociados a enfermedades utilizando una red de interacción de proteínas. Es sobre esta última que se realiza la difusión y la condición inicial viene dada por un conjunto de genes que se sabe que están asociados a una enfermedad. Los genes se corresponden a proteínas y estas a los nodos de la red. Este conjunto de genes se denomina conjunto semilla S y el vector u_0 contiene en la posición i asociada al gen o nodo i de la red un valor dado por $u_0^i = \frac{1}{|S|}$ en el caso en que

$i \in S$ y es cero sino, donde $|S|$ es el tamaño del conjunto S. Entonces la condición inicial es que todos los nodos semillas comiencen con la misma cantidad de recurso y normalizado a la unidad.

2.3. Evaluación

El problema de encontrar nuevas asociaciones entre genes y enfermedades corresponde a uno de clasificación y recomendación ya que lo que se intenta es proponer nuevos genes (recomendar) que pertenezcan a la categoría de estar asociados a una enfermedad (clasificar).

Es común en este tipo de problemas y en general cuando se hace aprendizaje automático utilizar metodologías de evaluación y control para poder medir y comparar la performance de los algoritmos.

Para el control se suele utilizar una técnica de separación de los datos llamada validación cruzada que consiste en separar a los datos en dos conjuntos llamados conjuntos de entrenamiento y de evaluación, normalmente siendo el primero de mayor tamaño, por ejemplo, una partición 80 % entrenamiento y 20 % evaluación.

Luego se aparta el conjunto de evaluación y se trabaja solamente sobre los datos del conjunto de entrenamiento dentro del cual es habitual utilizar una técnica llamada k-folds. Esta consiste en volver a particionar el conjunto de entrenamiento pero ahora varias veces (k veces) de forma aleatoria en conjuntos llamados de entrenamiento y validación como se muestra en el esquema de la figura (2.2), por ejemplo, una partición 80 %, 20 % como en el ejemplo de entrenamiento-evaluación.

Entonces son estas particiones las que se utilizan para entrenar y medir como responde el modelo, se entrena en el subconjunto de entrenamiento y se prueba su rendimiento en el conjunto de validación repitiendo esto en los k subconjuntos en los que se haya particionado el conjunto de entrenamiento original para finalmente obtener un valor medio de las métricas que se utilizan para medir la performance.

Una métrica muy utilizada para caracterizar el rendimiento del modelo es lo que se conoce como AUC (área por debajo de la curva) de la curva ROC (receiver-operating characteristic). Esta curva se utiliza para saber que tan bien un modelo distingue los casos favorables de los que no lo son en problemas de clasificación binario.

Para ver cómo se utiliza y se calcula supongamos un problema que consta de un total T de elementos a clasificar dentro de los cuales se sabe de antemano que P son positivos y T - P son negativos. En el caso particular de priorización se cuenta con una lista ordenada

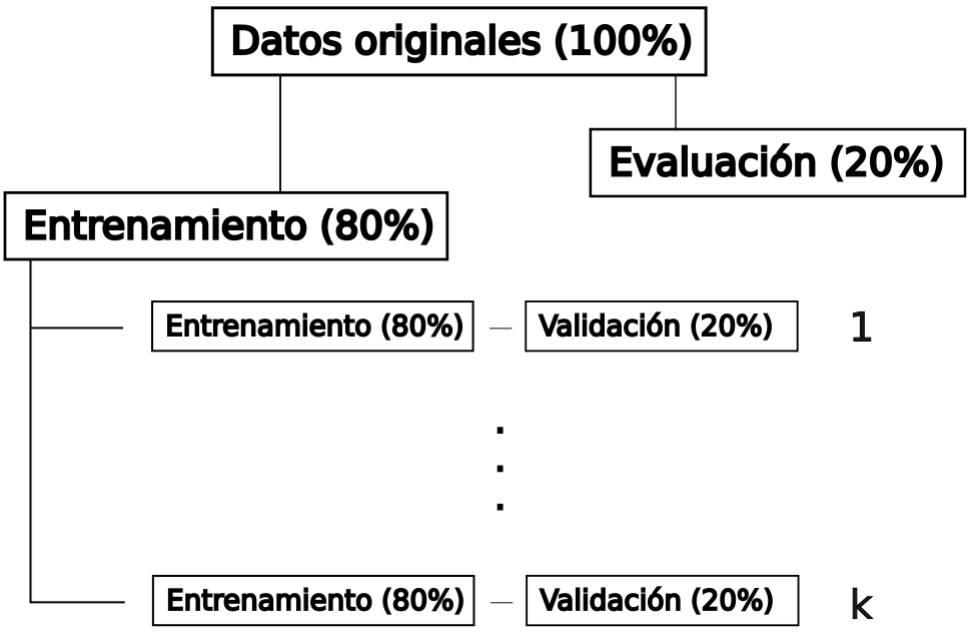


Figura 2.2: Esquema de la forma en la que se separan los datos llamada validación cruzada con k-folds.

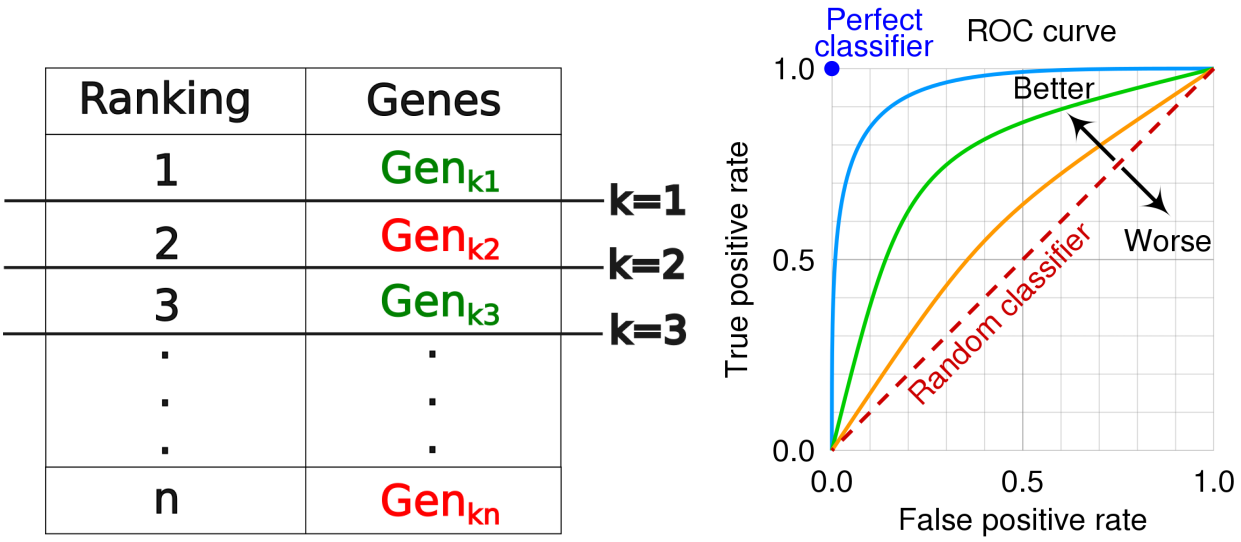
según la relevancia o confianza que se tiene de los elementos de ser positivos, esta lista es el resultado de algún modelo de priorización.

Ahora supongamos que se define un modelo que clasifica a los primeros k elementos como positivos, es decir, se fija un umbral en el elemento k del ranking para clasificar como positivos a todos los elementos que se encuentran por encima de k en el ranking. Dentro de estos k elementos entonces habrá alguna cierta cantidad de los que se conocía previamente como positivos, estos entonces serán clasificados por el modelo como verdaderos positivos (true positives) TP_k (los datos que el modelo clasificó como positivos y efectivamente lo eran) y también habrá una cantidad de falsos positivos (false positives) FP_k (los datos que el modelo clasificó como negativos pero que en realidad eran positivos). Con estas cantidades se pueden calcular las proporciones (true positives rate) $\mathrm{TPR}_k = \frac{TP_k}{P}$ y (false positives rate) $\mathrm{FPR}_k = \frac{FP_k}{T-P}$ para el modelo que toma como umbral el valor k.

La curva ROC consiste en variar el umbral de k desde 1 hasta el total de elementos del ranking y graficar los las cantidades de TPR vs. FPR para cada k. Estos ratios también se suelen llamar sensibilidad y 1-especificidad respectivamente. Este proceso por el cual se calculan las curvas se esquematiza en la figura (2.3a) en la que se muestra un ranking genérico de genes que puede ser resultado de algún modelo de priorización. En el se indican en color verde genes que son positivos y en rojo genes que son negativos. Además se muestran diferentes valores de k, osea diferentes umbrales. Así es como para cada umbral se calculan las cantidades TP_k y FP_k, por ejemplo, para $k = 1$ se tiene un solo gen verde y ninguno

rojo por lo que $\mathrm{TP}_{k=1} = 1$ y $\mathrm{FP}_{k=1} = 0$, para $k = 2$ se tiene uno de cada color por lo que $\mathrm{TP}_{k=1} = 1$ y $\mathrm{FP}_{k=1} = 1$, etc.

Para obtener una cantidad escalar con la que es más fácil comparar diferentes modelos se calcula el área bajo la curva ROC, abreviada como AUC. Mientras mayor sea el valor de esta métrica significa que el modelo utilizado fue capaz de posicionar a los elementos que verdaderamente eran positivos en los lugares mas altos del ranking. Cabe notar que estas medidas pueden verse fuertemente afectadas cuando la cantidad de positivos y negativos en los datos no está balanceada ya que si por ejemplo se tienen muchos negativos y pocos positivos será mas difícil para el modelo posicionar en mejores posiciones a los verdaderos positivos resultando así en valores bajos de AUC.



(a) Esquema de un ranking y como construir a partir de el la curva ROC. Los genes de color verde representan a positivos mientras que los rojos representan negativos. Los valores de k son los diferentes valores del umbral.

(b) Esquema de diferentes curvas ROC donde se indica el modelo aleatorio y curvas que se obtienen de mejores modelos.

Figura 2.3

En la figura (2.3b) se muestran varios ejemplos de curvas ROC para diferentes escenarios. La linea recta roja de pendiente 1 es la curva ROC esperada para un modelo donde el clasificador es aleatorio mientras que las curvas de color naranja, verde y celeste son curvas ROC para mejores modelos que el aleatorio ya que cada una tiene un valor cada vez mas grande del AUC. Finalmente el punto azul representa un modelo en el que el clasificador es perfecto teniendo un porcentaje del 100 % de aciertos.

Capítulo 3

Experimentos

En esta sección se explican los experimentos realizados para comprender el comportamiento de los algoritmos de difusión. La idea es poder hacer estadística aplicando los algoritmos repetidas veces y evaluando diferentes cantidades y métricas obtenidas de cada realización. Específicamente se introduce un modelo para producir redes sintéticas, que se utilizan en los experimentos y se definen los parámetros del algoritmo no lineal presentado en el capitulo anterior. Luego se presenta la forma en la que se eligen los nodos semillas a partir de los cuales se hace la difusión. Posteriormente se introducen las cantidades que se van a utilizar para estudiar los algoritmos y finalmente se presentan los resultados.

3.1. Redes sintéticas

Para poner a prueba los algoritmos de difusión que se quieren estudiar una buena idea para comprender sus funcionamientos es estudiarlos en casos pequeños y que sean controlados pero que se parezcan lo mas posible a los escenarios reales. En el caso particular de este trabajo eso significa usar redes de menor cantidad de nodos y enlaces que las redes reales en las que se pretende poner a prueba los algoritmos. Para ello se diseñó un modelo para producir redes sintéticas que se asemejan a las redes reales de interés en algunos aspectos y que presentan características de interés controlables mediante parámetros. Será en estas redes donde se hace un estudio preliminar de los algoritmos y donde se intentará comprender el funcionamiento de los métodos de difusión no lineales introducidos previamente.

La idea del modelo es poder generar redes que tengan una distribución de grado de tipo ley de potencias ya que está es la distribución de grado que aparece en las redes reales de interés siendo importante también poder controlar la modularidad. Para ello el modelo se basa en el uso de redes bipartitas y su posterior proyección sobre un tipo de nodo haciendo uso de la propiedad de una red bipartita de que si los nodos de un tipo en la red bipartita siguen una distribución de grado de ley de potencias, al proyectarla se mantiene esta distri-

bución con el mismo exponente [5] [4].

Primero, es necesario construir una red bipartita de ciertas características. Se comienza por definir el número de nodos n de la red final que se quiere generar, esta será la cantidad de nodos de un tipo, digamos de tipo A, en la red bipartita. Luego se toma una muestra de tamaño n a partir de una distribución de tipo ley de potencias con el exponente que se desee, es decir, se toman n números k_i con $i = 1, ..., n$ tales que su distribución sigue una ley de potencias con un dado exponente.

Esos números se asignan a cada nodo de tipo A y serán sus grados. El número de nodos del otro tipo de la red, digamos tipo B, es un parámetro a definir que se representa con m. Resta establecer los enlaces de la red respetando la distribución de grados asignada a los nodos de tipo A y también pensando en una forma en la que se pueda controlar la modularidad de la red.

Para ello primero se dividen a ambos tipos de nodos en g comunidades $c = 1, ..., g$ y se establece un parámetro q que corresponde a la probabilidad de que un nodo (de un tipo) de una dada comunidad tenga un enlace con otro de la misma comunidad (del tipo opuesto). Lo siguiente es recorrer cada nodo de tipo A y generar sus enlaces de acuerdo a su grado y su comunidad.

El nodo i-ésimo de tipo A tiene grado k_i y pertenece a la comunidad c_i. Para establecer sus k_i enlaces, se toma una muestra de ese tamaño de nodos de tipo B. Esta se realiza según una distribución en la que los nodos de tipo B que pertenecen a la comunidad c_i (la misma que el nodo i de tipo A) tienen una probabilidad de ser muestreados de q. Tomar la muestra de esta manera es lo que permite controlar la modularidad a través del parámetro q. Este proceso se repite para cada uno de los n nodos de tipo A.

De esta manera se arma una red bipartita con n nodos de tipo A que siguen una distribución de grado de tipo ley de potencias y m nodos de tipo B, ambos conjuntos de nodos divididos en g comunidades y con una probabilidad de enlace q entre nodos de una misma comunidad (de distinto tipo). Un esquema simplificado del procedimiento se puede ver en la figura (3.1).

El siguiente paso es proyectar esta red bipartita sobre los nodos de tipo A. Como se mencionó en la introducción hay muchas formas de proyectar una red bipartita, sin embargo, al proyectar de alguna de estas formas se pierde mucha información de la red bipartita por lo que en este caso la proyección a utilizar es la que se presentó en la introducción como probabilistic spreading y su objetivo es tratar de perder menor cantidad de información en la proyección. Este método produce una red pesada y dirigida pero para nuestro uso se transforma a una red pesada y no dirigida simetrizando la matriz mediante el peso medio de los enlaces, es decir, $w'_{ij} = (w_{ij} + w_{ji})/2$ donde w'_{ij} es la matriz de adyacencia pesada no dirigida. Además de esto, también se calcula la componente gigante de la red y es la que se

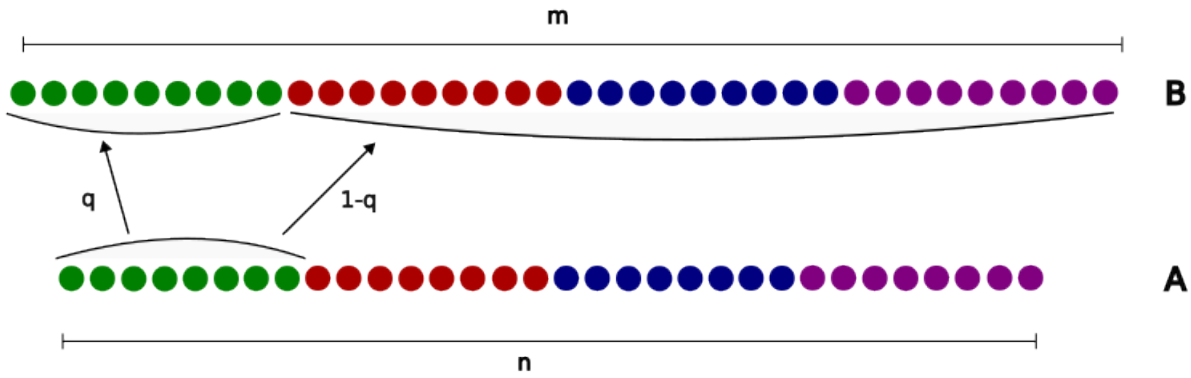


Figura 3.1: Esquema del proceso para construir la red bipartita para el modelo que produce redes sintéticas.

usa como la red que genera el modelo.

Teniendo un modelo para generar redes sintéticas lo siguiente que se hizo es generar un ensamble de redes variando los parámetros del modelo para estudiar cómo estos afectan en su estructura. El exponente elegido para la distribución de tipo ley de potencias ($P(k) \propto k^{-\alpha}$) usada para tomar una muestra de los grados de los nodos de tipo A es $\alpha = 2$ y el número de comunidades g en las que se dividen los conjuntos de nodos es $g = 4$, estos dos parámetros se mantienen fijos. Los valores tomados para la probabilidad q de enlace entre un par de nodos de la red bipartita de la misma comunidad (de distintos tipos) son $q = 0{,}25, 0{,}5, 0{,}7, 0{,}9, 0{,}99$. Finalmente para establecer las cantidades de nodos de cada tipo n y m se decidió tomar pares de valores de la forma (n, m) usando específicamente los valores $(400, 200), (400, 400), (200, 400), (400, 2000)$ dando lugar a los cocientes $m/n = 0{,}5, 1, 2, 5$. Entonces se tienen 5 valores de q y 4 valores del par (n, m) que dan en total 20 combinaciones. Además por cada una de las combinaciones se hicieron 200 redes por lo que en total se tienen 4000 redes.

La idea es encontrar dos cantidades que describan diferentes tipos de heterogeneidades que podrían ser relevantes. Se quiere formar una especie de diagrama de fases del ensamble de redes con el cual se pueda entender la estructura de la red a partir de su posición en este. Las cantidades que se encontraron son la modularidad y el coeficiente de variabilidad CV (valor medio / desviación estándar) de los pesos de los enlaces de la red. La modularidad describe heterogeneidad topológica mientras que el CV lo hace en los pesos.

El resultado de este diagrama se puede ver en la figura (3.2) donde se hace el gráfico de modularidad vs. CV para cada red del ensamble. En el se separa por color, tamaño y forma a los puntos según los diferentes parámetros utilizados, con colores se indican los valores de q, con formas se indican los cocientes m/n y con el tamaño de los puntos se indica el tamaño

de la componente gigante de la red respecto al de la red original. En el se puede ver cómo los valores de q tienen un efecto directo sobre la modularidad, incrementando cuando q lo hace y formando ciertos niveles de modularidad para cada valor del parámetro. Además se ve también como al aumentar el valor del cociente m/n aumenta la variabilidad de los pesos en los enlaces, el CV, ya que se ve que la forma de los puntos aparecen en general por secciones y en intervalos de CV. Por ultimo, parece verse que no hay demasiadas redes con un tamaño de la componente gigante menor al 75 % de la red original (se quiere que sean mayores a este tamaño) y las que hay aparecen en valores altos de CV. A partir de este diagrama se pueden tomar redes en diferentes zonas para explorarlas, en el gráfico se pueden ver algunos círculos rojos, estos corresponden a las redes que se muestran en las figuras (3.3 y 3.4).

Estas redes se tomaron para ejemplificar y mostrar cómo se ven los grafos que se van a utilizar posteriormente para realizar experimentos y la idea es mostrar algunos pertenecientes a diferentes zonas del diagrama. Los colores de los nodos en esas redes corresponden a las comunidades que fueron asignadas inicialmente en la red bipartita original y son las que se usan para calcular la modularidad. Se puede ver cómo las comunidades se separan muy claramente para las redes que tienen mayor modularidad como era de esperar, habiendo menos cantidad de enlaces entre comunidades diferentes para la red que tiene mayor CV. Además también sucede que cuando el número de nodos de tipo B utilizados es mayor a los de tipo A las redes resultan mas esparsas (de baja densidad) y sus componentes gigantes tienen un tamaño menor al de la red completa. Por último, aún utilizando valores del parámetro q mayores a 0.25 (enlaces entre nodos de la misma comunidad tienen la misma probabilidad que enlaces entre nodos de diferentes comunidades) se ve que la modularidad no aumenta rápidamente sino que el parámetro tiene que tomar valores mas cercanos a 1 para que las comunidades se separen claramente.

Para los experimentos que se realizan mas adelante se utilizan dos de estas redes ya que se quieren tomar unas que difieran en su estructura para poder comparar como funcionan los algoritmos sobre ellas. Las utilizadas serán las que se muestran en la figura (3.3), estas cuentan con un CV similar alrededor de 4 y difieren notablemente en su modularidad.

3.2. Elección de parámetros

Para terminar de explicar los detalles de la implementación de los algoritmos aún quedan parámetros por determinar. Estos son los valores de p (exponente de la difusión), el número de pasos temporales totales de la difusión y el tamaño del intervalo temporal h de cada paso.

Para p se tomó un valor para cada caso de difusiones no lineales, rápido y lento como se explicó en el capitulo anterior y también el caso lineal con el cual comparar. Específicamente se eligieron los valores $p = 0.5$, 1 y 2 que corresponden a cada uno de los casos mencionados

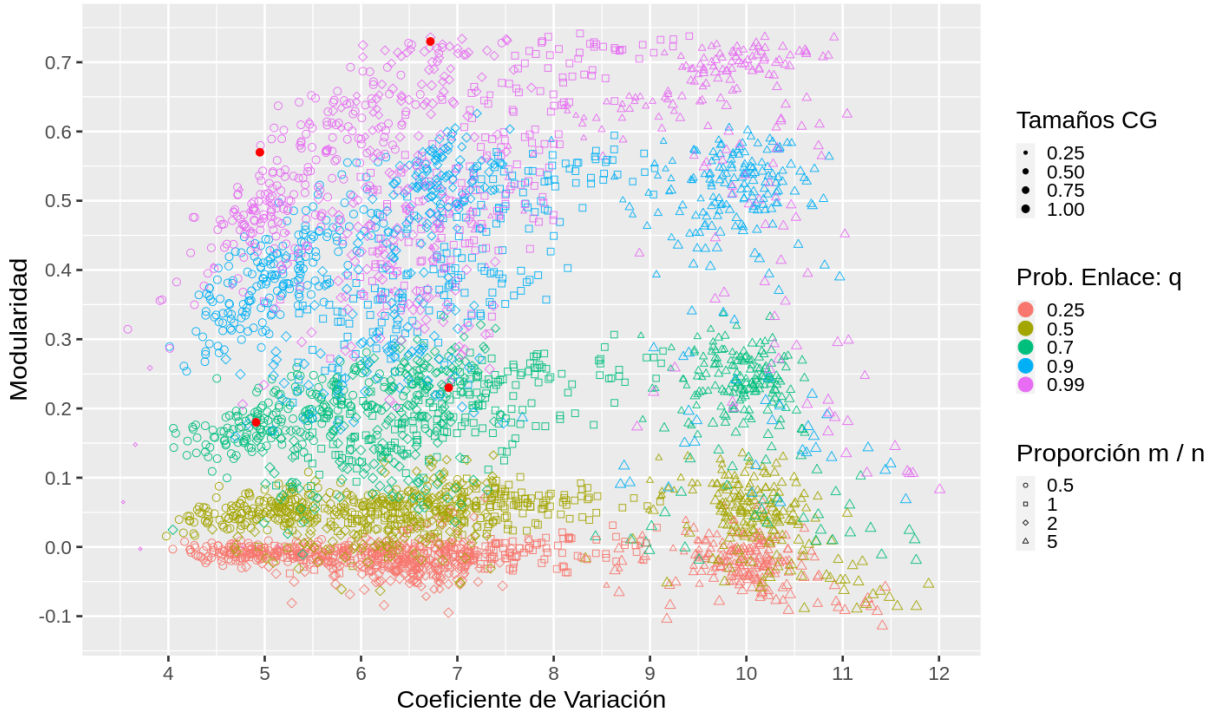
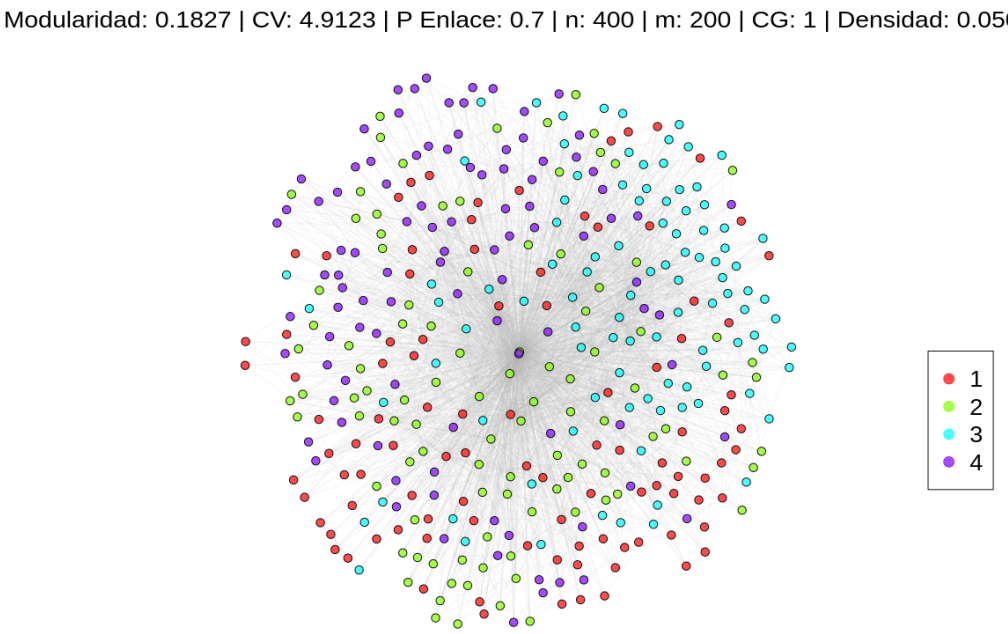


Figura 3.2: Gráfico de modularidad vs. coeficiente de variación de los pesos de los enlaces (CV) de las redes obtenidas en un ensamble donde se varían los parámetros q y m/n. Se separan los puntos por color, forma y tamaño según los valores de los parámetros.

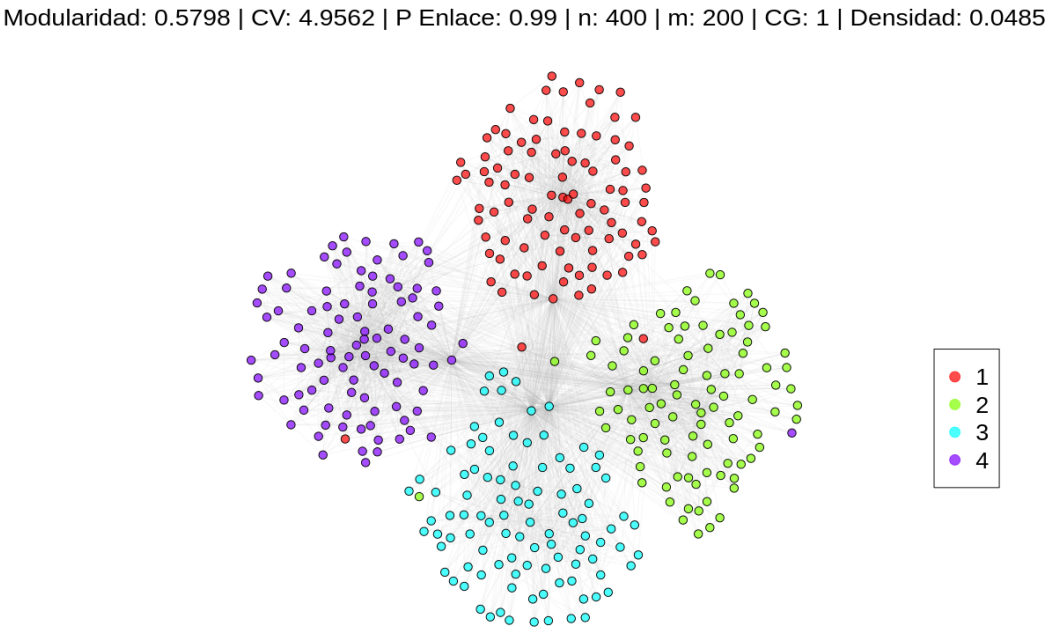
(rápido, lineal y lento respectivamente).

Para la elección de los dos parámetros restantes primero se hizo un estudio de las curvas obtenidas por difusión en diferentes casos. Recordando que el resultado de la difusión es un vector cuyos elementos representan la cantidad de recurso asignadas a los nodos de la red al finalizar la difusión, se puede ordenar este vector de forma decreciente para obtener el ranking de nodos y hacer un gráfico de recurso o 'score' o 'puntaje' vs. el ranking, son estos gráficos a los que nos referimos como curvas de difusión. Como ejemplo de estas curvas se puede ver la figura (3.5), donde hay 3 curvas cada una correspondiente con un valor de p de los valores mencionados previamente. Esta curva fue hecha difundiendo en una de las redes elegidas para los experimentos, la de menor modularidad, se tomaron dos nodos de la red de forma aleatoria para usarlos como semillas de la difusión y se usó una grilla de valores de la cantidad de pasos y el tamaño temporal del paso.

Lo que se está haciendo es resolver la ecuación diferencial de forma numérica. Valores pequeños de h permitirán que las soluciones aproximadas se acerquen mejor a la verdadera solución pero el problema con esto es que si h es chico se necesitan de muchos pasos para que avance el tiempo significativamente por lo que se quiere buscar un paso que sea suficientemente pequeño como para obtener una solución adecuada pero que no sea tan chico como para requerir muchos pasos y que sea computacionalmente costoso.

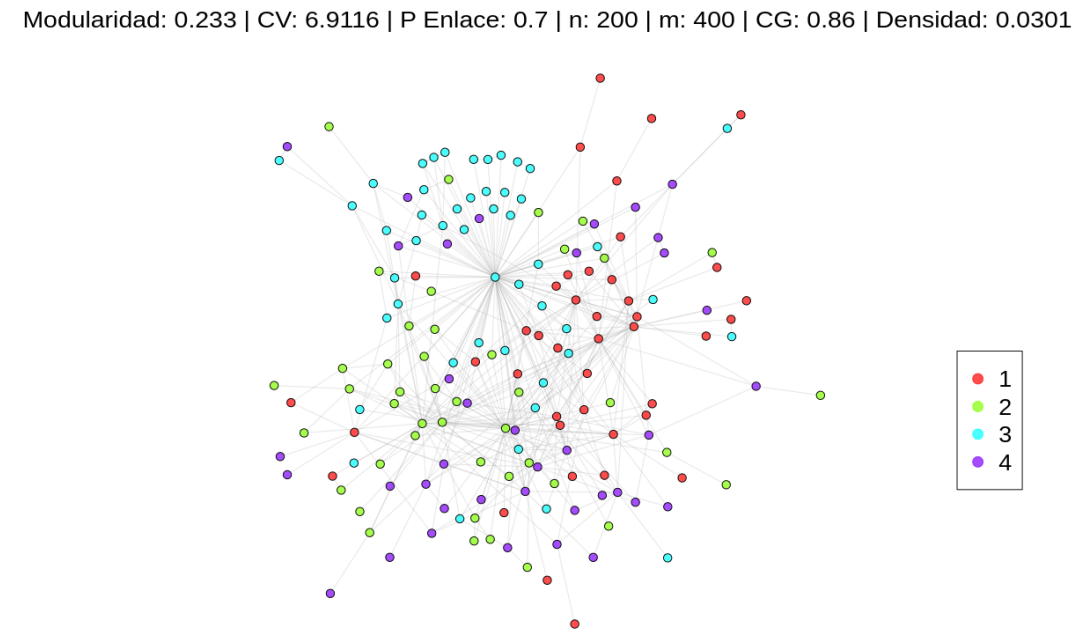


(a)

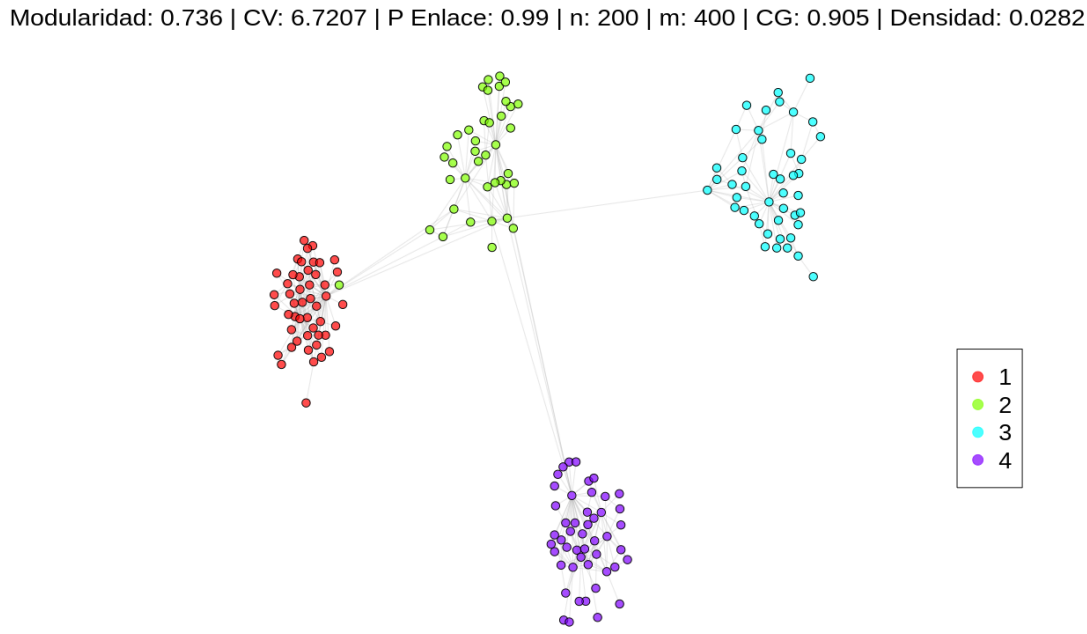


(b)

Figura 3.3: Redes sintéticas producidas con el modelo de red bipartita. Ambas tienen un CV similar, siendo CV= 4,2 para (a) y CV= 4,8 para (b) mientras que las modularidades si difieren significativamente siendo de 0,17 para (a) y 0,57 para (b).



(a)



(b)

Figura 3.4: Redes sintéticas producidas con el modelo de red bipartita. Ambas tienen un CV similar, siendo CV= 6,4 para (a) y CV= 6,5 para (b) mientras que las modularidades si difieren significativamente siendo de 0,23 para (a) y 0,74 para (b).

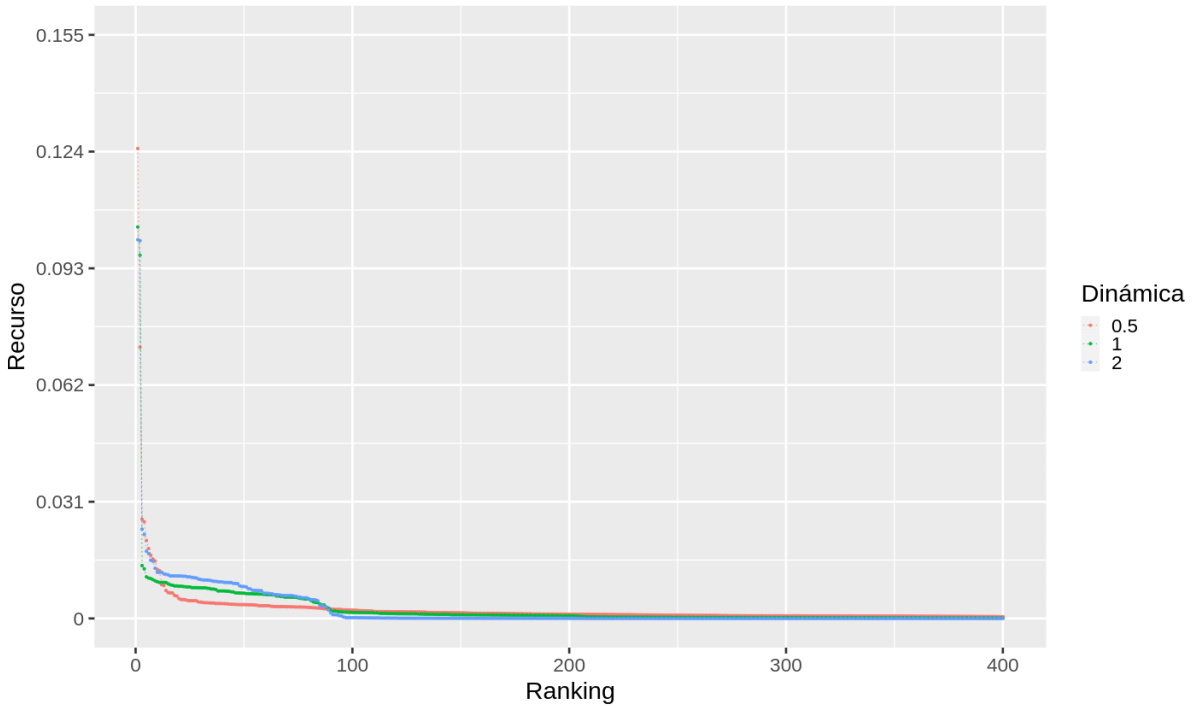


Figura 3.5: Ejemplo de una curva de difusión donde se grafica el recurso final asignado a cada nodo vs. el ranking de estos para los 3 valores de p utilizados.

Para definir los valores de estos parámetros se hizo una grilla para diferentes valores del tiempo total (pasos totales $\times$ h) de difusión variando ambos parámetros. La idea es poder ver a partir de ella para qué valores de h empiezan a diverger y separarse entre sí las curvas para lograr lo que se explicó previamente. La grilla se puede observar en la figura (3.6), en ella las filas representan los valores de p utilizados mientras que las columnas son el tiempo total T de la difusión. En cada gráfico de la grilla hay varias curvas de diferentes colores, cada uno indica el valor de h utilizado para esa curva que corresponde al número de pasos dado por T/h. A partir de la grilla se puede ver que para el caso de difusión lenta $p = 2$ no importa qué tamaño del paso h se utilice, las curvas siempre se mantienen juntas y se solapan para todo el ranking por lo que se puede utilizar un valor que sea conveniente y reduzca la cantidad de pasos totales a realizar. Para el caso de $p = 1$ sucede algo similar al caso de $p = 2$ aunque el solapamiento es menor pero todavía la separación entre curvas es muy pequeña, esto sin tener en cuenta el caso de $h = 1$ en el que se nota una separación mayor al resto. Finalmente para $p = 0.5$ vuelve a suceder lo anterior pero ahora las curvas que se separan considerablemente del resto son las que corresponden a los valores de $h = 1, 0.5$. En todos los casos los valores de mayor confianza son los mas pequeños y los que se toman de referencia como los mas cercanos a la solución.

Dados estos resultados se decide tomar como tamaño del paso el valor de $h = 0.01$ para los casos de difusión rápida y lineal ($p = 0.5, 1$) mientras que el valor de $h = 0.05$ para la

difusión lenta ($p = 2$). Para el número de pasos total, se decidió dejar libre y difundir hasta que la cantidad de recurso asignada a los nodos semillas llegue al 20 % del total.

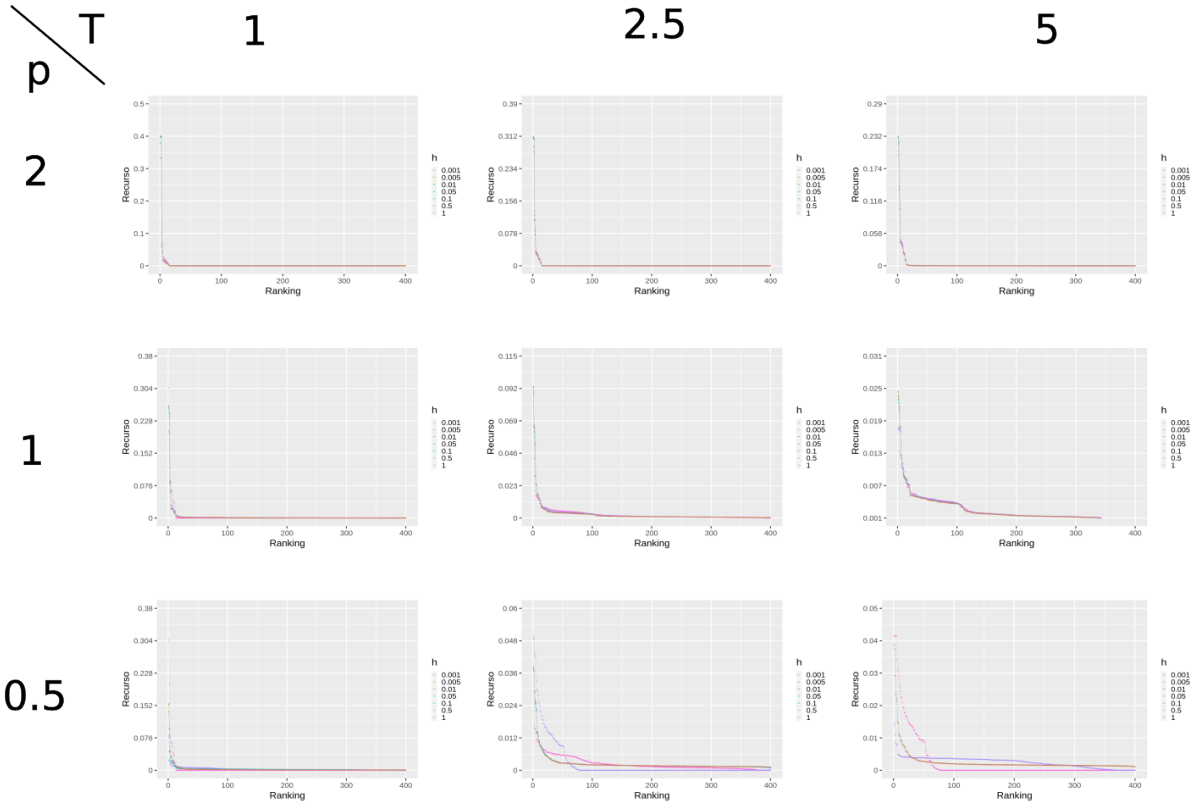


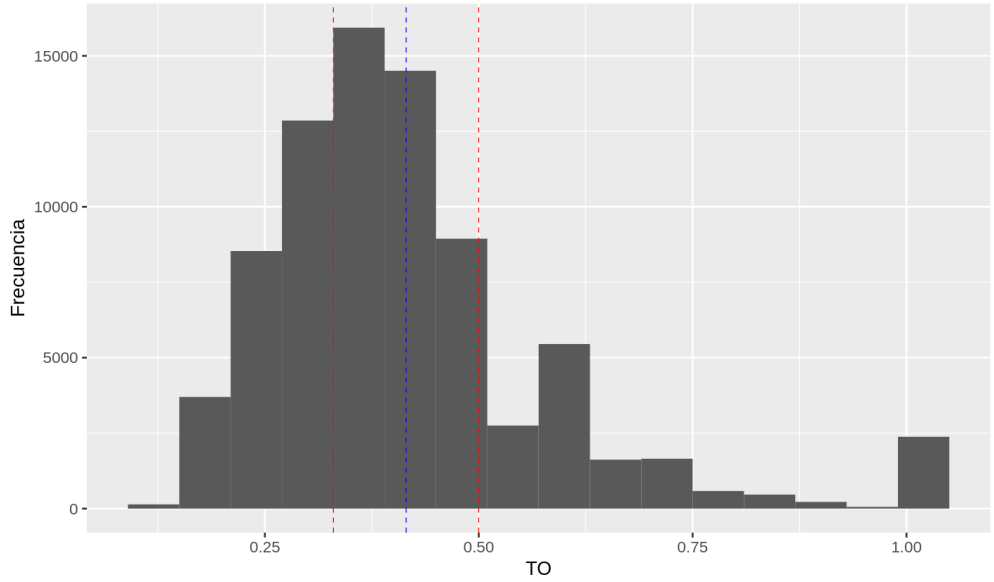
Figura 3.6: Grilla de gráficos con curvas de difusión para cada valor de p utilizado y para diferentes valores del tiempo total donde cada elemento contiene varias curvas con diferentes valores de h.

3.3. Semillas

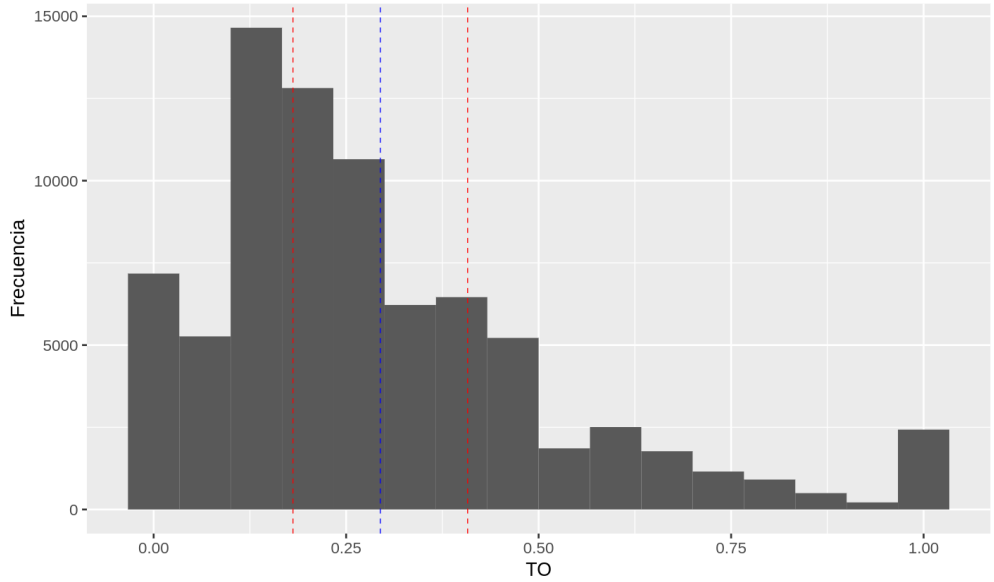
Para comenzar el proceso de difusión es necesario definir los nodos semillas, en nuestro análisis vamos a trabajar con pares de semillas. De esta manera consideraremos el caso más simple posible que al mismo tiempo nos permitirá estudiar la aparición de posibles efectos sinérgicos que separen las dinámicas lineales de las no lineales.

Será de interés considerar por separado casos en el que las semillas se encuentren en entornos cercanos unas de otra, respecto a casos donde ambas semillas se encuentren alejadas entre sí. Por eso se toman pares de nodos de la red según la distribución de TO que es una medida de similaridad topológica (solapamiento topológico visto en la introducción). Para ello se estudia cómo está distribuida esta cantidad para la red que utilicemos, en este caso para las redes elegidas anteriormente. Esto significa calcular el TO entre todos los pares de nodos de la red y armar un histograma para luego establecer umbrales que delimitan zonas de alta y baja conectividad. De nuevo, esto está pensado para intentar encontrar efectos sinérgicos entre las semillas en las no linealidades de los algoritmos. Los umbrales se toman como $t_{\pm} = \mu \pm \frac{\sigma}{2}$ donde μ es el valor medio de la distribución y σ su desviación estándar, de esta manera todos los pares de nodos que estén por debajo (encima) de t_- (t_+) se consideran pares de baja (alta) conectividad. Las distribuciones y los umbrales definidos se muestran en la figura (3.7).

Para hacer estadística se muestrean 1000 pares de nodos para cada valor de conectividad (rango de TO, bajo y alto) dando lugar a 2000 pares de nodos que se usarán como nodos semillas para la difusión. Además cada par de nodos es etiquetado según el grado de cada nodo en una tupla (x, y) donde $x, y \in \{$ alto, bajo $\}$ indicando si el grado de los nodos está por encima (alto) o por debajo (bajo) de la mediana de la distribución de grados de la red dando lugar a las combinaciones (bajo, bajo), (bajo, alto), (alto, alto) donde no importa el orden. En la figura (3.8) se muestran las distribuciones de grado de las redes seleccionadas y se indica la mediana. Esta división en categorías según los grados del par de nodos servirá más adelante para poder separar las distribuciones según ellas y ver cuál es el efecto del grado de los nodos semillas en la difusión.

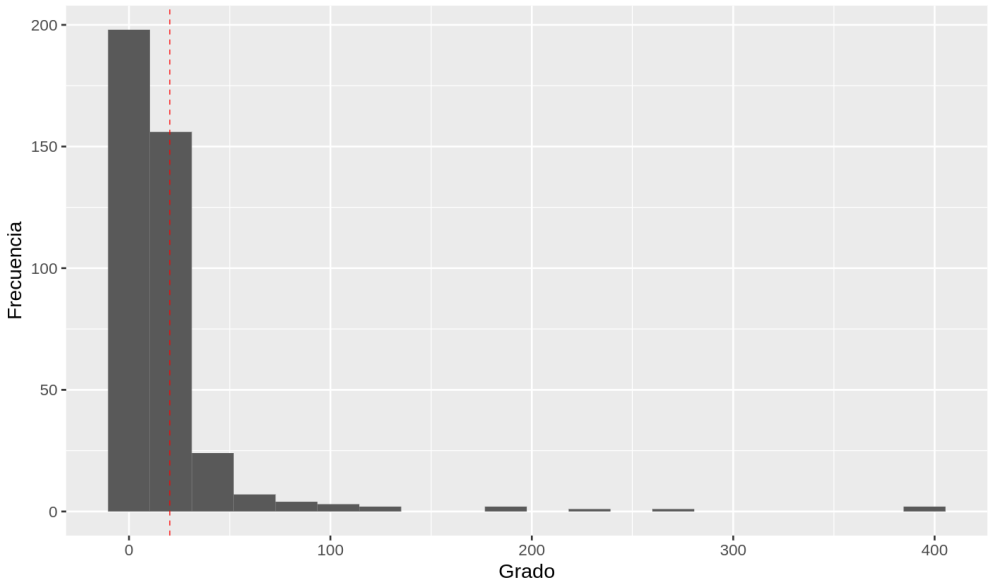


(a)

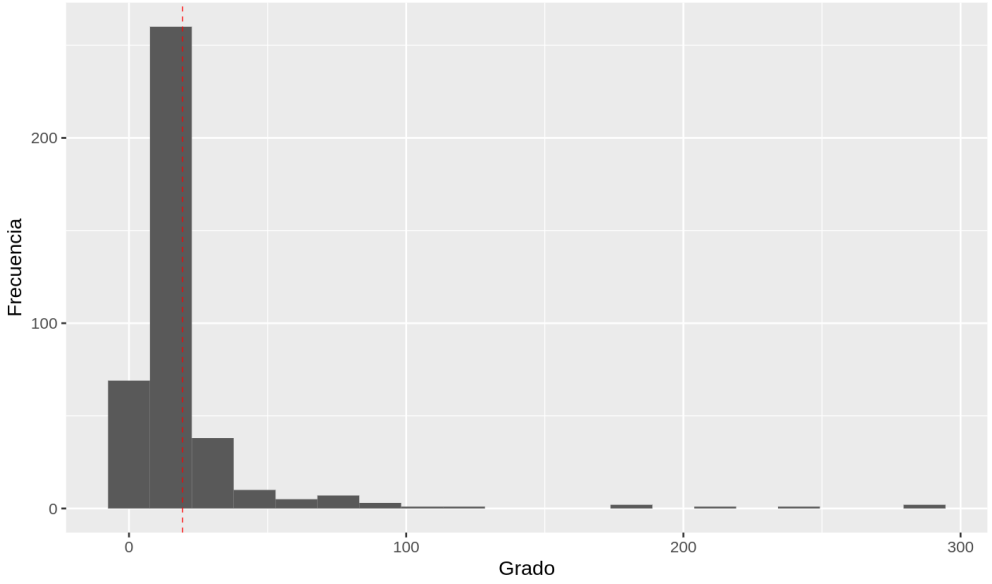


(b)

Figura 3.7: Distribuciones de TO para las redes de baja (a) y alta (b) modularidad. La línea vertical azul indica el valor medio mientras que las rojas indican un corrimiento de media desviación estándar desde este.



(a)



(b)

Figura 3.8: Distribuciones de grado de las redes de baja (a) y alta (b) modularidad. La línea vertical roja indica la mediana de la distribución.

3.4. Métricas

Para intentar comprender el comportamiento de los algoritmos de difusión se calculan algunas métricas que se creen que resumen las características del resultado de la difusión en una cantidad escalar.

La primera de estas se denomina acuerdo o agreement top 10 y se calcula entre dos rankings, es la proporción de elementos en común entre estos dos rankings dentro del top 10. Por ejemplo, si se tienen los rankings R_1 y R_2, se comparan los 10 primeros elementos (top 10 del ranking) y se cuentan cuántos elementos en común tienen (intersección) las dos listas, digamos $R_{1,2}$, para calcular la proporción de elementos en común $R_{1,2}/10$. En este caso particular se utiliza el agreement entre los rankings obtenidos a partir de los resultados de la difusión entre las dinámicas no lineales y las lineales. Es decir, se obtienen los rankings $R_{p=0,5}, R_{p=1}, R_{p=2}$ a partir de las difusiones inicializadas con el mismo par de nodos semillas y se calculan los acuerdos top 10 entre $R_{p=0,5}$ y $R_{p=1}$ por un lado y entre $R_{p=2}$ y $R_{p=1}$ por el otro.

Para definir las próximas cantidades es necesario antes introducir el concepto de disparity filter [7], que es un tipo de filtro introducido en el contexto de redes para identificar enlaces localmente relevantes. La idea es normalizar la suma de los pesos de los k_i enlaces del nodo i a la unidad para comparar ese conjunto de valores con el valor esperado al tomar el mismo número de pesos, también normalizados a la unidad, de una distribución aleatoria. Esta distribución nula permite definir un p-valor mínimo asociado a pesos observados en la red inusualmente grandes, que son los que consideraremos relevantes. Cabe resaltar que el calculo del umbral no depende de los valores del conjunto de números en si sino solamente del tamaño de este conjunto.

En nuestro caso utilizaremos esta metodología para identificar nodos con un score de difusión inusualmente grande. De esta forma se genera un umbral y todos los nodos a los que se haya asignado una cantidad de recurso en la difusión mayor al umbral se denominan como significativos.

Con esta definición del umbral, las dos métricas que restan definir son el número de nodos significativos y la cantidad de recurso que acumulan. Es decir, para un dado par de nodos semillas se difunde con cada una de las dinámicas elegidas (p = 0.5, 1, 2) a partir de el obteniendo un vector de recursos por cada una de las dinámicas a los que se le aplica el disparity filter. Así se obtiene un umbral con el que se calcula la cantidad de nodos con recurso por encima del umbral y también la suma de los recursos de esos nodos, esto para cada dinámica.

El umbral del disparity filter sirve en primera instancia para una distinción entre nodos

que resaltan frente al resto. Las cantidades mencionadas anteriormente calculadas a partir de este filtro pueden dar una intuición del comportamiento de cada dinámica. Para cada dinámica las semillas son las mismas y la cantidad de recurso a repartir también lo es, por lo que analizar que sucede con los nodos que reciben una cantidad de recurso por encima del umbral puede dar un indicio de la forma en que esa dinámica particular se propaga. Tener en cuenta la cantidad de este tipo de nodos y el recurso total que acumulan puede hablar sobre si una dinámica se comporta llegando a mayor cantidad de nodos pero dándoles menor cantidad de recurso o al contrario, resonando en una zona de la red con menor cantidad de nodos en los que concentra el recurso.

En la figura (3.9) se muestran como ejemplo las mismas curvas de difusión presentadas anteriormente para las 3 dinámicas y en el gráfico se indica con una linea horizontal roja el umbral resultante del disparity filter.

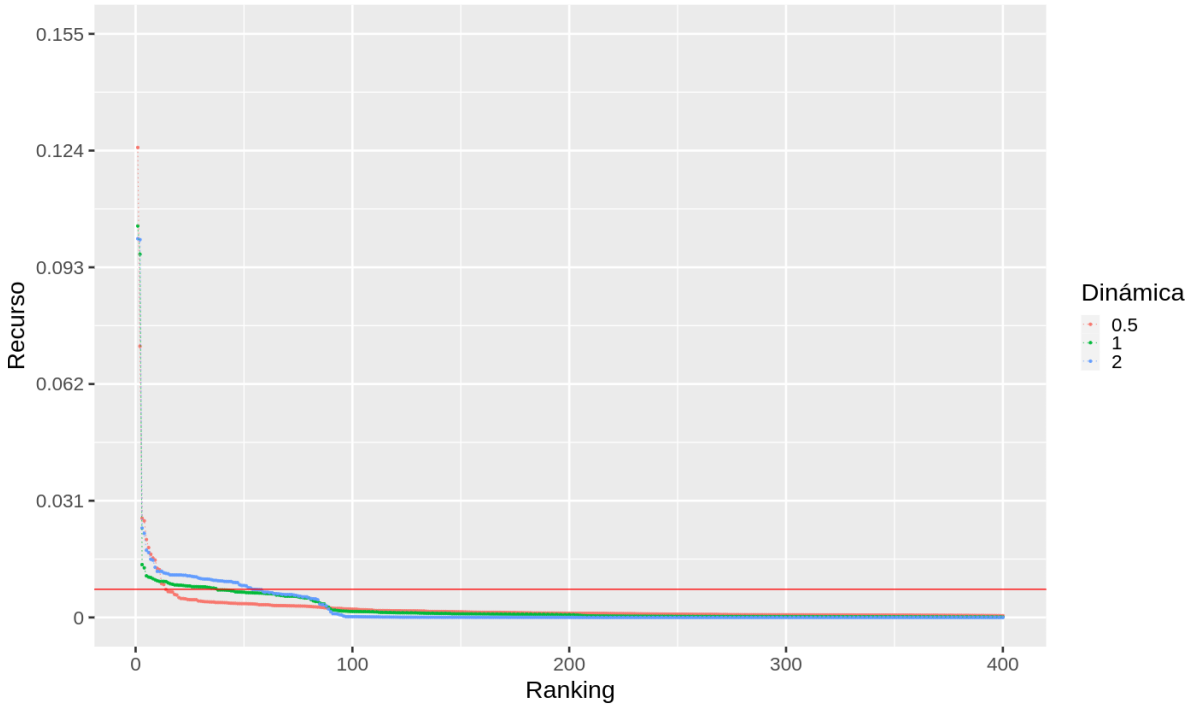


Figura 3.9: Curva de difusión de recurso vs. ranking para cada dinámica indicadas con color y linea roja horizontal indicando el umbral obtenido mediante el disparity filter.

De esta manera, por cada par de nodos tomados para usar como semillas de la difusión estimamos por un lado, el acuerdo top-10 de la difusión rápida vs lineal y la difusión lenta vs lineal. Por el otro, para cada dinámica, el número de nodos y el recurso acumulado para aquellos nodos que quedaron con recursos de recomendación por arriba del umbral.

3.5. Resultados

Antes de proceder con los resultados de los experimentos y las distribuciones obtenidas a partir de las varias realizaciones para hacer un estudio estadístico, es conveniente exponer un caso particular para comprender las diferentes dinámicas a un nivel cualitativo.

El ejemplo que se presenta aquí corresponde a una realización especifica de las que se hicieron para obtener las distribuciones de las métricas introducidas en la sección anterior y fue realizado sobre una de las redes elegidas en el capitulo anterior, la red de alta modularidad. El par de nodos semillas muestreado de la forma explicada en la sección 3.2 tiene un valor de TO en el rango que fue denominado como alto (es decir los nodos semillas podrían considerarse como cercanos) y además los grados de los nodos corresponden a la categoría (bajo-alto) según se explicó en la misma sección. En la figura (3.10) se muestra el gráfico con las curvas de difusión de recurso vs. rankings para las 3 dinámicas elegidas donde cada una se identifica por el color. Además también hay dos líneas horizontales, una roja que indica el umbral obtenido a partir del disparity filter introducido en la sección anterior, y la otra es una linea horizontal color azul que indica un valor casi nulo de recurso y se usa para identificar nodos por debajo de este como nodos de recurso nulo.

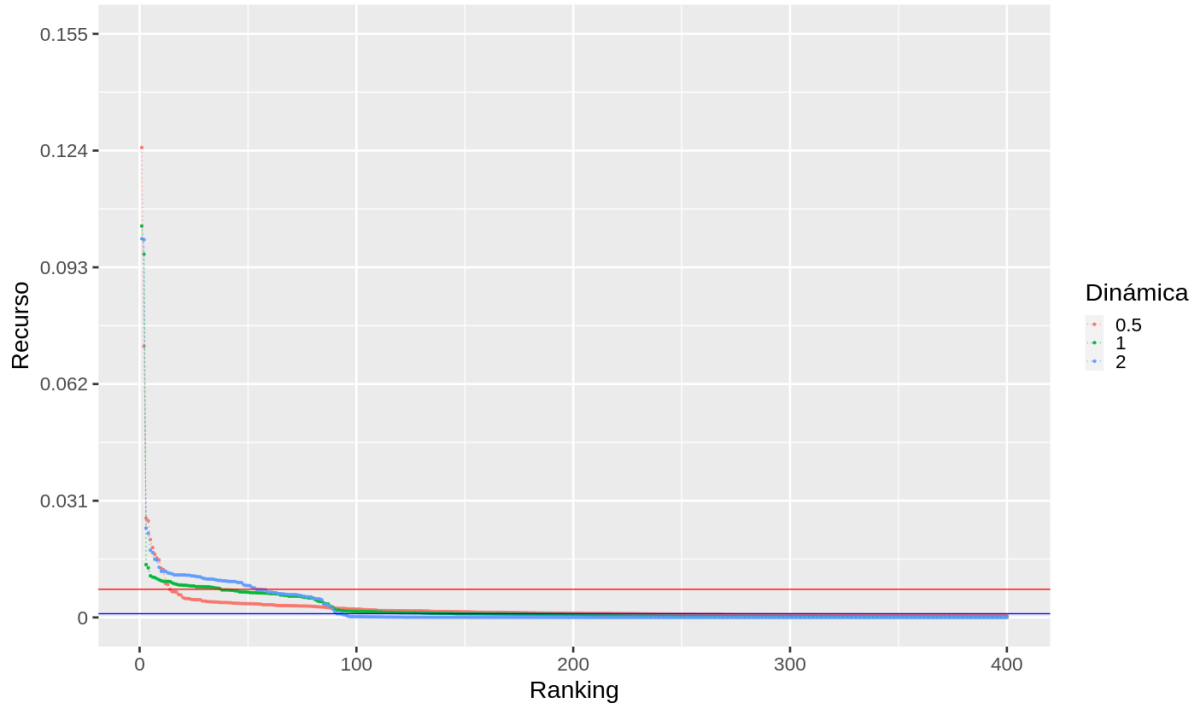


Figura 3.10: Curvas de difusión recurso vs. rankings para las 3 dinámicas. Se indican dos umbrales con lineas horizontales roja (disparity filter) y azul (recurso nulo).

Para poder visualizar y entender mejor cómo se ven los resultados de la difusión se hicie-

ron gráficos de la red utilizada pero coloreando los nodos con un mapa de colores según el recurso asignado al utilizar cada dinámica. Estos gráficos se encuentran en las figuras (3.11, 3.12, 3.13). Los dos nodos de mayor recurso de la red, los de color rojo, corresponden a las semillas desde las que se inicializa el proceso de difusión. Se puede ver que hay nodos de color naranja de diferentes intensidades que corresponde al recurso que hayan recibido, además estos nodos se encuentran por encima del umbral del disparity filter (linea roja horizontal en la figura 3.10). Hay otro rango de colores correspondiente a nodos que se encuentran entre el umbral de disparity filter y el umbral de recurso nulo (la linea horizontal azul de la figura 3.10), este rango es el que va decrecientemente en recurso asignado desde el verde al celeste. Finalmente todos los nodos de recurso menor al umbral de recurso nulo están identificados con el color blanco.

De esta forma es más sencillo comprender el comportamiento de las diferentes dinámicas y entender mejor el porqué de sus nombres.

Viendo la figura (3.11) para la difusión rápida se puede ver cómo el recurso es distribuido por toda la red aunque sea a niveles muy bajos indicando que este algoritmo explora la red en profundidad pero también resalta los valores de nodos cercanos a las semillas o nodos de alto grado y es además el que llega en menor cantidad de pasos a la condición establecida para finalizar la difusión.

Por otro lado la figura (3.12) correspondiente al caso lineal permite ver un campo de difusión mas concentrado alrededor de las semillas apenas explorando nodos por fuera de la comunidad y los nodos de alto grado que se encuentran en el centro de las comunidades, estos son nodos que siempre recibirán mucho recurso al ser hubs y conectar a una gran cantidad de nodos. Esta dinámica toma mayor cantidad de pasos en finalizar que la rápida pero menos que la lenta.

Por ultimo la figura (3.13) corresponde al caso lento y en ella se ve aun mayor concentración del recurso alrededor de los nodos semillas resaltando mas la estructura de conectividades dentro de la comunidad y sin explorar por fuera de ella salvo por los nodos de alto grado.

Algunas de estas características también pueden observarse en las curvas de la figura (3.10) donde se ve que el caso rápido decae mas rápidamente para los nodos de mayor recurso pero luego mantiene un nivel mayor de recurso para el resto de la red mientras que en las curvas de los casos lineal y lento se pueden ver ciertos cambios de escala, esto es, las curvas decrecen a un dado ritmo hasta llegar a un punto en el que rápidamente se hacen nulas. Este valor de corte corresponde en el caso lineal a nodos que se encuentran fuera de la comunidad mientras que en el caso lento corresponde a nodos fuera de una estructura intra comunidad que detecta esta dinámica.

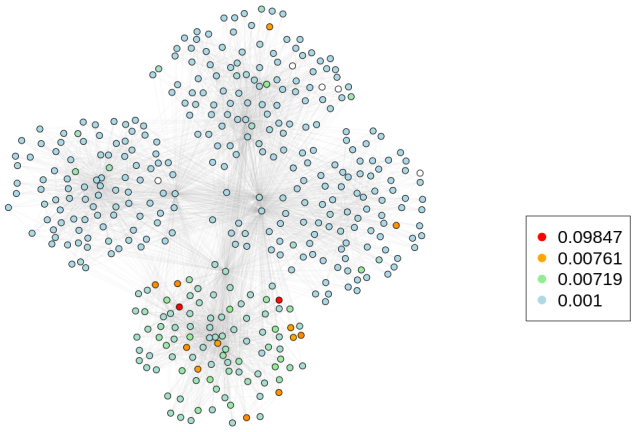


Figura 3.11: Gráfico de la red donde se indica con un mapa de color el recurso recibido por cada nodo para la dinámica rápida de $p = 0{,}5$.

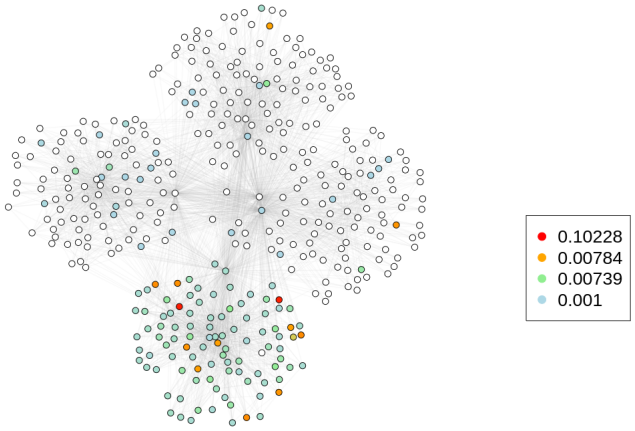


Figura 3.12: Gráfico de la red donde se indica con un mapa de color el recurso recibido por cada nodo para la dinámica lineal de $p = 1$.

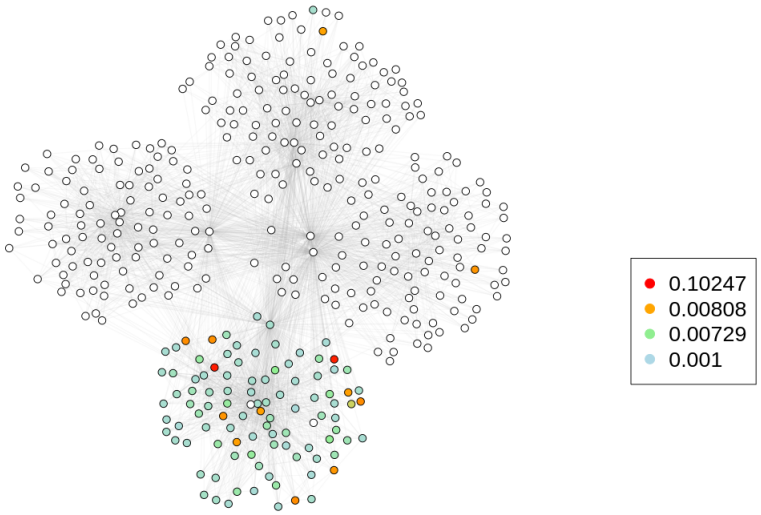


Figura 3.13: Gráfico de la red donde se indica con un mapa de color el recurso recibido por cada nodo para la dinámica lenta de $p = 2$.

Ahora se continúa presentando las distribuciones que se obtuvieron para las cantidades introducidas en la sección anterior. Como se explicó, estas distribuciones fueron calculadas realizando varias veces el proceso de muestrear un par de nodos semillas, difundir a partir de ellos con cada una de las dinámicas y finalmente evaluar las cantidades elegidas.

En la figura (3.14) se muestra la distribución del acuerdo top 10 en forma de gráficos de caja o boxplots. En rojo y verde se muestran las distribuciones asociadas a la cuantificación del acuerdo entre la dinámica rápida y la difusión lenta, respecto al caso lineal, respectivamente. En los paneles izquierdo y derecho de la figura se muestran las estimaciones calculadas para casos de baja y alta similaridad TO entre semillas respectivamente. Esto significa que los boxplots que se encuentran encima de la posición 'bajo' en el eje horizontal corresponden a todos los valores obtenidos a partir de pares de nodos semilla en el rango bajo de TO y de forma análoga para el rango alto.

De este gráfico se puede observar que en cualquiera de los rangos de TO la dinámica lenta tiene mayor coincidencia con la dinámica lineal que la rápida y esa diferencia se amplifica más cuando el par de semillas está en el rango bajo de TO. Lo que esto significa es que la dinámica lenta reconoce como importantes a los mismos o casi los mismos nodos que la lineal dándoles mayor relevancia mientras que la rápida explora nuevos nodos respecto a la lineal.

En la figura (3.15) se presentan las distribuciones en formato de boxplots para las canti-

dades que se introdujeron previamente como número de nodos significativos y el recurso total acumulado por estos nodos, ambas cantidades definidas a partir del disparity filter. En las distribuciones se indican con colores las dinámicas utilizadas y se separan según a qué rango de TO pertenece el par de nodos semillas como se explicó antes. En estas distribuciones se puede apreciar como ambas cantidades aumentan al aumentar p, es decir, con la reducción de la 'velocidad' siendo la tendencia más notable para la distribución del recurso acumulado. Esto significa que en el ranking generado por la dinámica lenta una mayor cantidad de nodos quedan por encima del umbral de disparity filter comparado con la dinámica rápida.

Esto se puede entender a partir de los ejemplos vistos al inicio de la sección donde se ve que la difusión lenta distribuye su recurso entre nodos cercanos a las semillas sin ir muy lejos obteniendo así mayor cantidad de nodos por encima del umbral. Por otro lado, la difusión rápida distribuye el recurso entre mayor cantidad de nodos porque tiene un alcance mayor en la red y al haber más nodos que reciben recurso solo algunos obtienen lo suficiente para superar el umbral. La mayor cantidad de nodos significativos por parte de la dinámica lenta también implica una mayor cantidad de recurso acumulado en estos nodos respecto de la dinámica rápida.

Por último, la separación por rangos de TO parece no influir notablemente en el comportamiento de estas cantidades ya que las distribuciones resultan muy similares.

En la figura (3.16) se presentan gráficos similares a los de la figura (3.15) pero además de separar las distribuciones en el eje horizontal por el rango de TO también se separan según las tres categorías a las que pueden pertenecer los grados del par de nodos semillas muestreados según se comentó en la sección (3.3). Esto se hace de la misma forma en la que se separa por rangos de TO pero con las categorías de grados.

Al hacer la separación según los grados de los nodos se puede ver un comportamiento que parece indicar alguna interacción entre los nodos semillas y que no es igual para todas las dinámicas. En la figura (3.16a) donde se encuentran las distribuciones del número de significativos se puede ver una dependencia con la combinación de grados del par de semillas tomando el orden de categorías de grado como bajo-bajo, bajo-alto y alto-alto.

Usando el orden mencionado (bajo-bajo, bajo-alto y alto-alto) como el orden del eje horizontal, se observa un comportamiento creciente para las dinámicas lineal y lenta. Por otro lado la dinámica rápida muestra una tendencia decreciente para el rango bajo de TO y levemente creciente en el rango alto de TO. En los dos rangos de TO las tres dinámicas tienen distribuciones similares para el caso de combinación de grados bajo-bajo.

Las distribuciones permiten observar que no solo el tipo de dinámica, sino también los grados de los nodos semilla juegan un papel importante en el proceso difusivo.

Siguiendo con la figura (3.16b), contiene las distribuciones de recurso acumulado por los nodos significativos y en ellas se marcan las mismas tendencias mencionadas para las distribuciones de número de nodos significativos aunque ahora con una mayor separación de

las distribuciones según las dinámicas. Además también se ve que no sucede para ninguna combinación de grados que las distribuciones se agrupen concentradas alrededor de un valor como lo hacen en el caso de grados bajo-bajo para la distribución del número de significativos. Esto produce por ejemplo que para casos donde la combinación de grado es bajo-bajo, las curvas de recurso vs. ranking decrezcan con mayor pendiente en el caso de la dinámica lenta ya que tiene una cantidad de significativos similar a las de las demás dinámicas pero mayor recurso acumulado en los nodos significativos.

De estos resultados se aprecia como el grado de los nodos juega un papel importante tanto en la velocidad como en el alcance de la difusión que a su vez interactúa con el tipo de dinámica formando así comportamientos complejos.

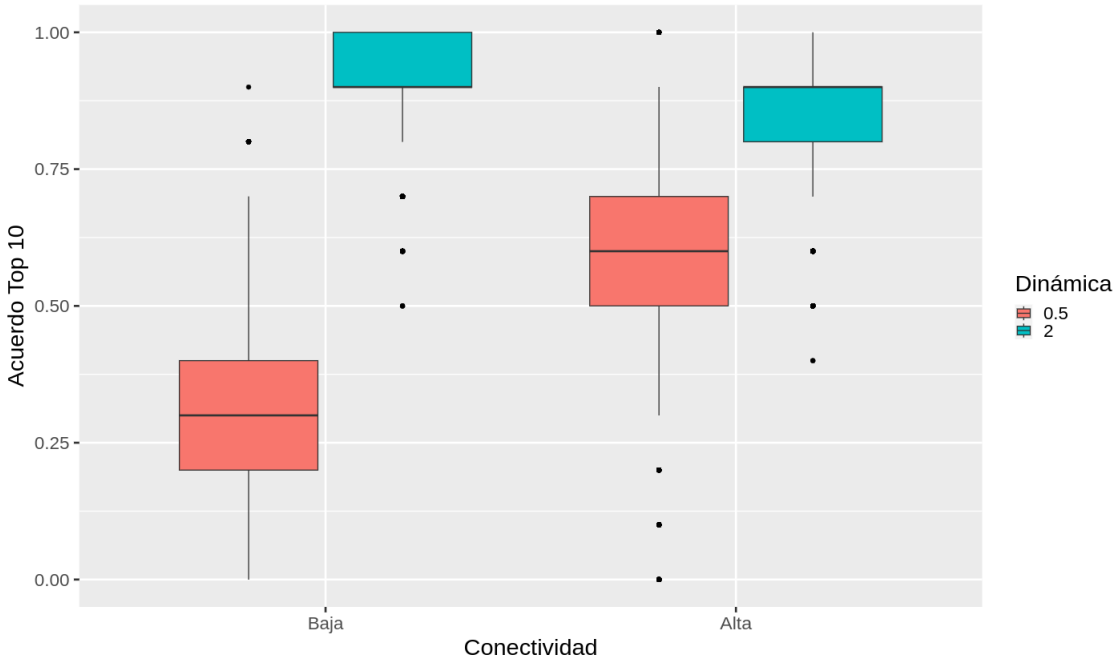
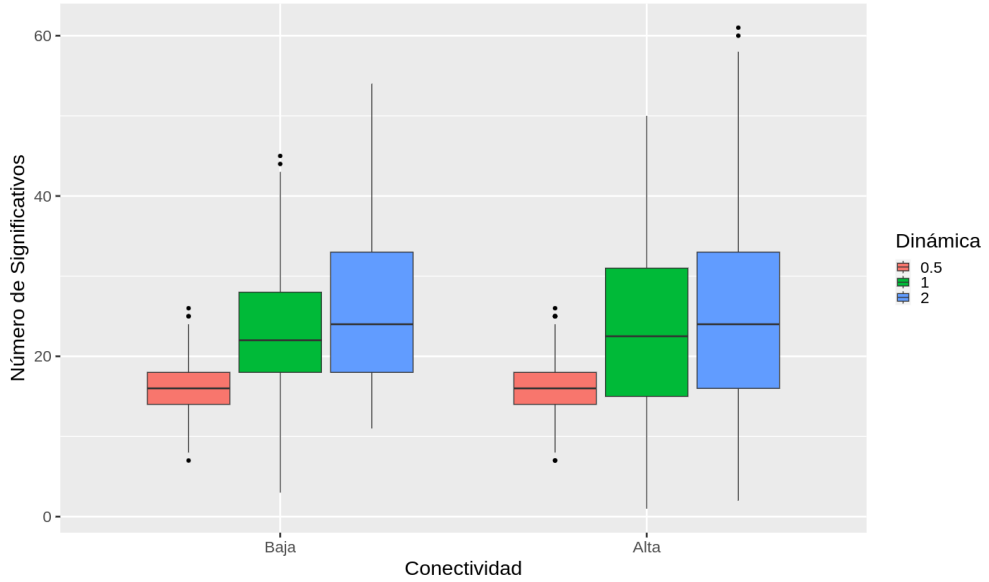
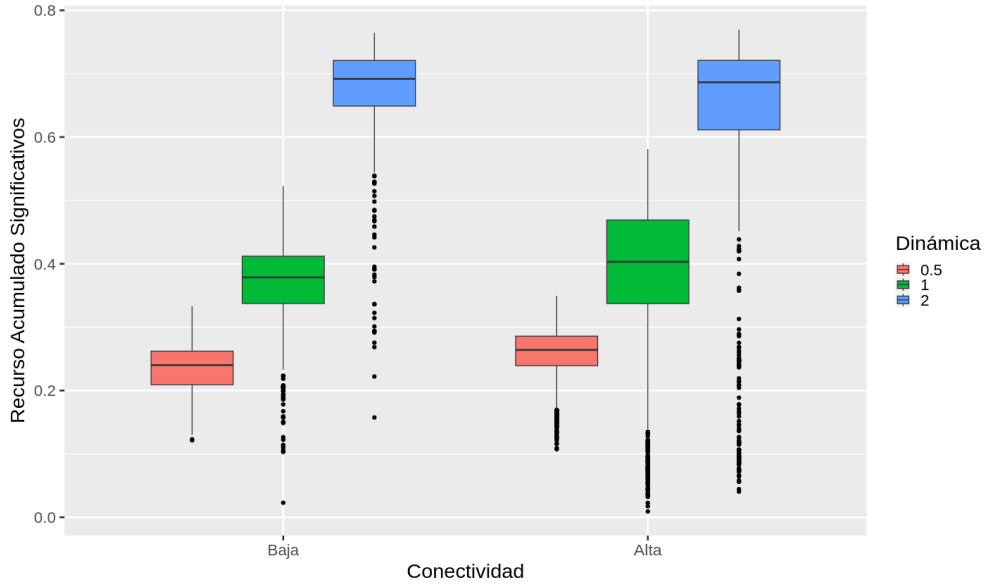


Figura 3.14: Distribuciones de acuerdo top 10 entre rankings para cada dinámica separadas por el rango de TO utilizado para tomar nodos semillas.

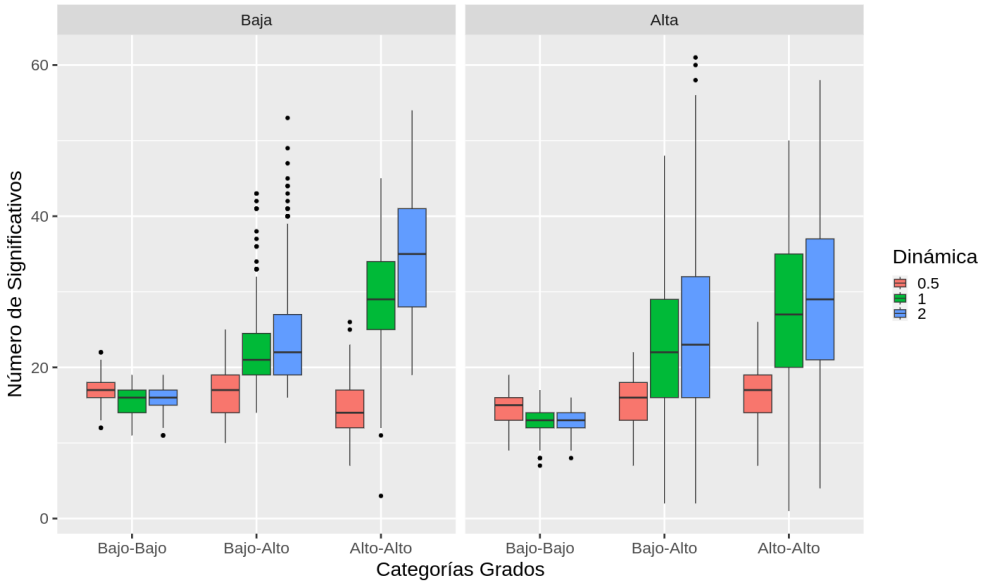


(a)

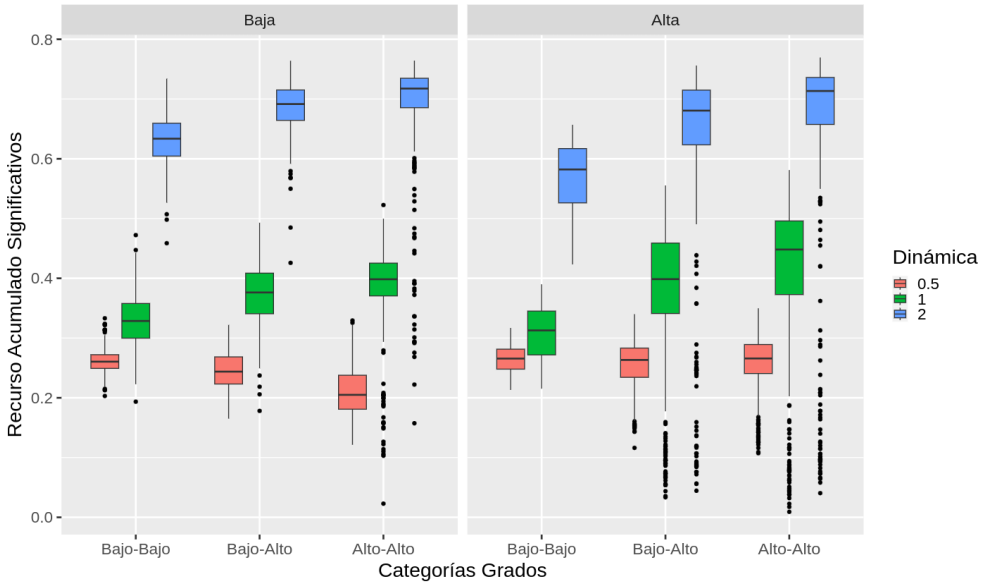


(b)

Figura 3.15: Distribuciones del número de nodos significativos (a) y del recurso acumulado por estos (b) utilizando el disparity filter. Están coloreadas por tipo de dinámica y separadas por rangos de TO para muestrear nodos semillas.



(a)



(b)

Figura 3.16: Distribuciones del número de nodos significativos (a) y del recurso acumulado por estos (b) utilizando el disparity filter. Están coloreadas por tipo de dinámica y separadas por rangos de TO para muestrear nodos semillas por un lado y en cada uno de esos rangos se separa también por las categorías de los grados del par de nodos muestreado.

Capítulo 4

Priorización en datos reales

Para poner a prueba los algoritmos de difusión presentados y estudiados en los capítulos anteriores se usan bases de datos reales que contienen información tanto de las proteínas y sus interacciones como de enfermedades y sus asociaciones a proteínas.

4.1. Redes biológicas

El organismo humano es un sistema complejo conformado por elementos que trabajan a diferentes escalas dando lugar a funcionalidades biológicas como resultado de incontables interacciones entre las diferentes partes. Por ejemplo, al nivel celular se llevan a cabo de forma continua reacciones químicas que sirven a varios fines pero que son reguladas por proteínas entre varias otras de sus funciones. Para realizar las tareas necesarias las proteínas interactúan entre si de diferentes formas para formar parte de estructuras mayores o participando en reacciones químicas. Todas estas interacciones entre proteínas pueden codificarse en un grafo formando así una red de interacciones de proteínas que describe de forma aproximada todo el complejo entramado de relaciones existentes entre ellas.

Estas redes se construyen a partir de varias fuentes de información recopiladas de forma independiente y comprenden relaciones que pueden ser de diferentes tipos. En particular las redes de tipo PPI (protein-protein interaction) describen interacciones físicas entre proteínas mientras que hay otros tipos de redes que por ejemplo describen la coexpresión génica. Otra característica de las redes PPI es que surgen de experimentos *ex-vivo*, lo que significa que se estudia la posibilidad de una interacción física entre proteínas en un contexto controlado diferente al contexto biológico en el que se encuentran al estar en el organismo cumpliendo sus funciones.

Este tipo de redes son de interés ya que para llevar a cabo funciones las proteínas deben interactuar entre sí además del hecho de que al estudiar las causas de una enfermedad de base genética, se ha observado que los genes asociados a esta suelen estar próximos en una red PPI lo que es de gran relevancia para los algoritmos utilizados [6] [13] [9]. Sin embargo

aunque los genes estén cerca entre sí no se encuentran en módulos bien definidos y esto afecta al rendimiento de los algoritmos que utilizan la cercanía de los genes en la red.

En este trabajo se utiliza la base de datos provista por el consorcio HIPPIE [20], esta consta de interacciones entre proteínas recopiladas de una combinación de reportes realizados mediante técnicas experimentales, analíticas y algorítmicas. Para integrar los datos el consorcio creó un índice para indicar la certeza de cada tipo de técnica utilizada y luego a cada interacción entre un par de proteínas se asigna un valor numérico basado en ese índice que funciona como un puntaje para representar el grado de confianza que se tiene en la existencia de la interacción.

De esta manera se obtiene una lista de interacciones con un número real asociado a ella con lo que es posible formar una red pesada, siendo el grado de confianza el peso de los enlaces entre proteínas. Esta red se puede transformar en una red no pesada binarizando los pesos de los enlaces, es decir, estableciendo un umbral para los pesos para eliminar aquellos enlaces cuyo peso sea menor a este umbral y mantener los enlaces cuyo peso es mayor pero sin conservar el peso de estos sino que solo utilizarlos como indicativos de la existencia de una interacción.

En este trabajo se utiliza la red de interacciones de proteínas no pesada obtenida a través de la binarización de los enlaces quedándonos con enlaces de alta confianza (score > 0.73 [19]).

Hasta aquí se explicó acerca de las interacciones entre proteínas y la naturaleza de su codificación en una red. Aún resta mencionar la fuente de donde se obtiene la información sobre las asociaciones entre genes y enfermedades conocidas que se utilizarán para el estudio y la evaluación de los algoritmos. Esta fuente se llama DisGeNet [1] y es una plataforma completa de descubrimiento diseñada para estudiar diversas cuestiones sobre la base genética de las enfermedades humanas. La plataforma integra bases de datos elaboradas por especialistas con datos minados de textos, abarca información sobre enfermedades mendelianas y complejas e incluye datos de modelos de enfermedades en animales. Estas fuentes las combina ofreciendo una puntuación basada en la evidencia a cada asociación entre gen y enfermedad.

En la figura (4.1) se muestra un esquema que representa las dos redes utilizadas y la relación entre ellas. En conjunto parece como que las redes formaran una red bipartita pero con la salvedad de que la red de proteínas tiene enlaces entre nodos de la misma red. En realidad, la red de enfermedades se utiliza simplemente para hallar las asociaciones conocidas y luego se pasa a utilizar solo la red de interacciones de proteínas haciendo uso de la información sobre las asociaciones.

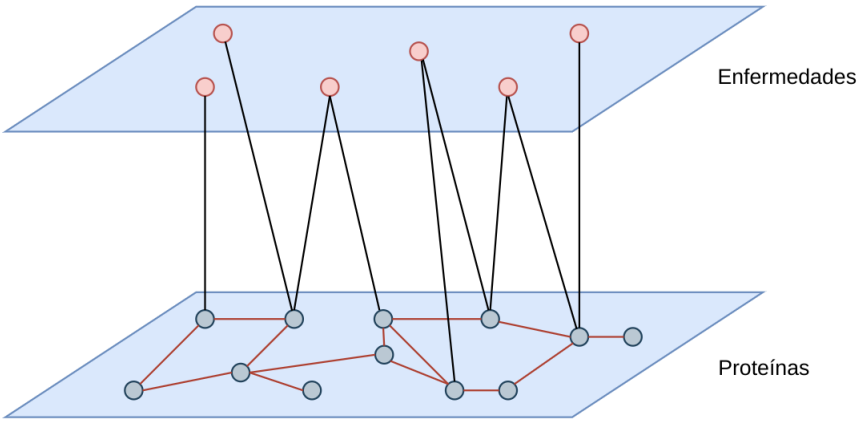


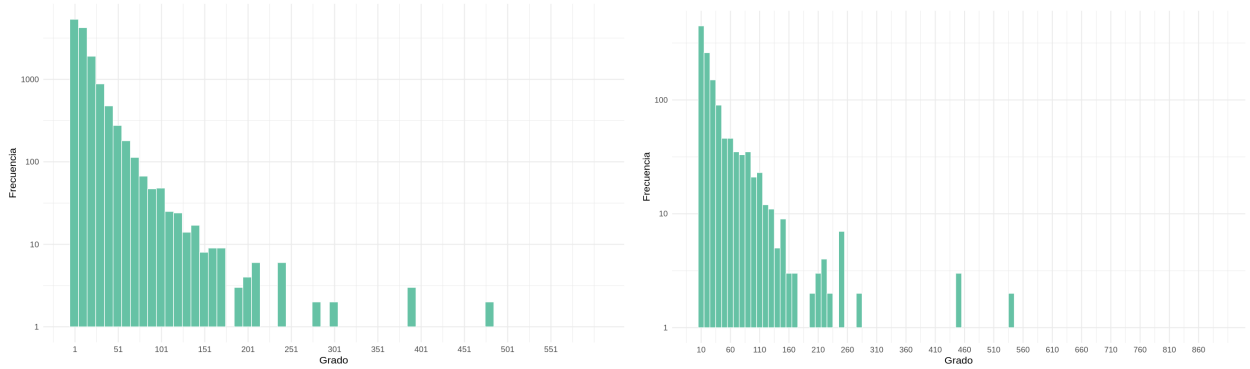
Figura 4.1: Diagrama de las redes reales y las relaciones entre ellas. La red de proteínas tiene enlaces entre ellas mientras que los enlaces de las enfermedades solo se conectan a proteínas.

4.2. Tratamiento de los datos y evaluación

En las versiones de las redes utilizadas se cuentan con 15.760 proteínas, 11.063 enfermedades, 110.320 enlaces proteína-proteína y 83.805 enlaces enfermedad-proteína. Sin embargo, la red de proteínas no es conexa y se utiliza la componente gigante, que cuenta con 13.712 nodos y 105.281 enlaces. Por otro lado, en la red de enfermedades se mantienen únicamente aquellas que tengan asociaciones en la componente gigante de la red de proteínas y que además tengan un mínimo de 10 asociaciones para poder hacer particiones en los conjuntos de entrenamiento, validación y evaluación. Así la red de enfermedades ahora cuenta con 1264 enfermedades y 54.977 asociaciones enfermedad-proteína. De esta forma las distribuciones de grado resultantes para ambas redes se presentan en la figura (4.2) en escala logarítmica

ya que son distribuciones de cola pesada (la mayoría tiene un grado bajo pero existen nodos de grado alto) y esta escala hace mas fácil apreciar las escalas. La distribución que se encuentra en (4.2a) corresponde a la red de proteínas mientras que (4.2b) es la de la red de enfermedades y por eso el grado mínimo es 10.

En esta parte del trabajo se continúan utilizando prácticas y elecciones de parámetros adoptadas previamente. Las dinámicas que se van a utilizar son las mismas tres que se presentaron antes y estas son rápida, lineal y lenta que se corresponden a los valores del parámetro $p = 0{,}5, 1, 2$ en la ecuación de difusión no lineal presentada en el capitulo de métodos. Además, la forma de llevar a cabo la difusión es realizando tantos pasos como sean necesarios para que la cantidad de recurso aun retenida por los nodos semillas sea del 20 % del total. Finalmente, el valor del paso temporal h se toma como $h = 0{,}01$ para las dinámicas rápida y lineal mientras que $h = 0{,}05$ para la dinámica lenta de igual forma que en la sección anterior.



(a) Proteínas (b) Enfermedades

Figura 4.2: Distribuciones de grado en escala logarítmica. (a) corresponde a la red de proteínas y (b) a la red de enfermedades.

Con las redes acondicionadas de esta manera se procedió a realizar las separaciones en conjuntos de entrenamiento, validación y evaluación como se explicó en la sección de métodos. La forma especifica en la que se hizo es tomando cada enfermedad individualmente y particionando el conjunto de genes asociados a esta en los conjuntos mencionados, es decir, primero se separa de forma aleatoria en dos conjuntos, el de entrenamiento con el 80 % de los genes y el de evaluación con el 20 % restante. Luego, el conjunto de entrenamiento se vuelve a separar en dos dando un subconjunto de entrenamiento con el 70 % del conjunto de entrenamiento original y otro conjunto llamado de validación con el 30 % restante. Esta separación de entrenamiento-validación se realiza 20 veces de forma aleatoria para así poder evaluar los algoritmos y tomar valores medios de las métricas.

La métrica utilizada para evaluar los algoritmos es el AUC de la curva ROC. El proceso especifico se lleva a cabo de la siguiente manera, para una enfermedad y una de las 20 particiones de entrenamiento-validación se hace la difusión utilizando como semillas a los genes del conjunto de entrenamiento lo que resulta en un campo de recurso sobre los nodos de la red. Este permite armar el ranking de relevancia con los genes de la red y junto al conjunto de validación se utiliza para armar la curva ROC a partir de la cual se calcula el AUC como se explicó en la introducción. De esta forma se obtienen 20 valores de AUC (uno por cada una de las particiones de entrenamiento-validación) que se promedian para obtener el AUC medio y esta es una cantidad escalar directamente relacionada con una enfermedad. Este proceso se realiza para cada una de las tres dinámicas por lo que se obtienen tres valores de AUC medio para cada enfermedad, estando cada uno asociado a una dinámica.

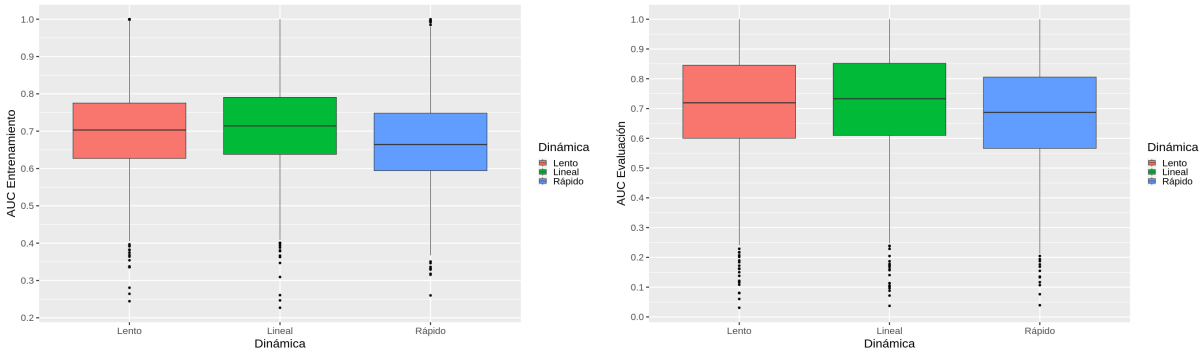
Finalmente, para comparar los métodos de evaluación y control también se evalúan los algoritmos utilizando solamente los conjuntos de entrenamiento y evaluación. En este caso

no se hacen subconjuntos para validación ni se calculan valores medios sobre ellos, sino que simplemente se hace la difusión partiendo del conjunto de entrenamiento obtenido originalmente y luego se evalúan las métricas con el conjunto de evaluación.

4.3. Resultados y discusión

En esta sección se presentan los resultados obtenidos de los experimentos y los métodos aplicados.

Para ver si las diferentes dinámicas presentan la misma capacidad de recuperar asociaciones conocidas analizamos la distribución de AUC. Para ello consideramos estas distribuciones en las dos etapas mencionadas en la sección anterior (entrenamiento y evaluación) para comparar los resultados. En la figura (4.3) se muestran las distribuciones de AUC en diferentes colores según la dinámica considerada. El gráfico de (4.3a) corresponde a la etapa de entrenamiento y las distribuciones son de los valores medios del AUC (promediado sobre las particiones de la etapa de entrenamiento-validación como se explico en la sección anterior) mientras que en (4.3b) se encuentran las distribuciones de AUC obtenidas en la etapa de evaluación. De los gráficos se puede ver que la mayor parte de las distribuciones se encuentran en general entre los valores de 0.6 y 0.8, que son valores un poco bajos pero aceptables del AUC. Además se ve que las distribuciones para las tres dinámicas no difieren considerablemente.



(a) (b)

Figura 4.3: Distribuciones del AUC medio separadas por dinámica. El gráfico de (a) corresponde a la etapa de entrenamiento mientras que el (b) a la de evaluación.

Al estar utilizando procesos difusivos es de esperar que nodos más cercanos a los nodos semillas reciban más recurso. Si esto sucede, enfermedades cuyos genes asociados se encuentren más localizados en la red tendrían un mejor rendimiento al intentar recuperarlos con los algoritmos utilizados.

Para profundizar este análisis nos preguntamos qué tipo de relación se podía establecer entre el resultado de la priorización por difusión, bajo las diferentes dinámicas, y la distancia topológica entre los genes que se desea recuperar y aquellos que se utilizan como punto de partida.

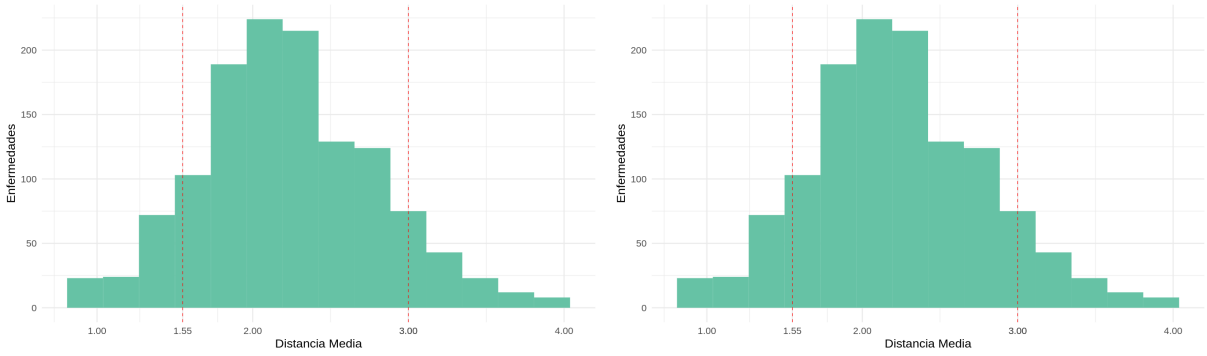
La forma en la que intentamos cuantificar esto y comprender en que casos el AUC resulta en valores más altos es estudiando su dependencia con una cantidad derivada de la distancia entre los genes de entrenamiento y los genes del conjunto utilizado para evaluar al algoritmo (pudiendo este ser el de validación o el de evaluación). Para explicar el concepto a continuación se supone que se utiliza el par entrenamiento-validación.

La cantidad mencionada se refiere desde aquí como distancia media y se calcula de la siguiente manera. Para una enfermedad se tiene el par de conjuntos de entrenamiento-validación. A partir de cada nodo del conjunto de validación se calcula la distancia en la red de este nodo a todos los nodos de entrenamiento y luego se toman las dos menores para promediarlas. De esta manera disponemos de una medida más o menos robusta de la cercanía sobre la red del nodo de validación al conjunto de entrenamiento utilizado. Por último, se promedian todas las distancias medias obtenidas para los nodos de validación. De esta forma se obtiene el valor medio de la distancia media entre los dos genes de entrenamiento más cercanos a cada gen de validación.

En la etapa de entrenamiento este cálculo se hace para cada una de las 20 particiones de entrenamiento-validación y luego se promedian para obtener una sola cantidad escalar por enfermedad con la idea de que esta sea una representación de la distribución en la red de los genes asociados a esa enfermedad. Por otro lado, en la etapa de evaluación esta cantidad se calcula una sola vez por enfermedad sin tomar valores medios.

En la figura (4.4) se muestran las distribuciones de la distancia media, en (4.4a) para la etapa de entrenamiento y en (4.4b) para la etapa de evaluación. Queremos separar entre enfermedades cuya distribución de genes se encuentra más concentrada o más dispersa en la red para estudiar si la distancia media juega un papel importante en el rendimiento de los algoritmos. Para ello, separamos las distribuciones de distancia media en tres intervalos indicados con líneas verticales rojas en los gráficos. Estos se toman de manera que los intervalos de los extremos contengan el 10 % de las enfermedades cada uno. Esta separación da lugar a tres regímenes de distancias siendo estas bajas, medias y altas.

Previo a separar los resultados según los intervalos de distancia se muestra en la figura (4.5) el gráfico de puntos del AUC en función de la distancia media, los colores de los puntos indican las diferentes dinámicas. Cada gráfico corresponde a una etapa de evaluación distin-

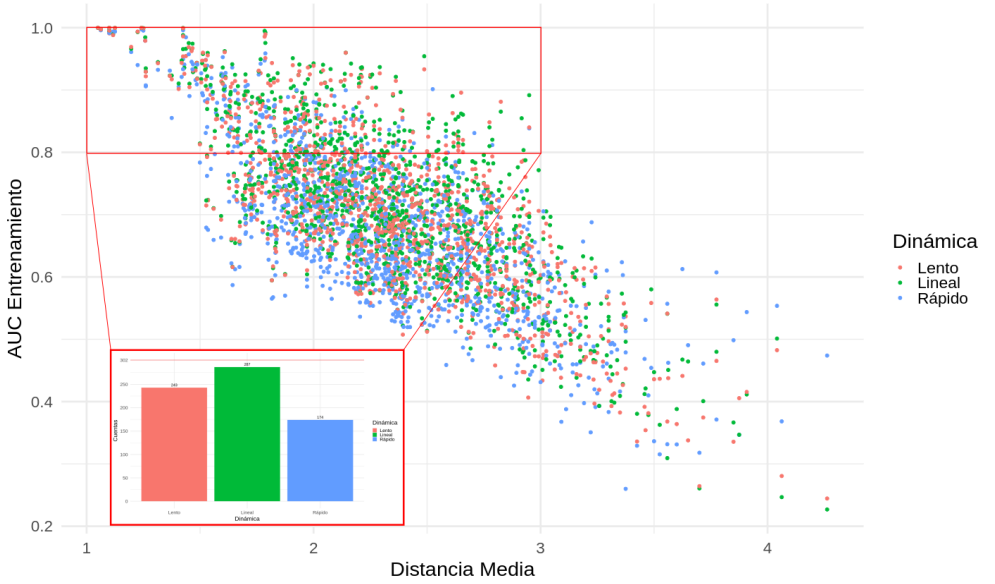


(a) (b)

Figura 4.4: Distribuciones de distancia media para cada etapa. Estas se encuentran divididas en intervalos por lineas verticales rojas para dar lugar a tres regímenes de distancia: baja, media y alta.

ta como ya se explicó. Como anticipamos, la capacidad de los algoritmos de recuperar a los genes asociados a una enfermedad depende de la distribución de estos en la red y que tan cerca se encuentren de los nodos semillas. Esto se ve de la tendencia en ambos gráficos del AUC a decrecer al aumentar la distancia media. En las figuras además se hace un zoom sobre una parte de los gráficos. En el se toma las enfermedades que mejor rendimiento tuvieron (AUC > 0.8) y con ellas se hace un gráfico de barras mostrando qué cantidad de estas hay y como se distribuyen según cada dinámica considerada. Se puede observar que las dinámicas no se comportan de la misma manera ya que son las lenta y lineal las que tienen mayor cantidad de puntos en la zona de mayor AUC cuando las distancias son bajas-medias. Estos son comportamientos que estudiamos en el capítulo anterior en donde vimos que la dinámica lenta entrega más recurso a nodos cercanos a las semillas, mientras que la dinámica rápida hace lo contrario explorando la red en mayor profundidad.

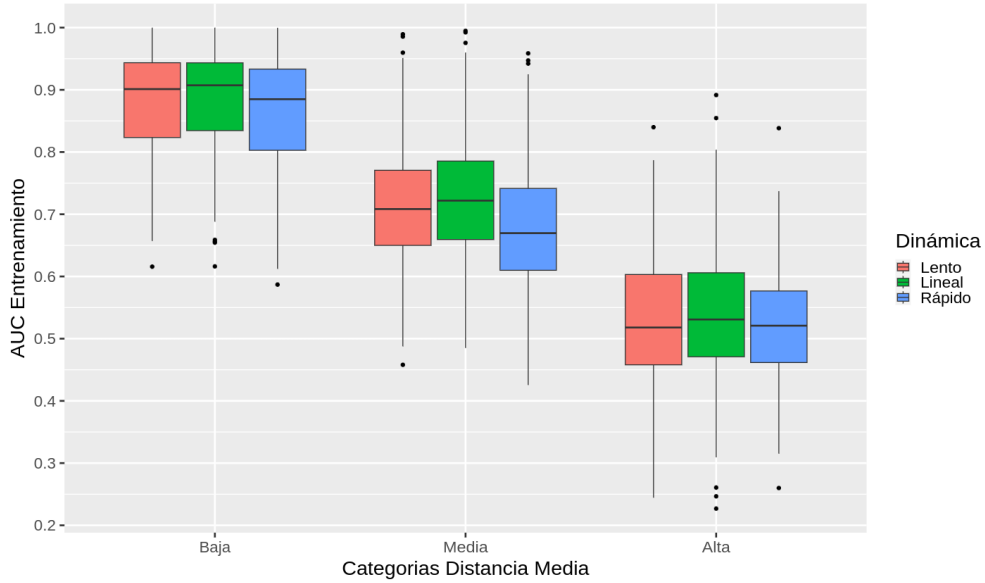
En la figura (4.6) se encuentran las distribuciones de AUC vistas anteriormente pero en estas se separan por los grupos de distancias medias mencionados anteriormente. Nuevamente se ve la tendencia decreciente que se observa en la figura (4.5) ya que estas distribuciones provienen de binear los puntos de ese gráfico según los intervalos de distancia. Como en la figura (4.3) sucede que dentro de cada intervalo de distancia media, las distribuciones no se diferencian significativamente entre las distintas dinámicas.



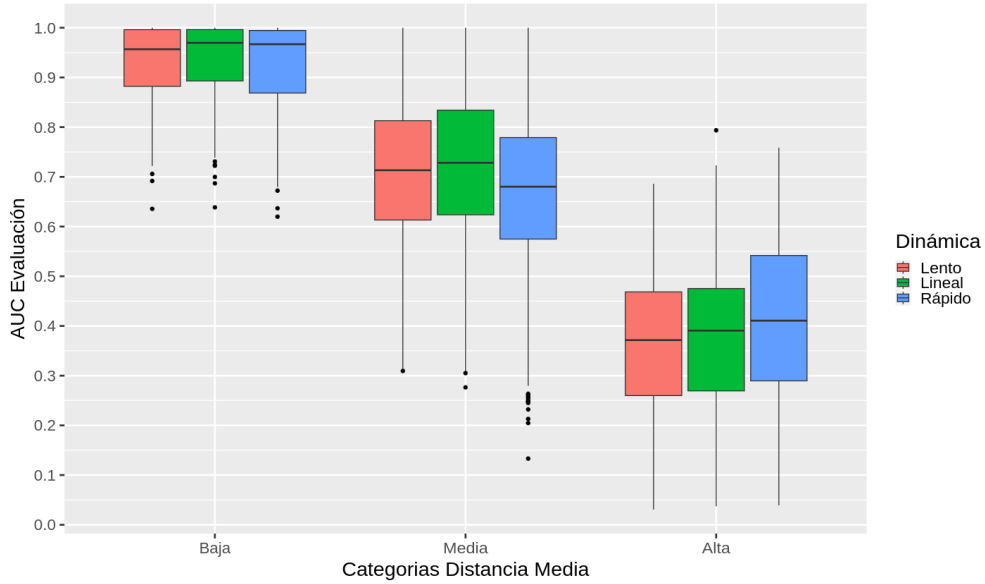
(a)

(b)

Figura 4.5: AUC en función de la distancia media donde los puntos se colorean por dinámica. El gráfico de (a) corresponde a la etapa de entrenamiento mientras que el (b) a la de evaluación. Se hace zoom sobre la zona de mayor rendimiento (AUC $>$ 0.8) para hacer un gráfico de barras con la cantidad de enfermedades en esta zona y como se distribuyen para cada dinámica.



(a)



(b)

Figura 4.6: Distribución del AUC medio separado por diferentes rangos de la distancia media entrenamiento-validación donde los colores indican la dinámica. El gráfico de (a) corresponde a la metodología entrenamiento-validación mientras que el (b) a entrenamiento-evaluación.

Ambas etapas en nuestro proceso de análisis involucran apartar un conjunto de nodos semilla para evaluar la capacidad del algoritmo de recuperarlos. Quisimos investigar cuánto recurso de priorización alcanzaba efectivamente a los nodos de este conjunto al finalizar el proceso de difusión bajo las diferentes dinámicas. Para ello consideramos las distribuciones del recurso medio acumulado por los nodos pertenecientes a este conjunto.

En la etapa de validación donde se utiliza el método de kfolds se calcula el recurso medio que reciben los nodos del conjunto de validación al finalizar el proceso de difusión y luego se toma el valor medio sobre las 20 particiones como se explicó anteriormente. En la etapa de evaluación simplemente se calcula para los nodos que pertenecen al conjunto de evaluación. En la figura (4.7) se encuentran las distribuciones del recurso medio acumulado. Están separadas por las categorías de distancia media y los colores indican la dinámica utilizada.

En estos gráficos se ve nuevamente la dependencia con la distancia media, recibiendo mayor recurso los nodos que se encuentran más cercanos a los nodos semillas. Además para cada rango de distancias se puede observar una tendencia para las dinámicas donde la cantidad de recurso es mayor para la dinámica lenta y menor para la rápida. Este es un comportamiento similar al que se obtuvo en los experimentos realizados en las redes sintéticas del capitulo anterior donde se expresa el carácter de cada dinámica. Se vuelve a ver que la dinámica lenta se concentra mayormente cerca de los nodos semillas en lugar de explorar la red en profundidad que es lo que parece suceder en la dinámica rápida.

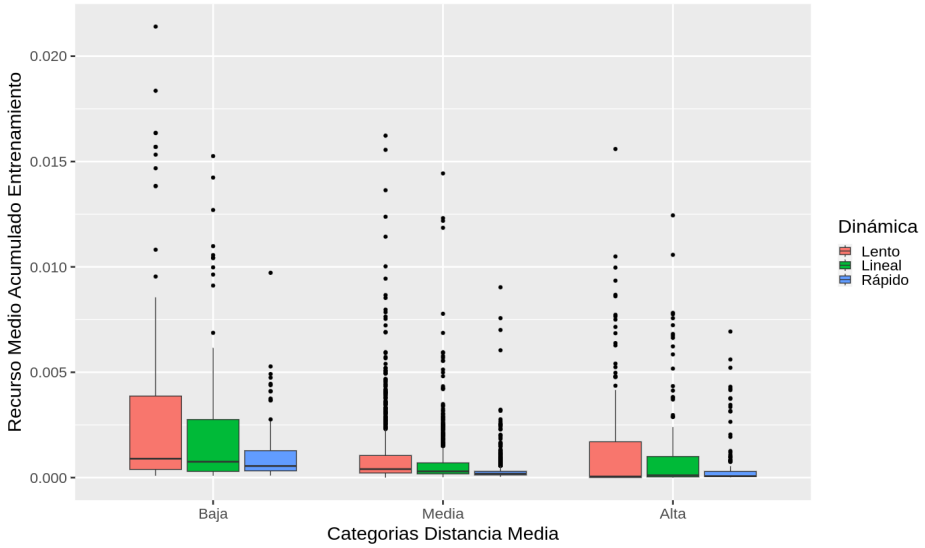
Asimismo quisimos investigar qué tan predictivo es el desempeño de los algoritmos observado durante la etapa de entrenamiento-validación con respecto a los resultados obtenidos para la etapa de evaluación. Esto es, verificar si se cumple que la distribución de genes sobre la red para una enfermedad se mantiene para las particiones en conjuntos de entrenamiento y evaluación realizadas.

Para ello se comparan los valores de AUC obtenidos por las dos metodologías de evaluación utilizadas. En la figura (4.8a) se muestra el gráfico de AUC-evaluación vs AUC-entrenamiento. Por otro lado, en (4.8b) se muestra un gráfico de boxplots para las distribuciones de AUC-evaluación pero donde se separan por intervalos de los valores de AUC-entrenamiento y también se separan por dinámica. En ambos se puede ver una tendencia creciente o de correlación positiva, sugiriendo que las enfermedades que presentan buenos resultados en la etapa de entrenamiento también lo hacen en la de evaluación y que los conjuntos de entrenamiento y evaluación son representativos de la distribución de los nodos en la red.

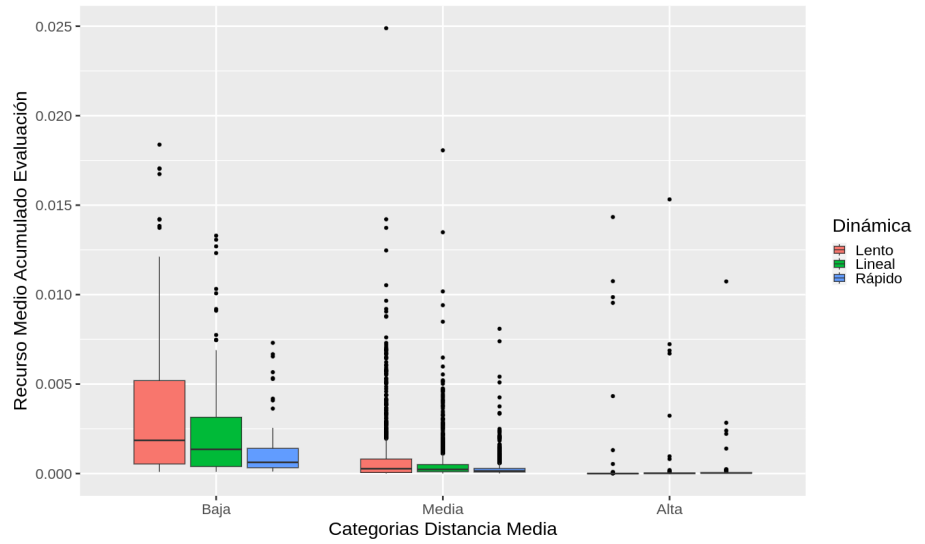
Además también se presenta el gráfico de la figura (4.9) que es el mismo de la figura (4.8a) en la que se tiene el gráfico de puntos pero se agregan curvas de suavizado que se utilizan para ayudar en la visualización de las tendencias.

Para complementar estos gráficos se quiere estudiar si existe una relación entre los re-

sultados que se obtuvieron en la etapa de entrenamiento y los que se obtuvieron en la de evaluación. Para ello se realiza un test estadístico de correlación entre el AUC-evaluación y el AUC-entrenamiento. Este consiste en calcular el coeficiente de correlación de Pearson (R) y aplicar un test estadístico para evaluar si el valor obtenido es distinto de cero de forma significativa. Específicamente la distribución que sigue el estadístico R es la distribución t de student con n-2 grados de libertad donde n es el número de datos. El valor obtenido para R es de 0.39 con un p-valor mucho menor a 0.05 (2.2e-16), lo que indica que, a pesar de que es un valor bajo, resulta significativamente diferente a cero. Esto sugiere que los resultados obtenidos en la etapa de entrenamiento son en cierta medida transferibles a la etapa de evaluación. Además, por ejemplo las enfermedades que tuvieron buenos resultados en la primer etapa podrían tenerlos en la segunda.

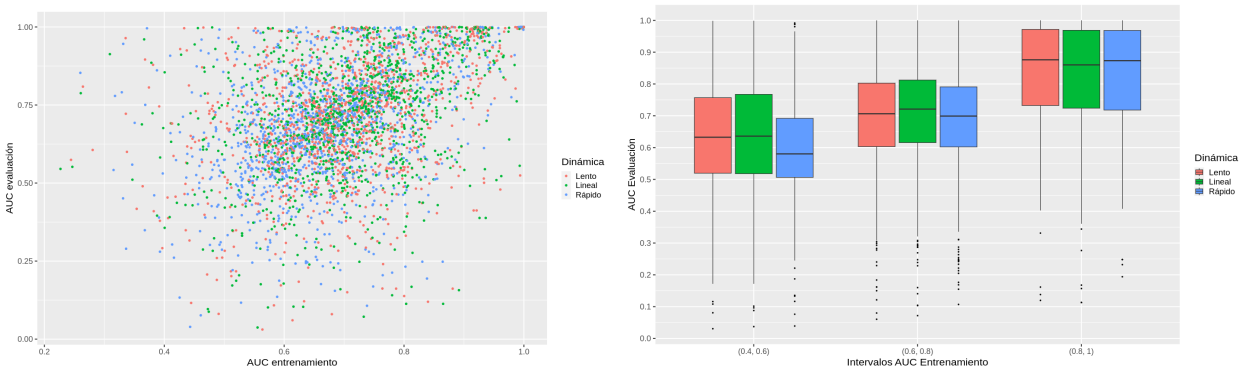


(a)



(b)

Figura 4.7: Distribución del recurso medio acumulado por los nodos de validación separado por diferentes rangos de la distancia media entrenamiento-validación donde los colores indican la dinámica. El gráfico de (a) corresponde a la metodología entrenamiento-validación mientras que el (b) a entrenamiento-evaluación.



(a) Gráfico del AUC obtenido en la etapa de evaluación vs el obtenido en la etapa de entrenamiento. El color de los puntos representa el tipo de dinámica.

(b) Gráfico en formato de boxplots para el AUC en la etapa de evaluación donde se separa por intervalos del AUC en la etapa de entrenamiento. El color de cada distribución corresponde al tipo de dinámica utilizada.

Figura 4.8

Figura 4.9: Gráfico del AUC obtenido en la etapa de evaluación vs el obtenido en la etapa de entrenamiento. Además sobre los puntos se encuentran las curvas de suavizado que ayudan a visualizar la tendencia de los puntos. El color representa el tipo de dinámica.

En los resultados obtenidos para la métrica de AUC es posible observar que las tres dinámicas utilizadas no presentan diferencias significativas que sugieran el uso de una de ellas por sobre las otras.

A pesar de que los diferentes tipos de difusión tienen comportamientos distintos como ya se ha estudiado, ninguno parece ofrecer ventajas a la hora de recuperar genes con asociaciones. Es decir, al eliminar algunos genes del conjunto de genes asociados a una dada enfermedad e intentar recuperarlos al utilizar cualquiera de los tres algoritmos, ninguno ofrece un resultado significativamente mejor al resto.

Sin embargo, los resultados muestran una característica fundamental de los procesos difusivos que está relacionado a la forma en la que estos exploran la red. Un proceso difusivo comienza localizado en un nodo de la red y se propaga hacia sus alrededores de forma progresiva en el tiempo a través de los nodos más cercanos a aquél desde el que inicia el proceso.

El hecho de que los resultados dependan de las distancias entre los nodos semillas y los nodos que se quieren recuperar es un comportamiento razonable debido a la naturaleza del fenómeno de difusión.

De todas maneras, estos métodos pueden funcionar bien dependiendo de la enfermedad que se está queriendo estudiar. Esto se debe a que parecen existir ciertas enfermedades cuyo conjunto de genes asociados tiene una correlación con la topología de la red ya que su distribución sobre esta no tiene gran dispersión resultando en conjuntos de genes que se encuentran concentrados en una zona de la red. Esto resulta en valores de AUC aceptables al ser aprovechados por los algoritmos la cercanía de los genes que inician la difusión y aquellos que se quieren recuperar por los algoritmos.

Por último, por más de que las diferentes dinámicas utilizadas no presentan diferencias significativas en los resultado estas podrían ser utilizadas en conjunto aprovechando las diferencias en sus comportamientos para intentar potenciarlas entre sí y mejorar la recuperación de genes con asociaciones y así la métrica de AUC. Por ejemplo, en el capítulo anterior se estudió como las recomendaciones de cada uno de los algoritmos difieren en los genes que mejor posicionados aparecen en los rankings generados. Una posibilidad sería intentar utilizar los diferentes rankings en conjunto e integrarlos para obtener uno que ofrezca una mayor capacidad para recuperar y recomendar posibles asociaciones.

Capítulo 5

Conclusión

En esta tesis se estudió el problema de priorización en redes gen - enfermedad utilizando la teoría de redes complejas. El método elegido a través del cual se hizo la priorización es el de difusión en redes. En cuanto a las redes utilizadas, estas provienen de diferentes bases de datos para integrar la información de las asociaciones entre genes y enfermedades por un lado y las interacciones de proteínas por el otro.

El método utilizado para la priorización fue el de difusión en redes y el objetivo fue estudiar y comprender el comportamiento de un método de difusión no lineal y la comparación de este con el lineal.

Para ello primero se estudiaron estos métodos en redes sintéticas armadas con ciertas características de su estructura de forma de que mantengan cierta similaridad con las redes del mundo real de interés. Para el armado de estas redes se diseñó un método basado en redes bipartitas que permite controlar cantidades estructurales de las redes a través de parámetros del modelo.

En las redes sintéticas se hicieron experimentos utilizando los modelos de difusión de interés y se estudiaron algunas características fundamentales de estos utilizando cantidades y métricas propuestas por nosotros ya que creemos que capturan los comportamientos estudiados. Allí se vio que por ejemplo la difusión llamada lenta se comporta de tal manera de quedarse concentrada y resonando cerca de los nodos semilla, mientras que la rápida se expande explorando gran parte de la red.

Finalmente se estudiaron los métodos en las redes reales. Estas redes fueron obtenidas de bases de datos e integradas para obtener una red de interacción de proteínas con información externa correspondiente con las asociaciones entre genes y enfermedades. Estas asociaciones se utilizaron para armar conjuntos de nodos semillas de los cuales parte la difusión y es la

información que se utiliza para evaluar los modelos. En cuanto a las métricas utilizadas, la principal fue el área bajo la curva ROC o por sus iniciales en inglés AUC. Es una medida utilizada en problemas de clasificación y mide qué tanto el modelo es capaz de recuperar información que le fue quitada durante la etapa de entrenamiento.

Los resultados obtenidos nos dicen que en general el rendimiento de los modelos no lineales y lineales no presenta diferencias significativas ya que los valores de AUC resultan similares. Sin embargo, cada modelo tiene características propias que pueden ser de interés y utilidad al combinarlas o en diferentes situaciones. En donde se encontraron diferencias es al tener en cuenta la distribución sobre la red de los nodos semillas y los nodos objetivos que el modelo debía recuperar. Al ser un modelo de difusión, aquellas enfermedades cuyos genes se encuentran en módulos más compactos de la red obtuvieron mejores valores de AUC mientras que las enfermedades cuyos genes están dispersos por toda la red de proteínas tuvieron un rendimiento bajo. Por ello es que estos modelos pueden ser de particular interés en algunas enfermedades donde las distribuciones de genes son más concentradas.

Bibliografía

[1] Janet Piñero et al. "DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes". En: (2015). DOI: 10.1093/database/bav028.

[2] Timothy J. Aitman Anne M. Glazier Joseph H. Nadeau. "Finding Genes That Underlie Complex Traits". En: (2002). DOI: 10.1126/science.1076641.

[3] Neil Risch David Botstein. "Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease". En: (2003). DOI: 10.1038/ng1090.

[4] Matthieu Latapy Jean-Loup Guillaume. "Bipartite graphs as models of complex networks". En: (2006). DOI: 10.1016/j.physa.2006.04.047.

[5] Matthieu Latapy Jean-Loup Guillaume. "Bipartite structure of all complex networks". En: (2004). DOI: 10.1016/j.ipl.2004.03.007.

[6] M A Huynen M Oti B Snel y H G Brunner. "Predicting disease genes using protein–protein interactions". En: (2006). DOI: 10.1136/jmg.2006.041376.

[7] Marián Boguñáb M. Ángeles Serranoa y Alessandro Vespignanic. "Extracting the multiscale backbone of complex weighted networks". En: (2009). DOI: 10.1073/pnas.0808904106.

[8] Tim Kacprowski Nadezhda T. Doncheva y Mario Albrecht. "Recent approaches to the prioritization of candidate disease genes". En: (2012). DOI: 10.1002/wsbm.1177.

[9] Saket Navlakha y Carl Kingsford. "The power of protein interaction networks for associating genes with diseases". En: (2010). DOI: 10.1093/bioinformatics/btq076.

[10] M.E.J Newman. *Networks an introduction*. Oxford university press, 2010.

[11] Alexander Zien Olivier Chapelle Bernhard Schölkopf. *Semi-supervised learning*. Cambridge, Massachusetts: MIT-press, 2006.

[12] David F. Gleich Rania Ibrahim. "Nonlinear Diffusion for Community Detection and Semi-Supervised Learning". En: (2019). DOI: 10.1145/3308558.3313483.

[13] Denise Horn Sebastian Köhler Sebastian Bauer y Peter N. Robinson. "Walking the Interactome for Prioritization of Candidate Disease Genes". En: (2008). DOI: 10.1016/j.ajhg.2008.02.013.

[14] Insuk Lee U. Martin Blom Peggy I. Wang Jung Eun Shim y Edward M. Marcotte. "Prioritizing candidate disease genes by network-based boosting of genome-wide association data". En: (2011). DOI: 10.1101/gr.118992.110.

[15] Rob M Ewing Sinan Erten Gurkan Bebek y Mehmet Koyutürk. "DADA: Degree-Aware Algorithms for Network-Based Disease Gene Prioritization". En: (2011). DOI: 10.1186/1756-0381-4-19.

[16] Matúš Medo Tao Zhou Jie Ren y Yi-Cheng Zhang. "Bipartite network projection and personal recommendation". En: (2007). DOI: 10.1103/PhysRevE.76.046115.

[17] Tao Zhoua Zoltán Kuscsika Jian-Guo Liua Matúš Medoa Joseph Rushton Wakelinga y Yi-Cheng Zhang. "Solving the apparent diversity accuracy dilemma of recommender systems". En: (2010). DOI: 10.1073/pnas.1000488107.

[18] Bruce Alberts Alexander Johnson Julian Lewis Martin Raff Keith Roberts Peter Walter. *Biología molecular de la célula*. Ediciones omega, 1983.

[19] Martin H. Schaefer Jean-Fred Fontaine Arunachalam Vinayagam Pablo Porras Erich E. Wanker y Miguel A. Andrade-Navarro. *HIPPIE Website*. URL: http://cbdm-01.zdv.uni-mainz.de/~mschaefer/hippie/information.php#score.

[20] Martin H. Schaefer Jean-Fred Fontaine Arunachalam Vinayagam Pablo Porras Erich E. Wanker y Miguel A. Andrade-Navarro. "HIPPIE: Integrating Protein Interaction Networks with Experiment Based Quality Scores". En: (2012). DOI: 10.1371/journal.pone.0031826.