



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Licenciatura En Tecnologías Para La Información En Ciencias

ESCUELA NACIONAL DE ESTUDIOS SUPERIORES
UNIDAD MORELIA

APRENDIZAJE NO SUPERVISADO PARA EL ESTUDIO DE REDES TEMÁTICAS DE TWITTER

T E S I S

Para obtener el título de:

LICENCIADO EN TECNOLOGÍAS PARA LA
INFORMACIÓN EN CIENCIAS

PRESENTA

Rodrigo Sebastián Cortez Madrigal

Directora de Tesis: Dra. Marisol Flores Garrido

Co-director de Tesis: Dr. Luis Miguel García Velázquez

Morelia, Michoacán. 2023



Universidad Nacional
Autónoma de México

Dirección General de Bibliotecas de la UNAM

Biblioteca Central



UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.



ESCUELA
NACIONAL
de ESTUDIOS
SUPERIORES
ENES
UNIDAD MORELIA

10
años
(2011-2021)

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
ESCUELA NACIONAL DE ESTUDIOS SUPERIORES UNIDAD MORELIA
SECRETARÍA GENERAL
SERVICIOS ESCOLARES

MTRA. IVONNE RAMÍREZ WENCE

DIRECTORA

DIRECCIÓN GENERAL DE ADMINISTRACIÓN ESCOLAR

P R E S E N T E

Por medio de la presente me permito informar a usted que en la **sesión ordinaria 25** del **Comité Académico de la Licenciatura en Tecnologías para la Información en Ciencias** de la Escuela Nacional de Estudios Superiores (ENES), Unidad Morelia, celebrada el día **12 de mayo de 2022**, se acordó poner a su consideración el siguiente jurado para la presentación del Trabajo Profesional del alumno **Rodrigo Sebastián Cortez Madrigal** de la Licenciatura en **Tecnologías para la Información en Ciencias**, con número de cuenta **416064845**, con el trabajo titulado: **"Aprendizaje no supervisado para el estudio de redes temáticas de Twitter "**, bajo la dirección como tutora de la **Dra. Marisol Flores Garrido** y como co-tutor el **Dr. Luis Miguel García Velázquez**.

El jurado queda integrado de la siguiente manera:

Presidente:	Dra. María del Río Francos
Vocal:	Dr. Miguel Raggi Pérez
Secretario:	Dra. Marisol Flores Garrido
Suplente:	Dra. Adriana Menchaca Méndez
Suplente:	Dr. Julián Alberto Fresán Figueroa


Sin otro particular, quedo de usted.

Atentamente
"POR MI RAZA HABLARÁ EL ESPÍRITU"
Morelia, Michoacán a 15 de noviembre de 2023.

DRA. YUNUEN TAPIA TORRES
SECRETARIA GENERAL

CAMPUS MORELIA

Antigua Carretera a Pátzcuaro N° 8701, Col. Ex Hacienda de San José de la Huerta
58190, Morelia, Michoacán, México. Tel: (443)689.3500 y (55)5623.7300, Extensión Red UNAM: 80614
www.enesmorelia.unam.mx



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

Redes y Graphlets. 2023

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

Lic. en Tecnologías para la Información en Ciencias

Universidad Nacional Autónoma de México

Escuela Nacional de Estudios Superiores, Unidad Morelia

Antigua Carretera a Pátzcuaro No. 8701, Col. Ex Hacienda de San José de la Huerta

C.P. 58190 , Morelia, Michoacán, México.

Agradecimientos Institucionales

Esta tesis se realizó bajo el PAPIIT IA106620, *Ciencia de Datos para las Humanidades Digitales*. Gracias a la Universidad Nacional Autónoma de México, la Licenciatura en Tecnologías para la Información en Ciencias y el cuerpo docente de la Escuela Nacional de Estudios Superiores, Unidad Morelia.

En especial a mis profesores,

Marisol Flores Garrido · Luis Miguel García Velázquez · **Miguel Raggi Pérez** ·
Adriana Menchaca Méndez · **Mario Alberto Duarte García** · Mercedes Martínez
Gonzales · **Daniele Colosi** · Fernando García García · **María del Río Francos**

Quiénes me han enseñado tanto dentro y fuera del salón de clases.

Agradecimientos Personales

“

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

— **Isaac Newton**

A mis padres, quienes fueron también mis primeros maestros y me enseñaron que el amor, la sabiduría y el amor a la sabiduría eran lo que hacía de esta existencia un poco más llevadera e interesante. A mis hermanos, quienes siempre me han acompañado en la difícil tarea de crecer y aprender. A mis amigos, quienes estuvieron presentes y me tendieron la mano siempre que lo necesité (Gracias Fer). A Sigrid Solbakk Raabe, quien fue mi más grande compañía en todas las noches en vela en las que realicé este trabajo.

Adicionalmente,

Este trabajo se realizó en el contexto de la Pandemia del COVID-19 causada por el virus SARS-CoV-2. Durante este acontecimiento fue más claro que nunca que la apertura del conocimiento y el libre acceso a la información es un Derecho Humano. Velemos por una ciencia transparente e inclusiva.

Gracias a todo el software libre y de código abierto con el que este trabajo se realizó. Я также хотел бы поблагодарить Библиотеку Генезис за демократизацию доступа к знаниям.

Gracias infinitas. Con amor, Rodrigo.

Resumen

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. Este método consiste en dos procesos generales: comienza indentificando 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 y posteriormente creamos *embeddings* de la red utilizando los perfiles anteriores creados y se establecen grupos dentro de la colección. Se muestra la aplicabilidad del método propuesto analizando 75 redes reales generadas en torno a *Trending 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*

Résumé

La capacité de Twitter à relier les utilisateurs autour d'un sujet donné permet de comprendre les mécanismes complexes qui confèrent des positions d'influence à un sous-groupe d'utilisateurs. Cet article se concentre sur le regroupement d'une collection de réseaux thématiques Twitter à l'aide d'une approche interprétable centrée sur les relations asymétriques de la plateforme. Cette méthode est composée par deux processus généraux : on commence par identifier les profils structurels des utilisateurs du réseau à partir d'une représentation du réseau basée sur la présence de sous-graphes dirigés de 2-4 nodes et ensuite on crée des *embeddings* de ce réseau en utilisant les profils créés précédemment et ensuite on établit des *clusters* à l'intérieur de la collection. L'applicabilité de la méthode proposée est illustrée par l'analyse de 75 réseaux réels générés autour de *Trending Topics* au Mexique et par la discussion des profils d'utilisateurs identifiés du point de vue de la dynamique du pouvoir social qu'ils reflètent.

Abstract

Twitter's ability to connect users around a given topic provides insights into the complex mechanisms that grant positions of influence to a subset of users. This paper focuses on clustering a collection of Twitter topic networks using an interpretable approach centred on the platform's asymmetric relationships. This method consists of two general processes: it starts by identifying the structural profiles of the network users from a representation of the network based on the presence of directed subgraphs of 2-4 nodes and then we create *embeddings* of the network using the previous profiles created and clusters are established within the collection. The applicability of the proposed method is shown by analysing 75 real networks generated around *Trending Topics* in Mexico and discussing the identified user profiles from the standpoint of the social power dynamics they reflect.

Índice general

Índice de figuras	xix
1. Introducción	1
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
2. Agrupamiento sobre Redes	13
2.1. Redes	13
2.2. Aprendizaje automático	15
2.2.1. Aprendizaje automático supervisado	16
2.2.2. Aprendizaje automático no supervisado	17
2.3. Agrupamientos en grafos	22
2.3.1. Agrupamientos de nodos	22
2.3.2. Agrupamientos de grafos	23
2.4. <i>Representaciones vectoriales para grafos</i>	25
2.4.1. <i>Embeddings</i> a nivel de nodo	25
2.4.2. <i>Embeddings</i> a nivel de grafo	27
3. <i>Graphlets</i>, Órbitas y Roles Estructurales	29
3.1. <i>Graphlets</i>	29
3.2. Órbitas y firma orbital	30
3.2.1. Ejemplo Karate Club	32

4. Método propuesto	35
4.1. Graphlets y órbitas dirigidas	35
4.2. Perfilar usuarios	37
4.2.1. MiniBatch KMeans	39
4.2.2. Análisis de los perfiles identificados	41
4.2.3. Estabilidad de los perfiles identificados	42
4.3. Agrupar redes	44
4.3.1. Agrupamiento jerárquico	44
4.4. Resumen	47
5. Experimentos y resultados	49
5.1. Conjunto de datos	49
5.2. Primer agrupamiento: perfilando usuarios	50
5.2.1. Estabilidad	52
5.2.2. Perfiles identificados	53
5.3. Segundo agrupamiento: estructura en redes	54
5.4. Visualización de resultados	57
5.5. Discusión	60
6. Conclusiones	67
A. Apéndice	69
Homofilia	69
Función Biyectiva	69
Línea base	69
Bibliografía	73

Índice de figuras

1.1.	Árbol de decisión para clasificar redes temáticas en Twitter, propuesto por Himelboim <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.	13
2.2.	Grafo Dirigido (<i>DiGraph</i>). Podemos observar que la dirección de las aristas esta representada por una flecha que indica en 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 agrupamiento (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.	23
2.7.	Dos agrupamientos de distintas redes de acuerdo a sus propiedades estructurales. En el primer grupo podemos observar <i>egonet-works</i> ; en el segundo se muestran redes más complejas.	24
3.1.	<i>Graphlets</i> de 2, 3 y 4 nodos.	30
3.2.	Ejemplo de roles distintos en los nodos que componen un <i>graphlet</i> de tamaño 3. Los nodos A y C pueden considerarse equivalentes, pero tienen un rol estructural distinto al de B . Este <i>graphlet</i> , G_1 , tendría dos órbitas: una representada en color naranja y otra en color azul.	30
3.3.	<i>Graphlets</i> y órbitas no dirigidas de 2 a 5 nodos.	31

3.4.	Grafo dirigido de 5 nodos.	32
3.5.	Matriz de conteo de órbitas para el grafo 3.4	33
3.6.	Red de Karate Club [Zac77]. Los nodos más influyentes, Mr. Hi, John A. y sus respectivos vecinos a distancia 1 han sido coloreados.	33
3.7.	Comparación del conteo de órbitas normalizado para 4 usuarios de la red <i>Karate Club</i>	34
4.1.	Órbitas en <i>graphlets</i> de hasta 4 nodos. El subgrafo G_i representa un <i>graphlet</i> en la colección; las órbitas dentro de cada <i>graphlet</i> están enumeradas 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.	Ejemplo de dendrograma asociado al agrupamiento jerárquico.	45
4.4.	Resumen de la metodología propuesta.	47
5.1.	Método Elbow o Codo para determinar el tamaño de K	51
5.2.	Exploración de la estabilidad del agrupamiento que se obtiene usando <i>KMeans</i> y $K = 5$. En este ejercicio se calculó la Información Mutua Normalizada entre distintas ejecuciones del algoritmo.	52
5.3.	Perfiles identificados mediante la metodología propuesta.	53
5.4.	Composición de las redes de acuerdo al porcentaje de usuarios de cada perfil encontrado.	55
5.5.	Agrupamiento jerárquico utilizando <i>Ward linkage</i>	56
5.6.	Agrupamiento jerárquico utilizando <i>complete linkage</i>	57
5.7.	Código QR para acceder a la herramienta web.	58
5.8.	Página principal de la sitio web.	58
5.9.	Ejemplo de la visualización de la composición de una red utilizando una gráfica de radar.	59
5.10.	Ejemplo de la visualización de una red (#Coco) dentro de la herramienta web.	60
5.11.	Red #SalarioRosa coloreada respecto al grupo al que pertenece cada nodo en la red.	61
5.12.	Red #Coco coloreada respecto al grupo al que pertenece cada nodo en la red.	62

5.13. Graphlet 10 y órbitas 29 y 30.	63
5.14. Graphlet 8 y órbitas 21 a 24.	64
5.15. Graphlet 0 y órbita 1.	64
5.16. Graphlet 0 y órbita 0.	65

Lista de Algoritmos

1.	Pseudocódigo <i>KMeans</i> [Kub17]	21
2.	Pseudocódigo <i>MiniBatchKMeans</i> [Scu10]	40

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 es particularmente interesante ya 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 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 Zhou 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 *KMeans* [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 ha sido 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 existe 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. Las 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 en específico definido por un *trending topics*. 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 son interesantes ya que contienen la configuración estructural de la discusión pública sobre un tema en específico. Con esta motivación, Himmelboim *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

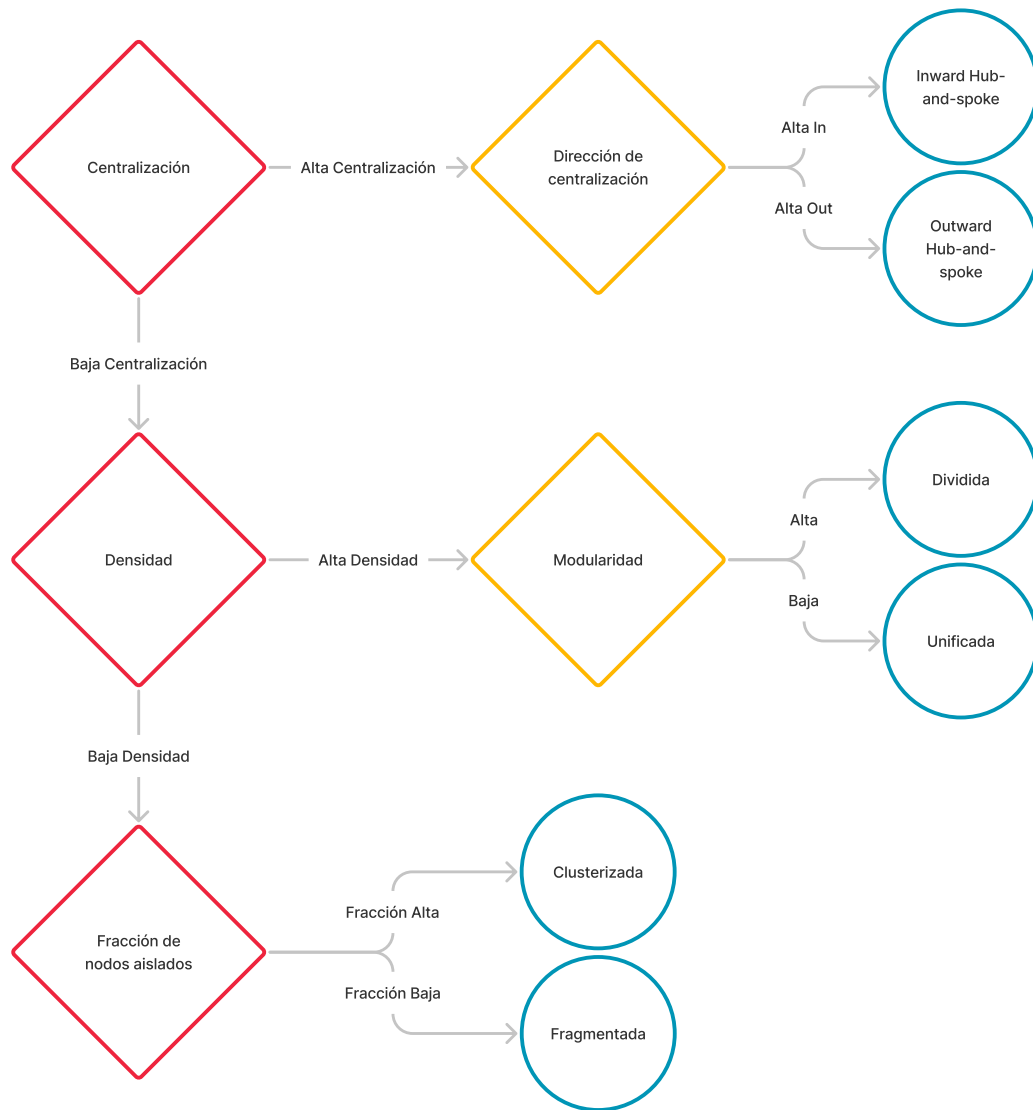


Fig. 1.1. Árbol de decisión para clasificar redes temáticas en Twitter, propuesto por Himelboim *et al.* [Him+17]

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 las redes dentro del contexto de investigación sobre gestión estratégica y cómo estas impactan directamente dentro de las organizaciones (Ver Fig.1.2).

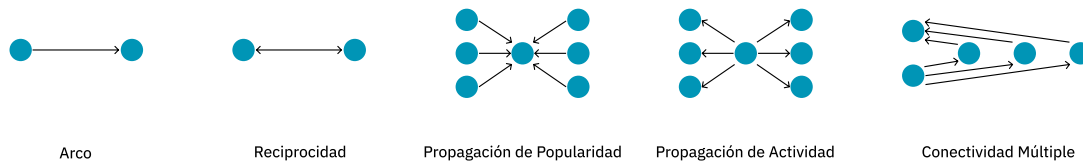


Fig. 1.2. Roles estructurales y su función según Kim *et al.* [Ros+16]

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, es decir, los que están conformados por la menor cantidad de aristas posible. Formalmente se define como

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

dónde σ_{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

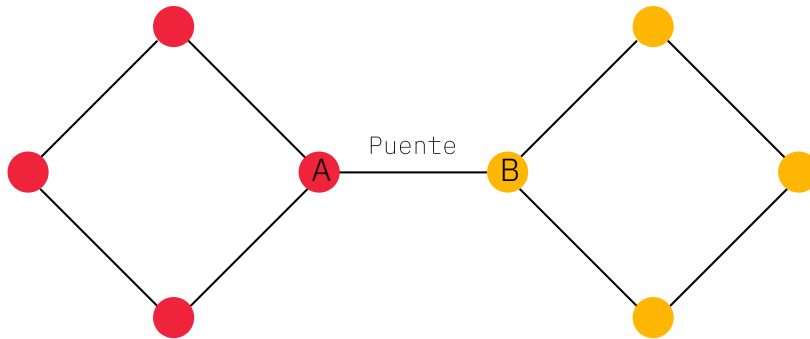


Fig. 1.3. En un grafo se conoce como puente a los nodos que conectan dos grupos, estos nodos tienen una alta intermediación ya que necesariamente por ellos pasan los caminos más cortos entre nodos de ambos grupos.

basándonos en la idea de dichos roles. Para ello, utilizamos las órbitas de *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 metodología propuesta. Posteriormente, en el capítulo 5 se exponen los resultados de los experimentos realizados. Finalmente, en el capítulo 6 encontramos las conclusiones del trabajo, algunas consideraciones del mismo y el trabajo futuro propuesto.

Agrupamiento sobre Redes

En este capítulo se presenta el problema principal y algunos elementos necesarios para comprender tanto su complejidad como una posible solución. Comenzamos con una definición formal de una red y de su representación matemática; posteriormente, se presenta el problema de agrupamiento y su relevancia dentro del aprendizaje automático. Finalmente, el capítulo se enfoca en analizar la tarea de agrupamiento en una colección de redes, describiendo brevemente enfoques y limitaciones para resolver el problema.

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 una red se denomina grafo y es estudiada en matemáticas discretas, específicamente en teoría de grafos. Un grafo (2.1) está formalmente definido como un par de conjuntos $G = (V, E)$, donde V es un conjunto no vacío de nodos (vértices) y E es un conjunto de aristas [Sao21].

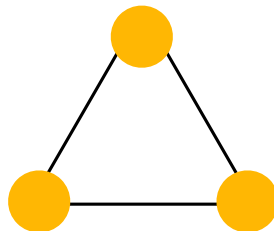


Fig. 2.1. Grafo no dirigido de tres nodos y tres aristas.

Un grafo dirigido o digrafo, es un grafo en el que las aristas tienen direcciones (Ver 2.2). En un sentido más formal, un grafo dirigido es una tripleta $G = (V, E, \phi)$ donde ϕ es una función de incidencia que asigna cada arista a un par ordenado de nodos, es decir, $\phi : E \rightarrow \{(x, y) \mid (x, y) \in V^2 \text{ y } x \neq y\}$

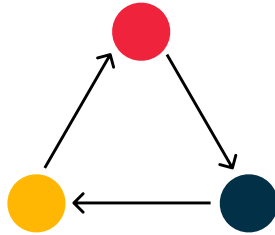


Fig. 2.2. Grafo Dirigido (*DiGraph*). Podemos observar que la dirección de las aristas esta representada por una flecha que indica en donde se origina la arista (inicio de la flecha).

Un subgrafo $H = (V', E')$ de un grafo $G = (V, E)$ es un grafo formado a partir de un subconjunto de nodos y un subconjunto de aristas de G en donde el subconjunto de aristas E' conectan solo los nodos del subconjunto V' y mantienen las mismas relaciones de adyacencia que tenían en el grafo original. Es decir, V' es un subconjunto de V ($V' \subseteq V$) y E' es un subconjunto de E que contiene solo aristas cuyos dos extremos están en V' .

Un subgrafo inducido $H = (V', E')$ de un grafo $G = (V, E)$ es aquel que en donde E' es el conjunto de todas las aristas de G cuyos nodos incidentes están todos en V' . Esto significa que en un subgrafo inducido, los nodos seleccionados forman un subconjunto completamente conectado del grafo original. Formalmente, V' es un subconjunto de V ($V' \subseteq V$) y E' consiste en todas las aristas en E que tienen ambos extremos en V' .

Un isomorfismo de grafos es una biyección de los nodos de un grafo sobre otro, de modo que se preserve la adyacencia de los nodos. Formalmente, el isomorfismo entre dos grafos G y H es una función biyectiva $f : V(G) \rightarrow V(H)$. Esta definición se extiende a las gráficas dirigidas si la función preserva el orden entre cada par de nodos asociados con una arista. Una simetría es un isomorfismo de una gráfica sobre sí misma.

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. Este

es un problema NP y se cree que es NP-Completo, aunque no está demostrado [KST93].

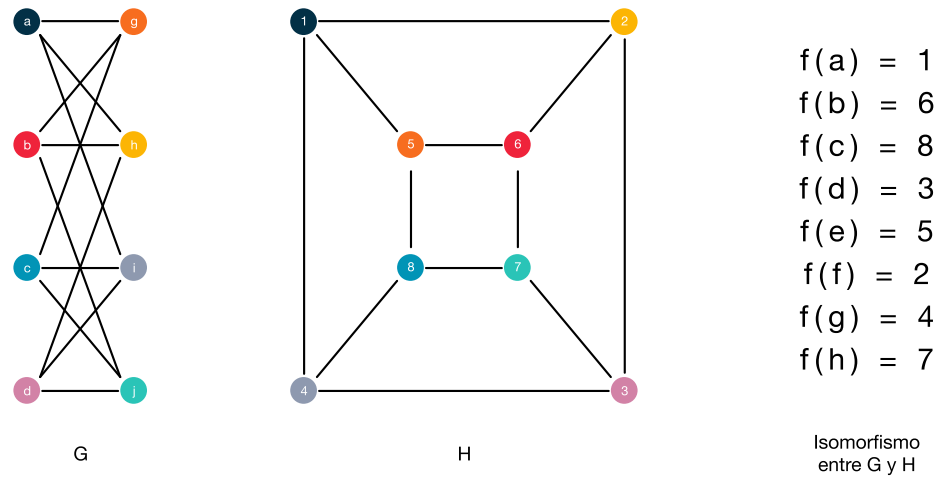


Fig. 2.3. Ejemplo de Isomorfismo entre G y H

2.2 Aprendizaje automático

El aprendizaje automático o aprendizaje de máquina, del inglés *Machine Learning* (ML), es una rama de la Inteligencia Artificial (IA) que estudia algoritmos y técnicas que permiten automatizar soluciones a problemas complejos. Esto se logra a partir del aprendizaje sobre conjuntos de datos.

Como subconjunto de la Inteligencia Artificial, un campo de estudio amplio y diverso que estudia 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 problemas complejos, a pesar de tener requerimientos claros y específicos, puede resultar complicado crear y programar conjuntos de reglas explícitas que representen una solución. Un ejemplo claro podría ser la tarea de detectar objetos en una imagen [RRC19].

Los algoritmos de aprendizaje automático son capaces de resolver problemas de manera implícita, aprendiendo estructuras y reglas a partir de un conjunto de datos en vez de tener una estructura y diseño explícito. La naturaleza de

estos algoritmos hace que dependan directamente de la calidad y cantidad de ejemplos en el conjunto de datos. Dependiendo de la naturaleza de la tarea planteada en el conjunto de datos, encontraremos distintas categorías de algoritmos dentro del aprendizaje automático [RRC19]. Principalmente, se establece una distinción entre los conjuntos de datos etiquetados y no etiquetados.

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 de ejemplos. En el caso de la tarea de clasificación, el conjunto de datos contiene información sobre la clase representada por cada objeto; por ejemplo, una imagen de un perro contiene la etiqueta perro.

Por otro lado, los datos no etiquetados son aquellos que no contienen una etiqueta, es decir que no han sido catalogados de ninguna manera y de los cuales no poseemos más información que el dato en sí. Los datos no etiquetados son, por ejemplo, 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 predecir las etiquetas de datos nuevos a partir del aprendizaje de la relación entre las características y la variable objetivo [RRC19]. En otras palabras, estos algoritmos resuelven problemas a partir de generar un modelo que aprende sobre un conjunto de datos con etiquetas conocidas (datos de entrenamiento) y que después se ejecuta sobre nuevos datos para predecir su etiqueta.

Durante la fase de entrenamiento, el modelo ajusta sus parámetros para minimizar la diferencia entre las predicciones y los valores reales de la variable objetivo en el conjunto de datos de entrenamiento. Para lograr esto, 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 modelo ha sido entrenado y se ha ajustado a los datos, este es capaz de etiquetar datos nuevos que no se habían visto previamente durante el entrenamiento.

Este tipo de algoritmos tienden a ser muy efectivos. No obstante, una consideración importante es que no siempre es clara la manera en la que el problema está siendo resuelto y por lo tanto es complicado interpretar el modelo [RRC19].

Aunque se sabe qué entradas se deben proporcionar y qué salidas se deben esperar, no es sencillo comprender cómo el modelo llega a esas salidas; a este problema se le conoce como el problema de interpretabilidad de un modelo [Zha+21; RRC19].

La interpretabilidad es un problema especialmente presente cuando se usa un enfoque de redes neuronales [Zha+21], que tienen millones de parámetros y vuelven complicado interpretar las decisiones o la serie de reglas que construyen para resolver un problema. Los modelos recientes tienen una cantidad creciente de capas y conexiones que pueden procesar grandes cantidades de datos, y pueden aprender patrones muy sutiles que son difíciles de detectar por los humanos.

La interpretabilidad puede ser un problema en algunas aplicaciones en donde es necesaria la transparencia. Un ejemplo podría ser el área médica, en donde es deseable que los algoritmos que asisten diagnósticos sean auditables y permitan revisar bajo qué criterios se realizó un diagnóstico.

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 de reglas de asociación y búsqueda de agrupamientos [Kub17]. En esta sección nos enfocaremos en la tarea de agrupar, que es quizá la tarea más representativa del aprendizaje no supervisado.

Llevar a cabo una tarea de agrupamiento, o *clustering*, consiste en dividir un gran conjunto de datos (puntos) de tal manera 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.

Los resultados de un agrupamiento pueden ser utilizados como clasificadores o predictores de valores de atributos desconocidos, e incluso como herramientas de visualización [Kub17].

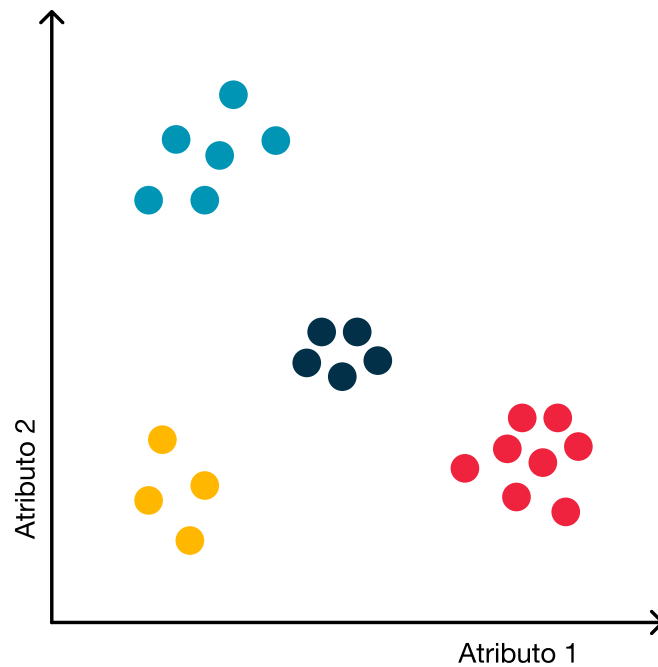


Fig. 2.4. Un ejemplo de agrupamiento (clustering) en R^2 .

Un ejemplo sencillo de agrupamiento en R^2 puede ser el que se muestra en la Fig. 2.4, en el que cada punto representa un ejemplo descrito por dos atributos. Aunque en este caso es sencillo encontrar los agrupamientos a simple vista, para cuatro dimensiones o más no es posible visualizar los datos ni los grupos. A medida que los datos tienen mayor complejidad, establecer grupos sólo puede lograrse mediante algoritmos diseñados para la tarea [Kub17].

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 convenientes.

Una elección común es la distancia Minkowski. Consideremos $X = (x_1, x_2, \dots, x_n)$ y $Y = (y_1, y_2, \dots, y_n) \in R^n$, la distancia Minkowski de orden p , con p un número entero, se define como

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

Uno de los algoritmos más conocidos de agrupamiento es *KMeans*. Este algoritmo agrupa los datos de entrada en K grupos, para una K predefinida por el usuario. La representación matemática de cada uno de los K grupos es un centroide, que es el punto promedio de la distancia entre los puntos del grupo que representa. Por ser el valor promedio de todos los elementos del grupo, un centroide permite una caracterización adecuada del grupo.

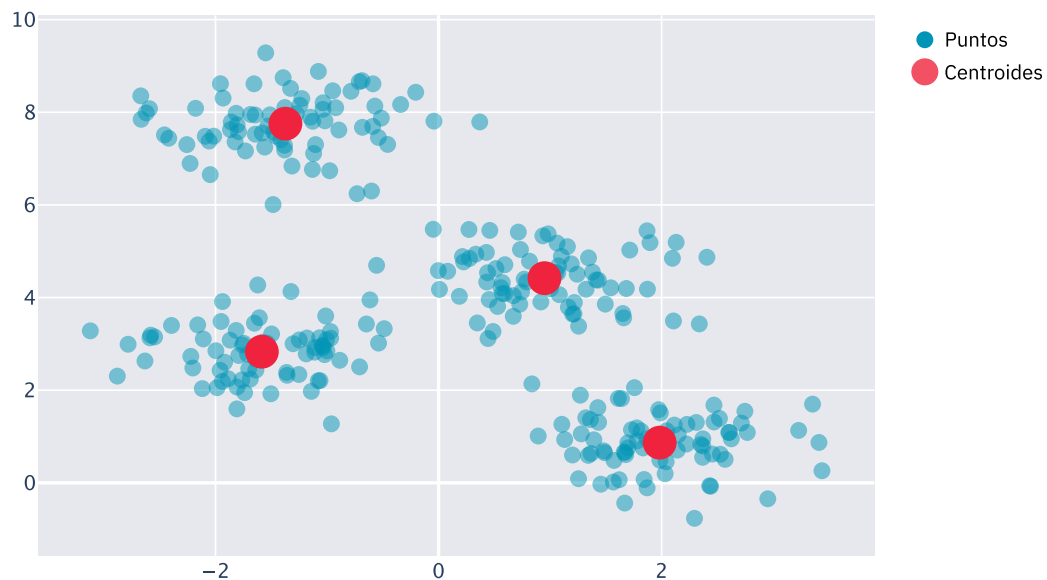


Fig. 2.5. Centroides.

El algoritmo *KMeans* busca minimizar la distancia promedio de cada punto al centroide del grupo al que fue asignado. De manera formal, dado un conjunto de ejemplos (x_1, x_2, \dots, x_n) donde cada ejemplo es un vector d – dimensional, *KMeans* busca agrupar los n ejemplos en $K (<= n)$ grupos $S = S_1, S_2, \dots, S_k$ de tal manera que se minimice el error cuadrático total entre los ejemplos de

entrenamiento y sus centroides correspondientes. Es decir, se busca resolver el problema

$$\arg \min_{\mathbf{s}} \sum_{i=1}^k \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2, \quad (2.1)$$

en donde μ_i representa el centroide del grupo S_i .

La función objetivo en la expresión anterior se conoce como *inertia* o *within-cluster sum-of-squares criterion*.

El problema 2.1 puede resolverse de manera iterativa, siguiendo el Algoritmo 2.2.2. En este enfoque, los centroides pueden ser inicializados de manera aleatoria o con algunas técnicas de inicialización que permitan al algoritmo converger más rápido. Posteriormente, el algoritmo itera recalculando los centroides y los puntos correspondientes a cada grupo, hasta que ya no haya un cambio significativo en la función objetivo o se alcance el número máximo de iteraciones.

Uno de los aspectos negativos de este algoritmo es que es muy susceptible a los centroides iniciales. En la práctica, es común ejecutar el algoritmo varias veces y considerar al mejor resultado, en términos de la función objetivo, como el agrupamiento final.

Algorithm 1: Pseudocódigo *K Means* [Kub17]

Input: Puntos y K

Output: K Grupos

Inicializar K Centroides y K Grupos;

Escoger un punto X y calcular su distancia a cada Centroide C_i ;

Definir el centroide más cercano como C_j ;

while *true* **do**

for *Cada X* **do**

if *X se encuentra en C_j* **then**

 | continuar

end

else

 | Mover X a C_j y recalcular los centroides

end

end

if *los Centroides no cambien de manera significativa o se ha alcanzado un número máximo de iteraciones* **then**

 | false

end

end

2.3 Agrupamientos en grafos

Algoritmos de agrupamiento como KMeans, diseñados para trabajar en conjuntos de vectores, no pueden ser utilizados directamente en grafos. Sin embargo, las numerosas aplicaciones y la necesidad en diferentes contextos de identificar grupos en datos estructurados han derivado en la propuesta de numerosos algoritmos para hacer agrupamientos en grafos.

Conviene señalar que al hablar de agrupamiento en grafos se abarcan dos tareas distintas. La primera consiste en la detección de comunidades, o grupos de nodos, dentro de un solo grafo. La segunda, significativamente menos explorada en la literatura, consiste en realizar agrupamientos a nivel grafo, es decir, identificar grupos de grafos con características en común dentro de una colección.

A continuación se describen brevemente algunas características de estos dos problemas.

2.3.1 Agrupamientos de nodos

Realizar agrupamientos de nodos dentro de una red ha sido un problema ampliamente explorado en años recientes, debido a la gran cantidad de aplicaciones. Este problema se subdivide en dos tareas, conocidas como partición de grafos y detección de comunidades. Aunque ambas tareas se refieren a la división de los nodos de una red en grupos según el patrón de aristas de la red, en la primera (partición) se conoce previamente el número de grupos que se busca, mientras que en la segunda (detección de comunidades), determinar el número de grupos es parte del problema [New10].

La partición de grafos es un problema estudiado desde 1960 [New10], y se enfoca en dividir los nodos de un grafo en n grupos de tal manera que las aristas entre grupos sean las menores posibles. Al número de aristas entre cada grupo se le llama *tamaño de corte* (*cut size*).

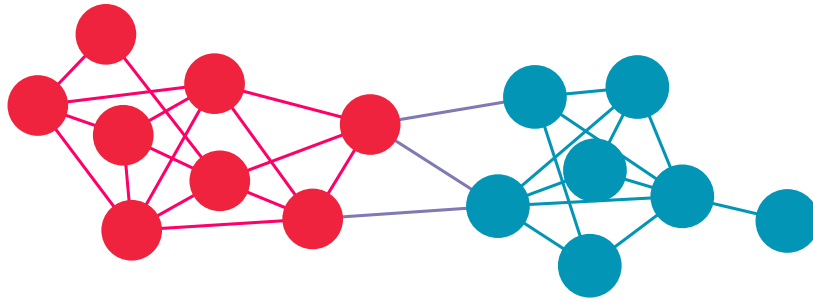


Fig. 2.6. Nodos de una red divididos en 2 grupos donde el color del nodo representa el grupo al que pertenece.

Por otro lado, 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. Esta tarea nos permite estudiar la estructura y organización de una red.

2.3.2 Agrupamientos de grafos

Encontrar grupos dentro de una colección de grafos es un problema menos estudiado que el agrupamiento de nodos.

Comparar redes que podrían tener diferente tamaño y orden requiere basarse en propiedades estructurales. Este es un problema con aplicaciones de importancia, pero que puede ser computacionalmente costoso.

En general, agrupar grafos requiere de dos herramientas: 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.

Entre 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]. A continuación se describen dos enfoques populares para comparar grafos.

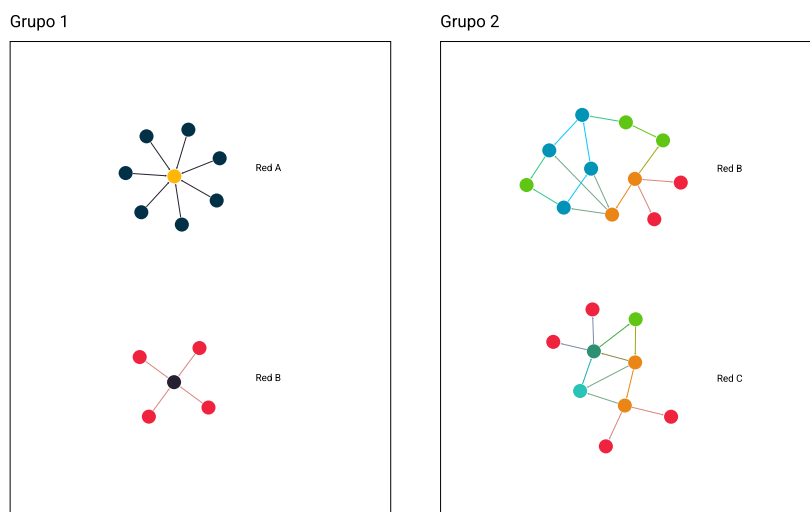


Fig. 2.7. Dos agrupamientos de distintas redes de acuerdo a sus propiedades estructurales. En el primer grupo podemos observar *egonetworks*; en el segundo se muestran redes más complejas.

Medidas estructurales. Es posible establecer una comparación entre dos grafos basada en propiedades de los mismos. Por ejemplo, el trabajo «*Classifying Twitter Topic-Networks Using Social Network Analysis*» [Him+17], relacionado muy de cerca a nuestra propuesta en cuanto a objetivo, utiliza las medidas de centralidad, centralización, densidad, modularidad y fracción de *clusters* e *Isolates* para establecer categorías de redes dentro de una colección, agrupando así conforme a características estructurales. Desafortunadamente este enfoque puede ser limitado en cuanto a la distinción que establece entre redes. Dependiendo de los indicadores considerados y del contexto del problema, el enfoque podría no capturar con suficiente detalle la estructura de las redes que se están comparando.

Distancia de Edición (Graph Edit Distance, GED). Esta es una de las alternativas más conocidas para comparar dos grafos. La GED mide el número de cambios (inserciones y remociones de nodos y/o aristas) necesarios para llegar a la estructura del grafo *B* partiendo desde el grafo *A*.

2.4 Representaciones vectoriales para grafos

Como se mencionó en la Sección 2.2, la mayoría de los algoritmos clásicos de aprendizaje automático no pueden ser utilizados directamente en redes, pues están diseñados para elementos de un espacio vectorial. Una de las estrategias recientes para resolver este problema es extraer características de los nodos o del grafo entero y utilizarlas para crear una representación vectorial. De esta manera es posible utilizar medidas clásicas de distancia en este espacio y aplicar algoritmos conocidos de aprendizaje de máquina.

El proceso de extracción de características es llamado *representation learning*, y la representación vectorial se conoce como *embedding*. En 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, las principales se basan en vecindario, rol estructural o atributos. En el caso del *embedding* de un grafo, la extracción de características puede dividirse en dos principales categorías: técnicas basadas en la estructura global de la red y técnicas basadas en subestructuras de la red.

2.4.1 Embeddings a nivel de nodo

Los *embeddings* a nivel nodo han sido ampliamente explorados en años recientes, por ejemplo, grandes empresas tecnológicas los han utilizado para perfilar enormes cantidades de usuarios dentro de distintas redes sociales [Ler+19]. Otras aplicaciones también abarcan el área de biología y química [Yue+20].

La idea principal en los *embeddings* de nodos basados en vecindario es obtener el *embedding* de cada nodo en función de los nodos vecinos y sus respectivos *embeddings*. Una lista de algoritmos que siguen esta idea puede encontrarse en la Tabla 2.1.

Por otro lado, los métodos basados en roles estructurales buscan generar un *embedding* a partir de la extracción de características de los roles que

toma cada nodo dentro de la red. Algoritmos que siguen este enfoque pueden encontrarse en la Tabla 2.2.

Embeddings de nodos basados en el vecindario	
Publicación	Algoritmo
“Relational Learning via Latent Social Dimensions” [TL09]	SocioDim
“Billion-scale Network Embedding with Iterative Random Projection” [Zha+18]	RandNE
“GLEE: Geometric Laplacian Eigenmap Embedding” [TCE20]	GLEE
“Diff2Vec: Fast Sequence Based Embedding with Diffusion Graphs” [RS18]	Diff2Vec
“NodeSketch: Highly-Efficient Graph Embeddings via Recursive Sketching” [Yan+19]	NodeSketch
“Network Embedding as Matrix Factorization: Unifying DeepWalk LINE PTE and Node2Vec” [Qiu+18]	NetMF
“Multi-Level Network Embedding with Boosted Low-Rank Matrix Approximation” [Li+19]	BoostNE
“Don’t Walk, Skip! Online Learning of Multi-scale Network Embeddings” [Per+17]	Walklets
“GraRep: Learning Graph Representations with Global Structural Information” [CLX15]	GraRep
“DeepWalk: Online Learning of Social Representations” [PAS14]	DeepWalk
“node2vec: Scalable Feature Learning for Networks” [GL16]	Node2Vec
“Alternating Direction Method of Multipliers for Non-Negative Matrix Factorization with the Beta-Divergence” [SF14]	NMFADMM
“Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering” [DBG02]	LaplacianEigenmaps

Tab. 2.2. Algunos ejemplos de algoritmos basados en el enfoque de embeddings de nodos basados en roles estructurales.

Embedding estructural de nodos	
Publicación	Algoritmo
“Learning Structural Node Embeddings Via Diffusion Wavelets” [Don+18]	GraphWave
“Learning Role-based Graph Embeddings” [Ahm+22]	Role2Vec

Tab. 2.3.

Embedding de grafos	
Publicación	Algoritmo
“Graph2Vec: Learning Distributed Representations of Graphs” [Nar+17]	Graph2Vec
“A Simple Baseline Algorithm for Graph Classification” [LP18]	SF
“NetLSD: Hearing the Shape of a Graph” [Tsi+18]	NetLSD
“GL2vec: Graph Embedding Enriched by Line Graphs with Edge Features”	GL2Vec
“Geometric Scattering for Graph Data Analysis” [GWH]	GeoScattering
“Invariant Embedding for Graph Classification” [GL]	IGE

2.4.2 Embeddings a nivel de grafo

Para construir representaciones vectoriales de un grafo es posible extraer características de la red completa o de estructuras locales dentro de ella.

En el primer caso, se utilizan propiedades enfocadas en la topología general de la red o características *globales*. La Tabla 2.3 muestra algunos ejemplos de algoritmos que utilizan este enfoque.

En contraste, existen algoritmos que están enfocados en identificar y resumir estructuras locales dentro de la red, es decir, relaciones entre nodos. Los algoritmos que extraen características de los nodos examinan las estructuras locales haciendo uso de subgrafos, por ejemplo, *EgoNetworks* o *graphlets*.

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.

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 grandes. No obstante al igual que otras algoritmos de esta clase heredan las problemáticas de interpretación de las redes neuronales.

El problema que se aborda en este trabajo tiene como principal objetivo explorar estructuras dentro de redes sociales, comprendiendo el papel que juegan diferentes usuarios. La interpretabilidad es clave en nuestro problema, pues no sólo queremos agrupar redes sino interpretar los motivos y entender qué características comparten redes que son agrupadas. Por esta razón, no usaremos *embeddings* generados con redes neuronales, sino basados en estructuras locales de la red. Específicamente, usaremos *graphlets*, que se detallan en el siguiente capítulo.

Graphlets, Órbitas y Roles Estructurales

3.1 Graphlets

Los *graphlets* son subgrafos que pueden identificarse de manera inducida en una red mas 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].

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].

A partir de ese trabajo, se ha utilizado la enumeración de *graphlets* de tamaño n para estudiar la estructura de redes. Es decir, dada una red G , se observa el subgrafo que se forma en cada posible combinación de n nodos conexos dentro de G y se realiza un conteo de las estructuras observadas.

Así, podemos pensar en los *graphlets* como una colección o diccionario de todas las clases de isomorfismo de subgrafos de hasta un tamaño fijo n . La Fig. 3.1 muestra los *graphlets* correspondientes a $n = 2$, $n = 3$ y $n = 4$. Analizar una red usando *graphlets* de tamaño máximo $n = 4$ consistiría en identificar cada una de las estructuras que se muestran en la figura y contar cuántas veces aparece cada una. Este conteo después puede utilizarse para analizar propiedades de la red, de la misma manera en que se utiliza el grado. De hecho, el perfil que tiene una red respecto al conteo de *graphlets* puede considerarse una generalización de la distribución de grado [Sar+16].

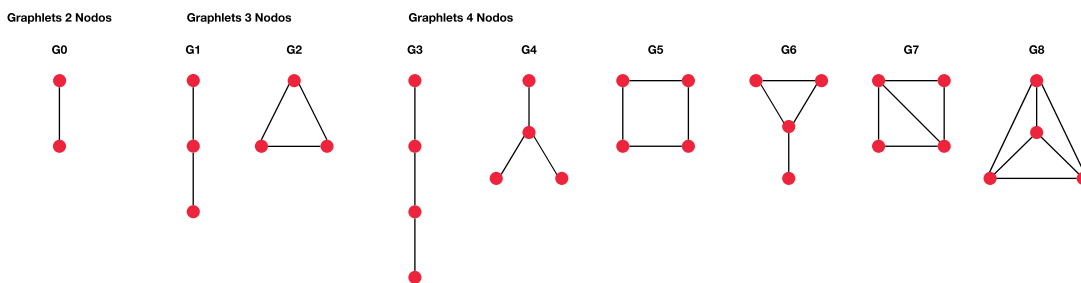


Fig. 3.1. *Graphlets* de 2, 3 y 4 nodos.

La relevancia de los *graphlets* va más allá de la comparación de 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 Órbitas y firma orbital

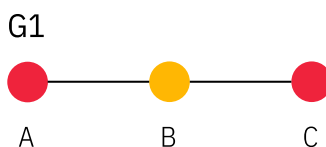


Fig. 3.2. Ejemplo de roles distintos en los nodos que componen un *graphlet* de tamaño 3. Los nodos *A* y *C* pueden considerarse equivalentes, pero tienen un rol estructural distinto al de *B*. Este *graphlet*, G_1 , tendría dos órbitas: una representada en color naranja y otra en color azul.

En los *graphlets* es posible reconocer diferentes roles de nodos. Por ejemplo, en la Fig. 3.2, el nodo *B* juega un papel claramente distinto al de *A* y *C*, pues es un extremo para cada arista que aparece en el *graphlet* y tiene un rol central. En contraste, *A* y *C* son nodos que tienen un papel equivalente, en la periferia del grafo. A cada papel, o rol estructural, que se puede identificar dentro de un *graphlet* se le llama órbita.

Dicho de una manera más formal, las órbitas son las posiciones posibles que un nodo puede tomar en un *graphlet* al reetiquetar todos sus nodos de forma que se preserven las relaciones ordenadas de adyacencia [Sar+16]. La Fig. 3.3 ilustra todas las posibles órbitas que existen en la colección de *graphlets*

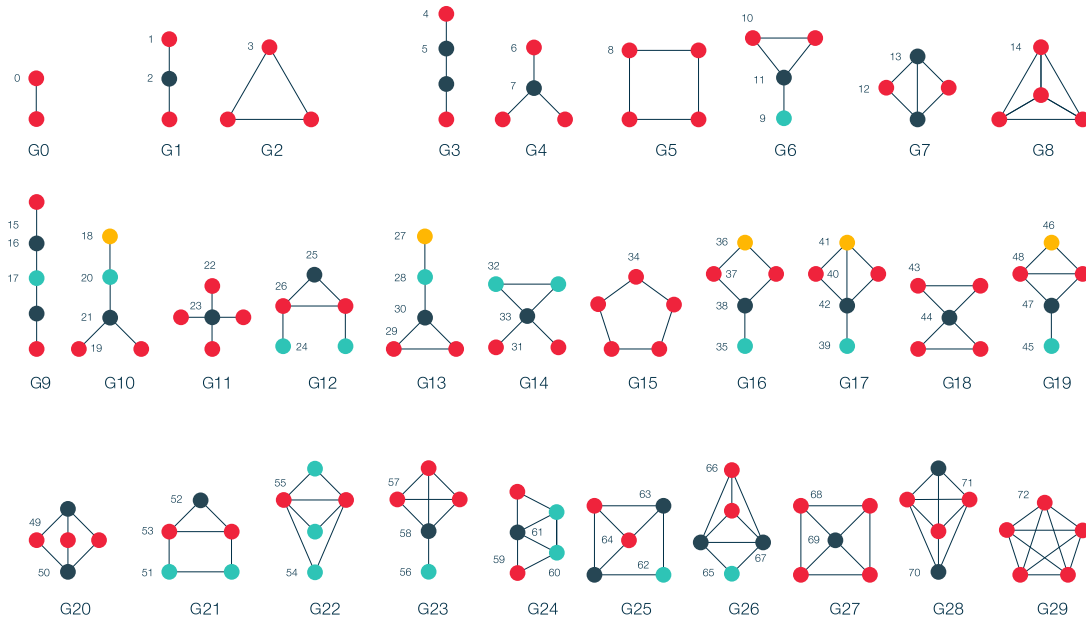


Fig. 3.3. *Graphlets* y órbitas no dirigidas de 2 a 5 nodos.

de dos o más puntos con tamaño máximo $n = 5$; los colores en cada nodo identifican roles distintos (o equivalentes) dentro de un *graphlet*.

Tomando en cuenta la lista de posibles órbitas, podríamos analizar un nodo v en una red y contar cuántas veces aparece en cada órbita al considerar todos los posibles *graphlets* de los que forma parte.

Una firma orbital es eso: el conteo de las posiciones orbitales de un nodo. Si, por ejemplo, consideramos *graphlets* de tamaño máximo $n = 5$, la firma orbital de un nodo v sería un vector con 73 componentes, de manera que la i -ésima componente del vector represente las veces que v aparece en la órbita i . Este vector, o firma orbital del nodo, logra describir la topología del nodo y su vecindario, y captura sus interconexiones hasta una distancia $n = 5$, incluso de puntos aislados que tendrán ceros en todas las entradas [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].

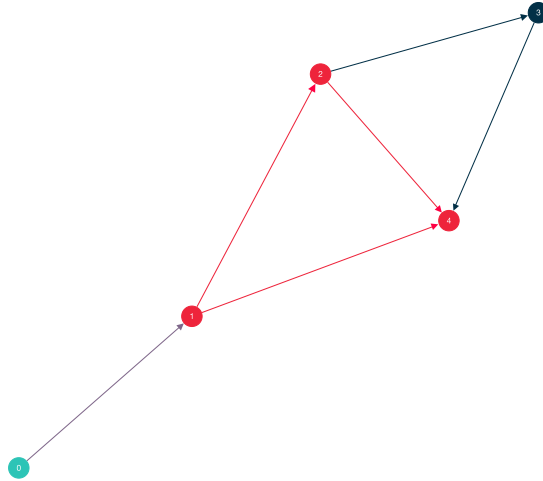


Fig. 3.4. Grafo dirigido de 5 nodos.

En la Fig. 3.4 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 Fig. 3.5.

3.2.1 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.3.6) se convirtió en un estándar para el estudio de algoritmos y a menudo se utiliza como referencia.

En la Fig. 3.7 podemos observar los *embeddings* para 4 nodos de la red Karate Club. Para contrastar los *embeddings* de distintos tipos de nodos en la red 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. 3.3. 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.

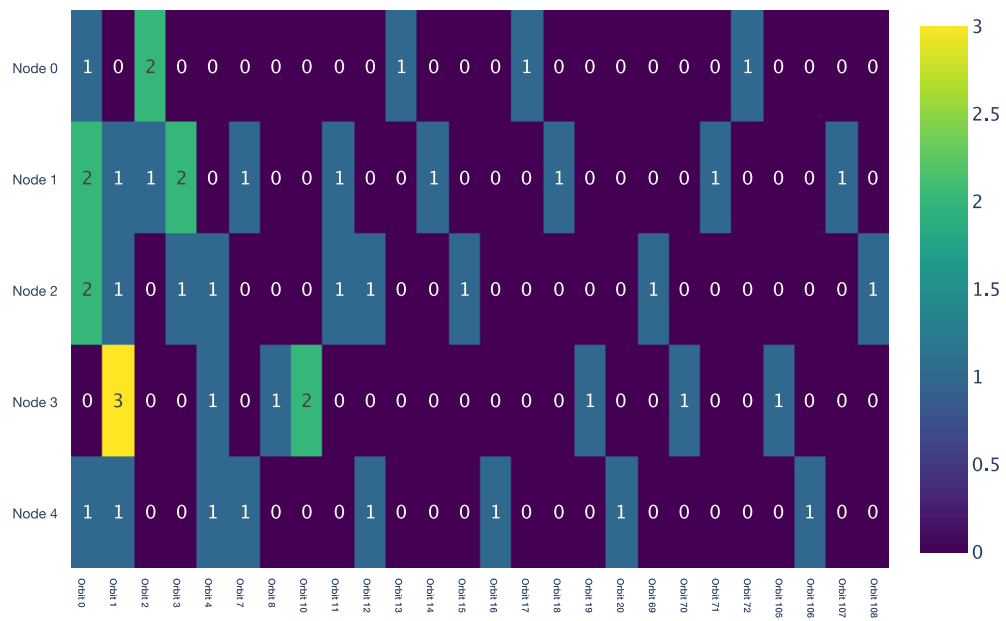


Fig. 3.5. Matriz de conteo de órbitas para el grafo 3.4

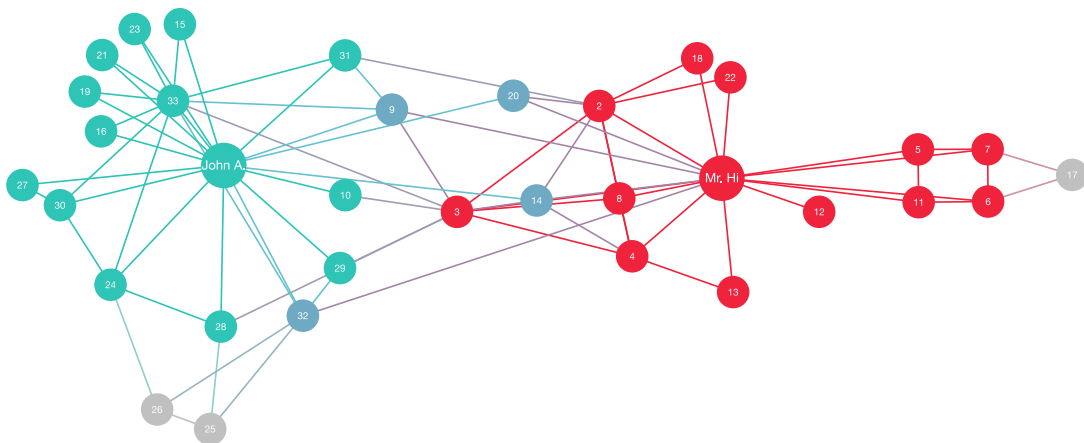


Fig. 3.6. Red de Karate Club [Zac77]. Los nodos más influyentes, Mr. Hi, John A. y sus respectivos vecinos a distancia 1 han sido coloreados.

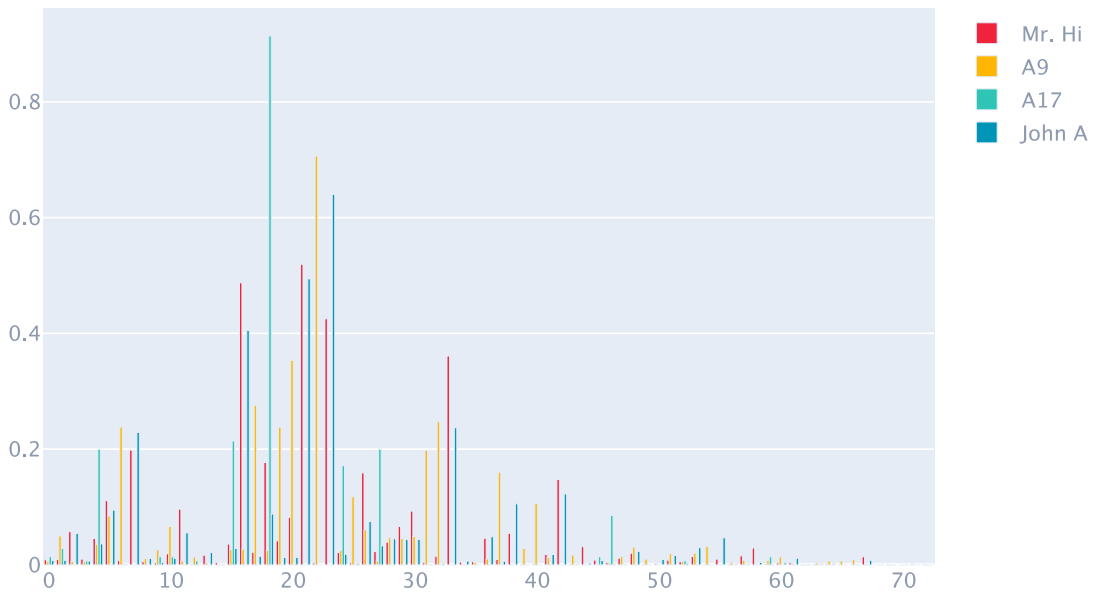


Fig. 3.7. Comparación del conteo de órbitas normalizado para 4 usuarios de la red *Karate Club*.

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.

Método propuesto

El agrupamiento de una colección de grafos no es un problema sencillo. El uso de algoritmos de agrupación populares, como KMeans, requiere representar los grafos en un espacio vectorial. Esta tarea puede llevarse a cabo mediante métodos que van desde la extracción de características 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 el método propuesto 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 orbital de cada nodo en un análisis basado en *graphlets*. Después, agrupa las redes con base en la distribución de perfiles de nodos que presentan.

4.1 Graphlets y órbitas dirigidas

Sarajilic *et al.* propusieron extender la idea una firma orbital 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. Con estas consideraciones, 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.

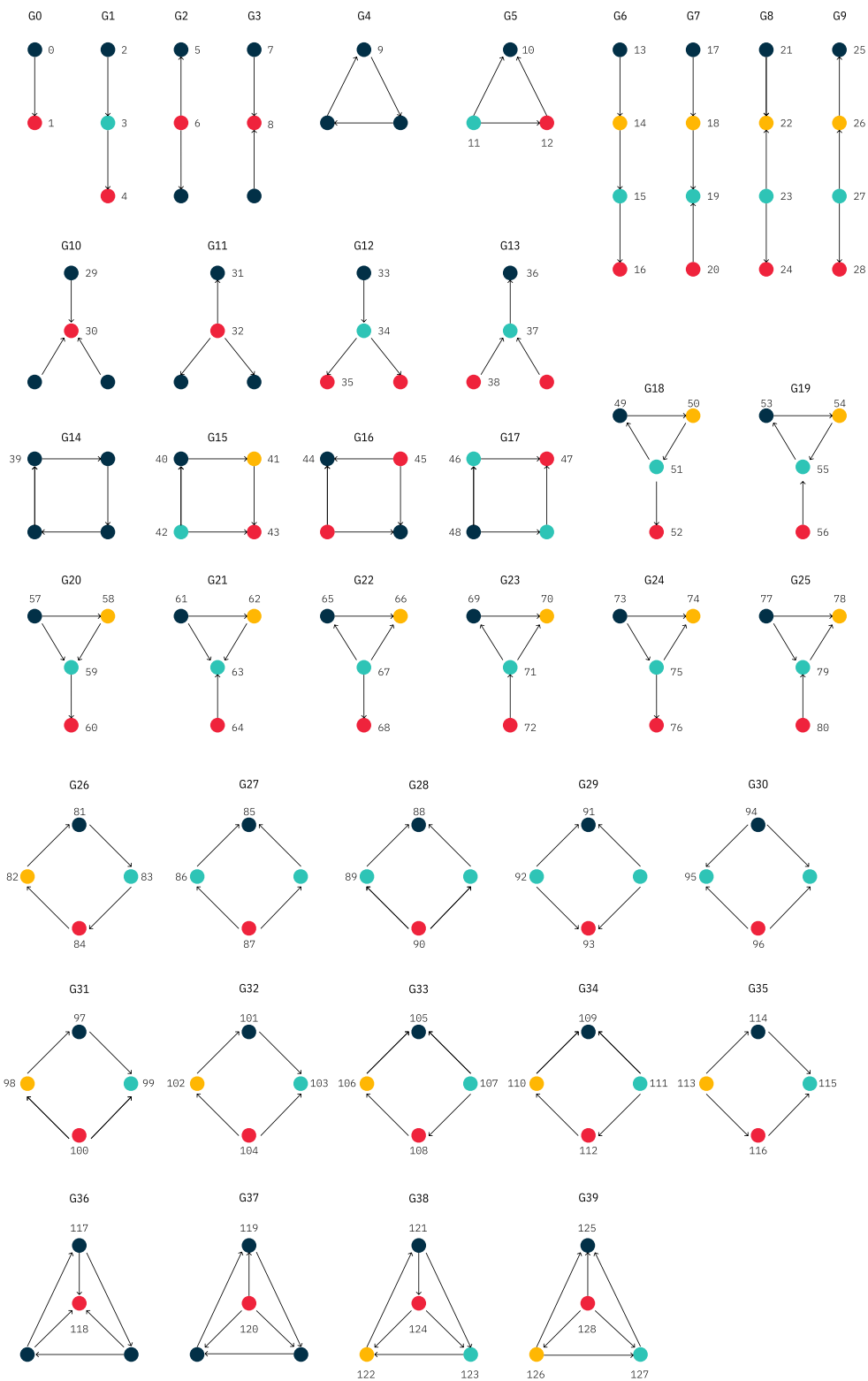


Fig. 4.1. Órbitas en *graphlets* de hasta 4 nodos. El subgrafo G_i representa un *graphlet* en la colección; las órbitas dentro de cada *graphlet* están enumeradas para futuras referencias en este trabajo.

4.2 Perfilar 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 hacen posible, 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 2.3.1, es posible realizar agrupamientos en una red basados en los diversos roles estructurales de los nodos que la componen. En el caso de una red social, 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, y 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 de esta tesis es que, en el contexto de redes sociales, la firma orbital de un nodo basada en *graphlets* 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

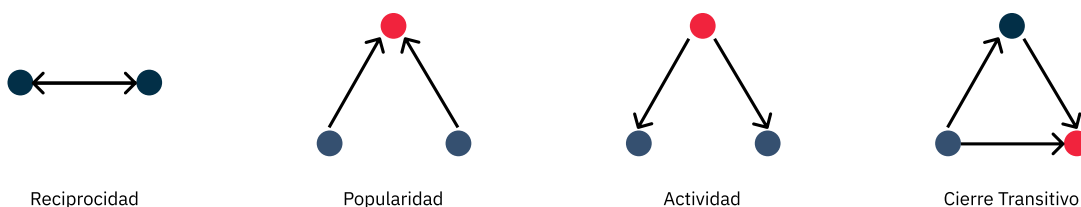


Fig. 4.2. Algunos patrones propuestos por Lusher y Robins para describir configuraciones sociales dentro de procesos colectivos [Lus]. Las aristas dirigidas permiten la distinción entre jerarquías y posiciones de poder dentro de la red.

sobre las dinámicas sociales, la jerarquía que se establece entre personas y la estructura general 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 KMeans (Algoritmo 2.2.2) 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 KMeans [Scu10], que es una de las distintas modificaciones de KMeans propuestas para lidiar con limitaciones de tiempo y memoria. El algoritmo se describe a continuación.

4.2.1 MiniBatch KMeans

A pesar de la enorme popularidad de KMeans 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 KMeans [Scu10] es una versión modificada de KMeans 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 KMeans 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 KMeans 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 4.2.1 muestra el pseudocódigo de MiniBatch KMeans y sus particularidades, entre ellas el muestreo M de ejemplos aleatorios y el cálculo de la función objetivo (distorsión).

Algorithm 2: Pseudocódigo *MiniBatchKMeans* [Scu10]

Input: Puntos $X = \{x_1, x_2 \dots x_n\}$, Cantidad de grupos K , Tamaño del MiniBatch b , iteraciones t

Output: K Grupos $C = \{c_1, c_2 \dots c_k\}$

Inicializar K centroides y K Grupos;
 N_{C_i} Inicializar el número de muestra para cada grupo;

for j en rango t **do**

- Definir M que es el batch con b ejemplos aleatorios de X ;
- for** m en rango b **do**
 - $C_i(x_m) = \frac{1}{|C_i|} \sum x_m$; /* Calcular centroide */
- end**
- for** m en rango b **do**
 - $c_i = C_i(x_m)$; /* Obtener centroide */
 - $N_{C_i} = N_{C_i} + 1$; /* Actualizar el número de muestra */
 - $lr = 1/N_{C_i}$; /* Calcular tasa de aprendizaje */
 - $C_i = (1 - lr)c_i + lr * x_m$; /* Utilizar la tasa de aprendizaje para actualizar el centroide */
- end**

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 de un grupo establecido.

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 n de un *graphlet*, el número de aristas dirigidas que terminan en él se denomina grado de entrada (*in-degree*) de n . Se denota como $indegree(n)$
- Grado de salida: El número de aristas dirigidas que comienzan en el nodo de grafo es su grado de salida (*out-degree*). Se denota como $outdegree(n)$.
- Fuente: Un nodo se considera una fuente en un grafo si tiene un grado *in-degree* de 0. Es decir, n tal que $indegree(n) = 0$.
- Pozo: Un nodo n tal que $outdegree(n) = 0$.
- Camino dirigido en un *graphlet*: Una secuencia finita de aristas en una secuencia de distintos nodos de tal manera que todas las aristas tengan 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 una simetría, podemos extender las definiciones de fuente y de pozo de los nodos a las órbitas.

Considerando las relaciones establecidas anteriormente respecto a las posibles interacciones en *Twitter* (Ver. 1.1 se pueden establecer roles en función de las órbitas.

Podemos decir que 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 otro nodo $n \in \mathcal{O}$, n es un pozo con $\text{indegree}(n) > 1$.

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 m en una arista que comienza en n es un hablante, es decir un pozo con $\text{indegree}(m) > 1$. 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 n participa en diferentes *graphlets* dentro de un red; cada *graphlet* nos da información sobre el 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 $\text{indegree}(n)$ o $\text{outdegree}(n)$. 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 (ver Fig. 4.1) que nos permitirá distinguir entre dos de los perfiles que se describirán en la sección de resultados.

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

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

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 está 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 las corridas.

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 en esa red.

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.

4.3.1 Agrupamiento jerárquico

Una vez que se tiene la representación vectorial de cada red, utilizamos agrupamiento 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 agrupamiento jerárquico, la estructura, o jerarquía de grupos, se determina de manera avara y comúnmente se presenta en un dendrograma (Fig. 4.3). Los resultados dependen de una medida de distancia entre las instancias del conjunto y un criterio de distancia para subconjuntos de datos.

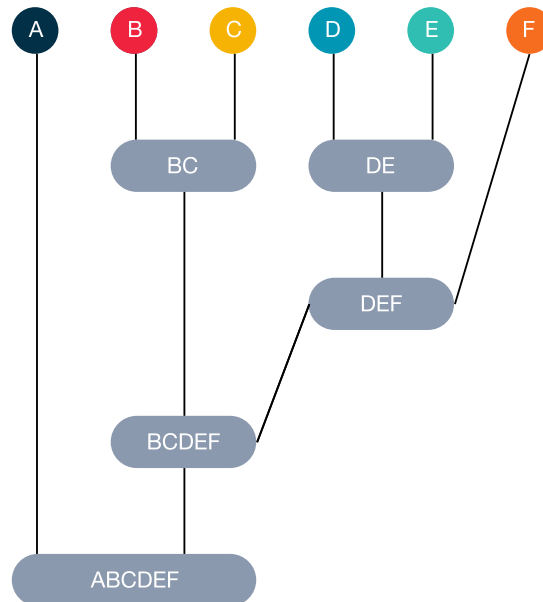


Fig. 4.3. Ejemplo de dendrograma asociado al agrupamiento jerárquico.

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 $d(u, v)$ entre dos grupos u y v ; los criterios más comunes 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 u y v con mínima distancia entre ellos se unirá para formar un nuevo cluster w . 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

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

Tab. 4.1. Criterios de encadenamiento (*linkage*) para calcular la distancia entre dos grupos en el agrupamiento jerárquico aglomerativo.

Nombre	Función
<i>Single</i>	$d(u, v) = \min(\text{dist}(u[i], v[j]))$
<i>Complete</i>	$d(u, v) = \max(\text{dist}(u[i], v[j]))$
<i>Average</i>	$d(u, v) = \sum_{ij} \frac{d(u[i], v[j])}{(u * v)}$
<i>Weighted</i>	$d(u, v) = \frac{\text{dist}(s, v) + \text{dist}(t, v)}{2}$
<i>Centroid</i>	$d(u, v) = \ c_s - c_t\ _2$
<i>Ward</i>	$d(u, v) = \sqrt{\frac{ v + s }{T} d(v, s)^2 + \frac{ v + t }{T} d(v, t)^2 - \frac{ v }{T} d(s, t)^2}$ <p>dónde $T = v + s + t$</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.

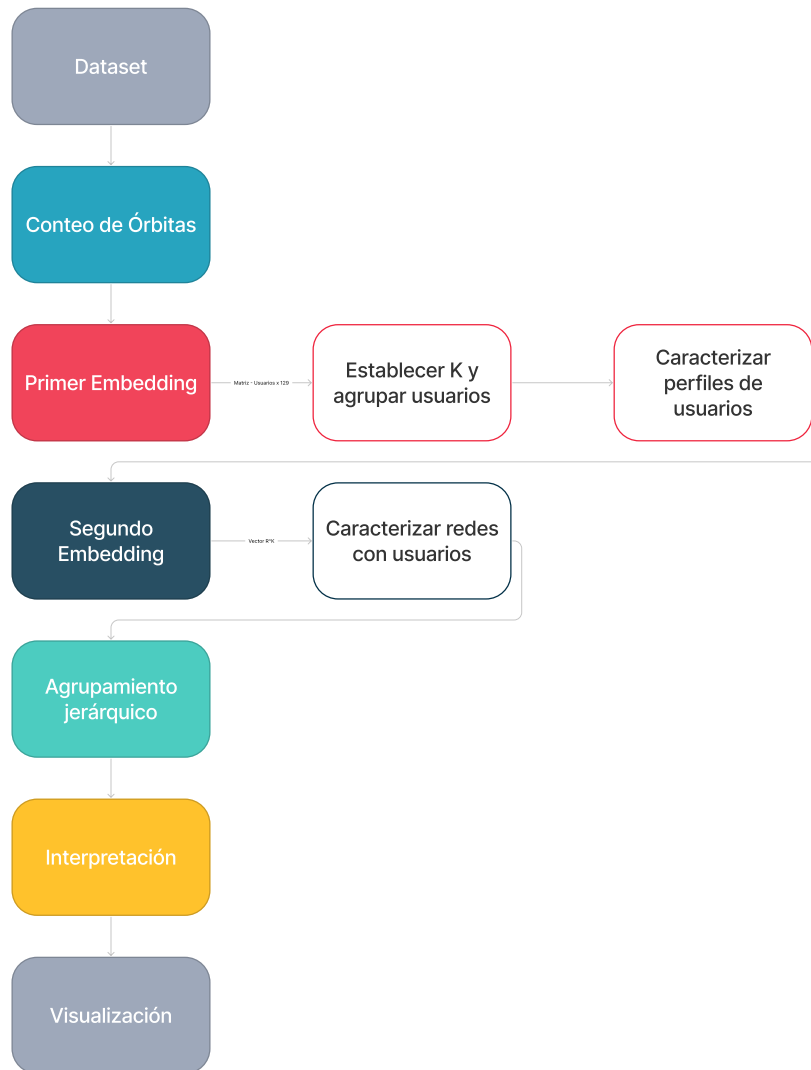


Fig. 4.4. Resumen de la metodología propuesta.

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 a 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. Consideramos que en una arista de A a B , el usuario A (origen) está reaccionando a una publicación del usuario B (destino). No se utilizan etiquetas para distinguir

entre reacciones, por lo que ambas interacciones, responder y retuitear, 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.

Posteriormente, decidimos cuántos perfiles establecer para los usuarios. Con este objetivo, analizamos el conjunto de vectores-usuario con el método del codo, centrándonos 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].

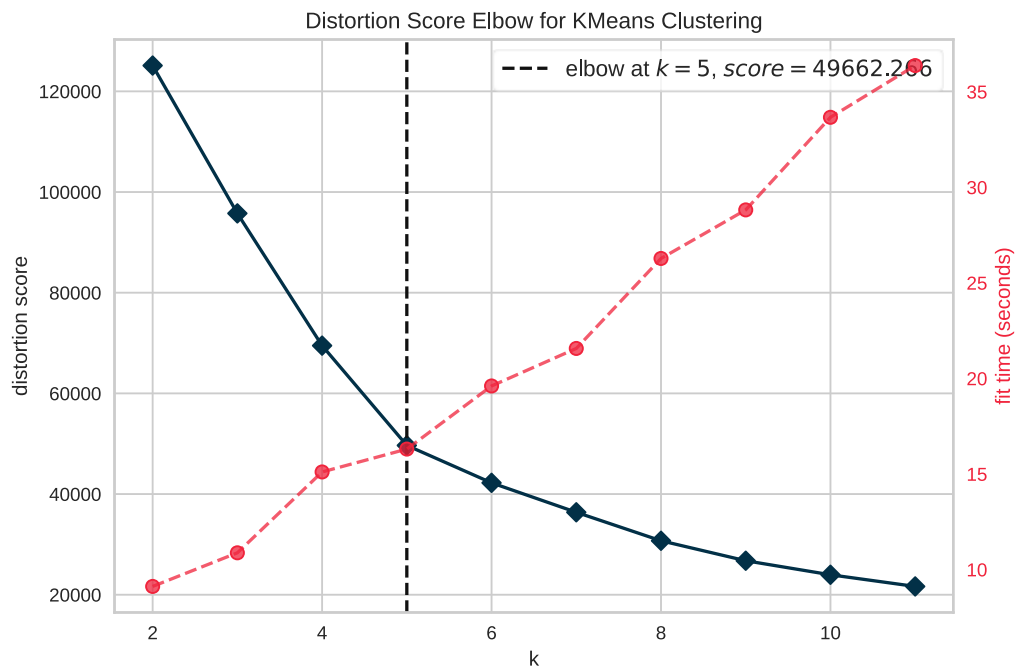


Fig. 5.1. Método Elbow o Codo para determinar el tamaño de K

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) (Ver. 4.2.3) por pares de las ejecuciones se muestra en la Fig. 5.2 donde los pares están denotados por los ejes x y y , los colores más claros representan valores más cercanos a 1 y de manera inversa los colores más oscuros representan valores más cercanos a cero. Es importante notar que la Información Mutua Normalizada de todos los pares en la diagonal de la matriz $R \times R$ es igual a 1 dado que es $I(C_i, C_j)$ en dónde $i = j$.

El valor medio del total de ejecuciones fue de 0.93.

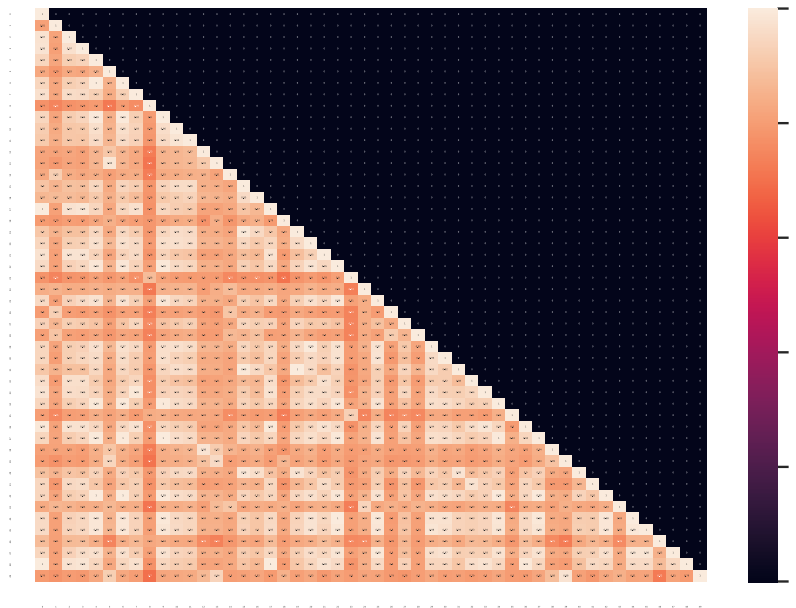


Fig. 5.2. Exploración de la estabilidad del agrupamiento que se obtiene usando *KMeans* y $K = 5$. En este ejercicio se calculó la Información Mutua Normalizada entre distintas ejecuciones del algoritmo.

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. 5.3 muestra las principales órbitas para cada grupo.

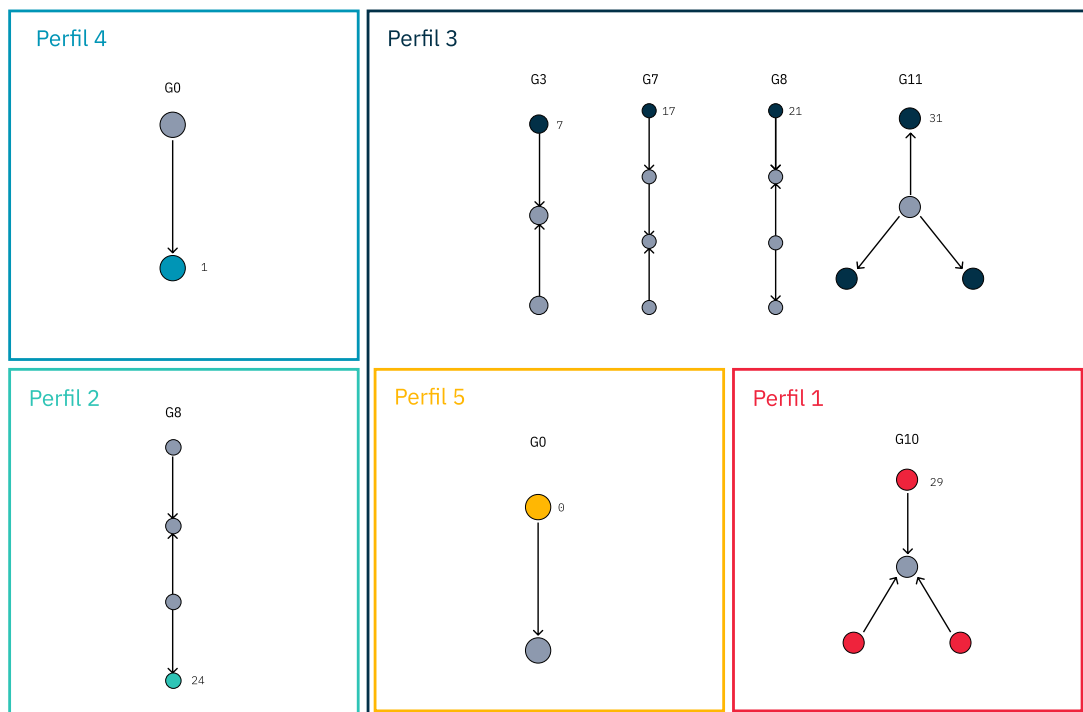


Fig. 5.3. Perfiles identificados mediante la metodología propuesta.

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.

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

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 1, 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.

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

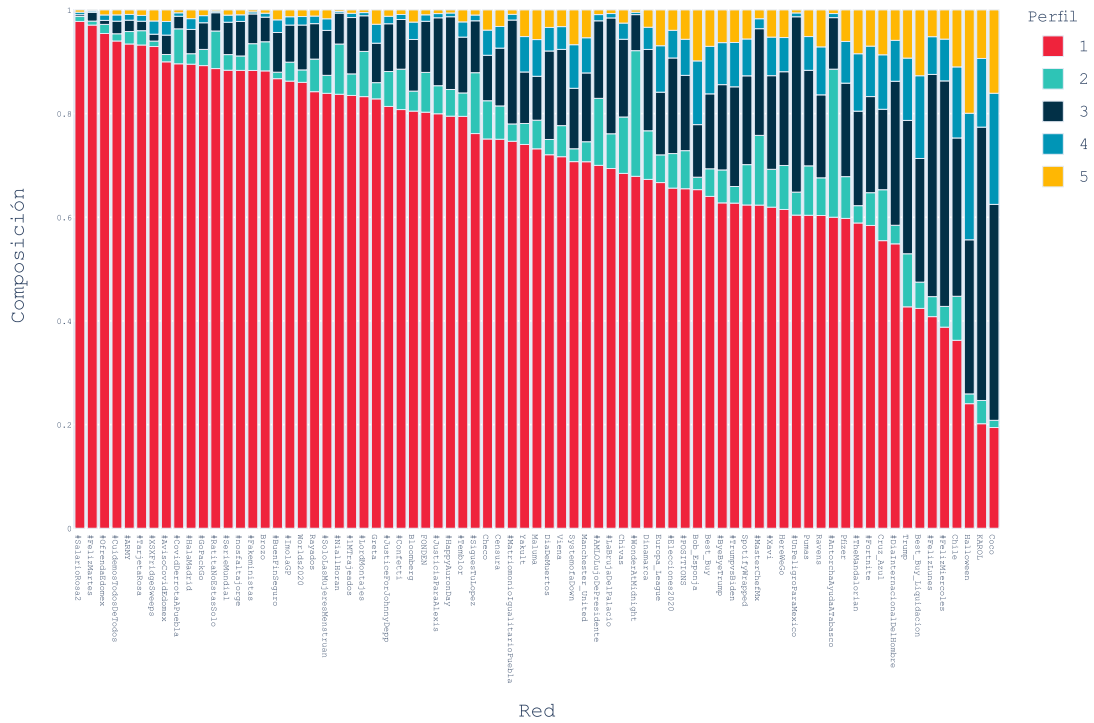


Fig. 5.4. Composición de las redes de acuerdo al porcentaje de usuarios de cada perfil encontrado.

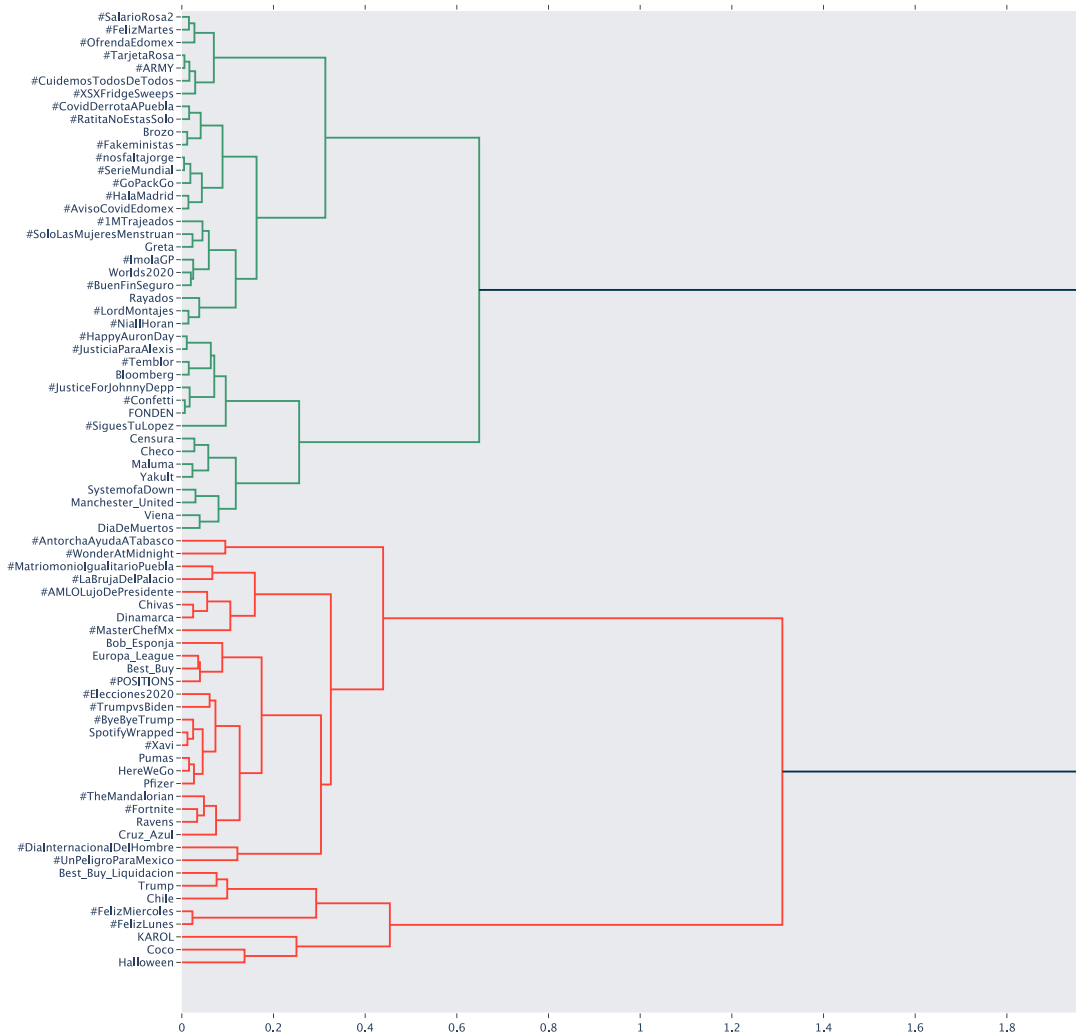


Fig. 5.5. Agrupamiento jerárquico utilizando *Ward linkage*

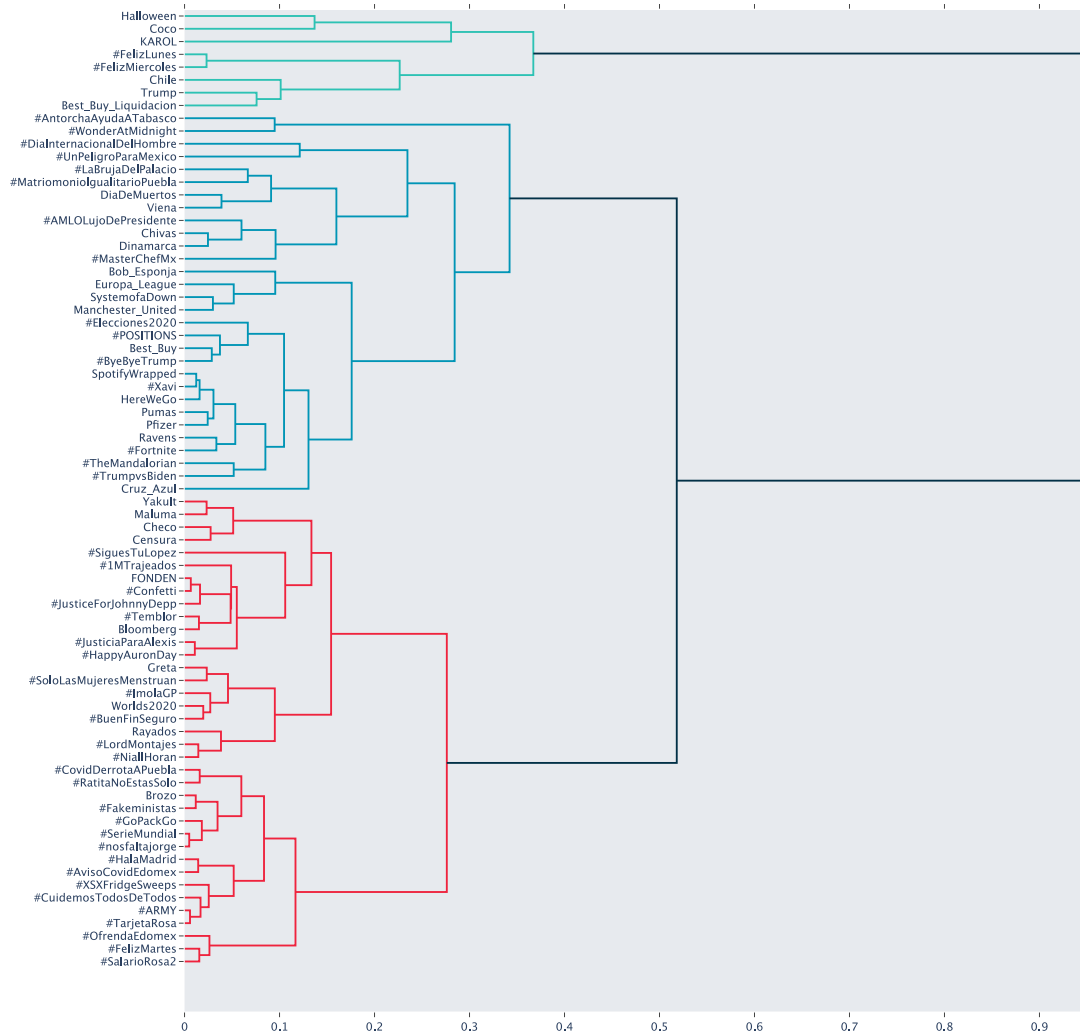


Fig. 5.6. Agrupamiento jerárquico utilizando *complete linkage*

5.4 Visualización de resultados

Para explorar visualmente los resultados, se desarrolló un sitio web utilizando las tecnologías de Docusaurus [Met22] y React [Met13]. El sitio web es estático y esta disponible en *Github Pages* (ver Figs. 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

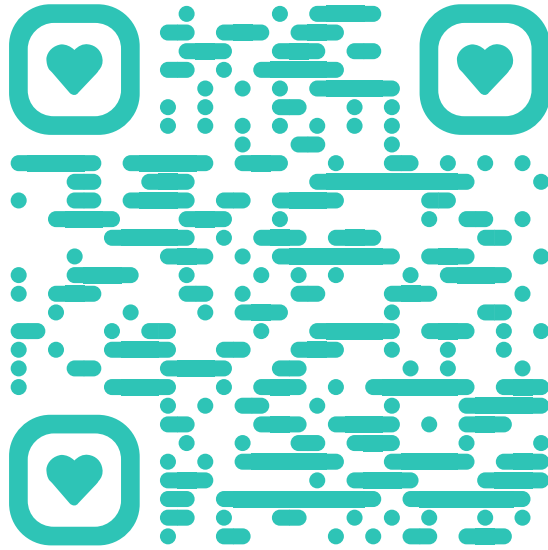


Fig. 5.7. Código QR para acceder a la [herramienta web](#).

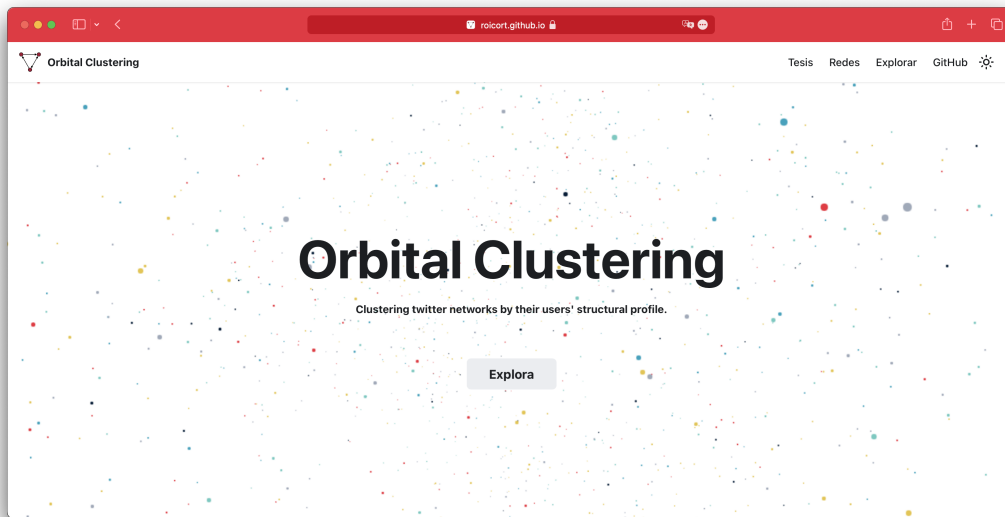


Fig. 5.8. Página principal de la sitio web.

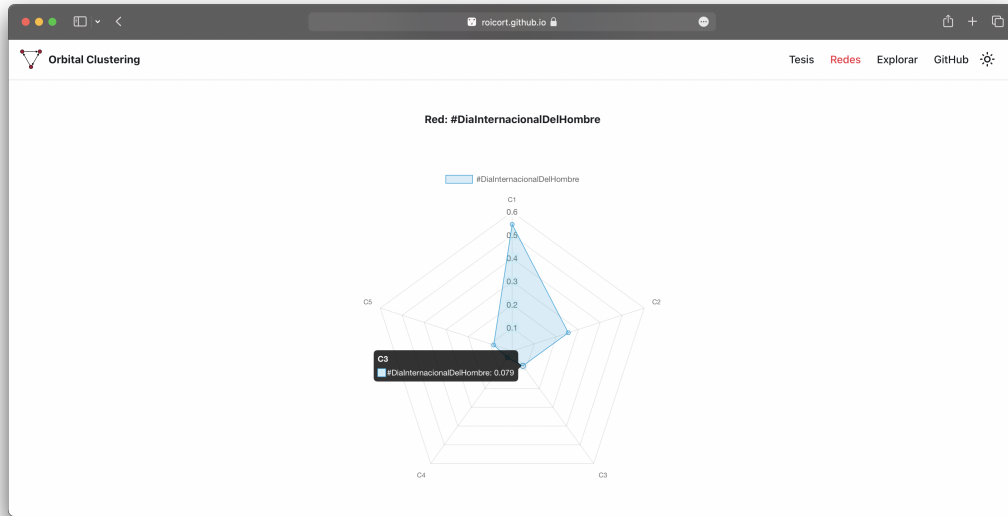


Fig. 5.9. Ejemplo de la visualización de la composición de una red utilizando una gráfica de radar.

radar la composición por tipo de usuarios de la red seleccionada. La Fig. 5.9 ejemplifica la composición de una de las redes temáticas en la colección.

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

Los grafos que se muestran a continuación (Figs. 5.11 y 5.12) fueron escogidos como ejemplos por ser los más lejanos de acuerdo al agrupamiento jerárquico. En la Fig. 5.11 observamos la red correspondiente al #SalarioRosa, en la 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 Fig. 5.12 muestra una red más bien fragmentada en la que no existe una conversación central. La Tabla 5.2 presenta el contraste entre los *embeddings* generados para ambas redes.

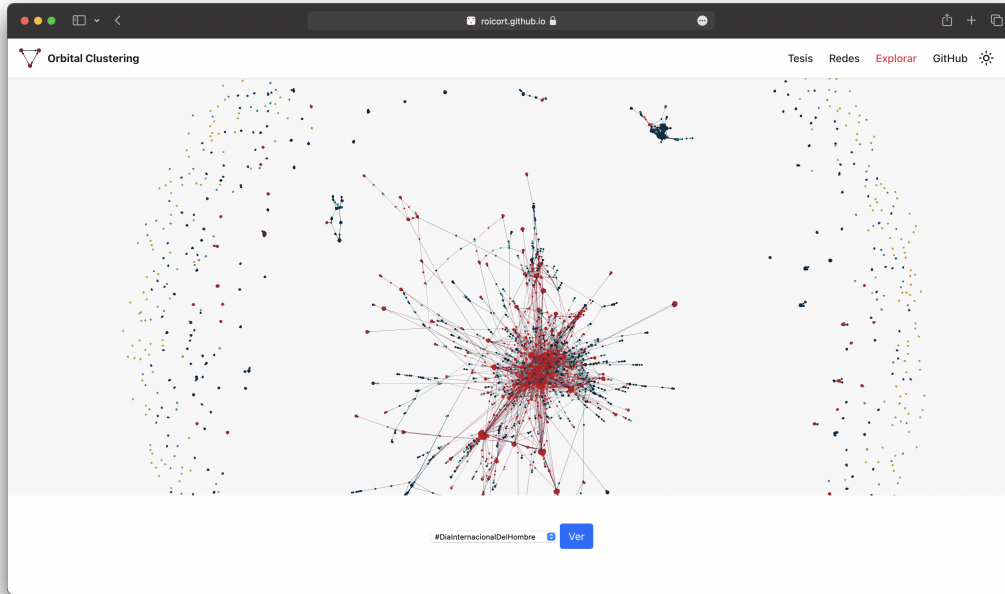


Fig. 5.10. Ejemplo de la visualización de una red (#Coco) dentro de la herramienta web.

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 #SalarioRosa

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 centroe representativo. En cambio, el grupo restante (3) tiene una distribución de órbitas más equilibrada en el vector de firmas de su centroe.

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 #SalarioRosa coloreada respecto al grupo al que pertenece cada nodo en la red.

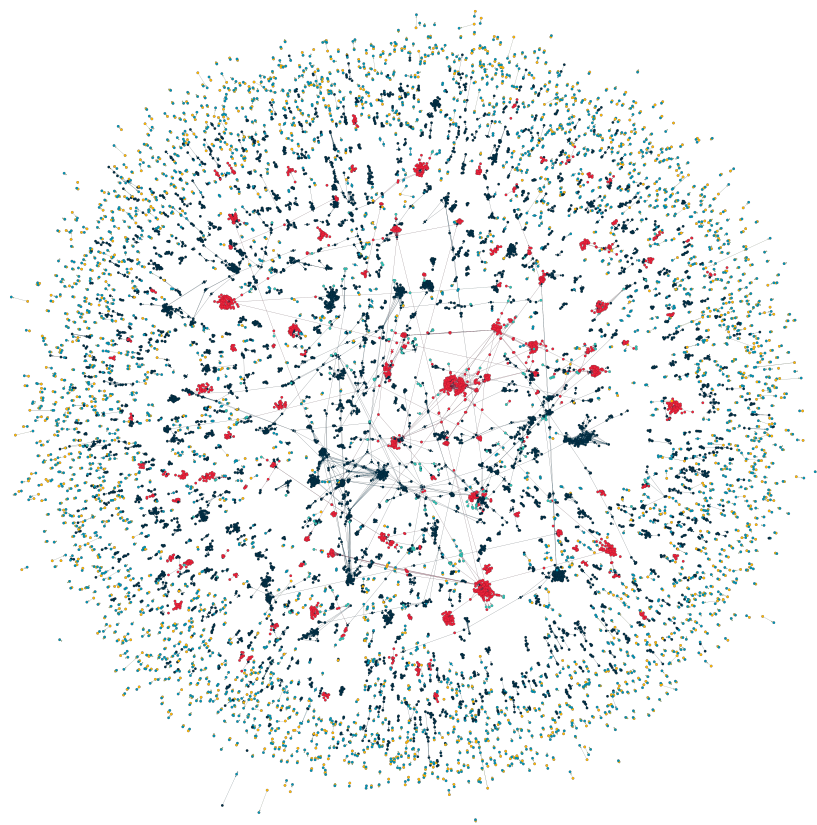


Fig. 5.12. Red #Coco coloreada respecto al grupo al que pertenece cada nodo en la red.

- *Perfil 1, 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.

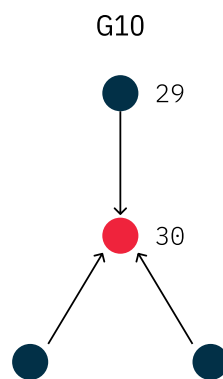


Fig. 5.13. Graphlet 10 y órbitas 29 y 30.

- *Perfil 2, 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.
- *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



Fig. 5.14. Graphlet 8 y órbitas 21 a 24.

de firmas asociados, lo que sugiere que este perfil permite el flujo de información hacia y desde los otros perfiles predominantes.

- *Perfil 4, 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.

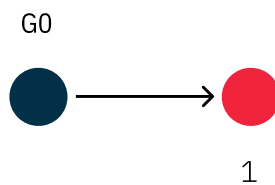


Fig. 5.15. Graphlet 0 y órbita 1.

- *Perfil 5, 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.

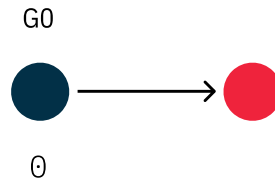


Fig. 5.16. Graphlet 0 y órbita 0.

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 Difusor, Conversador o Inconformista.

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 (difusores) de las que se hacen eco otros perfiles (reporteros), encontramos algunas iniciativas gubernamentales (#SalarioRosa, #OfrendaE-doMex, #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.

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

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.

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 $y \in Y$ existe un único $x \in X$ al que la función evaluada en x es igual a y .

Línea base

Utilizando el árbol de decisión de *Himmelboim 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 que resulta no es capaz de capturar la complejidad de la colección, y la mayoría de las redes quedan en un solo grupo (*clustered*).

Network	Label
LordMontajes	Clustered
MasterChefMx	Clustered
NiallHoran	Clustered
POSITIONS	Clustered
TheMandalorian	Clustered
WonderAtMidnight	OutWard Hub and Spoke
XSXFridgeSweeps	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
Temblor	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).
- [Ahm+22] Nesreen K. Ahmed, Ryan A. Rossi, John Boaz Lee y col. «Role-Based Graph Embeddings». en. En: *IEEE Transactions on Knowledge and Data Engineering* 34.5 (mayo de 2022), págs. 2401-2415. ISSN: 1041-4347, 1558-2191, 2326-3865. URL: <https://ieeexplore.ieee.org/document/9132694/> (visitado 25-01-2023) (vid. pág. 27).
- [AV] David Arthur y Sergei Vassilvitskii. «k-means++: The Advantages of Careful Seeding». en. En: (), pág. 11 (vid. pág. 50).
- [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).

- [CLX15] Shaosheng Cao, Wei Lu y Qionгкаi Xu. «GraRep: Learning Graph Representations with Global Structural Information». en. En: *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. Melbourne Australia: ACM, oct. de 2015, págs. 891-900. ISBN: 978-1-4503-3794-6. URL: <https://dl.acm.org/doi/10.1145/2806416.2806512> (visitado 25-01-2023) (vid. pág. 26).
- [DBG02] «Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering». en. En: *Advances in Neural Information Processing Systems 14*. Ed. por Thomas G. Dietterich, Suzanna Becker y Zoubin Ghahramani. The MIT Press, 2002. ISBN: 978-0-262-27173-8. URL: <https://direct.mit.edu/books/book/2485/chapter/66411/laplacian-eigenmaps-and-spectral-techniques-for> (visitado 25-01-2023) (vid. pág. 26).
- [Don+18] Claire Donnat, Marinka Zitnik, David Hallac y Jure Leskovec. «Learning Structural Node Embeddings via Diffusion Wavelets». en. En: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. London United Kingdom: ACM, jul. de 2018, págs. 1320-1329. ISBN: 978-1-4503-5552-0. URL: <https://dl.acm.org/doi/10.1145/3219819.3220025> (visitado 25-01-2023) (vid. pág. 27).
- [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).
- [GL] Alexis Galland y Marc Lelarge. «Invariant embedding for graph classification». en. En: () (vid. pág. 27).
- [GWH] Feng Gao, Guy Wolf y Matthew Hirn. «Geometric Scattering for Graph Data Analysis». en. En: (), pág. 10 (vid. pág. 27).

- [GL16] Aditya Grover y Jure Leskovec. «node2vec: Scalable Feature Learning for Networks». en. En: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco California USA: ACM, ago. de 2016, págs. 855-864. ISBN: 978-1-4503-4232-2. URL: <https://dl.acm.org/doi/10.1145/2939672.2939754> (visitado 20-08-2021) (vid. pág. 26).
- [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 (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, 24, 69, 71, 72).
- [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. 15).
- [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. xxiii, 17, 18, 21).

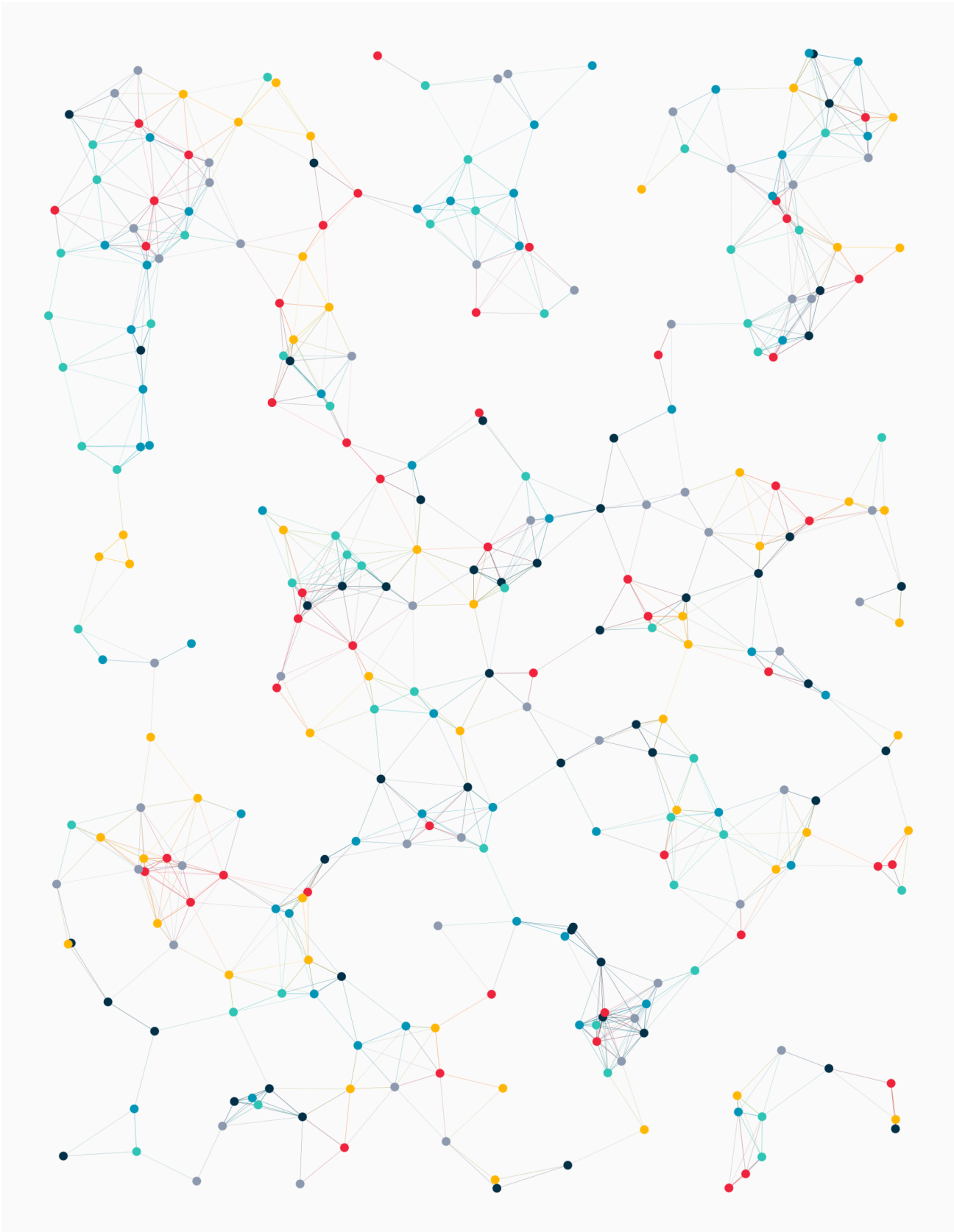
- [Kwa+10] Haewoon Kwak, Changhyun Lee, Hosung Park y Sue Moon. «What is Twitter, a social network or a news media?» en. En: *Proceedings of the 19th international conference on World wide web - WWW '10*. Raleigh, North Carolina, USA: ACM Press, 2010, pág. 591. ISBN: 978-1-60558-799-8. URL: <http://portal.acm.org/citation.cfm?doid=1772690.1772751> (visitado 02-10-2020) (vid. pág. 5).
- [LP18] Nathan de Lara y Edouard Pineau. «A Simple Baseline Algorithm for Graph Classification». en. En: *arXiv:1810.09155 [cs, stat]* (nov. de 2018). arXiv: 1810.09155. URL: <http://arxiv.org/abs/1810.09155> (visitado 31-07-2020) (vid. pág. 27).
- [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. 25, 28).
- [Li+19] Jundong Li, Liang Wu, Ruocheng Guo, Chenghao Liu y Huan Liu. «Multi-level network embedding with boosted low-rank matrix approximation». en. En: *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. Vancouver British Columbia Canada: ACM, ago. de 2019, págs. 49-56. ISBN: 978-1-4503-6868-1. URL: <https://dl.acm.org/doi/10.1145/3341161.3342864> (visitado 25-01-2023) (vid. pág. 26).
- [Lus] Dean() Lusher. «Exponential Random Graph Models for Social Networks». en. En: (), pág. 361 (vid. págs. 30, 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).
- [Met22] Meta. *DocuSaurus*. Ago. de 2022. URL: <https://docusaurus.io> (vid. pág. 57).
- [Met13] Meta. *React*. Mayo de 2013. URL: <https://react.dev> (vid. pág. 57).

- [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. 29, 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).
- [Nar+17] Annamalai Narayanan, Mahinthan Chandramohan, Rajasekar Venkatesan y col. «graph2vec: Learning Distributed Representations of Graphs». en. En: *arXiv:1707.05005 [cs]* (jul. de 2017). arXiv: 1707.05005. URL: <http://arxiv.org/abs/1707.05005> (visitado 31-07-2020) (vid. pág. 27).
- [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. 22).
- [Per+17] Bryan Perozzi, Vivek Kulkarni, Haochen Chen y Steven Skiena. «Don't Walk, Skip!: Online Learning of Multi-scale Network Embeddings». en. En: *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*. Sydney Australia: ACM, jul. de 2017, págs. 258-265. ISBN: 978-1-4503-4993-2. URL: <https://dl.acm.org/doi/10.1145/3110025.3110086> (visitado 25-01-2023) (vid. pág. 26).
- [PAS14] Bryan Perozzi, Rami Al-Rfou y Steven Skiena. «DeepWalk: online learning of social representations». en. En: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York New York USA: ACM, ago. de 2014, págs. 701-710. ISBN: 978-1-4503-2956-9. URL: <https://dl.acm.org/doi/10.1145/2623330.2623732> (visitado 25-01-2023) (vid. pág. 26).
- [Prz07] N. Przulj. «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).

- [Qiu+18] Jiezhong Qiu, Yuxiao Dong, Hao Ma y col. «Network Embedding as Matrix Factorization: Unifying DeepWalk, LINE, PTE, and node2vec». en. En: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. Marina Del Rey CA USA: ACM, feb. de 2018, págs. 459-467. ISBN: 978-1-4503-5581-0. URL: <https://dl.acm.org/doi/10.1145/3159652.3159706> (visitado 25-01-2023) (vid. pág. 26).
- [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. 15-17).
- [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).
- [RS18] Benedek Rozemberczki y Rik Sarkar. «Fast Sequence-Based Embedding with Diffusion Graphs». en. En: *Complex Networks IX*. Ed. por Sean Cornelius, Kate Coronges, Bruno Gonçalves, Roberta Sinatra y Alessandro Vespignani. Series Title: Springer Proceedings in Complexity. Cham: Springer International Publishing, 2018, págs. 99-107. ISBN: 978-3-319-73197-1 978-3-319-73198-8. URL: http://link.springer.com/10.1007/978-3-319-73198-8_9 (visitado 03-02-2023) (vid. pág. 26).
- [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 Nebil 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, 29-31, 35, 50).
- [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. xxiii, 38-40).
- [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).
- [SF14] Dennis L. Sun y Cedric Fevotte. «Alternating direction method of multipliers for non-negative matrix factorization with the beta-divergence». en. En: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Florence, Italy: IEEE, mayo de 2014, págs. 6201-6205. ISBN: 978-1-4799-2893-4. URL: <http://ieeexplore.ieee.org/document/6854796/> (visitado 25-01-2023) (vid. pág. 26).
- [TL09] Lei Tang y Huan Liu. «Relational learning via latent social dimensions». en. En: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. Paris France: ACM, jun. de 2009, págs. 817-826. ISBN: 978-1-60558-495-9. URL: <https://dl.acm.org/doi/10.1145/1557019.1557109> (visitado 25-01-2023) (vid. pág. 26).
- [TCE20] Leo Torres, Kevin S. Chan y Tina Eliassi-Rad. «GLEE: Geometric Laplacian Eigenmap Embedding». en. En: *Journal of Complex Networks* 8.2 (abr. de 2020). arXiv:1905.09763 [cs], cnaa007. ISSN: 2051-1310, 2051-1329. URL: <http://arxiv.org/abs/1905.09763> (visitado 25-01-2023) (vid. pág. 26).
- [Tsi+18] Anton Tsitsulin, Davide Mottin, Panagiotis Karras, Alex Bronstein y Emmanuel Müller. «NetLSD: Hearing the Shape of a Graph». en. En: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (jul. de 2018). arXiv: 1805.10712, págs. 2347-2356. URL: <http://arxiv.org/abs/1805.10712> (visitado 31-07-2020) (vid. pág. 27).
- [Twi] Twitter. *Twitter.com*. en. URL: twitter.com/about (visitado 25-05-2021) (vid. págs. 4, 5).

- [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).
- [Yan+19] Dingqi Yang, Paolo Rosso, Bin Li y Philippe Cudre-Mauroux. «NodeSketch: Highly-Efficient Graph Embeddings via Recursive Sketching». en. En: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Anchorage AK USA: ACM, jul. de 2019, págs. 1162-1172. ISBN: 978-1-4503-6201-6. URL: <https://dl.acm.org/doi/10.1145/3292500.3330951> (visitado 25-01-2023) (vid. pág. 26).
- [Yue+20] Xiang Yue, Zhen Wang, Jingong Huang y col. «Graph embedding on biomedical networks: methods, applications and evaluations». en. En: *Bioinformatics* 36.4 (feb. de 2020). Ed. por Lenore Cowen, págs. 1241-1251. ISSN: 1367-4803, 1367-4811. URL: <https://academic.oup.com/bioinformatics/article/36/4/1241/5581350> (visitado 03-02-2023) (vid. pág. 25).
- [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 Tiño, Aleš Leonardis y Ke Tang. «A Survey on Neural Network Interpretability». en. En: *IEEE Transactions on Emerging Topics in Computational Intelligence* 5.5 (oct. de 2021). arXiv:2012.14261 [cs], págs. 726-742. ISSN: 2471-285X. URL: <http://arxiv.org/abs/2012.14261> (visitado 25-01-2023) (vid. pág. 17).
- [Zha+18] Ziwei Zhang, Peng Cui, Haoyang Li, Xiao Wang y Wenwu Zhu. «Billion-Scale Network Embedding with Iterative Random Projection». en. En: *2018 IEEE International Conference on Data Mining (ICDM)*. Singapore: IEEE, nov. de 2018, págs. 787-796. ISBN: 978-1-5386-9159-5. URL: <https://ieeexplore.ieee.org/document/8594903/> (visitado 25-01-2023) (vid. pág. 26).



ESCUELA
NACIONAL
DE ESTUDIOS
SUPERIORES

UNIDAD MORELIA



TECNOLOGÍAS PARA
LA INFORMACIÓN
EN CIENCIAS