Zotenkopedia: Correlaciones entre la topología y el interés de los usuarios en la Wikipedia

Emanuel Ferreyra, Bruno Kaufman, Ariel Salgado

12 de diciembre de 2018

Resumen

Este trabajo presenta un análisis de la información recolectada de la Wikipedia, a través del empleo de un *scrapper* basado en los paquetes de R `wikipediR` y `curl`. Los datos corresponden a las páginas que incluyen los términos *"graph "* o *"graphs"* en sus nombres. Presentamos varios análisis involucrando tanto a características topológicas locales de la red como datos intrínsecos a los artículos (como el largo en caracteres o la cantidad de visitas mensuales). Estos incluyen análisis de asortatividad y homofilia, y correlaciones entre centralidades y información adicional de cada página. A través de esto se exploran conceptos de esencialidad y funcionalidad de artículos en la red, y su conexión con la topología.

1. Introducción

Casi desde el comienzo de la aplicación de la teoría de grafos a situaciones reales se ha buscado encontrar correlaciones entre la estructura observada de una red, e información adicional disponible. Por ejemplo, la mayoría de las aplicaciones a transmisión de enfermedades o de formación de opinión, toman una red de relaciones entre personas, y buscan determinar como se propagará una enfermedad o una idea, suponiendo que el proceso de contacto guarda alguna relación con las conexiones establecidas en la red. Por otro lado, nos motivamos con el trabajo de letalidad-centralidad de Zotenko et al., donde se analizan correlaciones entre la relevancia de un nodo de la red y el lugar que ocupa en la topología de la red [1].

En el presente trabajo el interés está en buscar correlaciones entre la estructura de una red de páginas de Wikipedia y la información disponible para cada página. Dado que la amplitud de la información disponible excede nuestro poder de adquisición y almacenamiento de datos, es útil parcializar la información disponible. Por esto consideraremos no toda la enciclopedia, sino únicamente las páginas correspondientes a ciertas categorías[1]. La red, que en este caso es dirigida, se construye tomando las páginas como nodos, y los vínculos entre ellos como los hipervínculos entre las páginas. Por cada hipervínculo que conecta la página A con la página B, establecemos un eje en el grafo que apunta de A hacia B. Para nuestro análisis no consideramos los autoejes. A cada página se le asocia una o más categorías, así como cierta información extra, como la longitud de la página o la cantidad de vistos en el último mes.

[1] Wikipedia indica para cada página algún conjunto de categorías a las que esta pertenece.

Buscaremos dar respuesta esencialmente a tres preguntas: ¿Las páginas importantes en términos de su metadata tienen un rol importante en la conectividad de la red? ¿Las páginas importantes en términos de la conectividad de la red son relevantes en términos de su metadata? ¿Cuán difícil sería eliminar un tema de Wikipedia?

2. Adquisición de los datos

Para obtener los datos construimos una serie de scripts en `R` que toma los información de Wikipedia en dos partes. Inicialmente hicimos uso del API de Wikipedia para `R` para obtener los vínculos salientes de una página (incluyendo las categorías a las que pertenece). Luego, basándonos en el paquete `curl` para `R`, obtuvimos datos asociados a las páginas individuales (metadata). El algoritmo de descarga a través del API es el siguiente:

1 Comenzamos con una página (por ejemplo Graph (discrete mathematics)), un conjunto de categorías de interés (por ejemplo Graph theory), y un criterio de incorporación de categorías (en nuestro caso, todas aquellas que cumplan con tener las palabras "graph " o "graphs" (no consideramos mayúsculas o minúsculas).

2 Empleando el API de Wikipedia, buscamos todas las páginas a las que apunta la página anterior. Guardamos todas estas conexiones en un archivo.

3 Para cada una de las páginas que tenían conexión con la página de inicio, revisamos si tiene alguna categoría que cumpla con el caso de interés. En caso afirmativo, la agregamos a una lista de búsqueda.

4 Repetimos el proceso desde [1] para cada página en la lista de búsqueda.

5 Rearmamos la lista de categorías de interés con todas las nuevas categorías encontradas que cumplan con el criterio.

Dado que conforme el algoritmo completa un ciclo aparecen nuevas categorías, realizamos multiples veces el scrappeo, hasta que ya casi no aparezcan páginas nuevas (en 1000 iteraciones no aparece ninguna página nueva).

Al finalizar este proceso, obtenemos una lista de conexiones, mediante la cual podemos construir nuestra red con las funcionalidades provistas por `igraph`. La red obtenida con este algoritmo siempre es conexa por la forma en que generamos los datos.

A continuación, para cada una de las páginas encontradas, obtenemos datos numéricos pertinentes al artículo, como su largo en caracteres o la cantidad de visitas que tuvo en el último mes. Esto se realiza con el paquete `curl`, con el cual descargamos el contenido de una página en código `html`. Se trata de la página de información que cada artículo posee, que contiene una tabla con varios datos que utilizan los editores de Wikipedia para informarse, y que era fácilmente accesible desde el navegador para corroborar nuestra información.

Luego, utilizando `stringr` y funciones base de `R`, encontramos hilos de texto que señalan la proximidad de los datos que nos interesan en el código, y conociendo el patrón con el cual se escriben, se obtiene dicho dato con el nombre que

le corresponde, totalizando 21 atributos diferentes. De entre estos, nos quedamos inicialmente con los que eran numéricos: el tamaño de la página en bytes, un ID, la cantidad de editores de wikipedia que participaron en ese artículo, cuántos de ellos revisaron las ediciones más recientes, el número de páginas que son redireccionadas a la entrada en cuestión, la cantidad de visitas que tuvo la página en los últimos 30 días y finalmente el número total de ediciones junto con las fechas de creación y de última edición.

En un análisis posterior, habiendo considerado diferentes formas de relacionar estas cantidades, observando similitudes y diferencias y el potencial de predicción que tenían, decidimos continuar con sólo cuatro de ellas: el largo (Length), las visitas mensuales (MonthCount), los revisores (Watchers) y las ediciones mensuales.

Realizamos la descarga de los datos partiendo de tres páginas y criterios de categorias:

- Páginas en la temática de teoría de grafos, partiendo de `Graph (discrete mathematics)`, y categoría `Graph Theory`, y considerando las categorías que incluyeran los textos ‘‘`graph `’’ o ‘‘`graphs`’’.
- Páginas en la temática de comunismo, partiendo de `Communism`, y categoría `Communism`,`Anarchism`, y `far-left` y considerando las categorías que incluyeran los textos ‘‘`communism`’’, ‘‘`anarchism`’’ y ‘‘`far-left`’’.
- Páginas en la temática de artes marciales, partiendo de `Martial arts`, y categoría `martial arts` y considerando las categorías que incluyeran los textos ‘‘`martial arts`’’.

3. Resumen topológico

Después de correr el scrapper basado en la API `wikipediR` obtuvimos tres redes cuyas cantidades de aristas y nodos se sintetizan en la tabla 1.

	Cantidad de Nodos	Cantidad de aristas
Communism	1106771	6751996
Graphs	484330	2303067
MartialArts	1153950	5892718

Tabla 1: Tamaño de la red

En las tablas 2 y 3 se muestra el solapamiento de las redes que relevamos. En la posición i,j se encuentra la fracción de nodos o ejes de la red i que pertenecen a la red j, según corresponda.

Podemos observar que la red de grafos es la que más contenida está en las otras dos, pero contiene alrededor del 20 % de los nodos de estas. Por otro lado, la intersección entre las redes de comunismo y artes marciales tiene una intersección del 38 %.

Este solapamiento tan grande de los nodos no se corresponde con el de las aristas. Podemos observar que a pesar de haber llegado a los mismos nodos a traves de los hipervínculos de la Wikipedia, los caminos que tomamos son muy diferentes, puesto que la contención de aristas de una red en otra solo en un

caso supera el 10 % (el 12 % de las aristas de la red de grafos está contenida en las de comunismo) mientras que por ejemplo las redes de comunismo y artes marciales tienen solo un 3 y 4 % de sus ejes contenidos en la de grafos. Es decir que a pesar de compartir bastantes nodos, la cantidad de aristas comunes entre estas redes es muy baja.

	Communism	Graphs	MartialArts
Communism	1.00	0.23	0.38
Graphs	0.54	1.00	0.47
MartialArts	0.37	0.20	1.00

Tabla 2: Fracción de nodos compartidos entre las distintas redes.

	Communism	Graphs	MartialArts
Communism	1.00	0.04	0.08
Graphs	0.12	1.00	0.08
MartialArts	0.09	0.03	1.00

Tabla 3: Fracción de ejes compartidos entre las distintas redes.

Otra información topológica que fue relevada inicialmente tiene que ver con las distribuciones de grados. Aquí tenemos en cuenta 3 valores: el grado in, es decir la cantidad de páginas que tienen un vínculo apuntando a un nodo determinado; el grado out, la cantidad de páginas que son señaladas; y el grado total, que proviene de sumar ambos grados.

A continuación están graficados los tres grados, con ambos ejes en escala logarítmica.

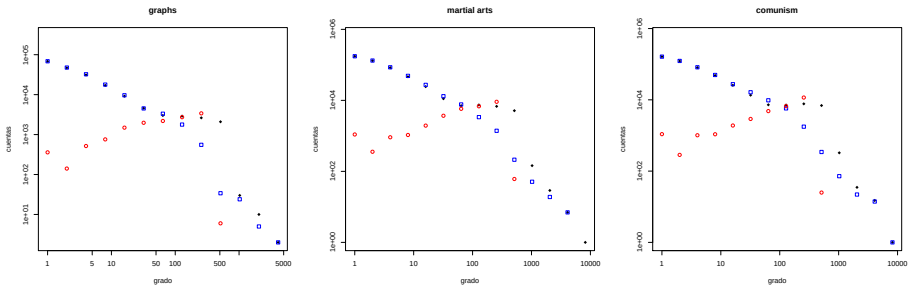
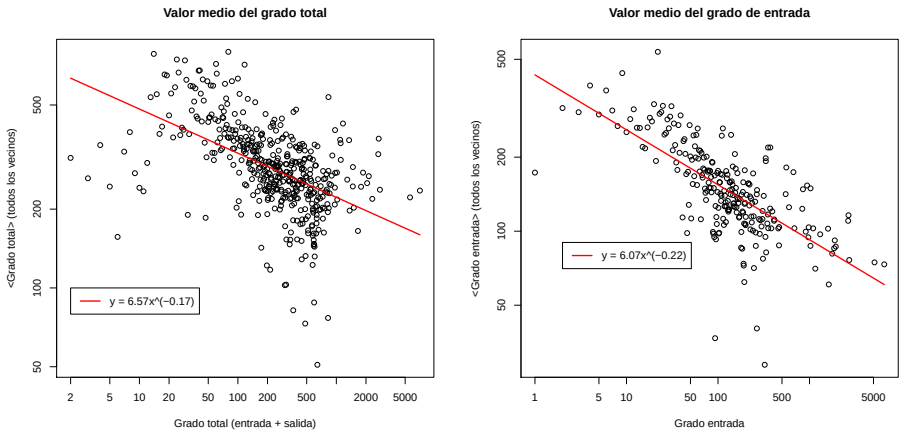


Figura 1: Distribución de grados

Se puede observar que el grado out, en rojo, es creciente hasta 500, mientras que el grado in, en azul, decrece en todo su soporte. Lo mismo sucede con el grado total. En este punto es pertinente aclarar que la API de `R` tiene un límite de 500 páginas en el atributo que indica cuántas páginas son apuntadas por una página determinada. También sabemos que este límite es bastante acertado, puesto que la cantidad de páginas con grado out mayor a 350 representa una fracción despreciable.



(a) (b)

Figura 2: Promedio del número de conexiones entrantes y salientes (a) y entrantes (b) de las páginas vecinas, en función de número de conexiones entrantes y salientes (a) y entrantes (b), de la página.

Podemos observar en los gráficos la característica libre de escala proveniente de una distribución de grado del tipo ley de potencia.

Dado que el objetivo inicial de relevar redes de distintas temáticas era considerar topologías distintas, pero encontramos una fuerte similaridad entre las distintas redes (tanto en términos de topología, como en las páginas particulares que comparten), decidimos dedicarnos únicamente al análisis de la red relacionada a teoría de grafos. En las siguientes secciones se presenta el análisis de la relación entre la metadata y la topología concerniente a esta red.

4. Correlaciones de corta escala: Asortatividad

Nos propusimos ver si a nivel local existía alguna relación entre la metadata de cada página y las de sus vecinos. Consideramos para esto el promedio de los valores vecinos, en función de distintas variables.

Partimos del grado total (número total de hipervínculos que conectan a la página) promedidado sobre los vecinos de un nodo, como función del grado de ese nodo, junto al valor medio del grado de entrada, en función del grado de entrada del nodo. Ambos se pueden observar en la figura 2. Estos gráficos muestran una asortatividad negativa, indicando que en la red las páginas con grados altos tienden a vincularse con páginas con grados bajos y viceversa. Este resultado es esperable considerando que la distribución de grados sigue una ley de potencia.

A continuación, consideramos el número medio de vistos mensuales de los vecinos, en función de los vistos mensuales de la página en cuestión. En este caso observamos poco cambio del número de vistos de los vecinos conforme cambia el número de vistos de la página. Sin embargo, bajo la hipótesis de que un usuario de Wikipedia luego de leer una página lee alguna de las que se vinculan con

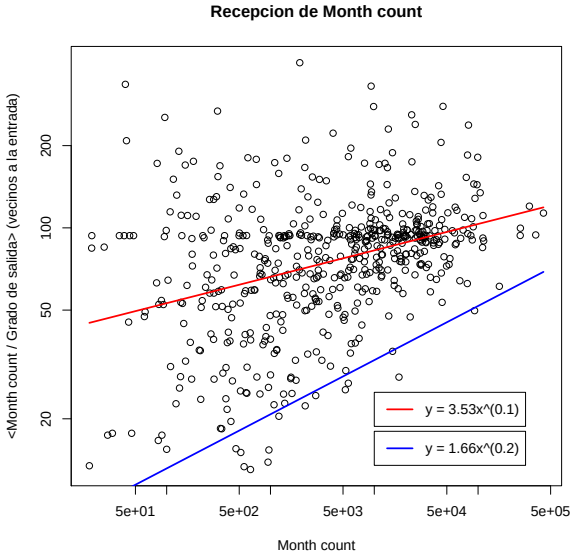


Figura 3: Vistos mensuales (MonthCount) normalizado por el grado de salida, promediado sobre las páginas que apuntan a una página, en función del Month Count de esa página. La linea roja es un ajuste del total de los puntos, y la línea azul ajusta los puntos valores mínimos en intervalos de a 50 puntos.

ella, analizamos otras dos relaciones: el número de vistos de las páginas vecinas a la entrada (i.e. páginas que permiten llegar al nodo en cuestión), normalizado por su grado de salida, en función del número de vistos de la página. En la figura 3 podemos observar el resultado. Allí vemos que, si bien la dispersión es grande, se puede observar una tendencia positiva entre ambas magnitudes. Más aun, podemos esbozar una línea límite entre el número de vistos de la página en cuestión y el número de vistos de las páginas que la apuntan, de donde podemos concluir que las páginas que alcanzan un alto número de vistos están apuntadas por páginas que emiten una fracción alta de usuarios visitantes. Esto nos da una noción de caminata aleatoria de visitantes de Wikipedia, que analizaremos más en la sección 6.

Por otro lado, en la figura 4 podemos observar una medida del traspaso de visitas mensuales. Para esto, consideramos para cada página el número medio de visitas mensuales dividido por su grado de salida de las páginas que le apuntan (las que enviarían visitas), y el número medio de visitas mensuales dividido su grado de entrada de las páginas a las que apunta (las que recibirían visitas). Vemos que existe entre ambas una correlación positiva, aunque bastante ruidosa. Interpretamos esto como una medida de la transferencia de visitas a través de las páginas. A mayor número de visitas de las páginas que me apuntan, más visitas recibiran las páginas a las que yo apunto.

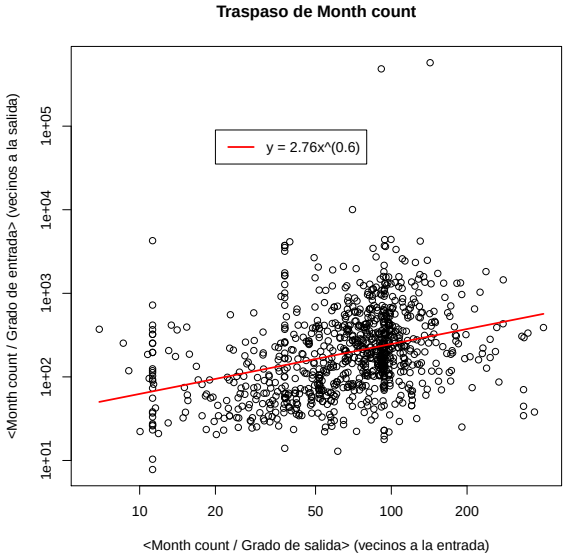


Figura 4: Vistos mensuales, normalizados por el grado de entrada y promediados sobre las páginas que son apuntadas por una, en función de los vistos mensuales normalizado por grado de salida promediado sobre las páginas que apuntan a una.

5. El destructor local: eliminando un tema de wikipedia

Durante el desarrollo del trabajo se nos planteó la pregunta de si sería dificil o no eliminar un tema de la Wikipedia. Para poder medir esto de alguna forma, lo primero que necesitamos es definir el concepto de tema. Dado que Wikipedia asigna categorías a sus páginas, nuestra propuesta fue considerar que los temas asociados a una página podían ser identificados con las categorías a las que esta pertenece. Nuestra red, que gira en torno a la temática graphs, toca 423900 categorías distintas a lo largo de sus 484326 nodos, con un número medio de 4.98 categorías por nodo, estando el segundo cuartil entre 1 y 4, y el tercero entre 4 y 8. De estas, únicamente 20 están relacionadas a la temática de grafos, con 967 página en alguna de esas temáticas. Vemos que las categorías son muy exhaustivas, siendo su número comparable al de páginas, principalmente debido a que una misma página puede pertenecer a múltiples categorías.

Sin embargo, la presencia de cada categoría en la red no es igual. En la figura 5 podemos observar la frecuencia de cada categoría en nuestra red, donde vemos que la distribución es aproximadamente exponencial. Dominan las páginas sin categoría, seguidas por Living people con 35000 páginas asociadas, seguido por Fellows of the royal society con 2700 páginas asociadas. Las páginas de la temática de grafos cuentan, por el contrario, con solo unas 1000 páginas asociadas, siendo la más importante Graph theorists con 169 páginas asociadas. Esto ya nos adelanta, partiendo de la forma en la que el algoritmo de descarga construye la red, que las páginas conectan temáticas muy distintas entre sí a través de sus

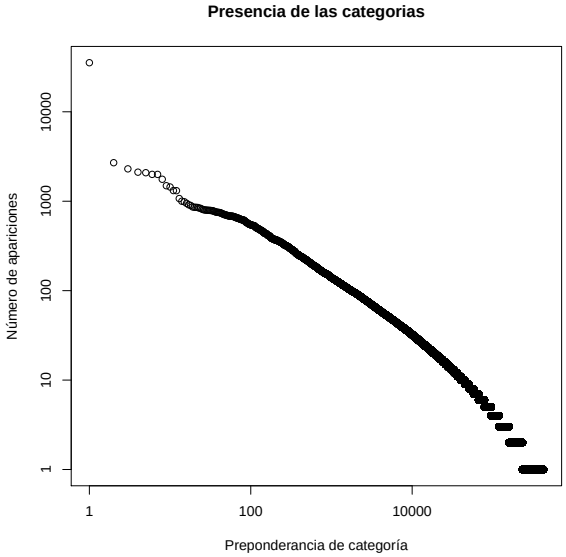


Figura 5: Presencia de cada categoría, ordenada por cantidad de apariciones. Vemos que la distribución semeja una ley de potencia.

hipervínculos dejando entrever que la red es poco redundante, y que gran parte de la red relevada no tiene páginas cuyo contenido sea estrictamente del tópico seleccionado inicialmente.

Para medir cuan similares son las temáticas de dos páginas conectadas con un hipervínculo, medimos para cada eje dos coeficientes, S_i y S_o, que representan cuan parecidas son las categorias de la página a la que llego a través del hipervínculo a las de la página de la que parto, y viceversa. Además consideramos un tercer coeficiente S_a que mide el parecido entre las categorías a ambos lados del hipervínculo. Estos se definen como

$$\begin{aligned} S_i &= \frac{|C \cap C'| - |C - C'|}{|C|} \\ S_o &= \frac{|C \cap C'| - |C' - C|}{|C'|} \\ S_a &= \frac{2|C \cap C'| - |C - C'| - |C' - C|}{|C| + |C'|} \end{aligned} \tag{1}$$

donde C son las categorias del nodo a la cola del hipervínculo y C' las categorías a la cabeza. Los tres coeficientes tienen rango entre -1 y 1, indicando total diferencia o total similitud respectivamente. Medimos el coeficiente para todos aquellos ejes que conectaban nodos con al menos una categoría. En la figura 6 podemos observar que los tres coeficientes toman valores cercanos a -1 en la mayoría de los casos, de donde podemos concluir que las categorías de las páginas suelen ser totalmente distintas entre sí. En base a esto planteamos el siguiente escenario: si considerásemos un robot que intenta eliminar un tema (i.e. una categoría) de Wikipedia, partiendo de una página con la categoría en cuestión y buscando las páginas que limitan con ella que tienen la misma

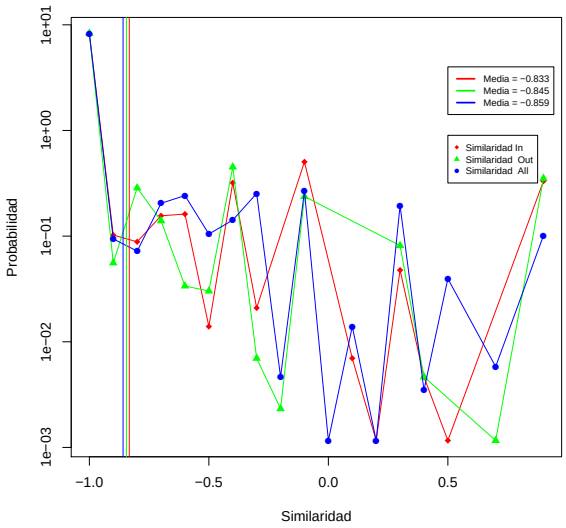


Figura 6: Distribución de las medida de similaridad a lo largo de los ejes. Vemos que la mayoría de los casos están concentrados cerca de -1, y el promedio total también es muy bajo, indicando que al atravezar un eje, lo esperable es cambiar completamente el set de categorías.

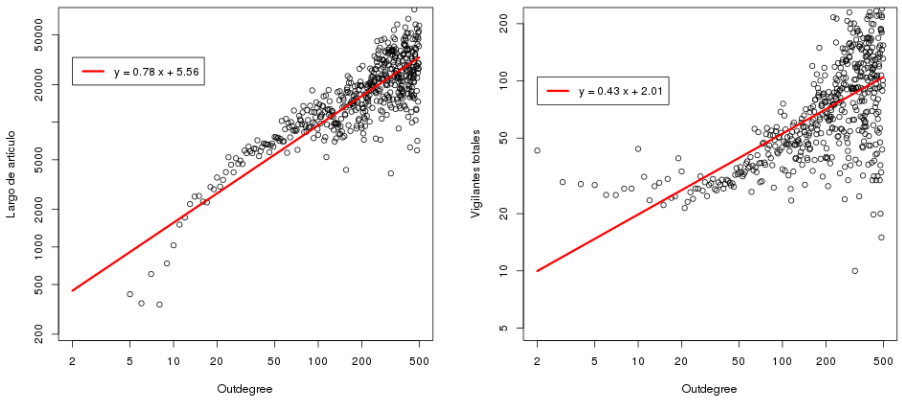
categoría para eliminarla, le resultaría muy difícil borrar completamente el tema, ya que se encontraría con muchas páginas de temas distintos.

6. Correlaciones con datos locales

Como vimos, cada artículo cuenta con una cantidad de datos generales asociados y de actividad reciente. Disponemos del largo de la página en caracteres, por ejemplo, o el número de visitas que tuvo en el último mes. Siguiendo la línea del trabajo [1] buscamos correlaciones entre las medidas de centralidad topológica (grado, page rank, etc.) y los datos asociados a la página. Entre estas, una de las más notables es el grado saliente. Podemos ver dos de dichas correlaciones en la figura 7.

Se ve que el grado saliente acompaña de forma creciente al largo del artículo y a la cantidad de usuarios vigilando el artículo. El primero se puede deber a una cuestión de espacio: cuanto más se referencie a otros artículos (o sea, cuanto mayor sea el grado saliente), mayor espacio deberá tener el artículo. Es interesante que también debe cumplirse el opuesto para verse esta correlación: cuanto más largo sea el artículo, mayor cantidad de referencias deberá tener a otros artículos. Esto es importante, y se puede deber a la *no-redundancia* de Wikipedia. Si se puede achicar el largo del artículo vinculando a un artículo que ya lo explique, efectivamente se lo hace.

La otra correlación es interesante de analizar. La cantidad de vigilantes debería ser desde un principio quienes están interesados en el mantenimiento de la página. Si se trata de quienes llegan a la página a verlo, debería correlacionarse



(a) (b)

Figura 7: Correlaciones del grado saliente con el largo del artículo (a) y con los usuarios activamente (b) siguiendo a la página (*watchers*) en escala logarítmica en ambas ejes, con una regresión lineal. Se pueden ver ciertas tendencias en ambos casos, aunque no necesariamente sean lineales en todo punto.

con el grado *entrante*, no el *saliente*. Sin embargo, como veremos pronto, este no es el caso. Una explicación es que tener una mayor cantidad de vigilantes se correlaciona con el largo del artículo (que como vimos anteriormente se correlaciona con el grado saliente). Otra explicación, posiblemente complementaria, es que al tener más vigilantes, se pueden encontrar más fácilmente vínculos con otros artículos, que como vimos en la sección 5, son de una gran variedad de temas, y por lo tanto requerirán una variedad de personas con puntos de vista distintos para plantear sus vínculos.

Por otro lado, en la figura 8 podemos ver que la cantidad de vigilantes no se correlaciona especialmente bien con el grado entrante de un artículo.

Algo que si se observa es que hay una gran zona "prohibida" hacia el lado izquierdo del gráfico. Esto indica que, pasada una cantidad de vigilantes, un aumento en el grado entrante está prácticamente garantizado. El inverso se ve sólamente en una región del gráfico, por lo que no es un comportamiento generalizable. Esto se puede deber a que los visitantes deben enterarse de la página para llegar, lo que se facilita con un mayor grado de entrada. Por el otro lado, es posible tener alto grado de entrada y que aún así nadie se interese en vigilar la página.

Además de estas correlaciones, existen algunas con mucha dispersión para visualizar, aún habiendo combinado valores en y con la función `aggregate()` de `R`, ya que se tienen puntos muy juntos en x. Por esto se acumulan puntos dentro de cierto espacio en x, considerado para describir un espaciamiento uniforme en escala logarítmica.

La más clara de estas correlaciones sucede entre el grado saliente y la cantidad de ediciones mensuales que tiene un artículo. Esto se puede ver en la figura 9.

Como se puede ver, la correlacion es positiva, y como suele ser el caso en estas

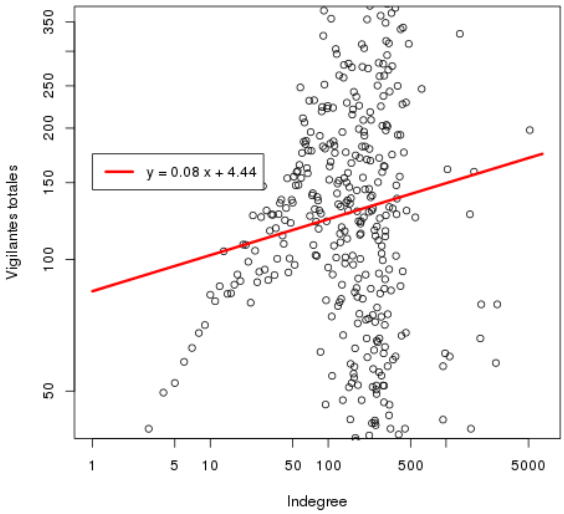


Figura 8: Correlación entre el grado entrante y la cantidad de vigilantes de un artículo, con un intento de regresión lineal que demuestra que la correlación efectivamente no es lineal. Se pueden ver áreas donde no existen puntos, que son de interés al analizar el gráfico.

correlaciones, se vuelve menos dispersa a medida que crece el grado saliente. La correlación se puede deber a algo similar a lo planteado entre el grado saliente y la cantidad de editores visto en la figura 7: podría ser que al tener más editores, se vean conexiones con más temas y que por lo tanto crezca el grado saliente.

Además de las correlaciones con el grado, existen correlaciones con la centralidad por autovector. Las observamos en relacion con el largo y con la cantidad de vigilantes, y se pueden ver en la figura 10.

Lo que se observa en estos gráficos es que tanto el largo de un artículo como la cantidad de vigilantes que tiene *disminuyen* a medida que uno se acerca a la centralidad máxima por autovalor. Esta centralidad, recordemos, se calcula como la suma de las centralidades de los vecinos, iterando un vector sobre la matriz de adyacencia. Este resultado entonces se puede entender como que los artículos con mayor largo y mayor cantidad de vigilantes son los que dependen menos de sus vecinos para establecer centralidad. Los artículos con mayor largo y cantidad de vigilantes tendran vecinos poco centrales, que asimismo tendrán largos y cantidades de visitantes similares (dado que de otra forma no se cumpliría una correlación). Esta es una forma de decirnos que hay cierta asortatividad en los largos de artículos y en las cantidades de vigilantes. Esto tiene sentido, dado que las páginas más largas tratarán de evitar escribir grandes cantidades de texto haciendo referencia a otros lugares donde esten escritas esas cantidades de texto. Por el lado de los vigilantes, también tiene sentido que algunos de los vigilantes de un artículo se conviertan también en vigilantes de algún vecino, por temáticas compartidas.

Una última correlación explorada es la de la centralidad PageRank con la

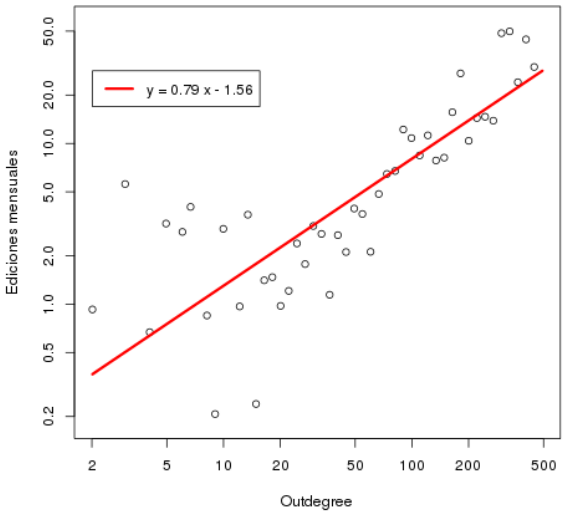


Figura 9: Correlación entre el grado saliente y la cantidad promedio de ediciones mensuales que tiene un artículo en cierto rango de grados.

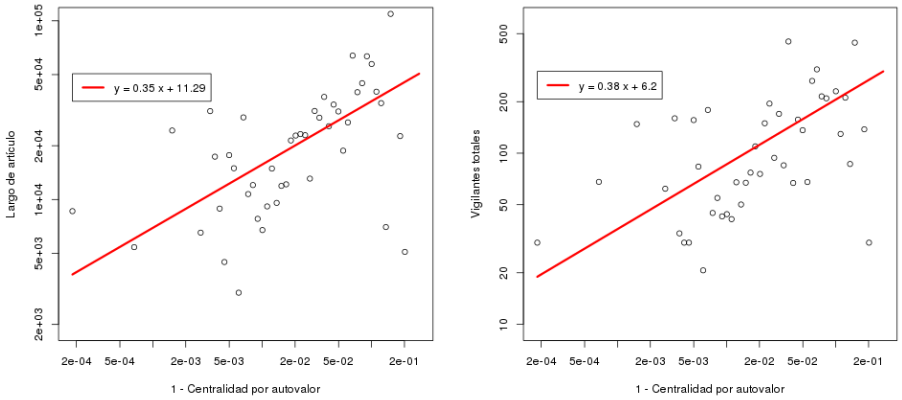
cantidad de visitantes mensuales. El PageRank se basa en una idea de caminata aleatoria, por lo que la centralidad indicaría cuantos caminantes terminan en un artículo habiendo caminado cierta cantidad de tiempo. El resultado de esta correlación se ve en la figura 11.

A partir de este gráfico podemos ver que hay cierta correlación de ley de potencia entre la cantidad de visitantes. Sin embargo, esta ley no es lineal, como debería ser para una caminata aleatoria. Esto nos sugiere tres posibilidades: por un lado, puede que la caminata aleatoria no tenga un factor de amortiguamiento $d = 0{,}66$ como supuesto a priori (esto daría un promedio de artículos visitados de 3 por caminata, suponiendo que este número sigue una distribución geométrica). Por otro lado, es posible que no sea una caminata aleatoria con probabilidad equivalente de seguir cualquier enlace (como supone el PageRank) sino que hayan enlaces más propensos a ser seguidos. Por último, es posible que hayan artículos más propensos a retener el interés de un visitante, y otros más propensos a perderlo, por lo que el factor de amortiguamiento no sería el mismo para todos los artículos.

En resumen de estas correlaciones, se ve que los grados tienen un rol muy importante en como predictores de características intrínsecas de las páginas, especialmente el grado saliente. Esto tiene sentido ya que el grado saliente se decide desde dentro de la edición de la misma página, en comparación con el grado entrante y las otras características topológicas que tienen un carácter más fuertemente no local. Como se ve en la sección 7, el grado total cumple un rol importante en la robustez topológica también.

Además, pudimos ver que la centralidad por autovalor se correlaciona inversamente con el largo de un artículo y con la cantidad de gente que lo vigila.

Finalmente, se vio cierta correlación entre la cantidad de visitantes mensua-



(a) (b)

Figura 10: Correlaciones de la centralidad por autovalor con el largo de un artículo (a) y la cantidad de vigilantes (b). Nótese que se grafica $1 - \textbf{eigen.centrality}$, y que se lo hace para los valores más cercanos a 1; al avanzar el eje x, uno se *aleja* de la mayor centralidad por autovalor. Esto se hace para poder tener una escala logarítmica rodeando a los nodos más centrales.

les y la centralidad PageRank. Esto sugiere que hay cierta semejanza con una caminata aleatoria, aunque ciertas hipótesis deben ser adaptadas o consideradas al no ser una equivalencia.

7. Destrucción topológica de la red

Como vimos en la sección anterior, ciertas características de centralidad topológica se correlacionan con características intrínsecas a los artículos. Esto nos hace preguntar qué papel pueden tener a la hora de establecer la esencialidad de los artículos como nodos.

Podemos ver en la figura 12 que el grado es la centralidad más rápida en destruir la red, seguida por las centralidades de autovalor y PageRank. Esto refuerza la importancia del grado como característica topológica de la red.

Además, una pregunta importante es si la destrucción por datos intrínsecos a artículos puede ser tan letal como las medidas topológicas (o inclusive si será más efectiva que el azar). Esto se puede ver en la figura 13.

Podemos ver que no son tan efecivos como el grado, y comparando con la figura 12 los vemos menos letales que las centralidades de autovalor y PageRank también (aunque no por tanto). Aunque cumplen cierto rol en la integridad topológica, no son tan importantes como las centralidades topológicas.

8. Conclusiones

A través de este trabajo, pudimos observar distintos efectos de la relación entre la topología y el uso de Wikipedia de parte de sus lectores.

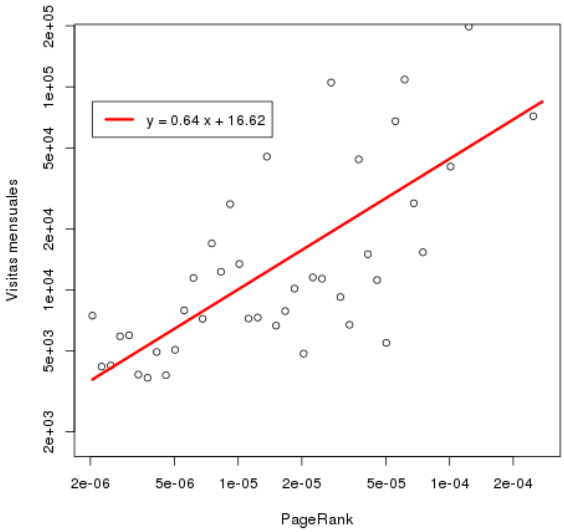


Figura 11: Correlación entre el PageRank y la cantidad mensual de visitantes. Se puede ver que a medida que crece la centralidad PageRank, la tendencia se dispersa. El PageRank es calculado con coeficiente de amortiguamiento $d = 0{,}66$.

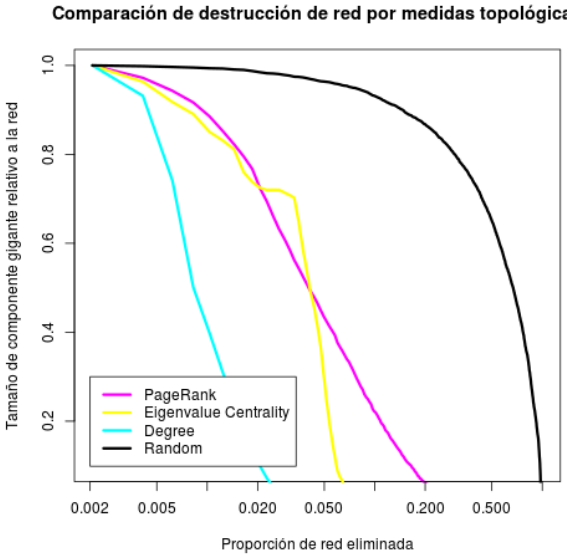


Figura 12: Destrucción progresiva de la red estudiada en órden de mayor centralidad por grado, autovalor, y PageRank, así como eligiendo de forma aleatoria a los nodos. Se puede ver que la destrucción por grado reduce a la componente gigante apreciablemente más rápido que las alternativas.

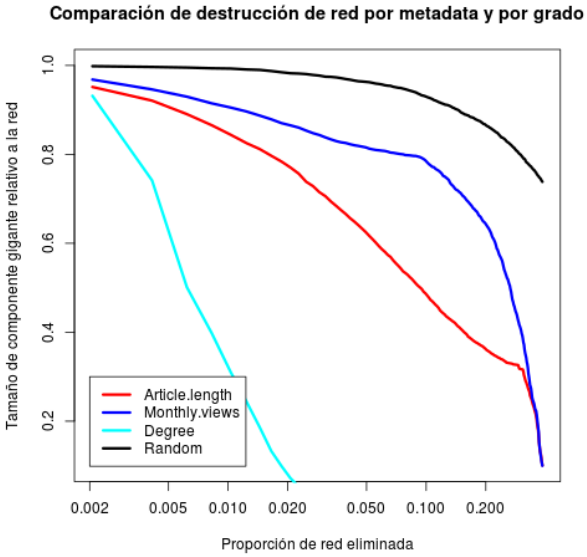


Figura 13: Destrucción progresiva de la red estudiada en órden de mayor largo de página y cantidad de visitantes, en comparación con la destrucción por grado y al azar (los casos más y menos eficientes conocidos). Se ve que son más efectivas que el azar, pero menos que el grado, y que de las dos el largo del artículo reduce la red más rápidamente.

El scrapper propuesto para relevar la red podría ser mejorado considerando los vecinos de las mismas categorías a segundo orden, de forma de asegurarse que uno no se pierde algo por una distancia muy pequeña.

Observamos que si bien las tendencias son ruidosas, hay evidencia de las caminatas que los usuarios realizan sobre la red, principalemente a través de la relación entre los vistos de una página y los de las páginas que limitan con ella, así como por la correlación entre las visitas de las redes y la medida de centralidad PageRank.

Pudimos ver que en términos de lo que definimos como tema (las categorías), atravezar un hipervínculo nos puede llevar a temáticas totalmente nuevas. Esto muestra la fuerte interrelación temática que hay en la Wikipedia, una cualidad excelente a nivel enciclopédico.

Finalmente, pudimos apreciar la gran importancia que tiene el grado total a nivel topológico, mucho mayor que otras medidas de centralidad. Además, el grado saliente tiene fuertes correlaciones con características internas de los nodos, lo que muestra una componente fuerte de decisión interna acerca de donde se conectará un artículo y que rol topológico tendrá, dandole gran importancia al grado como medida de centralidad.

Multiples posibilidades de análisis posteriores surgen, debido al gran número de variables involucradas en el tema. Una propuesta podría ser analizar la red que forman las categorías entre sí, y comparar las redes de categorías obtenidas para diferentes tópicos. Profundizar el análisis sobre la relación entre la topología y los vistos mensuales, en términos de caminatas aleatorias, también sería

interesante. Por último, comparar la similaridad entre dos páginas conectadas basandose en las categorías podría ser comparado con un análisis de tópicos de texto, usando la información de la página. Quizá a nivel categorías las páginas sean muy distintas, pero en términos de texto no tanto.

Referencias

[1] Zotenko, E. et al. *Why Do Hubs in the Yeast Protein Interaction Network Tend To Be Essential: Reexamining the Connection between the Network Topology and Essentiality*, PLOS Computational Biology. Agosto de 2008.