

# Informe Small World

January 12, 2025

## Trabajo entregado el 01-10-2024

El objetivo del presente trabajo es estudiar diferentes datasets disponibles en el paquete torch\_geometric de Python en búsqueda de ver cómo estos cumplen la **hipótesis de mundo pequeño (Small World)**, que consiste básicamente en decir que en los grafos que representan relaciones sociales o de comunicación, la distancia promedio entre dos nodos crece a una velocidad mucho menor (en la definición de métricas le daremos un significado numérico a esto) que el tamaño (cantidad de nodos) del grafo.

Afortunadamente, la librería antes mencionada nos proporciona una amplia variedad de grafos que podemos utilizar para estudiar esta hipótesis.

## Contents

<b>1 Definición de métricas</b>	<b>1</b>
<b>2 Estudio de distintos datasets</b>	<b>2</b>
2.1 Karate Club . . . . .	2
2.2 Actor . . . . .	2
2.3 Coautor . . . . .	2
2.4 Twitch . . . . .	3
2.5 FacebookPagePage . . . . .	4
2.6 GemsecDeezer . . . . .	4
<b>3 Análisis agregado</b>	<b>4</b>
<b>4 Persistencia de la propiedad de Mundo Pequeño</b>	<b>9</b>
<b>5 Comportamiento conjunto de las variables</b>	<b>12</b>
<b>6 Conclusiones</b>	<b>21</b>

## 1 Definición de métricas

Lo primero que debemos hacer para explorar la teoría de mundo pequeño es definir métricas medibles que plasmen la idea.

Vamos a explorar 3 métricas para los diferentes grafos:

- Diámetro: La máxima distancia entre dos nodos.

- Distancia promedio: La distancia promedio entre todos los posibles pares de nodos.
- Coeficiente de agrupamiento: Es la cantidad de triángulos (un clique de 3 nodos) dividida la cantidad de conjuntos {a,b,c} donde (a,b) y (b,c) son aristas del grafo.

La hipótesis de mundo pequeño dice que, para un grafo social de  $N$  nodos:

- \* Las dos primeras métricas deberían comportarse como  $O(\log(N))$
- \* La tercera métrica debe ser mayor o igual a  $\frac{M}{\binom{N}{2}}$ .

Otra forma de verlo es que estas 3 métricas se deben parecer a las de un grafo aleatorio de Erdos-Renyi con la misma cantidad de nodos y aristas, comportándose las 2 primeras como  $O(\log(N))$  y la tercera como  $O(\frac{M}{\binom{N}{2}})$ .

Utilizaremos la librería igraph por motivos de performance.

## 2 Estudio de distintos datasets

### 2.1 Karate Club

Este es un dataset bastante pequeño, utilizado como primer ejemplo en muchos trabajos sobre grafos, especialmente en el área de aprendizaje automático.

Los enlaces refieren a los miembros de un club de karate que interactúan (son amigos) fuera del club.

El dataset Karate Club tiene 34 nodos

```
Es conexo
Su diámetro es 5
La distancia promedio entre nodos es 2.408199643493761
Su coeficiente de agrupamiento es 0.2556818181818182
```

### 2.2 Actor

Este es un curioso conjunto de datos sobre actores donde las aristas indican que se los nombre en un mismo artículo de Wikipedia.

El dataset Actor tiene 7600 nodos

```
Es conexo
Su diámetro es 12
La distancia promedio entre nodos es 4.11027998836412
Su coeficiente de agrupamiento es 0.015701288192099157
```

### 2.3 Coautor

Este es un juego de dos datasets que contienen información sobre trabajos científicos de Ciencias de la Computación y de Física, donde los nodos son autores y las aristas indican que ha coproducido un trabajo científico.

El dataset Coautores de Ciencia de la Computación tiene 18333 nodos

```
Es conexo
Su diámetro es 24
```

La distancia promedio entre nodos es 5.427693716383878  
Su coeficiente de agrupamiento es 0.18255559535165766

El dataset Coautores de Física tiene 34493 nodos  
Es conexo  
Su diámetro es 17  
La distancia promedio entre nodos es 5.163814898581291  
Su coeficiente de agrupamiento es 0.18742917153412708

## 2.4 Twitch

Twitch nos ofrece un juego de conjuntos de datos donde los nodos son usuarios de la plataforma y las aristas representan seguimiento mutuo. Es interesante que está discriminado por el idioma de los streamers. Por eso hay 6 datasets, para alemán, inglés, español, francés, portugués y ruso.

El dataset Twitch DE tiene 9498 nodos  
Es conexo  
Su diámetro es 7  
La distancia promedio entre nodos es 2.7215711057961074  
Su coeficiente de agrupamiento es 0.0464708891573653

El dataset Twitch EN tiene 7126 nodos  
Es conexo  
Su diámetro es 10  
La distancia promedio entre nodos es 3.6776157289097005  
Su coeficiente de agrupamiento es 0.04243324947984254

El dataset Twitch ES tiene 4648 nodos  
Es conexo  
Su diámetro es 9  
La distancia promedio entre nodos es 2.883191439556992  
Su coeficiente de agrupamiento es 0.0842348748307571

El dataset Twitch FR tiene 6551 nodos  
No es conexo, y respecto de la componente más grande:  
Su diámetro es 7  
La distancia promedio entre nodos es 2.6809907139571783  
Su coeficiente de agrupamiento es 0.05412827315082585

El dataset Twitch PT tiene 1912 nodos  
Es conexo  
Su diámetro es 7  
La distancia promedio entre nodos es 2.5323791570055767  
Su coeficiente de agrupamiento es 0.1309809619261169

El dataset Twitch RU tiene 4385 nodos  
Es conexo

```

Su diámetro es 9
La distancia promedio entre nodos es 3.0210946408209804
Su coeficiente de agrupamiento es 0.04864801926174216

```

Notar que todos los datasets son conexos salvo el de francés.

## 2.5 FacebookPagePage

Este es un dataset donde los nodos representan páginas verificadas de Facebook y las aristas representan likes mutuos entre estas páginas.

El dataset FacebookPagePage tiene 22470 nodos

```

Es conexo
Su diámetro es 15
La distancia promedio entre nodos es 4.973703570580348
Su coeficiente de agrupamiento es 0.23232143653859755

```

## 2.6 GemsecDeezer

GemsecDeezer es una red social y, al igual que Twitich, nos ofrece un dataset con los usuarios discriminados por idioma que hablan y las relaciones de amistad entre hablantes de un mismo idioma

El dataset GemsecDeezer HU tiene 47538 nodos

```

Es conexo
Su diámetro es 14
La distancia promedio entre nodos es 5.3409423368686895
Su coeficiente de agrupamiento es 0.09292401767906112

```

El dataset GemsecDeezer RO tiene 41773 nodos

```

Es conexo
Su diámetro es 19
La distancia promedio entre nodos es 6.348893165004872
Su coeficiente de agrupamiento es 0.0752667042317545

```

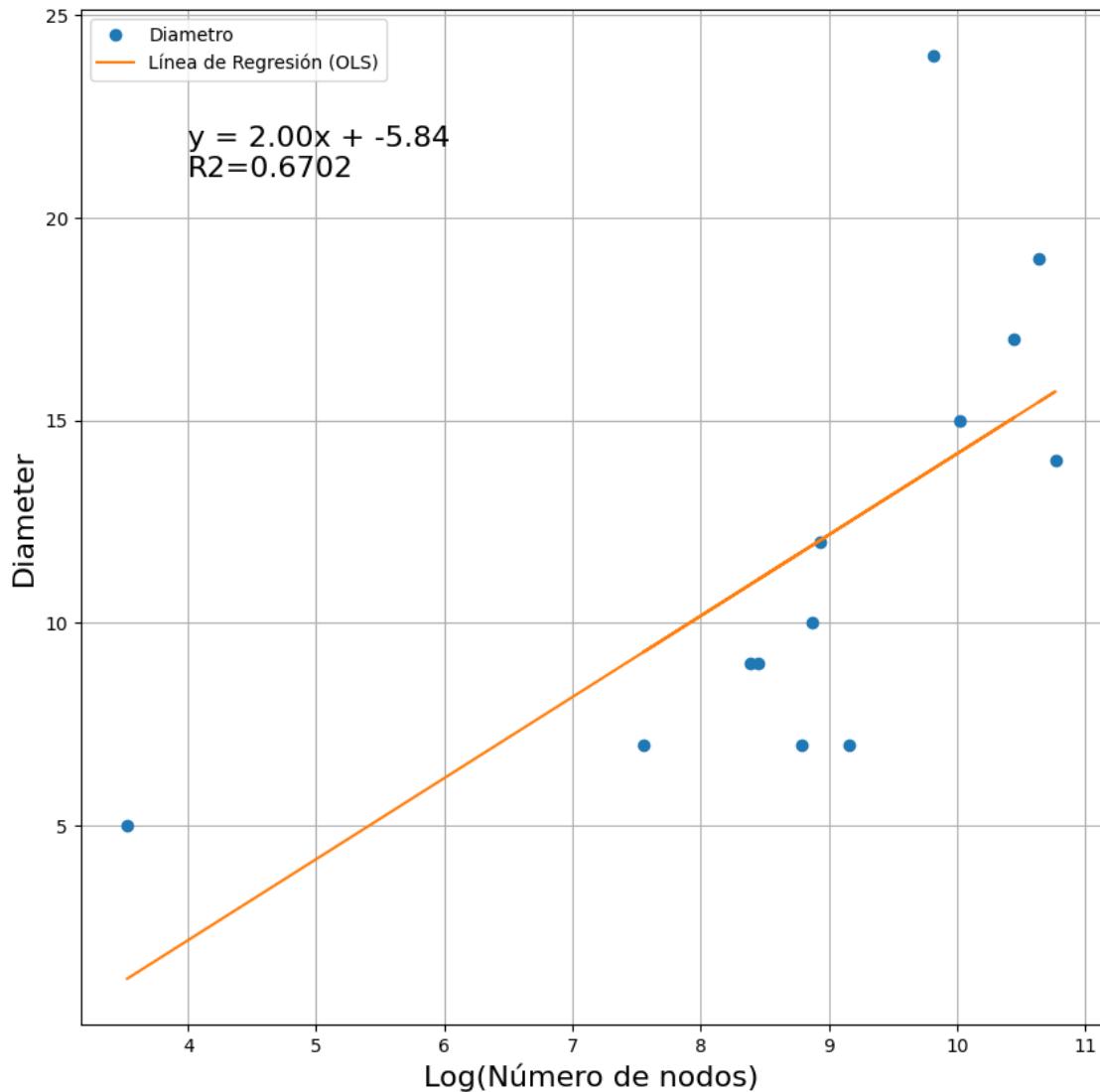
## 3 Análisis agregado

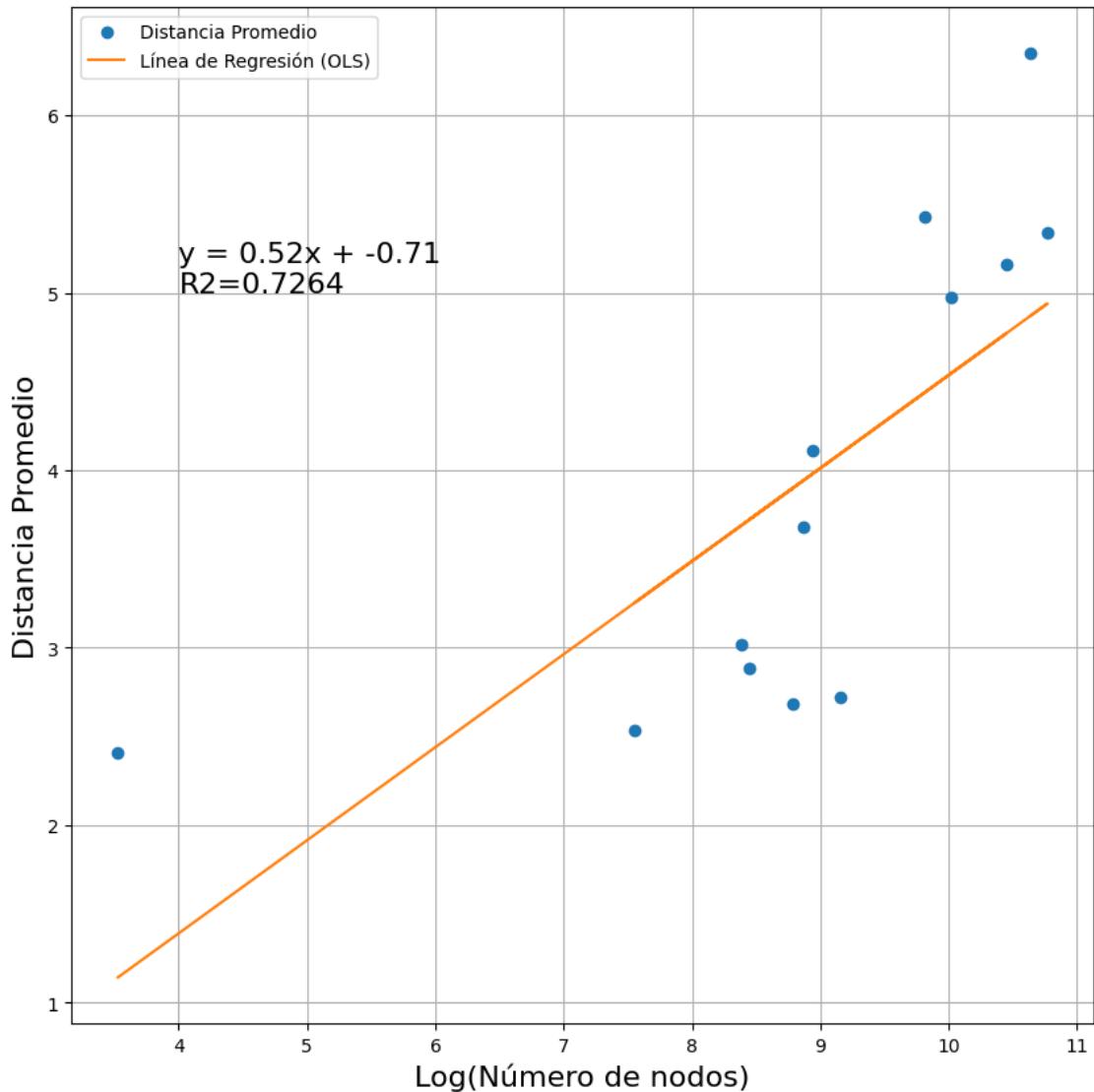
Dataset	Nodos	Aristas	Diámetro	Distancia Promedio	Coeficiente de Agrupamiento
Karate Club	34	78	5	2.41	0.26
Actor	7600	26752	12	4.11	

0.02							
Coautor CS	18333	81894	24	5.43			
0.18							
Coautor Física	34493	247962	17	5.16			
0.19							
Twitch DE	9498	162636	7	2.72			
0.05							
Twitch EN	7126	42450	10	3.68			
0.04							
Twitch ES	4648	64030	9	2.88			
0.08							
Twitch FR	6549	119215	7	2.68			
0.05							
Twitch PT	1912	33211	7	2.53			
0.13							
Twitch RU	4385	41689	9	3.02			
0.05							
FacebookPagePage	22470	171002	15	4.97			
0.23							
GemsecDeezer HU	47538	222887	14	5.34			
0.09							
GemsecDeezer RO	41773	125826	19	6.35			
0.08							
<hr/>							
<hr/>							

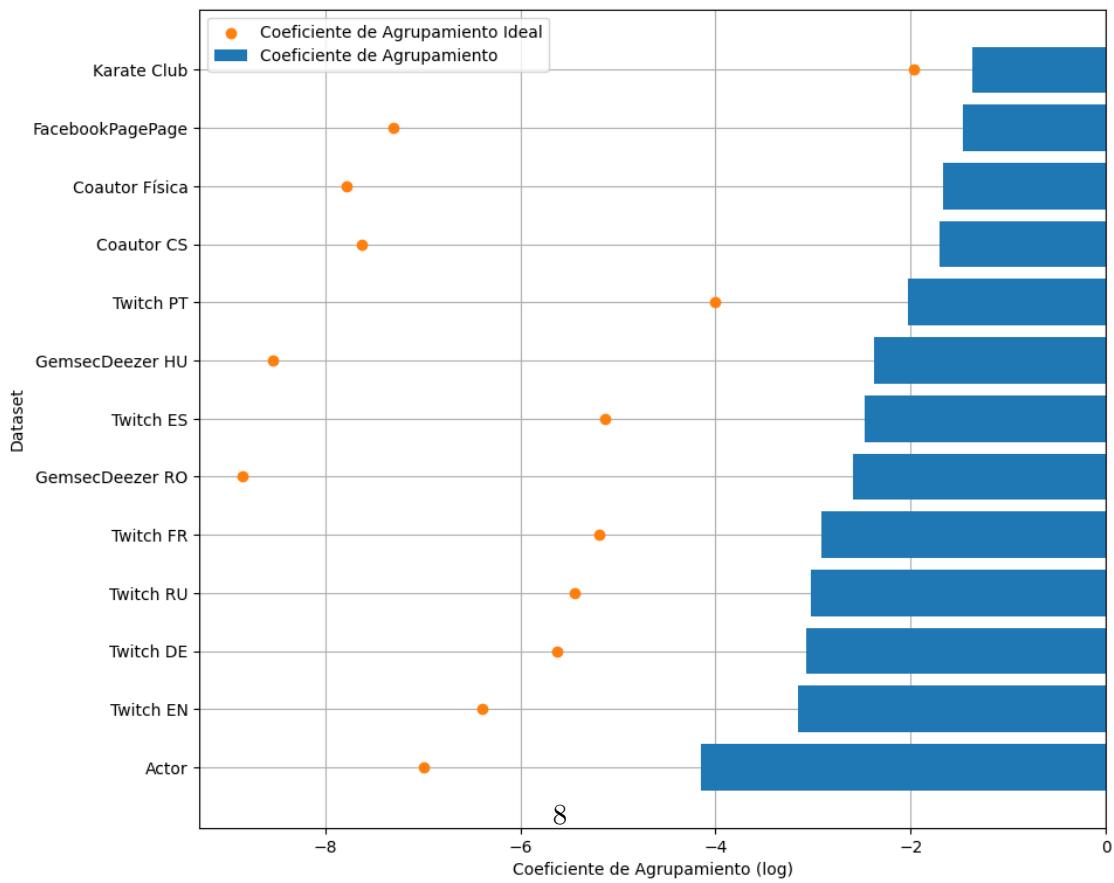
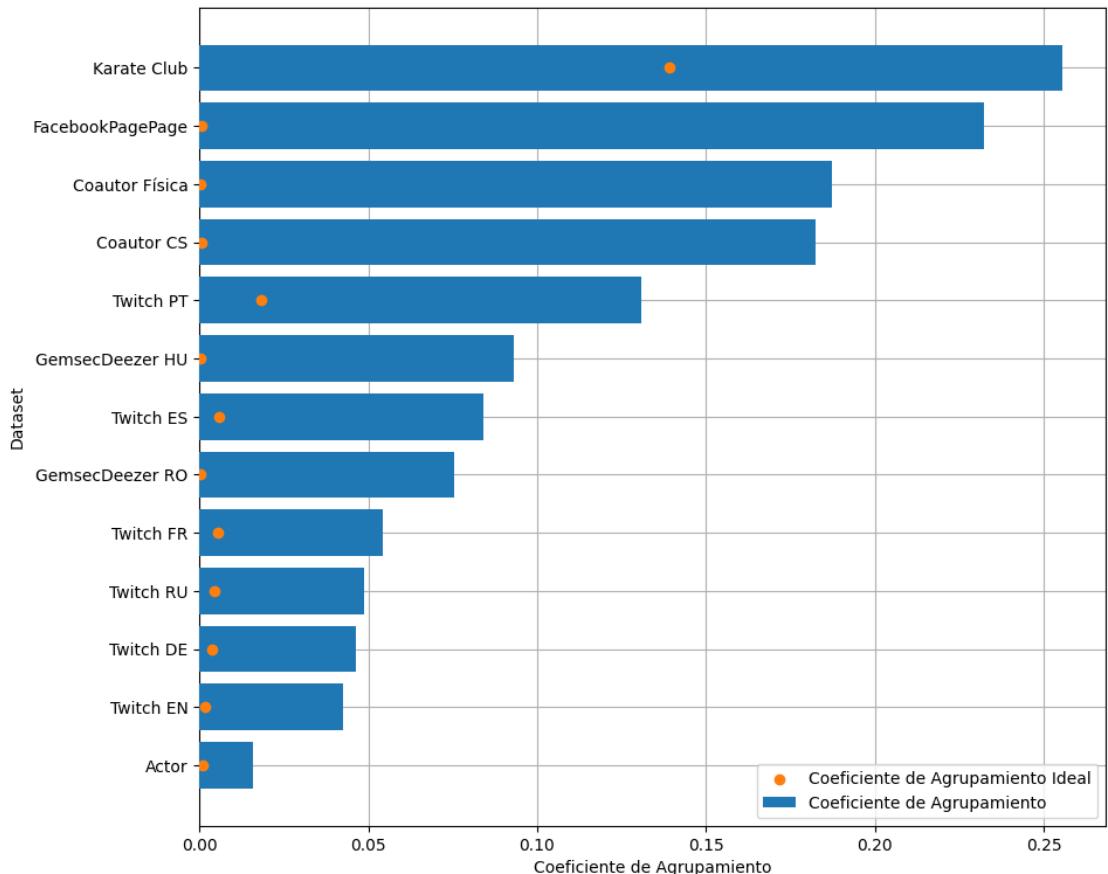
Con los datos generados en el apartado anterior, aunque a simple vista parece cumplirse la hipótesis (al menos yo al ver la tabla veo que el diámetro y la distancia promedio crecen muy lentamente al aumentar la cantidad de nodos y aristas, aunque esta es una afirmación *subjetiva* (\*)), podemos hacer más cosas. Por ejemplo, ver si efectivamente el logaritmo del número de nodos es un buen predictor lineal del diámetro y de la distancia promedio.

(*) Lo que quiero decir es que  $E[\text{diametro}|\text{nodos}]$  y  $E[\text{dist\_promedio}|\text{nodos}]$  crecen muy lentamente como funciones de nodos. Es notable que nosotros no tenemos acceso a las esperanzas, sino que en este caso digo que la estimación a ojimétro\** (la percepción subjetiva), influida por los ejemplos observados, de las mismas se comporta de esta manera.





Como puede verse anotado en ambos gráficos, el  $R^2$  como medida de capacidad predictiva de una variable sobre otra de forma lineal es de 0.67 en un caso y 0.73 en el otro. Lo que hace pensar que la función logaritmo es buen predictor de estas métricas, lo que es consistente con la hipótesis de mundo pequeño.



A su vez, vemos que todos los grafos tienen un coeficiente de agrupamiento mayor al de un grafo aleatorio de Erdos-Renyi con la misma cantidad de nodos y aristas, lo que nos sugiere que todos cumplen la propiedad de mundo pequeño.

El gráfico en escala logarítmica permite ver como hay órdenes de magnitud de diferencia entre el coeficiente de agrupamiento y el de un grafo aleatorio de Erdos-Renyi, y también entre los coeficientes normalizados de distintos grafos.

## 4 Persistencia de la propiedad de Mundo Pequeño

Como idea innovadora de este trabajo práctico propongo ver si la propiedad de ser mundo pequeño se mantiene al retirar aristas. Para eso, para los datasets chicos (menos de 100K aristas) voy a formar una grilla de proporciones de aristas y para cada proporción voy a generar una población de subgrafos con ese porcentaje de aristas y voy a realizar el mismo análisis para cada uno de esos subgrafos.

Por tanto, los grafos excluidos son los siguientes:

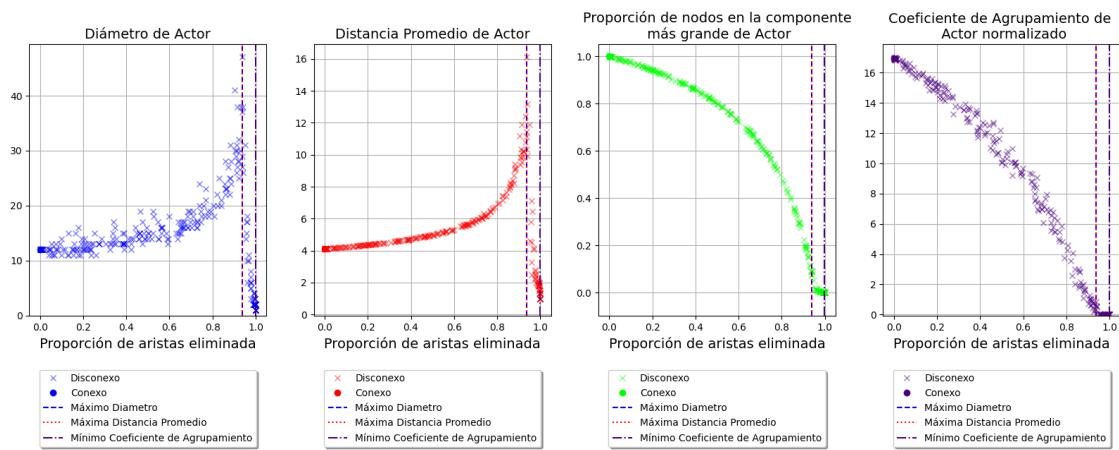
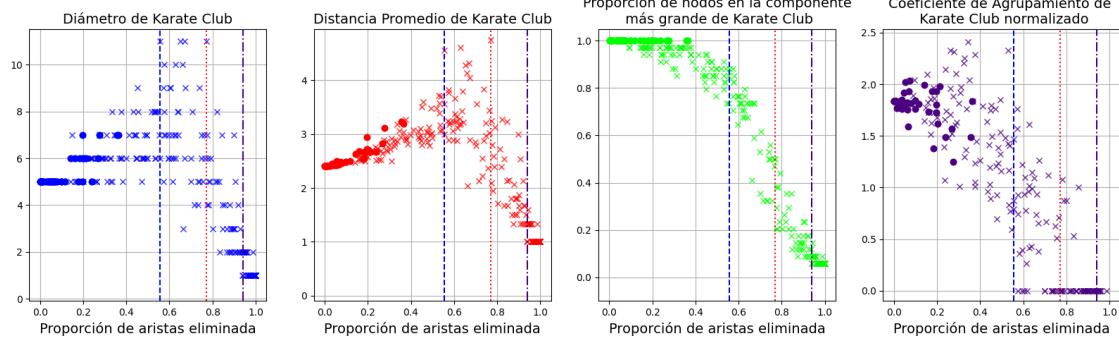
El grafo Coautor Física tiene 247962 aristas  
El grafo Twitch DE tiene 162636 aristas  
El grafo Twitch FR tiene 119215 aristas  
El grafo FacebookPagePage tiene 171002 aristas  
El grafo GemsecDeezer HU tiene 222887 aristas  
El grafo GemsecDeezer RO tiene 125826 aristas

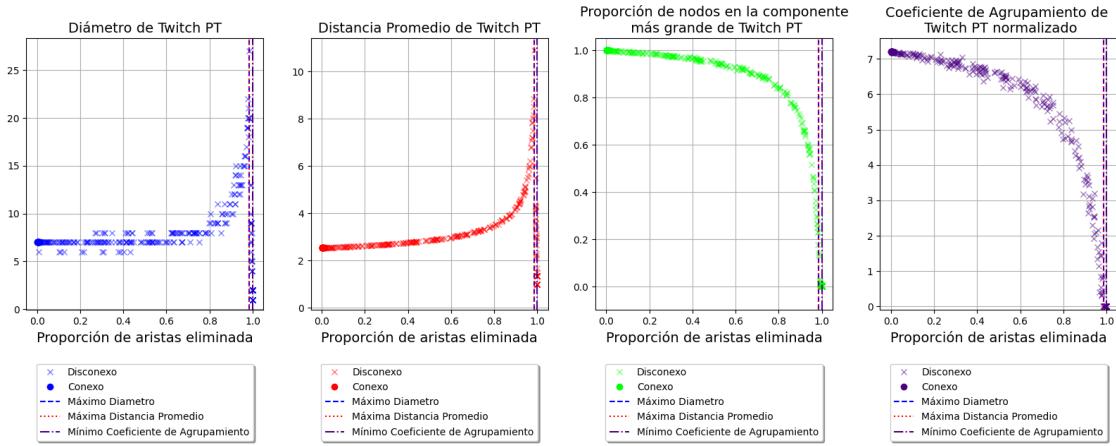
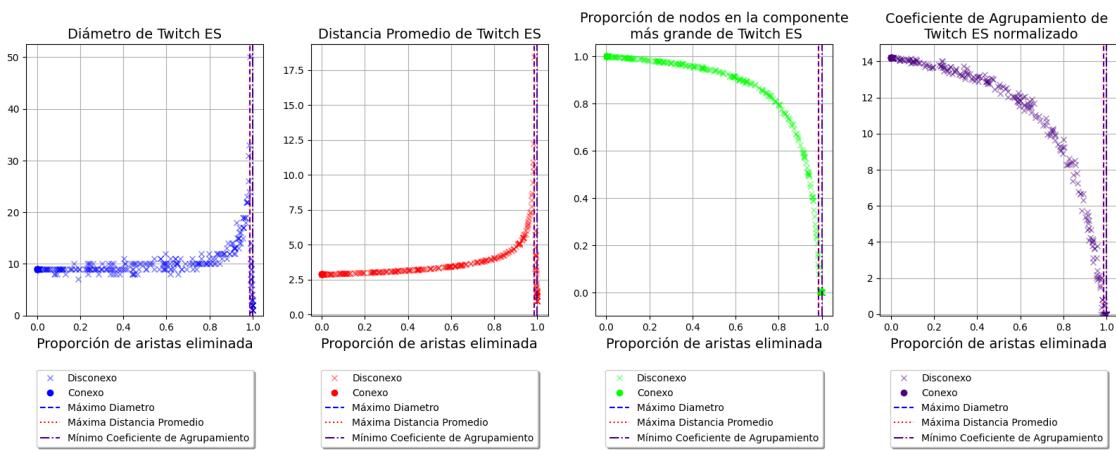
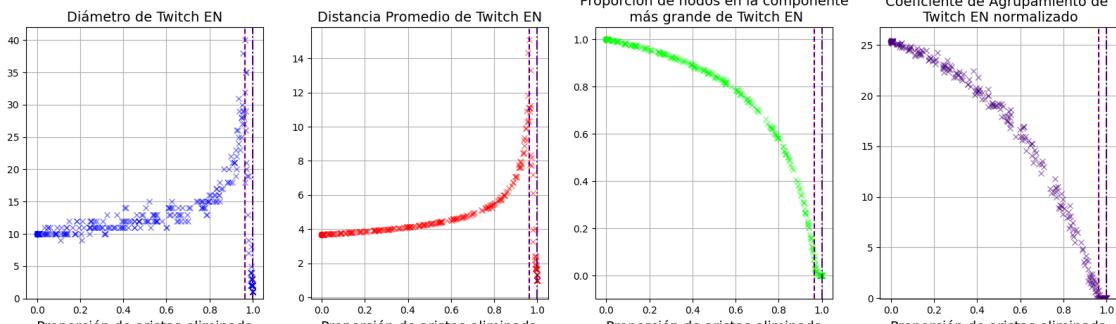
Y se utilizan en la siguiente parte del análisis los siguientes grafos:

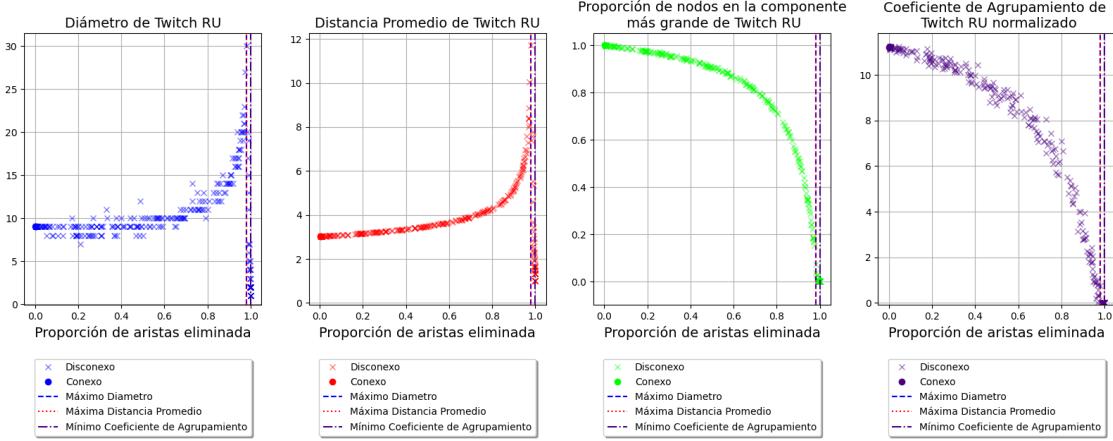
El grafo Karate Club tiene 78 aristas  
El grafo Actor tiene 26752 aristas  
El grafo Coautor CS tiene 81894 aristas  
El grafo Twitch EN tiene 42450 aristas  
El grafo Twitch ES tiene 64030 aristas  
El grafo Twitch PT tiene 33211 aristas  
El grafo Twitch RU tiene 41689 aristas

Hay que tener en cuenta que como al eliminar aristas el grafo puede dejar de ser conexo, se calcularán las métricas sobre la componente conexa más grande del subgrafo resultante. Además, considero que el tamaño de esta componente conexa es una métrica interesante a analizar y graficar.

```
['variables/calculados.pkl']
```







Lo que se puede apreciar en el gráfico anterior es que todos los grafos (salvo Karate por ser especialmente chico) siguen una dinámica similar donde el tamaño de la mayor componente conexa decrece más lentamente que la proporción de aristas conservadas, para posteriormente caer abruptamente. A su vez, el diámetro y la distancia promedio de la componente mayor crecen lentamente hasta que cerca del punto donde se conservan 0 aristas tienen un pico abrupto y luego caen rápidamente cuando la componente conexa mayor es muy chica.

Como se puede ver, las líneas punteadas azul y roja coinciden en todos los casos salvo Karate, es decir que los picos de las métricas Diámetro y Distancia Promedio se dan en el mismo caso para todos los grafos. Es decir, de todos los subgrafos generados eliminando subconjuntos aleatorios de las aristas, el que tiene mayor diámetro es el mismo que el que tiene mayor distancia promedio entre nodos.

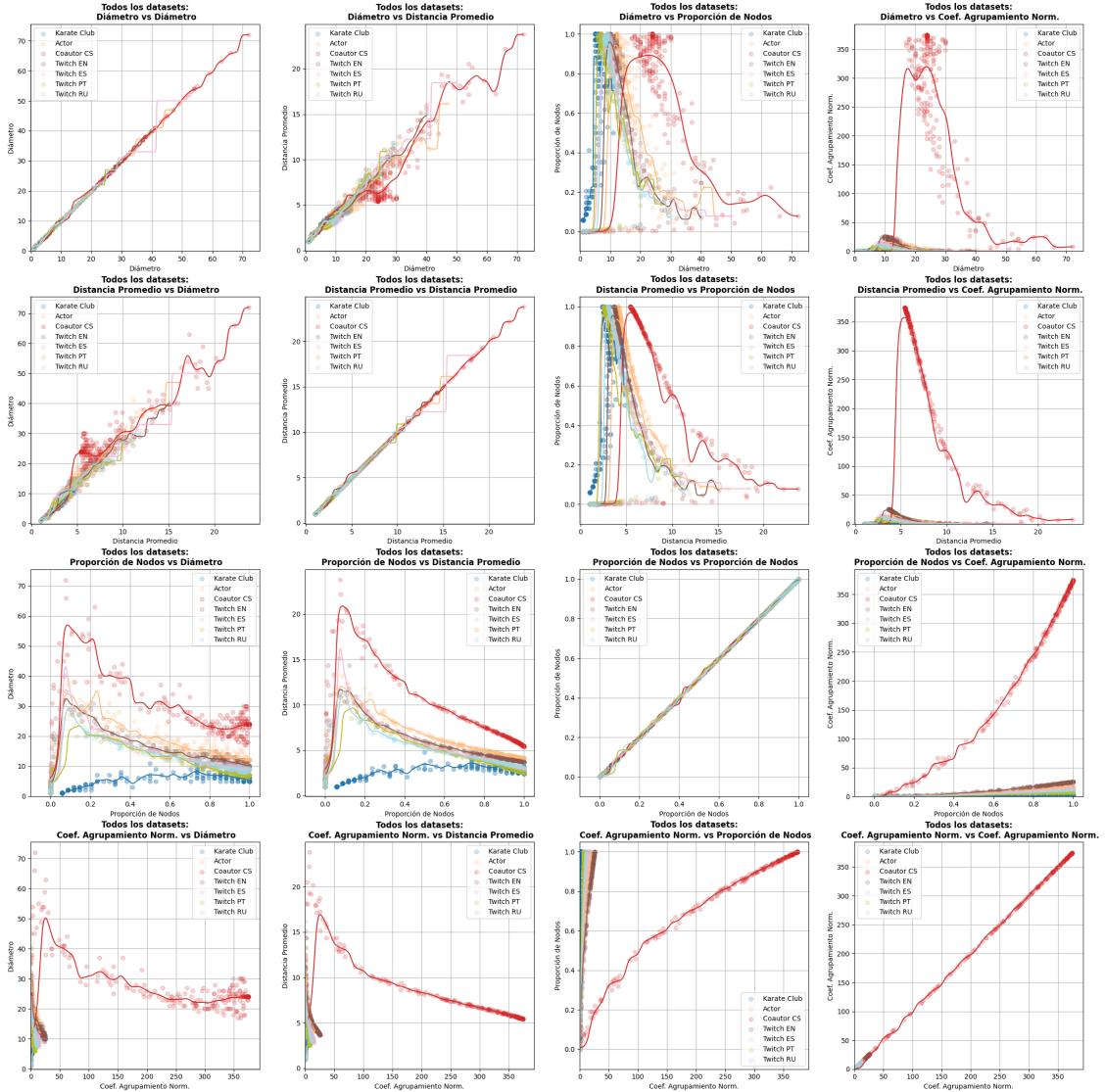
La línea violeta, que indica en cuál caso se obtuvo el mínimo coeficiente de agrupamiento normalizado, no coincide con las líneas roja y azul pero queda cerca en todos los casos. Es razonable dado que coincide con que al perder muchas aristas se deterioran todas las características del mundo pequeño. Como se puede ver, el coeficiente de agrupamiento tiene un comportamiento curiosamente similar al tamaño de la componente más grande.

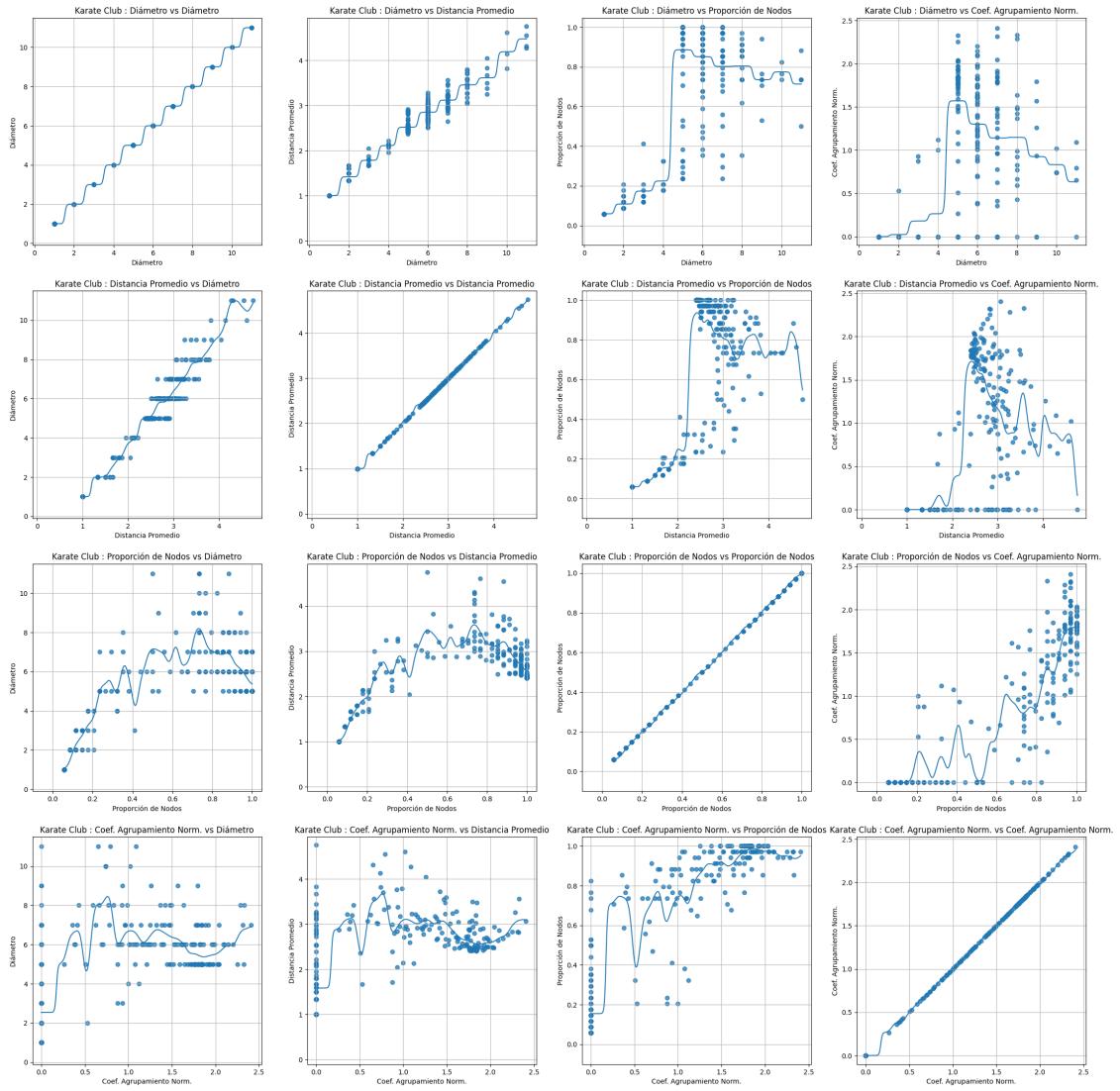
## 5 Comportamiento conjunto de las variables

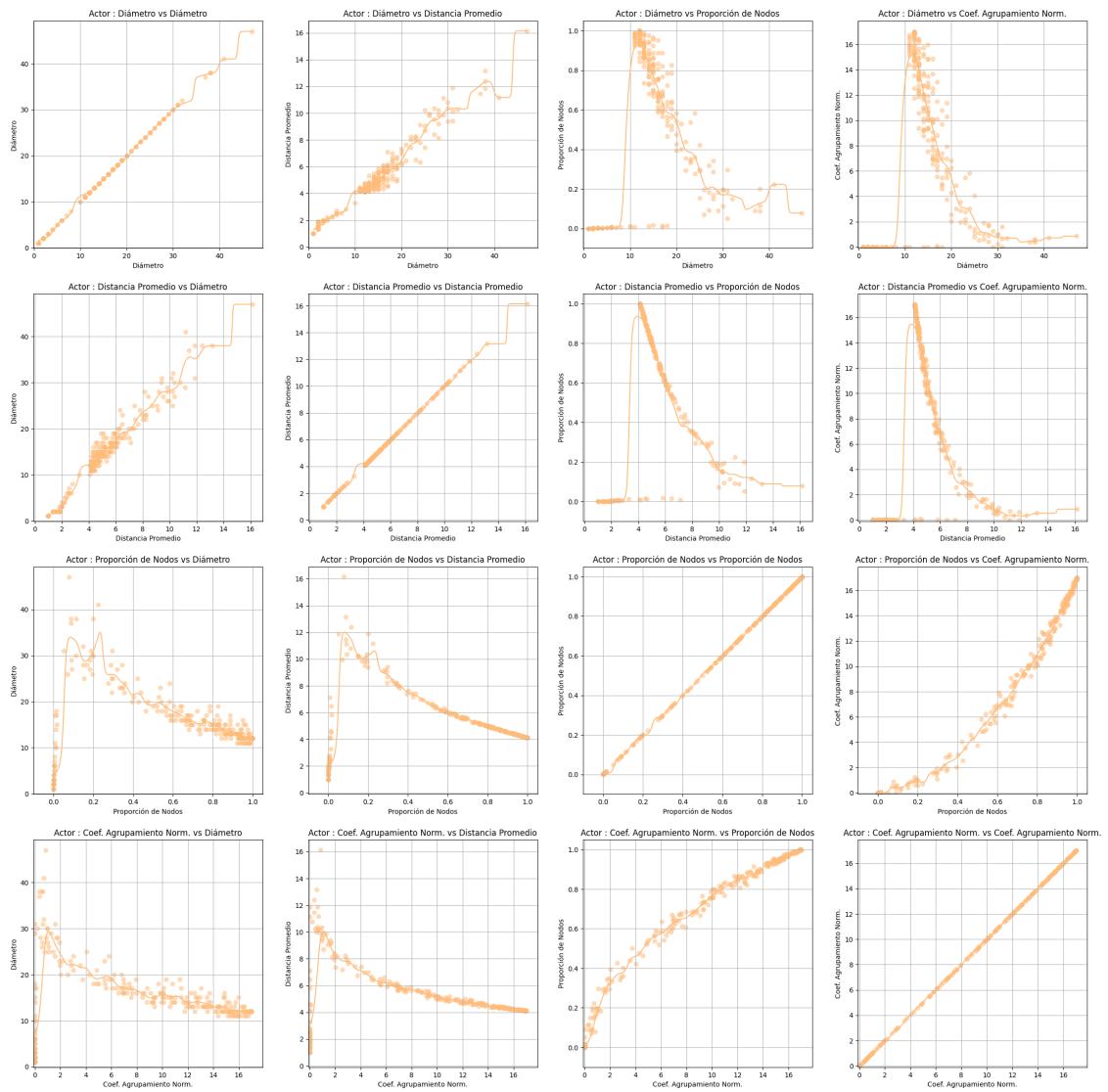
Otra cosa que puede ser interesante ver es la distribución conjunta de las 4 variables calculadas, dando lugar a una grilla de 4x4 graficos para cada gráfico (es decir, el conjunto de subgráficos de ese gráfico):

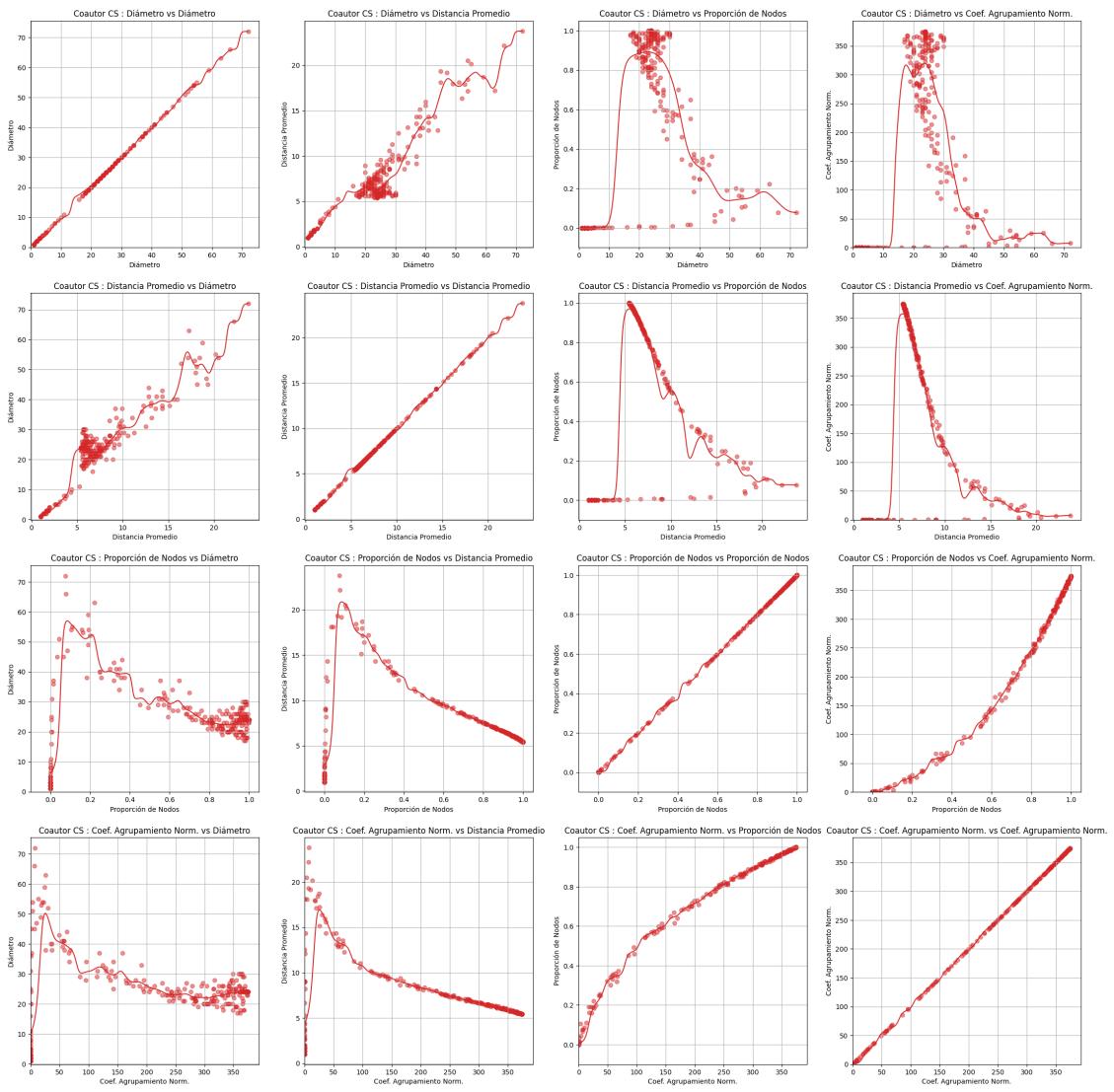
- **Diámetro**
- **Distancia Promedio**
- **Tamaño componente mayor**
- **Coeficiente de agrupamiento normalizado**

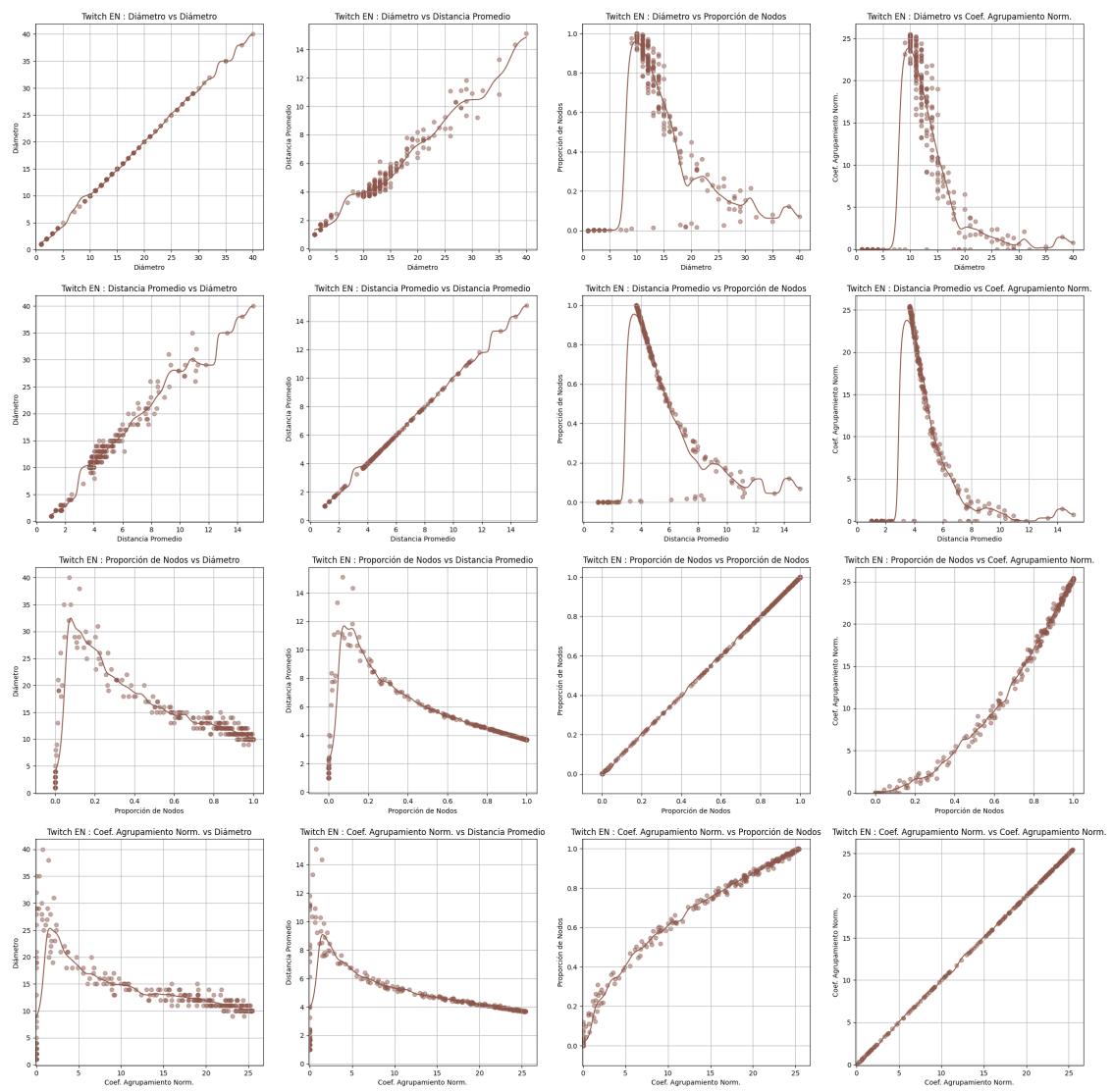
Como al eliminar aristas es esperable que el grafo deje de ser conexo, por un lado es importante tener en cuenta que el diámetro y la distancia promedio se calcularan sobre

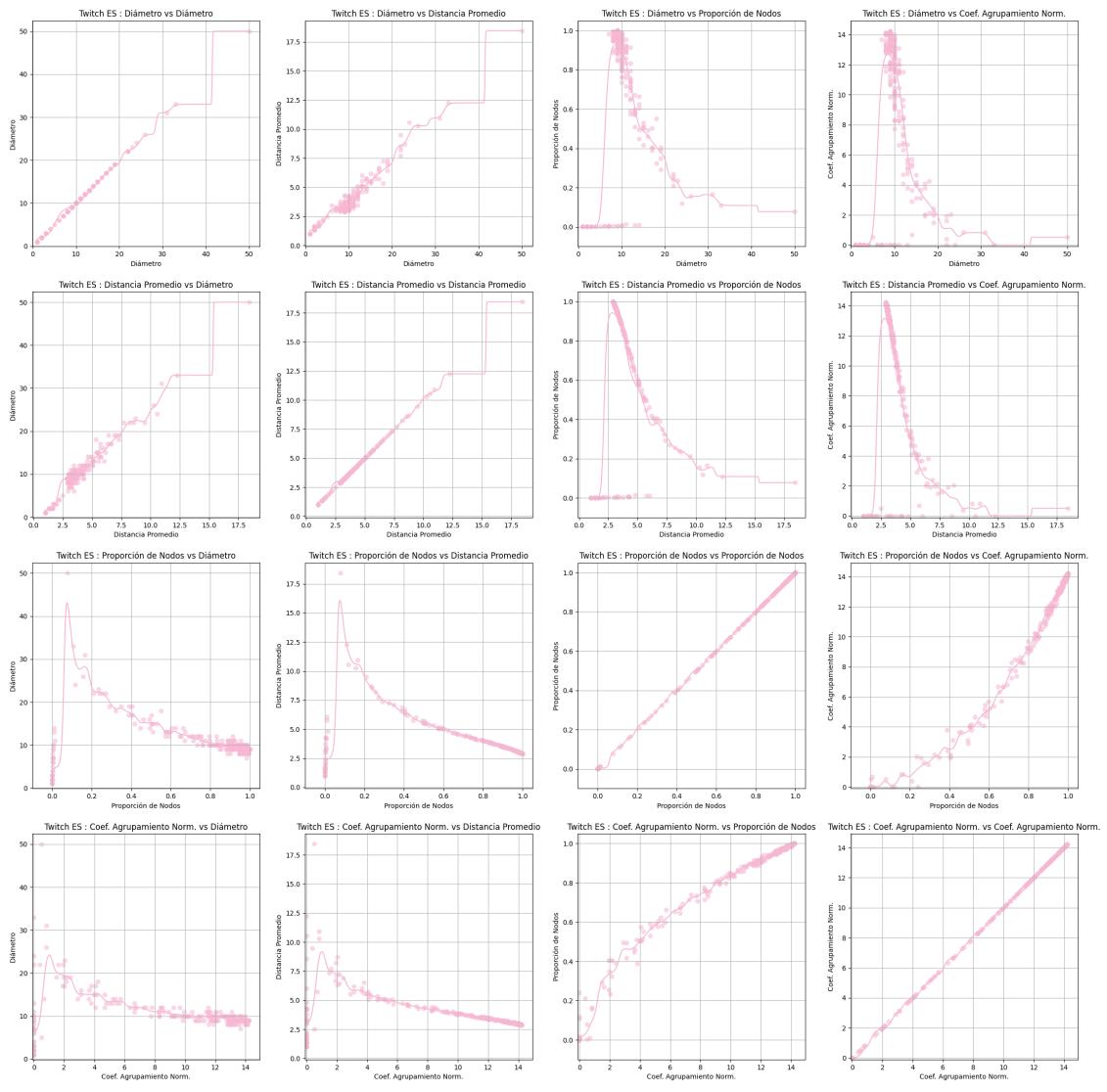


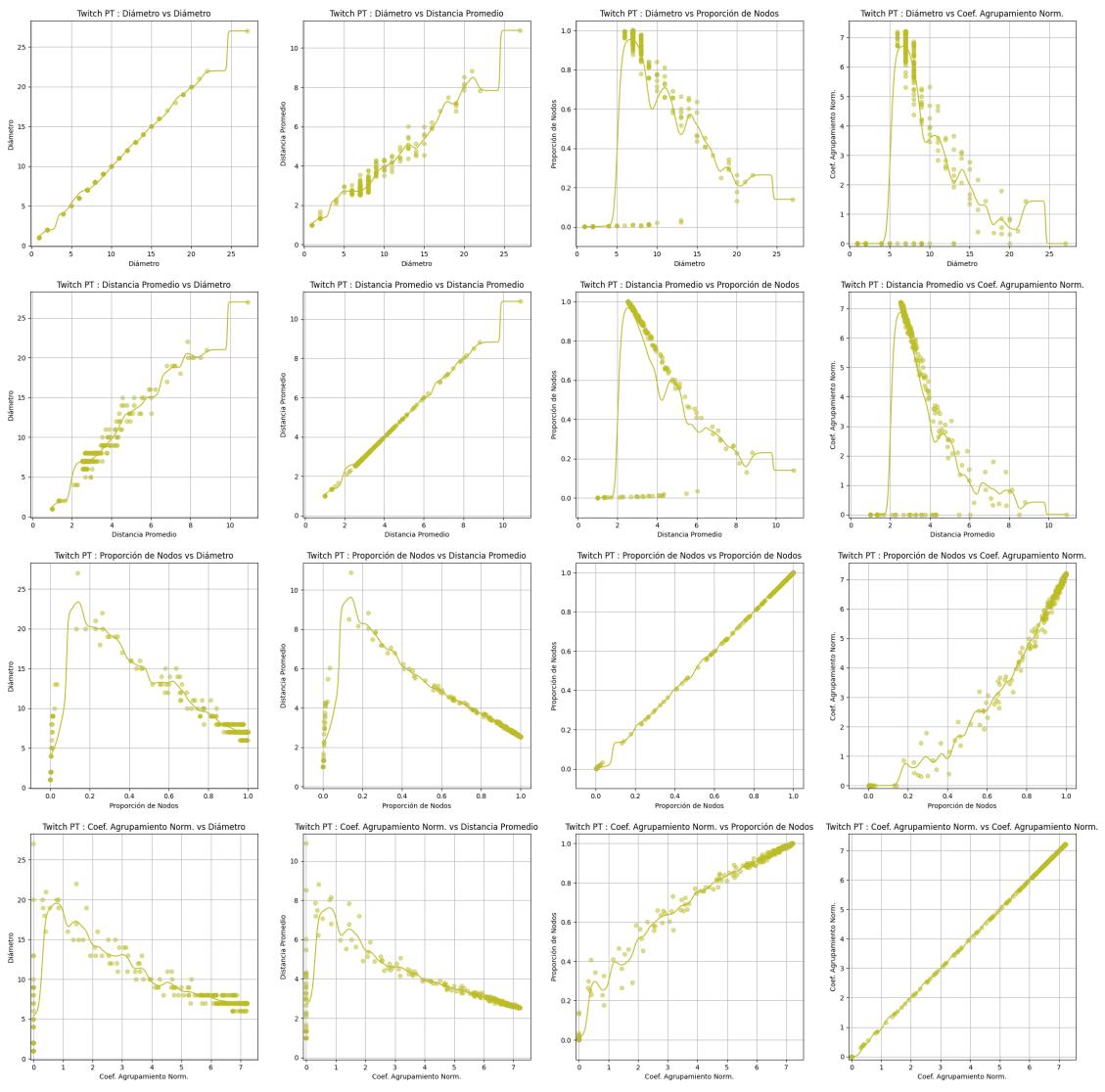


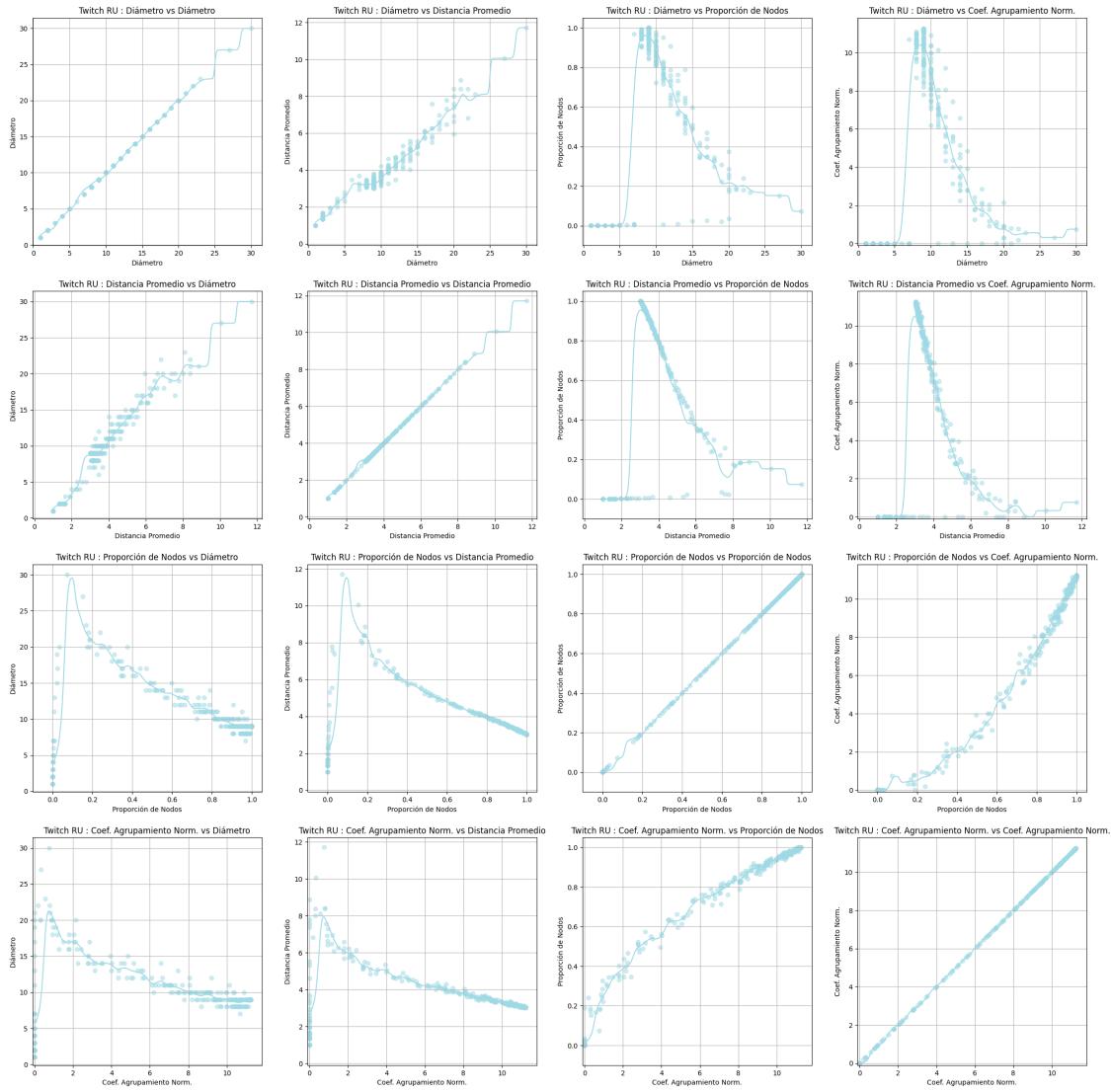












En las gráficas anteriores la línea de aproximación se calculó con un ajuste no paramétrico utilizando un kernel normal, la varianza del mismo se calculó mediante validación cruzada sobre una grilla de valores.

Respecto de las baterías de gráficos de arriba, es notable como en los distintos grafos las curvas generadas son similares.

También resalta como **Coautor CS** tiene un coeficiente de agrupamiento normalizado en una escala completamente distinta a los demás grafos.

Se ve que en general, como cabría esperar, las características de Small World se correlacionan bien (al menos viendo los gráficos). Es decir, tienen a coincidir:

- Valor bajos (altos) del diámetro.
- Valor bajos (altos) de la distancia promedio.
- Valores altos (bajos) del coeficiente de agrupamiento normalizado.

- Valor altos (bajos) del tamaño de la componente conexa más grande.

Esto es consistente con pensar que son métricas que reflejan un mismo fenómeno, en este caso la propiedad de mundo pequeño.

*En lo anterior se ha ignorado el dataset de Karate, ya que aunque las observaciones le aplican parcialmente, al ser un grafo tan chico hay mucho más ruido en las observaciones.*

## 6 Conclusiones

Lo que se desprende del presente estudio no es solo una verificación empírica del hecho de que los grafos sociales tienen la propiedad de mundo pequeño, sino además que esta propiedad es robusta al retirar una cantidad moderada de aristas.

Al retirar la mitad de las aristas ni el diámetro ni la distancia promedio aumentan significativamente y la componente más grande conserva más del 80% de los nodos totales.

Esto tiene sentido si se tiene en cuenta que son grafos donde la cantidad de aristas es significativamente mayor a la cantidad de nodos, por lo cual hay mucha redundancia, lo que explica su resistencia a la pérdida de aristas. Aunque no sea una cantidad cuadrática de aristas, alcanza para que al reducir en una proporción fija la cantidad de aristas, la mayor componente conexa mantenga a una amplia mayoría de los nodos.

Estas dos cosas pueden explicarse también al tener en cuenta la semántica de las aristas. Como estas modelan interacciones sociales, es esperable que no se den de forma aislada, sino que, generalmente, como parte de un grupo de aristas que vinculan fuertemente a un grupo de nodos entre sí.