



CENTRO DE
INVESTIGACIÓN EN
COMPLEJIDAD SOCIAL

Prueba

Teoría de Redes

Aníbal Olivera Morales

Fecha: 13 de Enero de 2024

1. Scale-Free Networks are Rare

El estudio del cual se basa este desafío pretende poner en entredicho la afamada ley de potencia (*power law*) que supuestamente siguen una gran variedad de redes empíricas, al hacerse con los datos de cerca de 1000 estudios publicados en investigaciones anteriores y ofrecer una manera de discriminar si una red tiene evidencia suficiente como para decir que la misma es una red libre de escala o no:

1. Sin evidencia, **Not Scale Free**: Hay distribuciones alternativas que se ajustan mejor que la *power law* para al menos el 50 % de los grafos.
2. Evidencia **Super Débil**: No hay una distribución alternativa mejor que la *power law* para el 50 % de los grafos.
3. Evidencia **Más Débil**: La distribución *power law* no puede ser rechazada ($p \geq 0.1$) para el 50 % de los grafos.
4. Evidencia **Débil**: La distribución *power law* no puede ser rechazada para al menos 50 nodos.
5. Evidencia **Fuerte**: Además, que $2 < \hat{\alpha} < 3$ para al menos el 50 % de los grafos.
6. Evidencia **Más Fuerte**: Que $2 < \hat{\alpha} < 3$ para al menos el 90 % de los grafos, y que no haya una distribución mejor que la *power law* para el 95 % de los grafos.

Antes de concentrar la atención a la supuesta arbitrariedad de estas categorías, veamos los principales resultados del estudio.

1.1. Resultados

El primer resultado es mostrado en la Imagen 1 e indica que casi la mitad de los estudios (49 %) no tienen evidencia para proponer la *power law* como el mejor modelo para explicar la distribución de los nodos. Un 46 % de ellos sí son explicados mejor por una *power law* a una cierta batería establecida de opciones, pero el nivel de significancia es muy bajo como para no inspeccionar otras distribuciones que pueden ajustar mejor. Más allá del 48 % que cae en las categorías Débil o Más Débil, es preocupante el bajo número de investigaciones que están agrupadas en Fuerte o Más Fuerte.

Además, resulta llamativo el descenso escalonado de las representaciones de cada categoría (de aproximadamente 10 puntos entre categorías) que resulta en una aparente coincidencia que puede levantar sospechas o no de la búsqueda de categorías que entregaran una frecuencia equidistante entre ellas. Este descenso escalonado en la robustez de la evidencia corre de acuerdo a los datos agregados de los cerca de 1000 estudios analizados.

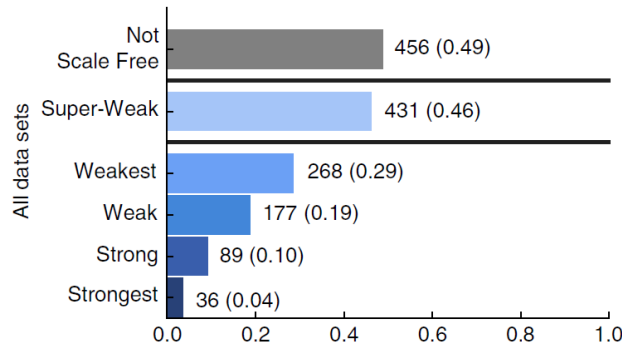


Imagen 1

Frecuencia de estudios ubicados en cada categoría. Las frecuencias caen equi-espaciadamente en cada una de estas categorías, lo que puede sugerir definiciones levantadas con tal propósito.

Sin embargo, los datos utilizados para este meta-análisis no están equilibrados según el dominio al cual la investigación original. De ellos, compararemos constantemente los dominios **Biológico** (53 % de los datos), **Social** (16 %) y **Tecnológico** (22 %). Salta a la vista además que la cantidad de datos provenientes de la biología más que triplica a aquellos provenientes de estudios aplicados a problemas sociales.

Domain	Number (Prop.)	Multiplex	Bipartite	Multigraph	Weighted	Directed	Simple
Bio.	495 (0.53)	273	41	378	29	37	39
Info.	16 (0.02)	0	0	4	0	5	7
Social	147 (0.16)	7	0	6	8	0	129
Tech.	203 (0.22)	122	0	3	1	195	5
Trans.	67 (0.07)	48	0	65	3	2	0
Total	928 (1.00)	450	41	456	41	239	180

Imagen 2

*Distribución de grafos por dominios y sus características. De entre los dominios **Biológico**, **Social** y **Tecnológico**, el Social es el menos representado, y la gran mayoría son modelados con grafos simples, a diferencia de los otros dos dominios principales.*

Notar además que la abrumadora mayoría de los estudios en el dominio **Social** son modelados utilizando grafos simples, a diferencia de los otros dos dominios, donde los grafos simples son los menos frecuentes, tiendo que pasar por transformaciones matemáticas para tener un indicador comparable de distribuciones de grados. Por lo tanto, la comparación entre las dominios pierde un fundamentos metodológico sólido, ya que involucra elementos heterogéneos, que poseen características inherentes distintas entre dominios, lo cual impide una evaluación objetiva y significativa.

Naturalmente, estas proporciones distintas de datos entre los dominios de interés exige ver los resultados de las frecuencias de cada categoría separado por cada dominio. La próxima Imagen 3 entrega esta separación.

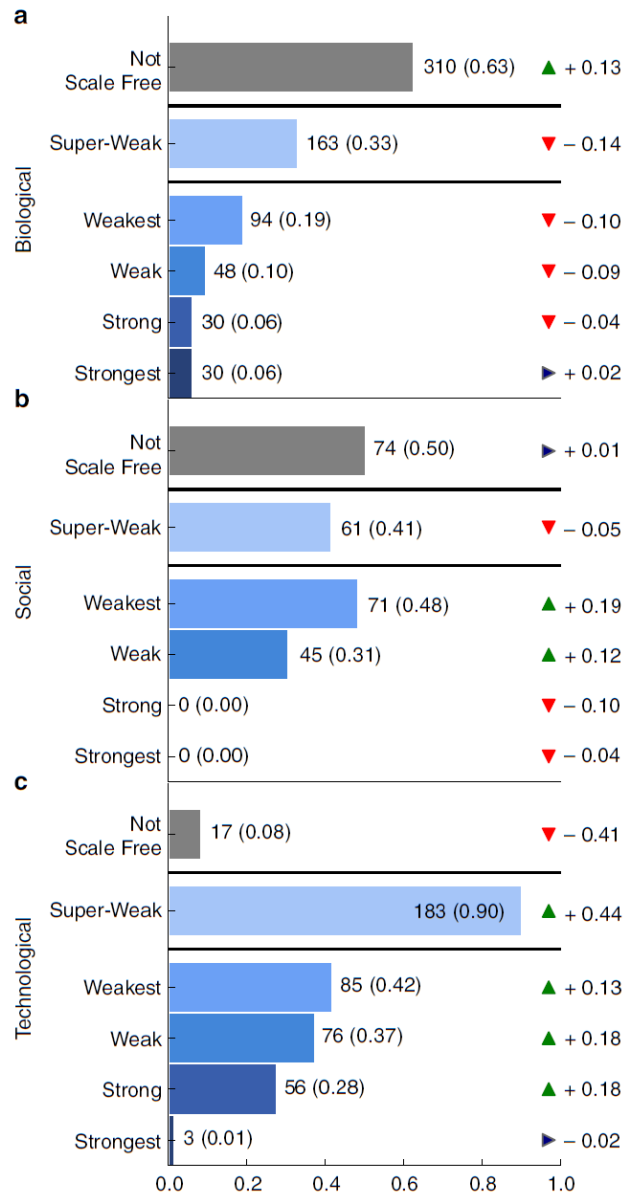


Imagen 3

Una vez que se separan solo entre las redes levantadas con datos perteneciente al dominio **Social**, los resultados se aglomeran más en las categorías **Débil** y **Más Débil**. Notar que el dominio en los que la evidencia **Más Fuerte** contribuye en mayor proporción es el **Biológico**.

De estas gráficas se desprende que los resultados de las discriminaciones varían tremendamente entre dominios. Algunos ven aumentados su proporción de Más Fuerte (Biología), mientras que otros datos caen con mayor frecuencia a mostrar evidencias Super Débiles (Tecnología). Impresionantemente, ninguna de las 147 investigaciones realizadas con miras a un problema Social mostró un exponente $2 < \hat{\alpha} < 3$ para el 50% de los grafos, pero sin embargo, en el mismo dominio, hay más grafos para

los cuales no se puede rechazar una *power law*.

Estos resultados parecen indicar que, para el dominio Social, es poco probable que la distribución de grados sea exactamente una *power law* (Fuerte y Más Fuerte en 0 %), pero que sin embargo la distribución que sigue debe ser aproximada a la *power law* (may más que no se pueden rechazar como *power law*). En la próxima Imagen 4 se indagan distribuciones alternativas: exponencial, log-normal, Weibull y una *power law* con *cutoff*:

Table 1 Comparison of scale-free and alternative distributions				
Alternative	$p(x) \propto f(x)$	Test outcome		
		M_{PL}	Inconclusive	M_{Alt}
Exponential	$e^{-\lambda x}$	33%	26%	41%
Log-normal	$\frac{1}{x} e^{-\frac{(\log x - \mu)^2}{2\sigma^2}}$	12%	40%	48%
Weibull	$e^{-\left(\frac{x}{b}\right)^a}$	33%	20%	47%
Power law with cutoff	$x^{-\alpha} e^{-\lambda x}$	-	44%	56%

The percentage of network data sets that favor the power-law model M_{PL} , alternative model M_{Alt} , or neither, under a likelihood-ratio test, along with the form of the alternative distribution $f(x)$

Imagen 4

En todas las comparaciones los modelos alternativos se ajustan mejor a una mayor proporción de las redes. Sobresale entre ellas la *power law* con *cutoff* (límite superior en la distribución de grados de la red) lo que significa que la *power law* que describe la distribución de grados puede dejar de ser válida para nodos con grados muy altos.

De las 4 alternativas presentadas para ser contrastadas ante la distribución *power law*, todas ellas mostraron un don de mayor ajuste a los datos agregados en comparación a la *power law* ($M_{ALT} > M_{PL}$ en todos los casos). Sin embargo, una de ellas llama la atención: la mayoría de los set de datos mostró un mayor ajuste en la distribución *power law* con *cutoff* en comparación a la estándar ($M_{ALT} = 56\%$), y en el resto de los datos el test se mostró no concluyente, pero en ningún caso la *power law* performó mejor que aquella con *cutoff*.

Lamentablemente, no reportan estos análisis de comparaciones separados por cada dominio específico. De esta manera no podemos saber si para el dominio Social el M_{alt} es mayor o no al promedio 56 % de todos los datos.

De esta forma, el artículo solo siembra las semillas de **1)** dudas en la pertinencia de la distribución *power law* para representar la distribución de grados en datos reales, que **2)** esta preocupación se acrecienta en el dominio Social, y que **3)** puede ser una buena idea probar distribuciones similares en el espectro inicial de los grados, pero diferentes funcionalmente para número de grados altos.

Más específicamente, el punto **2)** indica que si bien no hay ningún set de datos del dominio Social que apoye con robustez la *power law*, tampoco se puede rechazar la posibilidad que sí lo sea, o que sea algo similar a ella. Por desgracia, los resultados reportados por el artículo tampoco entregan mayores detalles de estas distribuciones de frecuencias de las categorías de acuerdo al número de

nodos de una red del dominio Social, ni tampoco de la naturaleza específica de estas. Así como es posible que se trate de una red infectados en una pandemia global (millones de nodos), también se puede tratar de la red de amistades en un colegio (pocos nodos en total).

La siguiente sección presenta una propuesta para abordar este problema evidenciado en el dominio Social, indicando los supuestos tomados, y mostrando la forma funcional de la distribución de probabilidad sugerido para tales casos, comparándolo numéricamente con la alicaída *power law*.

2. Consideraciones previas

En líneas de presentar una propuesta para analizar por qué el dominio Social pareciera ajustarse en peor medida que aquellos datos provenientes de los dominios Biológicos o Tecnológicos, resulta necesario averiguar si el número de nodos de las redes varía de dominio a dominio. La siguiente imagen describe la cantidad de nodos de todas las redes, y su parámetro $\hat{\alpha}$ estimado.

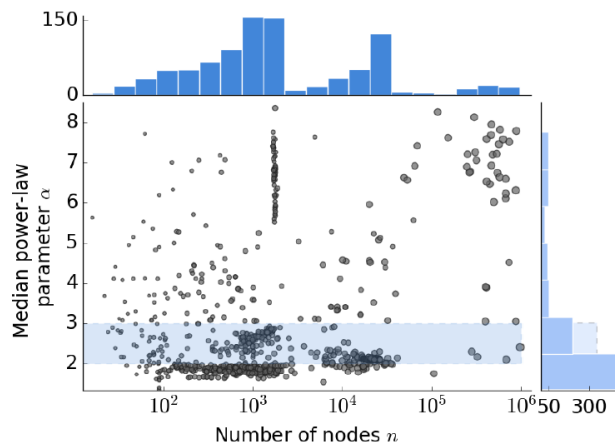


Imagen 5

Parámetro $\hat{\alpha}$ medio versus tamaño de red n . La banda horizontal destaca el rango canónico [2,3] e ilustra la amplia diversidad de los parámetros estimados de ley de potencia en diferentes trabajos empíricos.

Por desgracia, estos datos siguen siendo agregados, y no podemos ver si el número de nodos varía sustancialmente entre los dominios bajo consideración. Afortunadamente, sí es posible visualizar esta separación en el sitio web donde están alojados los datos del meta-análisis: <https://icon.colorado.edu/#!/networks>.

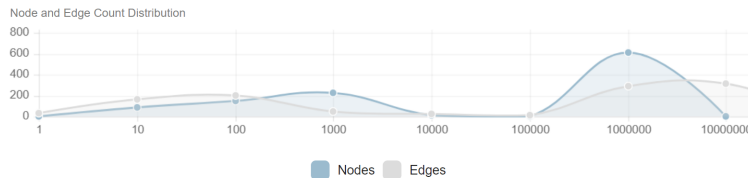


Imagen 6

Distribución de nodos (grises) y enlaces (azules) para aquellos estudios pertenecientes al dominio **Biológico**.

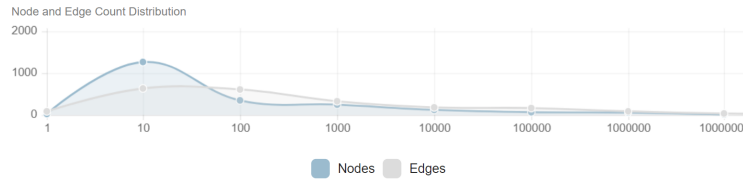


Imagen 7

*Distribución de nodos (grises) y enlaces (azules) para aquellos estudios pertenecientes al dominio **Social**.*

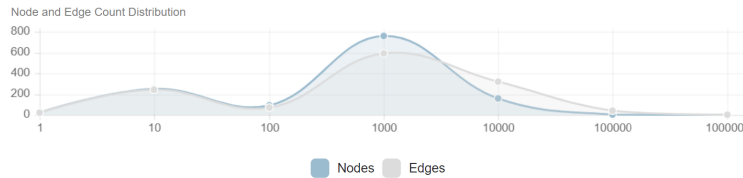


Imagen 8

*Distribución de nodos (grises) y enlaces (azules) para aquellos estudios pertenecientes al dominio **Tecnológico**.*

Como se puede apreciar, las distribuciones de nodos entre cada uno de los dominios varía marcadamente. Podemos rescatar que para redes biológicas, el número de nodos está centrado en 10^6 , para redes sociales cerca de 10, y para redes tecnológicas se halla una conglomeración en los 10^3 nodos.

El hecho que para redes Biológicas hayan más estudios del orden de millones de nodos, y que a su vez sea aquel dominio que más evidencia Más Fuerte entrega, sugiere que una distribución *power law* es adecuada para ellas. Por otro lado, la cantidad nula de experimentos que se encuentran entre las evidencias Fuertes y Más Fuertes en el dominio Social, pero que a su vez aumente la frecuencia de aparición en las evidencias Débiles o Más Débiles, puede sugerir que: **1)** o se necesitan más sensibilidad en el número de nodos para lograr un ajuste a una *power law* continua, o **2)** la distribución es similar a una *power law*, si bien no idéntica, para grandes números de nodos.

La propuesta abarca la sugerencia **2)**, pues la primera no es susceptible a modificarse con trabajo posterior:

Si tenemos en presente que la mayoría de las redes del dominio Biológico son modelos de interacciones de proteínas u otras moléculas de tamaños similares, y que en general estas interacciones se dan en redes de grandes cantidades de nodos, entonces podemos aceptar que una parte significativa de lazos totales del dominio Biológico están mediados en última instancia por la fuerza electromagnética. Esta es la única fuerza fundamental relevante en esa escala, que a su vez esta mediada por fotones emitidos espontáneamente en direcciones más o menos arbitrarias (dependiendo de la geometría del espacio circundante), cuya cantidad de fotones únicos emitidos es varios órdenes de magnitud superior a la cantidad máxima de lazos mostrados por las redes de mayores lazos en el este dominio (aproximadamente 10^7 lazos), entonces podemos deducir que para la basta mayoría estudios de este dominio los enlaces no tienen costo alguno para ninguna de los nodos bajo interacción (pese a que la tasa de emisión de fotos sí puede cambiar a niveles atómicos dependiendo de si el espacio circundante está modificado o no por la presencia de otro átomo, esta alteración es no

significativa a niveles moleculares de los tamaños involucrados en las interacciones entre proteínas [1]).

Esta conceptualización del problema requiere añadir un techo presupuestario para estas interacciones en el caso que sean costosas. Para la interacción materia-materia mediada por fuerzas electromagnéticas (o gravitacionales, aunque en estas escalas no es relevante), es equivalente plantear el problema como interacciones de costo nulo o presupuestos de interacción arbitrariamente altos.

Por el lado del dominio Social podemos encontrar interacciones costosas y no costosas. Diremos que una interacción no costosa puede ser la cantidad de seguidores en una cuenta de red social, ya que en términos generales no hay un límite superior de seguidores, y quien sigue a otro no gasta una cantidad limitada de cuentas a los que tiene permitido seguir. Diremos que una interacción costosa es la cantidad de personas reales con las que una persona puede encontrarse y establecer relaciones de cualquier naturaleza. El techo presupuestario lo demarca la cantidad finita de horas del día disponible para establecer estos vínculos. Por lo tanto, no es posible armar una red social de encuentros físicos con hubs arbitrariamente altos. Por último, podemos decir que hay interacciones costosas o no costosas dependiendo de las características específicas que aplican a un estudio en particular. Tal puede ser el caso de la acumulación de riqueza: en una nación sin impuestos graduales de acuerdo al nivel de ingresos, no hay cotas superiores de acumulación (tener dinero no cuesta dinero). Sin embargo, en sociedades donde esta acumulación sí implica impuestos graduales, tener dinero sí cuesta dinero, y las penalizaciones geométricas pueden llevar a un límite finito máximo.

Con fines pedagógicos, centrémonos en el ejemplo de una red social con personas reales. Supongamos que hay una red establecida con n nodos. Como la literatura indica, deberíamos encontrar hubs. Supongamos que, por el presupuesto temporal finito, una persona solo es posible que esté conectada a otras N personas, y que la red con n nodos tiene dos hubs, uno con $N - 1$ enlaces y otra con $N - q$, donde $q \gg 1$. Si seguimos el modelo de *preferential attachment* planteado en primer lugar por Price y luego redescubierto de Barabási, la probabilidad que un nuevo nodo, etiquetado como el nodo $n + 1$, cree un enlace con aquel que está conectada a N personas es linealmente más alta que con aquella que tiene $N - q$ nodos. Pero claramente en este caso el modelo de Barabási fallará en replicar la red social bajo estudio.

3. Propuesta

Este problema se puede atacar de una manera sencilla: el *preferential attachment* no puede ser lineal con el número de lazos de un nodo en la iteración anterior. Debe existir un parámetro, que llamaremos γ , que modula esta preferencia de tal modo que hay una tasa marginal decreciente en la probabilidad de acoplamiento.

Para formalizar la propuesta, digamos que el *preferential attachment* original se puede escribir como

$$a_k = k, \quad (1)$$

donde a_k es el kernel¹ la probabilidad de crear un lazo con un nodo de grado k . Una preferencia que

¹El kernel de una probabilidad busca simplemente mantener la forma funcional de la distribución de probabilidad,

muestra un patrón similar pero que agrega una tasa marginal decreciente se puede escribir como

$$a_k = k^\gamma \quad (2)$$

donde $\gamma < 1$, de manera que se obtiene

$$\frac{d^2 a_k}{dk^2} = (\gamma - 1)\gamma k^{\gamma-2} < 0 \quad \forall x \iff \gamma < 1. \quad (3)$$

Además, sea $p_k(n)$ la fracción de nodos con grado k cuando la red tiene n nodos. Nuestro objetivo será armar una distribución p_k en su límite asintótico $n \rightarrow \infty$. La siguiente formulación seguirá aquella presentada por Newman [2].

...

Supongamos que cada nodo nuevo de la red se acopla a otros c nodos existentes en la red. La probabilidad que un nuevo nodo se acople a un nodo particular con k lazos es

$$\frac{a_k}{\sum_i a_{k_i}}, \quad (4)$$

por lo tanto, si un nodo i tiene k lazos, el número esperado de nuevas lazos para ese nodo i al sumarse este nuevo nodo es

$$c \cdot \frac{a_k}{\sum_i a_{k_i}} \quad (5)$$

Como hay $np_k(n)$ nodos con k lazos en la red de n nodos, entonces el número esperado de nodos de grado k que reciben un nuevo enlace cuando se agrega un solo nodo nuevo a la red es:

$$np_k(n) \cdot c \cdot \frac{a_k}{\sum_i a_{k_i}}. \quad (6)$$

Ahora, si definimos una cantidad μ que represente al valor promedio del kernel de acoplamiento sobre todos los nodos,

$$\mu(n) = \frac{1}{n} \sum_{i=1}^n a_{k_i} \quad (7)$$

entonces la expresión 6 se puede escribir como

$$\frac{c}{\mu(n)} a_k p_k(n).$$

De esta manera, podemos deducir una Ecuación Maestra² para la evolución de la red al permitir la llegada de nuevos nodos que se acoplarán a otros c nodos. La cantidad de nodos con k lazos una

sin agregar una normalización.

²Las Ecuaciones Maestras capturan la dinámica elemental de un proceso, considerando las pérdidas del sistema en caso de sistemas abiertos.

vez que la red pasa de tener n nodos a $n + 1$ nodos viene dado por

$$(n + 1)p_k(n + 1) = np_k(n) + \frac{c}{\mu(n)}a_{k-1}p_{k-1}(n) - \frac{c}{\mu(n)}a_kp_k(n). \quad (8)$$

El primer término de la izquierda da cuenta del número de nodos que tenían k lazos en la iteración anterior, el segundo término de la cantidad de nuevos nodos con k lazos, y el último de los nodos que ya no tienen k lazos sino que más. En orden de obtener el límite asintótico de la probabilidad de encontrar nodos con k lazos, podemos reordenar la Ec. 8 para obtener

$$p_k(n + 1) = n[p_k(n) - p_k(n + 1)] + \frac{c}{\mu(n)}[a_{k-1}p_{k-1}(n) - a_kp_k(n)], \quad (9)$$

que, en el límite $n \rightarrow +\infty$, y definiendo $p_k = p_k(+\infty)$ y $\mu = \mu(+\infty)$, queda como

$$p_k = \frac{c}{\mu}[a_{k-1}p_{k-1} - a_kp_k]. \quad (10)$$

Ahora, como la probabilidad p_k está tanto a la izquierda como a la derecha de la ecuación, se puede reordenar para obtener

$$p_k = \frac{a_{k-1}}{a_k + \mu/c}p_{k-1}. \quad (11)$$

Esta expresión es una forma iterativa de presentar p_k . Es decir, de la misma forma podemos escribir

$$p_{k-1} = \frac{a_{k-2}}{a_{k-1} + \mu/c}p_{k-2}, \quad p_{k-2} = \dots \quad (12)$$

Por lo tanto, la función generatriz de esta distribución se puede expresar como

$$p_k = \frac{\mu}{ca_k} \frac{a_k \dots a_c}{(a_k + \mu/c) \dots (a_c + \mu/c)} \quad (13)$$

$$= \frac{\mu}{ca_k} \frac{a_k \dots a_c}{a_k (1 + \mu/a_k c) \dots a_c (1 + \mu/a_c c)} \quad (14)$$

$$= \frac{\mu}{ca_k} \prod_{r=c}^k \left[1 + \frac{\mu}{ca_r} \right]^{-1}, \quad (15)$$

donde asumimos que en la configuración inicial no hay nodos una cantidad de lazos menor a c . La pregunta ahora a qué valor converge esa multiplicatoria. La tarea de hallar la forma funcional de μ se ve complicada, pero podemos obviar esto y encontrar aproximaciones que en un buen caso serán suficientes para ilustrar el impacto de considerar *preferential attachment* no lineales. Por ahora, simplemente reemplacemos $a_k = k^\gamma$ para obtener

$$p_k = \frac{\mu}{ck^\gamma} \prod_{r=c}^k \left[1 + \frac{\mu}{cr^\gamma} \right]^{-1} = \frac{\mu}{ck^\gamma} \exp \left[- \sum_{r=c}^k \ln \left(1 + \frac{\mu}{cr^\gamma} \right) \right]. \quad (16)$$

La última igualdad es sencilla de derivar siguiendo propiedades de logaritmos. Ya estamos cerca de hallar la forma funcional final de la distribución p_k . Todo lo que nos resta por considerar es una

aproximación del término entre paréntesis. Para los casos en que $1/2 < \gamma < 1$ (cuya restricción no es potente y admite una significativa disminución del valor del kernel no lineal $a_k = k^\gamma$ en relación al lineal de Barabási $a_k = k$, como veremos en un gráfico dentro de poco), es posible encontrar la aproximación asintótica [2],

$$\sum_{r=c}^k \ln \left(1 + \frac{\mu}{cr\gamma} \right) \simeq A + \frac{\mu k^{1-\gamma}}{c(1-\gamma)}, \quad (17)$$

donde A es una constante que no es relevante a la hora de comparar las formas funcionales de p_k , resultando por fin en una aproximación de p_k en el límite $n \rightarrow +\infty$:

$$p_k \sim k^{-\gamma} \exp \left(-\frac{\mu k^{1-\gamma}}{c(1-\gamma)} \right). \quad (18)$$

Esta expresión nos dice que la simple modificación de considerar una tasa marginal decreciente en la probabilidad de acoplamiento (siguiendo una lógica de presupuestos no arbitrarios para permitir seguir agregando lazos indefinidamente) nos condujo a tener una distribución con forma de *power law* con un *cutoff* que depende de una combinación de los parámetros $\{\mu, k, \gamma, c\}$. Este resultado concuerda con la sugerencia de la tabla de la Imagen 4.

Para ver ejemplo de cómo se diferencial el *differential attachment* lineal con este modelo no-lineal, veamos primeramente cómo se comporta un valor de $\gamma = 0.8$ en el rango de los millones de lazos.

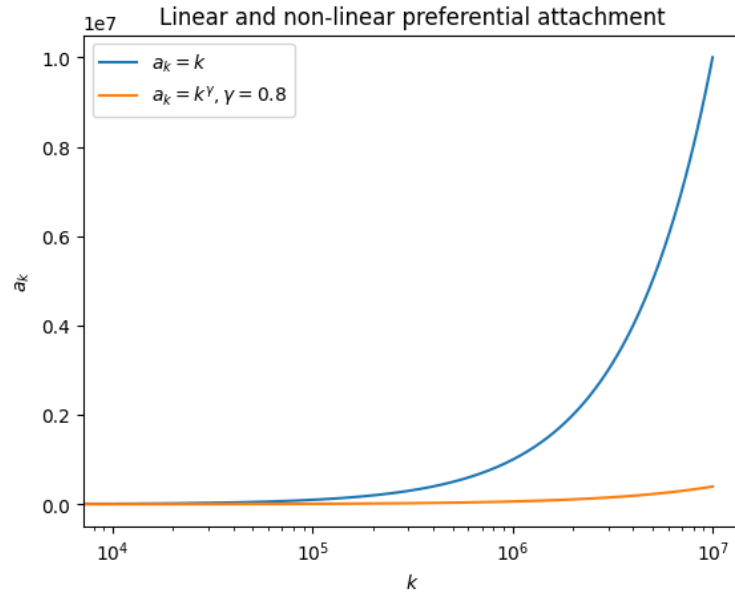


Imagen 9

Comparación de los kernel a_k . Para $a_k = k$ se está asumiendo que la probabilidad que un nuevo nodo forme un lazo con un nodo de grado k es lineal con el grado. Para $a_k = k^\gamma$ hay un efecto marginal decreciente: el atractivo de los nodos con k lazos aumenta significativamente más lento.

Supongamos ahora que cada nuevo nodo creará un lazo con exactamente otros tres nodos, $c = 3$, de

tal manera que solo queda fijar el valor de μ . Para ello, simplemente se buscó aquel valor numérico que permitiera una clara comparación entre las *power law* y esta modificación. Luego ajustar su valor mediante algoritmos de minimización de cuadrados, se fijó $\mu = 7/4$. La siguiente gráfica permite visualizar el efecto de un *preferential attachment* con efecto marginal decreciente:

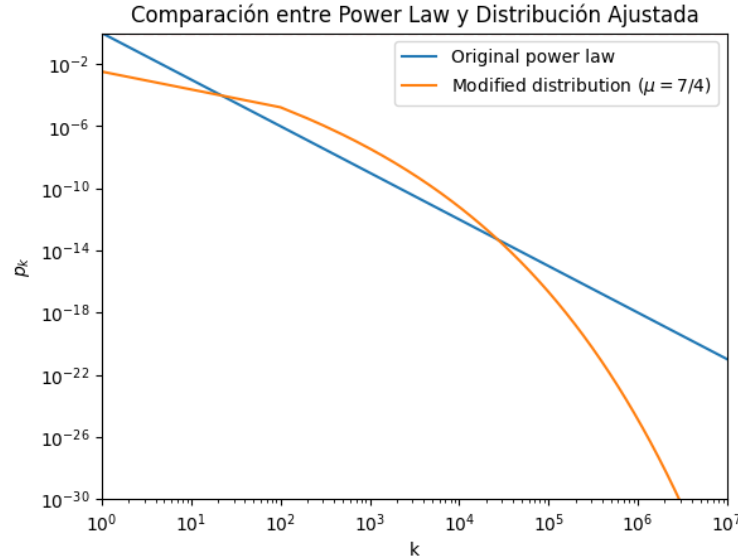


Imagen 10

Probabilidad p_k de encontrar un nodo con k lazos en una red no dirigida para una cantidad de nodos $n \rightarrow +\infty$. Debido a la validez exclusivamente para valores altos de k (en comparación con n), la discrepancia para k bajos no debe tomarse como una preocupación. Para este gráfico, se utilizó $\mu = 7/4$, y como referencia, se escogió una distribución *power law* con parámetro $\alpha = 3$, de tal manera que $p_k = k^{-\alpha}$.

La aproximación de $n \rightarrow +\infty$ exige que nos centremos en la diferencia funcional de la distribución para valores altos de k . En tal caso, el efecto es el buscado al requerir una disminución de la tendencia a preferir el acoplamiento a nodos con alta centralidad de grado de una forma lineal con k . De esta forma, la preferencia por hubs se estanca en valores aproximadamente iguales para un cierto valor de k , como muestra la Imagen 9, lo que deviene en una interrupción de la tendencia de la *power law*, como muestra la Imagen 10. En términos simples, esto plasma una saturación de los nodos a partir de los 10^5 lazos: más allá de eso, el techo presupuestario para lazos se agota y comienza una disminución de la tendencia en aceptar nuevos lazos.

Este framework permite explicar en parte el rendimiento pobre del dominio Social en las categorías propuestas por el artículo, solo en aquellos estudios que en su modelación utilizan del orden de 10^5 nodos. Es importante recalcar que este procedimiento arroja formas funcionales similares a la *power law* estándar, lo que va en línea con la frecuencia de los datos en las categorías: no son exactamente *power law*, pero no se puede rechazar que sí lo sean; al menos algo funcionalmente similar en gran parte de su espectro.

Referencias

1. S. Y. Buhmann, Dispersion Forces II (Springer-Verlag, Berlin, Heidelberg, 2012).
2. M. Newman, Networks, Second Edition (Oxford University Press, 2018).