



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

ESCUELA NACIONAL DE ESTUDIOS SUPERIORES  
UNIDAD MORELIA

APRENDIZAJE NO SUPERVISADO PARA EL ESTUDIO DE  
REDES TEMÁTICAS DE TWITTER

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

LICENCIADO EN TECNOLOGÍAS PARA LA  
INFORMACIÓN EN CIENCIAS

PRESENTA:

RODRIGO SEBASTIÁN CORTEZ MADRIGAL

TUTORES:

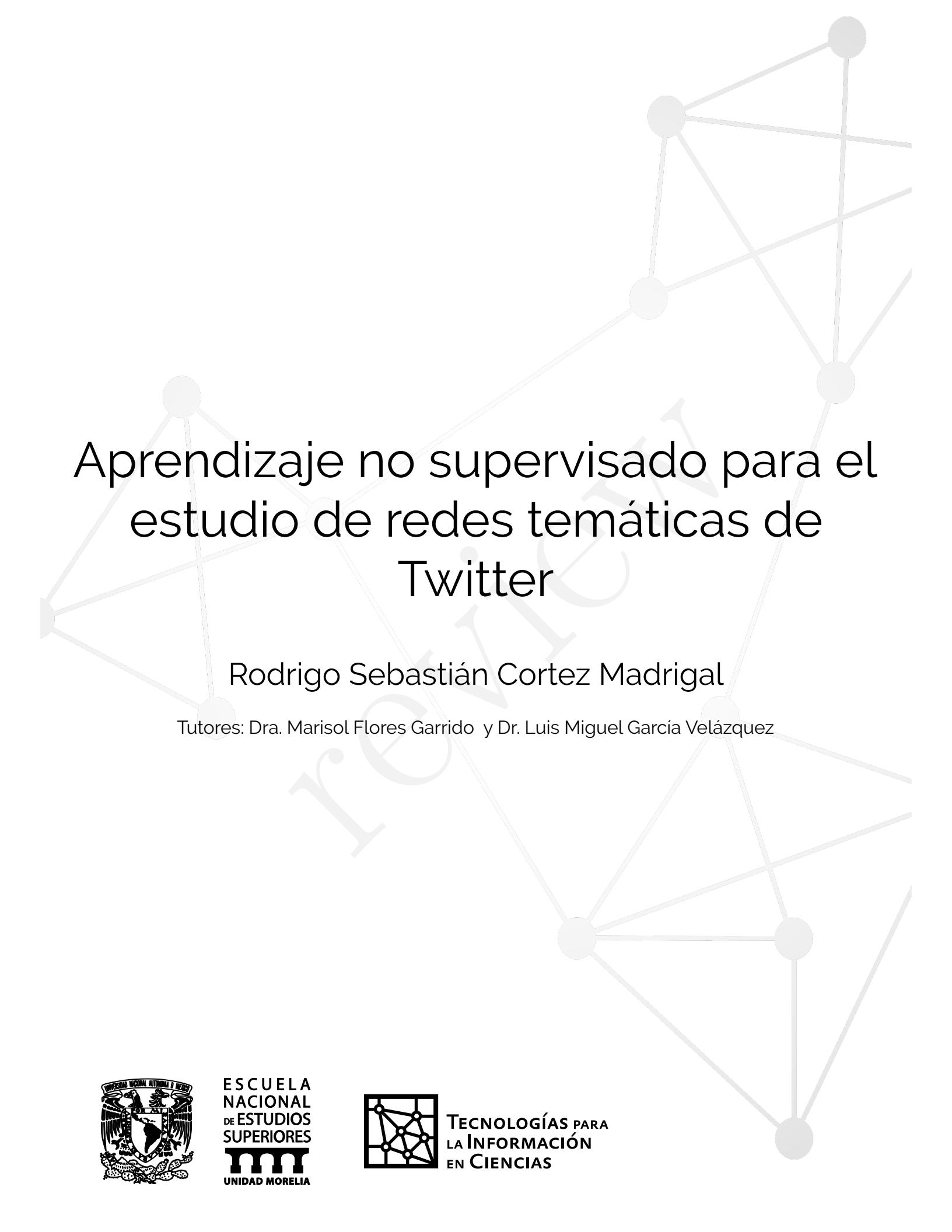
DRA. MARISOL FLORES GARRIDO  
DR. LUIS MIGUEL GARCÍA VELÁZQUEZ

ESCUELA  
NACIONAL  
DE ESTUDIOS  
SUPERIORES



Morelia, Michoacán, 2022





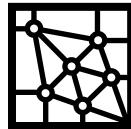
# Aprendizaje no supervisado para el estudio de redes temáticas de Twitter

Rodrigo Sebastián Cortez Madrigal

Tutores: Dra. Marisol Flores Garrido y Dr. Luis Miguel García Velázquez



ESCUELA  
NACIONAL  
DE ESTUDIOS  
SUPERIORES  
**UNIDAD MORELIA**



TECNOLOGÍAS PARA  
LA INFORMACIÓN  
EN CIENCIAS

**Rodrigo Sebastián Cortez Madrigal**

*Aprendizaje no supervisado para el estudio de redes temáticas de Twitter*

Tesis de Licenciatura. 2022

Tutores: Dra. Marisol Flores Garrido y Dr. Luis Miguel García Velázquez

**Licenciatura en Tecnologías para la Información en Ciencias**

*Universidad Nacional Autónoma de México*

Escuela Nacional de Estudios Superiores, Unidad Morelia  
Antigua Carretera a Pátzcuaro  
58000, Morelia

# Abstract

La capacidad de Twitter para conectar a los usuarios en torno a un tema determinado permite conocer los complejos mecanismos que otorgan posiciones de influencia a un subconjunto de usuarios. Este trabajo se centra en el agrupamiento de una colección de redes temáticas de Twitter mediante un enfoque interpretable centrado en las relaciones asimétricas de la plataforma. Nuestro método consiste en dos pasos generales: primero, identificamos los perfiles estructurales de los usuarios de la red a partir de una representación de la red basada en la presencia de subgrafos dirigidos de 2 a 4 nodos. A continuación, creamos *embeddings* de la red utilizando los perfiles anteriores creados y establecemos grupos dentro de la colección. Mostramos la aplicabilidad del método propuesto analizando 75 redes reales generadas en torno a *Trendings Topics* en México y discutiendo los perfiles de usuarios identificados desde el punto de vista de las dinámicas de poder social que reflejan.

**Keywords —** Graphlets, Órbitas, Embeddings, Clustering, Redes Sociales, Roles Estructurales



# Agradecimientos

” ” *Si he visto a lo lejos ha sido porque me he subido a hombros de gigantes.*

— Isaac Newton

Спасибо библиотеке Генезис за демократизацию доступа к знаниям.



# Índice general

<b>Índice de figuras</b>	<b>xiii</b>
<b>1. Introducción</b>	<b>1</b>
1.1. Twitter . . . . .	4
1.2. Agrupamiento de redes temáticas . . . . .	6
1.3. Roles estructurales y <i>graphlets</i> . . . . .	6
1.4. Presentación del problema y objetivos . . . . .	9
1.4.1. Metodología . . . . .	10
1.5. Estructura del trabajo . . . . .	11
<b>2. Agrupamiento sobre Redes</b>	<b>13</b>
2.1. Redes . . . . .	13
2.2. Aprendizaje automático ( <i>Machine Learning</i> ) . . . . .	15
2.2.1. Aprendizaje Automático Supervisado . . . . .	16
2.2.2. Aprendizaje Automático No-Supervisado . . . . .	17
2.2.3. Agrupamientos y grafos . . . . .	20
2.3. Agrupamientos de Nodos . . . . .	21
2.3.1. Agrupamientos de Grafos . . . . .	22
2.4. <i>Representation Learning y Embeddings</i> . . . . .	23
2.4.1. <i>Embedding</i> Nivel Nodo . . . . .	24
Neighbourhood-based Node Embedding . . . . .	25
Structural Node Embedding . . . . .	26
2.4.2. <i>Embedding</i> Nivel Grafo . . . . .	26
Características de la Red . . . . .	27
Características de los Nodos . . . . .	28
2.5. Interpretabilidad . . . . .	28

<b>3. Graphlets, Órbitas y Roles Estructurales</b>	<b>29</b>
3.1. Graphlets . . . . .	29
3.2. Órbitas y firma orbital . . . . .	29
3.2.1. Relevancia de las órbitas y su relación con los roles estructurales . . . . .	30
3.2.2. Conteo de órbitas . . . . .	31
3.2.3. Ejemplo Karate Club . . . . .	32
<b>4. Método propuesto</b>	<b>35</b>
4.1. Graphlets y órbitas dirigidas . . . . .	35
4.2. Perfilar usuarios . . . . .	37
4.2.1. MiniBatch K-Means . . . . .	39
4.2.2. Análisis de los perfiles identificados . . . . .	40
4.2.3. Estabilidad de los perfiles identificados . . . . .	41
4.3. Agrupar Redes . . . . .	43
4.4. Resumen . . . . .	46
<b>5. Experimentos y resultados</b>	<b>47</b>
5.1. Conjunto de datos . . . . .	47
5.2. Primer agrupamiento: perfilando usuarios . . . . .	48
5.2.1. Estabilidad . . . . .	49
5.2.2. Perfiles identificados . . . . .	49
5.3. Segundo agrupamiento: estructura en redes . . . . .	51
5.4. Visualización de resultados . . . . .	52
5.5. Discusión . . . . .	57
<b>6. Conclusiones</b>	<b>65</b>
<b>A. Apéndice</b>	<b>67</b>
A.1. Capítulo 1 . . . . .	67
Homofilia . . . . .	67
Centralidad de Intermediación . . . . .	67
A.2. Capítulo 2 . . . . .	68
Función Biyectiva . . . . .	68
A.3. Capítulo 5 . . . . .	68
Línea base . . . . .	68

review



# Índice de figuras

1.1.	Árbol de decisión para clasificar redes temáticas en Twitter, propuesto por Himmelboim <i>et al.</i> [Him+17] . . . . .	7
1.2.	Roles estructurales y su función según Kim <i>et al.</i> [Ros+16] . . . . .	8
1.3.	En un grafo se conoce como puente a los nodos que conectan dos grupos, estos nodos tienen una alta intermediación ya que necesariamente por ellos pasan los caminos más cortos entre nodos de ambos grupos. . . . .	9
2.1.	Grafo no dirigido de tres nodos y tres aristas. . . . .	14
2.2.	Grafo Dirigido (DiGraph). Podemos observar que la dirección de las aristas está representada por una flecha que indica de donde se origina la arista (inicio de la flecha) . . . . .	14
2.3.	Ejemplo de Isomorfismo entre $G$ y $H$ . . . . .	15
2.4.	Un ejemplo de Clustering en $R^2$ . . . . .	18
2.5.	Centroides . . . . .	19
2.6.	Nodos de una red divididos en 2 grupos donde el color del nodo representa el grupo al que pertenece. . . . .	22
2.7.	. . . . .	22
3.1.	Graphlets de 2, 3 y 4 nodos. . . . .	29
3.2.	Graphlets y órbitas no dirigidas de 2 a 5 nodos. . . . .	30
3.3.	Grafo dirigido de 5 nodos. . . . .	31
3.4.	Matriz de conteo de órbitas para el grafo 3.3 . . . . .	32
3.5.	Red de Karate Club [Zac77]. Los nodos más influyentes, Mr. Hi, John A. y sus respectivos vecinos a distancia 1 han sido coloreados. . . . .	33
3.6.	Comparación del conteo de órbitas normalizado para 4 usuarios de la red <i>Karate Club</i> . . . . .	34

4.1.	Órbitas de hasta 4 nodos. $G_i$ representa un graphlet en la colección; las órbitas dentro de cada graphlet están ennumeradas para futuras referencias en este trabajo. . . . .	36
4.2.	Algunos patrones propuestos por Lusher y Robins para describir configuraciones sociales dentro de procesos colectivos [Lus]. Las aristas dirigidas permiten la distinción entre jerarquías y posiciones de poder dentro de la red. . . . .	38
4.3.	Clustering Jerárquico . . . . .	44
4.4.	Resumen de metodología . . . . .	46
5.1.	Método Elbow o Codo para determinar el tamaño de K . . . . .	48
5.2.	Estabilidad del agrupamiento con K=5 utilizando Normalized Mutual Information . . . . .	50
5.3.	Perfiles encontrados . . . . .	51
5.4.	Composición de las redes de acuerdo al porcentaje de usuarios de cada perfil encontrado. . . . .	52
5.5.	Agrupamiento jerárquico utilizando <i>Ward Linkage</i> . . . . .	53
5.6.	Agrupamiento jerárquico utilizando <i>Complete Linkage</i> . . . . .	54
5.7.	Código qr para acceder a la herramienta web. . . . .	55
5.8.	Página principal de la herramienta web. . . . .	55
5.9.	Ejemplo de la visualización de la composición de una red utilizando una gráfica de rada. . . . .	56
5.10.	Ejemplo de la visualización de una red dentro de la herramienta web. . . . .	57
5.11.	Red #SalarioRosa2 coloreada respecto al grupo al que pertenece cada nodo en la red. . . . .	58
5.12.	Red #Coco coloreada respecto al grupo al que pertenece cada nodo en la red. . . . .	59
5.13.	Graphlet 0 y órbita 1. . . . .	60
5.14.	Graphlet 0 y órbita 0. . . . .	60
5.15.	Graphlet 10 y órbitas 29 y 30. . . . .	61
5.16.	Graphlet 8 y órbitas 21 a 24. . . . .	62

# Introducción

En un mundo cada vez más conectado, las interacciones de los usuarios en los espacios digitales crean una inmensidad de conexiones que a la vez reflejan complejas estructuras sociales. Estudiar estas redes y sus estructuras es de gran interés para distintas disciplinas ya que permiten extraer información valiosa que permite comprender distintos procesos sociales, como pueden ser el flujo de información, las interacciones y jerarquías entre los usuarios.

Dado que las redes sociales como Twitter generan intensos debates relacionados con cuestiones sociopolíticas clave y tienen una gran capacidad para proyectar diversos discursos en el ámbito público, es de particular interés para muchos científicos la configuración de dichas redes en Twitter. Esta plataforma de *microblogging* se ha señalado como una pieza crítica en la construcción de debates políticos y movimientos sociales [BR15] e incluso de gran influencia en la configuración de la opinión pública sobre temas de salud [Sha+22].

Una búsqueda en la plataforma especializada *Science Direct* arroja más de 32,449 artículos que involucran estudios de redes en Twitter, con ángulos que van desde los mecanismos de creación de redes, hasta los ecosistemas de poder creados en torno al flujo de información. Las redes sociales como Twitter generan intensos debates relacionados con cuestiones sociopolíticas clave y tienen una gran capacidad para proyectar diversos discursos en el ámbito público. A continuación se describen algunos ejemplos de estudios realizados sobre Twitter en distintas disciplinas.

- **Política y movimientos sociales.** Twitter ha demostrado ser un importante actor dentro de recientes movimientos sociales y políticos. Algunos ejemplos interesantes de estudios que se han hecho en Twitter son *Misinformation warnings: Twitter's soft moderation effects on COVID-19 vaccine belief echoes* donde Sharevski *et al.* estudian los efectos de la moderación

de Twitter en las creencias sobre la efectividad de las vacunas durante la pandemia de COVID-19 [Sha+22] y en *Twitter: A useful tool for studying elections?* Ivon Gaber estudia la correlación entre la actividad en Twitter y el desempeño electoral de los candidatos del Partido Laborista y el Partido de la Independencia en el Reino Unido. [Gab17]

- **Salud pública.** En cuestiones de salud pública, Twitter es un herramienta útil para modelar las concepciones que los usuarios tienen sobre ciertos temas. En específico, Han *et al.* propone una metodología para modelar las ideas y el marketing detrás del uso de cigarrillos electrónicos en Estados Unidos [HK16].
- **Economía.** En *Twitter mood predicts the stock market*, Bollen *et al.* utilizan economía del comportamiento (*behavioral economics*) y Twitter para predecir el estado de ánimo colectivo en Twitter y estudiar la correlación con el mercado de valores [BMZ11]. De manera similar, Aharon *et al.* miden el impacto de las *Twitter Uncertainty Measures (TMU & TEU)* sobre criptomonedas [Aha+22].
- **Psicología, marketing e influencers.** En *Hashtag homophily in twitter network: Examining a controversial cause-related marketing campaign*, publicado en *Computers in Human Behavior*, Sifan Xua y Alvin Zhoub estudiaron redes de campañas de marketing controversiales para analizar la tendencia a la homofilia de los usuarios que utilizaron ciertos #hashtags. Los resultados del estudio muestran que a pesar de que la discusión se dio principalmente dentro de los discursos de la campaña, los usuarios reaccionaron más fuertemente ante los influencers. Además, la red de menciones de estos usuarios mostró una tendencia a la homofilia basada en los hashtags ideológicos y no conceptuales [XZ20].
- **Lingüística, noticias y fake news.** En 2020, Medford *et al.* analizaron los sentimientos colectivos en Twitter sobre la pandemia de COVID-19. La mitad de los Tweets expresaron miedo mientras que un tercio expresó sorpresa. Al analizar los Tweets más retuiteados, el contenido se enfocaba en las formas de transmisión, los esfuerzos de prevención y la cuarentena, mientras que el miedo disminuía. En la cohorte completa, el impacto económico y político de COVID-19 fue el tema más discutido [Med+20].

Los procesos por los que las *fake news* se diseminan y afectan la conversación pública también pueden ser estudiados en Twitter. En *Modeling the spread of fake news on Twitter* se propone que las noticias falsas se diseminan como una noticia ordinaria hasta que los usuarios se dan cuenta de la falsedad y eso se convierte en otra noticia [Mur+21].

En 2017, Himelboim *et al.* [Him+17] se enfocaron en el estudio de redes temáticas en Twitter. Es decir, analizaron las interacciones que surgen entre usuarios de la plataforma cuando se aborda un tema específico. Su trabajo no utiliza aprendizaje automático, pero propone una serie de reglas que les permite caracterizar diferentes redes temáticas. Este problema es interesante porque busca distinguir, en un conjunto de redes, las distintas configuraciones estructurales que pueden surgir. De manera intuitiva, los autores tratan de establecer similitudes y diferencias entre redes, de forma que puedan compararlas y crear grupos.

Debe señalarse que el problema de agrupamiento de redes implica distintos retos computacionales. Debido a la naturaleza de los grafos, no se puede utilizar directamente los métodos convencionales de aprendizaje automático, como *K-Means* [Béj], sobre los mismos; es necesario primero crear una representación vectorial. Además, tratándose de un trabajo de exploración, la representación debería poder interpretarse para que los resultados tengan significado para especialistas en otras áreas.

En este trabajo de tesis se propone una metodología que, organizada en dos etapas principales, permite estudiar redes temáticas en Twitter a partir de sus estructuras locales utilizando como base la idea de órbitas [Sar+16] en graphlets. Dichas órbitas corresponden a los roles de nodos en la colección de todos los posibles grafos de cierto orden dado (típicamente se consideran sólo 2-5 nodos), conocidos como *graphlets* y originados en estudios de bioinformática [Prz07]. Con estas órbitas, que se describen con detalle más adelante, este trabajo construye una representación vectorial (*embedding*) con el objetivo final de realizar un agrupamiento que tome en cuenta los roles estructurales de usuarios.

Es importante mencionar que dicha metodología ha sido aceptada en distintos congresos y será publicada en la *Mexican Conference on Pattern Recognition (Proceedings)*.

En el resto de este capítulo se presenta una descripción de términos importantes relacionados con Twitter. Después, se motiva el estudio de redes temáticas con una perspectiva de roles estructurales. Finalmente, se establecen los objetivos y la metodología de esta investigación.

## 1.1 Twitter

Cada medio digital en el que usuarios interactúan define los canales y las estructuras del flujo de información. Tanto las estructuras de flujo como las jerarquías sociales en una plataforma reflejan patrones interesantes que nos permiten entender la relación que existen entre las mismas. Uno de los ejemplos más claros dentro de los medios digitales y las redes sociales donde se dan este tipo de interacciones y jerarquías es Twitter. Twitter es un servicio de *microblogging* y red social en la que los usuarios publican e interactúan con posts conocidos como “tweets” [Twi].

Un tweet es la unidad mínima de Twitter, se trata de mensajes de hasta 280 caracteres, son públicamente visibles por defecto y cualquier usuario puede responder a los demás, creando de esta manera una discusión pública que se puede modelar con una red dirigida.

La forma en que se propaga la información en Twitter se asemeja a cómo se propaga la información en la vida real. Las comunicaciones humanas suelen caracterizarse por una asimetría entre los productores de información (medios de comunicación, empresas, personas influyentes, entre otros) y los consumidores de contenidos [GRL14]. El papel de los usuarios en la propagación de la información a través de la red está intrínsecamente relacionado con la topología de la misma. Entender estos roles puede proporcionar una valiosa visión de los debates públicos en la plataforma.

A continuación se describen algunos términos relevantes para analizar el funcionamiento de Twitter.

Trending Topics. Twitter hace un seguimiento de las frases, palabras y hashtags que se mencionan con mayor frecuencia y los publica bajo el título de "Trending Topic". Un hashtag es una etiqueta por convención entre los usuarios de Twitter para crear y seguir un hilo de discusión prefijando una palabra con el símbolo "#". Los Trending Topics ayudan a Twitter y a sus usuarios a entender lo que está ocurriendo en la red social e invitarles a unirse a la discusión [Twi]. Los Trending Topics se representan filtrados por país dependiendo de la configuración de la cuenta y son calculados en tiempo real a lo largo del día.

Interacciones. La mayor parte de las interacciones dentro de Twitter corresponden a la práctica común de responder o reaccionar a un tweet [Kwa+10]. Las más comunes están definidas por las siguientes acciones:

- RT que la abreviatura "*retweet*" es la práctica de replicar el tweet de otro usuario. El mecanismo de *retweet* permite a los usuarios difundir la información que deseen más allá del alcance de los seguidores del tweet original.
- '@' seguido de un identificador (username) se refiere a una mención y se utiliza para etiquetar y responder directamente a un usuario.

Red temática. Una red temática es aquella que captura las interacciones anteriormente mencionadas dentro de un tema específico definido por un Trending Topic. Es decir, los nodos de la red representan usuarios que han escrito un tweet sobre un tema en tendencia (TT) y las aristas representan las interacciones entre ellos, ya sea un RT o una mención. Es importante mencionar que las aristas son dirigidas y representan el sentido de la interacción.

## 1.2 Agrupamiento de redes temáticas

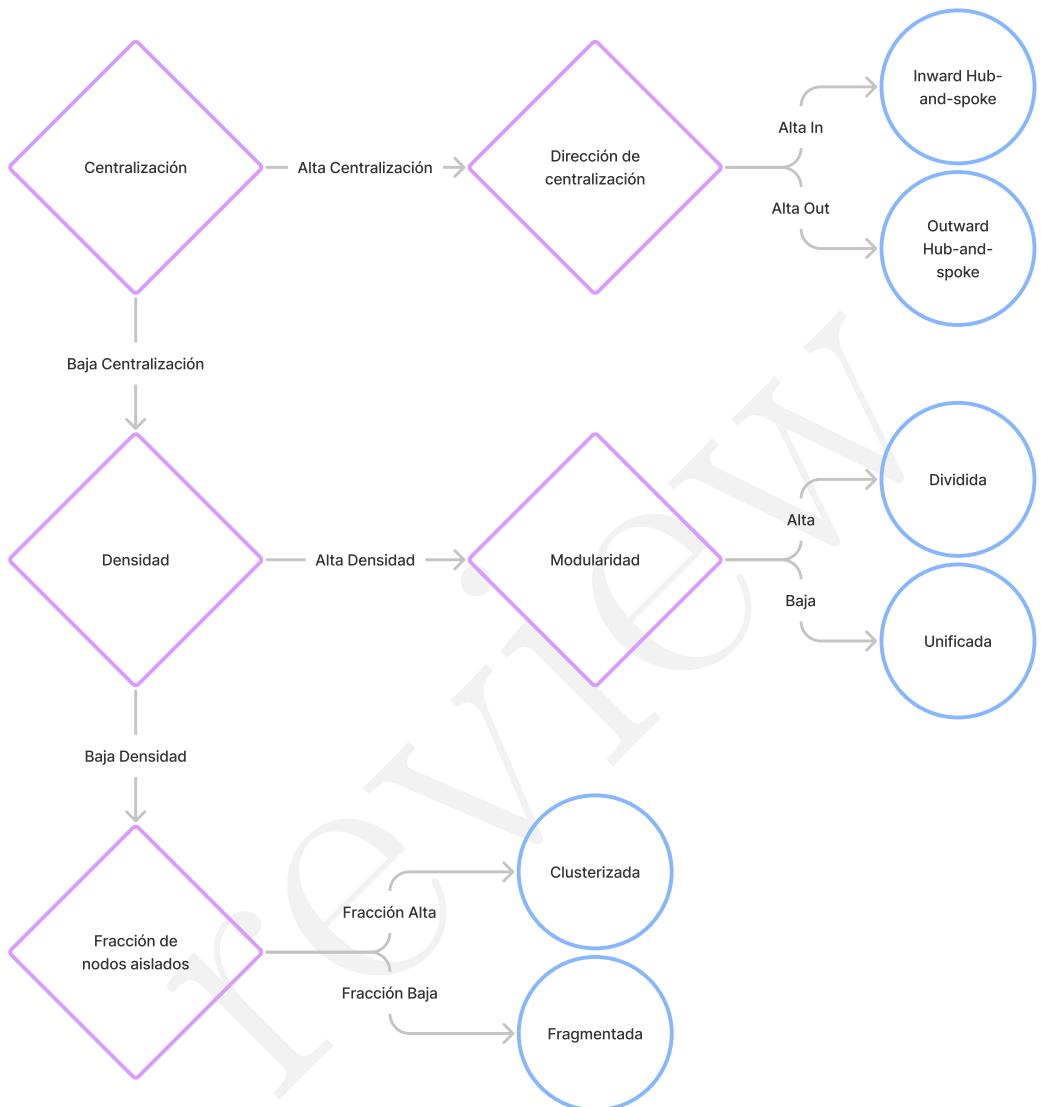
Como se mencionó anteriormente, las redes temáticas pueden ser interesantes ya que contiene la configuración estructural de la discusión pública sobre un tema en específico. Con esta motivación, Himelboim *et al.* propusieron un estudio de redes temáticas usando criterios que ellos mismos definieron con base en su experiencia desde el campo de la sociología.

En su trabajo, estos autores hacen clasificación, aunque no en el sentido de aprendizaje automático, pues no utilizan datos etiquetados ni siguen una metodología basada en los datos. Más bien proponen que hay 6 clases importantes para el estudio de redes, que son: dividida, unificada, fragmentada, clusterizada, *in hub-and-spoke* y *out hub-and-spoke*. Después, utilizando distintas medidas de las redes crean un árbol de decisión para clasificar cada una en los grupos predefinidos, como se puede observar en la Fig. 1.1.

Aunque este trabajo se considera una aportación importante al estudio de redes en Twitter, utilizar grupos predefinidos podría llevar consigo algunos problemas, como limitar la clasificación a sólo las categorías concebidas por los autores, desestimando otros criterios que permitirían diferenciar entre redes. Preguntas interesantes que pueden plantearse a partir de este trabajo son: ¿Es posible llevar a cabo un agrupamiento basado directamente en los datos? ¿De qué forma puede hacerse si además se requiere que los resultados sean interpretables? Quizá los algoritmos de aprendizaje automático no-supervisado para agrupamiento no son directamente una opción, pero extrayendo características de las redes para crear un *embedding* podría ser una alternativa viable.

## 1.3 Roles estructurales y graphlets

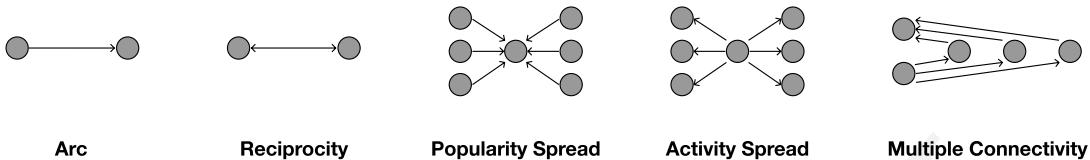
Los roles estructurales han sido estudiados por distintas disciplinas desde hace algunos años. Un rol estructural en redes puede entenderse como las funciones que tiene un nodo dentro de un grafo. La importancia de estos roles



**Fig. 1.1.:** Árbol de decisión para clasificar redes temáticas en Twitter, propuesto por Himelboim *et al.* [Him+17]

estructurales reside en su correlación con las estructuras y jerarquías sociales así como su comportamiento.

Desde distintas disciplinas se ha intentado mapear las estructuras en grafos a estructuras sociales. En *Understanding Network Formation in Strategy Research* [Ros+16] se estudia la composición de la redes dentro del contexto de investigación sobre gestión estratégica y cómo estas impactan directamente dentro de las organizaciones (Ver Fig.1.2).



**Fig. 1.2.:** Roles estructurales y su función según Kim *et al.* [Ros+16]

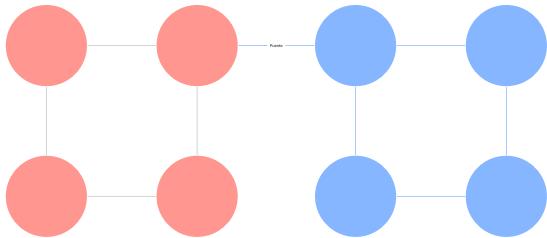
Otro ejemplo muy interesante es el de *Structural Holes and Good Ideas* [Bur04], donde se describe el mecanismo por el que la intermediación influye directamente en el capital social. Esto debido en gran parte a que la opinión y el comportamiento son más homogéneos dentro de los grupos que entre todos ellos, por lo que las personas que conectan grupos (puentes) están más familiarizadas con formas alternativas de pensar y comportarse.

En la Fig. 1.3 podemos observar un ejemplo en el que encontramos un puente entre dos grupos. Estos nodos (*A* y *B*) también pueden ser encontrados en la literatura con el nombre de *brokers* y tienen una alta intermediación. La centralidad de intermediación es una medida de centralidad en grafos basada en los caminos más cortos. Formalmente se define como

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

dónde  $\sigma_{st}$  es el número total de caminos más cortos desde el nodo *s* al nodo *t* y  $\sigma_{st}(v)$  es el número de esos caminos que pasan por *v* (donde *v* no es un nodo final)

Dada la relevancia que, en sociología, ha tenido el análisis de roles estructurales, en este trabajo exploramos la posibilidad de agrupar redes temáticas en Twitter basándonos en la idea de dichos roles. Para ello, utilizamos las órbitas de



**Fig. 1.3.:** En un grafo se conoce como puente a los nodos que conectan dos grupos, estos nodos tienen una alta intermediación ya que necesariamente por ellos pasan los caminos más cortos entre nodos de ambos grupos.

graphlets, que son diccionarios de grafos de orden fijo, descritos con mayor detalle en el capítulo 3.

## 1.4 Presentación del problema y objetivos

Los objetivos de este trabajo son los siguientes:

Objetivo general. Dada una colección de redes de Twitter definidas por la interacción de los usuarios sobre temas concretos (redes temáticas), agrupar redes dentro de la colección según el perfil de los usuarios que conforman cada red, tomando como base el rol estructural de los usuarios.

### Objetivos específicos

OE1 Crear una colección de redes temáticas en Twitter en México.

OE2 Identificar perfiles de usuarios en las redes mediante una representación vectorial a nivel nodo, basada en la firma orbital de graphlets.

OE3 Construir una representación vectorial para las redes temáticas basada en la caracterización de usuarios y roles estructurales, y usarla para agrupar las redes en la colección.

## 1.4.1 Metodología

OE1 Crear una colección de redes temáticas en Twitter en México.

- a) Determinar un criterio para elegir temas que permitan la construcción de redes.
- b) Descargar tweets con los criterios previamente determinados de tal manera que las redes temáticas puedan ser construidas.
- c) Preprocesar los datos y construir las redes a partir de la discusión pública.
- d) Guardar las redes en un formato apropiado para trabajar con la colección.

OE2 Identificar perfiles de usuarios en las redes mediante una representación vectorial a nivel nodo, basada en la firma orbital de graphlets.

- a) Calcular los graphlets y la firma orbital de cada nodo para cada red
- b) Llevar a cabo clustering usando la firma orbital de los nodos que se calculó en el paso anterior.
- c) Identificar los distintos perfiles de usuario que se distinguen de acuerdo a la firma orbital.

OE3 Construir una representación vectorial para las redes temáticas basada en la caracterización de usuarios y roles estructurales, y usarla para agrupar las redes en la colección.

- a) Representar cada red de acuerdo al tipo de usuarios que emergen en la conversación.
- b) Agrupar las redes temáticas utilizando la representación anterior, de modo que puedan identificarse grupos basados en un criterio interpretable: el rol estructural de los usuarios.

## 1.5 Estructura del trabajo

En el capítulo 2 revisaremos algunos conceptos útiles relacionados con el agrupamiento en grafos. Después en el capítulo 3, se discutirán los graphlets y las órbitas que pueden definirse a partir de ellos. Tomando como base los capítulos 2 y 3, el capítulo 4 describe la la metodología propuesta. Posteriormente, en el capítulo 4 se exponen los resultados de los experimentos realizados. Finalmente, en el capítulo 5 encontramos las conclusiones del trabajo, algunas consideraciones del mismo y el trabajo futuro propuesto.



# Agrupamiento sobre Redes

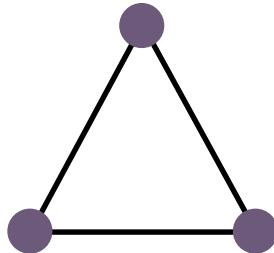
En este capítulo se presentará el problema principal y los elementos necesarios para comprender la complejidad del mismo y una posible solución. Comenzaremos con una definición formal de un red así como su representación matemática, posteriormente se presenta el problema de agrupamiento y su relevancia dentro del aprendizaje automático. Finalmente este capítulo se enfoca en analizar las diferencias y retos de las tareas de agrupamiento en redes y agrupamiento sobre redes, así como de los diferentes enfoques y limitaciones para resolver estos problemas. Los elementos claves se presentan en esta sección sin embargo existe un apéndice que podría ayudar a resolver dudas adicionales.

## 2.1 Redes

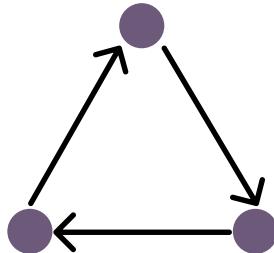
Una red es un conjunto de nodos unidos por aristas que representan relaciones. Los nodos y aristas los podemos encontrar en distintas disciplinas con distintos nombres, por ejemplo en física se denominan sitios y vínculos y en sociología actores y vínculos.

La representación matemática de un red se denomina grafo y es estudiada en matemáticas discretas y más específicamente en teoría de grafos. Un grafo esta formalmente definido como un par ordenado de conjuntos  $G = (V, E)$ , donde  $V$  es un conjunto de nodos (vértices) y  $E$  aristas (edges). [Sao21]

Un grafo dirigido o digrafo es un grafo en el que las aristas tienen direcciones. En el sentido más formal un grafo dirigido es una tripleta  $G = (V, E, \phi)$  donde  $\phi$  es una función de incidencia que asigna cada arista a un par ordenado de nodos, es decir,  $\phi : E \rightarrow \{(x, y) \mid (x, y) \in V^2 \text{ and } x \neq y\}$



**Fig. 2.1.:** Grafo no dirigido de tres nodos y tres aristas.

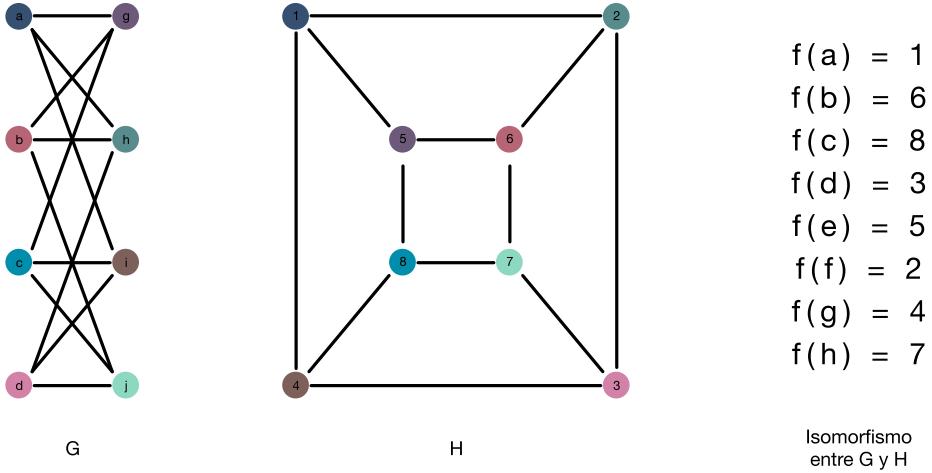


**Fig. 2.2.:** Grafo Dirigido (DiGraph). Podemos observar que la dirección de las aristas esta representada por una flecha que indica de donde se origina la arista (inicio de la flecha)

Un subgrafo  $H$  de un grafo  $G$  es otro grafo formado a partir de un subconjunto de nodos y un subconjunto de aristas de  $G$ . El subconjunto de nodos debe incluir todos los extremos del subconjunto de aristas, pero también puede incluir otros nodos. Un subgrafo inducido  $H$  de un grafo  $G$  es aquel que incluye todas las aristas del grafo  $G$  cuyos puntos extremos pertenecen al subconjunto de nodos que define al subgrafo  $H$

Un isomorfismo de grafos es una biyección de los nodos de un grafo sobre otro, de modo que se preserva la adyacencia de los nodos. Formalmente, el isomorfismo entre dos grafos  $G$  y  $H$  es una función biyectiva que se define de la siguiente manera  $f : V(G) \rightarrow V(H)$ .

Determinar si dos grafos con el mismo número de vértices  $n$  y aristas  $m$  son isomorfos o no, se conoce como el problema del isomorfismo de grafos. Se considera a este problema un problema NP ya que no hay prueba de que sea NP-Completo. [KST93] [Ver Apéndice] Por otro lado, se trata de un caso especial del problema de isomorfismo de subgrafos que sí se ha demostrado que es un problema NP-Completo. Resolver el problema de isomorfismo de grafos requeriría probar si las  $n!$  biyecciones posibles preservan la adyacencia,



**Fig. 2.3.:** Ejemplo de Isomorfismo entre  $G$  y  $H$

sin embargo hasta ahora no se conoce un algoritmo general para resolver el problema y por lo tanto se considera un problema no resuelto dentro de la computación.

## 2.2 Aprendizaje automático (Machine Learning)

El Aprendizaje Automático o Aprendizaje de Máquinas, en inglés Machine Learning (ML), es una rama de la Inteligencia Artificial que estudia algoritmos y técnicas que permitan automatizar soluciones a problemas complejos a partir del aprendizaje sobre conjuntos de datos en vez de los métodos convencionales de programación.

Dentro de la Inteligencia Artificial, que es un campo de estudio muy amplio y utiliza distintas técnicas para crear algoritmos inteligentes, el Aprendizaje Automático se enfoca principalmente en imitar el aprendizaje humano y gradualmente mejorar la precisión sobre una tarea [IBM]. En la programación convencional dados ciertos requerimientos se diseña un programa que siga una serie de pasos para resolver un problema. No obstante en problemas complejos, a pesar de tener requerimientos claros y específicos este enfoque puede resultar complicado para crear y programar grandes conjuntos de re-

glas, pensemos por ejemplo en la tarea de detectar objetos en una imagen [RRC19].

Los algoritmos de Aprendizaje Automático son capaces de resolver problemas de una manera un tanto más genérica aprendiendo estructuras y reglas a partir de un conjunto de datos en vez de tener una estructura y diseño explícito. Por esta razón este tipo de algoritmos dependen directamente de la calidad y cantidad de ejemplos en el conjunto de datos. Estos ejemplos pueden tener etiquetas o ser datos crudos y dependiendo de la naturaleza del conjunto de datos encontraremos distintas categorías de algoritmos dentro del Aprendizaje Automático [RRC19]. Un conjunto de datos etiquetado es aquel cuyos ejemplos tienen la respuesta a la pregunta que se hace. Podemos pensar al conjunto de datos etiquetado como una guía de estudio, a partir de la cual el estudiante (en este caso la máquina) puede aprender a partir de ejemplos. Un claro ejemplo es el caso de la tarea de clasificación, para la cual el conjunto de datos contiene información sobre la clase a la que representa cada objeto, por ejemplo, una imagen de un perro contiene la etiqueta perro. Por otro lado los datos no etiquetados (crudos) son aquellos que no contienen una etiqueta, es decir que no ha sido preprocesados de ninguna manera y de los cuales no poseemos más información que el dato en sí. Los datos no etiquetados son aquellos que podemos recolectar de un sensor a partir de observaciones de algún entorno.

### 2.2.1 Aprendizaje Automático Supervisado

El objetivo del Aprendizaje Supervisado es crear un modelo sobre un conjunto de datos etiquetados para posteriormente poder predecir las etiquetas de datos crudos [RRC19]. La manera en la que estos algoritmos resuelven problemas es a partir de un modelo generado aprendiendo (con un entrenamiento) sobre un conjunto de datos con etiquetas conocidas (ejemplos) que después se ejecuta sobre nuevos datos para predecir su etiqueta. Durante la fase de entrenamiento el conjunto de datos etiquetados es dividido, una parte del conjunto se utiliza para que el algoritmo aprenda (como una guía con ejemplos) y a la vez otra parte más pequeña es utilizada para poner a prueba el entrenamiento (como un examen). Una vez que el entrenamiento se ha completado y el modelo se ha

ajustado a los datos, este es capaz de etiquetar datos nuevos que no se habían visto previamente durante el entrenamiento.

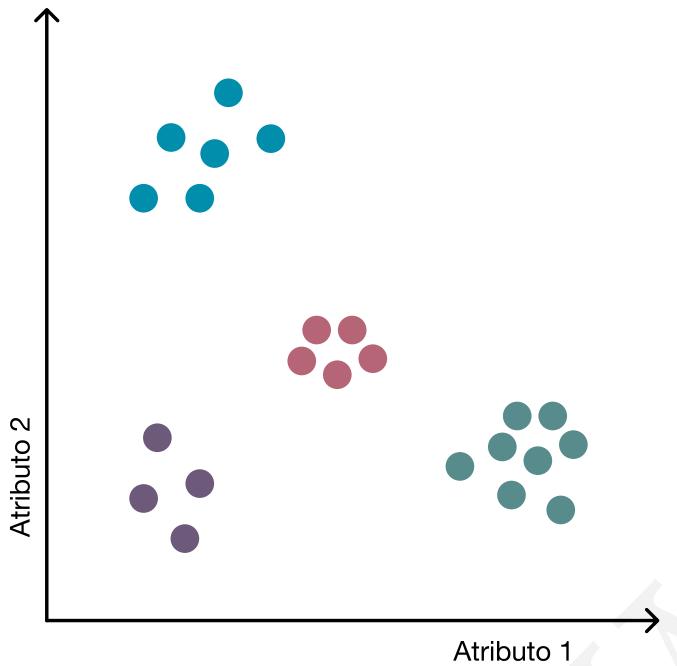
Estos algoritmos tienden a ser más efectivos que los modelos creados por humanos ya que pueden considerar mas atributos sobre un ejemplo y pueden procesar una cantidad superior de datos, no obstante una consideración importante es que no siempre es clara la manera en la que el problema esta siendo resuelto y por lo tanto es complicado interpretar el modelo [RRC19].

## 2.2.2 Aprendizaje Automático No-Supervisado

En el Aprendizaje no Supervisado el objetivo principal es aprender patrones a partir de conjuntos de datos no etiquetados. Dentro del Aprendizaje Automático No-Supervisado existen tareas como la identificación de patrones frecuentes, creación reglas de asociación y búsqueda de agrupamientos [Kub17]. En este capítulo nos enfocaremos en la tarea de agrupar.

**Agrupamientos** La tarea quizás mas representativa de Aprendizaje No-Supervisado es el *Clustering* o Agrupamiento, se trata de dividir un gran conjunto de datos (puntos) de tal manera en que los puntos con propiedades o patrones en común se encuentren en un mismo grupo. La complejidad de esta tarea radica en que los grupos no se conocen previamente y la cantidad de los mismos es desconocida. Posteriormente los resultados de esta tarea pueden ser utilizados como clasificadores o predictores de valores de atributos desconocidos, e incluso como herramientas de visualización. [Kub17]

Un ejemplo sencillo de agrupamiento en  $R^2$  puede ser el de la Fig. 2.4. Aquí cada punto representa un ejemplo descrito por dos atributos. En este caso es sencillo encontrar los agrupamientos a simple vista (ojímetro), sin embargo para cuatro dimensiones o más, no es posible para los humanos visualizar los datos ni los grupos; estos casos solo pueden ser resueltos por los algoritmos de agrupamiento. [Kub17]

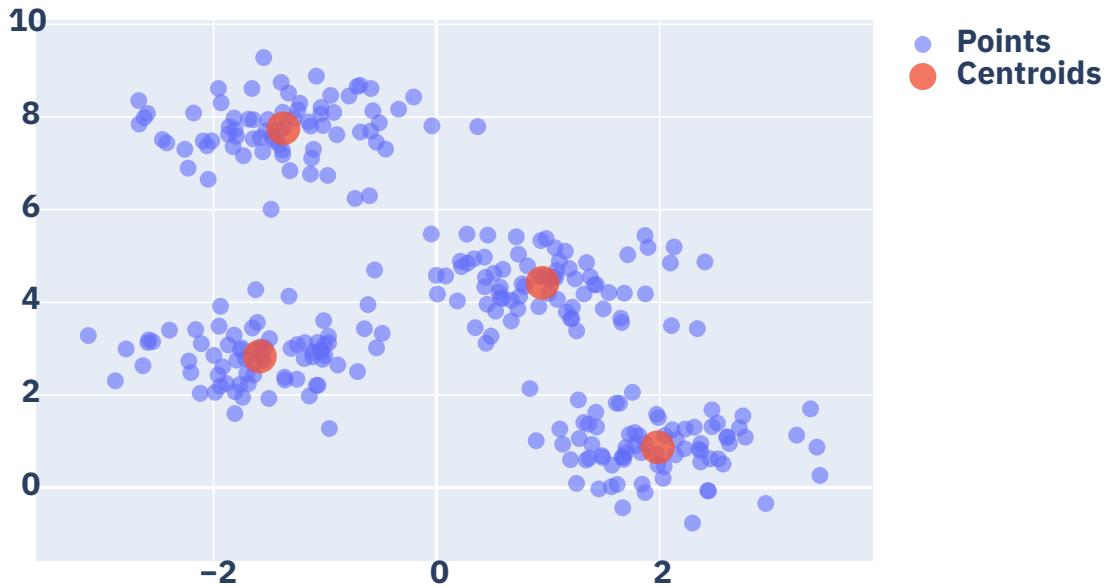


**Fig. 2.4.:** Un ejemplo de Clustering en  $R^2$

Los algoritmos de agrupamiento frecuentemente requieren definir una función de distancia entre un ejemplo y todos los elementos del grupo. Dependiendo de la naturaleza de los atributos distintas medidas pueden ser propuestas; cuando se trata de puntos numéricos en el espacio comúnmente se utiliza la distancia Minkowski.  $X = (x_1, x_2, \dots, x_n)$  y  $Y = (y_1, y_2, \dots, y_n) \in R^n$

$$D(X, Y) = \left( \sum_{i=1}^n |x_i - y_i|^p \right)^{\frac{1}{p}}$$

Quizás uno de los algoritmos más conocidos de Agrupamiento es *K-Means*. Este algoritmo agrupa los datos de entrada en  $K$  grupos para una  $K$  predefinida por el usuario. Como cada ejemplo no incluye una etiqueta de la clase o grupo al que pertenece se trata inherentemente de un algoritmo de Aprendizaje No-Supervisado. La representación matemática de los  $K$  grupos se conoce como *centroide*, que es el punto promedio de la distancia a cada punto del grupo que representa. La interpretación de cada grupo, representado por su centroide, puede ser que su valor promedio es la caracterización de todos los elementos del grupo.



**Fig. 2.5.:** Centroides

Por lo tanto, el algoritmo de *K-Means* busca minimizar la distancia promedio de cada centroide a los puntos de su grupo. De manera formal, el criterio de optimización es minimizar el error cuadrático total entre los ejemplos de entrenamiento y sus centroides correspondientes. La función objetivo se conoce como *inertia* o *within-cluster sum-of-squares criterion*.

$$\arg \min_{\mathbf{S}} \sum_{i=1}^k \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2$$

Dado un conjunto de ejemplos  $(x_1, x_2, \dots, x_n)$  donde cada ejemplo es un vector  $d - dimensional$ , K-Means busca agrupar los  $n$  ejemplos en  $K ( \leq n )$  grupos  $S = S_1, S_2, \dots, S_k$  de tal manera que se minimice la suma de las distancias cuadradas para cada grupo.

Los centroides pueden ser inicializados de manera aleatoria o con algunas técnicas de inicialización que permitan al algoritmo converger más rápido. El algoritmo se itera recalculando los centroides y los puntos correspondientes hasta que ya no hay un cambio significativo o se ejecuta el número máximo de iteraciones. Uno de los aspectos negativos de este algoritmo es que es muy

susceptible a las condiciones iniciales y por lo tanto se recomienda ejecutar el algoritmo varias veces para quedarse con el mejor resultado o promediarlos.

#### List. 2.1: Pseudocódigo $K - Means$ [Kub17]

```
1 Input: Puntos y K
Output: K Grupos

5 Inicializar K Centroides y K Grupos \cite{kubat_introduction_2017}
repetir:
    Escoger un punto  $X$  y calcular su distancia a cada
    Centroide  $C_i$ 
    Notar el centroide más cercano como  $C_j$ 
    Si  $X$  se encuentra en  $C_j$  no hacer nada.
    De otra manera mover  $X$  a  $C_j$  y recalcular los centroides.

10 hasta:
    Para cuando los Centroides no cambien de manera
    significativa
    O se haya alcanzado un número máximo de iteraciones
```

### 2.2.3 Agrupamientos y grafos

A pesar de que existen numerosos algoritmos de agrupamiento, estos no pueden ser utilizados directamente en grafos. Debido a la representación y la dificultad de encontrar una función de distancia o similitud entre dos nodos o dos grafos, encontrar un agrupamiento no es un problema trivial.

Debido a las aplicaciones y a la necesidad de extraer información de este tipo de estructuras de datos, recientemente se han propuesto numerosos algoritmos para hacer agrupamientos en grafos. La primera tarea que podemos encontrar dentro de estos algoritmos es la detección de grupos o comunidades dentro de la misma red. La segunda tarea que es un tanto más compleja y ha sido menos explorada es la de realizar agrupamientos a nivel grafo, es decir dentro de una colección de grafos agrupar aquellos que tengan características en común.

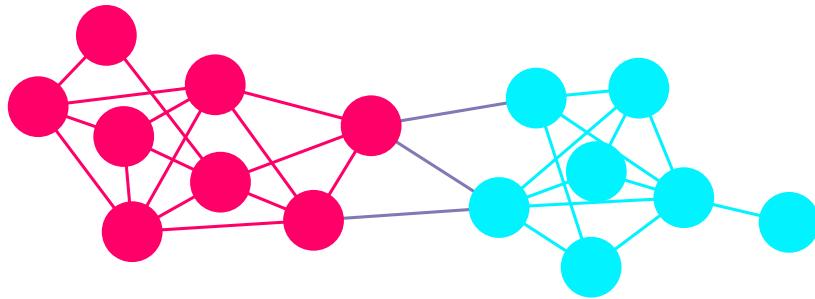
## 2.3 Agrupamientos de Nodos

Realizar agrupamientos de nodos dentro de la red ha sido un problema ampliamente explorado en años recientes debido a la gran cantidad de aplicaciones. Este problema se subdivide en dos conocidos como partición de grafos y detección de comunidades, ambos problemas se refieren a la división de los nodos de una red en grupos, clusters o comunidades según el patrón de aristas de la red.[\[New10\]](#) Podemos diferenciar uno del otro a partir de si conocemos previamente el número de grupos para la agrupación (partición) o es parte del problema y es desconocido (detección de comunidades).

La partición de grafos es un problema estudiado desde 1960 [\[New10\]](#) y se trata de dividir los nodos de un grafo en  $n$  grupos de tal manera que las aristas que entre ellos sean las menores posibles, a este número de aristas entre cada grupo se llama Tamaño de Corte (Cut Size).

El caso más sencillo se llama Bisección en el que se divide la red en dos grupos y así recursivamente. A pesar de que es bastante sencillo de entender este problema, no es nada fácil de resolver. La idea mas intuitiva quizás podría ser analizar todas las particiones posibles de la red en dos grupos, calcular el tamaño del corte y quedarse con aquella que tenga el menor (fuerza bruta). No obstante esto es computacionalmente costoso para grafos muy grandes ya que las posibles particiones en dos grupos  $n_1yn_2$  para una red de  $n$  nodos es de  $\frac{n!}{n_1!n_2!}$  [\[New10\]](#) por lo que encontrar la solución óptima es complicado, distintos algoritmos heurísticos que aproximen una solución óptima han sido ampliamente estudiados. Uno de los mas algoritmos heurísticos más sencillos para resolver la bisección de un grafo es el algoritmo *Kernighan-Lin*.

En el caso del problema de Detección de Comunidades, un reto adicional es el encontrar también el número y tamaño de grupos adecuado. La Detección de Comunidades busca grupos que ocurren naturalmente en la estructura de una red independientemente de la cantidad de grupos y el número de nodos en ellos. Estos algoritmos nos permiten descubrir y estudiar la estructura y organización de una red independientemente de su naturaleza.



**Fig. 2.6.:** Nodos de una red divididos en 2 grupos donde el color del nodo representa el grupo al que pertenece.

### 2.3.1 Agrupamientos de Grafos

Como se mencionaba anteriormente, un problema aún más complejo, menos estudiado y por lo tanto, un reto más grande, es el de agrupar grafos completos. Comparar propiedades estructurales entre redes muy complejas es un problema importante con distintas aplicaciones científicas, sin embargo también es un problema computacionalmente costoso.



**Fig. 2.7.**

En general agrupar dos grafos requiere de dos pasos, una función de distancia que nos permita comparar grafos entre sí y un algoritmo de agrupamiento que haga uso de estas distancias para asignar cada grafo a un grupo determinado. Dentro de las aproximaciones populares para comparar dos grafos podemos

encontrar el Isomorfismo de Grafo, la Distancia de Edición, el Alineamiento de Redes y la Extracción de Características.[SKB19]. Idealmente el Isomorfismo de Grafo sería la aproximación más adecuada, no obstante como vimos en 2.1 se trata de un problema NP y por lo tanto existen limitaciones mas que considerables en la práctica.

A pesar de que existen distintas medidas estructurales y de distancia, no suelen haber sido pensadas teniendo en mente la tarea de agrupamiento de grafos. Algunos ejemplos que podemos encontrar en la literatura al respecto son los siguientes.

**Medidas Estructurales** Se han propuesto distintas medidas estructurales para capturar patrones dentro de una red. Este trabajo parte de hecho de las ideas propuestas en *Classifying Twitter Topic-Networks Using Social Network Analysis* [Him+17] En este trabajo se utilizan medidas como la Centralidad, la Centralización, la Densidad, la Modularidad y la Fracción de *Clusters* e *Isolates* para clasificar redes enteras dependiendo de sus características estructurales. Aplicando esta metodología al conjunto de datos con el que se trabajó inicialmente nos dimos cuenta de que existían ciertas limitaciones respecto a la información que capturan estas medidas estructurales sobre la red.

**GED** La Distancia de Edición (Graph Edit Distance, GED) puede ser una de las alternativas más conocidas para comparar dos grafos. La GED mide el número de cambios necesarios para llegar a la estructura del grafo *B* partiendo desde el grafo *A*.

## 2.4 Representation Learning y Embeddings

Como discutíamos anteriormente los algoritmos clásicos de Aprendizaje de Máquina 2.2 no puede ser utilizados directamente para realizar agrupamientos en redes debido a las dificultades anteriormente mencionadas. Una de las estrategias recientes para resolver este problema es extraer características

de los nodos o el grafo entero y utilizarlas para crear de una representación vectorial y de esta manera poder utilizar medidas clásicas de distancia en este espacio y algoritmos clásicos de aprendizaje de máquina. Este proceso de extracción de características es llamado Representation Learning y a la representación de las mismas *embedding*, pero para el fin de este documento utilizaremos estos dos términos de manera intercambiable.

El *embedding* puede obtenerse para cada nodo o para representar un grafo entero. Existen numerosas técnicas para representar un nodo, puede ser a partir de basado en su vecindario, por su estructura topológica o sus atributos. En el caso de el *embedding* de un grafo la extracción de características puede dividirse en dos principales categorías: técnicas basadas en la topología global de la red y técnicas basadas en subestructuras a nivel de los nodos de la red.

### 2.4.1 Embedding Nivel Nodo

Los *embedding* a nivel nodo han sido ampliamente explorados en años recientes, en gran medida debido a los retos que enfrentan las grandes tecnológicas para perfilar enormes cantidades de usuarios dentro de distintas redes [Ler+19] y por muchas más aplicaciones, por ejemplo en el área de biología y química. A continuación se presentan algunos conceptos útiles para comprender muchos algoritmos para este tipo de tarea.

**Random Walks** Una Caminata Aleatoria (Random Walk) es un proceso estocástico en el espacio matemático que describe una trayectoria de pasos aleatorios. Las Caminatas Aleatorias son utilizadas para analizar y simular procesos aleatorios así como para calcular correlaciones entre los objetos de estudio [XK19]. En Grafos, las Caminatas Aleatorias permiten calcular la distancia entre nodos y extraer características de la topología. Cada paso en la trayectoria se da de acuerdo a cierta distribución de probabilidad, esta probabilidad de transición entre nodos es un factor relevante para la magnitud de su correlación. Es decir, mientras mas asociados se encuentran dos nodos, mayor es su probabilidad de transición. Uno de los algoritmos más famosos que hace uso de esta técnica

es PageRank, que hace caminatas aleatorias dentro del grafo de páginas web para calcular la importancia de cada una de ellas.

## Neighbourhood-based Node Embedding

Esta familia de algoritmos extrae características de la vecindad de un grafo para obtener una representación. Existen distintas técnicas para obtener un vector como *embedding*, a continuación una breve reseña de algunos algoritmos y los métodos que utilizan.

Neighbourhood-Based Node Embedding	
Paper	Algoritmo
“Relational Learning via Latent Social Dimensions”	SocioDim
“Billion-scale Network Embedding with Iterative Random Projection”	RandNE
“GLEE: Geometric Laplacian Eigenmap Embedding”	GLEE
“Diff2Vec: Fast Sequence Based Embedding with Diffusion Graphs”	Diff2Vec
“NodeSketch: Highly-Efficient Graph Embeddings via Recursive Sketching”	NodeSketch
“Network Embedding as Matrix Factorization: Unifying DeepWalk LINE PTE and Node2Vec”	NetMF

Neighbourhood-Based Node Embedding	
“Multi-Level Network Embedding with Boosted Low-Rank Matrix Approximation”	BoostNE
“Don’t Walk, Skip! Online Learning of Multi-scale Network Embeddings”	Walklets
“GraRep: Learning Graph Representations with Global Structural Information”	GraRep
“DeepWalk: Online Learning of Social Representations”	DeepWalk
“node2vec: Scalable Feature Learning for Networks”	Node2Vec
“Alternating Direction Method of Multipliers for Non-Negative Matrix Factorization with the Beta-Divergence”	NM-FADMM
“Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering”	LaplacianEigenmaps

## Structural Node Embedding

Structural Node Embedding	
Paper	Algoritmo
“Learning Structural Node Embeddings Via Diffusion Wavelets”	GraphWave
“Learning Role-based Graph Embeddings”	Role2Vec

### 2.4.2 Embedding Nivel Grafo

En general los algoritmos que hacen uso de *embeddings* para agrupar redes siguen cuatro pasos principales que se describen a continuación.

- Extracción de características: Se extraen patrones o características de la estructura topológica de los grafos a agrupar.

- Agregación de características: Se agregan estas características a los vectores que caracterizarán el grafo para de esta manera componer los *embeddings* de los grafos.
- Cálculo de la distancia: Calcular la distancia entre los vectores de los grafos para cuantificar la similitud entre los mismos.
- Agrupar grafos: Agrupar los grafos más cercanos.

De las características a extraer dentro de un grafo se pueden extraer características de la red completa o del estructuras locales dentro de ella.

## Características de la Red

Este tipo de propiedades se centran en la topología general de la red y tratan de extraer características globales. Los algoritmos enfocados a este tipo de extracción de características buscan compilar y resumir estructuras importantes dentro de la red utilizando por ejemplo,

Graph Embedding	
Paper	Algoritmo
“Graph2Vec: Learning Distributed Representations of Graphs”	Graph2Vec
“Hunt For The Unique, Stable, Sparse And Fast Feature Learning On Graphs”	FGSD
“A Simple Baseline Algorithm for Graph Classification”	SF
“NetLSD: Hearing the Shape of a Graph”	NetLSD
“GL2vec: Graph Embedding Enriched by Line Graphs with Edge Features”	GL2Vec
“Geometric Scattering for Graph Data Analysis”	GeoScattering
“Invariant Embedding for Graph Classification”	IGE

## Características de los Nodos

Este tipo de propiedades a diferencia de las propiedades globales de la red se centran en características locales entre los nodos que la conforman. Los algoritmos que extraen características de los nodos examinan las estructuras locales haciendo uso de subgrafos, por ejemplo, EgoNetworks o Graphlets.

## 2.5 Interpretabilidad

A pesar de que los algoritmos de Aprendizaje de Máquina funcionan muy bien, muchas veces es difícil conocer las reglas que el algoritmo ha creado en el modelo aprendido y por lo tanto no es posible comprender como es que un problema esta siendo resuelto, a este problema se le conoce como el problema de interpretabilidad de un modelo. [Zha+21] [RRC19]

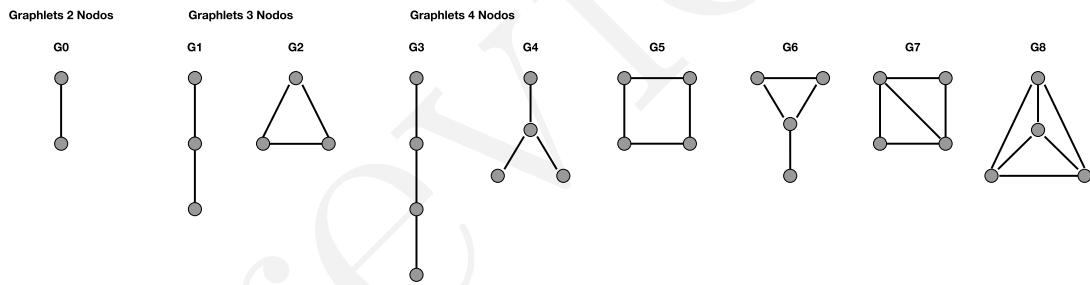
Este problema esta especialmente presente en las Redes Neuronales que tienen millones de parámetros y por lo tanto es complicado interpretar las decisiones o la serie de reglas que llevan a cabo para resolver un problema. En algunas áreas es igual de importante la interpretabilidad del modelo que su precisión, un ejemplo claro es el de la medicina en donde los médicos deben ser capaces de interpretar y confirmar los resultados del diagnóstico de un algoritmo. En el contexto de Twitter es igualmente importante interpretar los motivos por los que las redes son agrupadas.

Recientemente se han adaptado las Redes Neuronales para trabajar con grafos. Algoritmos como Pytorch:BiGraph [Ler+19] son extremadamente eficientes a la hora de obtener *embeddings* para nodos en redes enormes. No obstante al igual que otras algoritmos de esta clase heredan las problemáticas de interpretación de las redes neuronales.

# Graphlets, Órbitas y Roles Estructurales

## 3.1 Graphlets

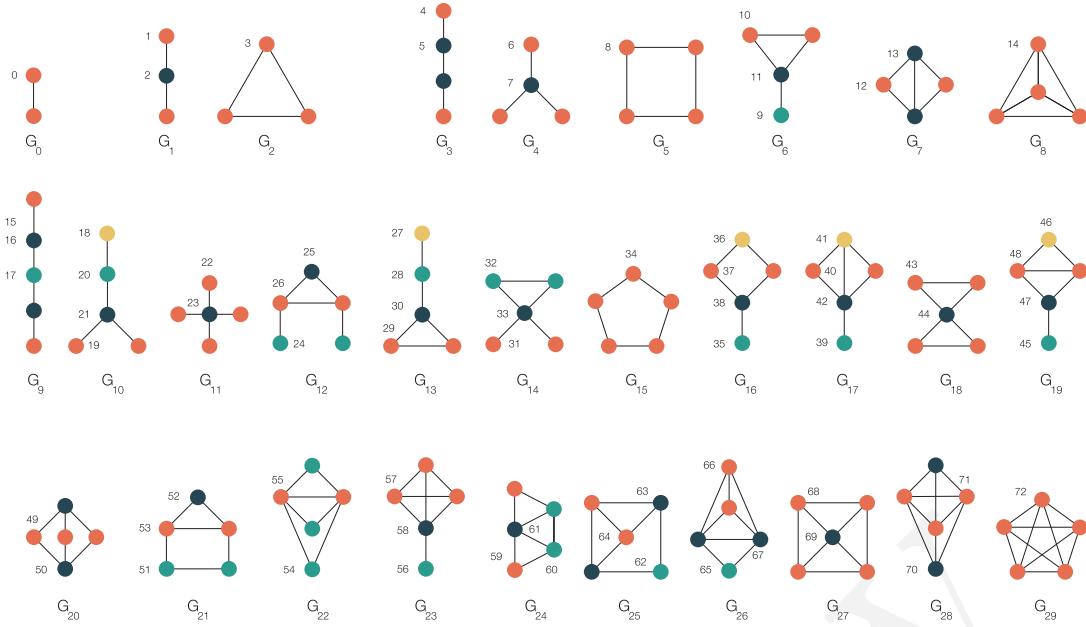
Los *graphlets* pueden ser entendidos como una colección o diccionario de todas las clases de isomorfismo de subgrafos de hasta un tamaño fijo  $n$ . Estos subgrafos pueden identificarse de manera inducida en un red más grande. En teoría de grafos, un subgrafo inducido de un grafo  $G$  se conforma a partir de un subconjunto de vértices de  $G$  y de todas las aristas incidentes a pares de vértices del subconjunto [Prz07].



**Fig. 3.1.:** Graphlets de 2, 3 y 4 nodos.

## 3.2 Órbitas y firma orbital

Los graphlets pueden ser extendidos a la noción de órbitas, que son las clases de nodos que resultan de la acción de los grupos de simetría de los graphlets, es decir, las posiciones posibles que un nodo puede tomar en un graphlet dirigido. De manera formal, las órbitas son grupos de simetría de nodos (automorfismos)



**Fig. 3.2.:** Graphlets y órbitas no dirigidas de 2 a 5 nodos.

[Sar+16] que describen los distintos roles topológicos en los que un nodo puede participar dentro del graphlet. (Ver Fig. 4.1)

Una firma orbital es el conteo de las posiciones orbitales de un nodo; el vector resultante de los conteos es la firma orbital del nodo. Este vector describe la topología del nodo y su vecindario y captura sus interconexiones hasta una distancia  $n$  [Sar+16].

Originalmente las firmas orbitales se utilizaron en el contexto de biología para analizar redes e identificar grupos de nodos topológicamente similares que por lo tanto compartieran propiedades biológicas [MP08], de tal manera que se pudieran predecir propiedades biológicas de nodos no caracterizados.

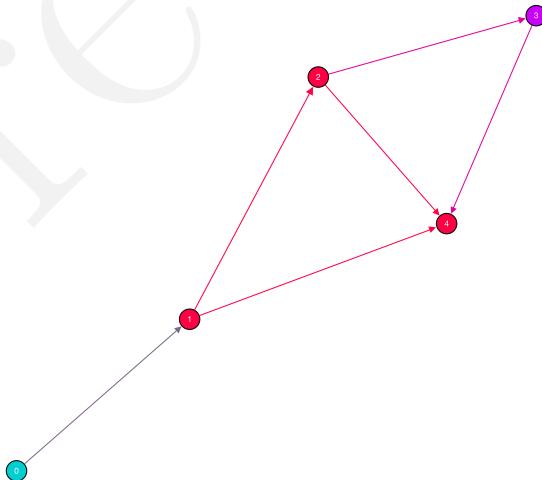
### 3.2.1 Relevancia de las órbitas y su relación con los roles estructurales

Los graphlets fueron introducidos por primera vez dentro del contexto biológico con la idea de comparar grafos. Milenkovic *et al.* crearon un diccionario de todos

los posibles subgrafos con 2-5 nodos considerando las clases de isomorfismo [MP08]. Mientras que una comparación basada en la distribución de grado cuenta el número de nodos unidos a las aristas  $i$ , los autores comparan los grafos contando el número de nodos unidos al graphlet  $i$  [Sar+16]. No obstante, la relevancia de los graphlets y las órbitas van más allá de la mera utilidad para comparar grafos, recientemente se ha sugerido que la presencia, o ausencia, de ciertas estructuras locales dentro de una red podría tener un impacto crítico en la estructura general de la red [Lus].

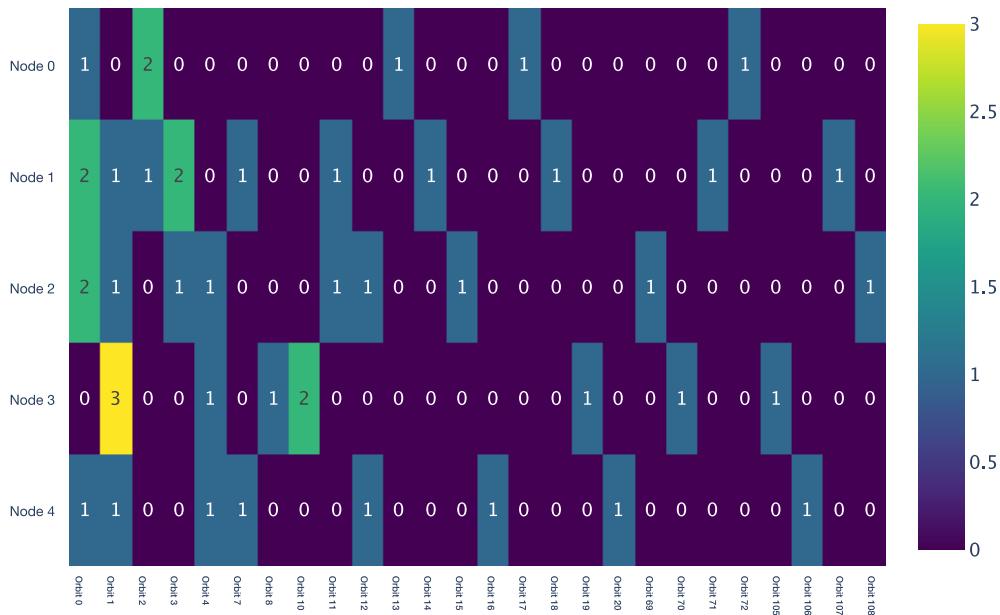
### 3.2.2 Conteo de órbitas

Relacionado con el conteo de graphlets en una red, Sarajlíc et al. (2016) [Sar+16] propusieron hacer el conteo de órbitas para un nodo. El conteo se representa en forma de un vector en  $\mathbb{R}^n$  donde el componente  $n$  representa la cantidad de veces que el nodo  $u$  aparece en la órbita  $n$ . Las órbitas permiten distinguir los diferentes roles que puede tener un nodo dentro de un mismo graphlet, por lo tanto el cálculo de la firma orbital de un nodo proporciona información sobre su grado generalizado, basado en el graphlet, y sobre las diferentes formas en las que interactúa con sus vecinos.



**Fig. 3.3.:** Grafo dirigido de 5 nodos.

En la Fig. 3.3 encontramos un grafo dirigido para el cual identificamos las órbitas en las que aparecen sus nodos, la matriz de órbitas lo podemos encontrar en la siguiente figura 3.4.

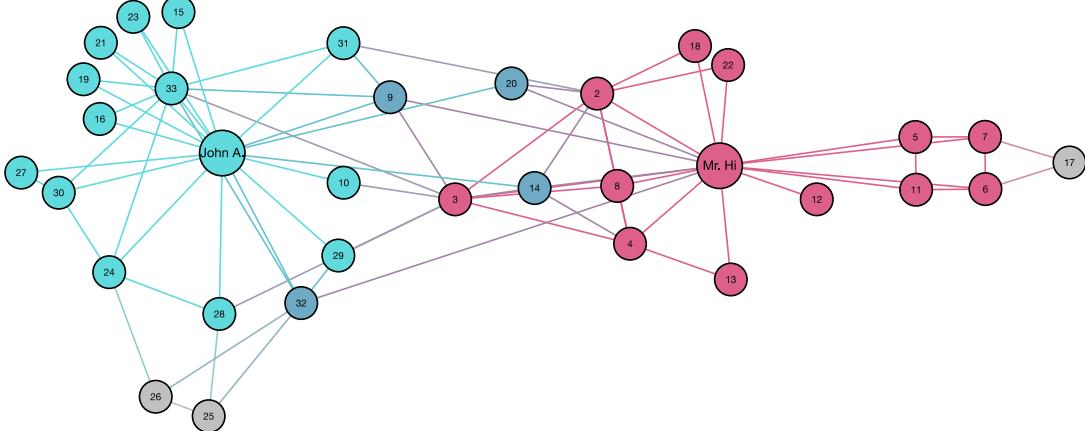


**Fig. 3.4.:** Matriz de conteo de órbitas para el grafo 3.3

### 3.2.3 Ejemplo Karate Club

La red Karate Club estudiada por Wayne W. Zachary en [Zac77] describe las interacciones de 34 miembros de un club de karate de 1970 a 1972, periodo durante el cual surgió un conflicto entre el administrador John A. y el instructor Mr. Hi. y el club se dividió en dos grupos al rededor de cada uno de ellos. Esta red (Ver Fig. 3.5) se convirtió en un estándar para el estudio de algoritmos y a menudo se utiliza como referencia.

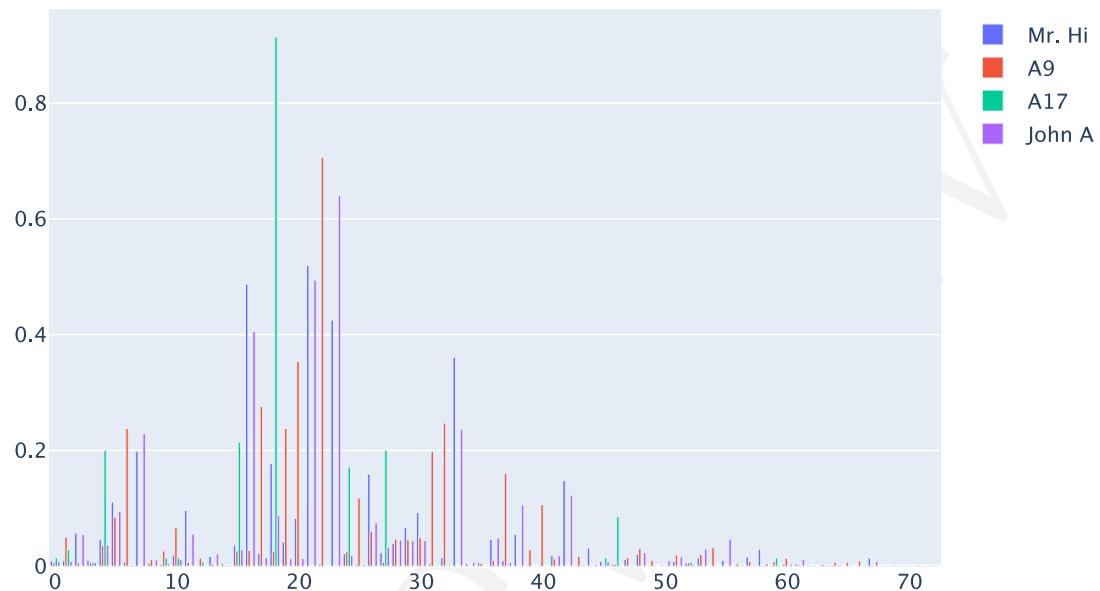
En la Fig. 3.6 podemos observar los *embeddings* para 4 nodos de la red Karate Club. Para contrastar los *embeddings* de distintos tipos de nodos en la red



**Fig. 3.5.:** Red de Karate Club [Zac77]. Los nodos más influyentes, Mr. Hi, John A. y sus respectivos vecinos a distancia 1 han sido coloreados.

tomamos como referencia a los más influyentes, Mr. Hi y John A., y a los menos conectados, los nodos 9 y 17. En el caso de los nodos más influyentes podemos observar que comparten órbitas dominantes, que son las órbitas 16, 21 y 23 descritas en la Fig. 3.2. Por otro lado las órbitas dominantes del nodo 9 son las 17, 20 y 22 y del nodo 17 son las 4, 15 y 18. Es importante notar que las órbitas 16, 21 y 23 son órbitas centrales pertenecientes a graphlets de 5 nodos.

Mediante este ejemplo observamos que la firma orbital de los nodos de una red puede ser una herramienta útil para diferenciar los roles en los que participan. En el caso de una red social, este rol puede referirse a distintas jerarquías sociales y niveles de influencia en el flujo de la información.



**Fig. 3.6.:** Comparación del conteo de órbitas normalizado para 4 usuarios de la red *Karate Club*.

## Método propuesto

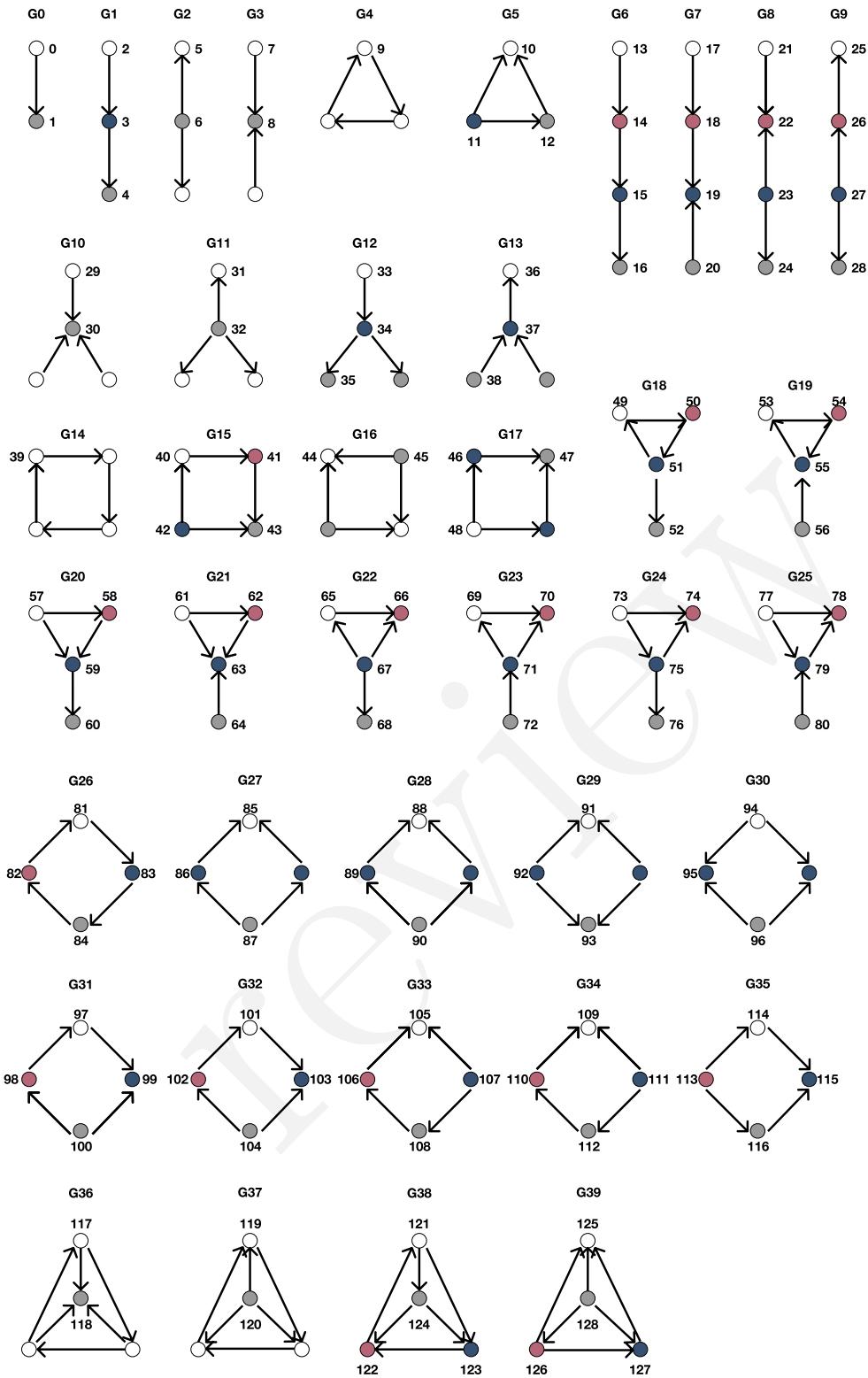
El agrupamiento de una colección de grafos no es un problema sencillo. El uso de algoritmos de agrupación populares, como K-Means, requiere representar los grafos en un espacio vectorial. Esta tarea puede llevarse a cabo mediante métodos que van desde la extracción de características - por ejemplo graphlets - hasta *embeddings* más sofisticados generados a través de redes neuronales.

Priorizando la interpretación de los resultados, proponemos usar el conteo de órbitas en graphlets dirigidos para hacer una caracterización de los usuarios en la colección analizada y crear un *embedding* de las redes que brinde información sobre el tipo de comportamiento que genera un determinado tema.

En este capítulo se presenta un método para agrupar redes a través de la firma orbital de sus nodos y que, así, toma en cuenta los roles estructurales de los usuarios. El método tiene dos etapas principales. Primero construye perfiles de usuarios utilizando la firma de la órbita asignada a cada nodo en un análisis de la red basado en graphlets. Después, agrupa las redes con base en la distribución de perfiles que presentan.

### 4.1 Graphlets y órbitas dirigidas

A partir de los conteos de órbitas vistos en el capítulo anterior, Sarajilic *et al.* propusieron extender las órbitas a grafos dirigidos [Sar+16]. Dada la cantidad de posibles configuraciones para las órbitas en un graphlet dirigido, los autores limitan el conteo a graphlets de hasta 4 nodos. En este caso, la firma orbital resultante para cada nodo es un vector en  $R^{129}$  donde el componente  $i$  representa el conteo de la órbita  $i$ , de acuerdo a la descripción presentada en la Fig. 4.1.



**Fig. 4.1.:** Órbitas de hasta 4 nodos.  $G_i$  representa un graphlet en la colección; las órbitas dentro de cada graphlet están enumeradas para futuras referencias en este trabajo.

## 4.2 Perfillar usuarios

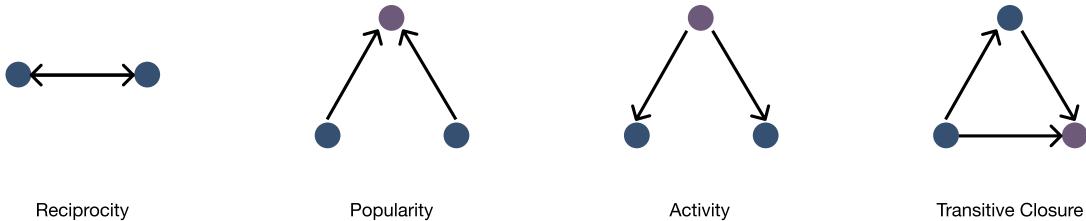
La creación de perfiles de usuario (*user profiling*) ha tenido numerosas aplicaciones dentro y fuera de las ciencias computacionales. Existen metodologías que permiten encontrar perfiles de usuario a partir de minería de datos en redes sociales, de modo que los perfiles representan ciertos rasgos psicológicos con sus conductas asociadas y permiten, entre otras cosas, campañas de marketing dirigidas [Hu20]. Estos métodos para crear perfiles o grupos de usuarios comúnmente se basan en los metadatos de las interacciones entre usuarios.

Como se ilustró en la sección 2.3, es posible realizar agrupamientos en una red basados en los diversos roles estructurales de los nodos que la componen. Esta tarea nos permite agrupar los distintos comportamientos de los usuarios a partir del papel que desempeñan y, por lo tanto, crear perfiles de usuarios con comportamientos y funciones en la red similares.

En las redes sociales, específicamente en Twitter, las funciones y las interacciones que realiza un nodo dentro de una red inciden directamente en la composición y topología de la misma. Estudiar roles estructurales permite caracterizar nodos de acuerdo a su función, obtener información sobre los tipos de comportamiento de los usuarios y estudiar la composición de la red.

De hecho, diferentes trabajos en ciencias sociales se centran en los patrones de asociación en una red para entender los procesos dentro de un sistema. Por ejemplo, Lusher y Robins sugieren la presencia de configuraciones a lo largo de las líneas de "huellas arqueológicas" impresas en los mecanismos sociales a través del tiempo y ejemplifican su idea sugiriendo los arreglos mostrados en la Fig. 4.2.

Una propuesta importante de esta tesis es que, en el contexto de redes sociales, la firma orbital basada en graphlets que puede obtenerse para un nodo, podría analizarse como extensión del trabajo de Lusher y Robins. Es decir, proponemos considerar estructuras que van más allá de las triadas de usuarios, con el fin de capturar información sobre las dinámicas sociales, la jerarquía que se establece entre personas y la estructura general de la red.



**Fig. 4.2.:** Algunos patrones propuestos por Lusher y Robins para describir configuraciones sociales dentro de procesos colectivos [Lus]. Las aristas dirigidas permiten la distinción entre jerarquías y posiciones de poder dentro de la red.

Así, consideramos que en el caso de Twitter es posible identificar perfiles de usuarios similares dentro de las redes temáticas. Debido a la capacidad de las órbitas de capturar información sobre las posiciones y roles estructurales de un nodo dentro de una red, proponemos agrupar los nodos (usuarios) de la red temática utilizando la firma orbital como un *embedding* para crear perfiles de usuarios.

Las redes temáticas de Twitter son redes con aristas dirigidas. En el estudio propuesto de las órbitas, consideramos aquellas que aparecen en graphlets de orden 2-4, de acuerdo al trabajo de Sarajilic *et al.* descrito en la sección anterior. Por lo tanto, al realizar el conteo de órbitas dirigidas para cada nodo, se obtiene una matriz  $M$  de tamaño  $n_{users} \times 129$ , en donde cada fila representa un nodo en la red.

Identificar los distintos perfiles a partir de la matriz  $M$  requiere una tarea de agrupamiento. Aunque K-Means (Algoritmo 2.1) es conveniente por motivos como la interpretabilidad, el volumen de datos en nuestro problema demanda un método más eficiente, considerando que se desea analizar la representación vectorial de todos los usuarios en todas las redes en la colección. Por esta razón, proponemos el uso de MiniBatch K-Means [Scu10], que es una de las distintas modificaciones de K-Means propuestas para lidar con limitaciones de tiempo y memoria. El algoritmo se describe a continuación.

### 4.2.1 MiniBatch K-Means

A pesar de la enorme popularidad de K-Means por su simplicidad y buen desempeño, el algoritmo se ve limitado frente a la cada vez más grande cantidad de datos a analizar. Esto se debe a restricciones como tener que mantener todo el conjunto de datos en memoria.

MiniBatch K-Means [Scu10] es una versión modificada de K-Means que busca reducir la complejidad computacional del algoritmo original utilizando únicamente una fracción del conjunto de datos en cada iteración. Esta estrategia reduce el número de cálculos de distancias por iteración y por lo tanto la complejidad total, pero con un costo asociado de un agrupamiento de menor calidad [Béj].

La idea principal de MiniBatch K-Means es utilizar pequeños lotes (mini batches) aleatorios de un tamaño fijo del conjunto de datos para poder almacenarlos en la memoria. En cada iteración se obtiene una nueva muestra aleatoria del conjunto de datos y se utiliza para actualizar los grupos (clusters) hasta la convergencia.

MiniBatch K-Means hace uso de una tasa de aprendizaje que disminuye con el número de iteraciones. La tasa de aprendizaje es inversa del número de ejemplos asignados a un grupo durante el proceso y por lo tanto a medida que aumenta el número de iteraciones se reduce el efecto de nuevos ejemplos. La convergencia del algoritmo se puede detectar cuando no se producen más cambios en los grupos durante un número definido de iteraciones continuas.

El Algoritmo 4.1 muestra el pseudocódigo de MiniBatch K-Means y sus particularidades, entre ellas el muestreo  $M$  de ejemplos aleatorios y el cálculo de la función objetivo (distorsión).

**List. 4.1:** Pseudocódigo *MiniBatchK – Means* [Béj]

```
1 Input: Puntos (X), K, Tamaño del MiniBatch (b), iteraciones (t)
Output: K Grupos
5 Inicializar K Centroides y K Grupos de manera aleatoria.

for i en rango(t):
    M = b ejemplos aleatorios de X
```

```

10      for x in M:
11          d[x] = f(C,x)
12      end
13      for x in M:
14          c = d[x]
15          v[c] = v[c] + 1
16          \eta = 1/v[c]
17          c = (1-\eta)c+\eta x
18      end
19  end

```

## 4.2.2 Análisis de los perfiles identificados

Para caracterizar el rol de los usuarios, consideramos las propiedades topológicas de las órbitas dominantes de los grupos y el papel que desempeñan en el graphlet al que pertenecen. También proponemos revisar las órbitas ausentes en los grupos, es decir, las órbitas ausentes en todos los usuarios del grupo.

Algunas definiciones serán útiles para interpretar el rol que desempeñan las órbitas en un graphlet específico. Es conveniente recordar que cada arista dirigida indica una relación entre dos nodos, donde el nodo inicial representa a un usuario que ha mencionado, respondido o retuiteado al usuario representado por el nodo final.

- Grado de entrada: Para un nodo de un graphlet, el número de arcos dirigidos que comienzan en él se denomina grado de entrada (*indegree*) de  $n$ . Se denota como  $\deg - (n)$
- Grado de salida: El número de arcos dirigidos que terminan en el nodo de un graphlet es su grado de salida (*outdegree*). Se denota como  $\deg + (n)$ .
- Fuente: Un nodo  $n$  tal que  $\deg - (n) = 0$ .
- Pozo: Un nodo  $n$  tal que  $\deg + (n) = 0$ .
- Camino dirigido en un graphlet: Una secuencia finita de aristas que une una secuencia de distintos nodos de tal manera que todas las aristas tenga la misma dirección. Es fácil observar que cada camino maximal en un graphlet comienza en una fuente y termina en un pozo.

Dado que el grado de entrada y el grado de salida de un nodo son invariantes bajo un automorfismo, podemos extender las definiciones de fuente y de pozo de los nodos a las órbitas.

Podemos decir una órbita fuente  $\mathcal{O}$  es un oyente (*listener*) si para cada nodo  $n \in \mathcal{O}$ , la longitud de cada camino maximal que contiene un nodo comenzando en  $n$  es igual a 1. Las órbitas 0, 6, 7, 21, 23, y 29 son ejemplos de órbitas de oyentes, pero las órbitas 11 y 17 no lo son. (Ver Fig. 4.1)

De manera similar podemos decir que una órbita pozo  $\mathcal{O}$  es un hablante (*speaker*) si para cada nodo  $n \in \mathcal{O}$ ,  $n$  es un pozo con  $\deg - (n) > 1$ . Finalmente podemos decir que una órbita  $\mathcal{O}$  es una audiencia (*audience*) si para cada nodo  $n \in \mathcal{O}$ ,  $n$  es un oyente y cada otro nodo en una arista que comienza en  $n$  es un hablante. Las órbitas 7, 21 y 29 son ejemplos de órbitas de audiencia, pero la órbita 23 no lo es. (Ver Fig. 4.1)

Cada nodo en  $n$  participa en diferentes graphlets dentro de un red; cada graphlet nos da información sobre su vecindario local de 2, 3, o 4 nodos en los que  $n$  participa. Adicionalmente, la información proporcionada por distintos graphlets es diferente a aquella dada únicamente por el  $\deg - (n)$  o  $\deg + (n)$ . Por ejemplo, es posible distinguir las órbitas 0 y 29 reconociendo que pueden frecuentemente participar en distintos roles dentro de la estructura general de la red, que en el caso nos permite distinguir entre los perfiles 1 y 2. (Ver Fig. 4.1)

#### 4.2.3 Estabilidad de los perfiles identificados

A la similitud entre distintas particiones generadas para un conjunto de datos, la llamaremos la estabilidad de la solución. Mientras más robusta es una estructura de organización en una colección, más parecidos son los agrupamientos resultantes de distintas corridas, con distintas inicializaciones.

Se puede estimar la estabilidad de la solución utilizando la Información Mutua Normalizada (NMI).

La información mutua de dos variables aleatorias mide la dependencia estadística entre ambas variables. Es decir, mide la información o reducción de la

incertidumbre (entropía) de una variable aleatoria,  $X$ , debido al conocimiento del valor de otra variable aleatoria  $Y$ .

Consideremos dos variables aleatorias  $X$  e  $Y$  con posibles valores  $x_i, i = 1, 2, \dots, n, y_j, j = 1, 2, \dots, m$  respectivamente. Dónde

$$P(X = x_i|Y = y_i) = P(x_i|y_j)$$

y

$$P(X = x_i) = P(x_i)$$

De manera formal la Información Mutua está definida como

$$I(x_i; y_j) = \log \frac{P(x_i|y_j)}{P(x_i)}$$

y se puede obtener a partir de la entropía que esta definida como

$$H(X) = - \sum_i p(x_i) \log_2 p(x_i)$$

$$H(X, Y) = - \sum_{x,y} p(x, y) \log_2 p(x, y)$$

$$H(X|Y) = - \sum_y p(y) \sum_x p(x|y) \log_2 p(x|y)$$

Para obtener la estabilidad de la solución en nuestro problema, corremos el algoritmo de agrupamiento de perfiles  $R$  veces y obtenemos el promedio de los valores NMI para cada par de corridas del modelo. Es decir obtenemos una matriz de tamaño  $R \times R$  donde  $C_1, \dots, C_r$ , representan cada  $Corrida_i$ .

Formalmente,

$$Stability(C_1, \dots, C_r) = \frac{1}{r(r-1)} \sum_{i,j, i \neq j}^r NMI(C_i, C_j) \quad (4.1)$$

$$= \frac{1}{r(r-1)} \sum_{i,j, i \neq j}^r \frac{\mathbb{I}(C_i, C_j)}{\sqrt{H(C_i)H(C_j)}} \quad (4.2)$$

donde  $I(C_i, C_j)$  es la NMI entre corridas  $i, j$  y  $H(C_i)$  denota la entropía de la  $i$ -ésima asignación.

Tomar en cuenta lo robusto del agrupamiento para diferentes valores iniciales de los centros en el algoritmo de agrupamiento permite estimar la confianza en los perfiles identificados para usuarios considerados en la colección. Esto es importante porque dichos perfiles representan la base de la siguiente fase.

## 4.3 Agrupar Redes

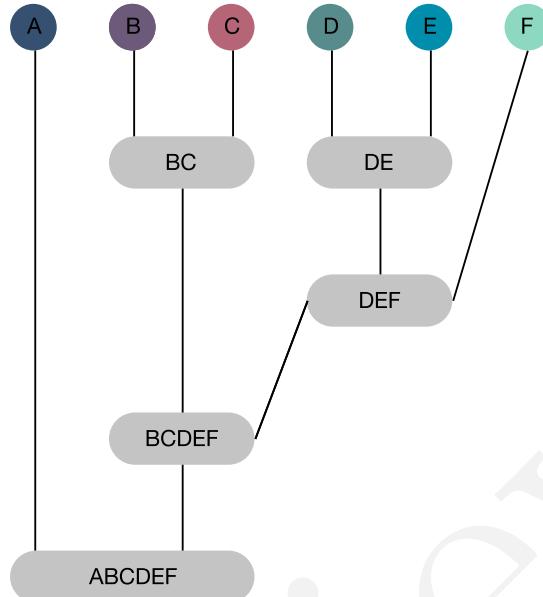
La segunda parte de la metodología se centra a agrupar las redes temáticas de la colección. Para ello, utilizamos una representación vectorial basada en los perfiles identificados en la primera parte de nuestro trabajo. De este modo, una vez que los perfiles de usuario se han establecido, creamos un segundo *embedding* a partir de la frecuencia de aparición de cada tipo de usuario en cada una de las redes. Nuestra hipótesis es que la frecuencia de aparición de cada perfil de usuario en la red podría variar en función del interés suscitado por un tema y de la naturaleza de la discusión pública (colectiva) en Twitter. Así, cada red es representada por un vector  $v$  en  $R^k$  donde  $k$  es el número de perfiles encontrados en el paso anterior y el componente  $v_i$  es el conteo de usuarios con el perfil  $i$ .

Al representar cada red de acuerdo a la distribución de frecuencia de los tipos de usuario identificados en la fase 1, estamos sugiriendo que un criterio que permite diferenciar las redes en la colección es la dinámica que generan.

**Clustering jerárquico** Una vez que se tiene la representación vectorial de cada red, utilizamos clustering jerárquico para establecer la comparación entre redes. Este método permite analizar la estructura, en términos de distancia, de los grupos que surgen dentro del conjunto de datos considerando la representación basada en perfiles de usuario.

En el clustering jerárquico, la estructura, o jerarquía de grupos, se determina de manera avara y comúnmente se presenta en un dendrograma. Además,

los resultados dependen de una medida de distancia entre las instancias del conjunto y un criterio de distancia para subconjuntos de datos.



**Fig. 4.3.:** Clustering Jerárquico

Una vez calculada la matriz de distancias entre instancias, los grupos se forman de acuerdo a alguno de los distintos criterios para calcular la distancia  $d(s, t)$  entre dos clusters  $s$  y  $t$ ; dichos criterios se muestran en la Tabla 4.1. El algoritmo que utilizamos, con un enfoque aglomerativo, comienza considerando cada instancia un grupo. En cada paso, la pareja de clústers  $s$  y  $t$  con mínima distancia entre ellos se unirá para formar un nuevo cluster  $u$ . El algoritmo termina cuando solo queda un único cluster al que llamamos raíz.

Para nuestro problema, la distancia entre instancias se calcula usando la norma L2, definida formalmente como

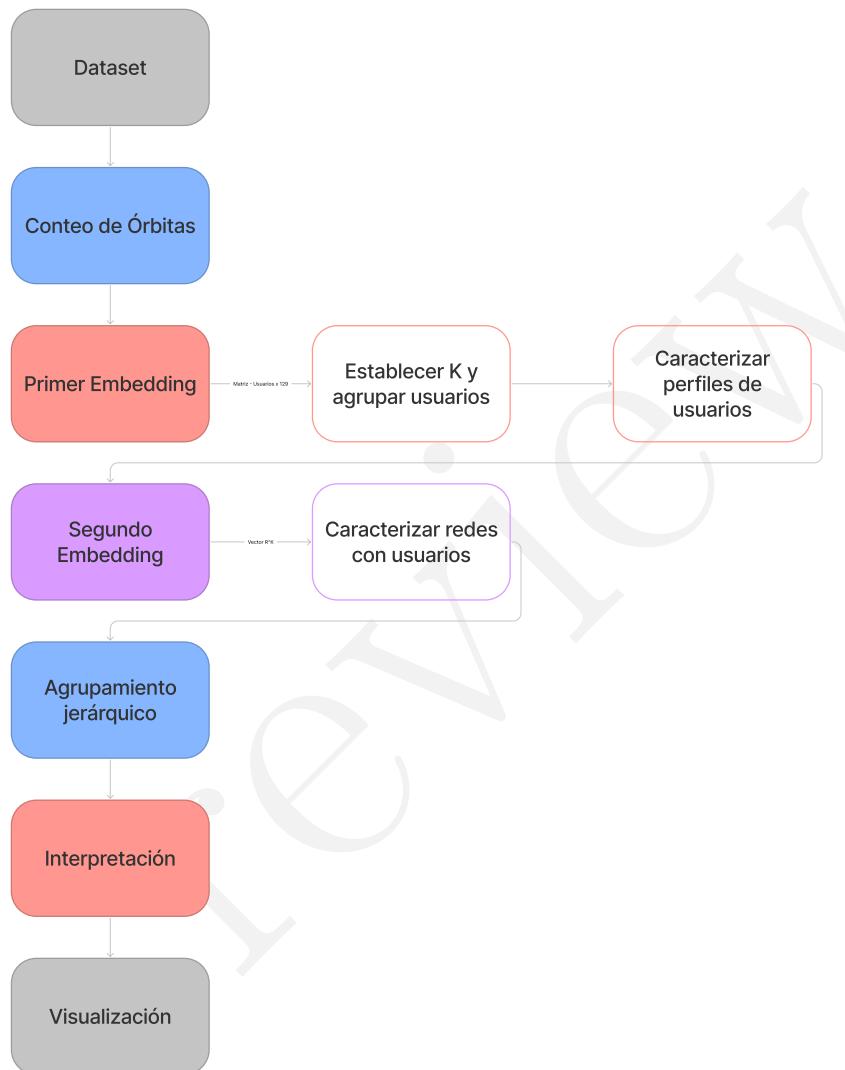
$$\|\boldsymbol{x}\|_2 := \sqrt{x_1^2 + \cdots + x_n^2}.$$

**Tab. 4.1.:** Criterios para calcular la distancia entre dos grupos en el agrupamiento jerárquico agolmerativo.

Nombre	Función
single	$d(u, v) = \min(\text{dist}(u[i], v[j]))$
complete	$d(u, v) = \max(\text{dist}(u[i], v[j]))$
average	$d(u, v) = \sum_{ij} \frac{d(u[i], v[j])}{( u  *  v )}$
weighted	$d(u, v) = \frac{\text{dist}(s, v) + \text{dist}(t, v)}{2}$
centroid	$d(u, v) = \ c_s - c_t\ _2$
Ward	$d(u, v) = \sqrt{\frac{ v  +  s }{T} d(v, s)^2 + \frac{ v  +  t }{T} d(v, t)^2 - \frac{ v }{T} d(s, t)^2}$ <p>dónde <math>T =  v  +  s  +  t </math></p>

## 4.4 Resumen

La metodología propuesta en este capítulo permite agrupar redes temáticas en Twitter de una forma guiada por los datos, interpretable y basada en el comportamiento que cada tema genera. La Fig. 4.4 muestra un esquema general del método propuesto.



**Fig. 4.4.:** Resumen de metodología

# Experimentos y resultados

En este capítulo presentamos los resultados del análisis de 75 redes temáticas reales asociadas a *trending topics* de Twitter en México durante 2020. En el primer paso de la metodología propuesta, se logró agrupar a los usuarios en 5 tipos de perfiles distintos de acuerdo las funciones estructurales inferidas de la firma orbital basada en graphlets. Posteriormente, una vez contados los perfiles de usuarios en cada red, se organizaron las diferentes redes a través del agrupamiento jerárquico aplicado en la colección. Al final del capítulo se discuten los resultados obtenidos, describiendo los diferentes perfiles de usuario identificados en términos de los patrones de comportamiento sugeridos por la frecuencia de sus órbitas.

## 5.1 Conjunto de datos

Uno de los principales retos en este trabajo fue obtener los datos necesarios para formar las redes temáticas. Se construyeron 75 redes temáticas a partir de *Trending Topics* (TTs) en Twitter haciendo un *scrapping* de tweets. Todos los temas elegidos están entre los primeros cinco TTs reportados por Twitter con más de 20K tweets en México durante noviembre de 2020. Las redes fueron preprocesadas para eliminar los bucles (gente que se responde a sí misma en la plataforma) y los nodos aislados (gente que decide no interactuar).

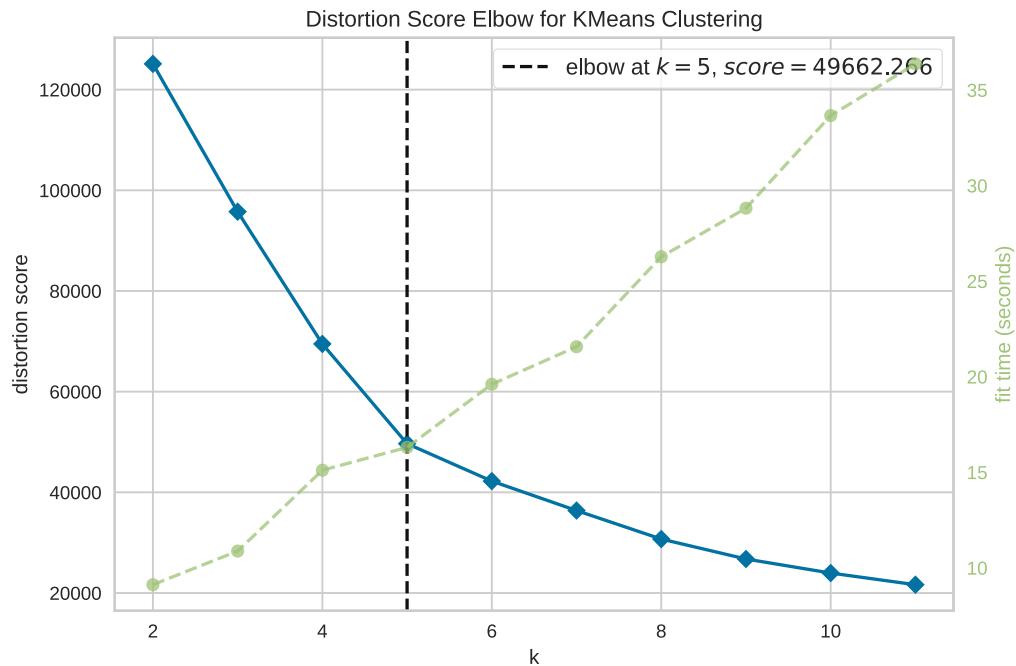
Todas las redes se crearon siguiendo la misma metodología, dando como resultado un conjunto de usuarios (nodos) y aristas dirigidas que corresponden a las interacciones de responder (incluyendo las menciones) y retuitear. No se utilizan etiquetas, por lo que ambas interacciones están igualmente representadas por una arista dirigida.

El orden y el tamaño de las redes están dentro del rango de [1952, 24876] y [9515, 35508] respectivamente. El conjunto de datos representa un total de 925896 nodos (usuarios) en la colección.

## 5.2 Primer agrupamiento: perfilando usuarios

Los perfiles de usuarios se basan en la firma orbital de cada nodo dentro de cada red considerando graphlets dirigidos. Las firmas orbitales de los nodos se calcularon con el software desarrollado por Anida Sarajlic et al. [Sar+16].

Después de realizar el cálculo de las firmas orbitales, obtenemos un primer *embedding* en  $R^{129}$  para cada nodo. Cada componente del vector representa el número de veces que un usuario (nodo) aparece en esa órbita. De esta manera, los vectores proveen información sobre los roles estructurales de los nodos (usuarios) dentro de la red.



**Fig. 5.1.:** Método Elbow o Codo para determinar el tamaño de K

Posteriormente, decidimos cuántos perfiles establecer para los usuarios. Con este objetivo, analizamos el conjunto de vectores-usuario con el método del codo, centrandonos en la suma de los errores cuadrados (*SSE* o *distortion*), i.e., experimentamos con un número diferente de grupos, tratando de identificar el punto de máxima curvatura (método del codo) en el cambio de *SSE*. Con este procedimiento, elegimos  $k = 5$  (ver Fig. 5.1).

Para agrupar las firmas orbitales en todo el conjunto de datos de la red, se utilizó la implementación de scikit-Learn de MiniBatch KMeans. El algoritmo de clustering se ejecutó con 500 inicializaciones aleatorias. En todos los casos, los centroides iniciales se calcularon utilizando el método de inicialización *K-means++* [AV].

### 5.2.1 Estabilidad

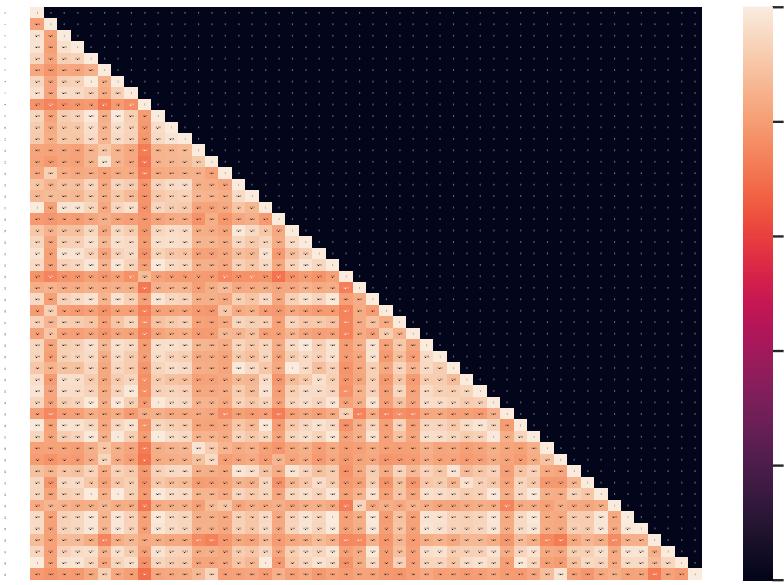
Se utilizaron 50 ejecuciones de la tarea de agrupamiento para estimar la estabilidad de los grupos identificados. La Información Mutua Normalizada (NMI) por pares de las ejecuciones se muestra en la Figura 5.2; el valor medio fue de 0.93.

### 5.2.2 Perfiles identificados

Para analizar los diferentes tipos usuarios identificados, consideramos los centroides como representantes de grupo.

En los datos analizados, los grupos 1, 2, 4 y 5 están definidos por una órbita claramente dominante, mientras que el grupo 3 corresponde a una distribución más balanceada en los roles de sus usuarios. La Fig. 4.1 muestra las principales órbitas para cada grupo.

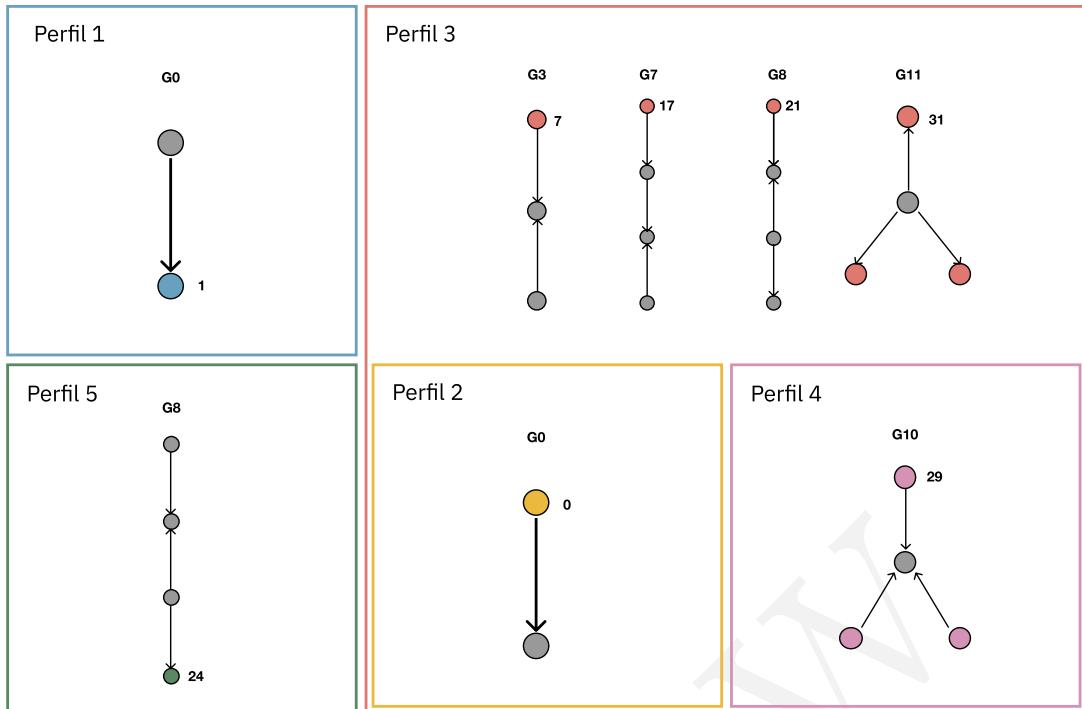
La tabla 5.1 expande la descripción de cada perfil al mostrar todas las órbitas con frecuencia relativa arriba de un umbral  $\Delta = 0.06$ , i.e., con un valor indicando que los usuarios en ese grupo participan en ese rol particular más del 5% de las veces.



**Fig. 5.2.:** Estabilidad del agrupamiento con  $K=5$  utilizando Normalized Mutual Information

**Tab. 5.1.:** Caracterización de los perfiles identificados de acuerdo a sus órbitas. Para las órbitas principales (segunda columna), solo se muestran los componentes con magnitud mayor que  $\Delta = 0.06$ .

Perfil	(Órbita, Puntuación)	Órbitas ausentes
1	(1, 0.85)	2, 3, 6, 7, 9, 11-18, 20, 21, 23-29, 31-62, 64, 65, 67-90, 92-124, 126-128
2	(0, 0.96)	1, 3-5, 7-10, 12-128
3	(29, 0.13), (7, 0.11), (31, 0.11), (17, 0.09), (0, 0.08), (21, 0.08)	Ninguna
4	(29, 0.94)	Ninguna
5	(24, 0.83)	111

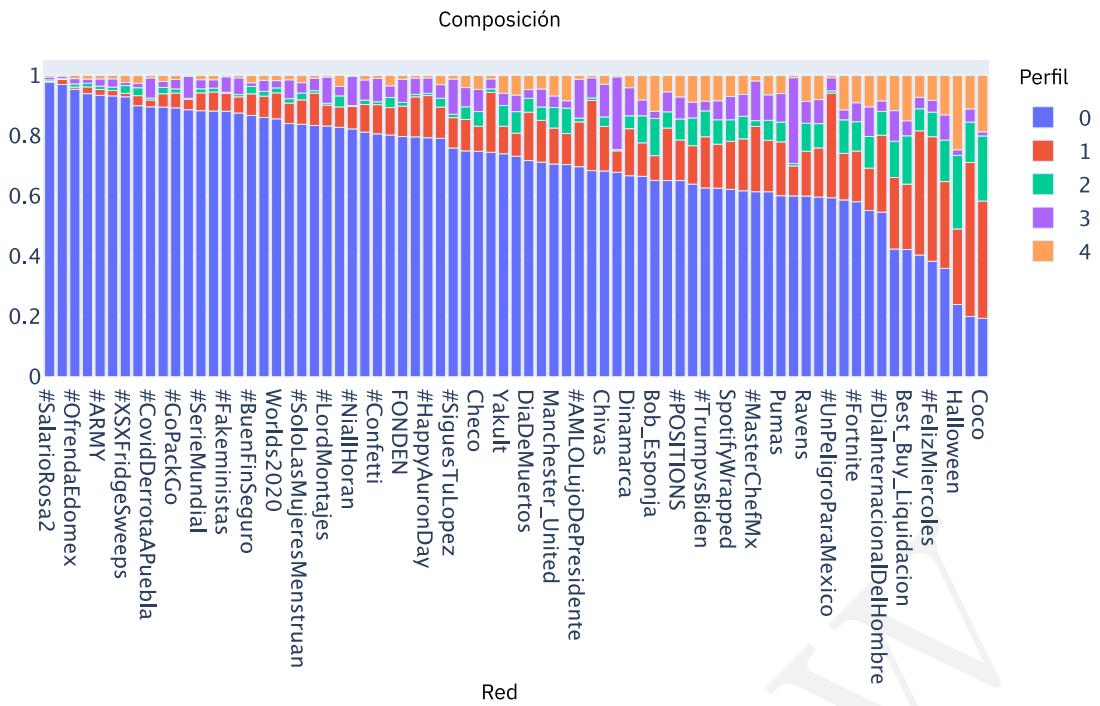


**Fig. 5.3.:** Perfiles encontrados

### 5.3 Segundo agrupamiento: estructura en redes

Usando los perfiles identificados en el paso anterior, podemos determinar la composición de las redes en la colección observando el porcentaje de los tipos de usuarios que se encuentran en cada red. En la Fig. 5.4 se observa que en la colección analizada existe una asimetría en la dinámica de comunicación de Twitter, con un gran grupo de usuarios que se involucra en la conversación principalmente a través de responder/apoyar lo que proponen unas pocas voces establecidas.

La Fig. 5.4 muestra la composición de cada red en términos de los cinco perfiles de usuario y revela diferentes dinámicas dentro de las redes. La mayoría de ellas están compuestas principalmente por usuarios con el perfil 4, lo que indica una dinámica muy jerarquizada en la que unos pocos usuarios tienen autoridad y fijan las ideas que circulan sobre el tema. El segundo perfil más común es el 3, seguido de los perfiles 1, 5 y 2.



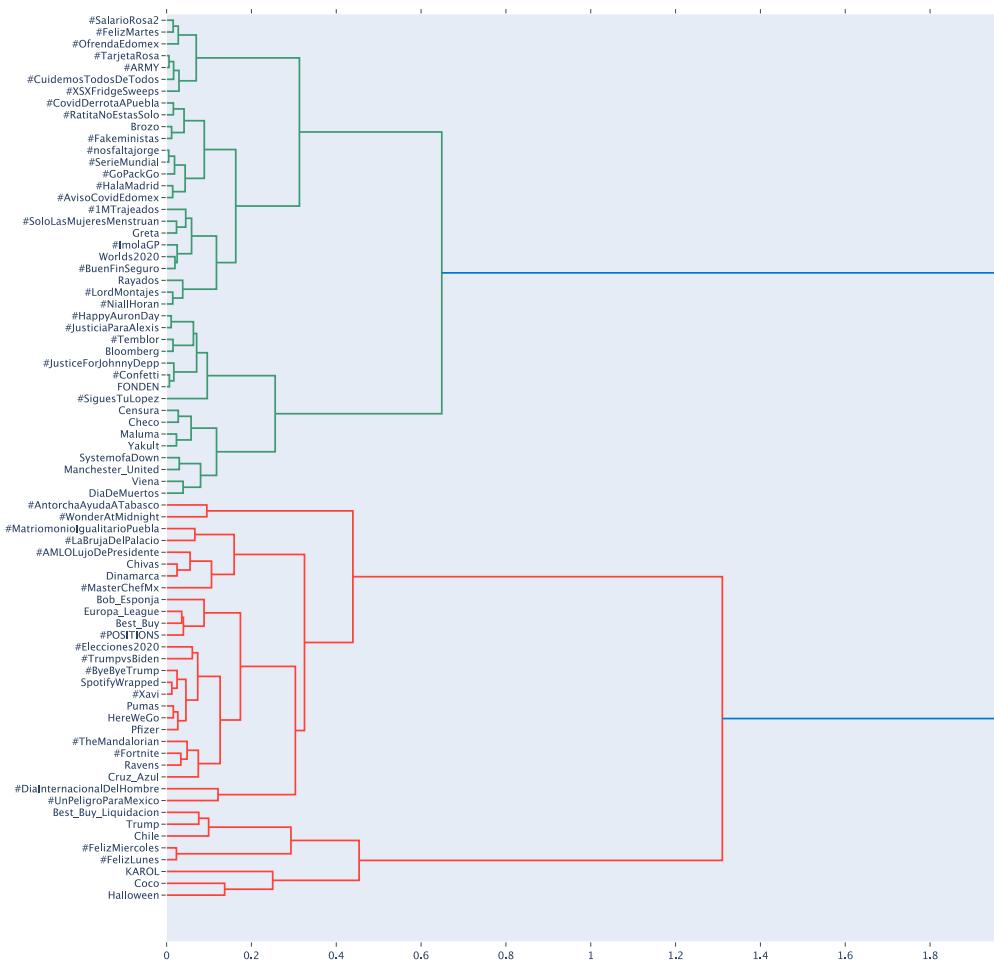
**Fig. 5.4.:** Composición de las redes de acuerdo al porcentaje de usuarios de cada perfil encontrado.

Con estos vectores, se utilizó agrupamiento jerárquico aglomerativo para buscar grupos. En la Figura 5.5 podemos observar el dendrograma que resulta al utilizar *Ward Linkage* y en la Figura 5.6 el dendrograma correspondiente a *Complete Linkage*.

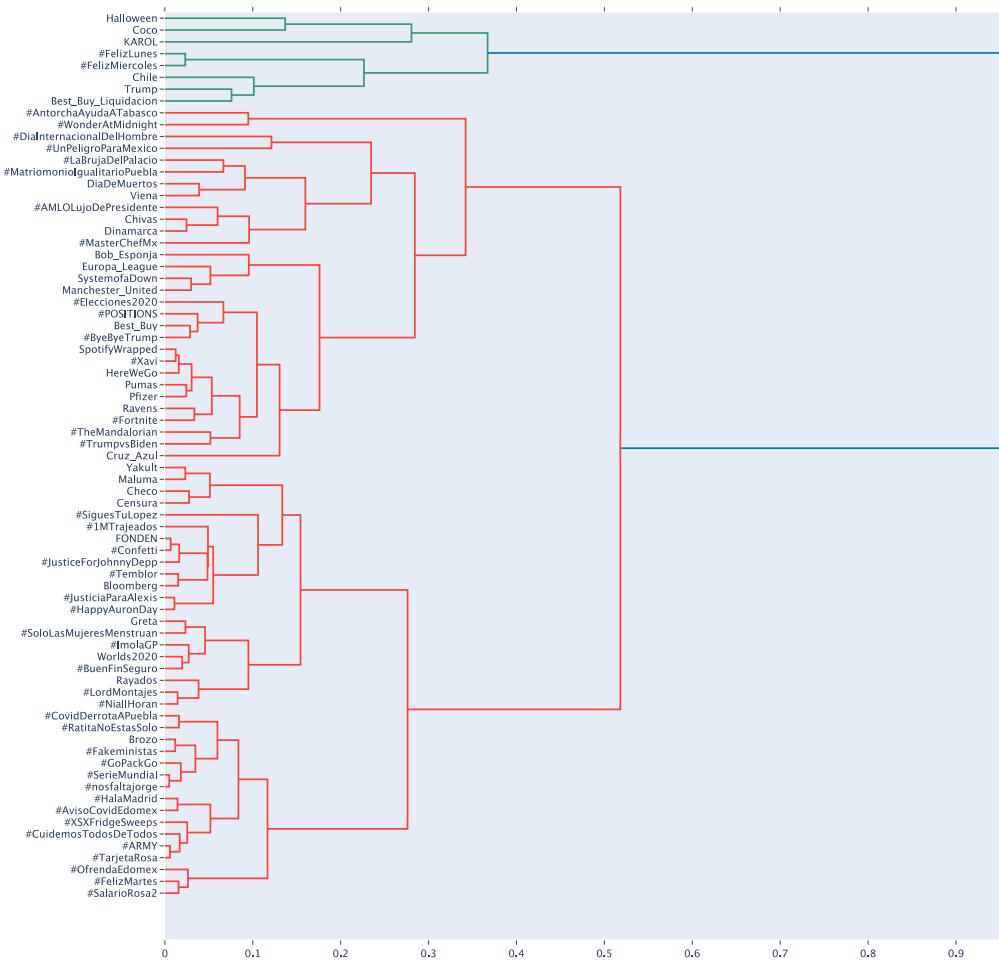
## 5.4 Visualización de resultados

Para explorar visualmente los resultados, se desarrolló una herramienta web utilizando las tecnologías de Docusaurus y React. El sitio web es estático y esta disponible en GithubPages (ver Figuras 5.7 y 5.8).

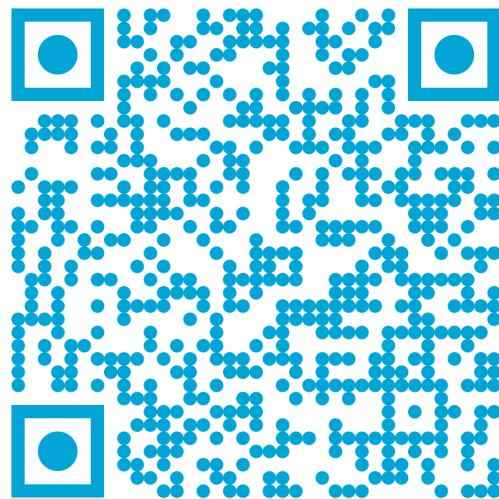
La herramienta web tiene distintas pestañas que permiten explorar distintos aspectos de nuestro trabajo. La primera, permite explorar con una gráfica de radar la composición por tipo de usuarios de la red seleccionada. La 5.16 ejemplifica la composición de una de las redes temáticas en la colección.



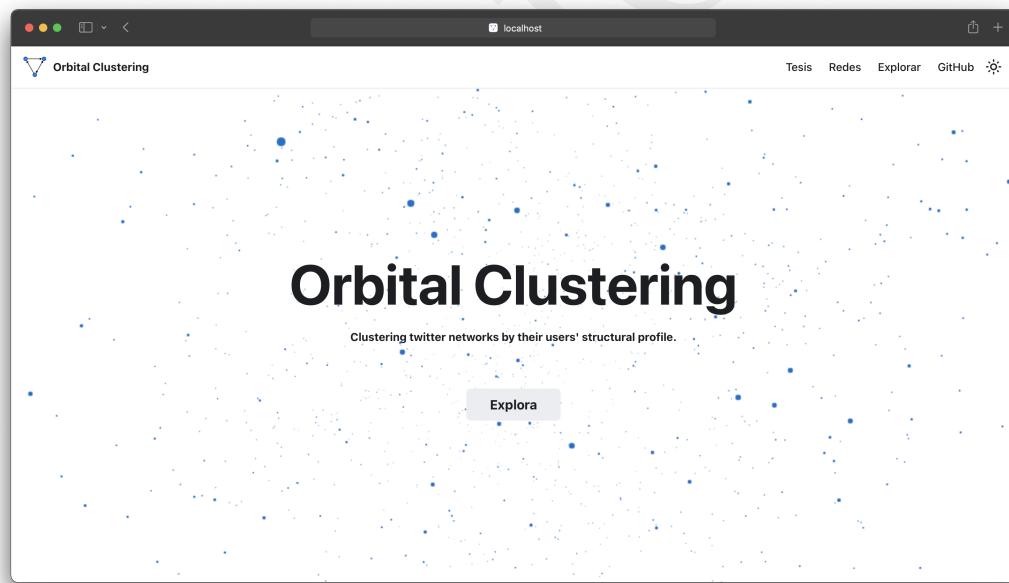
**Fig. 5.5.:** Agrupamiento jerárquico utilizando *Ward Linkage*



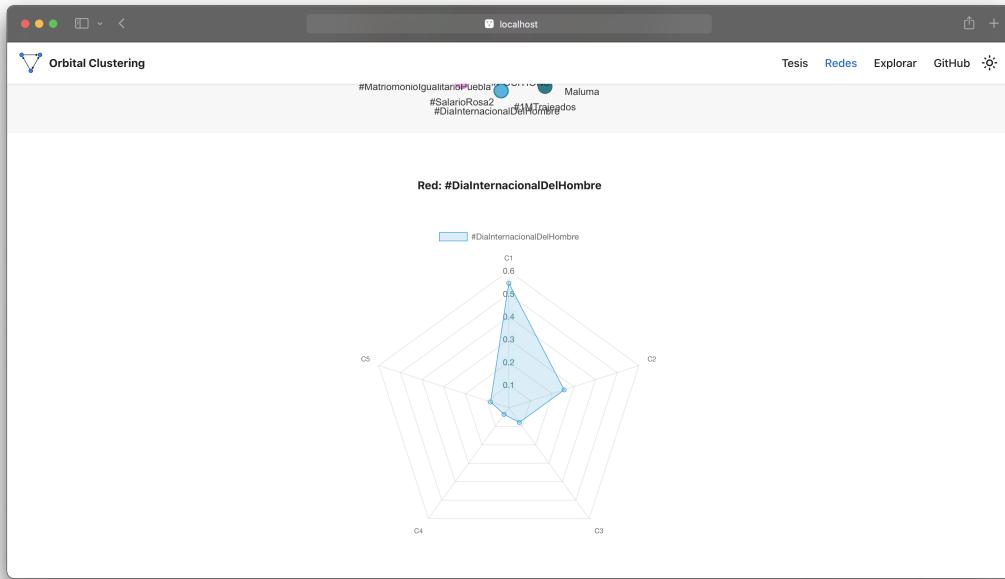
**Fig. 5.6.:** Agrupamiento jerárquico utilizando *Complete Linkage*



**Fig. 5.7.:** Código qr para acceder a la herramienta web.



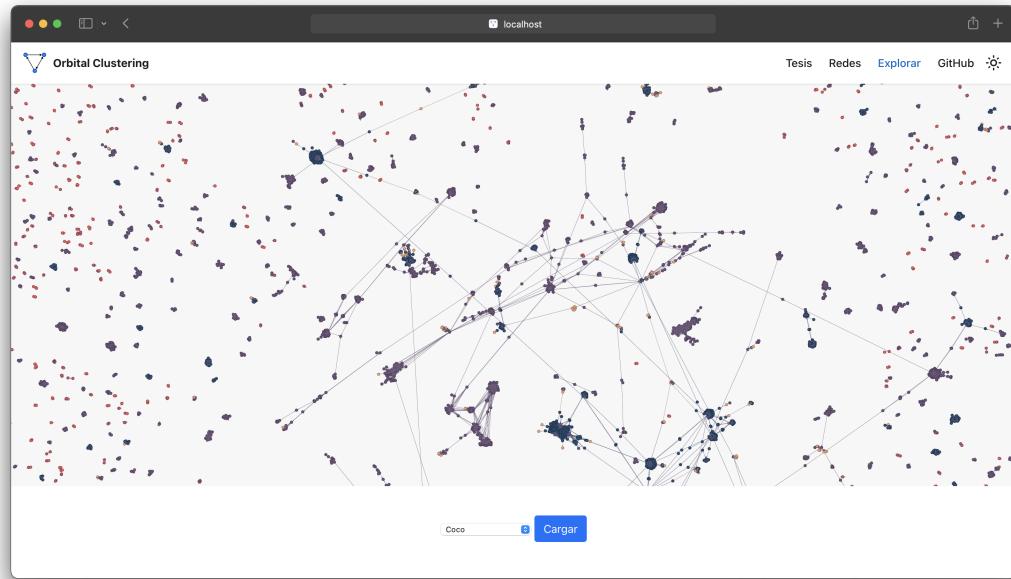
**Fig. 5.8.:** Página principal de la herramienta web.



**Fig. 5.9.:** Ejemplo de la visualización de la composición de una red utilizando una gráfica de rada.

Otra de las funciones principales es la visualización de los grafos con sus respectivos nodos coloreados de acuerdo al perfil que pertenecen. La Figura 5.10 nos muestra la red de #Coco, visibilizando algunas interacciones dentro de la red.

Los grafos que se muestran a continuación fueron escogidos como ejemplos por ser los más lejanos de acuerdo al agrupamiento jerárquico. En 5.11 observamos la red correspondiente al #Salario Rosa en las que la mayoría de las cuentas interactúan con un único usuario. Este tipo de comportamiento podría sugerir que el hashtag (#) nace a partir de un gran influenciador o que se trata de cuentas automatizadas que tienen el objetivo de hacer central a un usuario en la red. En contraste, la 5.12 muestra una red más bien fragmentada en la que no existe una conversación central.



**Fig. 5.10.:** Ejemplo de la visualización de una red dentro de la herramienta web.

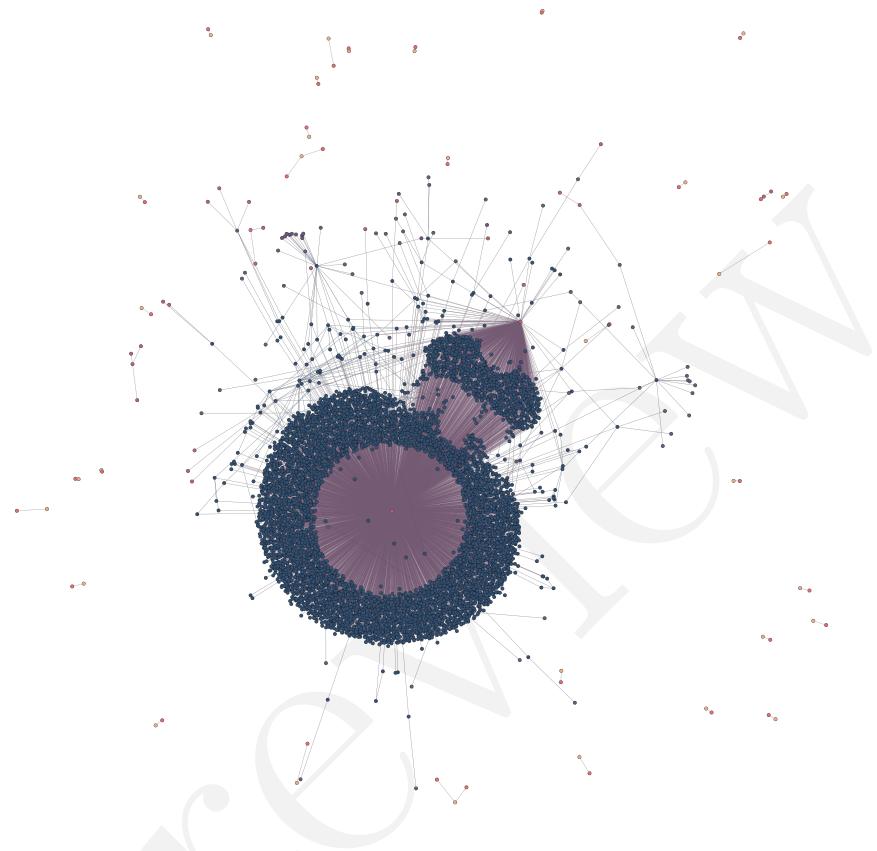
Red	Perfil 1	Perfil 2	Perfil 3	Perfil 4	Perfil 5
SalarioRosa2	0.977	0.009	0.003	0.004	0.004
Coco	0.194	0.013	0.416	0.214	0.160

**Tab. 5.2.:** Comparación de los *embeddings* de las redes de #Coco y #SalarioRosa2

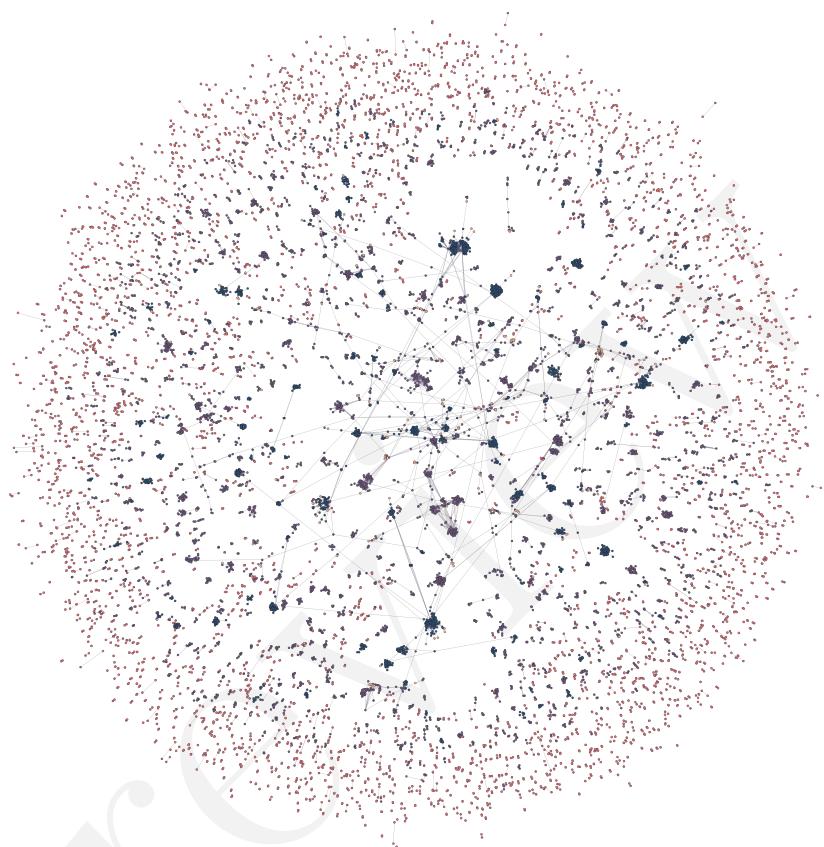
## 5.5 Discusión

En el conjunto de datos analizado, cuatro de los perfiles (1, 2, 4, 5) se distinguen por la presencia de una órbita dominante en el vector centroide representativo. En cambio, el grupo restante (3) tiene una distribución de órbitas más equilibrada en el vector de firmas de su centroide.

A continuación, se presenta una caracterización para cada uno de los perfiles de usuario identificados. Aunque la discusión se centra en los perfiles específicos identificados para esta colección, ejemplifica el tipo de análisis que puede derivarse de la metodología propuesta en este trabajo.

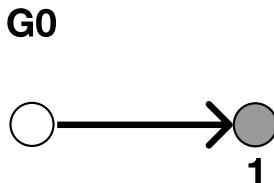


**Fig. 5.11.:** Red #SalarioRosa2 coloreada respecto al grupo al que pertenece cada nodo en la red.



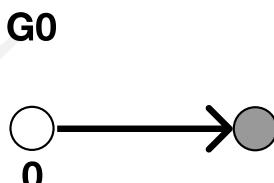
**Fig. 5.12.:** Red #Coco coloreada respecto al grupo al que pertenece cada nodo en la red.

- *Perfil 1, Difusor.* La órbita dominante es la 1, que desempeña el papel de un pozo en el graphlet compuesto por un solo arco. Las órbitas 2, 6 y 11 (todas ellas órbitas fuente) nunca aparecen en los vectores de firmas de estos usuarios. Analizando los vecindarios con tres nodos, las pocas veces que este perfil desempeña el papel de oyente, también lo hace de audiencia. Dada la alta frecuencia de la órbita dominante, es razonable suponer que estos usuarios producen información que motiva a los lectores a responder.



**Fig. 5.13.:** Graphlet 0 y órbita 1.

- *Perfil 2, Repetidor.* La órbita dominante es la 0, que desempeña el papel de oyente en un graphlet de arco, pero no tiene el papel de audiencia. La mayoría de las otras órbitas no aparecen asociadas a este tipo de usuario. En particular, si observamos todos los vecindarios con dos y tres nodos, este perfil nunca es retuiteado o mencionado por otro usuario. Además, observamos que el usuario no participa en graphlets de tamaño cuatro y, por tanto, tampoco en vecindarios más grandes. A partir de los roles recurrentes encontrados en esta órbita, podríamos decir que estos usuarios tienden a repetir los mensajes en la mayoría de sus interacciones sin impactar significativamente en la conversación.

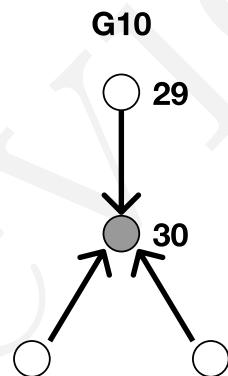


**Fig. 5.14.:** Graphlet 0 y órbita 0.

- *Perfil 3, Conversador.* En este perfil aparecen todas las órbitas incluyendo aquellas dominantes de los otros cuatro perfiles. Las órbitas dominantes

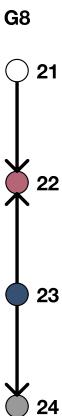
en este perfil son la 29, 7, 17, 21 y 31. Las mayoría de las órbitas son oyentes, pero la órbita 31 desempeña todos los papeles de pozo en un graphlet de 4 nodos. La variedad de roles que puede adoptar este grupo de usuarios, se ve reflejada en la composición equilibrada de los vectores de firmas asociados, sugiere que este perfil permite el flujo de información hacia y desde los otros perfiles predominantes.

- *Perfil 4, Reportero.* La órbita dominante es la 29, que desempeña todos los papeles de oyente en el graphlet de un triodo. Esta órbita dominante desempeña el papel de oyente. Analizando los vecindarios con tres nodos, es infrecuente que este perfil participe en rutas con una longitud superior a uno o que responda a tweets de dos nodos diferentes, pero es habitual que el usuario responda a tweets que están siendo contestados por una o dos personas más. Así, podríamos decir que este tipo de usuario tiende a responder a tweets y usuarios que son populares. Dado que este perfil incluye todas las órbitas, podríamos decir que estos usuarios tienen más impacto en la conversación que los repetidores.



**Fig. 5.15.:** Graphlet 10 y órbitas 29 y 30.

- *Perfil 5, Inconformista.* La órbita dominante es la 24, que desempeña el papel de hablante en un graphlet de 4 nodos. La particular arquitectura de este graphlet sugiere la presencia de nodos que recogen información de diferentes fuentes y que no interactúan entre sí. El comportamiento sugiere que este usuario participa en una discusión más amplia con un punto de vista parcial.



**Fig. 5.16.:** Graphlet 8 y órbitas 21 a 24.

Las órbitas 30, 63, 85, 91, 105, 118 y 125 son hablantes con un grado de salida igual a 3, que aparecen con muy poca frecuencia en las firmas de los perfiles identificados. Es de esperar que estas órbitas aparezcan en usuarios reconocidos como *Influencers* de la red. La presencia de la órbita 29 en el perfil de Reportero sugiere que la órbita 30 aparece varias veces en una red. Curiosamente, la órbita 30 aparece de forma distribuida, sin ser la órbita principal en los perfiles Locutor, Conversador o NonConformer.

En cuanto a la agrupación de las redes, la metodología propuesta permite ordenar la colección y definir grupos interpretables que proporcionan una visión de la dinámica originada por los diferentes temas. Los grupos no responden a una diferenciación temática, lo que refuerza la idea de que los procesos de difusión en Twitter no dependen sólo del contenido. No obstante, el análisis revela diferencias entre las redes sugiriendo una clara variación en cuanto a roles que emergen entre los usuarios y el efecto que esto tiene en la circulación de ideas a través de Twitter.

En el grupo de las redes que muestran una alta inequidad en las opiniones propagadas (redes más a la izquierda en la Fig. 5.4), con unas pocas voces autorizadas (locutores) de las que se hacen eco otros perfiles (reporteros), encontramos algunas iniciativas gubernamentales (#SalarioRosa, #OfrendaEndoMex, #TarjetaRosa). Podría darse el caso de que algunos tweets sean lanzados y manejados estratégicamente para aumentar su importancia. En el otro lado del espectro (instancias más a la derecha en la Fig. 5.4), encontramos

redes temáticas relacionadas con películas y temas generales (Coco, Karol, FelizMiercoles) que abarcan un intercambio de información más distribuido, lo que sugiere un tema con un mayor nivel de participación y menos voces predominantes sobre el tema.



## Conclusiones

Con el uso de modelos basados en redes en diferentes disciplinas del conocimiento, el agrupamiento en conjuntos de redes se vuelve una tarea muy importante. Sin embargo, no todos los métodos existentes proporcionan resultados que puedan traducirse fácilmente a nuevas interpretaciones de los datos.

En este trabajo se presenta una alternativa para agrupar redes sociales. El método propuesto tiene dos etapas principales: detectar el perfil de usuarios con base en su firma orbital en graphlets, y agrupar las redes de acuerdo a la caracterización de usuarios que las conforman. Nuestro enfoque es interpretable y capaz de captar la estructura de la red mediante el uso de graphlets.

La metodología presentada utiliza algoritmos computacionales ampliamente conocidos con implementaciones eficientes que permiten el desarrollo de cada paso propuesto. De este modo, nuestro enfoque aprovecha la utilidad de los graphlets y de sus órbitas asociadas para capturar información sobre la estructura de una red y llevar a cabo tareas de agrupamiento.

Mostramos la utilidad de la metodología propuesta a través de una aplicación real con redes temáticas de Twitter. Encontramos que los perfiles establecidos en el primer paso del método nos dan información útil sobre las estructuras de la red y las dinámicas sociales dentro de ellas. Esta descripción de perfiles puede considerarse una extensión de trabajo propuesto en sociología que sólo consideraba triadas de nodos. El método también reconoce que un usuario puede tener varios roles dentro de la discusión sobre un cierto tema en Twitter.

Consideramos que nuestro enfoque tiene al menos dos ventajas. En primer lugar, proporciona un método para agrupar redes temáticas de Twitter de forma

explicable, capturando las diferencias entre ellas que van más allá de las métricas generales de la red. En segundo lugar, produce una caracterización de los usuarios de la red que puede ayudar a comprender la estructura, las relaciones y los patrones latentes creados por la compleja dinámica de Twitter.

Desde el punto de vista sociológico, la utilización de perfiles de usuario sobre las redes temáticas, permite explorar las interacciones y las dinámicas que surgen durante una conversación pública en Twitter. Como vimos en el primer capítulo, el análisis de este tipo de redes permite modelar y comprender fenómenos asociados a este tipo de discusiones. Seguramente aún quedan distintas posibilidades de análisis por explorar a partir de estudio de las órbitas.

Entre las líneas de trabajo futuro que se proponen, existe la posibilidad de explorar la generalidad de los perfiles de usuario detectados. Es decir, queda por hacer un análisis más detallado de estos perfiles para extender la discusión con herramientas y metodologías de otras áreas afines.

# Apéndice

## A.1 Capítulo 1

### **Homofilia**

En sociología se denomina homofilia (del griego «amor a los iguales») a la tendencia de las personas por la atracción a sus homónimos. Esta atracción puede ser respecto a distintos atributos como edad, género, creencias, educación, estrato social, etc.

### **Centralidad de Intermediación**

La distancia de intermediación o *Betweenes Centrality* es una medida de centralidad basada en geodésicas o caminos más cortos [Wik].

Formalmente esta definida como:

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

donde  $\sigma_{st}$  es el número total de caminos mas cortos desde el nodo  $s$  al nodo  $t$  y  $\sigma_{st}(v)$  es el número total de esos caminos que pasan a través de  $v$  (dónde  $v$  no es el nodo final de un camino).

## A.2 Capítulo 2

### Función Biyectiva

Una función es biyectiva es aquella que es a la vez inyectiva y suprayectiva. Es decir, una función entre los elementos de dos conjuntos, donde cada elemento de un conjunto se empareja con exactamente un elemento del otro conjunto, y cada elemento del otro conjunto se empareja con exactamente un elemento del primer conjunto.

Formalmente, dada una función  $f$

$$\begin{aligned} f : X &\longrightarrow Y \\ x &\longmapsto y = f(x) \end{aligned}$$

Es biyectiva si para todo  $y$  de  $Y$  existe un único  $x$  de  $X$  al que la función evaluada en  $x$  es igual a  $y$

$$\forall y \in Y : \exists! x \in X / f(x) = y$$

## A.3 Capítulo 5

### Línea base

Utilizando el árbol de decisión de *Himelboim et. al* [Him+17] se clasificó el conjunto de datos de redes temáticas de Twitter para establecer una línea base. En este caso el agrupamiento no es óptimo ya que la mayoría de las redes quedan en un solo grupo.

Network	Label
LordMontajes	Clustered
MasterChefMx	Clustered
NiallHoran	Clustered
POSITIONS	Clustered
TheMandalorian	Clustered
WonderAtMidnight	OutWard Hub and Spoke
XSFridgeSweeps	Fragmented
Bob Esponja	Clustered
Coco	Clustered
KAROL	Clustered
Maluma	Clustered
SpotifyWrapped	Clustered
SystemofaDown	Clustered
Yakult	Clustered
AMLOLujoDePresidente	Clustered
AntorchaAyudaATabasco	Clustered
ByeByeTrump	Clustered
CuidemosTodosDeTodos	Clustered
Elecciones2020	Clustered
JusticiaParaAlexis	Clustered
LaBrujaDelPalacio	Clustered
MatriomonioIgualitarioPuebla	Clustered
RatitaNoEstasSolo	OutWard Hub and Spoke
SalarioRosa2	Clustered
SiguesTuLopez	Clustered
SoloLasMujeresMenstruan	Clustered
TarjetaRosa	Fragmented
TrumpvsBiden	Clustered
UnPeligroParaMexico	Clustered
nosfaltajorge	Clustered
Brozo	Clustered
Censura	Clustered
Chile	Clustered
FONDEN	Clustered
Trump	Clustered

**Tab. A.1.:** Resultado del agrupamiento realizado utilizando el árbol de decisión de [Him+17] para el conjunto de redes temáticas.

Network	Label
AvisoCovidEdomex	Fragmented
BuenFinSeguro	Fragmented
CovidDerrotaAPuebla	OutWard Hub and Spoke
DiaInternacionalDelHombre	Clustered
Fakeministas	Clustered
FelizLunes	Clustered
JusticeForJohnnyDepp	Clustered
OfrendaEdomex	Clustered
Tremblor	Clustered
Best Buy Liquidacion	Clustered
Best Buy	Clustered
Bloomberg	Clustered
DiaDeMuertos	Clustered
Dinamarca	Clustered
Greta	Clustered
Halloween	Clustered
Pfizer	Clustered
Viena	Clustered
GoPackGo	OutWard Hub and Spoke
HalaMadrid	Clustered
ImolaGP	Clustered
SerieMundial	Clustered
Xavi	Clustered
Checo	Clustered
Chivas	Clustered
Cruz Azul	Clustered
Europa League	Clustered
HereWeGo	Clustered
Manchester United	Clustered
Pumas	Clustered
Ravens	Clustered
Rayados	OutWard Hub and Spoke
Worlds2020	Clustered
1MTrajeados	OutWard Hub and Spoke
ARMY	OutWard Hub and Spoke
Confetti	OutWard Hub and Spoke
FelizMartes	OutWard Hub and Spoke
FelizMiercoles	Clustered
Fortnite	Clustered
HappyAuronDay	Clustered

**Tab. A.2.:** Resultado del agrupamiento realizado utilizando el árbol de decisión de [Him+17] para el conjunto de redes temáticas.

# Bibliografía

- [Aha+22] David Y. Aharon, Ender Demir, Chi Keung Marco Lau y Adam Zaremba. “Twitter-Based uncertainty and cryptocurrency returns”. en. En: *Research in International Business and Finance* 59 (ene. de 2022), pág. 101546. ISSN: 02755319. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0275531921001677> (visitado 25-01-2022) (vid. pág. 2).
- [AV] David Arthur y Sergei Vassilvitskii. “k-means++: The Advantages of Careful Seeding”. en. En: (), pág. 11 (vid. pág. 49).
- [BR15] Pablo Barberá y Gonzalo Rivero. “Understanding the Political Representativeness of Twitter Users”. en. En: *Social Science Computer Review* 33.6 (dic. de 2015), págs. 712-729. ISSN: 0894-4393, 1552-8286. URL: <http://journals.sagepub.com/doi/10.1177/0894439314558836> (visitado 25-01-2022) (vid. pág. 1).
- [Béj] Javier Béjar. “K-means vs Mini Batch K-means: A comparison”. en. En: (), pág. 12 (vid. págs. 3, 39).
- [BMZ11] Johan Bollen, Huina Mao y Xiaojun Zeng. “Twitter mood predicts the stock market”. en. En: *Journal of Computational Science* 2.1 (mar. de 2011), págs. 1-8. ISSN: 18777503. URL: <https://linkinghub.elsevier.com/retrieve/pii/S187775031100007X> (visitado 25-01-2022) (vid. pág. 2).
- [Bur04] Ronald S. Burt. “Structural Holes and Good Ideas”. en. En: *American Journal of Sociology* 110.2 (sep. de 2004), págs. 349-399. ISSN: 0002-9602, 1537-5390. URL: <http://www.journals.uchicago.edu/doi/10.1086/421787> (visitado 21-01-2022) (vid. pág. 8).
- [Gab17] Ivor Gaber. “Twitter: A useful tool for studying elections?” en. En: *Convergence: The International Journal of Research into New Media Technologies* 23.6 (dic. de 2017), págs. 603-626. ISSN: 1354-8565, 1748-7382. URL: <http://journals.sagepub.com/doi/10.1177/1354856516646544> (visitado 25-01-2022) (vid. pág. 2).

- [GRL14] Maksym Gabielkov, Ashwin Rao y Arnaud Legout. “Studying social networks at scale: macroscopic anatomy of the twitter social graph”. en. En: *The 2014 ACM international conference on Measurement and modeling of computer systems - SIGMETRICS '14*. Austin, Texas, USA: ACM Press, 2014, págs. 277-288. ISBN: 978-1-4503-2789-3. URL: <http://dl.acm.org/citation.cfm?doid=2591971.2591985> (visitado 02-10-2020) (vid. pág. 4).
- [HK16] Sifei Han y Ramakanth Kavuluru. “Exploratory Analysis of Marketing and Non-marketing E-cigarette Themes on Twitter”. En: *Social Informatics*. Ed. por Emma Spiro y Yong-Yeol Ahn. Vol. 10047. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016, págs. 307-322. ISBN: 978-3-319-47873-9 978-3-319-47874-6. URL: [http://link.springer.com/10.1007/978-3-319-47874-6\\_22](http://link.springer.com/10.1007/978-3-319-47874-6_22) (visitado 25-01-2022) (vid. pág. 2).
- [Him+17] Itai Himelboim, Marc A. Smith, Lee Rainie, Ben Shneiderman y Camila Espina. “Classifying Twitter Topic-Networks Using Social Network Analysis”. en. En: *Social Media + Society* 3.1 (mar. de 2017), pág. 205630511769154. ISSN: 2056-3051, 2056-3051. URL: <http://journals.sagepub.com/doi/10.1177/2056305117691545> (visitado 02-10-2020) (vid. págs. 3, 7, 23, 68-70).
- [Hu20] Margaret Hu. “Cambridge Analytica’s black box”. en. En: *Big Data & Society* 7.2 (jul. de 2020), pág. 205395172093809. ISSN: 2053-9517, 2053-9517. URL: <http://journals.sagepub.com/doi/10.1177/2053951720938091> (visitado 09-03-2022) (vid. pág. 37).
- [IBM] IBM. *What is Machine Learning?* en. Blog. URL: <https://www.ibm.com/cloud/learn/machine-learning> (visitado 23-08-2021) (vid. pág. 15).
- [KST93] Johannes Köbler, Uwe Schöning y Jacobo Torán. *The Graph Isomorphism Problem*. en. Boston, MA: Birkhäuser Boston, 1993. ISBN: 978-1-4612-6712-6 978-1-4612-0333-9. URL: <http://link.springer.com/10.1007/978-1-4612-0333-9> (visitado 22-08-2021) (vid. pág. 14).
- [Kub17] Miroslav Kubat. *An Introduction to Machine Learning*. en. Cham: Springer International Publishing, 2017. ISBN: 978-3-319-63912-3 978-3-319-63913-0. URL: <http://link.springer.com/10.1007/978-3-319-63913-0> (visitado 21-08-2021) (vid. págs. 17, 20).

- [Kwa+10] Haewoon Kwak, Changhyun Lee, Hosung Park y Sue Moon. “What is Twitter, a social network or a news media?” en. En: *Proceedings of the 19th international conference on World wide web - WWW '10*. Raleigh, North Carolina, USA: ACM Press, 2010, pág. 591. ISBN: 978-1-60558-799-8. URL: <http://portal.acm.org/citation.cfm?doid=1772690.1772751> (visitado 02-10-2020) (vid. pág. 5).
- [Ler+19] Adam Lerer, Ledell Wu, Jiajun Shen y col. “PyTorch-BigGraph: A Large-scale Graph Embedding System”. en. En: *arXiv:1903.12287 [cs, stat]* (abr. de 2019). arXiv: 1903.12287. URL: <http://arxiv.org/abs/1903.12287> (visitado 06-10-2021) (vid. págs. 24, 28).
- [Lus] Dean() Lusher. “Exponential Random Graph Models for Social Networks”. en. En: (), pág. 361 (vid. págs. 31, 38).
- [Med+20] Richard J Medford, Sameh N Saleh, Andrew Sumarsono, Trish M Perl y Christoph U Lehmann. “An “Infodemic”: Leveraging High-Volume Twitter Data to Understand Early Public Sentiment for the Coronavirus Disease 2019 Outbreak”. en. En: *Open Forum Infectious Diseases* 7.7 (jul. de 2020), ofaa258. ISSN: 2328-8957. URL: <https://academic.oup.com/ofid/article/doi/10.1093/ofid/ofaa258/5865318> (visitado 25-01-2022) (vid. pág. 2).
- [MP08] Tijana Milenković y Nataša Pržulj. “Uncovering Biological Network Function via Graphlet Degree Signatures”. en. En: *Cancer Informatics* 6 (ene. de 2008), CIN.S680. ISSN: 1176-9351, 1176-9351. URL: <http://journals.sagepub.com/doi/10.4137/CIN.S680> (visitado 21-01-2022) (vid. págs. 30, 31).
- [Mur+21] Taichi Murayama, Shoko Wakamiya, Eiji Aramaki y Ryota Kobayashi. “Modeling the spread of fake news on Twitter”. en. En: *PLOS ONE* 16.4 (abr. de 2021). Ed. por Kazutoshi Sasahara, e0250419. ISSN: 1932-6203. URL: <https://dx.plos.org/10.1371/journal.pone.0250419> (visitado 25-01-2022) (vid. pág. 3).
- [New10] M. E. J. Newman. *Networks: an introduction*. OCLC: ocn456837194. Oxford ; New York: Oxford University Press, 2010. ISBN: 978-0-19-920665-0 (vid. pág. 21).
- [Prz07] N. Pržulj. “Biological network comparison using graphlet degree distribution”. en. En: *Bioinformatics* 23.2 (ene. de 2007), e177-e183. ISSN: 1367-4803, 1460-2059. URL: <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btl301> (visitado 19-10-2021) (vid. págs. 3, 29).

- [RRC19] Gopinath Rebala, Ajay Ravi y Sanjay Churiwala. *An Introduction to Machine Learning*. en. Cham: Springer International Publishing, 2019. ISBN: 978-3-030-15728-9 978-3-030-15729-6. URL: <http://link.springer.com/10.1007/978-3-030-15729-6> (visitado 21-08-2021) (vid. págs. 16, 17, 28).
- [Ros+16] Ji Youn Rose Kim, Michael Howard, Emily Cox Pahnke y Warren Boeker. “Understanding network formation in strategy research: Exponential random graph models: Understanding Network Formation in Strategy Research: ERGMs”. en. En: *Strategic Management Journal* 37.1 (ene. de 2016), págs. 22-44. ISSN: 01432095. URL: <http://doi.wiley.com/10.1002/smj.2454> (visitado 17-06-2021) (vid. pág. 8).
- [Sao21] Karin R. Saoub. *Graph theory: an introduction to proofs, algorithms, and applications*. en. Textbooks in mathematics. Boca Raton: CRC Press, 2021. ISBN: 978-1-138-36140-9 978-0-367-74375-8 (vid. pág. 13).
- [Sar+16] Anida Sarajlić, Noël Malod-Dognin, Ömer Nabil Yaveroğlu y Nataša Pržulj. “Graphlet-based Characterization of Directed Networks”. en. En: *Scientific Reports* 6.1 (dic. de 2016), pág. 35098. ISSN: 2045-2322. URL: <http://www.nature.com/articles/srep35098> (visitado 02-10-2020) (vid. págs. 3, 30, 31, 35, 48).
- [SKB19] Rakhi Saxena, Sharanjit Kaur y Vasudha Bhatnagar. “Identifying similar networks using structural hierarchy”. en. En: *Physica A: Statistical Mechanics and its Applications* 536 (dic. de 2019), pág. 121029. ISSN: 03784371. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0378437119306399> (visitado 20-08-2021) (vid. pág. 23).
- [Scu10] D. Sculley. “Web-scale k-means clustering”. en. En: *Proceedings of the 19th international conference on World wide web - WWW '10*. Raleigh, North Carolina, USA: ACM Press, 2010, pág. 1177. ISBN: 978-1-60558-799-8. URL: <http://portal.acm.org/citation.cfm?doid=1772690.1772862> (visitado 03-06-2021) (vid. págs. 38, 39).
- [Sha+22] Filipo Sharevski, Raniem Alsaadi, Peter Jachim y Emma Pieroni. “Misinformation warnings: Twitter’s soft moderation effects on COVID-19 vaccine belief echoes”. en. En: *Computers & Security* 114 (mar. de 2022), pág. 102577. ISSN: 01674048. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0167404821004016> (visitado 25-01-2022) (vid. págs. 1, 2).
- [Twi] Twitter. *Twitter.com*. en. URL: [twitter.com/about](http://twitter.com/about) (visitado 25-05-2021) (vid. págs. 4, 5).

- [Wik] Wikipedia. *Betweenness centrality*. en. URL: [https://en.wikipedia.org/wiki/Betweenness\\_centrality](https://en.wikipedia.org/wiki/Betweenness_centrality) (vid. pág. 67).
- [XK19] Feng Xia y Xiangjie Kong. “Random Walks: A Review of Algorithms and Applications”. en. En: *IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTATIONAL INTELLIGENCE* 0.0 (2019), pág. 13 (vid. pág. 24).
- [XZ20] Sifan Xu y Alvin Zhou. “Hashtag homophily in twitter network: Examining a controversial cause-related marketing campaign”. en. En: *Computers in Human Behavior* 102 (ene. de 2020), págs. 87-96. ISSN: 07475632. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0747563219302936> (visitado 25-01-2022) (vid. pág. 2).
- [Zac77] Wayne W. Zachary. “An Information Flow Model for Conflict and Fission in Small Groups”. en. En: *Journal of Anthropological Research* 33.4 (1977), págs. 452-473. URL: <http://www.jstor.org/stable/3629752> (vid. págs. 32, 33).
- [Zha+21] Yu Zhang, Peter Tiňo, Aleš Leonardis y Ke Tang. “A Survey on Neural Network Interpretability”. en. En: *arXiv:2012.14261 [cs]* (jul. de 2021). arXiv: 2012.14261. URL: <http://arxiv.org/abs/2012.14261> (visitado 30-11-2021) (vid. pág. 28).

review