



PROYECTO ANALISIS DE DATOS

Introducción

El conjunto de datos seleccionado proviene de la plataforma de transmisión Twitch y se centra en las redes sociales entre usuarios que comparten contenido relacionado con videojuegos. Twitch es una plataforma líder en la transmisión en vivo de videojuegos y actividades creativas, donde los usuarios, conocidos como "streamers", interactúan con sus seguidores a través de transmisiones en tiempo real. Este conjunto de datos específico se enfoca en las redes de usuarios de Twitch que transmiten en diferentes idiomas, con un enfoque especial en el español. La recopilación de datos se realizó en mayo de 2018 y aborda aspectos fundamentales de las interacciones sociales, como las amistades mutuas entre los streamers. La información de los nodos incluye características como los juegos jugados, preferencias de ubicación y hábitos de transmisión, brindando así una perspectiva rica sobre las dinámicas de la comunidad de Twitch.

Dataset: Preparacion y Limpieza

Aristas

from y to: Estas columnas representan nodos en la red social de Twitch, específicamente streamers que transmiten en español. Cada fila indica una conexión bidireccional (amistad mutua) entre dos streamers, permitiendo la construcción de la red de usuarios en la plataforma.

	from	to
0	0	1819
1	0	2840
2	1	1565
3	1	1309
4	1	1397
...
59377	2225	3811
59378	4638	940
59379	940	4645
59380	940	3921
59381	3184	3921
59382 rows × 2 columns		

Nodos

- id: Identificador único asociado a cada nodo en la red social de Twitch, específicamente para los streamers que transmiten en español.
- days: Número de días que ha estado activo el streamer.
- mature: Indicador que especifica si el contenido del streamer es para audiencias maduras.
- views: Número total de vistas acumuladas por el streamer.
- partner: Indicador que especifica si el streamer es un socio oficial de Twitch.
- new_id: Nuevo identificador asociado al streamer.

	id	days	mature	views	partner	new_id
0	68458707	1522	False	4405	False	3558
1	133928858	768	False	164810	True	3372
2	46892468	1895	False	4953	False	818
3	128745923	828	True	12262	False	236
4	84422595	1317	False	4937	False	2255
...
4643	94945228	1199	False	2740	False	4613
4644	86149235	1292	False	512	False	2673
4645	44028163	1949	False	442	False	2430
4646	63400827	1594	False	4877	False	2830
4647	62989435	1604	False	31719	False	655

4648 rows × 6 columns

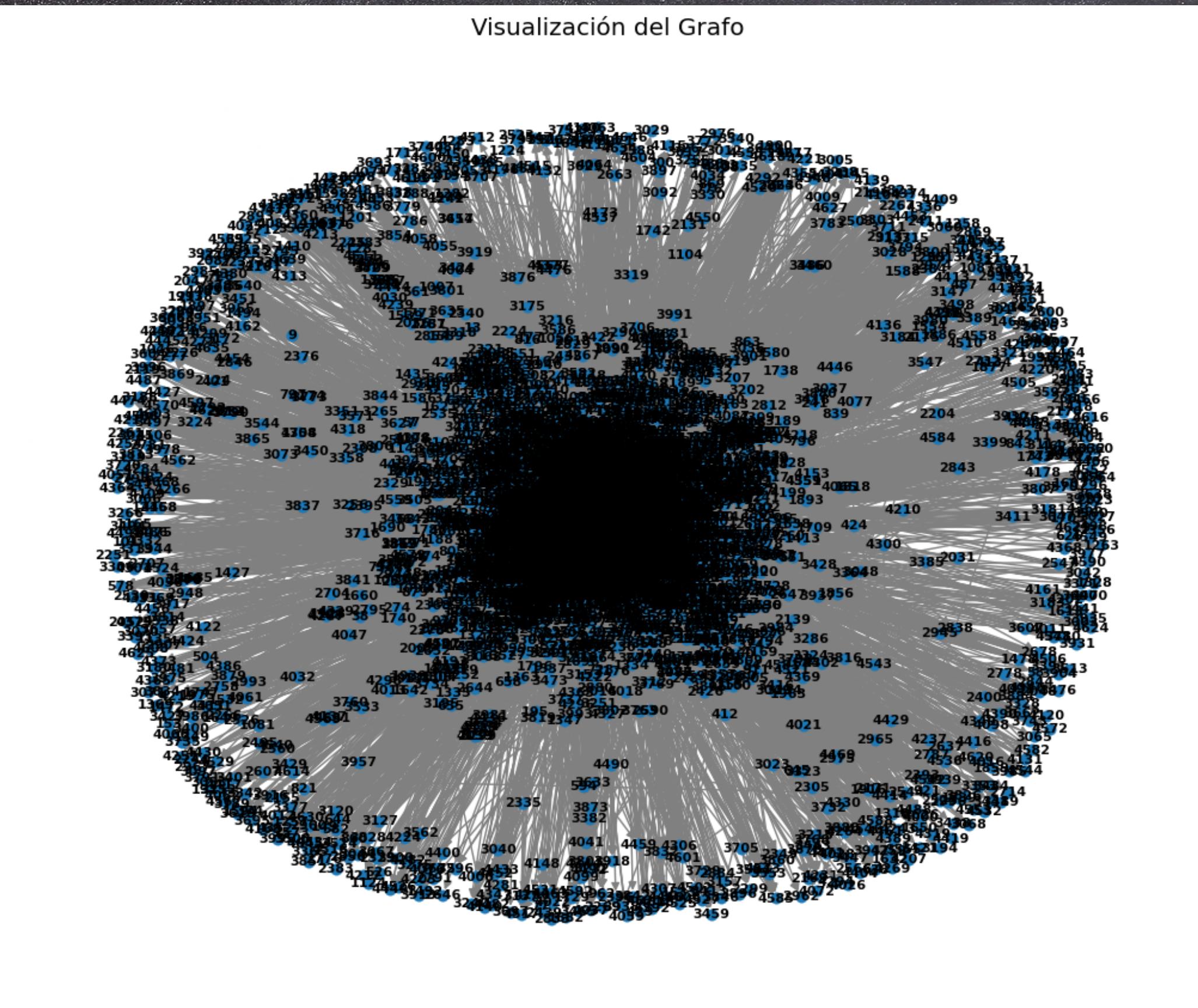
Atributos

Atributos del Nodo en formato JSON: Este archivo contiene características adicionales específicas de cada nodo en la red de Twitch, incluyendo información sobre los juegos jugados, preferencias de ubicación y hábitos de transmisión. Los números en el JSON representan valores asociados a estas características.

Ejemplos de Atributos:

```
1412: [89, 166, 1040, 846, 2987, 1649, 920, 224, 3097, 400, 569, 822, 2362, 802, 2728, 2734]
3032: [515, 1943, 289, 3084, 1575, 3164, 920, 224, 3097, 400, 1391, 635, 569, 821, 2645, 1147, 440]
4032: [1948, 421, 586, 202, 2024, 846, 45, 3164, 920, 224, 2798, 2064, 2534, 139, 2664, 2362]
3945: [438, 2464, 967, 861, 152, 1649, 920, 1907, 2185, 2986, 1607, 1895, 1013, 928, 569, 139, 608, 2362, 802, 1530, 1028, 1147, 763, 2734]
949: [2598, 1713, 1053, 2928, 473, 846, 920, 224, 3097, 706, 1525, 2912, 2362]
1051: [2554, 1417, 2, 846, 1203, 653, 920, 224, 3097, 329, 569, 2384, 2764]
3156: [1726, 2864, 1778, 1433, 653, 811, 2063, 920, 224, 3097, 615, 1013, 1362, 2645, 2737, 2677, 569, 3054, 2764, 802, 3044, 1245, 2424,
4193: [2352, 1535, 28, 2035, 2928, 131, 2063, 920, 224, 810, 553, 2549, 2524, 929, 83, 139, 2168, 704, 440, 276, 2645, 2501, 69, 1876]
1803: [670, 216, 160, 2374, 1998, 653, 1225, 48, 920, 1144, 1761, 2386, 2645, 2424, 2737, 1895]
4437: [1550, 216, 2307, 861, 1709, 861, 920, 1144, 1761, 760, 2335, 1876]
2468: [3048, 378, 333, 653, 2331, 2003, 2814, 119, 1644, 2906, 326, 1606, 569, 2783, 276, 1548, 2676]
289: [1449, 1943, 2212, 861, 68, 1649, 920, 224, 3097, 615, 2912, 1013, 635, 139, 569, 395, 2764, 3161, 276, 2645, 2424, 2737, 1525]
3055: [1554, 2864, 1294, 861, 1955, 48, 920, 224, 810, 1428, 704, 2362]
922: [308, 3152, 677, 861, 184, 2063, 920, 224, 810, 14, 821, 569, 2645, 2384, 440]
2811: [308, 3152, 642, 3109, 653, 2853, 3139, 653, 920, 1907, 1612, 2327, 1656, 326, 588, 257, 635, 1213, 569, 1796, 2645, 2676, 2130]
1689: [1199, 1713, 1358, 2928, 2738, 846, 920, 224, 3097, 400, 569, 821, 2645, 608]
3242: [2248, 1191, 2927, 861, 1594, 2282, 920, 224, 810, 464, 139, 276, 2424, 2737]
3324: [1234, 3152, 496, 846, 1881, 1085, 920, 224, 3097, 1574, 2384, 2645]
221: [1951, 165, 2583, 2950, 796, 653, 1085, 920, 1907, 2185, 1043, 569, 257, 3044]
```

Visualización del Gra



Depuracion del dataset

Exploracion inicial

Visualizamos las primeras filas del DataFrame para entender la estructura de los datos y verificar si hay problemas obvios

	<code>id</code>	<code>days</code>	<code>mature</code>	<code>views</code>	<code>partner</code>	<code>new_id</code>
0	68458707	1522	False	4405	False	3558
1	133928858	768	False	164810	True	3372
2	46892468	1895	False	4953	False	818
3	128745923	828	True	12262	False	236
4	84422595	1317	False	4937	False	2255

Depuracion del dataset

Valores nulos

Verificamos la presencia de valores nulos en el conjunto de datos.

En este caso no existen datos nulos.

```
id          0
days        0
mature      0
views       0
partner     0
new_id      0
dtype: int64
```

Depuracion del dataset

Duplicados

Buscamos y eliminamos posibles filas duplicadas en el conjunto de datos. No existen datos duplicados.

```
[9] aristas.duplicated().sum()  
objetivo.duplicated().sum()
```

0

Depuración del dataset

Exploración Estadística

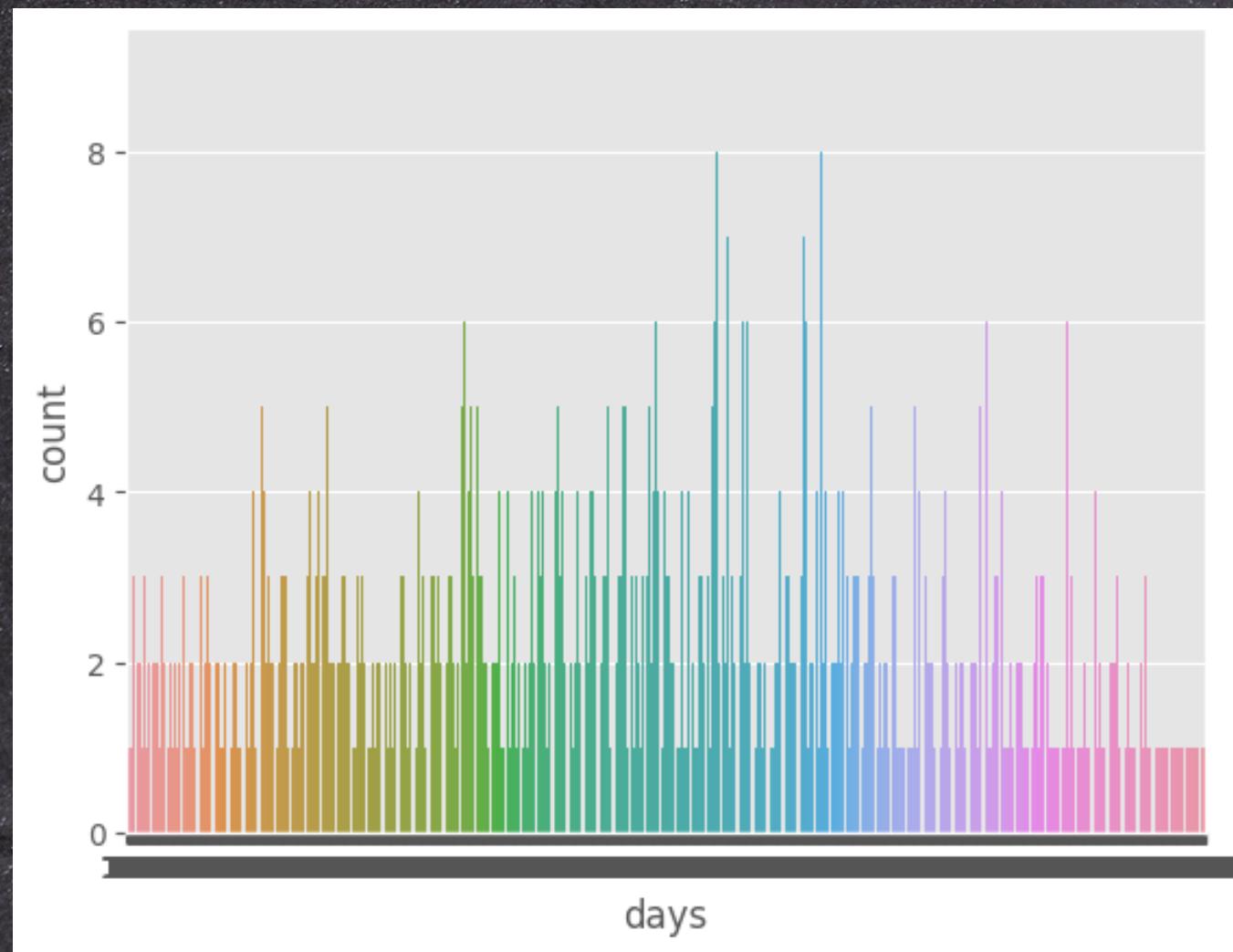
Realizamos un análisis estadístico básico para entender la distribución de las variables numéricas.

	id	days	views	new_id
count	4.648000e+03	4648.000000	4.648000e+03	4648.000000
mean	9.130725e+07	1341.602625	1.032367e+05	2323.500000
std	4.768005e+07	648.824912	8.852851e+05	1341.906355
min	5.933900e+04	118.000000	8.000000e+00	0.000000
25%	5.083369e+07	830.000000	1.532750e+03	1161.750000
50%	8.290819e+07	1323.500000	4.404500e+03	2323.500000
75%	1.280544e+08	1799.250000	1.509250e+04	3485.250000
max	2.177764e+08	4083.000000	3.054557e+07	4647.000000

Depuracion del dataset

Visualizaciones Exploratorias

Utilizamos visualizaciones, como gráficos de barras para explorar la distribución de las variables y detectar posibles anomalías.



Medida de Grado (In-Degree y Out-Degree)

La medida de out-degree se aplica en grafos dirigidos para calcular el número de aristas que salen de un nodo específico. En un grafo no dirigido, las aristas no tienen una dirección específica, es decir, no hay nodos "de salida" o "de llegada".

Medidas de Grado:

{0: 2, 1819: 1022, 2840: 22, 1: 9, 1565: 595, 1309: 35, 1397: 357, 2677: 101, 3497: 3, 357: 5,

Todos los nodos están conectados de manera bidireccional.

Medida de Cercanía (Closeness Centrality)

La cercanía de un nodo mide qué tan cercano está un nodo respecto a los demás nodos en términos de distancia geodésica. La cercanía de un nodo es alta si está cerca de todos los demás nodos. Es útil para identificar nodos que podrían comunicarse de manera más eficiente con el resto de la red.

```
Closeness Centrality: {0: 0.34409477971121805, 1819: 0.5168501835168502, 2840: 0.395691416893733, 1: 0.34801168276791733, 1565: 0.48936394271272116,
```

```
1309: 0.39612991219844856, 1397: 0.4466122056703508, 2677: 0.43670707640259376, 3497: 0.29154903067946547, 357: 0.3256482130343378,
```

Medida de Intermediación (Betweenness Centrality)

La cercanía de un nodo mide qué tan cercano está un nodo respecto a los demás nodos en términos de distancia geodésica. La cercanía de un nodo es alta si está cerca de todos los demás nodos. Es útil para identificar nodos que podrían comunicarse de manera más eficiente con el resto de la red.

```
Betweenness Centrality: {0: 0.0, 1819: 0.10946565514740338, 2840: 0.00018388202132763547, 1: 0.0005271177451024974, 1565: 0.03924203341079563,
```

```
1309: 0.0002970660873101713, 1397: 0.012837208636540557, 2677: 0.002258825136606329, 3497: 1.8452572785910947e-06, 357: 4.361716217889616e-05, 1492: 0.0,
```

PageRank

La cercanía de un nodo mide qué tan cercano está un nodo respecto a los demás nodos en términos de distancia geodésica. La cercanía de un nodo es alta si está cerca de todos los demás nodos. Es útil para identificar nodos que podrían comunicarse de manera más eficiente con el resto de la red.

```
PageRank: {0: 4.6849930262989706e-05, 1819: 0.009161052057663354, 2840: 0.00018769412615564569, 1: 0.00016359731937688084, 1565: 0.004606983392185898,  
1309: 0.0002734178400613355, 1397: 0.002716713502427387, 2677: 0.0007613339576488589, 3497: 6.882733730955611e-05, 357: 0.00012404025023037636, 1492: 4.710251456}
```

Linear Threshold Model

Es un modelo en el que la adopción de un comportamiento (o nodo) en una red depende de la suma ponderada de las influencias de los nodos vecinos.

Resultado del Modelo de Dispersión de Influencia:

```
{0: 0, 1819: 0, 2840: 0, 1: 0, 1565: 0, 1309: 0, 1397: 0, 2677: 0, 3497: 0, 357: 0, 1492: 0, 4125: 0, 1351: 0, 2: 0, 1437: 0, 1728: 0, 214: 0, 485: 0,  
2753: 0, 676: 0, 3719: 0, 3: 0, 3830: 0, 2504: 0, 3387: 0, 4528: 0, 246: 0, 1266: 0, 982: 0, 596: 0, 291: 0, 2480: 0, 1676: 0, 1450: 0, 79: 0, 2112: 0, 222: 0,
```

Dispersión de influencia sobre el top-10 nodos

Inicialmente, se crea el grafo no dirigido a partir de un conjunto de datos que describe las relaciones entre usuarios de Twitch. Se asignan pesos aleatorios a las aristas para simular la presencia de interacciones más fuertes o débiles entre los usuarios. Se define un umbral para la adopción del comportamiento en el modelo de dispersión de influencia. Se inicializa el estado de los nodos, donde cada nodo comienza en un estado no adoptado. Luego, se elige un nodo de inicio aleatorio y se establece en un estado adoptado.

Se itera sobre los nodos del grafo, calculando una suma ponderada de los estados de los nodos vecinos, teniendo en cuenta los pesos de las aristas. Si la suma supera el umbral y el nodo está en un estado no adoptado, se marca como adoptado, y viceversa.

$$S_i = \sum_{j \in N(i)} w_{ij} \cdot x_j$$

Después de calcular las medidas de centralidad (grado, cercanía, intermediación y PageRank) y ordenar los nodos según estas medidas, se seleccionan los top-10 nodos para cada medida. Finalmente, se ejecuta el modelo de dispersión de influencia para cada conjunto de top-10 nodos y se muestra el resultado, indicando qué nodos han adoptado el comportamiento bajo la influencia de sus vecinos.

Resultado del Modelo de Dispersión de Influencia (Grado)

```
{0: 0, 1819: 0, 2840: 0, 1: 0, 1565: 0, 1309: 0, 1397: 0, 2677: 0, 3497: 0, 357: 0, 1492: 0, 4125: 0, 1351: 0, 2: 0, 1437: 0, 1728: 0, 214: 0, 485: 0, 2753: 0,
```

Resultado del Modelo de Dispersión de Influencia (Closeness Centrality)

```
{0: 0, 1819: 0, 2840: 0, 1: 0, 1565: 0, 1309: 0, 1397: 0, 2677: 0, 3497: 0, 357: 0, 1492: 0, 4125: 0, 1351: 0, 2: 0, 1437: 0, 1728: 0, 214: 0, 485: 0, 2753: 0,
```

Resultado del Modelo de Dispersión de Influencia (Betweenness Centrality)

```
{0: 0, 1819: 0, 2840: 0, 1: 0, 1565: 0, 1309: 0, 1397: 0, 2677: 0, 3497: 0, 357: 0, 1492: 0, 4125: 0, 1351: 0, 2: 0, 1437: 0, 1728: 0, 214: 0, 485: 0, 2753: 0,
```

Resultado del Modelo de Dispersión de Influencia (PageRank):

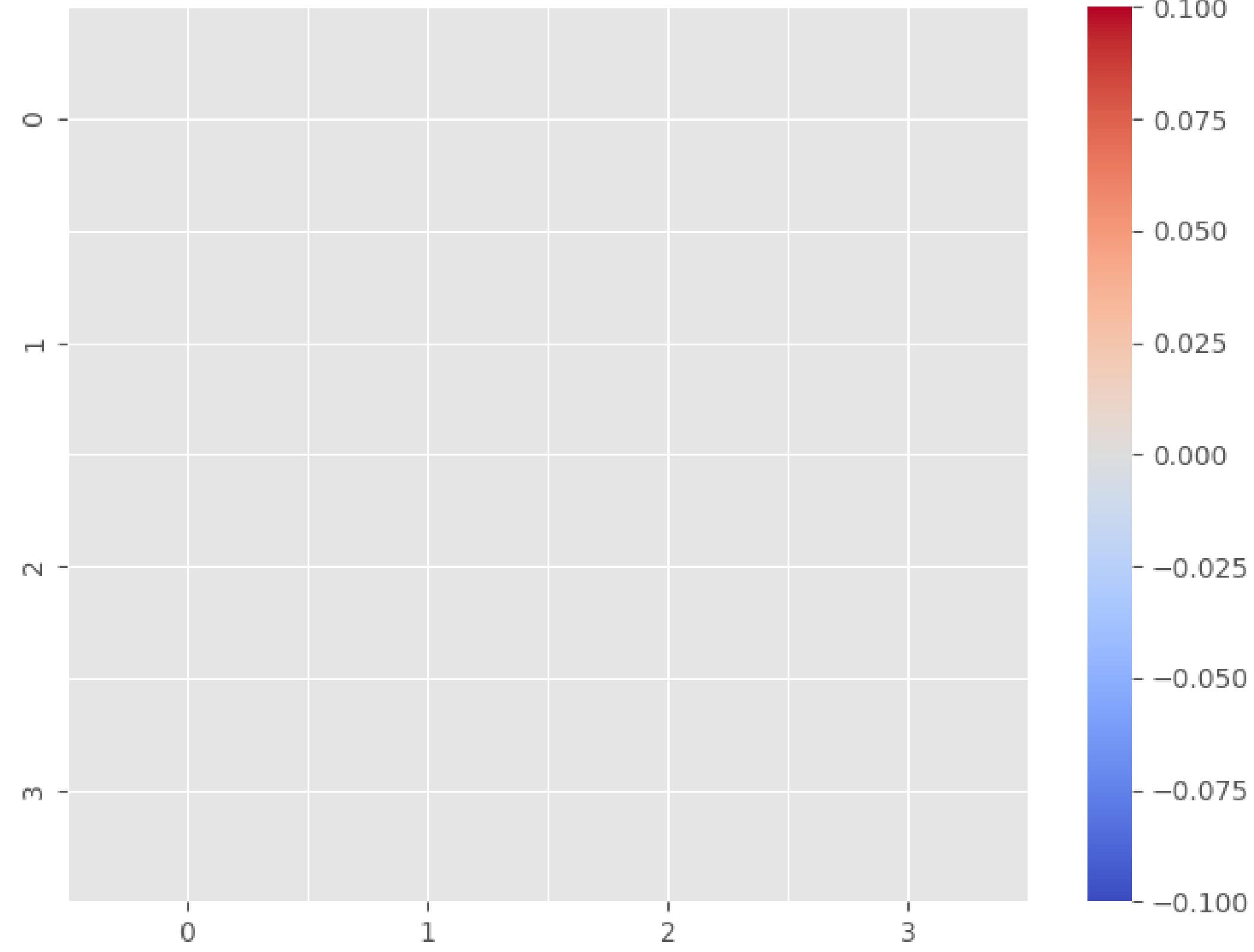
```
{0: 0, 1819: 0, 2840: 0, 1: 0, 1565: 0, 1309: 0, 1397: 0, 2677: 0, 3497: 0, 357: 0, 1492: 0, 4125: 0, 1351: 0, 2: 0, 1437: 0, 1728: 0, 214: 0, 485: 0, 2753: 0,
```

Correlación de Spearman

El código proporcionado realiza un análisis de la correlación entre los resultados obtenidos para las distintas medidas de centralidad (grado, cercanía, intermediación y PageRank) aplicadas al modelo de dispersión de influencia en un grafo no dirigido. La correlación se evalúa utilizando el coeficiente de correlación de Spearman, que es una medida estadística no paramétrica que evalúa la relación monotónica entre dos variables.

Los resultados de esta visualización permiten identificar patrones de correlación entre las diferentes medidas de centralidad, lo que ayuda a comprender cómo la influencia en la propagación del comportamiento se relaciona entre nodos según distintas métricas. Un coeficiente de correlación cercano a 1 indica una fuerte correlación positiva, mientras que un valor cercano a -1 indica una fuerte correlación negativa. Un valor de 0 indica ausencia de correlación.

Matriz de Correlaciones de Spearman (p-value < 0.05)



Interpretación de Resultados

En el análisis de datos utilizando el Modelo de Dispersión de Influencia, la obtención de valores 0 puede atribuirse a diversas razones interrelacionadas. En primer lugar, la resistencia a la adopción por parte de los nodos sugiere que, a pesar de la presencia de influencia proveniente de sus vecinos, la suma ponderada de estas influencias no es suficiente para superar los umbrales individuales de los nodos. Además, la falta de conexiones fuertes entre los nodos indica que las interacciones entre ellos carecen de la fuerza necesaria para generar un cambio generalizado en la adopción del comportamiento. Por último, la presencia de umbrales altos implica que se requiere una influencia considerable para que un nodo modifique su estado, y si la suma ponderada no alcanza este umbral, se obtendrán valores 0. En conjunto, estos factores sugieren que la red exhibe una resistencia colectiva al cambio de comportamiento, destacando la importancia de comprender la configuración de ponderaciones, la distribución de umbrales y la naturaleza de las conexiones en el análisis de la dinámica de adopción en la red.