Red Bibliográfica y Big Data

Lucas Alonso, Emanuel Chironi, Francisco Correa, Federico Sevlever

Diciembre 2018

1 Introducción

El estudio de las redes bibliográficas comenzó en la década de los años 60, con el trabajo de Dereck de Sola Price y Garfield [1], quienes realizaron análisis de frecuencias e indexación a partir de redes de publicaciones científicas. Con el auge de la informática, la minería de datos y el análisis de redes, estos estudios comenzaron a multiplicarse debido a la aparición de repositorios y bases de datos creadas a partir de motores de búsqueda o procesos manuales en conjunto con machine learning. Los estudios bibliométricos también incluyen las redes de colaboraciones científicas, en las cuáles los nodos de la red no son ya los artículos, sino sus autores. Este tipo de estudio permite detectar, por ejemplo, la existencia de "colegios invisibles".

La importancia de las redes bibliográficas se debe a que aportan datos acerca del modo como una o varias disciplinas han evolucionado a través del tiempo. Esto incluye la conformación de nuevos campos de estudio, el cambio en las características de las relaciones que mantienen distintos campos, así como la influencia de artículos o autores individuales en procesos de avance científico. En su estudio pionero, de Sola Price analizaba la evolución de los artículos científicos en cuanto a la cantidad de referencias que un artículo puede tener, así como la "vida útil" de un artículo científico y proponía que, en general, la vida útil de un árticulo es muy corta. De acuerdo con su estudio, ya para 1961, en el lapso de diez años un artículo podía volverse obsoleto (la probabilidad de un paper con antiguedad mayor a diez años de ser citado por otro paper es muy pequeña). Esto conforma una especie de frente de investigación activo, tal que los papers que van saliendo citan mayoritariamente a los papers que han salido en los últimos tiempos, de modo que un artículo, salvo con excepción de los clásicos, pierde vigencia rápidamente.

En este trabajo nos proponemos analizar la evolución en el tiempo de una de tales redes bibliográficas, a partir de herramientas de teoría de redes. En particular, esperamos no solamente poder reproducir y reinterpretar en términos de redes algunos de los resultados de De Sola Price, sino que analizaremos algunos puntos adicionales tales como el nivel de interdisciplina en las publicaciones especializadas y la conexión entre componente de la red.

2 Base de Datos

Para este trabajo hemos utilizado una base de datos elaborada por el proyecto "Open Accademic Graph" [2], denominada "Microsoft Accademic Graph". Esta consta de 166 archivos, cada uno de los cuáles tiene alrededor de 1 000 000 de documentos. En líneas generales, la base de datos contiene no solamente artículos de revista, sino también artículos de magazines, libros, capítulos de libros y artículos periodísticos. La información se encuentra en formato JSON a modo de diccionario, donde cada línea del archivo se corresponde con un artículo que contiene los siguientes campos: id (código único identificador de cada artículo), campos de estudio, palabras clave, título, abstract, autores (nombre y filiación), revista, año de publicación, idioma, tipo de artículo (si es paper, libro, etc) y referencias (como lista de ids).

3 Filtrado de la Base de Datos y Armado de la Red

Es evidente que una red del orden de 160 000 000 de nodos no es manejable. Por consiguiente, se decidió aplicar algunos criterios de selección tales que permitieran reducir el volumen de datos a un nivel razonable, pero teniendo el cuidado de no introducir ningún sesgo.

Es importante tener en cuenta que la base de datos fue repartida entre los 166 archivos de un modo aleatorio por sus creadores. Esto se determinó a partir de un análisis de la distribución de distintas características de los documentos entre los distintos archivos y se encontró que todos los valores son muy similares. Esto se hizo tanto para la distribución de años de publicación, de cantidad de referencias y de campos de estudio. La figura 1 muestra la distribución de referencias para 9 archivos diferentes. Gracias a esa homogeneidad fue posible seleccionar una submuestra de documentos del tamaño deseado. En este caso puntual, debido a razones de complejidad computacional se decidió tomar solamente 40 de los 166 archivos. Dado que eso constituye poco menos del 25% del total, se estimó que resultaba una muestra razonable.

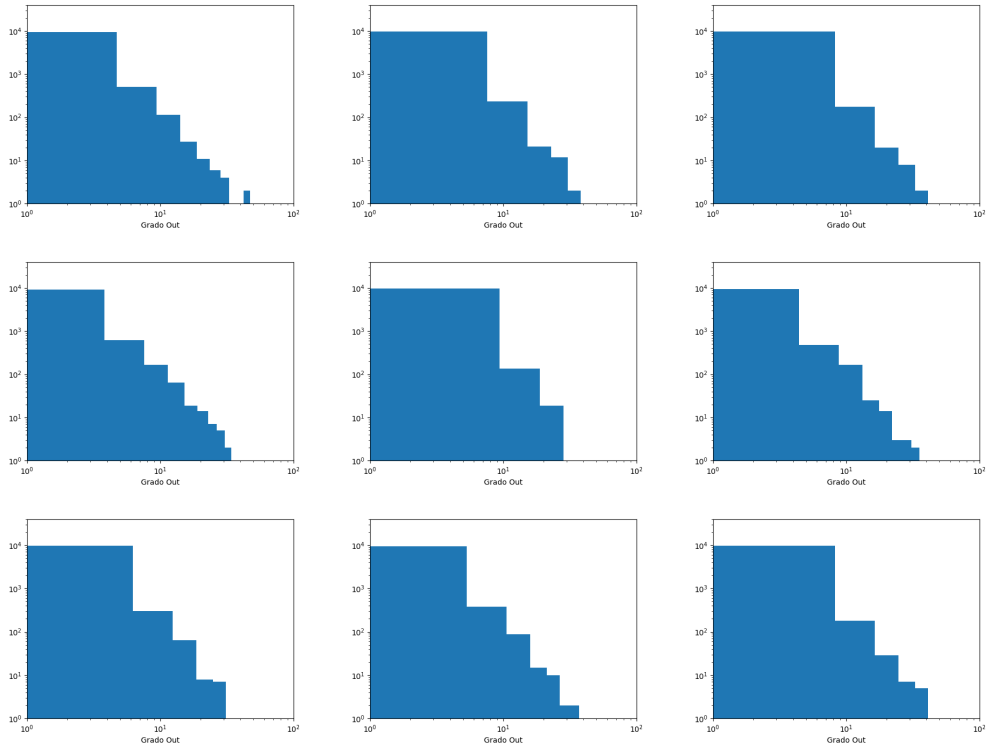

Figura 1: Distribuciones de grado out (referencias) para 9 redes de 10000 papers cada una provenientes 9 archivos diferentes de la base de datos. Puede verse que las distribuciones son similares de manera que no se introducirá un sesgo poblacional seleccionando una muestra de archivos del total.

3.1 Primer filtrado

En primera instancia, se aplicaron algunos criterios conceptuales para reducir aún más el tamaño de la red. Dado que se proponía el estudio de artículos científicos, se descartaron todos aquellos documentos que no eran papers. Además, dado que resultaba relevante el estudio de los campos temáticos y las revistas de las cuáles provenían los artículos, se descartaron todos aquellos artículos que estuvieran incompletos en los campos de palabras clave, campos de estudio y revista. Por último, también se filtraron solo aquellos papers escritos en inglés.

3.2 Segundo filtrado y construcción de la red

Luego del primer filtrado se procedió a construir la red dirigida donde cada nodo corresponde a un paper y cada enlace dirigido a una referencia (el enlace sale del paper que refiere y llega al paper que es citado). Una muestra de la red obtenida puede verse en la figura 2. El problema fue que en lugar de una gran red interconectada

se tenía una componente disconexa y pequeña por cada paper. A partir de esto lo primero que se intentó fue agregar la mayor cantidad de papers de la base de datos a la red, esperando aumentar la probabilidad de encontrar alguna de las referencias en el archivo y conectar las distintas componentes, pero antes de obtener una componente gigante de tamaño considerable para estudiar se agotaba la memoria RAM de la PC con la que se trabajó. Para resolver este problema, se aplicó el segundo filtro que consiste en seleccionar solamente los artículos que tuvieran referencias, pero que, además, hubieran sido citados al menos una vez por otros artículos dentro de la base. El objetivo de este filtro en particular, era procurar que la red no saturara con papers aislados o que se presumían poco relevantes. En efecto, es posible considerar que, un artículo que no es citado nunca, no contribuye al avance del campo de estudio y que, evidentemente, la comunidad de su tiempo no ha considerado relevante. Esta manera de filtrar los nodos garantiza que cada uno de los nodos esté vinculado con al menos otro nodo; es decir, no hay nodos aislados en esta red.

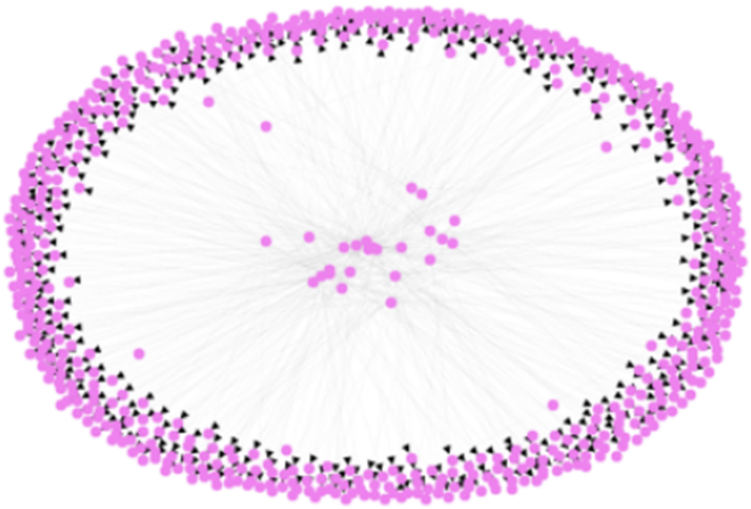

Figura 2: Muestra de la red sin aplicar el segundo filtrado. Puede verse que no hay papers que citen y sean citados y por ende no es más que un conjunto de redes disconexas donde cada una tiene un solo paper central conectado a todas sus referencias

3.3 Tercer filtrado y corrección manual de los campos de estudio

Para el tercer y último filtrado, se trabajó con los campos de estudio. Se encontró que los distintos artículos podían tener varios campos simultáneamente, donde existían mezclas de campos bien generales como "Física" o "Geología" con campos muy específicos como "Espectrometría de masas" o "Corteza Continental". La distribución de los campos puede verse en la figura 3. Para simplificar el trabajo se eligió un solo campo por artículo. El procedimiento que se usó para hacerlo fue el siguiente: para cada paper se tomaron todos los campos que contenía y se los ordenó en una gran lista. A esa lista se le realizó un análisis de frecuencias para determinar la frecuencia con que cada valor posible de campo de estudio aparecía en la base filtrada. Una vez determinadas las frecuencias, los campos se ordenaron en forma decreciente. Entonces, para cada artículo se conservó aquel campo que tuviera la mayor frecuencia, de modo que el campo de estudio más frecuente elegido fuera también el más general.

Con todo, la cantidad de campos diversos de estudio resultaba grande. El problema es que hay una gran dispersión de valores con muy poca incidencia. Por ende, lo que se hizo fue estudiar qué proporción de la red podría mantenerse si solamente se consideraran los artículos correspondientes a los N campos de estudio más frecuentes, con N un número entero. Se encontró que para apenas 25 campos de los más de 1400 que había en total, se conservaba el 97 % de la red ya filtrada. En consecuencia, lo que se hizo fue eliminar los artículos correspondientes a los demás campos. Por otra parte, era simple ver cualitativamente que, de los 25 campos que se conservaron, algunos de ellos eran sub-campos fácilmente identificables. Por ejemplo, tal es el caso de "Óptica" y "Física", o de "Química Orgánica" y "Química". Uniendo los artículos de los sub-campos a los campos asociados, los 25 campos se redujeron a 13: biología, medicina, química, física, psicología, matemática, economía, ciencias de la computación, geología, ingeniería, sociología, y los interdisciplinarios bioquímica y ciencia de los materiales.

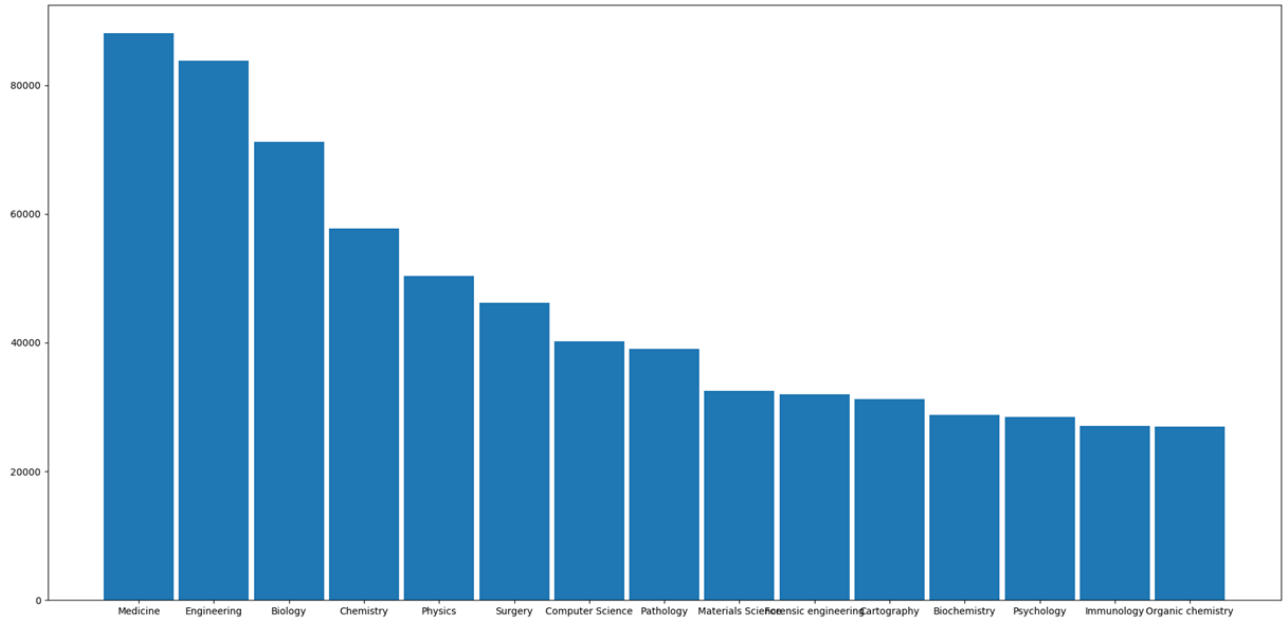

Figura 3: Distribución de los campos más frecuentes sin aplicar el tercer filtrado

4 Características generales de la red

Una vez filtrados los archivos se procedió a elaborar la red de citas y a realizar una caracterización general de sus propiedades. La red total constaba de 2 751 733 papers (nodos) y 9 908 260 citas (enlaces dirigidos). Las relaciones entre nodos están definidas de modo tal que el enlace simboliza la relación "A cita a B". Esta relación es asimétrica, de modo tal que la red resultante es *dirigida*.

4.1 Crecimiento

En su estudio previo, De Sola Price había encontrado algunas condiciones interesantes en cuanto al comportamiento de las redes de artículos. En primer lugar, el mencionado autor, al trabajar con una extensa base de artículos anteriores a 1962, detectó la existencia de un crecimiento de perfil exponencial en la cantidad de artículos publicados. Además, determinó que cada paper tenía un promedio de 15 referencias. En la misma línea, detectó que la cantidad de papers con n referencias decrecía con $1/n^2$. En términos de redes, eso equivale a decir que el grado de salida de los nodos presenta una distribución de cola pesada, libre de escala, con exponente -2. En este trabajo, se graficó la cantidad de artículos publicados en función del tiempo. Mostramos los resultados de uno de los 166 archivos en la figura 4. Constatamos que los distintos archivos poseen idéntica distribución. La forma de las curvas está de acuerdo con lo que había planteado De Sola Price. Más aún, los pozos que se observan, asociados con las guerras mundiales (franjas rojas), habían sido oportunamente indicados por el autor. Es interesante notar que luego de ambas guerras la tasa de crecimiento se incrementa hasta alcanzar el ritmo previo a cada guerra en lugar de retrasarse la cantidad de años correspondiente.

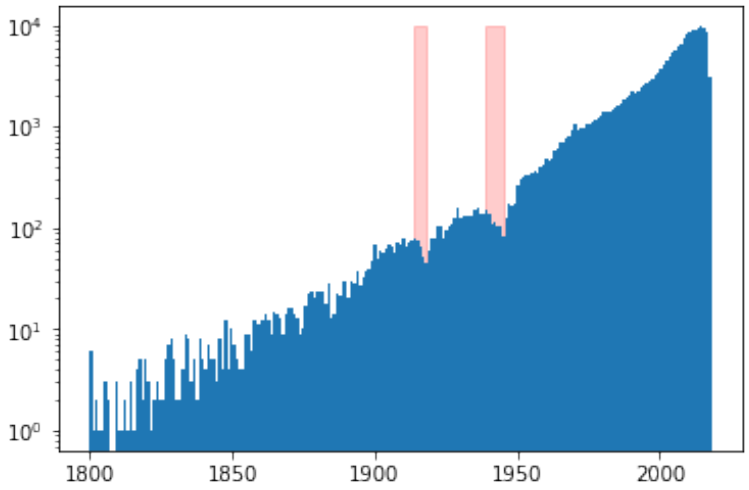

Figura 4: Muestra para la distribución de papers por año. En bandas rojas, los períodos correspondientes a las dos guerras mundiales.

4.2 Distribución de Grado

En cuanto a la cantidad media de referencias por artículo, se encontró que este valor es de 25. Finalmente, se analizó la distribución de grado de entrada, salida y total para esta red. El análisis de grados de entrada equivale a ver cuán citados son los distintos artículos, mientras que el análisis de grados de salida equivale a ver cuántas referencias tiene cada artículo. Ambas distribuciones de grado se analizaron a efectos de determinar sus exponentes característicos, asumiendo y ajustando con una distribución de la forma $p(k) \sim k^\gamma$ donde γ sería el exponente negativo. Sin embargo, existe el problema de que mientras más papers se incluyen en la red, también aumenta el número de citas y la probabilidad de que un dado paper sea citado. Es por esto que estos ajustes se realizaron en función de la cantidad de papers de la red, seleccionando submuestras al azar y llegando a la red completa, para luego sacar conclusiones con respecto a si la cantidad total de papers incluídos en la red es suficiente.

Las distribuciones y los resultados del análisis se pueden consultar en la Fig. 5.

En todos los casos, se observa que los ajustes son buenos a partir de un $K_{mín} \sim 10^3$ y que para baja cantidad de papers, γ del grado total coincide con el γ para las referencias, pero a partir de $\sim 10^6$ papers, γ se aleja del

exponente de las referencias y se acerca al de la citas. Esto se explica con el hecho de que si tenemos poca cantidad de papers elegidos aleatoriamente, la probabilidad de que se citen entre ellos es baja y sólo se observan sus referencias, pero, a medida que aumentamos la cantidad de papers de la red, comienza a haber citas entre ellos y aparecen hubs (existen papers con casi 1000 citas, pero no hay papers con más de 300 referencias). Para la red total, podemos observar que es libre de escala tanto la distribución de grado total como la de citas, econtrándose en el rango correspondiente a "Ultra Small World", lo que significa que el camino entre dos papers cualquiera es increíblemente corto comparado con el tamaño de la red. Sin embargo, puede verse que ambos exponentes no se estabilizaron para la cantidad de papers de la red total e incluso parecerían hacerse más pequeños aún si tomaramos una mayor cantidad de papers. No ocurre lo mismo para la distribución de grado de las referencias, la cual sí parece estabilizarse en el régimen "Small World" en concordancia con lo visto en la materia.

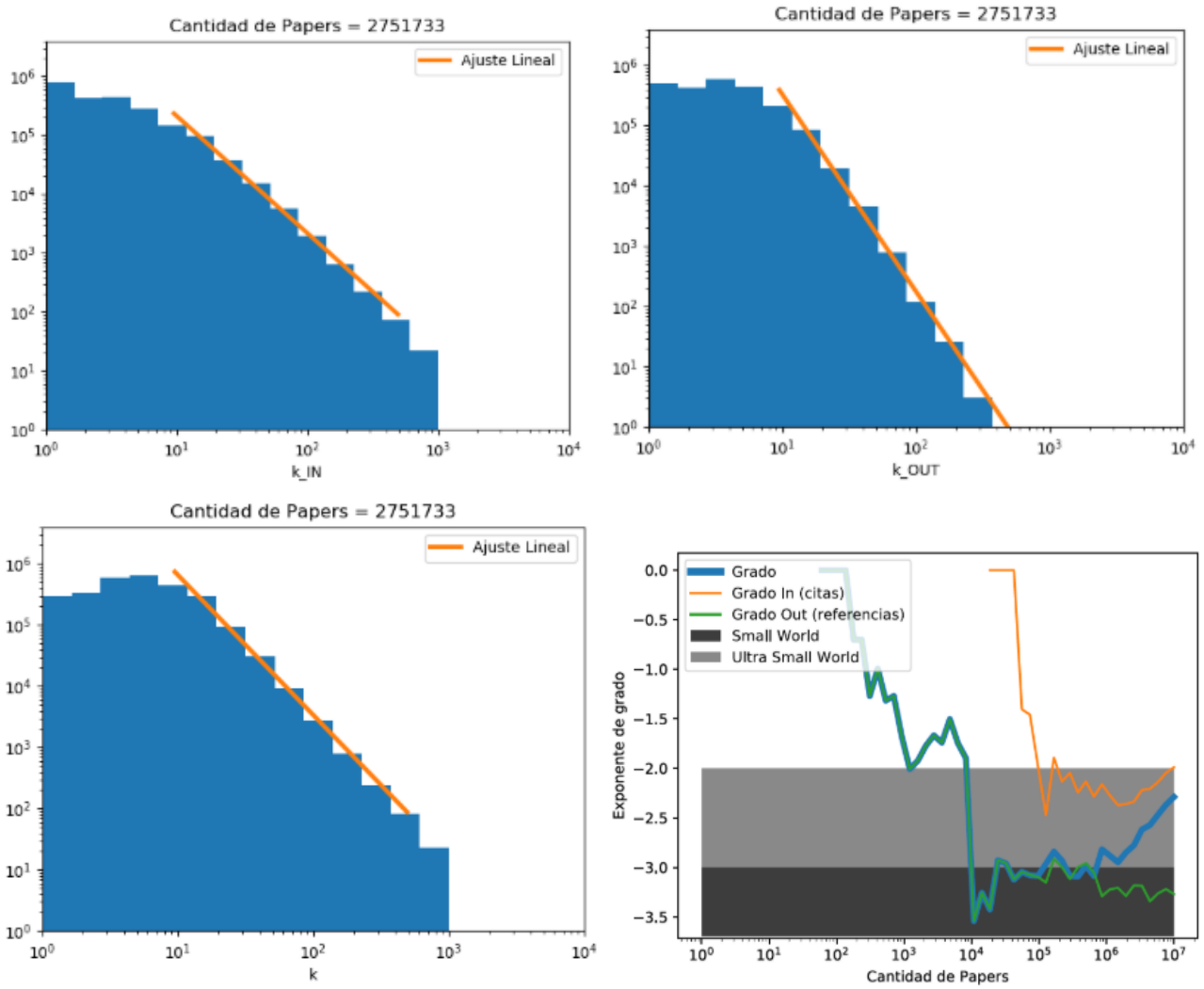

Figura 5: Los gráficos que se muestran en esta figura son histogramas de las citas (k_{in}) en el panel superior izquierdo, referencias (k_{out}) en el panel superior derecho y grado total k de cada artículo (suma de k_{in} y k_{out}) en el panel inferior izquierdo. En el panel derecho inferior se muestra la evolución de la pendiente del ajuste lineal de los histogramas previos a medida que la cantidad de papers aumenta.

4.3 Asortatividad

Una medida que permite detectar si los papers hubs correspondientes a los más citados tienden a conectarse más entre sí o no es la asortatividad. Se espera que, al ser los papers clásicos fundadores de distintas áreas de estudio, los papers hubs tiendan a conectarse más con nodos de grado bajo que entre sí. La asortatividad fue calculada a partir del ajuste del gráfico de la figura 6, de donde se obtuvo $\mu = 0.12$ a partir del ajuste. Este valor es coherente con lo esperado para una red de citas de acuerdo a lo visto en la materia, sin embargo, para extraer conclusiones en cuanto a si los papers hubs se conectan más o menos entre sí hace falta comparar este valor con una distribución del mismo para un conjunto de redes recableadas aleatoriamente. Este trabajo queda

pendiente debido a los tiempos computacionales que llevó recablear y obtener la distribución de grados medios para muchas versiones recableadas de la red.

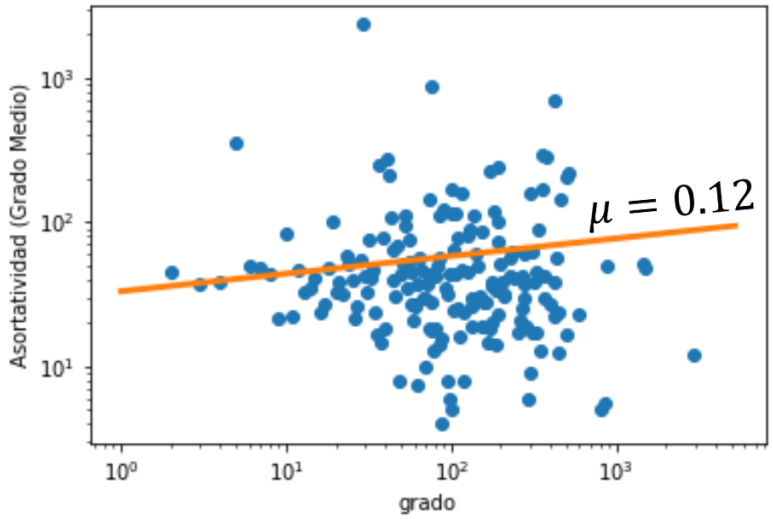

Figura 6: Asortatividad de la red medida de acuerdo al grado medio de los vecinos en función del grado. En naranja, el ajuste para obtener μ.

5 Análisis temporal de las componentes conectadas

La red total fue separada en sus distintas componentes conectadas: si bien hay más de 10^6 componentes en total, son todas despreciables respecto de la componente gigante, como se muestra en la figura 7. En efecto, consta de casi 3 000 000 de nodos, mientras que la siguiente componente conectada no llega a 30 nodos. Esto significa en la práctica (y teniendo en cuenta que usamos sólo el 25% del total de la base de datos) que se puede llegar de cualquier paper a otro recorriendo nada más que sus citas y referencias.

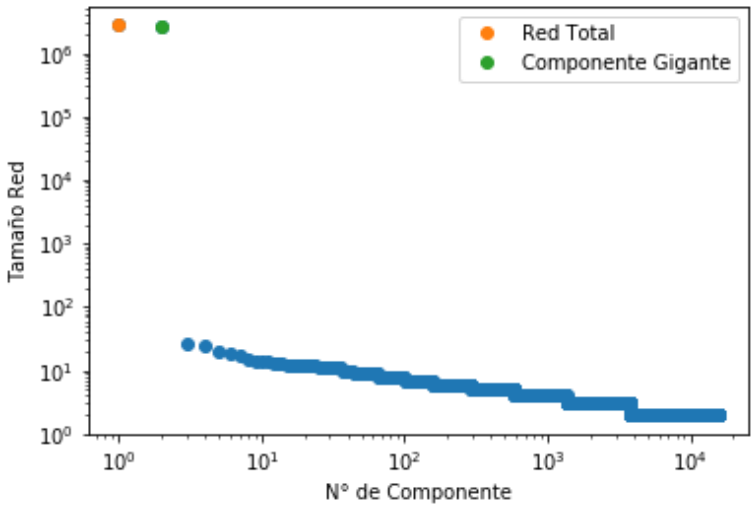

Figura 7: Tamaño de la red y de las distintas componentes.

De todas maneras, las componentes han ido cambiando con el tiempo. Por esa razón, se decidió estudiar con más detalle este comportamiento teniendo en cuenta que esto no era así en el 1900 y buscando cómo las distintas componentes se fueron conectando. La estrategia fue establecer el tamaño de las primeras cinco componentes conectadas en función del tiempo, para todo el rango de tiempos del que se disponía en la red (1850-2018). El resultado se puede consultar en la figura 8. Es importante efectuar una aclaración en cuanto al modo de interpretarla. Cada color representa una posición en el ranking de las componentes conectadas, en lugar de representar conjuntos específicos de nodos. Entonces, cuando una cierta componente se combina con otras, a traves de la aparición de un paper en determinado momento que cite dos papers de componentes distintas cambia su posición en el ranking. Esa es la razón por la cual las curvas de color en la Fig. 8 experimentan decrecimientos abruptos: la unión de dos componentes hace que las que estaban por debajo en el ranking suban y cambien de color. Podemos observar que hasta el año 1960, todas las componentes tenían un tamaño similar, y luego de esa década, la componente gigante crece muchísimo más que las siguentes componentes debido no sólo a los papers que aparecen esos años sino a que éstos papers conectan componentes más pequeñas con la gigante.

Cabe destacar la unión de las componentes 1 y 2 a la componente 0 (gigante) en el año 1960. Esta observación llevó a la pregunta sobre si se estaba observando la unión de distintos campos de estudio. Es por esto que se procedió a desglosar la cantidad total de papers de la componente gigante para los distintos campos de estudio, en función del tiempo. Los resultados pueden verse en la figura 9. Nótese que los campos más frecuentes son "Biología", "Medicina", "Química" y "Física". En particular, es interesante notar que Medicina domina sobre Biología hasta los años 1950 donde la cantidad de papers en Biología pega un salto debido a la conexión de una componente menor ligada a la Biología. Este detalle es muy interesante ya que es precisamente la época en la cual se descubrió la estructura helicoidal del ADN, por parte de Watson y Crick [4]. Por tal motivo, no parece que este cruce sea accidental. De todas formas, explicar las circunstancias de este cambio requeriría un análisis más profundo. Esto se hizo para el caso de 1960 donde se vé que las componentes 1 y 2 que se unen a la gigante estaban más relacionadas a la física y la química (curvas roja y verde). A partir de este descibrimiento surge la pregunta sobre cuáles son los papers que aparecen en 1960 citando tanto a uno en la componente gigante como a otro en la componente 1 y 2. Se realizó un código que haciendo un barrido en cada año buscaba los papers que citaban al menos un paper en la componente gigante y otro en la componente 1 del año anterior. Se encontró que desde 1850 a 1959 no existe ningún paper con estas características mientras que en 1960 aparecen 2, y entre éste año y 1965 aparecen 8 que son responsables del "despegue" de la componente gigante respecto de las demás. Sus campos de estudio varían entre electroforesis, genética, bioquímica, microscopía electrónica y

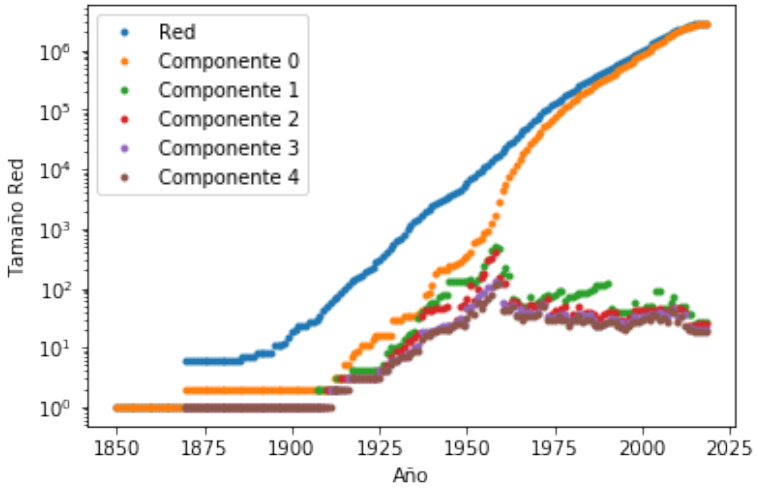

Figura 8: Evolución del tamaño de la red total y de las primeras cinco componentes conectadas, en función del tiempo. El color corresponde a la posición que la componente tiene en el ranking de tamaños en un dado momento en el tiempo, en vez de ser solidaria de un conjunto específico de nodos.

cristalografía. En cuanto a los dos papers que aparecen en 1960 [3; 5], ambos papers son sobre electroforesis, una técnica desarrollada por la Física y la Química ampliamente utilizada en Biología que consiste en la separación de sustancias por su tamaño molecular utilizando un campo eléctrico. Esta técnica fue refinada lo suficiente como para utilizarse en macromoléculas de la biología como el ADN y las proteínas recién a principios de los '60.

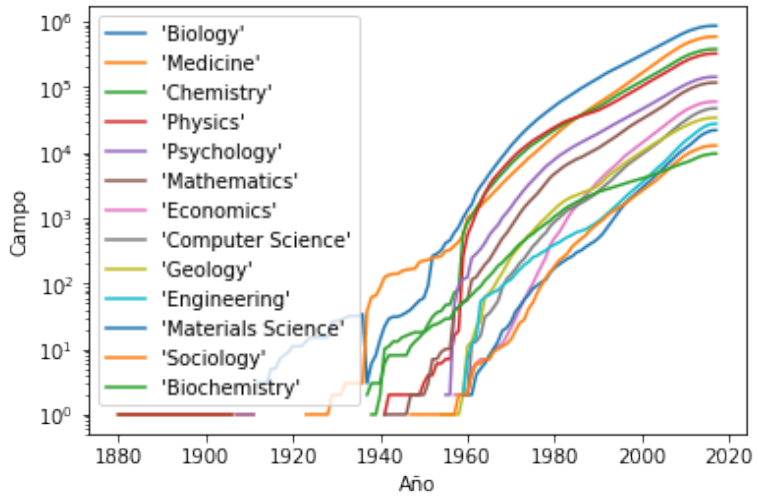

Figura 9: Evolución de la cantidad de artículos por campo de estudio, dentro de la componente gigante, en función del tiempo.

6 Interdisciplina

Un aspecto importante a estudiar en la componente gigante, es la relación de los distintos campos de estudio entre sí. En concreto, se planteó la siguiente pregunta: ¿La ciencia se ha vuelto más interdisciplinaria con el paso de los años? Para contestar la pregunta se calculó, para todos los años, qué proporción de enlaces son interdisciplinarios, definiendo como enlace interdisciplinario aquel que conecta dos papers de distintos campos o donde al menos uno de estos campos es interdisciplinario (como por ejemplo "Bioquímica" o "Ciencia de Materiales"). Los resultados se pueden observar en la Fig. 10. Esto sugiere que, si bien ambos tipos de enlaces (inter e intradisciplinarios) crecen en el tiempo, la intradisciplina comenzó antes y entonces, teniendo en cuenta la ventaja temporal y el crecimiento exponencial, la diferencia logarítmica se mantiene haciendo que la proporción de enlaces interdisciplinarios se haga cada vez más pequeña.

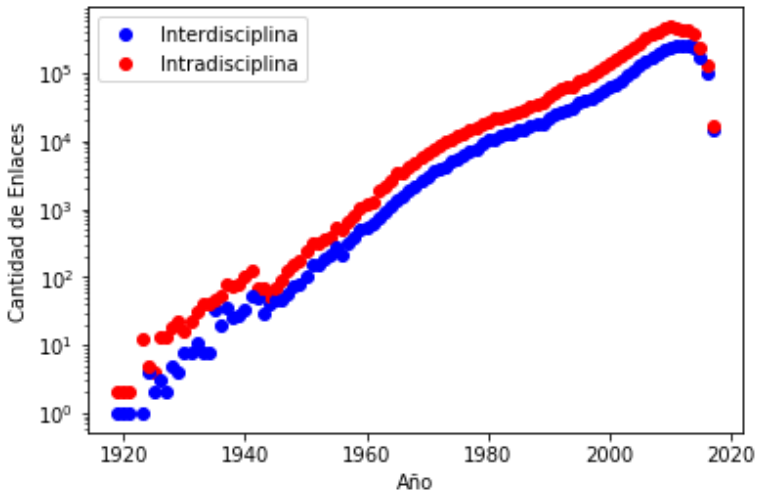

Figura 10: Evolución de la cantidad de artículos con tereferncias inter e intra campos.

Otro análisis que se hizo fue estudiar la relación entre los años de los papers que citan y los citados por inter e intra disciplina. Los resultados pueden verse en la figura 11, donde además de observar también que la cantidad de citas intradisciplinaria es mayor, se observa que los papers casi no son citados pasados los 10 años de publicación, exactamente como había encontrado de Sola Price.

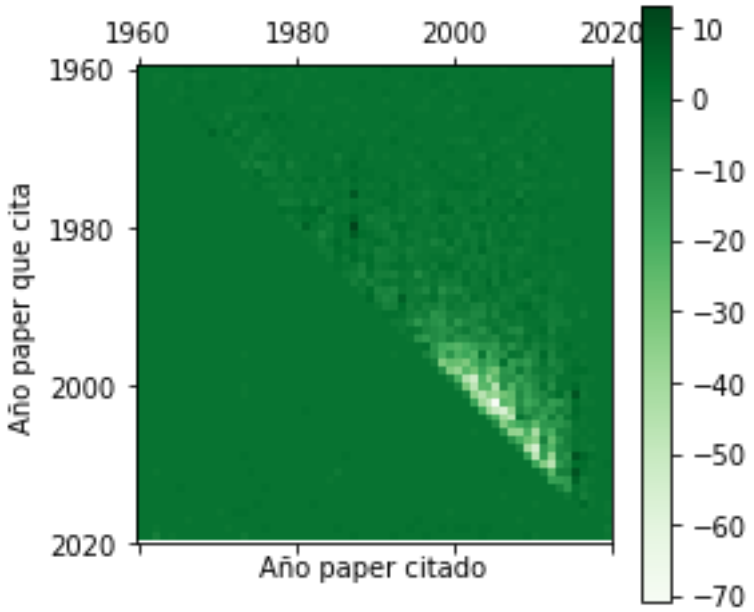

Figura 11: Campo de color de citas entre papers para los distintos años. El código de colores indica la diferencia entre citas interdicsiplina e intradisciplina. Puede verse que en general son más la cantidad de citas intradisciplina (colores claros) y la existencia de un frente de alrededor de 10 años de la actualidad hacia atrás a partir del cual casi no existen citas hacia papers más antiguos.

7 Conclusiones y perspectivas a futuro

En conclusión, se logró construir una red de papers y citas no sesgada en campos de estudio a partir de la extensa base de datos con la que se trabajó. Se estudió la distribución de grado total, grado *in* y *out* y se comprobó que se trata de una red libre de escala en el régimen de "Ultra Small World". Por otro lado, se encontró que las distintas disciplinas como la Física y Química conformaban redes disconexas respecto a la de Biología y Medicina, pero que en 1960 se unen gracias a la implementación de nuevas tecnologías experimentales y es en esa década que la red de papers se hace casi completamente conexa. A futuro sería ideal repetir el análisis para el total de los archivos de la base de datos y buscar los papers que unen la componente gigante a cada una de las otras antes de 1960, así como aplicar algoritmos de separación en comunidades los cuales no logramos correr a tiempo por falta de poder computacional.

References

[1] Derek J. de Solla Price. Networks of Scientific Papers. In SCIENCE, VOL. 149, 1965.

[2] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. ArnetMiner: Extraction and Mining of Academic Social Networks. In Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'2008). pp.990-998.

[3] Rossan R., Serum Proteins of Animals Infected with *Leishmania donovani* with Special Reference to Electrophoretic Patterns. Department of Zoology, Rulgers-The State University, New Jersey. 1959.

[4] J.D. Watson, F.H.C. Crick. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. Nature. 1953.

[5] B. S. Blumberg. Genetics and Rheumatoid Arthritis. National Institute of Arthritis and Metabolic Diseases, National Institutes of Health, Bethesda, Md. Journal of American College of Rheumatology. 1960.