CENTRO DE INVESTIGACION Y DE ESTUDIOS
AVANZADOS DEL INSTITUTO POLITÉCNICO NACIONAL

UNIDAD ZACATENCO

DEPARTAMENTO DE

COMPUTACIÓN

# Uso de Atributos para Detectar Comunidades de Calidad en Redes Sociales

## T E S I S

Que presenta

**BELLA CITLALI MARTÍNEZ SEIS**

Para obtener el grado de

**DOCTORA EN CIENCIAS EN COMPUTACIÓN**

Directora de la Tesis:

**Dra. Xiaoou Li**

Ciudad de México Marzo, 2018

CENTRO DE INVESTIGACION Y DE ESTUDIOS
AVANZADOS DEL INSTITUTO POLITÉCNICO NACIONAL

UNIDAD ZACATENCO

COMPUTER SCIENCE DEPARTMENT

# Use of Attributes to Detect Communities of Quality in Social Networks

## T E S I S

Dissertation submitted by

**BELLA CITLALI MARTÍNEZ SEIS**

for the degree of

**Doctor of Philosophy in Computer Science**

Supervisor:

**Dra. Xiaoou Li**

Mexico City March, 2018

A mi esposo, que me acompañó de principio a fin en esta travesía.

A mi hermana, por las sonrisas en los momentos más difíciles.

A mis padres, por la educación y los valores que me inculcaron.

A mis familiares y amigos, que estuvieron para escucharme y apoyarme.

# Agradecimientos

Inicialmente, agradezco a mi asesora de Tesis, la Dra. Xiaoou, por compartir sus experiencias y conocimientos, además del apoyo y paciencia que me brindó durante el trabajo de tesis realizado.

Agradezco a mis sinodales por las observaciones realizadas y por aquellas preguntas que abren más puertas.

Agradezco a Sofy por su carisma y apoyo en los diversos procesos durante mi estancia y al departamento de computación por brindarme la oportunidad de concluir este ciclo.

Finalmente, agradezco al Conacyt por el apoyo inicial brindado, así como al IPN.

# Abstract

Community is a group of vertices densely coneected. They are often studied structurally, nevertheless some real-world networks contain attributes. Those are important because a node is in the same community as its neighbors, but it should also share the community with similar nodes. A member of a social network belongs to multiple communities, so there methods to detect overlapping communities.

Just recent works merge attributes and graph structure. There are two main methods: the model based and the distance based. The first ones are usually probabilistic and they find communities similar to the ground-truth communities but they require previous knowledge of the network and do not consider basic network properties. The second ones use measures based on one or more properties but there is not an approximation to the ground-truth communities. We propose a **mixed method** based on model (RMOCA) and distance (BAS) to get advantages of both process.

**R**egression **MO**del for **C**ommunities in **A**ttributed networks (**RMOCA**) models that two vertices with a connection tend to be in the same community, and a node includes its attributes to the community in order to integrate nodes with same attributes. We used regression models with ordinary least squares to define a function that is minimized. Experimental results demonstrate mostly the entropy (attribute similarity) and purity (precision to recover ground-truth communities) improvement.

We propouse $Q_A$, a **community quality measure based on attributes**. It stores structural information, local and global importance of attributes, node degree, and attribute degree. We integrate it with conductance to **B**alance **A**ttributes and **S**tructure (**BAS**). The measure is used by the **mixed method** to improve communities generating overlapping communities. The global importance of attributes is also used to **rank attributes** and to select the best ones to reduce the complexity. Experimental results shows better communities with the expansion of BAS and a time reduction with the pre-process without affecting the quality of the communities.

# Resumen

Las comunidades son grupos de nodos densamente conectados y son estudiadas de manera estructural, sin embargo las redes sociales reales contienen atributos. Si se consideran ambos tipos de datos, un nodo pertenece a la misma comunidad que sus vecinos y éstos son similares. Además los nodos pertenecen a múltiples grupos por lo que existen algoritmos enfocados a las comunidades sobrepuestas.

Trabajos recientes mezclan la estructura y los atributos. Existen dos tipos de métodos: basados en modelo y basados en distancia. Los primeros, usan elementos probabilísticos y detectan comunidades como las reales pero requieren de conocimiento previo de la red y descuidan propiedades básicas. Los segundos, generalmente optimizan una medida basada en una o más propiedades, pero su aproximación a las comunidades reales no es buena. En esta tesis se propone un **método mixto** basado en modelo (RMOCA) y en distancia (BAS) para obtener las ventajas de ambos.

El MOdelo de Regresiones para Comunidades en redes con Atributos (**RMOCA**) considera que dos vértices con una conexión tienden a pertenecer a la misma comunidad y que un nodo incluye a sus atributos a la comunidad para integrar posteriormente nodos con los mismos atributos. Se usa un modelo de regresiones con mínimos cuadrados para definir una función que es minimizada. Los resultados demuestran el incremento del RMOCA en similitud entre nodos y precisión a comunidad reales.

Se propone $Q_A$ como una **medida de calidad de comunidades basada en atributos**. Ésta considera información estructural, importancia local y global de atributos, grado del nodo y densidad de atributo. $Q_A$ es integrada a la conductividad y es balanceada (**BAS**). Éstas medidas son usadas por el **método mixto** para mejorar las comunidades y generar comunidades sobrepuestas. La importancia global de los atributos es usada para seleccionar los mejores atributos y reducir la complejidad sin afectar la calidad de las comunidades. Los resultados muestran un incremento de la entropía y calidad en traslape de las comunidades detectadas con la expansión.

# Publicaciones

- Martinez-Seis, B., & Li, X. (2014, October). Topic hierarchy in social networks. In Systems, Man and Cybernetics (SMC), 2014 IEEE International Conference on (pp. 2013-2018). IEEE.

- Martinez-Seis, B., & Li, X. (2016, April). Ranking features in Facebook to detect overlapping communities. In Networking, Sensing, and Control (ICNSC), 2016 IEEE 13th International Conference on (pp. 1-6). IEEE.

- Martínez-Seis, B. (2017, April). RELNA: Ranking Attributes in Social Networks to Detect Overlapping Communities Efficiently. In Data Engineering (ICDE), 2017 IEEE 33rd International Conference on (pp. 1431-1435). IEEE.

- Martinez-Seis, B., & Li, X. A new community quality measure using attributes to improve structure-based communities. Physica A: Statistical Mechanics and its Applications, Elsevier (en proceso)

- Martinez-Seis, B., & Li, X. Use of attributes to detect communities in on-line social networks. Expert Systems with Applications, Elsevier. Special Issue: Big Data Analytics for Business Intelligence. (por someter)

# Índice

# Símbolos y Notaciones

| | | |
|---|---|---|
| $V$ | Conjunto de vértices o nodos en un sociograma | 14 |
| $n$ | Número de vértices en un grafo tal que $n = \|V\|$ | 14 |
| $E$ | Conjunto de aristas que son conformadas por un par de vértices $(v_i, v_j)$. | 14 |
| $m$ | Número de aristas en un grafo tal que $m = \|E\|$ | 14 |
| $G(V, E)$ | Sociograma o grafo $G$ de una red con vértices $V$ y aristas $E$ | 14 |
| $A$ | Conjunto de atributos en un grafo provenientes de características de los nodos. | 18 |
| $k$ | Número de atributos en un grafo tal que $k = \|A\|$ | 18 |
| $\Lambda$ | Función de asociación de vértices y atributos | 18 |
| $G'(V, E, A)$ | Grafo $G$ con un conjunto de atributos $A$ relacionados con los vértices $V$ a través de la función de asociación $\Lambda$ | 18 |
| $V_A$ | Vértice de atributos: conjunto de atributos representados como vértices | 19 |
| $E_A$ | Conjunto de aristas de pares vértices y vértice-atributo | 19 |
| $V'$ | Conjunto de vértices de entidades $V$ y vértices de atributos $V_A$ | 19 |

| | | |
|---|---|---|
| $E'$ | Conjunto de aristas del grafo $E$ y aristas vértice-atributo $E_A$ | 19 |
| $G''(V', E')$ | Grafo $G$ con vértices aumentados $V'$ y aristas aumentadas $E'$ | 19 |
| $k_v$ | Grado de un nodo $v$ | 21 |
| $D(G)$ | Densidad de un grafo $G$ | 24 |
| | | |
| $C$ | Conjunto de comunidades disjuntas o sobrepuestas | 40, 50 |
| $q$ | Número de comunidades en un grafo $G$ tal que $q = \|C\|$ | 40, 50 |
| $Q$ | Modularidad | 43 |
| $D(c)$ | Densidad de una comunidad $c$ | 47 |
| $D(C)$ | Densidad del conjunto de comunidades $C$ | 48 |
| $\phi(c)$ | Conductividad de una comunidad $c$ | 49 |
| $C^*$ | Conjunto de comunidades reales (*ground-truth communities*) | 50 |
| $purity(C, C^*)$ | Pureza entre comunidades detectadas $C$ y las reales $C^*$ | 50 |
| $F_1$ | Medida $F_1$ para comparar las comunidades detectadas $C$ con las reales $C^*$ ($F_1$-*measure*) | 50 |
| $M_{jaccard}$ | Medida Jaccard para evaluar comunidades | 50 |
| $F$ | Matriz de pertenencia de nodos a comunidades usadas por Bigclam y CESNA | 55 |
| $\Omega(C, C^*)$ | Índice Omega para evaluar comunidades sobrepuestas | 56 |
| $OC$ | Medida $OC$ para comunidades sobrepuestas | 57 |
| $W$ | Matriz de pertenecía de nodos a aristas usada por CESNA | 60 |

| | | |
|---|---|---|
| $H(G)$ | Entropía del grafo $G$ | 64 |
| $M$ | Matriz de adjacencia que contiene las relaciones entre pares de nodos | 71 |
| $X$ | Matriz de atributos que contiene la pertenencia de atributos a un nodos | 72 |
| $C_S$ | Conjunto de comunidades de nodos basada en la estructura | 109 |
| $C_A$ | Conjunto de comunidades de atributos | 75 |
| $M_S$ | Matriz de pertenencia de nodos a comunidades basada en la estructura | 109 |
| $M_A$ | Matriz de pertenencia de atributos a comunidades | 75 |
| $W_a(G)$ | Medida de importancia global de un atributo $a$ | 104 |
| $W_a(c)$ | Medida de importancia local: importancia de un atributo $a$ en una comunidad $c$ | 108 |
| $H_a(c)$ | Densidad de una atributo $a$ en una comunidad $c$ | 109 |
| $Q_A(c)$ | Medida de calidad de comunidad $c$ basada en atributos | 110 |
| $Q_S(c)$ | Medida de calidad de comunidad $c$ basada en estructura | 110 |
| $\alpha$ | Parámetro tal que esta en un rango de $[0, 1]$ que regula el balance entre estructura y atributos | 111 |
| $Q(c)$ | Medida para el balance de la calidad basada en atributos $Q_A(c)$ y la calidad basada en estructura $Q_S(c)$ | 111 |

# Capítulo 1

# Introducción

En el devenir histórico de los seres vivos han existido diversas formas de interacción entre animales, personas o incluso neuronas que han permitido el modelado de redes sociales. Hemos asociado dicho término a las redes sociales en línea como Facebook, sin embargo, se tienen redes sociales desde el comienzo de la vida, cuyo análisis tuvo auge en 1930 principalmente por sociólogos [95]. En la actualidad, Internet ha permitido la comunicación entre elementos como empresas e individuos, lo que ha llevado a crear redes sociales en línea con objetivos productivos, comerciales, económicos, de conocimiento, académicos, periodísticos, deportivos, entre otros. En este sentido, las redes sociales en línea como Dogster, Facebook o Sermo constituyen una plataforma de comunicación entre elementos virtuales que representan a perros, personas y personal médico, respectivamente.

Las redes sociales son una abstracción que permite el estudio de las relaciones entre entidades, ya sea a nivel psicológico, antropológico o matemático. Se ha establecido un modelado con grafos llamado *sociograma* para el análisis de estos sistemas complejos. Esta representación permite el estudio de las relaciones entre individuos donde se atienden las características específicas de los sistemas sociales.

Por naturaleza, los seres conviven, se interrelacionan y comparten experiencias con grupos de intereses afines, por lo que las redes sociales se dividen naturalmente en pequeñas comunidades. Una comunidad es un conjunto de nodos que están densamente conectados entre ellos y poco conectados con el resto de la red, la detección de comunidades se ha complicado por la estructura de las redes sociales en línea [84].

Las entidades en la redes sociales se caracterizan por pertenecer a múltiples comunidades, por ejemplo una persona interactúa en su círculo familiar, de amistades y el laboral o escolar. Esto también sucede en otras redes complejas [117], por lo que se ha incrementado el estudio para la detección de comunidades sobrepuestas.

Aunado al sociograma, se cuenta con información adicional dada por características de los individuos, lo cual ha aportado ventajas significativas en estudios dentro del área de Análisis de Redes Sociales (*SNA, Social Network Analysis*) [101]. La mayoría de los algoritmos de detección de comunidades se basan en la estructura dada por el grafo, pero es necesaria la integración de información adicional como atributos.

El presente trabajo propone una evaluación de atributos en redes sociales, así como la integración de éstos para el desarrollo de un método mixto que realice la detección de comunidades sobrepuestas mejorando la calidad de éstas.

## 1.1 Redes sociales

Entre 1930 y 1970 aumentó el número de antropólogos y sociólogos estudiando la estructura social debido a la migración de alemanes a Estados Unidos, uno de ellos fue Jacob Moreno que introdujo el término de *sociograma* como una manera formal de representar las propiedades de las configuraciones sociales. El sociograma (o grafo) relaciona entidades (nodos) a través de relaciones (aristas).

En los años 50 el uso de la teoría de grafos involucró a matemáticos que ayudaron

al desarrollo formal de modelos de cohesión de grupos, presión social, cooperación, poder y liderazgo. En los años 70 se tenían dos corrientes matemáticas: una referente a los modelos algebraicos de grupos, usando la teoría de conjuntos y la segunda en trasladar las relaciones en distancias sociales; ambas fueron unificadas con el análisis de redes sociales. El análisis de redes sociales provee el vocabulario y un conjunto de medidas para una análisis relacional, pero no implica la aceptación de una teoría en particular relacionada a la estructura social.

Dado el crecimiento del análisis de redes sociales, causado por Internet a principios de este siglo, se involucraron físicos en esta área que retomaron modelos basados en grafos aleatorios [31], que fueron confrontados con las propiedades de mundo pequeño como la de seis grados de separación [112] y la distribucción *power law* [111]. Los cuales establecen que se requieren pocos elementos intermedios para conectar a dos nodos y pocos nodos tienen un mayor número de enlaces, mientras que la mayor parte de las entidades tienen pocos enlaces.

Derivado de la oposición al Marxismo[1], durante 1930 y 1940 en la Universidad de Harvard, se tuvieron estudios empíricos que buscaban obtener la estructura de subgrupos de sistemas sociales con datos relacionales, definiendo las comunidades como un conjunto de relaciones a través de las cuales las entidades interaccionan unas con otras. Una persona puede ser miembro de diferentes comunidades dando origen al concepto de comunidades sobrepuestas.

## 1.2 Detección de comunidades

La detección de comunidades o clasificación en grafos tiene como idea principal separar en clases o categorías [87] con base en distancia o similitud, definiendo las

[1]Modelo teórico sobre la lucha de clases, la crítica a la economía capitalista, la dominación mental bajo el concepto de ideología, y el comunismo

Tabla 1.1: Algoritmos para la detección de comunidades en redes usando atributos

| Algoritmos | Tipo de Método | Tipo de Comunidades |
|---|---|---|
| SA-Cluster[132], Codicil[91], CME[101], SANS[83] | Basado en Distancia | Disjuntas |
| PICS [2], SAC[23], NAS[100], TCPI[48] | Basado en Modelo | Disjuntas |
| EDCAR[46], CoPaMs[71], DB-CSC[45], JCDC[129] | Basado en Distancia | Sobrepuestas Baja |
| GenClus [102], GBAGC [119] | Basado en Modelo | Sobrepuestas Baja |
| CESNA[122] | Basado en Modelo | Sobrepuestas Alta |

comunidades o *clusters* como un conjunto de elementos que son similares con respeto a alguna propiedad relacional. Existen dos familias principales en el análisis de *clusters*: aglomerativos y divisivos, donde ambos establecen una jerarquía de similitud. Uno de los algoritmos divisivos más importantes es el de Girvan-Newman [40] con la medida de modularidad; otros relevantes son METIS [55] y *clustering* espectral [79]. Estos se han enfocado en comunidades disjuntas de tal forma que un nodo sólo pertenece a una comunidad usando la estructura de la red.

Como se ha mencionado, las personas en una red pertenecen de manera natural a varias comunidades, con mayor intensidad sucede en redes en línea convirtiéndose en una característica importante de las redes sociales reales [56][86] por lo que ha incrementado el estudio de **comunidades sobrepuestas**. Xie et. al. [117] presentan una clasificación para este tipo de algoritmos, además distinguen entre baja [62] [42] y alta densidad [116] [17] de sobreposición en las comunidades.

Las **comunidades usando estructura y atributos** surgen debido a que se han generado una gran cantidad de información de objetos reales, por lo que además de las conexiones en el sociograma, se tiene una lista de atributos asociada a los nodos que describen las características y propiedades de los objetos, dando origen a las redes con atributos. Por ejemplo, se ha usado la información geográfica de Foursquare para encontrar comunidades que agrupan tanto lugares cercanos como características similares [110].

Según un estudio realizado por Ding [26], las comunidades detectadas usando la estructura difieren significativamente de las comunidades obtenidas usando sólo los

atributos de la misma red. La obtención de grupos en este tipo de grafos representa una mayor complejidad. Se distinguen dos tipos de métodos para detectar comunidades usando atributos: basadas en distancia o basadas en modelos, como se muestra en la Tabla 1.1, los cuales serán detallados en el Capítulo 3.

Los **métodos basados en distancia** requieren definir una distancia o similitud que integre ambos elementos, donde las comunidades se adecuan a la definición de la medida y no al modelado real de las comunidades, lo que es subsanado por los **métodos basados en modelo** que, aunque requieren un conocimiento previo de la red, suelen descuidar propiedades esperadas de las comunidades en redes sociales, como la densidad. El algoritmo de CESNA [122] se autoproclamó como el primero en detectar comunidades densas sobrepuestas en redes sociales con atributos; éste se basa en la probabilidad de conexión entre nodos según las comunidades y atributos que comparten generando comunidades parecidas a las reales, pero descuidando la densidad interna.

La definición de una o varias medidas permite obtener comunidades acordes a propiedades de las redes sociales, mientras que elaborar un modelo obtiene comunidades que describen las comunidades reales o esperadas. Por ello surge la pregunta: ¿Cómo obtener comunidades basadas en medida y modelo considerando estructura y atributos?

## 1.3 Motivación

Internet ha permitido la interacción y creación de estructuras sociales en línea que están jugando un papel importante para la generación de recomendaciones e influencias. La popularidad de las redes sociales ha generado altas expectativas tanto para los negocios y como para la comunidad científica. Específicamente, la detección de

comunidades sobrepuestas en redes con atributos es uno de los problemas principales y, comparado con el gran estado del arte de comunidades tradicionales, aún son pocos los algoritmos propuestos para resolverlo.

Los principales tipos de datos en el análisis de redes sociales son los atributos (*attribute data*) y la estructura (*relational data*) [95]. Se ha demostrado que la integración de ambos mejora la calidad de las comunidades. En este ámbito han surgido dos formas de resolver esta detección: basado en distancia y basado en modelo. Para la definición de la distancia se hacen modificaciones de pesos en las arista o se transforman los sociogramas a grafos aumentados o a varias dimensiones.

El principal problema es integrar en la medida la estructura y los atributos con el balance adecuado, sin embargo, estos métodos logran comunidades densas y parecidas (alta entropía). Por otro lado, los algoritmos basados en modelos suelen ser probabilísticos y requieren de un conocimiento previo de la red para su diseño, lo cual permite encontrar comunidades naturales o con mayor similitud a las reales, aunque descuidan aspectos como la densidad. Por ello, se ve la oportunidad de obtener las ventajas de los métodos basado en modelo y en distancia generando comunidades con:

- **Cohesión estructural.** Los vértices dentro de una comunidad están altamente conectados.
- **Homogeneidad de atributos.** Los vértices dentro de una comunidad tienen valores similares en sus atributos, mientras que vértices en diferentes comunidades tienen diferentes valores.
- **Naturalidad.** Las comunidades detectadas son similares a los grupos reales.

De tal forma que el objetivo de esta investigación es: *mejorar la calidad de las comunidades detectadas en redes sociales, haciendo uso de atributos a través de la*

*propuesta de un nuevo modelo y medidas que busquen el balance entre el sociograma y los atributos relevantes.*

Se debe tener una medida que considere tanto la estructura de la red como los atributos de los nodos. En las redes sociales existen atributos que son más importantes que otros, pero cómo determinar dicha importancia y cómo vincular a los atributos con la red. Estos y otros aspectos deben ser considerados en la medida propuesta. Por otro lado, la definición del modelo debería considerar tanto la estructura como los atributos; hemos partido de la premisa de que los atributos pueden formar parte de las comunidades, de tal forma que los nodos compartirían comunidad con los atributos cuando los nodos tengan los atributos de esa comunidad.

Aunado a esto, los usuarios de redes sociales en línea pertenecen a más de un grupo o comunidad, por lo que se deben detectar comunidades con traslapes basadas en un grafo con atributos. En este contexto, las comunidades se conocen como comunidades sobrepuestas (*covers* en inglés) $C = \{c_1, c_2, ..., c_q\}$ para $q$ número de comunidades tal que $q = |C|$, donde un nodo $v_i$ se asocia a las comunidades con un factor de pertenencia $p_{v_i c_j}$ que determina la pertenencia del nodo $v_i$ con la comunidad $c_j$. Se asume que $0 \leq p_{v_i c_j} \leq 1$ para $\forall v_i \epsilon V, \forall c_j \epsilon C$.

## 1.4 Contribuciones

En este trabajo se detectan comunidades sobrepuestas en redes sociales con atributos por medio de un método mixto: basado en modelo y basado en distancia. Estos pueden ser usados de manera independiente o integral. Además, se seleccionan los atributos más importantes para reducir la complejidad del algoritmo. A continuación se da una breve descripción de las propuestas generadas:

1. **Método basado en modelo (RMOCA)**. El modelo propuesto define una

función objetivo, la cual considera que las comunidades se generan a partir de los enlaces y a partir de la relación entre nodos y atributos. Las comunidades se estiman haciendo uso de un modelo de regresiones con mínimos cuadrados, el cual toma como entradas la matriz de adyacencia del grafo y una matriz de atributos que mapea la relación de nodos con atributos. A través de la optimización de la función objetivo obtenida en el modelo, se generan comunidades de nodos ($C_S$) y comunidades de atributos ($C_A$). Esta propuesta mejora la precisión y recuperación de las comunidades, comparadas con las comunidades esperadas, es decir encuentra las comunidades reales o naturales de las redes sociales.

2. **Medida de calidad de comunidades considerando importancia de atributos (BAS)**. Se define una nueva calidad de comunidades que considera que los atributos tienen diferente influencia en la generación de las comunidades, de tal forma que la importancia de los atributos varía no sólo con respecto a la red, sino también con respecto a cada una de las comunidades que se generan, por lo que se proponen importancias globales y locales para cada atributo. Esta importancia del atributo es probabilística y está basada en la estructura. La importancia global de un atributo también nos permite hacer un *ranking*[2] para seleccionar los atributos más importantes. Las comunidades son detectadas con la medida de calidad propuesta que balancea cohesión estructural y homogeneidad de atributos, considerando probabilidades y pesos para los atributos según la estructura, además de la conductividad. Considerar los atributos incrementa sobre todo la similitud entre los nodos de una misma comunidad, es decir la entropía, sin afectar la densidad.

3. **Método mixto**. RMOCA y BAS son independientes, sin embargo han sido

[2]Según la RAE ranking es una palabra usada en español, proveniente de una voz inglesa, cuyo significado es la clasificación de mayor a menor, útil para establecer criterios de valoración.

integrados a través de un proceso de **expansión de comunidades** que transforma las comunidades disjuntas en sobrepuestas, lo que permite obtener un método mixto basado en modelo y distancia. El proceso de expansión ha sido planteado para realizarse en algoritmos base como Girvan-Newman así como en nuestra propuesta de RMOCA. Dado que el algoritmo basado en modelo obtiene comunidades naturales y mejora la entropía, la expansión de BAS permite mejorar las comunidades, aumentando la entropía y mejorando la densidad. El *ranking* dado por la importancia global de un atributo ($W_a(G)$), es usado para seleccionar los atributos más importantes como un pre-proceso al método mixto para reducir la complejidad.

## 1.5 Organización de la tesis

En los siguientes capítulos se encuentran definiciones y el estado del arte correspondiente a redes sociales (Capítulo 2) y a la detección de comunidades (Capítulo 3). Como se ha mencionado, las redes sociales presentan características que las distinguen de cualquier grafo, se tienen diversas formas de representarlas y se han hecho muchos estudios derivados de éstas en el área de SNA (*Social Network Analysis*). En cuanto a las comunidades, se muestra un estudio que va desde los algoritmos clásicos hasta aquellos enfocados en comunidades sobrepuestas y aquellos que integran atributos.

El método basado en modelo, RMOCA (*Regression Model for Overlapping Community with Attributes*) se describe en el Capítulo 4, en donde también se presentan los experimentos que comparan a éste con los del estado del arte donde se obtienen comunidades similares a las reales.

En el Capítulo 5 se propone una nueva medida de calidad de comunidades basada en atributos $Q_A$ y la medida de similitud BAS (*Balanced Attribute and Structure mea-*

*sure*) que balancea el peso de la estructura y el peso de los atributos en determinada comunidad. Se describen y se hacen experimentos de su uso tomando como base el método tradicional de Girvan-Newman (GN) donde se obtiene comunidades densas con nodos similares.

La integración del método basado en modelo y la distancia genera el método mixto descrito en el Capítulo 6. Inicialmente, se compara la integración de RMOCA+BAS con los del estado del arte en precisión, densidad y similitud interna. Posteriormente, se integra la selección de atributos relevantes dados por la importancia global $W_A(G)$ en nuestro modelo mixto y en CESNA donde se reduce el tiempo de ejecución.

Finalmente en el Capítulo 7 se presentan las conclusiones y trabajo a futuro.

# Capítulo 2

# Redes sociales

El análisis de redes sociales está ligado a análisis sociológicos y a un conjunto de técnicas metodológicas, cuyo objetivo es describir y explorar los aparentes patrones en relaciones sociales que forman los individuos. En los años 1930, sociólogos y antropólogos comenzaron a utilizar las matemáticas para formalizar el estudio de redes sociales. En los 1970 los trabajos técnicos y las aplicaciones especializadas dieron pauta a la formalización de los conceptos para el análisis de redes sociales.

Los tipos de datos usados en ciencias sociales son las relaciones (*relational data*), los atributos (*attribute data*) y los conceptos (*ideational data*). Los datos relacionales se refieren a contactos, vínculos y conexiones estudiados por el análisis de relaciones (*relational analysis*) cuya base es la teoría de grafos, que es el que ha dado mayor terminología y métricas al análisis de redes sociales. Los atributos son los relacionados con actitudes, opiniones y comportamientos vistos tradicionalmente como propiedades, cualidades o características que pertenecen a un grupo o a un individuo; estos elementos son mayormente estudiados con análisis univariable o multivariable. Los datos conceptuales describen el significado, motivos, definiciones o tipificaciones de las acciones, que son los menos estudiados con técnicas de análisis topológico.

En esta tesis no sólo nos enfocamos en los datos relacionales, como tradicionalmente se hace, sino también consideramos los atributos dado que son variables que influyen en la formación de grupos en las redes sociales. En esta sección, damos algunos conceptos de redes sociales como las variables, su representación, las métricas y las propiedades que las distinguen de grafos aleatorios.

## 2.1 Variables que componen las redes sociales

Las redes sociales son estructuras compuestas por un conjunto de participantes con relaciones entre ellos. El elemento principal de una red social, y el más estudiado, es la estructura. Sin embargo, las redes sociales tienen tres variables según Cruz et. al. [21]: **estructura**, **composición** y **afiliación**. Estos elementos permiten la representación de los grafos en redes sociales, mostrando la interacción entre individuos, así como las características de ellos. El objetivo es analizar desde diferentes perspectivas la misma red aumentada. Por ejemplo, en la Figura 2.1 tenemos una red de políticos mexicanos [39] que representa a 35 políticos entre los que se encuentran los presidentes de México posteriores a la Revolución Mexicana.

**Estructural.** Esta información muestra las conexiones entre elementos de la red, estas pueden ser: relaciones como amistad en Facebook, interacciones como escribir en el muro de un usuario de Google+, latentes como la búsqueda de perfiles en Linked-in, o de subscripción como seguir un perfil en Snapchat. Este tipo de datos permite un análisis estructural y basado en enlaces que estudia el comportamiento de los vínculos en una red para identificar los nodos que tienen más influencia, las comunidades, las sugerencias de conexión, entre otras. En la Figura 2.1, los nodos representan al político, mientras que las aristas son relaciones de amistad, de negocios o parentescos.

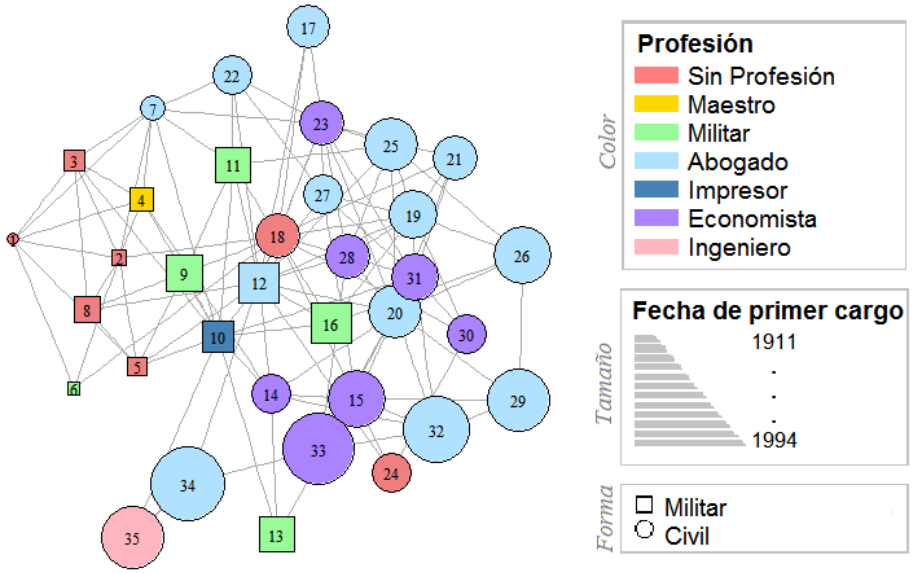



Figura 2.1: Red de políticos mexicanos, entre los que se encuentran los presidentes de México de 1911 a 1994. Los nodos representan a los políticos, las aristas son relaciones familiares, de amistad o negocios y los atributos se visualizan en relación al tamaño para el año de primer cargo público, el color varía según su profesión y forma para la corriente post-revolucionaria.

**Composición.** Se refiere a la información de cada nodo de la red de tal forma que describe cada elemento en un contexto en particular. Tenemos información individual y meta-información. La primera se refiere a las características de un usuario como pueden ser los *pinboards*[1] de un usuario en Pinterest o los horarios de mayor ocupación de un lugar en Foursquare, y la segunda hace referencia al uso y datos asociados a un elemento en particular como la dirección MAC. Por ejemplo, en la Figura 2.1 los atributos son: el año en que ocuparon su primer puesto político (tamaño) y la profesión que ejercieron (color). Este tipo de información permite el análisis basado en contenido, ya sea categórico como la profesión o numérico como el año.

**Afiliación.** Esta información representa la pertenencia de un nodo a un grupo o comunidad. Por ejemplo, en las comunidades de cooperación, se podrían estudiar

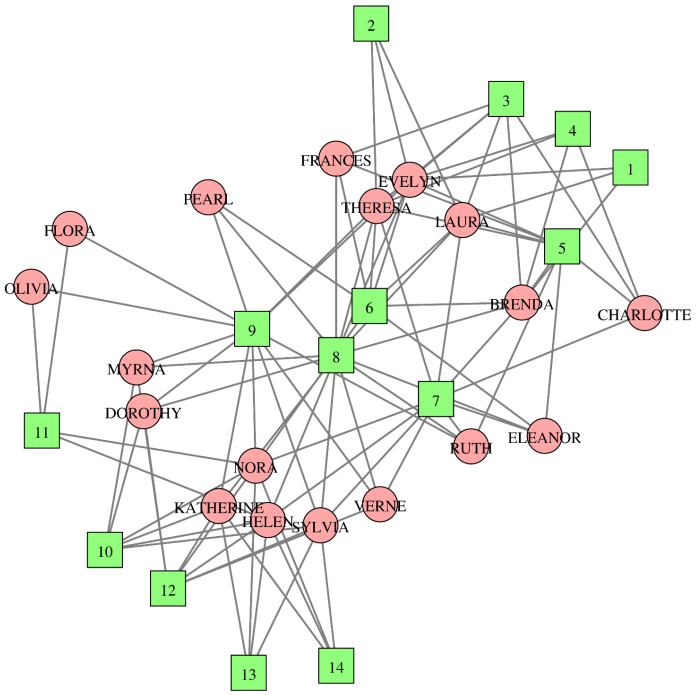
[1]Los *pinboards* siguen la metáfora de los tableros en los que se fijan con *pines* las fotos y temas que interesan.

las citas de artículos científicos de un área para obtener la línea de investigación en la que cada científico se desarrolla. Cabe destacar que los perfiles podrían presentar más de un aspecto de organización. En el estudio que se realizó sobre la red social de la Figura 2.1, los autores determinaron que tras la revolución había un grupo militar (nodos cuadrados) y uno civil (nodos circulares), lo que mostraba su afinidad política ya que el grupo militar fue asociado a políticos revolucionarios.

## 2.2 Representación de redes sociales

En minería de datos, las redes sociales se definen como un conjunto heterogéneo y multirelacional de datos representado por un grafo [47]. En un grafo $G(V, E)$ las entidades son denotadas por los nodos o vértices tal que $V$ representa el conjunto de vértices $V = \{v_i, v_2, ..., v_n\}$ para $n = |V|$, mientras que el conjunto de aristas $E$ son los vínculos que denotan interacciones o relaciones entre dos entidades, de tal forma que es un conjunto de pares de la forma $(v_i, v_j)$ tal que $v_i, v_j \in V$. Las entidades pueden ser neuronas en redes biológicas, nodos de telefonía o de personas representados por usuarios en redes sociales en línea. Las interacciones en las redes sociales se relacionan a la conectividad y la distancia en grafos, de tal forma que se tienen interacciones entre elementos dadas por llamadas, transmisión de enfermedades, relaciones de amistad, hypervínculos, búsquedas, entre otras. En una red social como Twitter, los nodos son los usuarios y las interacciones como seguir (*follow*), ser seguido (*followed by*) y menciones son diferentes tipos de aristas en la red. En Facebook los nodos son los usuarios y las aristas son las relaciones de amistad.

La representación de una estructura social como una red o grafo facilita la comprensión y el análisis de ésta, permitiendo identificar características locales y globales, obtener los participantes que son una influencia, detectar las agrupaciones y estudiar la evolución de la red [114]. Esta representación se conoce como sociograma, la cual




(a) Red bipartita donde los círculos representan mujeres y los cuadrados eventos a los que asistieron

| | E1 | E2 | E3 | E4 | E5 | E6 | E7 | E8 | E9 | E10 | E11 | E12 | E13 | E14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Evelyn | 1 | 1 | 1 | 1 | 1 | 1 | | 1 | 1 | | | | | |
| Laura | 1 | 1 | 1 | | 1 | 1 | 1 | 1 | | | | | | |
| Theresa | | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | | | | | |
| Brenda | 1 | | 1 | 1 | 1 | 1 | 1 | 1 | | | | | | |
| Charlotte | | | 1 | 1 | 1 | | 1 | | | | | | | |
| Frances | | | 1 | | 1 | 1 | | 1 | | | | | | |
| Eleanor | | | | | 1 | 1 | 1 | 1 | | | | | | |
| Pearl | | | 1 | | | 1 | | 1 | 1 | | | | | |
| Ruth | | | | 1 | 1 | | 1 | 1 | 1 | | | | | |
| Verne | | | | | | | 1 | 1 | 1 | | | 1 | | |
| Myra | | | | | | | | 1 | 1 | 1 | | 1 | | |
| Katherine | | | | | | | | 1 | 1 | 1 | | 1 | 1 | 1 |
| Sylvia | | | | | | | 1 | 1 | 1 | 1 | | 1 | 1 | 1 |
| Nora | | | | | | 1 | 1 | | 1 | 1 | 1 | 1 | 1 | 1 |
| Helen | | | | | | | 1 | 1 | | 1 | 1 | 1 | | |
| Dorothy | | | | | | | | 1 | 1 | | | | | |
| Olivia | | | | | | | | | 1 | | 1 | | | |
| Flora | | | | | | | | | 1 | | 1 | | | |

(b) Matriz de adjacencia entre dos tipos de nodos (nodos de mujeres y nodos de eventos)

Figura 2.2: Red bipartida de mujeres asistiendo a eventos estudiada en 1942 con su representación matricial [12]

podemos observar en la Figura 2.2(a) que corresponde a un grafo bipartita[2] cuyos enlaces representan la asistencia de los "nodos mujeres" a los "nodos eventos". Davis et. al. [22] mapearon la asistencia de las 18 mujeres a 14 eventos representados con matrices donde las columnas denotaban los eventos y las filas representaban a las mujeres como se puede observar en la Figura 2.2(b). Esta red bipartida sentó la base para la representación matricial de las redes sociales a través de matrices de adjacencia, donde cada elemento de la matriz representa una relación (arista) entre los nodos.

### 2.2.1 Redes sociales con atributos

Existen elementos pertenecientes al perfil (datos del usuario, mensajes privados, fotos, etc.) y metadatos asociados a ellos (marcas de tiempo, localización, etc.). Dichos elementos, que llamaremos atributos, son agregados sobre el sociograma. Supongamos que la Figura 2.3 muestra un fragmento de la red social de Facebook; los nodos representan los usuarios, las aristas son las amistadas entre los usuarios y podemos observar la lista de atributos que representan nombre, apellido, educación, cumpleaños, edad, género, entre otros. Actualmente, los atributos son usados para diversos fines, por ejemplo, para mostrar anuncios, Facebook considera los atributos como localidad, edad, género, nivel de educación, afinidad étnica, campo de estudio, vivienda, aniversarios, nuevos empleos o relaciones, cumpleaños cercanos, autos e intereses, además del rastreo de actividad [3].

La integración de atributos al sociograma en un solo modelo se encuentra en redes reales representado por el grafo con atributos $G^{+}(V, E, A)$, el cual cuenta con vértices

[2]Grafo cuyos vértices se pueden separar en dos conjuntos disjuntos, es decir que las aristas sólo pueden conectar vértices de un conjunto con vértices del otro

[3]Listado completo en https://www.washingtonpost.com/news/the-intersect/wp/2016/08/19/98-personal-data-points-that-facebook-uses-to-target-ads-to-you/?utm_term=.ae30309545bd

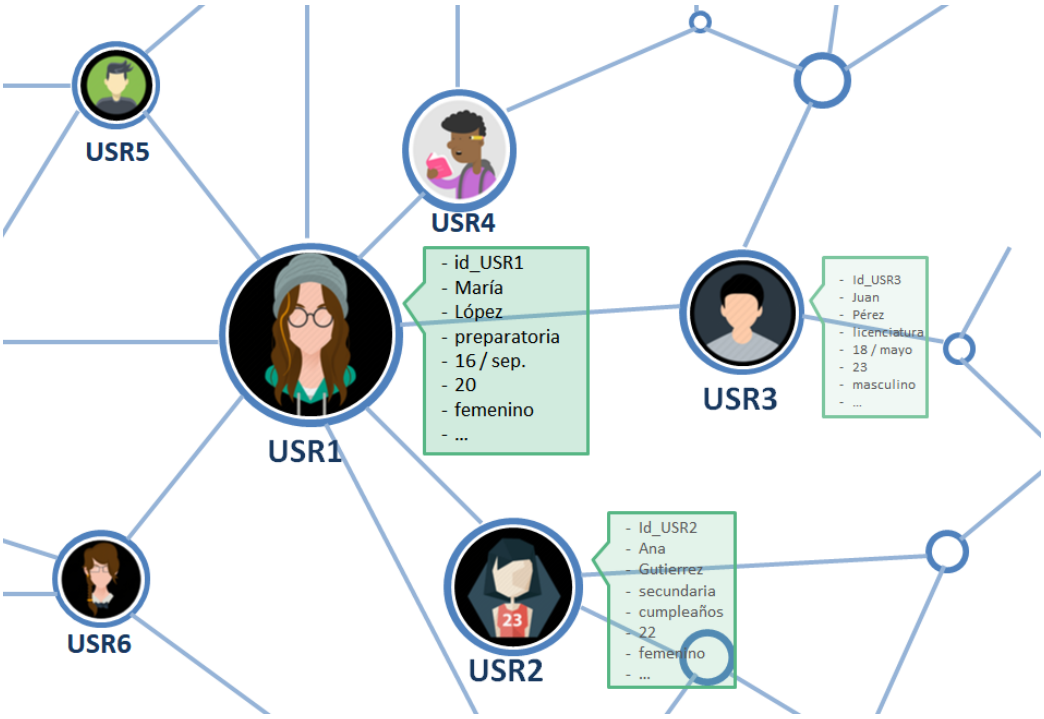


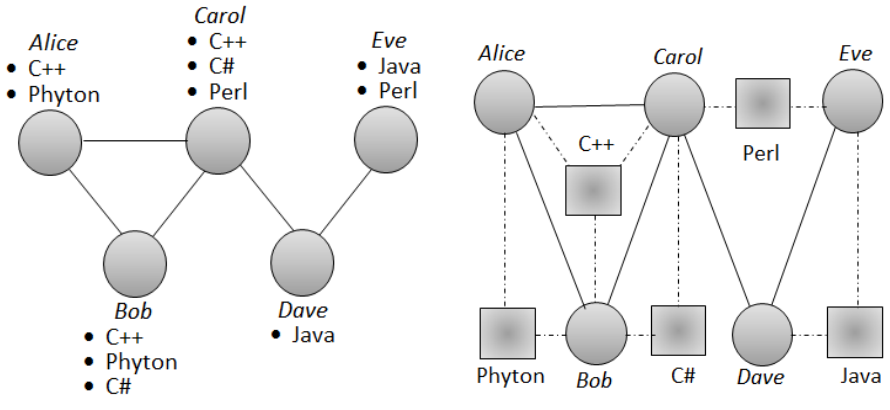

Figura 2.3: Ejemplo de fragmento de la red social Facebook

$V$, aristas $E$ y atributos $A$. Los nodos o vértices $V$ representan entidades sociales como personas, cosas o lugares. Las aristas $E$ representan la conexión entre un par de nodos $(v_u, v_w)$ que puede ser dirigida o no dirigida como amistad, compra o visitas. Los atributos $A$ son características de las entidades como el perfil, contenido generado o comportamiento. Por ejemplo, en la Figura 2.4, $V = \{\mathbf{A}lice, \mathbf{B}ob, \mathbf{C}arol, \mathbf{D}ave, \mathbf{E}ve\}$, $E = \{(A, C), (A, B), (B, C), (C, D), (D, E)\}$ y $A = \{C++, Phyton, Perl, Java\}$.

Dos de los principales modelos de sociogramas con atributos son: grafo con vectores de atributos y grafo aumentado, los cuales se explican a continuación:

#### 2.2.1.1 Grafo con vector de atributos

Dado un sociograma $G(V, E)$, donde $V$ es el conjunto de nodos y $E$ el conjunto de aristas, cada nodo $v_i \in V$ representa un elemento en la red y cada arista $e_j \in E$




(a) Grafo con vector de atributos (b) Grafo aumentado

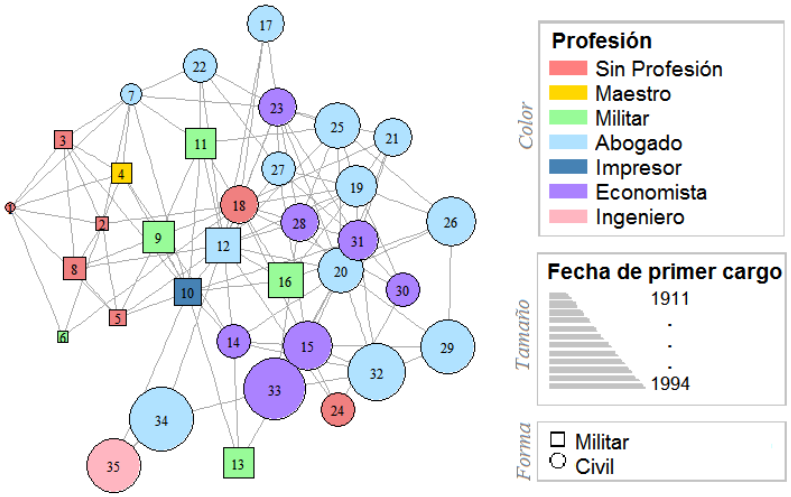
Figura 2.4: Representación de redes sociales con atributos

representa un enlace entre dos nodos. La inclusión de los atributos a la representación de la red social está dada por:

$$G'(V, E, A) \tag{2.1}$$

donde $A$ representa todos los $k$ atributos para $k = |A|$ tal que existe una función $\Lambda$ que relaciona los atributos $A$ con los nodos $V$ dada por $\Lambda : V \longrightarrow A$. Por lo tanto cada nodo $v_i$ tiene su propia lista de atributos $\Lambda_i$ tales que $\Lambda_i \subseteq A$, donde $\Lambda_i = \{a_1, a_2, ..., a_l\}$, $l \leq k$. Este modelo (ver Figura 2.4(a)) es usado para detectar comunidades disjuntas [119] y comunidades sobrepuestas [124]; y será el mismo modelo que usaremos para el desarrollo de esta tesis.

Estos vectores tienen una representación binaria de los atributos, por ejemplo, en la Figura 2.5 tenemos atributos categóricos como la profesión y su antecedente revolucionario, así como atributos numéricos como la fecha del primer cargo. En la parte inferior de la imagen, vemos que los valores de los atributos del nodo 10 son representados por unos y ceros; los valores numéricos fueron categorizados por décadas y los valores categóricos se representan en el vector de atributos de manera directa. Para el caso del Nodo 10 observamos que tiene UNO en el atributo de impresor; en la década de los 30 toma su primer cargo y tenía una postura revolucionaria.




| Dimensión | Profesión | | | | | | | Fecha de primer cargo | | | | | | | | | Rev. | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Atributo | | | | | | | | '10s | '20s | '30s | '40s | '50s | '60s | '70s | '80s | '90s | □ | ○ |
| Nodo 10 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

Figura 2.5: Binarización de atributos

#### 2.2.1.2 Grafo aumentado

Dado el sociograma $G(V, E)$, los atributos $A = \{a_1, a_2, ..., a_k\}$, para $k = |A|$, se introducen en un nuevo grafo $G''(V', E')$. Para cada nodo $v_i \in V$ en el grafo $G$ se crea un nodo correspondiente en el grafo $G''$. Por cada atributo $a_j$, se crea adicionalmente un *nodo atributo* $v_A$ en $G''$, tal que $V' = V \cup V_A$. Para cada atributo que tenga un nodo, se crea una arista $e_A$ entre el nodo $v_i$ y el nodo atributo $v_{Aj}$ con un peso correspondiente a $w(v_i, v_{A_j})$. Por cada arista $e_i \in E$, se crea una arista en el grafo $G''$ tal que $E' = E \cup E_A$ (ver Figura 2.4(b)).

Zhou et. al. [131] propusieron este modelo conocido como *Social-Attribute Network (SAN)*, como una red social aumentada con atributos para integrar los atributos a la red social; posteriormente Yin et. al. [125] dieron algunos pesos específicos para las aristas de este modelo.

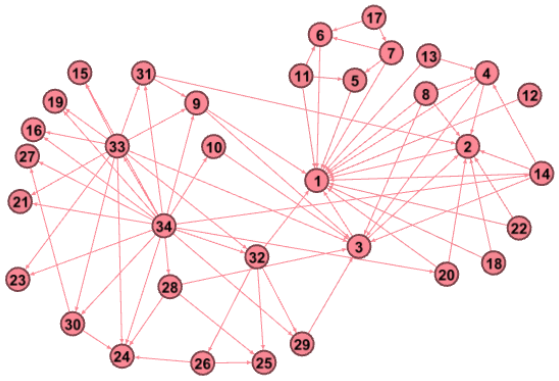



Figura 2.6: Red del Club de Karate de Zachary

## 2.3 Medidas en redes

Para entender el comportamiento de los sistemas complejos se tienen medidas locales que se evalúan a nivel actor o medidas globales que se evalúan a nivel red. Estas medidas de redes sociales serán explicadas a continuación; para ello haremos uso de la conocida Red del Club de Karate estudiada por Wayne W. Zachary [127] que se muestra en la Figura 2.6, donde el tamaño de la red es de 34 nodos con 78 aristas.

### 2.3.1 Medidas locales

Las medidas locales se basan en el concepto de centralidad (en redes no dirigidas) o prestigio (en redes dirigidas). Son usadas para evaluar a los actores por lo que una de las mayores aplicaciones es encontrar a los actores clave o influyentes. A continuación se explican el grado, intermediación (*betweenness*) y cercanía (*closeness*) propuestas con Freeman [38], centralidad de vector propio [9], excentricidad y coeficiente de *clustering*.

- **Grado**. Corresponde al número de aristas para un vértice. En un grafo no

dirigido, el grado $k_v$ de un nodo $v$ se define en $\{0, g-1\}$, siendo $g$ el número de nodos de la componente conexa. El nodo de mayor grado en la Red de Karate es el nodo 34 con un grado de 17 si no consideramos la dirección de las aristas.

En un grafo dirigido, se define un grado de entrada $k_v^{in}$ de un nodo $v$ conocido también como prestigio de entrada y el grado de salida $k_v^{out}$ de un nodo $v$ también conocido como prestigio de salida. Por ejemplo, en la Figura 2.6 los prestigios del nodo 26 son: $k_v^{in} = 1$ y $k_v^{out} = 2$ que corresponden a la aristas $\{(32, 26)\}$ para el prestigio de entrada y a las aristas $\{(26, 24), (26, 25)\}$ para el prestigio de salida.

- **Intermediación (*betweenness*)**. La centralidad de intermediación (*betweenness*) de un vértices $v$ es el número de caminos más cortos de todos los pares de la red que pasan por $v$. Entre más grande sea el grado de *betweenness*, mayor influencia tiene ese nodo en la red. La intermediación puede ser medida para vértices o para aristas, en el caso de los vértices se conoce como *betweenness centrality*. La siguiente ecuación calcula la intermediación del nodo $v_n$ en la trayectoria de $v_i$ y $v_j$.

$$C_B(v) = \sum_{v \neq v_i \neq v_j} \frac{d_{v_i v_j}(v)}{d_{v_i v_j}} \tag{2.2}$$

  donde $d_{v_i v_j}(v)$ representa el número de caminos más cortos de $v_i$ a $v_j$ que cruzan por $v$, y $d_{v_i v_j}$ el número de caminos más cortos que van de $v_i$ a $v_j$. Por ejemplo, en la Figura 2.6 los nodos 1, 32 y 34 son los que tienen un grado más alto de centralidad, ya que por estos nodos pasan los caminos más cortos de los nodos de la parte derecha a los nodos de la parte izquierda.

- **Cercanía**. La centralidad de cercanía (*Closeness*) mide los pasos requeridos para acceder a cada vértice de un vértice dado $v$. Esta medida se basa en el uso del camino más corto entre un nodo y los demás nodos de la red. Se define como la inversa de la longitud promedio de los caminos cortos con los demás

vértices de la red:

$$C_C(v) = \frac{1}{\sum_{i=1}^{|V|} d(v, v_i)} \tag{2.3}$$

donde $d(v, v_i)$ es el camino más corto entre el nodo $v$ y el nodo $v_i$. Por ejemplo, para la red de karate el nodo más lejano al resto de los nodos es el nodo 17, por lo que tiene la cercanía más alta, ya que recorre más nodos en los caminos más cortos; mientras que el nodo 1 es el que requiere distancias más cortas para llegar al resto de los nodos.

- **Centralidad de excentricidad**. La excentricidad es la máxima distancia geodésica entre un actor y cualquier otro actor de la red. La centralidad de excentricidad $C_E(v)$ de un nodo $v$ se define como:

$$C_E(v) = \frac{1}{\max_{v_i \in V} d(v, v_i)} \tag{2.4}$$

donde los actores con un mayor valor de excentricidad se denominan actores periféricos y los de menor valor forman el centro de la red. En el ejemplo de la red de karate hay nueve nodos con la mayor excentricidad, dados por los nodos 15, 19, 16, 27, 21, 23, 30, 24 y 17 que están en la periferia izquierda de la Figura 2.6. Mientras que el de menor valor de excentricidad es el nodo 1 que coincide con el de la centralidad de cercanía. Los otros tres nodos con baja excentricidad son los nodos 3, 2 y 4 que se conectan con el nodo 1.

- **Centralidad de vector propio**. La centralidad de vector propio $C_{VP}(v)$ se basa en que la centralidad de un nodo $v$ depende de qué tan centrales sean sus vecinos. Se define como una combinación lineal de los valores de sus actores:

$$C_{VP}(v) = \sum_{\{v_i | v_i \in V, (v, v_i)\}} C_{VP}(v_i) \tag{2.5}$$

En la red de karate, el nodo con la mayor centralidad de vector propio es el

nodo 34, seguido de los nodos 1, 33, 3 y 2; mientras que el que tiene la menor centralidad de vector propio es el nodo 17.

- **Coeficiente local de agrupamiento o transitividad**. Las redes sociales son transitivas por naturaleza, es decir, los amigos de un actor también suelen ser amigos entre sí. Esta puede ser local o global. El coeficiente local de agrupamineto (*clustering*) de un nodo $v$ está dada por:

$$C_v = \frac{2L_v}{k_v(k_v - 1)} \tag{2.6}$$

donde $L_v$ es el número de enlaces entre los vecinos del nodo $v$ y $k_v$ es el grado de un nodo $v$.

### 2.3.2 Medidas globales

La mayoría de las medidas globales para análisis de redes sociales son similares a las empleadas en grafos. A continuación se detallan las más importantes:

- **Diámetro**. El camino más corto en grafos es el camino entre cualesquiera dos vértices que pasa por el menor número de aristas. El **diámetro** de un grafo es el máximo de los caminos más cortos. Está dado por el valor máximo de excenticidad para todos los nodos de la red:

$$d_{max}(v) = \max\{E(v) : v \in V\} \tag{2.7}$$

Por ejemplo, en la el grafo de la Figura 2.6 el diámetro es de 5, correspondiente al camino dado por los nodos 15, 33, 3, 1, 6 y 7.

- **Distancia media**. La distancia media de un grafo dirigido se define como:

$$d = \frac{1}{2L_{\max}} \sum_{v_i, v_j \neq v_i} d_{v_i v_j} \tag{2.8}$$

tal que $L_{v_i}$ es el número de enlaces entre los vecinos del nodo $v_i$ y $d_{v_i v_j}$ es la distancia geodésica entre los nodos $v_i$ y $v_j$. Esto permite saber qué tan lejos están los distintos actores en promedio, lo cual representa la eficiencia del flujo de información en la red.

- **Grado medio**. El grado medio se define como la media del grado de los nodos. En un grafo no dirigido, el grado medio se define como:

$$k = \frac{1}{|N|} \sum_{v_i=1}^{|N|} k_i \tag{2.9}$$

tal que $|N|$ es el número de nodos. En el caso de la red de karate, la media de los grados es de 4.5882 *aristas/vertices* sin considerar la dirección de las aristas. En un grafo dirigido tenemos un grado medio de entrada y un grado medio de salida dados por:

$$k^{in} = \frac{1}{|N|} \sum_{i=1}^{|N|} k_{v_i}^{in}, k^{out} = \frac{1}{|N|} \sum_{i=1}^{|N|} k_{v_i}^{out} \tag{2.10}$$

- **Densidad**. La densidad mide el grado de conectividad de la red social a nivel global. Está dado por:

$$D = \frac{L}{L_{\max}} \tag{2.11}$$

tal que $L$ es el número de enlaces. En el ejemplo de la red estudiada por Zachary, la densidad es de 0.7 lo que significa que es muy alta.

- **Coeficiente de agrupamiento global**. El coeficiente de agrupamiento medio o transitividad indica la probabilidad de que dos vecinos de un nodo seleccionado

aleatoriamente esté conectados entre sí. Es decir, si los vértices $A$ y $B$ son vecinos, y los vértices $B$ y $C$ son vecinos, entonces existe una alta probabilidad de que $A$ y $C$ también estén conectados.

$$C_G = \frac{1}{N} \sum_{i=1}^{|N|} C_i \tag{2.12}$$

El coeficiente de agrupamiento $C_G$ también se puede medir basado en el número de tripletas en una red (conjunto de tres nodos conectados) y se calcula:

$$C_G = \frac{3 \times num\_de\_triangulos}{num\_de\_tripletas\_conectadas} \tag{2.13}$$

donde *num_de_tripletas_conectadas* hace referencia a tres nodos conectados por dos aristas, de tal forma que un triángulo entre nodos $A$, $B$ y $C$ tiene tres tripletas conectadas. En el ejemplo de la red de karate, el coeficiente global de agrupamiento es de 25.6% dados 45 triángulos que existen en dicha red.

Existen otras propuestas para el cálculo de la transitividad como la dada por Watts y Strogatz, quienes definen un *clique* como un subgrafo en el cual cada vértice está conectado a cada otro vértice del grafo, es decir, el subgrafo puede ser considerado como un grafo completo. Los *cliques* derivan del concepto de transitividad. Por ejemplo, en la Figura 2.7 podemos ver un *clique* de tamaño 4 ente los nodos en verdes y un *clique* de tamaño 6 entre los nodos naranja.

En las redes sociales se tiene una transitividad parcial, la cual se refiere a que no se garantiza que $A$ conozca a $C$ dados los enlaces entre $(A, B)$ y $(B, C)$, pero sí es más probable que se conozcan, de tal forma que existe una mayor probabilidad de que $A$ conozca a $C$ comparado con la probabilidad de que conozca algún otro miembro de la red seleccionado de manera aleatoria.

Las redes sociales son redes de mundos pequeños y suelen tener valores altos, con una mayor transitividad y aumentando la probabilidad de *cliques*.

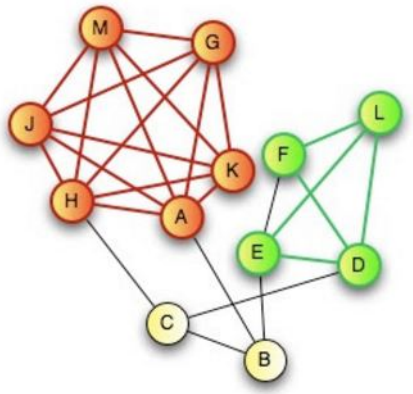



Figura 2.7: *Clique* de tamaño 6 y 4 [133]

- **Conectividad**. La conectividad es el grado en que los nodos están directamente conectados. Cuando la red tiene una alta conectividad significa que que el número de aristas es mucho mayor que el número de nodos. Se calcula usando la siguiente ecuación:

$$C_O = \frac{m}{n(n-1)} \tag{2.14}$$

donde $m$ es el número de aristas tal que $m = |E|$ y $n$ es el número de nodos en la red tal que $n = |V|$. Para la red de karate, la conectividad es de 0.06951 dado que existen solo 78 aristas de las 1122 posibles dadas por $n(n-1)$.

## 2.4 Propiedades de redes sociales

Las propiedades de las redes sociales, que se modelan como grafos, difieren de las redes tradicionales. Existen varios estudios enfocados en distinguir las propiedades de las redes sociales, sin embargo, aún no existe un criterio unificado. Algunos de los elementos destacados que comparten las redes sociales en línea y fuera de línea son: organización en cuatro capas jerárquicas, el tamaño de las capas escala en un factor cercano a tres, y que el número de relaciones sociales activas es cercano al

número de Dunbar. Las propiedades que distinguen a las redes sociales de otros tipos de redes son: efecto del mundo pequeño [112] [70] también conocida como seis grados de separación, comportamiento libre de escala [5] [35], distribución *power law* y transitividad parcial.

### 2.4.1 Efecto del mundo pequeño

El diámetro de una red es el camino más largo entre dos par de nodos. En redes sociales, el diámetro es pequeño y oscila según la propiedad de los seis grados de separación, por lo que la información se esparce mucho más rápido. El problema del mundo pequeño fue plateado en 1967 [70] en busca de la probabilidad de que cualesquiera dos personas en el mundo se conozcan, y del número de enlaces intermedios entre dos personas. Se realizó un estudio del monitoreo de grupos de personas que interactuaban enviando cartas, lo cual determinó que los individuos están a seis grados de separación [113], es decir que el número de personas intermedias entre dos personas en el mundo es menor a cinco. En fechas más recientes se han realizado estudios similares con correos electrónicos [27], con redes de computadores y aplicaciones de mensajería intantánea [63] y se han obtenido resultados similares.

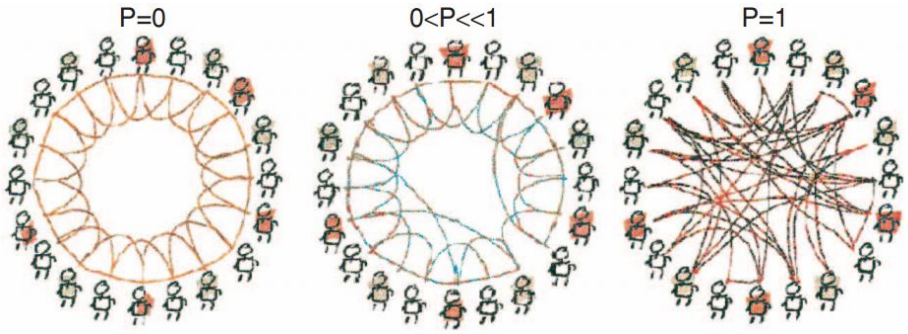



Figura 2.8: Poceso de rescritura aleatoria de aristas para anillo regular a una red aleatoria [108]

Además de los diámetros pequeños, las redes con efecto de mundo pequeño tienden a ser altamente conectadas, lo cual se mide con el promedio del coeficiente de

agrupamiento de una red. De tal forma que, a pesar de que la mayoría de los nodos no son vecinos de otros, la mayoría de ellos puede ser alcanzado desde cualquier otro nodo con pocos intermediarios. En el modelo de Watts y Strogatz [112] se tiene la longitud $L(p)$ que mide la distancia entre dos nodos en una red, donde $p$ es la probabilidad de que las aristas de un grafo regular sean enlazadas a otro nodo. Por ejemplo, en la Figura 2.8 podemos ver una red en la que las personas (nodos) son amigos de sus cuatro vecinos más cercanos, conforme aumenta la probabilidad de reescribir las aristas vemos que, en promedio, siguen conociendo a cuatro personas, pero tiene algunos amigos distantes; cuando el grafo es aleatorio pocas personas tiene amigos en común. De tal forma que cuando tenemos un anillo, cuando $p$ está cerca de la media, tenemos una red de mundo pequeño y cuando $p = 1$ la red es aleatoria. Para una red de mundo pequeño, la longitud $L$ crece proporcionalmente al logaritmo del número de nodos $|N|$ en la red: $L \propto log|N|$.

### 2.4.2 Distribución *power law*

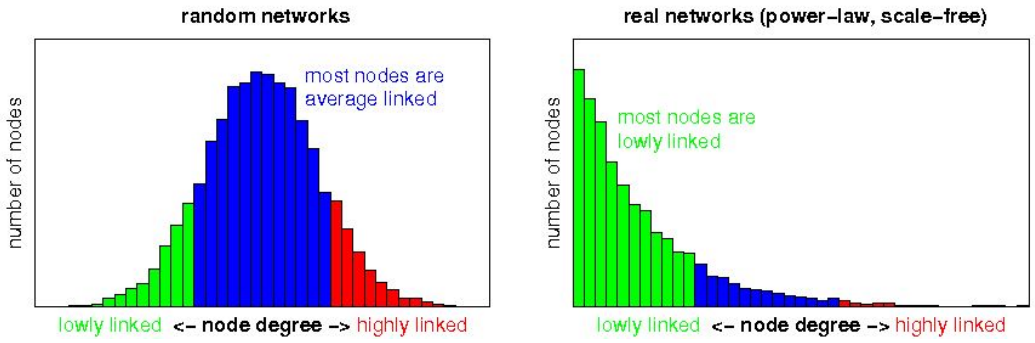



Figura 2.9: Grado de distribución en redes aleatorias y redes sociales [94]

La distribución de probabilidad que resulta de simulaciones numéricas de un grafo aleatorio corresponde al valor esperado de una distribución de Poisson, como se muestra en la Figura 2.9(a).

Las redes sociales en línea son redes a gran escala con un alto orden. El grado

de distribución en redes sociales y de información es de ley de potencias (*power law*), lo cual significa que muchos de los nodos tienen un grado muy bajo, mientras que pocos nodos tienen altos grados como se observa en la Figura 2.9(b). A estas redes también se les conoce como redes de libre escala. Por ejemplo, en Twitter, los nodos de celebridades llegan a tener más de 40 millones de seguidores. De tal forma que hay más nodos con pocos enlaces que con muchos enlaces, lo que garantiza una alta conectividad. Esto permite un comportamientos de anexo preferencial que consiste en que un nuevo nodo se agregue a la red con un enlace a un nodo con alto grado. Si se grafica el grado de distribución de la red de karate (ver Figura 2.6) obtenemos una distribución de libre escala como se ve en la Figura 2.10.

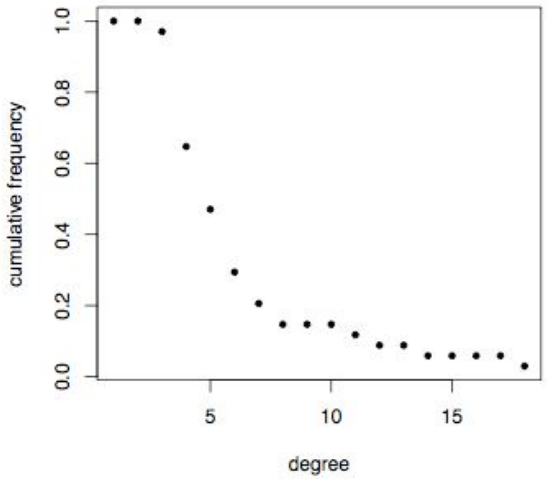



Figura 2.10: Grado de distribución de red de karate

### 2.4.3 Comunidades

Como se ha mencionado, en las redes sociales, los actores tienden a formar grupos con conexiones muy fuertes; estos grupos se conocen como comunidades, de tal forma que un grupo de entidades tiene una alta densidad dentro del grupo y escasa densidad entre grupos. Las comunidades también son llamadas grupos, agrupamien-

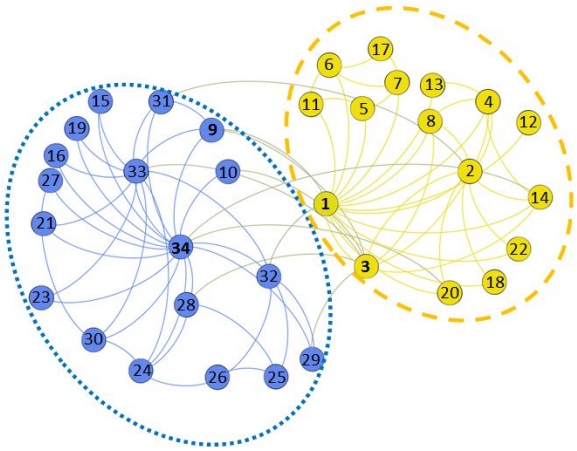



Figura 2.11: Comunidades de la red de karate tras el conflicto entre el administrador y el instructor.

tos, subgrupos o módulos. El enfoque de una comunidad varía según su área de estudio, por ejemplo, para los sociólogos las comunidades permiten una construcción cultural; para los antropólogos es importante la interacción entre miembros de la comunidad como símbolos; los economistas se interesan en cómo la organización de la comunidad contribuye a la producción, distribución y consumo, mientras que a los políticos les interesan las prácticas colectivas; esto permite múltiples interpretaciones de comunidades, lo cual constituye uno de los principales problemas.

Por ejemplo, en la Figura 2.11 se muestra la red de karate antes presentada, en la cual se formaron dos grupos después de un conflicto entre el administrador y el instructor. Aún enfocándonos únicamente en redes sociales en línea hemos encontrado múltiples definiciones, ya que depende del contenido de cada red social la definición que mejor aplica a ella.

## 2.5 Tipos de redes sociales

Actualmente no existe una clasificación genérica de redes sociales, ya que dependen del enfoque de estudio. Pueden ser clasificadas con base en diversos factores como se

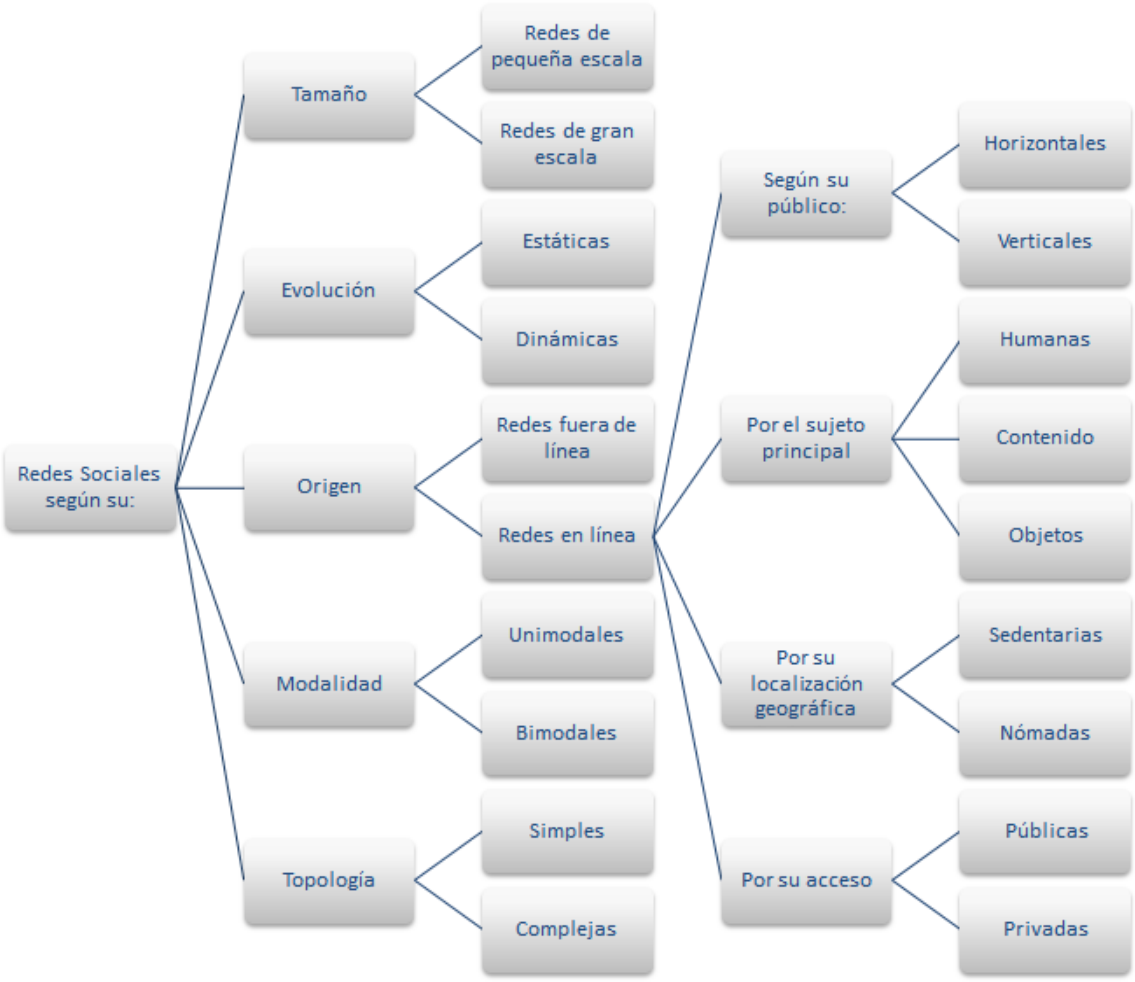



Figura 2.12: Clasificación de redes sociales

observa en la Figura 2.12. En esta tesis, nos enfocamos en el estudio de redes sociales complejas, estáticas y unimodales, de cualquier tamaño u origen. Las redes sociales presentan estructuras complejas según su estructura; aunado a ello se consideran las características de los nodos en un grafo con vectores (el uso de un grafo aumentado implicaría el uso de redes bimodales). Las redes sociales por naturaleza son dinámicas, sin embargo, se pueden estudiar en un lapso de tiempo determinado por lo que nos centraremos en la redes estáticas; por otra parte, la variación en tamaño nos permite evaluar la escalabilidad de la propuesta.

### 2.5.1 Redes según su tamaño

La clasificación basada en el tamaño, generalmente, depende del diámetro de la red, sin embargo, no existe un parámetro exacto para determinar si una red social es grande o pequeña.

- **Pequeña escala**. Se pueden analizar en sistemas de visualización de redes. Un ejemplo sería la colaboración entre investigadores dadas sus publicaciones o como la que hemos estudiado de los políticos mexicanos, que se muestra en la Figura 2.1.

- **Gran escala**. Estos son difíciles de analizar por lo que se toman proporciones representativas, como todos los correos electrónicos. Por ejemplo, la red social de Facebook es una red a gran escala que suele estudiarse por fracciones. En la Figura 2.13 se muestra un fragmento de las 9626 relaciones entre las 547 amistades de una persona (llamado ego-red) que corresponden a los nodos; como vemos ésta ya es difícil de interpretar y consta de una fracción muy pequeña de dicha red social.

### 2.5.2 Redes según su evolución

Depende de los cambios a través del tiempo, por lo que pueden ser estáticas o dinámicas. Sin embargo, esta clasificación se acota a las muestras de estudio, de tal forma que una red social dinámica en la vida real puede ser estudiada de manera estática si se toma un instante de tiempo de ésta.

- **Estáticas**. No sufren cambios mientras se estudian. En esta tesis se estudian este tipo de redes; si las redes sufren cambios en un periodo de tiempo, dichos cambios no son considerados para el estudio.

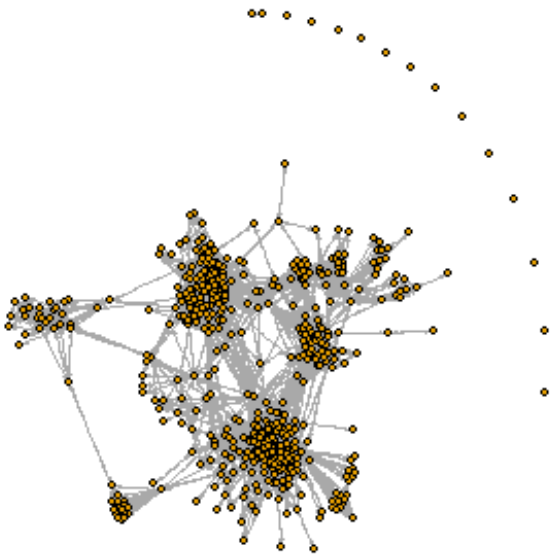


Figura 2.13: Ejemplo de red a gran escala según su tamaño; la imagen muestra una fracción correspondiente a la ego-red social de un usuario de Facebook con 547 nodos y 9626 relaciones entre las amistades del usuario estudiado.

- **Dinámicas**. Cambian su estructura por la incorporación o eliminación de nodos y/o enlaces. Por ejemplo, en la Figura 2.14 se muestra el curso temporal de una enfermedad simulada con brote inicial en Hong Kong, dicho modelo consideró el SARS originario de China en el 2003 y el H1N1 originario en México en el 2009.

### 2.5.3 Redes según su origen

Dependen de dos factores principalmente: de donde se originan y de donde se desarrollan. Y pueden ser de dos tipos: fuera de línea y en línea.

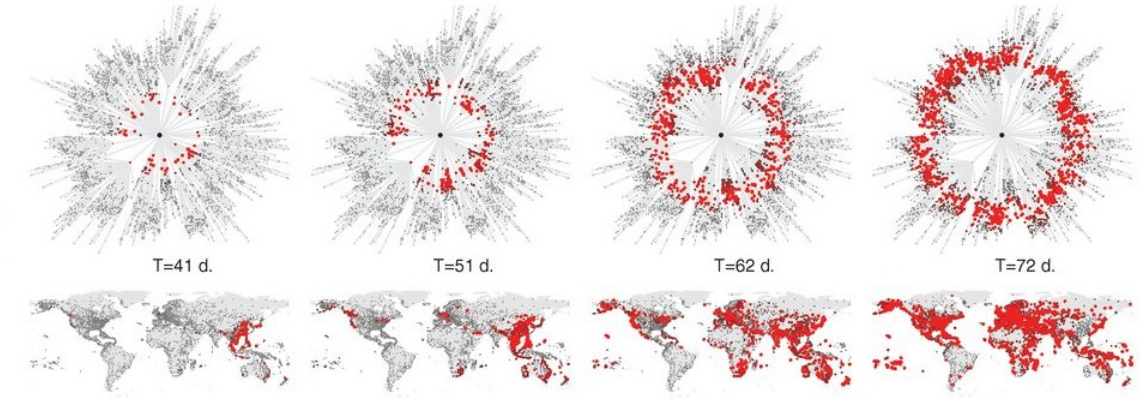



Figura 2.14: Ejemplo de red dinámica que muestra la difusión de una enfermedad. Cada panel compara el estado del sistema en la representación geográfica convencional (abajo) con la representación de distancia efectiva (arriba). [10]

#### 2.5.3.1 Fuera de línea (*off-line*)

En éstas, las relaciones sociales, sin importar su origen, se desarrollan sin mediación de aparatos o sistemas electrónicos. Por ejemplo, la red social de karate mostrada en la Figura 2.6 surge de la interacción entre un grupo de personas relacionadas con el karate.

#### 2.5.3.2 En línea (*on-line*)

Se originan y desarrollan a través de medios electrónicos, como las que se muestran en la Figura 2.15. Debido al auge tecnológico de nuestros días, éstas se clasifican además según su público, por el sujeto principal, por su localización y por su acceso.

- **Según su público objetivo y temática**. Se clasifican en **horizontales** y **verticales**. Las horizontales son aquellas en la que hay una participación libre sin tema ni tipo de usuario en particular como Facebook y Twitter. Las verticales se originan sobre un eje temático y pueden ser: verticales profesionales

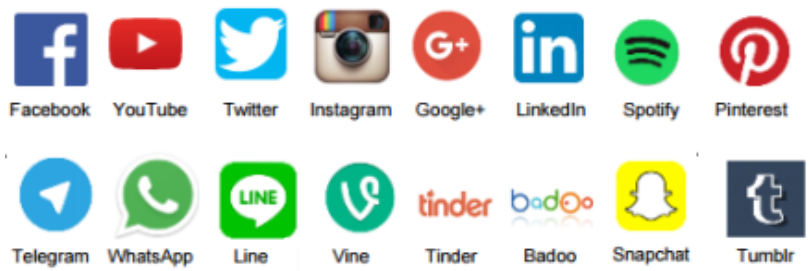



Figura 2.15: Logotipos de algunas redes sociales en línea de acceso público

como Linked In (orientada a empresas, negocios y empleo), verticales de ocio como Dogster o Spotify (enfocadas a mascotas o música) y verticales mixtas como Yuglo que es dirigida a artistas (desde un punto de vista profesional) pero pueden desarrollar actividades personales (de ocio).

- **Por el sujeto principal de la relación**. Pueden ser **humanas**, de **contenido** o de **objetos**. Las primeras fomentan relaciones entre personas según su perfil o intereses, por ejemplo, Dopplr permite interacciones entre viajeros frecuentes. Las segundas se forman con respecto al contenido publicado como Scribd, en donde se comparten documentos o Pinterest en donde se comparte contenido multimedia. Finalmente las inertes para unir objetos como marcas o como el caso de Colnect para artículos coleccionables.

- **Por su localización geográfica** pueden ser **sedentarias** o **nómadas**. Las sedentarias cambian en función del contenido como la mayoría de las redes de contenido, tal es el caso de Twitter. Mientras que las nómadas cambian en función de la ubicación geográfica como Foursquare que permite compartir tu ubicación y asistencia a lugares.

- **Por su acceso.** Tenemos redes sociales **públicas** en las que pude acceder cualquier persona como Facebook o **privadas** que tienen un acceso restringido como Sermo a la que sólo acceden médicos.

### 2.5.4 Redes según su modalidad

Las redes según su modalidad dependen de los tipos de entidades, tanto de los actores como de la relaciones.

- **Unimodales**. Consideran un único conjunto de actores, como Twitter, donde todos los usuarios son iguales y los tipos de relaciones sólo están relacionadas al seguimiento (seguir o ser seguido).
- **Bimodales**. Se centran en dos conjuntos de actores, como el mostrado en la Figura 2.2, donde tenemos por un lado a mujeres y por otro a eventos; otro ejemplo para las redes sociales en línea sería *airbnb* que relaciona a anfitriones con huéspedes.

### 2.5.5 Redes según su topología

Las redes según su topología dependen de la complejidad de la red y hay dos tipos:

- **Simples**. Son estructuras sencillas y se acoplan fácilmente a la teoría de grafos.
- **Complejas**. Presentan propiedades no triviales como el efecto del mundo pequeño, estructuras jerárquicas y estructuras comunitarias.

## 2.6 Análisis de Redes Sociales (SNA)

El Análisis de Redes Sociales (SNA por sus siglas en inglés de *Social Network Analysis*) [114] es el estudio de relaciones entre individuos incluyendo el análisis de estructuras sociales, posiciones sociales, roles, entre otras. SNA es interdisciplinario, incluye elementos de pscicología social, sociología, estadística y minería de grafos.

Los últimos dos son elementos de la minería de datos. SNA involucra una variedad de tareas [106]:

- **Análisis de centralidad.** Identifica los actores más importantes en la red. La centralidad se utiliza para determinar esta importancia a través de la influencia social. Por ejemplo, si se analiza una red de llamadas telefónicas entre delincuentes se podría obtener al líder o a alguien muy allegado a él, de dicha organización basado en el análisis de centralidad.

- **Detección de comunidades.** Los elementos en las redes sociales forman grupos, principalmente se buscan en la topología de la red, aunque también se pueden incluir características de los nodos. En el siguiente capítulo se detallará más al respecto. La detección de comunidades ha sido usada para campañas electorales, para inserción de productos en *marketing*, o incluso para estimar indicadores socio-económicos en entidades.

- **Análisis de roles.** Identifica la posición o rol asociado con diferentes actores mediante interacciones. Por ejemplo, el estudio de genes (ver Figura 2.16) a través de redes permitió analizar el rol de éstos en las enfermedades celulares, de tal forma que se encontró que los genes del cáncer desempeñan papeles críticos en el desarrollo y crecimiento celular.

- **Modelado de redes.** Simula las redes sociales reales a través de mecanismos simples. Por ejemplo, los sistemas de comunicación que van desde el uso de algún transporte público hasta mensajería instantánea.

- **Difusión de información.** Estudia la forma en que se propaga la información en la red, ésta ha sido útil en el estudio de esparcimiento de enfermedades como la mostrada en la Figura 2.14.

- **Clasificación.** Algunos actores son etiquetados con cierta información, esto

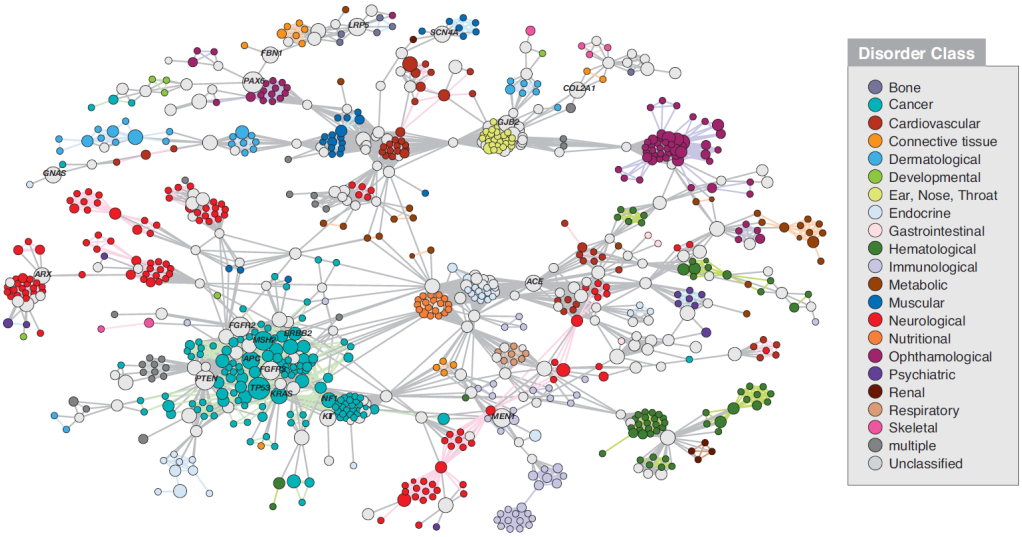



Figura 2.16: Red de genes, en donde los genes se conectan si ambos se presentan en un mismo desorden o enfermedad [41]

permite deducir que personas podrían compartir esa información, como para detectar tendencia en elecciones de productos o de políticos.

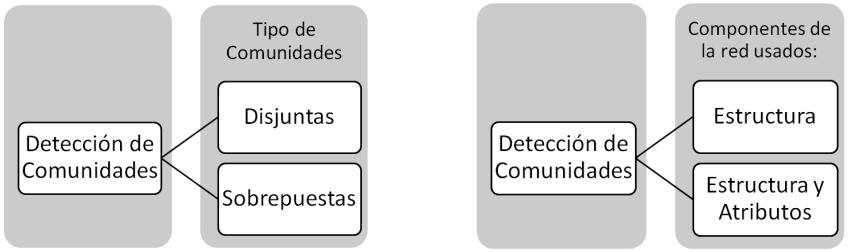
- **Predicción de enlaces (*links*).** El modelado de información, el proceso de difusión, el análisis de centralidad y la detección de comunidades en conjunto ayudan al *marketing* viral. A través de ésta se podrían dar recomendaciones, ya sea para la adquisición de nuevos productos o para establecer nuevas colaboraciones en trabajos científicos.

# Capítulo 3

# Detección de comunidades

Las redes sociales son una abstracción teórica usada en las ciencias sociales para estudiar las relaciones entre individuos, grupos u organizaciones. Las redes sociales en el mundo se dividen de manera natural en pequeñas comunidades. De manera general, una comunidad en redes sociales es un conjunto de nodos que están densamente conectados en el interior y muy poco conectados con el resto de la red. Sin embargo, no existe una definición generalizada de comunidad. Existen muchos algoritmos orientados a la detección de comunidades, muchos de ellos han sido propuestos recientemente. Abordaremos la detección de comunidades desde dos perspectivas (Figura 3.1):

La primera se refiere al tipo de comunidades detectadas según el número de comunidades a las que pertenezca un nodo, por lo que pueden ser disjuntas o sobrepuestas. En las comunidades disjuntas los actores sólo pertenecen a una comunidad; en la Sección 3.1 abordamos los algoritmos clásicos en este rubro enfocados a redes con estructura. Mientras que en las comunidades sobrepuestas los actores pueden pertenecer a más de una comunidad y serán detallados en la Sección 3.2, donde se muestra la clasificación para estos algoritmos enfocados desde la estructura del grafo.




(a) Detección de comunidades según el tipo de comunidades

(b) Detección de comunidades según los datos de la red usados

Figura 3.1: Detección de comunidades abordadas en el presente capítulo

La segunda perspectiva desde la que se abordará el estado del arte de detección de comunidades tiene origen en el hecho de que el uso de atributos en redes sociales puede detectar mejores comunidades que si sólo se considera el sociograma. Existen modificaciones a los algoritmos clásicos de agrupación en grafos [93][7][67], sin embargo existen pocos métodos que combinan ambas fuentes de información. Por lo que en la Sección 3.3 se abordarán algunos algoritmos para este fin. Éstos se clasifican en métodos basados en distancia o basados en modelo, por lo que la investigación se muestra desde esta perspectiva.

## 3.1 Detección de comunidades disjuntas

En la detección de comunidades se espera una alta frecuencia dentro de la comunidad y una cercanía entre los nodos. Las comunidades $C$ son un conjunto de $q$ agrupaciones de nodos $C = \{c_1, c_2, ..., c_q\}$ tales que $c_1 \bigcup c_2 \bigcup ... \bigcup c_q = N$. Existen algoritmos clásicos y nuevas propuestas que hemos clasificado en: jerárquicos, particionales, modulares, espectrales y dinámicos.

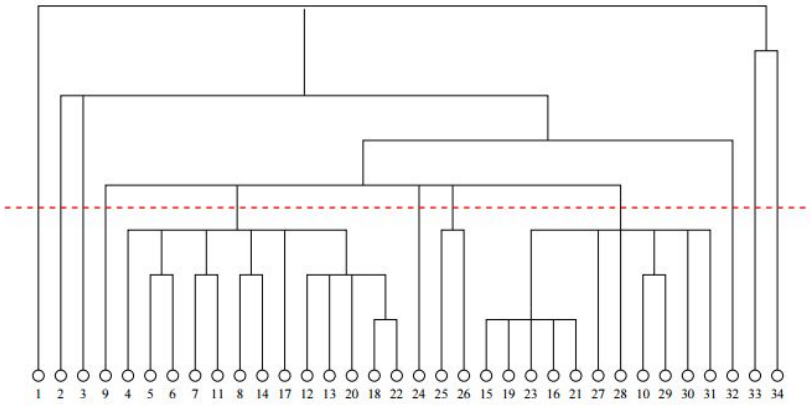



Figura 3.2: Dendograma del club de karate de Zachary [77]

### 3.1.1 Métodos jerárquicos

La clasificación con jerarquías genera un dendograma o árbol de agrupaciones (ver Figura 3.2), lo que permite la exploración a diferentes niveles de granularidad. Tiene la ventaja que se pueden aplicar medidas de similitud o distancias por lo que puede aplicar a estructura o atributos, pero se debe definir la métrica para determinar la similitud entre dos *clusters*. Existen dos tipos: aglomerativos y divisivos. Los **aglomerativos** empiezan con un nodo y recursivamente mezclan dos o más para obtener la agrupación más apropiada dado un criterio. La mayoría usan medidas clásicas de similitud: el algoritmo *hclust* usa la disimilitud Lance–Williams, **AGNES** [54] usa con coeficiente aglomerativo que mide la agrupación de la estructura y que destacó por la representación gráfica y **ROCK** emplea una medida que involucra el número de enlaces entre un para de puntos para cada *cluster* (toma como base la modularidad).

Por otro lado, los **divisivos** inicialmente se tiene un *cluster* que contiene todos los puntos y recursivamente se separa según sea más apropiado. Se usan coeficientes de agrupamiento, *cliques* o intermediación.

### 3.1.2 Métodos particionales

Se define el número de comunidades y una distancia entre los nodos. El objetivo es separar los nodos en las comunidades tal que se maximice o minimice una función de coste basada en la distancia. Pueden ser probabilísticos o de optimización iterativa.

En los **probabilísticos** se asume que los datos provienen de una mezcla de grupos cuya distribución se quiere encontrar. Uno de los métodos más sobresalientes en este aspecto es EM (*Expectation-Maximization*) que estima la probabilidad de un nodo en una comunidad y después se busca una aproximación al modelo de mezcla (*mixture model*).

En los de **optimización iterativa**, las funciones que más se suelen ocupar son las de *minimun k-cluster*, *k-cluster sum*, *k-center* y *k-median*, destacando *k-means* y k-*menoids*. Una de las más estudiadas es ***k-means*** [67] en el que cada iteración se estima el centroide. Se han propuesto múltiples extensiones para este algoritmo, las cuales han probado ser efectivas pero no es la óptima. Otro de los más utilizados es ***k-menoids*** que también ha sido mejorado, por ejemplo, **CLARA** [54] (*Clustering LARge Applications*) compara algunas vecindades, mientras que **CLARANS** (*Clustering Large Applications based upon RANdomized Search*) usa búsqueda aleatoria para generar vecinos y evaluar su partición.

Dentro de los algoritmos paticionales tenemos algunos que particionan el grafo buscando el menor número de cortes (*cut*).

### 3.1.3 Métodos modulares

La modularidad $Q$ dada por **Girvan-Newman** [76] es una medida que califica las comunidades, la cual se ha vuelto esencial en muchos métodos de agrupamiento,

es una de las medidas más usadas y se calcula:

$$Q = \Sigma_i(e_{ii} - a_i^2) \tag{3.1}$$

donde $i$ es el número de comunidades, $e_{ii}$ es el número de aristas dentro de las comunidades y $a_i$ es la fracción de aristas que tienen al menos uno de los nodos dentro de la comunidad, como ya se había explicado. Encontrar la máxima modularidad es un problema NP duro, por lo que existen múltiples algoritmos de optimización al respecto, destacando el algoritmo de Girvan-Newman [40] que es un algoritmo divisivo que utiliza la medida de *betweenness*. Posteriormente se propuso **OSLOM** [62] maximizando la modularidad, pero con un enfoque a redes sociales.

### Algoritmo de Girvan-Newman

El algoritmo involucra el cálculo de intermediación (*betweenness*) de todas las aristas de la red y remueve aquella con el valor más alto, repitiendo este proceso iterativamente. La centralidad de intermediación (*edge betweenness centrality*) se define como el número de caminos más cortos que pasan por una arista para conectar cualesquiera dos nodos. La eliminación de las aristas en los valores más altos permite la detección de comunidades.

Los pasos en el algoritmo de **Girvan-Newman** son:

1. Calcular la intermediación de todas las aristas existentes en la red.
2. La arista con la más alta intermediación es removida.
3. La intermediación de todas las aristas afectadas por la eliminación se recalculan.
4. Los pasos 2 y 3 se repiten hasta que no queden aristas.

Este algoritmo permitió la separación casi perfecta de la red de karate de Zachary que se muestra en la Figura 3.3, donde sólo el nodo 3 fue mal clasificado.

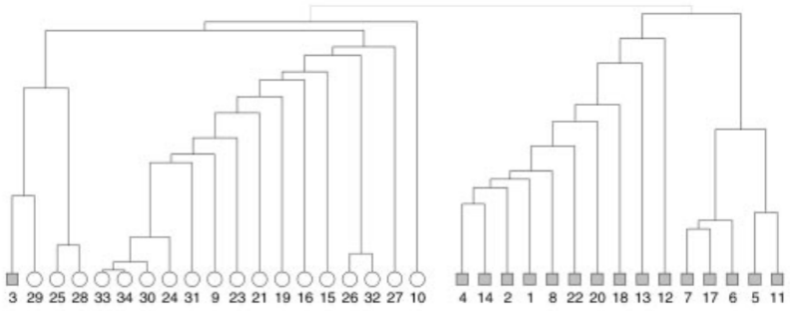



Figura 3.3: Árbol jerárquico del red de karate de Zachary usando modularidad [40]

Aunado a la modularidad tenemos los algoritmos basados en la densidad, los cuales pueden encontrar agrupaciones con figuras arbitrarias y suelen tener buena escalabilidad. Algunos se basan en la densidad de la conectividad como **DBSCAN** [32], mientras que otros se basan en funciones de densidad.

### 3.1.4 Métodos espectrales

El agrupamiento espectral consiste en transformar el conjunto inicial de objetos en un conjunto de puntos en el espacio usando *eigenvectors* y matrices (típicamente Laplacianas) que luego son agrupados con alguna otra técnica estándar. Es decir, el espectro de *eigenvectors* de varias matrices gráficas consiste típicamente en una gran cantidad de *eigenvectors* estrechamente espaciados, más algunos periféricos separados del volumen por un espacio significativo. Los *eigenvectors* correspondientes a estos valores atípicos contienen información sobre la estructura a gran escala de la red, como la estructura de la comunidad. La agrupación espectral consiste en generar una proyección de los vértices del gráfico en un espacio métrico, utilizando las entradas de esos vectores propios como coordenadas. Las $i$-ésimas entradas de los *eigenvectors* son las coordenadas del vértice $i$ en un espacio euclidiano $k$-dimensional, donde $k$ es

el número de *eigenvectors* utilizados. Los puntos resultantes se pueden organizar en grupos mediante el uso de técnicas de agrupamiento de particiones estándares.

La idea principal es que la representación de baja dimensionalidad inducida por los *eigenvectors* muestra la estructura de los grupos del grafo original con mayor claridad; lo que permite disminuir el problema de cortes, aunque su gran desventaja es la complejidad computacional.

Uno de los primeros métodos fue propuestos por Donath y Hoffmann [28], mientras que el vector de Fiedler [36] es la base para grafos bipartitas. Dada la complejidad es éstos, algunas implementaciones recientes usan algoritmos iterativos como Lanczos [20]. Para este mismo problema de complejidad, Dhillon et al. [25] mostraron que algunas medias de corte como corte normalizado (*normalized cut*) que suelen optimizarse con agrupamiento espectral, también se pueden optimizar utilizando un algoritmo equivalente de kernel ponderado *k-means* (*weighted kernel k-means*) que dio origen a **Graclus** [59].

### 3.1.5 Métodos dinámicos

En esta categoría encontramos métodos que emplean procesos que se ejecutan sobre sel grafo. Lo enfocaremos a caminos aleatorios (*random walks*) ya que debido a la alta densidad se pueden seguir múltiples caminos. La distancia entre dos nodos [130] se define como el promedio de aristas que pasa el caminante entre esos dos nodos, los pares de vértices con caminos más cercanos tienden a estar asociados a la misma comunidad.

**Infomap** [89] optimiza la ecuación llamada *map equation* que mide el promedio de longitud de un código dado por pasos que describen el movimiento de un camino aleatorio (*random walker*) dentro del grafo. Infomap captura los patrones de flujo de primer y segundo orden dinámico usando memoria en los nodos y logra agrupar los

nodos por vecindarios.

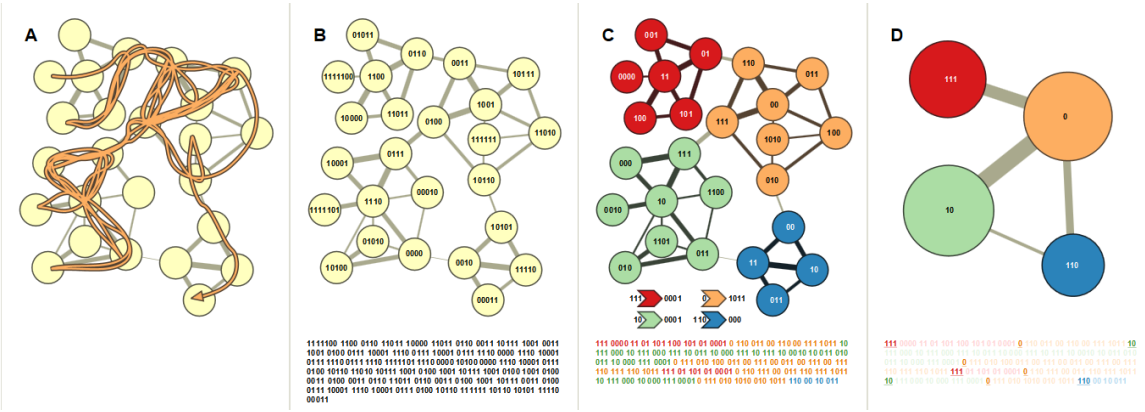



Figura 3.4: **A**. describe el camino de un camino aleatorio **B**. para obtener la descripción de un nivel se usaron el código de Huffman **C**. usan una descripción a dos niveles para reducir en un 32% el código, el libro de código (*codebook*) es usado como un índice para indicar los cambios (entrada y salida) entre bloques **D**. Se muestran los nombres de los módulos. [89]

### 3.1.6 Medidas de comunidades disjuntas

La detección de comunidades consta de dos etapas: primero, la detección de comunidades significativas, y segundo, la evaluación de las comunidades detectadas. Debido a que no existe una definición genérica de comunidades, cada una de las definiciones llega a justificarse en términos de la formulación de determinadas medidas.

Se han propuesto medidas que cuantifican propiedades para evaluar la calidad de la comunidad. Muchas de estas **medidas no sólo son usadas para evaluación sino que también se usan para la detección de comunidades**, tal es el caso de la modularidad [76] que ya hemos estudiado en la Sección 3.1.3.

Las medidas para descubrir comunidades disjuntas se basan en conexiones ya sean internas, externas, combinadas o basada en modelo. Otra forma de medir las comunidades detectadas es comparándolas con comunidades del mundo real ya existentes

que usan como base la precisión y recuperación.

#### 3.1.6.1 Medidas de calidad de comunidades

Las medidas de calidad de las comunidades son la base de varias tareas como la predicción de enlaces, la recomendación y el agrupamiento. En el agrupamiento de redes, la similitud es definida basada en la estructura, donde medir la calidad de un grupo es una tarea no supervisada que depende de las aristas [6]. Como se ha mencionado, no existe una definición precisa de agrupación (*cluster*), por lo que existen una variedad de medidas. Las diferentes medidas buscan medir la similitud o calidad de un grupo basados en conexiones entre objetos, para ello consideran la densidad intra-grupos [51] y la desconexión inter-grupo [30].

Existen muchas medidas enfocadas a la evaluación de la calidad de las comunidades detectadas. Yang et. al. [121] y Chakraborty [15] las clasificaron en cuatro grupos: basadas en conexiones internas, basadas en conexiones externas, combinando conexiones internas y externas y basadas en modelo.

Dentro de las **basadas en conexiones internas** tenemos: densidad interna (relación entre posibles y existentes aristas), número de aristas internas, promedio de aristas internas, FOMD (fracción sobre el grado de mediana), TPR (*triangle participation ratio*), coeficiente de agrupación (proporción de triángulos con cierto nodo) y volumen (suma de grados de los nodos en una comunidad)

La ***densidad*** de una comunidad es la relación del número de aristas en la comunidad $C$ comparada con el número de posibles aristas como se define en la Ecuación 3.2.

$$D(c) = \frac{|(v_i, v_j) \in E : v_i, v_j \in c|}{|v_j : v_j \in c|(|v_j : v_j \in c| - 1)} \tag{3.2}$$

La densidad nos permite evaluar qué tan bien conectada está la comunidad; cabe

destacar que estamos buscando comunidades con alta densidad. Esta medida es la se usará para obtener la densidad de las comunidades detectadas en los experimentos de este documento, de tal forma que se obtendrá el promedio de densidad de todas las comunidades detectadas en $G$ tal que,

$$D(C) = \frac{1}{|C|} \sum_{c_i \in C} D(c_i) \tag{3.3}$$

Otra medida que considera las aristas internas es la ***covertura*** [11] que está dada por la fracción del peso de las aristas intra-grupo con respecto al peso total de las aristas en todo el grafo, como se observa en la Ecuación 3.4, donde $w(C) = \sum_{k=1}^{|C|} w(v_i, v_j); v_i, v_j \in c_k, (v_i, v_j) \in E$. En este caso, se considera la separación entre grupos y no la densidad dentro del grupo:

$$coverage(c) = \frac{w(C)}{w(G)} \tag{3.4}$$

En las medidas **basadas en conexiones externas** tenemos: expansión (aristas por nodo que tienen un nodo externo), *cut ratio* (fracción de aristas que salen del grupo) y cortes de aristas (*edges cut*, cuenta las aristas que serían removidas para desconectar un grupo).

Dentro de las medidas que **combinan conexiones externas e internas** están: conductividad, cortes normalizados (*normalized cut*), Máximo de ODF[1], promedio de ODF, Pequeño ODF (*Flake-ODF*), separabilidad (proporción entre aristas internas y externas) y cohesión (conductividad de cortes internos).

Se han propuesto varias medidas o funciones de calidad para capturar qué tan buena es la división en los grafos; dos de las más importantes dado que combinan

[1]ODF: *Out Degree Fraction*, es la fracción de aristas que salen de un nodo

aristas internas y externas son cortes normalizados y conductividad.

Los ***cortes normalizados*** [98] de un grupo de vértices $c$ se define como se muestra en la Ecuación 3.5 tal que $Ncut(c)$ es la suma de las aristas que conectan $c$ con el resto del grafo, normalizado por el total de aristas de $c$ y con el resto del grafo. Entre menor sea al valor se tendrá una mejor comunidad:

$$Ncut(c) = \frac{|\{(v_i, v_j) : v_i \in c, v_j \notin c\}|}{|\{(v_i, v_j) : v_i \in c, v_j \in V\}|} + \frac{|\{(v_i, v_j) : v_i \in c, v_j \notin c\}|}{|\{(v_i, v_j) : v_i \notin c, v_j \in V\}|} \tag{3.5}$$

La ***conductividad*** ($\phi$)[53] corresponde a la premisa de que una comunidad es un conjunto de nodos que está más conectado internamente que con el exterior. La Ecuación 3.6 muestra la conductividad de una comunidad $c$:

$$\phi(c) = \frac{|\{(v_i, v_j) : v_i \in c, v_j \notin c\}|}{min(|\{(v_i, v_j) : v_i \in c, v_j \in V\}|, |\{(v_i, v_j) : v_i \notin c, v_j \in V\}|)} \tag{3.6}$$

Finalmente, las medidas **basadas en el modelo** se enfocan principalmente a la modularidad [40], que es la diferencia de la fracción de enlaces internos de un grupo y la fracción esperada de aristas en un grafo aleatorio. La modularidad $M$ para un grupo $c$ está dada por la Ecuación 3.1 detallada en la Sección 3.1.3.

#### 3.1.6.2 Medias para comparar con comunidades reales

Para comparar con las comunidades reales usamos dos conceptos básicos: precisión y recuperación. La precisión son los valores positivos recuperados con respecto a los esperados, de tal forma que se define como la fracción de los elementos relevantes de los recuperados. La recuperación, también conocida como sensibilidad, es la fracción de elementos relevantes que han sido recuperados sobre el total de todos los elementos relevantes.

Basadas en estas se definen algunas otras métricas como pureza, F1 y Jaccard. Dadas las comunidades detectadas $C$ y la comunidades reales $C^*$, la **pureza** [101] de la partición obtenida se define como se ve en la Ecuación 3.7:

$$purity(C, C^*) = avg_{c_i \in C}\{max_{c_j^* \in C^*} \frac{|c_i \cap c_j^*|}{|c_i|}\} \tag{3.7}$$

donde $|c_i \cap c_j^*|/|c_i|$ es la taza de precisión y $|c_i \cap c_j^*|/|c_j^*|$ la de recuperación.

Una medida que combina la precisión y la recuperación, a través de la media armónica, son las medidas $F - measure$. Específicamente, la **medida** $F_1$ usada para la evaluación de las comunidades en este documento se define como se muestra en la Ecuación 3.8, la cual es la media armónica de la precisión (*precision*) y la recuperación (*recall*).

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{3.8}$$

La medida de Jaccard es la relación entre la intersección y la unión de los grupos reales y los detectados. De tal forma que se calcula como se muestra en la Ecuación 3.9:

$$M_{jaccard} = \frac{precision \cdot recall}{precision + recall - precision \cdot recall} \tag{3.9}$$

## 3.2 Detección de comunidades sobrepuestas

En las redes sociales un nodo pertenece a más de una comunidad, esto permite la detección de comunidades sobrepuestas, tal que se tiene una agrupación $C = \{c_1, c_2, ..., c_q\}$, para $q = |C|$ número de comunidades, que no son necesariamente disjuntas de tal forma que cada nodo $v_i$ se asocia a una comunidad $c_j$ con cierto factor de pertenencia $p_{v_i c_j}$, donde $0 \leq p_{v_i c_j} \leq 1$, $\forall v_i \in V, \forall c_j \in C$ tal que $\sum_{j=1}^{q} p_{v_i c_j} = 1$ tal que $C = \{c_1, c_2, ..., c_q\}$ donde $c_1 \cup ... \cup c_q \subseteq V$, que no son necesariamente *clusters*

disjuntos. La posibilidad de que un vértice pertenezca a más de una comunidad da origen a las comunidades sobrepuestas, esto incrementa el problema de detección por el incremento de posibilidades de pertenencia de un nodo a múltiples comunidades.

Palla et. al. [80] introdujeron el concepto de comunidades sobrepuestas usando el conceptos de *k-cliques*, de tal forma que un nodo pertenece a más de una comunidad generado las comunidades sobrepuestas. Recientemente este problema ha tenido mucho auge [40]. Se distinguen dos tipos de traslapes, el primero es con sobreposición de baja densidad mostrado en el centro de la Figura 3.5 y el segundo es la sobreposición de alta densidad que se muestra del lado derecho de la misma figura. Existen enfoques que combinan la partición de nodos y enlaces [52][91], otros que usan las propiedades de la redes [122][109] o que extienden el algoritmo de **Girvan-Newman** como **CONGA** [43] dividiendo a un nodo en copias. Xie et. al. [117] presenta un estudio amplio de comunidades sobrepuestas y los clasificó como se muestra a continuación:

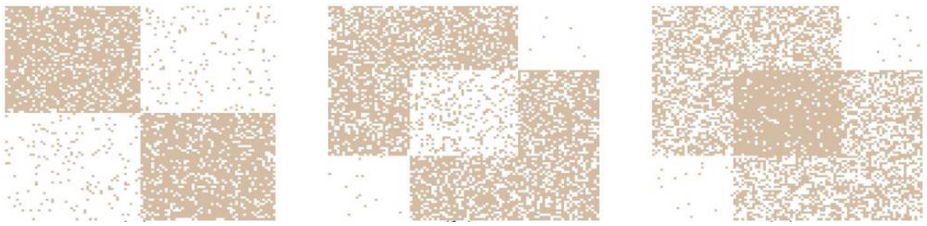


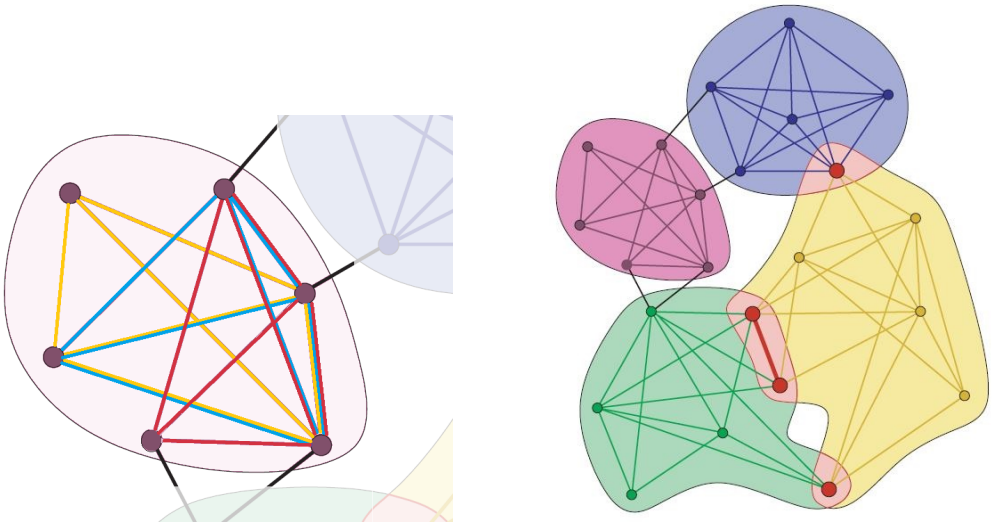
Figura 3.5: Del lado izquierdo se muestran comunidades disjuntas, al centro tenemos comunidades sobrepuestas pero cuya densidad en la intersección es menor que la propia de cada comunidad y del lado derecho tenemos una comunidad sobrepuesta con mayor densidad en el traslape [122]

### 3.2.1 Métodos de *clique percolation*

Detecta comunidades sobrepuestas en $k$ subgrafos completos ($k$-*cliques*), tal que $k$ corresponde al número de nodos; el traslape se presenta dada la combinación de dos o más *cliques* si comparten $k-1$ nodos. Entre ellos destaca **CPM** [80] cuya

implementación es **CFinder**, mejorado por **SCP** [60].

**CPM (*Clique Percolation Method*) [80]** fue uno de los iniciales abordando comunidades sobrepuestas y una de las más populares. Se basa en la premisa de que las aristas internas de una comunidad es probable que formen *cliques* debido a la alta densidad. El algoritmo encuentra todos los *cliques* de tamaño $k$, construye el grafo y dos *cliques* serán adjacentes si comparten $k-1$ vértices. Cada componente conectado es una comunidad en el grafo. Por ejemplo, en la Figura 3.6(a) podemos observar tres *cliques* de tamaño 4 marcados por las aristas de colores: rojo, azul y amarillo; como el rojo y el azul comparten $k-1$ nodos, es decir tres nodos, entonces se unirán en una comunidad; lo mismo sucede con los marcados en amarillo por lo que constituirán una comunidad, la cual se observa en color morado en la Figura 3.6(b).



(a) Unión de $k-1$ *cliques* para $k=4$ (b) Comunidades sobrepuestas

Figura 3.6: Ejemplo de comunidades sobrepuestas con *cliques* de k=4.

### 3.2.2 Métodos de particionamiento de enlaces

Dada una red social, supongamos que un nodo $v_a$ es amigo de un nodo $v_b$ y trabaja con el nodo $v_c$, de tal forma que podría estar en el traslape de los grupos de amigos en los que está el nodo $v_b$ y el grupo de compañeros de trabajo en el grupo $v_c$. Por lo tanto, ha surgido la idea de separar aristas para generar comunidades sobrepuestas, de tal forma que la arista $(v_a, v_b)$ pertenecerá a un grupo diferente a la arista $(v_a, v_c)$ generando un vértice en el traslape.

El particionamiento por enlaces es uno de los más intuitivos para la detección de comunidades sobrepuestas ya que los enlaces son particionados de manera disjunta pero los nodos, al tener varios enlaces, llegan a pertenecer a varias comunidades obteniendo comunidades sobrepuestas, incluso si solo se tiene un tipo de enlaces. De tal forma que se separan los enlaces en lugar de los nodos, permitiendo el uso de algoritmos de detección de comunidades disjuntas tradicionales como los jerárquicos o buscando similaridad entre aristas a través de la construcción de un grafo de línea (*line graph*) en donde los nodos son las aristas del grafo original [33][115].

Algunas otras propuestas extienden algoritmos como **Infomap** [57], modularidad [78] o *cliques* [34]. A pesar de la naturalidad de este concepto no se garantiza una detección de alta calidad que las basadas en nodos por la ambigüedad de la definición [37], además de que podrían existir aristas que conecten dos tipos diferentes de grupos que tendrían que se asignadas a dos comunidades o podría crearse una comunidad sólo de esos dos nodos.

Uno de las mayores deventajas de éstos métodos es la complejidad, dado que típicamente existen muchas más aristas que nodos y, al agruparse aristas en lugar de nodos, los cálculos se incrementan significativamente.

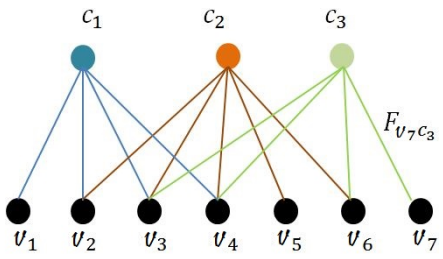



Figura 3.7: Pertenecia $F$ de que un nodo $v$ pertenezca a una comunidad $c$

### 3.2.3 Métodos de expansión local y optimización

Se basa en el crecimiento natural de las comunidades. La mayoría de ellos considera una función de beneficio local basada en alguna propiedad de la red como densidad. Destacaremos **RankRemoval** [4] que primero encuentra comunidades disjuntas y después expande las semillas iterativamente, usando una nueva medida de radio de intensidad que considera las aristas dentro de una comunidad con respecto a las externas, sin embargo depende de la calidad de las semillas. En **LFM** [61], la medida definida es similar, pero considerando un parámetro que controla el tamaño de las comunidades que posteriormente es mejorado con **OSLOM** [62], el cual prueba el significado estadístico de un *cluster*.

**EAGLE** [96] encuentra todos los *cliques* máximos y los va expandiendo con una medida de similitud que toma como base la modularidad, produciendo un dendograma. Aunado a estos, **BIGCLAM** [122] maximiza la probabilidad de que dos nodos estén conectados. BIGCLAM detecta comunidades significativas en grandes redes, con base en la premisa de que entre más comunidades compartan un par de nodos es más probable de exista un enlace entre ellos. Por ejemplo, en la Figura 3.7 podemos ver siete nodos y tres comunidades: el nodo $v_1$ tiene una baja probabilidad de estar conectado con $v_6$ porque no comparten ninguna comunidad; por otra parte, los nodos $v_3$ y $v_4$ tienen una alta probabilidad de tener una arista entre ellos porque pertenecen a las mismas tres comunidades. Adicionalmente cada nodo tiene difer-

entes niveles de pertenencia dadas por $F_{v_ic}$, por lo que la probabilidad de conexión entre los nodos $v_i \in N$ y $v_j \in N$ es:

$$P(v_i, v_j) = 1 - exp(-F_{v_i} \cdot F_{v_j}^T) \tag{3.10}$$

Para detectar las comunidades de la matriz $F$ se maximiza $l(F) = logP(G \mid F)$ de una matriz $F$ inicial dada por vecindades mínimas, por lo que la función a maximizar es:

$$l(F) = \Sigma_{(v_i,v_j)\in E} log(P(v_i, v_j)) - \Sigma_{(v_i,v_j)\notin E} log(1 - P(v_i, v_j)) \tag{3.11}$$

Para resolverlo, aplicaron la derivada de $l(F_{v_i})$ e iteraron.

### 3.2.4 Métodos de detección con *fuzzy*

Califica la asociación entre todos los pares de nodos, calculando un vector de pertenencia para cada nodo a las comunidades [42]. ***Fuzzy C-Means*** (**FCM**) [128] extiende el concepto de modularidad en un algoritmos basado en agrupación espectral.

Por otro lado, hay varios algoritmos mejorando **NMF** (*Nonnegative Matrix Factorization*), la cual es una técnica que ha sido adaptada para la detección de comunidades, aunque no tiene tan buenos resultados como otros algoritmos con este tipo de detección, tal es el caso de **MOSES** [69] que combina una optimización local con la asociación de un nodo a un vector latente.

### 3.2.5 Métodos basados en agentes y algoritmos dinámicos

Utilizan algoritmos de propagación de etiquetas. Los algoritmos con mejores resultados en esta categoría son **COPRA** [42] donde cada nodo actualiza su coeficiente de pertenencia y **Game** [17] donde cada comunidad es asociada con equilibrio mínimo,

se asocian funciones de ganancia y perdida con agente independiente.

***Speaker listener propagation algorithm* (SLPA)** [116] es una extensión de el Algoritmo de Propagación de Etiquetas (*Label Propagation Algorithm LPA*) donde los nodos seleccionan la etiqueta más popular de acuerdo a ciertas reglas, de tal forma que los nodos pueden ser un *speaker* o un *listener* para enviar o recibir etiquetas respectivamente. Los nodos tienen una memoria para guardar las actualizaciones, acumulando conocimiento de las diversas etiquetas que almacenan en su memoria. Dado que involucra un proceso estocástico, se tomó el resultado del mejor de cinco intentos.

### 3.2.6 Medidas de comunidades sobrepuestas

Existen algunas medidas para comunidades disjuntas (ver Sección 3.1.6) que son aplicables para comunidades sobrepuestas. Adicionalmente, existen medidas que evalúan los traslapes de comunidades, las cuales suelen ser extensiones de otras medidas adaptadas a comunidades sobrepuestas.

El **Índice Omega** es la versión para comunidades sobrepuestas del *Adjusted Rand Index* (ARI) [49]. Se basa en contabilizar los pares de nodos que concuerdan en do comunidades sobrepuestas y se define como se muestra en la Ecuación 3.12:

$$\Omega(C, C^*) = \frac{\omega_u(C, C^*) - \omega_e(C, C^*)}{1 - \omega_e(C, C^*)} \tag{3.12}$$

donde el índice omega ajustado $\omega_u$ se define como:

$$\omega_u(C, C^*) = \frac{2}{n(n-1)} \sum_{j=0}^{x} |t_j(C) \cap t_j(C^*)| \tag{3.13}$$

tal que $n = |V|$, $x = max(|v : v \in C|, |v : v \in C^*|)$ y $t_j(C)$ es el conjunto de pares

que aparecen exactamente $j$ veces $C$. El índice omega esperado $\omega_e$ está dado por:

$$\omega_e(C, C^*) = \frac{4}{n^2(n-1)^2} \sum_{j=0}^{x} |t_j(C)| \cdot |t_j(C^*)| \tag{3.14}$$

Otra de las medidas para comunidades sobrepuestas es $O\mathbf{C}$ [13] que se basa en un enfoque probabilístico intuitivo. Se define como la relación entre la probabilidad de encontrar dos elementos agrupados en ambas soluciones y la probabilidad máxima de encontrarlos en una de las soluciones dadas, como podemos ver en la Ecuación 3.15:

$$OC = \frac{\tilde{t}}{max(\tilde{p}, \tilde{p}^*)} \tag{3.15}$$

donde $\tilde{t}$ es la probabilidad de encontrar un par de vértices $(v_a, v_b)$, la cual es estimada como:

$$\tilde{t} = \frac{\sum_{i=1}^{q} \sum_{j=1}^{q^*} \left( \binom{|c_i \cap c_j^*|}{2} \right)}{\binom{n}{2} \frac{1}{n} min(q, q^*)} \tag{3.16}$$

y $\tilde{p}$ estima la probabilidad de encontrar un par de elementos en cualquier comunidad $c_i$ para todos las comunidades existentes, donde:

$$\tilde{p} = \frac{\sum_{i=1}^{q} \binom{|v_j : v_j \in c_i|}{2}}{q\binom{n}{2}} \tag{3.17}$$

En esta tesis se utilizan Omega Index y $O\mathbf{C}$ para evaluar las comunidades obtenidas.

## 3.3 Detección de comunidades considerando atributos

La mayoría de los algoritmos existentes emplean la estructura del grafo para encontrar comunidades, sin embargo agregar los atributos ayuda a mejorar la calidad

del *cluster*. En la Tabla 3.1 se describen brevemente algunos algoritmos que integran atributos. En esta se distinguen cuáles algoritmos, de los que integran atributos, detectan comunidades disjuntas, cuáles comunidades jerárquicas y cuáles comunidades sobrepuestas. En este último distinguimos dos tipos de sobreposición, las comunidades de baja densidad y las comunidades de alta densidad explicadas en la Figura 3.5.

Los algoritmos de detección de comunidades se pueden separar en [131]: basados en distancia (Sección 3.3.2) y basados en modelo (Sección 3.3.1). Dado que usamos atributos, adicionalmente, se tienen medidas que permiten ponderar los nodos y los atributos para extraer los más relevantes y así reducir la complejidad por lo que un estudio de éstos se muestra en la Sección 3.3.4.

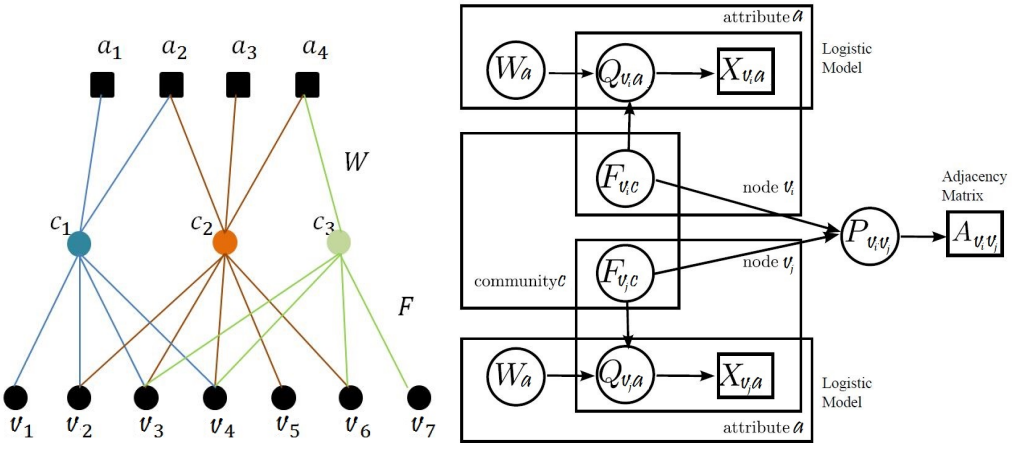
### 3.3.1 Métodos basados en modelo

Los métodos basados en modelo evitan el diseño artificial de una medida de distancia. Se basan principalmente en modelos probabilísticos que mezclan la información estructural y la de los atributos bajo el principio de que los vértices del mismo grupo se comportan de manera similar mientras que los de diferentes grupos se comportan diferente. Por lo general, dicho modelo es usado para definir un problema de inferencia probabilística [119] [66] [65].

De los métodos basados en modelo y para comunidades sobrepuestas detallamos tres de los principales algoritmos: **GenClus** [102], **GBAGC** [119] y **CESNA** [122]. Los dos primeros permiten un traslape entre comunidades suave, mientras que CESNA fue el primer algoritmo en la detección de comunidades con un traslape fuerte que lo logra con un modelado basado en observaciones de grupos en redes sociales reales, por lo que las comunidades detectadas tienen una alta precisión, aunque descuida aspectos como la cohesión y similitud interna.

| Algoritmo | Comunidad | Descripción |
|---|---|---|
| CoPaM's [71] | Jeráquico | Medida *cohesive patern* que mezcla atributos, usa elementos como candidatos y los expande mezclandolos. |
| SA-Cluster [132] | Disjuntas | Usa función que evalúa la influencia según la distancia dada por *random walk*, inicializa centroides y va seleccionando los mejores |
| PICS [2] | Disjuntas | Minimiza una función de costo de compresión basada en el número total de bits en una transmisión dentro de un mapeo. |
| GenClus [102] | Sobrepuestas Baja Densidad | Modelo probabilístico que agrupa objetos diferentes en un espacio oculto para aprender, de forma iterativa, la fuerza de los enlaces entre nodos. |
| Codicil [91] | Disjuntas | Introducen una medida de fuerza de conexión entre dos nodos que mezcla la fuerza del enlace con la similitud de contenido. |
| EDCAR [46] | Jeráquico | Evalúa el grado de similitud basada en densidad de *cliques* para crear un *set enumeration tree* que genera una cola de prioridades. (Basado en GAMer [44]) |
| CESNA [122] | Sobrepuestas Alta Densidad | Usa modelo probabilístico para determinar si existe un enlace entre dos nodos y entre nodos y atributos. Para detectar las comunidades maximiza las probabilidades. |
| DB-CSC [8] | Sobrepuestas Baja Densidad | Usa una densidad basada en similitud de atributos en subespacios y densidad local en el grafo. Usa $k$-vecindades para crear un grafo enriquecido. |
| GBAGC [119] | Sobrepuestas Baja Densidad | Esquema bayesiano que etiqueta aspectos de atributos y estructurales en un modelo probabilístico de distribución sobre el espacio de todas las posibles particiones. |
| SENC [88] | Disjuntas | Parte semillas definidas por *k-clique*, luego maximiza la probabilidad con EM. Mezcla conductividad con un vector de atributos. |
| *Complete Graph Model* [103] | Disjuntas | Mide el interior y exterior de la comunidad con una medida y la maximiza la diferencia. |

Tabla 3.1: Algoritmos de detección de comunidades en redes sociales con atributos.




(a) Modelo de CESNA (b) Representación de algoritmo de CESNA

Figura 3.8: Algoritmos de CESNA. (a) modela la pertenencia de un nodo a una comunidad, dado por $F$, y la pertenencia de un atributo a una comunidad dado por $W$. (b) Representación de CESNA donde $X_{v_i a}$ es el $k$ésimo atributo del nodo $v_i$, $W_a$ el vector de pesos para el atributo $a$, $Q_{v_i a}$ es la probabilidad de que el nodo $v_i$ tenga el atributo $a$, $F_{v_i c}$ es la pertenecia del nodo $v_i$ a la comunidad $c$ y $P_{v_i v_j}$ es la probabilidad de que exista un enlace entre $v_i$ y $v_j$ [122]

**CESNA** es un modelo que extendió **Bigclam**, agregándole la probabilidad de que un nodo $v_i$ tenga un atributo $a$. Como se puede observar en la Figura 3.8(a) comparada con la Figura 3.7, CESNA amplía el modelo de Bigclam con los atributos. En la Figura 3.8(a) la probabilidad de que el nodo $v_2$ tenga el atributo $a_4$ es mucho menor a que tenga el atributo $a_2$ por el número de comunidades que comparten. Esta probabilidad está dada por $Q_{v_i a}$ considerando el factor de peso $W_{ac}$ para cada atributo $a$ en la comunidad $c$. Por lo que $Q_{v_i a}$ está dada por:

$$Q_{v_i a} = \frac{1}{1 + exp(-\sum_c W_{ac} F_{v_i c})} \tag{3.18}$$

Para obtener las comunidades se debe maximizar la probabilidad de $F$, como es realizado en Bigclam, así como la probabilidad $W$, por lo que la función a maximizar

está dada por:

$$L(F,W) = logP(G,X|F,W) = l(F) + l(W) = logP(G \mid F) + logP(X|F,W) \quad (3.19)$$

como se explicó previamente $l(F)$ se obtiene derivando y actualizando $F_{v_ic}$; usando el mismo procedimiento (ver Ecuación 3.11) se obtiene $l(W)$ con la derivada y la actualización de $W_a$ usando *backtracking*:

$$W_{ac}^{new} = W_{ac}^{old} + \alpha(\Sigma_{v_i}\frac{\delta logP(X_{v_ia}|F,W_a)}{\delta W_{ac}} - \lambda \cdot Sin(W_{ac})) \quad (3.20)$$

Ellos realizaron experimentos con varios valores para los hyperparámetros $\alpha$ y $\lambda$ en un rango de $\alpha \in \{25; 0:5; 0:75\}$, $\lambda \in \{0.1, 1.0\}$ basado en los datos de probabilidad. Para su evaluación usaron $F_1$-score y similaridad Jaccard.

### 3.3.2 Métodos basados en distancia

La detección de comunidades con base en distancia utiliza medidas de similitud definidas para los elementos [93][81][24] los cuales se enfocan mayormente en la estructura de enlaces del grafo. Cualquier medida para grafos con pesos en aristas podría ser usada para redes con atributos, si los pesos representaran las conexiones entre atributos, de tal forma que se tenga un valor alto en la arista si comparten los mismos atributos, aumentando la probabilidad de pertenecer a la misma comunidad [74]. Por otro lado, existen métodos que definen una distancia con la información estructural para usar algoritmos de agrupamiento tradicionales [19] [131].

La selección de la medida de similitud adecuada depende de la tarea que se vaya a realizar [93]. Para la detección de comunidades en redes sociales con atributos la distancia debería incorporar el contenido de la información can base en la estructura. Neville et. al. [74] usan un **coeficiente de emparejamiento** (*matching coefficient*)

que busca el número de atributos que tienen en común dos nodos, dado por:

$$S_{v_i v_j} = \begin{cases} \sum_a \delta_a(v_i, v_j) & \text{if } (v_i, v_j) \in E \text{ or } (v_j, v_i) \in E, \\ 0 & \text{otherwise} \end{cases} \tag{3.21}$$

donde $\delta_a(v_i, v_j)$ es 1 si $\delta \in \Delta(v_i)$ y $\delta \in \Delta(v_j)$ para un grafo $G'$. Esta medida cuenta el número de atributos en común, dándoles a todos la misma importancia.

Cruz et. al. [21] consideran que no todos los atributos son relevantes, pero usa una distancia euclidiana, lo que hace que pierda habilidad al discriminar nodos e implica un problema de dimensionalidad; ellos usan la entropía para medir la similitud semántica en un subconjunto de atributos; esta discriminación de atributos la llevan a cabo dependiendo de un concepto que introducen como ***Point of View***.

Dichas distancias considera únicamente los nodos adjacentes. Algunos transforman los atributos a nodos usando un grafo aumentado, por ejemplo ***Neighborhood Random Walk Distance*** [131] mide la distancia entre dos vértices $v_i$ y $v_j$ con la distancia:

$$d(v_i, v_j) = \sum_{\tau: v_i \rightsquigarrow v_j}^{l} p_M(\tau) r (1-r)^{length(\tau)} \tag{3.22}$$

donde $P_M$ es la matriz de probabilidad de transición para un grafo aumentado, $\tau$ es el camino de $v_i$ a $v_j$ cuya longitud es $length(\tau)$ y $r \in (0, 1)$ es la probabilidad de reinicio. Esta es usada para calcular el número de caminos que conectan $v_i$ and $v_j$; entre más caminos existan más conectados están.

Algunos otros combinan la información estructural y la información de los atributos, de tal forma que la medida permita un balance entre atributos y estructura, como también ha sido considerado en este trabajo:

$$d(v_i, v_j) = \alpha \dot{d}_S(v_i, v_j) + \beta \dot{d}_A(v_i, v_j) \tag{3.23}$$

donde $d_S(v_i, v_j)$ es la **distancia estructural** y $d_A(v_i, v_j)$ es la **distancia de los atributos**. En este tipo de medidas lo complicado es la configuración de los parámetros. Estas medidas serán profundizadas en la Sección 3.3.3.

Para comunidades sobrepuestas con baja densidad se tiene **DB-CSC** [44] que detecta *clusters* basado en su densidad, debido a la complejidad observada al realizarlo en datos combinados, hacen uso de subespacios. Para ello primero define que un *cluster* combinado basado en su densidad en un grafo $G(V, E, l)$ con los parámetros $k$, $\epsilon$ y $minPts$ es un conjunto de vértices $O \subseteq V$ que tiene las propiedades de: densidad local alta $\forall v \in =:| \ N^O(v) \ | \geq minPts$ y conectividad local $\forall u, v \in O : \exists w_1, ..., w_l \in O : w_1 = u \wedge w_l = v \wedge \forall i \in 1, ..., l-1 : w_i \in N^O(w_{i+1})$.

También definen subgrafos enriquecidos que parten de un conjunto de vértices $O \subseteq V$, un subespacio $S$ y el grafo original $G = (V, E, l)$; el subgrafo enriquecido $G_s^O = (V', E')$ es definido por $V' = O$ y $E' = \{(u, v) \mid v \in N_S^O(u) \wedge v \neq u\}$ usando una función de distancia. Dicha distancia considera la similitud de los atributos en subespacios y la densidad local del grafo. A partir de éste, los autores hicieron algunas mejores con respecto a la duplicidad de comunidades que existía generando **EDCAR** [46].

El algoritmo ordena en una cola de prioridades los subespacios, toma el primer elemento de la cola y revisa que sea un *cluster*, rectificando que no sea redundante con los existentes. Si no existe, lo agrega, pero si agrupa nodos que ya habían sido considerados se considera que intersecta y existen dos casos: que sea o no un subárbol de alguno existente. Si es un subconjunto (subárbol) lo descarta, en el otro caso remueve los candidatos, genera el grafo enriquecido, determina los puntos y calcula los grupos para aumentarlos a la cola, en caso de ser necesario. Con base en este algoritmo se han desarrollado varias mejoras, destacando **HSC** [8].

### 3.3.3 Medidas de comunidades con atributos

Son aquellas medidas que cuantifican propiedades para evaluar la calidad de la comunidad. Estas medidas no sólo son usadas para evaluación, sino que también se usan para la detección de comunidades.

En esta sección se presenta el estudio de algunas de estas medidas, las cuales las hemos dividido en medidas basadas en estructura, basadas en atributos o basadas en ambos, a estas últimas las llamaremos medidas mixtas.

#### 3.3.3.1 Medidas basadas en atributos

La **entropía** [58] es, originalmente, para evaluar qué tanta información se puede transmitir ente conjuntos de pares de nodos. Sin embargo, la definición de entropía, que es considerada a lo largo del documento, permite evaluar el contenido de las comunidades en un grafo. La entropía del grafo $G$ está dada por la Ecuación 3.24:

$$H(G) = \sum_{i=1}^{k} \left( \frac{1}{k} \sum_{j=1}^{q} \frac{|v : v \in c_j|}{n} H(a_i, c_j) \right) \tag{3.24}$$

para $n = |V|$, $k = |A|$ y $q = |C|$, tal que la entropía de un atributo $H(a_i, c_j)$ es la proporción de vértices en $c_j$ con el atributo $a_i$ dado por la Ecuación3.25:

$$H(a_i, c_j) = \sum_{j=1}^{q} p_{a_i c_j} log_2 p_{a_i c_j} \tag{3.25}$$

La entropía permite evaluar qué tan similares son los nodos dentro de una comunidad. El máximo valor para un atributo es $log|v : v \in c_x|$, que depende del tamaño de $c_x$. Para comparar los resultados, en esta tesis, se normalizaron los valores.

#### 3.3.3.2 Medidas mixtas: considerando atributos y estructura

Las medidas de similitud pueden ser clasificadas en: basadas en atributos o basadas en enlaces [97]. La primera considera similitud en los valores de los atributos como los $k$ vecinos más cercanos, distancias euclidianas, distancia por coseno o coeficiente de Jaccard. La segunda se basa en los enlaces en un grafo como *PageRank* que evalúa la probabilidad de llegar, de un punto a otro, por caminos aleatorios. Otro ejemplo es modularidad que es una de las medidas más usadas, sin embargo no busca pequeños grupos ni considera atributos.

La calidad de un grupo, en redes sociales con atributos, debería **balancear la estructura y los atributos** [130][19][75][23] tal que se tenga una medida que defina la distancia entre $v_i$ y $v_j$ como:

$$d(v_i, v_j) = \alpha \cdot d_S(v_i, v_j) + \gamma \cdot d_A(v_i, v_j) \tag{3.26}$$

donde $d_S(v_i, v_j)$ y $d_A(v_i, v_j)$ sean medidas estructurales y de atributos respectivamente para los factores de peso $\alpha$ y $\gamma$, sin embargo definir estas distancia y parámetros aún representa un reto. Por ejemplo, una medida simple de peso dada por Neville et al. [75] define:

$$d(v_i, v_j) = \alpha \cdot \frac{1}{a} \sum_{a \in A} s_a(v_i, v_j) + (1 - \alpha) \cdot l \tag{3.27}$$

donde $s_a(v_i, v_j) = 1$ si $v_i$ y $v_j$ tienen al atributo $a$, mientras que $l = 0$ si existe una arista entre $v_i$ y $v_j$. Esta medida tiene una distancia lineal y sólo considera el vecino más próximo lo que no resulta en alta dimensionalidad y en redes grandes.

Gunnemann et. al. [45] maximizan la calidad de un grupo en el espacio de los atributos, considerando sub-gafos densos basado en k-vecinos con distancias lineales, sin embargo sólo considera pesos locales a cada atributo, así como densidades locales. **CoPaMs** [71] también considera un umbral de densidad, pero agrega un umbral para

las dimensiones dado por la presencia, o no, de los atributos, donde todos ellos tienen la misma importancia. Cruz et. al. [21] usan una medida con base en la entropía con distancia euclidiana. Sin embargo, cuando el número de atributos aumenta, todas las distancias tienden a converger, por lo que considera que no todos los atributos podrían ser importantes y usan la entropía para medir la similitud semántica de los atributos seleccionados, según su definición llamada ***Point of View***.

En algunos métodos de detección de comunidades sobrepuestas que consideran atributos, como el de Yang, et. al. [122], se otorga un peso a cada atributo en cada comunidad, dando una importancia local. Es este caso, falta una importancia global, ya que afectaría a su modelo, dado que la cantidad de atributos es mucho menor que el número de nodos. El indicador propuesto no sólo considera la entropía, sino también la importancia local y global del atributo. Además, se usa similitud coseno, descartando algunos problemas dados por la distancia euclidiana, como la dimensionalidad.

Otros métodos como **SENC** [88], además de enfocarse en comunidades sobrepuestas, buscan propiedades estructurales como la densidad y el diámetro, evaluando qué atributos son importantes para cada comunidad, sin embargo estos atributos no son consideraros desde el principio del proceso.

### 3.3.4 Ponderación de atributos

El etiquetado de importancia en redes sociales puede estar orientado a vértices o atributos. El algoritmo más importante en el *ranking* de vértices es PageRank ($PR$) [82] que evalúa la importancia de una página $A$ en Internet para las $i$ páginas que enlazan con $A$. El *ranking* de nodos ha sido usado en muchas tareas, Muller y Sanchez [72] [92] han desarrollado algoritmos orientados al grado de desviación, según las propiedades del grafo y los atributos. También se han buscado patrones correla-

cionados usando medidas estadísticas [99] y algunos otros trabajos se han enfocado en la importancia según las búsquedas [107] [90]. Cabe destacar que estos *rankings* están enfocados a nodos.

Para el *ranking* de atributos podemos asociar la topología del grafo con los atributos para obtener la importancia de los atributos con el objeto de reducir la complejidad y mejorar el proceso de clasificación. Uno de los más rápidos utilizado en minería de textos es el de Ganancia de Información (*Information Gain*) basado en entropía; otro es Componentes Principales (*Principal Component*) basado en técnicas estadísticas y CFS (*Correlation-based Feature Selection*) que fue el primero en evaluar conjunto de atributos. Los anteriores se enfocan en el conjunto de atributos, sin tomar en cuenta la estructura del grafo. Uno de los algoritmos que obtiene los atributos más importantes, usando la estructura, es *Global Weighting* propuesto para el algoritmo de LINKREC [125]. Otra forma de obtener la importancia de los atributos es por un cálculo simultáneo al proceso de comunidades como lo hace SA-Cluster. Estos cuatro se detallan a continuación:

**Ganancia de información**. Se basa en la entropía $H(C)$, en un conjunto de comunidades $C$, la entropía está dada por la Ecuación 3.24. Entre mayor sea la entropía, mayor es la impureza del conjunto de datos, de tal forma que si la entropía es cero, todos los nodos pertenecen al mismo grupo. Dado que queremos separar grupos, la mejor entropía es la más alta, ya que nos permitirá hacer clasificación porque no todos los nodos pertenecerán al mismo grupo. Para los atributos $A$ y las comunidades $C$, la entropía está dada por la Ecuación 3.25. La diferencia de entropía refleja la información adicional de una clase dado un atributo, a esto se le conoce como Ganancia de Información. A cada atributo $a$ se le asigna una calificación con base en la información ganada entre ella y la clase:

$$IF_i = H(C) - H(C \mid A_i) = H(A_i) - H(A \mid C) = H(a) + H(C) - H(a, C) \quad (3.28)$$

**Componentes principales.** Es utilizada para reducir la dimensionalidad de un conjunto de datos, además de hallar las causas de la variabilidad de un conjunto de datos y ordenarlas por importancia. Se calcula la covarianza de los atributos originales y se extraen sus *eigenvectors*, éstos son una transformación lineal de los atributos originales a un nuevo espacio en el que los atributos no están correlacionados. Los *eigenvectors* pueden tener un *ranking* de acuerdo a la variación con respecto a los datos originales.

***Global Weighting.*** El grafo aumentado mencionado en el Capítulo 2 fue mejorado por Yin. et al. dando pesos a los atributos. A este peso lo llaman Peso Global (*Global Weighting*) $g(a)$ para un atributo $a$ en $G$ tal que:

$$g(a) = \frac{\sum_{(v_i,v_j)\in E} e^a_{v_i v_j}}{\binom{n_a}{2}} \tag{3.29}$$

donde $n_a$ es el número de atributos que tienen el atributo $a$, $e^a_{v_i v_j} = 1$ si tanto el nodo $v_i$ como el nodo $v_j$ tienen el atributo $a$, y $e^a_{v_i v_j} = 0$ de otra forma. La importancia global de un atributo $a$ mide el porcentaje de que existan enlaces entre varios pares de nodos con $a_k$.

***Weigth Self-Adjutment*** definido en SA-Cluster como un peso a los atributos que se va actualizando en cada iteración $t$ del algoritmo.

$$w_a^{t+1} = \frac{1}{2}(w_a^t + \Delta w_a^t) \tag{3.30}$$

# Capítulo 4

# Uso de atributos en detección de comunidades

La integración de atributos al proceso de detección de comunidades ha dado pie a una gran cantidad de algoritmos que buscan sacar provecho de esta información disponible. Éstos se clasifican principalmente en algoritmos basados en modelo y los basados en distancia. En esta tesis, se propone un nuevo método basado en modelo, al cual llamaremos RMOCA por sus siglas en inglés, ***R****egression* ***Mo****del for* ***C****ommunity Detection in* ***A****ttributed Networks*.

RMOCA está basado en el principio de que dos nodos están en la misma comunidad si comparten aristas y, por otro lado, dos nodos con los mismos atributos pertenecen a la misma comunidad, dado que el atributo pertenece a la comunidad. El modelo RMOCA se basa en el modelo de regresiones, que es un proceso estadístico para estimar las relaciones entre variables. En el análisis de regresión es de interés caracterizar la variación de la variable dependiente en torno a la función de regresión, la cual es obtenida a partir del modelo que hace uso de mínimos cuadrados y que se describirá en este capítulo.

Las comunidades son estimadas haciendo uso de la minimización de dos errores obteniendo dos conjuntos de comunidades $C_S$ y $C_A$. La primera son las comunidades de nodos dadas por la estructura y las segundas son comunidades de atributos; éstas son calculadas simultáneamente estableciendo un vínculo entre ellas a partir de la relación entre las aristas (matriz de adyacencia) y las propiedades de ellos (matriz de atributos).

## 4.1 Representación matricial de la red social

Supongamos un grafo $G'(V, E, A)$ como el descrito en la Sección 2.2.1, tal que $V = \{v_1, v_2, ..., v_n\}$ representa el conjunto de $n$ nodos o vértices, $E \subseteq \{(v_i, v_j)|v_i, v_j \in V, v_i \neq v_j\}$ denotan el conjunto de $m$ aristas y $A = \{a_1, a_2, ..., a_k\}$ representa todos los posibles $k$ atributos, tales que $\Delta : V \longrightarrow A$ donde $\Delta_{v_i} \subseteq A$. Este grafo es representado con dos matrices: la matriz de adyacencia $M$ y la matriz de atributos $X$, como se observa en la Figura 4.1, de tal forma que cualquier red social es representada por un sociograma con atributos y es procesada a través de las matrices, que son explicadas a continuación.

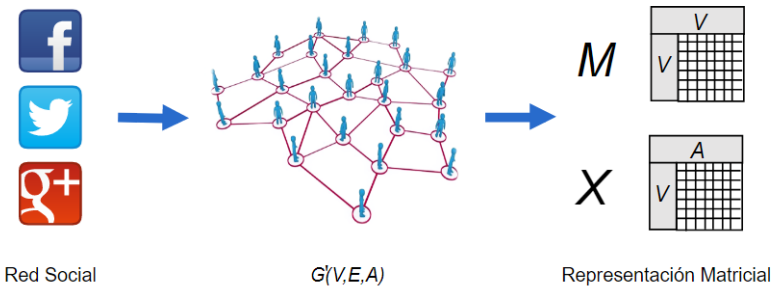



Figura 4.1: Representación matricial del grafo con atributos de redes sociales

### 4.1.1 Matriz de adyacencia

Para representar los grafos con atributos se pueden usar Listas de Adyacencia o Matrices de Adyacencia. En este caso se hace uso de la **Matriz de Adyacencia** $M$ que muestra las aristas, dada la relación de vértices.

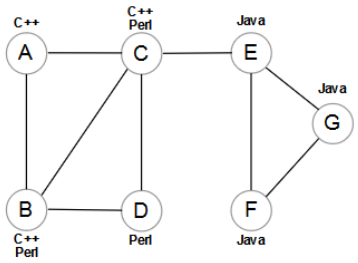
**Definición 4.1** ***Matriz de Adyacencia*** $M$. *Dado un grafo* $G'(V, E, A)$ *con un conjunto de vértices* $V$*, aristas* $E$ *y atributos* $A$ *la matriz de adyacencia* $M$ *contiene las aristas entre pares de vértices, tal que su tamaño está dado por* $n \times n$ *donde* $n = |V|$ *es el número de vértices en* $G$ *y cada elemento* $m_{ij}$ *de dicha matriz especifica el número de aristas del nodo* $i$ *al nodo* $j$*.*

Para el grafo de la Figura 4.2(a), se tienen 7 nodos representados por las primeras siete letras del abecedario, desplegados en las filas y columnas de la matriz de adyacencia $M$ de la Figura 4.2(b). La matriz de adyacencia resultante despliega las aristas existentes entre cualquier par de vértices; si existe una arista entre el nodo $i$ y $j$ entonces la posición $m_{ij} = 1$, tal que 1 representa la existencia de una arista y 0 la ausencia de ésta. En el caso de los grafos no dirigidos $m_{ij} = m_{ji}$, de tal forma que el número de unos en la matriz dividido por dos corresponderá al número de aristas en el grafo; por ello se tienen 9 aristas representadas por *unos* en la matriz $M$.

### 4.1.2 Matriz de atributos

Además de los vértices y aristas tenemos el conjunto de atributos que están relacionados con los vértices, para lo cual haremos uso de la **Matriz de Atributos** $X$.

**Definición 4.2** ***Matriz de Atributos*** $X$. *Dado un grafo* $G'(V, E, A)$ *con un conjunto de vértices* $V$*, aristas* $E$ *y atributos* $A$ *la matriz de atributos* $X$ *contiene la*




(a) Grafo con atributos

M

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| A | - | 1 | 1 | 0 | 0 | 0 | 0 |
| B | 1 | - | 1 | 1 | 0 | 0 | 0 |
| C | 1 | 1 | - | 1 | 1 | 0 | 0 |
| D | 0 | 1 | 1 | - | 0 | 0 | 0 |
| E | 0 | 0 | 1 | 0 | - | 1 | 1 |
| F | 0 | 0 | 0 | 0 | 1 | - | 1 |
| G | 0 | 0 | 0 | 0 | 1 | 1 | - |

(b) Matriz de Adyacencia $M$

X

| | C++ | Java | Perl |
|---|---|---|---|
| A | 1 | 0 | 0 |
| B | 1 | 0 | 1 |
| C | 1 | 0 | 1 |
| D | 0 | 0 | 1 |
| E | 0 | 1 | 0 |
| F | 0 | 1 | 0 |

(c) Matriz de Atributos $X$

Figura 4.2: Ejemplo de Representación Matricial del Grafo con Atributos de Redes Sociales

*relación de vértices con atributos, de tal forma que es de tamaño* $n \times k$ *donde* $n = |V|$ *es el número de vértices en* $G$ *y* $k = |A|$ *el número de posibles atributos en* $G$ *de tal forma que el elemento* $x_{ij}$ *es* 1 *si el nodo* $i$ *tiene el atributo* $j$*, en caso contrario es* 0.

Por ejemplo, en la Figura 4.2(a) se tiene un grafo con atributos, el cual tiene 7 nodos representados en las filas de la matriz $X$, 3 atributos representados en las columnas de la matriz $X$ y 8 relaciones nodo-atributo representadas por unos en la matriz $X$ de la Figura 4.2(c). Estas últimas indican que un nodo tiene ciertos atributos, por ejemplo el nodo $B$ tiene los atributos *C++* y *Perl* por lo que tiene un *uno* en esa posición de la matriz.

## 4.2 Comunidades basadas en estructura y atributos

La propiedad de separación por baja densidad (*Low Density Separation Assumption*) establece que los bordes que separan a las comunidades pasan por regiones de baja densidad, siendo éste el principio de detección de comunidades en las redes. Por lo que las comunidades $C = \{c_1, c_2, ..., c_q\}$, para $q = |C|$, son un grupo de vértices

tales que $c_i \subseteq V$. El modelo propuesto obtiene comunidades de estructura $C_S$ y comunidades de atributos $C_A$, las cuales serán descritas a continuación.

### 4.2.1 Comunidades basadas en estructura

Las comunidades de nodos basadas en la estructura de la red las llamaremos **comunidades de estructura** $C_S$, las cuales agrupan a los nodos según sus enlaces.

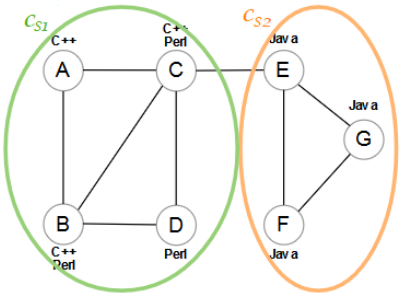
**Definición 4.3** ***Comunidades de Estructura*** $C_S$. *Dado un grafo* $G'(V, E, A)$ *con un conjunto de vértices* $V$*, aristas* $E$ *y atributos* $A$*, el conjunto de comunidades* $C_S = \{c_{S1}, c_{S2}, ..., c_{Sq}\}$ *está dado por* $q = |C|$ *número de comunidades* $c_S$ *que agrupan un conjunto de nodos, de tal forma que* $c_{Si} = \{v_1, v_2, ..., v_l\}$ *para* $v_1, v_2, ..., v_l \in V$ *y* $l \leq n$ *tal que* $n = |N|$.

En la Figura 4.3(a) se muestran dos comunidades de nodos denotadas por $c_{S1}$ y $c_{S2}$ las cuales agrupan los nodos según su estructura, ya que se observa que la separación de las dos comunidades sólo corta un enlace. De tal forma que $c_{S1} = \{A, B, C, D\}$ y $c_{S2} = \{E, F, G\}$. Para representar las comunidades usamos la matriz de estructura cuya definición se da a continuación.

**Definición 4.4** ***Matriz de Estructura*** $M_S$. *Dado el conjunto de comunidades* $C_S = \{c_{S1}, c_{S2}, ..., c_{Sq}\}$ *la matriz de estructura* $M_S$ *muestra la pertenencia de nodos* $V$ *a comunidades* $C_S$ *de tal forma que tiene un tamaño* $n \times q$ *donde* $n = |V|$ *es el número de nodos y* $q = |C|$ *el número de comunidades. Cada* $m_{S_{ij}}$ *representa la incidencia del nodo* $i$ *en la comunidad* $j$*, de tal forma que si el nodo* $i$ *pertenece a la comunidad* $j$ *entonces* $m_{S_{ij}} = 1$.

Por ejemplo, en la Figura 4.3(b) se muestra la matriz de estructura $M_S$ correspondiente al conjunto de comunidades de la Figura 4.3(a). Se observa que el nodo $B$

tiene un 1 en $c_{s1}$ indicando que pertenece a dicha comunidad; por otra parte el nodo $F$ tiene 0 en $c_{s1}$ por lo que no pertenece a esa comunidad, sin embargo tiene un uno en $c_{s2}$ para mostrar su pertenencia a esta comunidad.




(a) Comunidades de Estructura $C_S$

$\mathcal{M}_S$

| $\mathcal{V}$ \ $C_S$ | $c_{S1}$ | $c_{S2}$ |
|---|---|---|
| A | 1 | 0 |
| B | 1 | 0 |
| C | 1 | 0 |
| D | 1 | 0 |
| E | 0 | 1 |
| F | 0 | 1 |
| G | 0 | 1 |

(b) Matriz de Estructura $M_S$

Figura 4.3: Representación de Comunidades de Estructura $C_S$ y su Matriz de Estructura $M_S$

### 4.2.2 Comunidades basadas en atributos

Al conjunto de atributos agrupados según los enlaces de los nodos que los poseen lo llamaremos **comunidad de atributos** $C_A$.

**Definición 4.5** ***Comunidades de Atributos*** $C_A$. *Dado un grafo* $G'(V, E, A)$ *con un conjunto de vértices* $V$, *aristas* $E$ *y atributos* $A$, *el conjunto de comunidades* $C_A = \{c_{A1}, c_{A2}, ..., c_{Aq}\}$ *está dado por* $q = |C|$ *número de comunidades* $c_A$ *que agrupan un conjunto de atributos, de tal forma que* $c_{Ai} = \{a_1, a_2, ..., a_l\}$ *para* $a_1, a_2, ..., a_l \in V$ *y* $l \leq k$ *tal que* $k = |A|$.

En la Figura 4.4(a) se muestran dos comunidades de atributos denotadas por $c_{A1}$ y $c_{A2}$, las cuales agrupan atributos según la estructura de los nodos. Para visualizar esta agrupación de los atributos según la estructura, en la Figura 4.4(b) se muestra una representación de grafo en donde los nodos representan los atributos y para cada

arista $(i, j) \in G$ tal que $i, j \in V$ se tiene una arista que relaciona a los atributos de $i$ y $j$. Cabe destacar que esta transformación sólo se muestra para ejemplificar las comunidades de atributos, pero no es realizada en el cómputo del algoritmo propuesto.

**Definición 4.6** ***Matriz de Atributos*** $M_A$. *Dado el conjunto de comunidades de atributos* $C_A = \{c_{A1}, c_{A2}, ..., c_{Aq}\}$ *la matriz de atributos* $M_A$ *muestra la pertenencia de atributos* $A$ *a comunidades* $C_A$ *de tal forma que tiene un tamaño de* $k \times q$*, donde* $k = |A|$ *es el número de atributos y* $q = |C|$ *el número de comunidades. Cada* $m_{Aij}$ *representa la pertenencia de un atributo* $i$ *a una comunidad* $j$ *en la matriz* $M_A$.

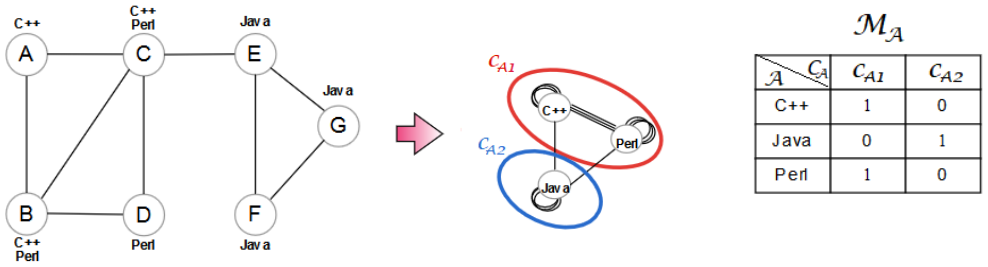
En la Figura 4.4(c) se observa la representación matricial de la pertenencia de los atributos $A$ a las comunidades $C$ en la matriz $M_A$, de tal forma que el atributo $C++$ pertenece a $c_{A1}$ al igual que el atributo $Perl$.

Es importante notar que $c_{Si}$ corresponde a $c_{Ai}$, por lo que se podría decir que $c_{Ai}$ contiene los atributos de la comunidad $c_{Si}$. Idealmente, éstos deberían corresponder totalmente, sin embargo en redes sociales reales difieren reducidamente; por ejemplo en la Figura 4.3 vemos que los nodos en $c_{S2}$ tienen el atributo $Java$ y la comunidad de atributos correspondiente sería $c_{A2}$ que contiene ese mismo atributo, sin embargo en $c_{A1}$ se tienen los atributos $C++$ y $Perl$ y existen nodos en $c_{S1}$ que no tienen alguno de estos atributos, como es el caso del nodo $A$ que no tiene el atributo $Perl$.

## 4.3 RMOCA: un nuevo método basado en modelo

Se propone un método de detección de comunidades basado en modelo. El modelo, llamado RMOCA, parte de dos premisas:

1. Los vértices que comparten un enlace tienden a estar en la misma comunidad.

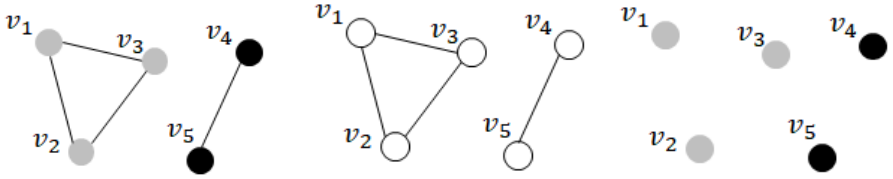



$\mathcal{M}_{\mathcal{A}}$

| $\mathcal{A}$ \ $C_A$ | $C_{A1}$ | $C_{A2}$ |
|---|---|---|
| C++ | 1 | 0 |
| Java | 0 | 1 |
| Perl | 1 | 0 |

(a) Grafo con atributos (b) Grafo de Comunidades de Atributos $C_A$ (c) Matriz de Atributos $M_A$

Figura 4.4: Ejemplo de representación de Comunidades de Atributos $C_A$




(a) Grafo con atributos (b) Estructura del grafo (c) Atributos del grafo

Figura 4.5: Desglose del grafo según su estructura y según sus atributos

2. Los atributos de los vértices son parte de la comunidad. Un vértice y sus atributos comparten la misma comunidad.

Considerando estas premisas, el modelo plantea obtener las comunidades $C_S$ y $C_A$ a partir de estimaciones para determinar una función objetivo que es optimizada; partiendo de que el grafo con atributos $G'(V, E, A)$ se puede ver desde su parte estructural (ver Figura 4.5(b)) o desde los atributos (ver Figura 4.5(c)) que lo componen.

### 4.3.1 Modelo de detección de comunidades basado en estructura

Se parte de la observación de que dos nodos, que comparten aristas, pertenecen a la misma comunidad. En la Figura 4.6(a) podemos ver la estructura de un grafo con 5 nodos y 4 aristas; supongamos que un nodo cualquiera $v_i$ pertenece a una comunidad $c_{S_j}$ como es el caso de del nodo $v_1$ a la comunidad $c_{S_1}$ en la Figura 4.6(b), $v_2$ podría pertenecer a cualquier comunidad, sin embargo tiene un enlace a $v_1$ que ya pertenece a $c_{S_1}$, por lo que se agrega a esa comunidad, como se ve en la Figura 4.6(c). Lo mismo sucede con el resto de los nodos, de tal forma que las comunidades resultantes se muestran en la Figure 4.6(d).

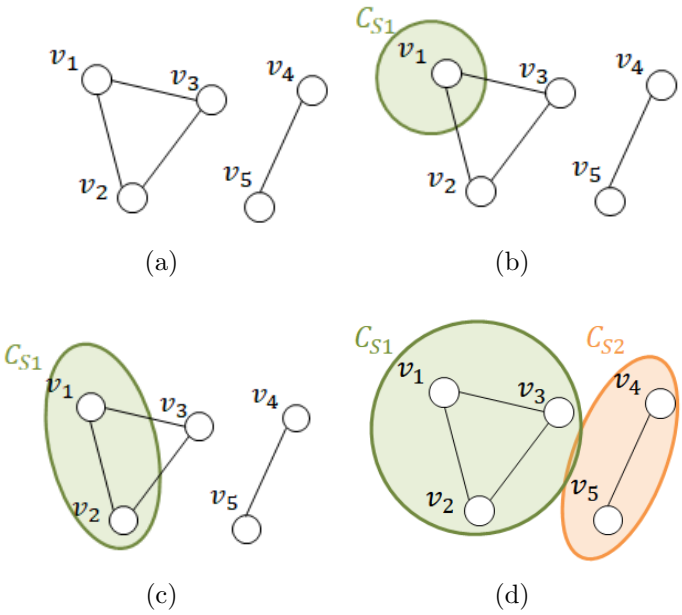



Figura 4.6: Obtención de comunidades con base en su estructura

Como vemos existe una intuitiva relación entre los enlaces que están representados en la matriz de adyacencia $M$ y el conjunto de comunidades $C_S$ que se generan. Esta relación nos lleva a que la multiplicación de la matriz de estructura $M_S$, que corresponde al conjunto de comunidades $C_S$, por su transpuesta puede obtener la

matriz de adyacencia $M$ como se muestra en la Figura 4.7 para el grafo de la Figura 4.6.

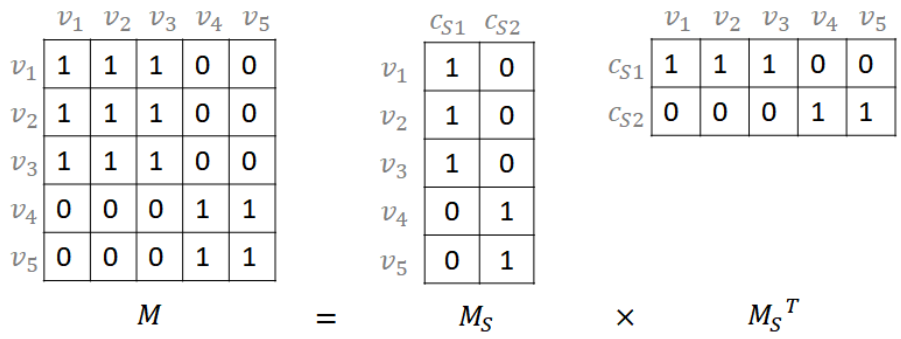



Figura 4.7: Generación de la matriz de adyacencia $M$ a partir de la multiplicación de la matriz de estructura $M_S$ con su transpuesta

En las redes sociales reales la separación no es tan evidente, al menos que las comunidades no tengan enlaces entre ellas, por lo que la multiplicación no genera la matriz de adyacencia $M$ original. De tal forma que la multiplicación generará una matriz de adyacencia estimada $M'$ que deberá ser lo más parecida posible a la matriz de adyacencia real $M$.

Entonces, definimos que la matriz de adyacencia estimada $M'$ es el producto de la multiplicación de la matriz de estructura $M_S$ por su transpuesta, tal que $M' \approx M$ como se muestra en la Ecuación 4.1. De tal forma que la matriz de adyacencia estimada $M'$ depende de la relación entre cualesquiera dos nodos $v_i$ y $v_j$ en cada una de las comunidades:

$$M' = M_S \cdot M_S^T \tag{4.1}$$

De tal forma que cada elemento $m_{ij}$ de la matriz de adyacencia $M$ representa la presencia o ausencia de una arista en la red; su estimación $m'_{ij}$ proviene de la suma de todas las pertenencias de los nodos $i$ y $j$ al conjunto de comunidades $C$, es

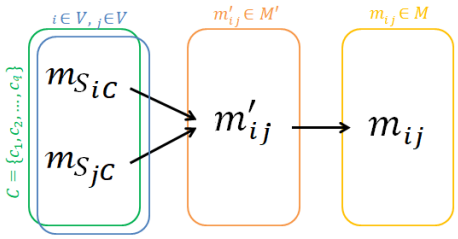



Figura 4.8: La matriz de adyacencia estimada $M'$ es obtenida evaluando la pertenencia o impacto de un nodo $v_i$ y $v_j$ simultáneamente en una comunidad $c_i$ para todas las comunidades $C$, tal que la matriz $M'$ sea similar a la representación matricial de la matriz de adyacencia original $M$ del red $G'$

decir de la relación entre el vector $m_{S_{iC}} = [m_{S_{i1}}, m_{S_{i2}}, ..., m_{S_{iq}}]$ y el vector $m_{S_{jC}} = [m_{S_{j1}}, m_{S_{j2}}, ..., m_{S_{jq}}]$ para $q = |C|$ comunidades, como se observa en la Figura 4.8, lo cual es obtenido a partir de la suma de la pertenencia de cada par de nodos a cada una de las comunidades $c_h$, como se ve en la Ecuación 4.2, para cada elemento $m_{S_{ij}}$ de la matriz de estructura $M_S$.

$$m'_{ij} = \sum_{h \in C} m_{S_{ih}} \cdot m^T_{S_{jh}} \tag{4.2}$$

Retomando los ejemplos de las Figura 4.3 y 4.4, si multiplicamos la matriz de estructura $M_S$ por su transpuesta, la matriz de adyacencia estimada $M'$ sería:

$$\begin{bmatrix} - & 1 & 1 & \mathbf{1} & 0 & 0 & 0 \\ 1 & - & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & - & 1 & \mathbf{0} & 0 & 0 \\ \mathbf{1} & 1 & 1 & - & 0 & 0 & 0 \\ 0 & 0 & \mathbf{0} & 0 & - & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & - & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & - \end{bmatrix}$$

que es similar más no igual a la matriz de adyacencia real $M$ de la Figura 4.2. En

negritas se distinguen las dos aristas que no corresponden, la que está en ceros (**0**) sugiere una arista entre los nodos $A$ y $D$ para tener un subgrafo completo entre los nodos $\{A, B, C, D\}$; mientras que la arista que está en 1, en la matriz estimada $M'$, es la que une al nodo $C$ con el nodo $E$, cuya ausencia permitiría la separación de los dos conjuntos. Como se observa, la estimación de la matriz $M'$ podría no ser la misma, pero se espera que el error sea lo mínimo posible.

Se requiere entonces, que la matriz de adyacencia estimada $M'$ sea lo más similar posible a $M$. Para ello se hace uso de regresiones para minimizar el error entre las originales y las estimadas. Se tiene que la estimación es $E(M|M') + e_M$ para la Ecuación 4.1 tal que $e_M$ es el error esperado; usando el modelo de regresiones tenemos que $M = \beta_M M' + e_M$, tal que la matriz $M$ es el valor condicional para la matriz estimada $M'$ y sea $\beta_M$ el efecto de la matriz estimada $M'$ sobre la matriz real $M$, de tal forma que se expresa como se muestra en la Ecuación 4.3:

$$e_M = M - \beta_M M' \tag{4.3}$$

Este error es minimizado como se muestra en la Sección 4.3.3.

### 4.3.2 Modelo de detección de comunidades basado en atributos

Se considera que las comunidades no sólo agrupan nodos, sino que también agrupan atributos; para ello veamos el conjunto de nodos con atributos en la Figura 4.9(a) correspondiente al grafo de la Figura 4.5(a), los cuales son convertidos a un grafo aumentado (ver Sección 2.2.1) de tal forma que los atributos se convierten en nodos-atributo (mostrados como cuadrados en la Figura 4.9) y los nodos que tenían un atributo crean un enlace a esos nodos-atributo, resultando la representación de la

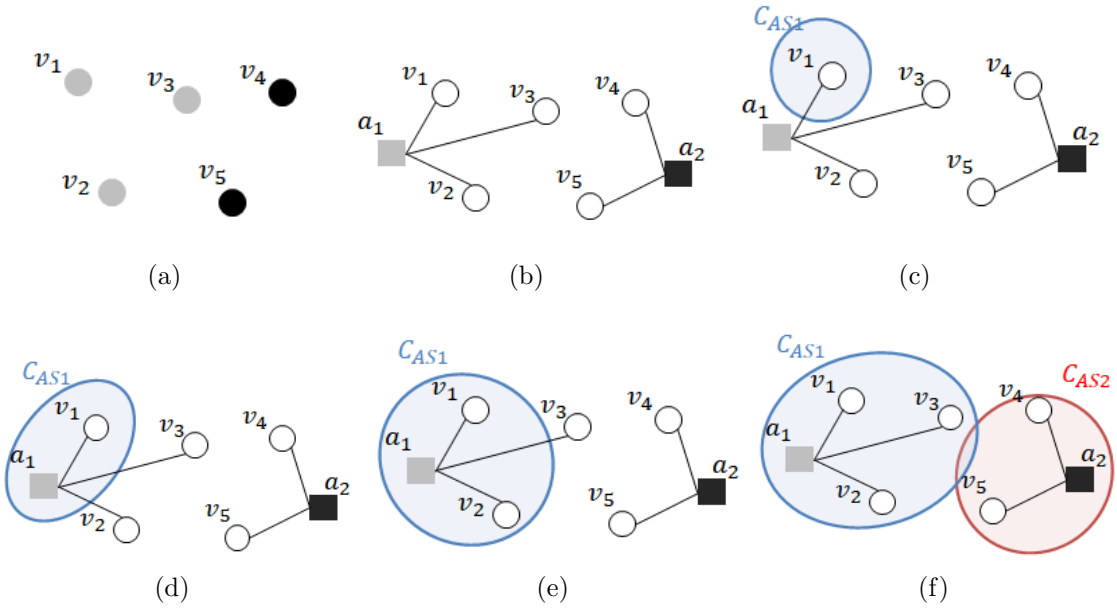



Figura 4.9: Obtención de comunidades con base en atributos

Figura 4.9(b).

Supongamos un conjunto de comunidades $C_{AS}$ cuyas comunidades contienen tanto nodos como atributos. Se parte de que un nodo $v_i$ pertenece a una comunidad $c_{AS_j}$ como es el caso del nodo $v_1$ a la comunidad $c_{AS_1}$ en la Figura 4.9(c). El atributo $a_1$ al tener un enlace a $v_1$, hace que ambos pertenezcan a $c_{AS_1}$, como se muestra en la Figura 4.9(d). Cuando otro nodo con el mismo atributo $a_1$ tiene un enlace a éste, obliga al nodo a pertenecer a la misma comunidad que el atributo, como sucede con el nodo $v_2$ en la Figura 4.9(e). Lo mismo sucede con el resto de los nodos, de tal forma que el conjunto de comunidades resultantes $C_{AS}$ se muestran en la Figura 4.9(f).

Como vemos, existe una relación entre la pertenencia de atributos a nodos representados en la matriz de atributos $X$ y el conjunto de comunidades $C_{AS}$ que se generan, de manera estricta al conjunto de comunidades de atributos $C_A$. Esta relación nos lleva a que la multiplicación de la matriz de atributos $M_A$ y la matriz de estructura $M_S$, que corresponden al conjunto de comunidades $C_A$ y $C_S$ respectivamente,

obtiene la matriz de atributos $X$, cuya transpuesta es $X^T$, como se muestra en la Figura 4.10 para el grafo de la Figura 4.9.

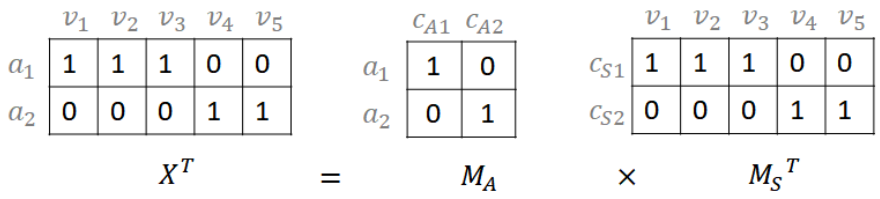



Figura 4.10: Generación de la matriz de adyacencia $M$ a partir a partir de la multiplicación de la matriz de estructura $M_S$ con su transpuesta

En redes reales, la separación no siempre es tan evidente, incluso existe un traslape ya que los nodos suelen tener más de un atributo por lo que la multiplicación no genera la matriz de atributos $X$ original. De tal forma que la multiplicación genera una matriz de atributos estimada $X'$, que deberá ser lo más parecida posible a la matriz de atributos real $X$.

La matriz estimada $X'$ depende de la relación entre un nodo $i \in V$ y un atributo $j \in A$ en una comunidad $c_h$, como se observa en la Figura 4.11, de tal forma que la multiplicación de $M_A$ por la transpuesta de $M_S$ nos dará $X'$, como se observa en la

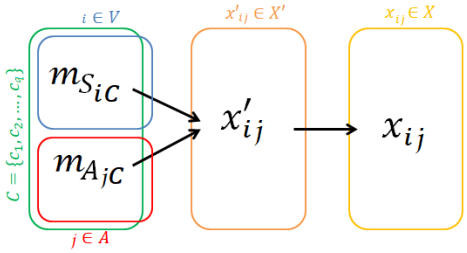



Figura 4.11: $X'$ es estimado evaluando la pertenencia o impacto de un nodo $i$ y de un atributo $j$ en el conjunto de comunidades $C$ de tal forma que $X'$ sea similar a la representación matricial de nodos-atributos $X$ de la red social original $G'$

Ecuación 4.4.

$$X' = M_A \cdot M_S^T \tag{4.4}$$

En la Figura 4.11, $x_{ij}$ representa la presencia o ausencia de un atributo $j$ en un nodo $i$ en la red $G$. La estimación $x'_{ij}$ proviene de la suma de todas las pertenencias de los atributos $j$ a los nodos $i$ en todo el conjunto de comunidades $C$, es decir de la relación entre el vector $m_{S_{iC}} = [m_{S_{i1}}, m_{S_{i2}}, ..., m_{S_{iq}}]$ y el vector $m_{A_{jC}} = [m_{A_{j1}}, m_{A_{j2}}, ..., m_{A_{jq}}]$ para $q = |C|$ comunidades, lo cual es obtenido a partir de la suma de la pertenencia en cada comunidad $c_h$ como se ve en la Ecuación 4.5:

$$x'_{ij} = \sum_{c_h \in C} m_{A_{jh}} \cdot m_{S_{ih}}^T \tag{4.5}$$

Se requiere que la matriz de atributos estimada $X'$ sea lo más similar posible a la matriz de atributos real $X$, al igual que con la matriz de adyacencia $M$ y su estimado $M'$, se hace uso de estimación del error tal que $E(X|X') + e_X$ para la Ecuación 4.4, de tal forma que $X = \beta_X X' + e_X$, tal que la matriz de atributos $X$ es el valor condicional para la matriz de atributos estimada $X'$ y $\beta_X$ mida el cambio de valor de la matriz de atributos estimada $X'$ desde la media de la matriz de atributos real $X$. De tal forma que el error entre la matriz de atributos real $X$ y la matriz de atributos estimados $X'$ se expresa como se muestra en la Ecuación 4.6:

$$e_X = X - \beta_X X' \tag{4.6}$$

### 4.3.3 Detección de comunidades usando el modelo RMOCA

Con el uso de mínimos cuadrados para los errores $e_M$ y $e_X$ de las Ecuaciones 4.3 y 4.6 dados por los modelos de detección de comunidades basados en estructura y atributos respectivamente, se genera la función objetivo de la Ecuación 4.7:

$$O(M_S, M_A) = (M - \beta_M M')^2 + (X - \beta_X X')^2 \tag{4.7}$$

Sustituyendo en la Ecuación 4.7, las ecuaciones del modelo (Ecuación 4.1 y 4.4), la función objetivo se puede representar con la Ecuación 4.8:

$$O(M_S, M_A) = (M - \beta_M M_S M_S^T)^2 + (X - \beta_X M_A M_S^T)^2 \tag{4.8}$$

De acuerdo a [117][122] la fase de aprendizaje para obtener las matrices $M_S$ y $M_A$ podría ser minimizando la función, de tal forma que el problema de optimización es:

$$min_{M_S, M_A \geq 0} O(M_S, M_A) \tag{4.9}$$

Para la optimización se utiliza el método del gradiente, por lo que se aplica el gradiente a la Ecuación 4.8 con respecto a la matriz de estructura $M_S$, como se observa en la Ecuación 4.10

$$\frac{\partial O}{\partial M_S} = -4\beta_M M M_S + 4\beta_M M_S M_S^T M_S - 2\beta_X X^T M_A + 2\beta_X M_A M_A^T M_S \tag{4.10}$$

y el gradiente con respecto a la matriz de atributos $M_A$ que se muestra en la Ecuación 4.11):

$$\frac{\partial O}{\partial M_A} = -2\beta_X X M_S + 2\beta_X M_A M_S M_S^T \tag{4.11}$$

Usando la metodología de [52] se pueden dividir las ecuaciones 4.10 y 4.11 en componentes positivos y negativos: $[\cdot]_+$ y $[\cdot]_-$, tales que $\frac{\partial O}{\partial M_S} = [\cdot]_+ - [\cdot]_-$ donde $[\cdot]_+ = 4\beta_M M_S M_S^T M_S + 2\beta_X M_A M_A^T M_S$, y $[\cdot]_- = 4\beta_M M M_S + 2\beta_X X^T M_A$. Para el proceso iterativo de aprendizaje se tiene: $m_{S_{ij}} = m_{S_{ij}} - \eta_{ij}([\cdot]_+ - [\cdot]_-)_{ij}$ donde $\eta$ es la taza de aprendizaje con $\eta_{ij} = \frac{m_{Sij}}{([\cdot]_+)_{ij}}$. De tal forma que la regla de actualización está dada por la Ecuación 4.12:

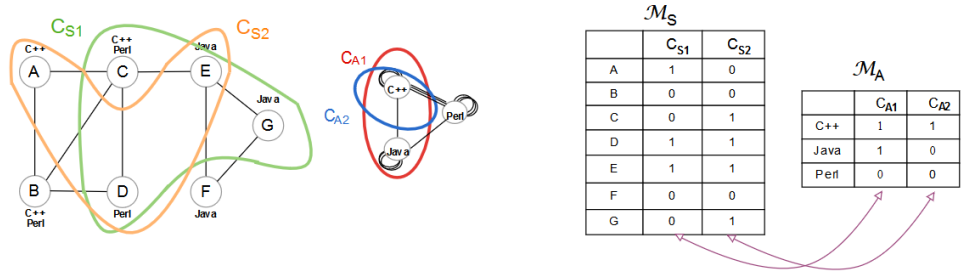
$$m_{S_{ij}} = m_{S_{ij}} \frac{([\cdot]_-)_{ij}}{([\cdot]_+)_{ij}} = m_{S_{ij}} \frac{(2\beta_M M M_S + \beta_X X^T M_A)_{ij}}{(2\beta_M M_S M_S^T M_S + \beta_X M_A M_A^T M_S)_{ij}} \tag{4.12}$$

La multiplicación converge cuando $([\cdot]_+)_{ij} = ([\cdot]_-)_{ij}$ tal que $\frac{\partial O}{\partial M_S} = 0$ es el punto estacionario en la función objetivo. Se aplicó el mismo procedimiento para actualizar $M_A$ mientras $M_S$ era constante, de tal forma que se tienen las Ecuación 4.13 y 4.14:

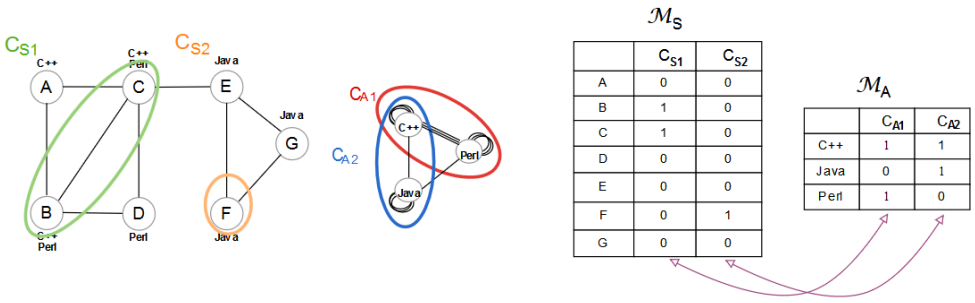
$$\frac{\partial O}{\partial M_A} = ([\cdot]_+ - [\cdot]_-) = 2\beta_X M_A M_S M_S^T - 2\beta_X X M_S \tag{4.13}$$

$$m_{A_{ij}} = m_{A_{ij}} - \frac{m_{A_{ij}}}{([\cdot]_+)_{ij}} = m_{A_{ij}} \frac{(X M_S)_{ij}}{(M_A M_S M_S^T)_{ij}} \tag{4.14}$$

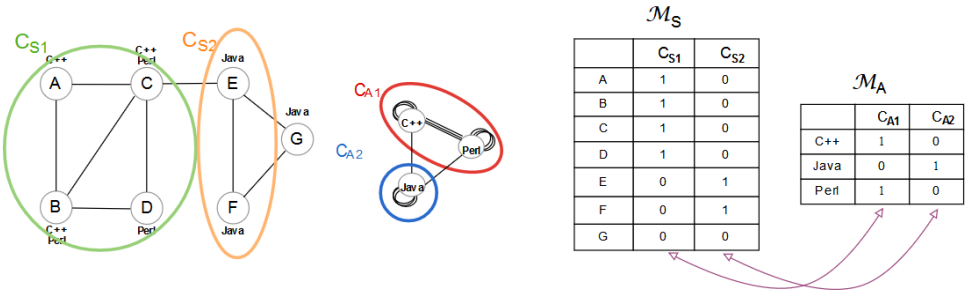
Para optimizar se itera de $M_A$ a $M_S$ hasta que convergen o llegan a un número máximo de iteraciones, de tal forma que se obtiene las matrices $M_A$ y $M_S$ por el proceso de minimización de errores entre $M'$ y el original $M$, así como el error entre $X'$ y $X$ como se observa en el Algoritmo 1.



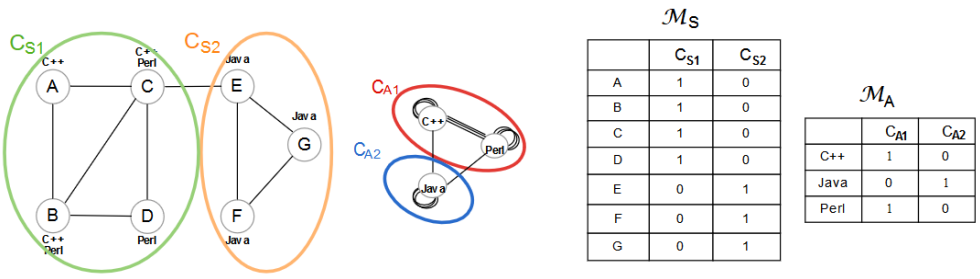

(a) Inicialización con valores aleatorios




(b) Primera iteración de la aproximación




(c) Segunda iteración de la aproximación




(d) Tercera iteración de la aproximación

Figura 4.12: Despliegue de resultados parciales dados en cada iteración del ciclo del algoritmo.

**Input**: $G$, $|C|$

**Output**: $C_S, C_A$

**while** $\frac{\partial O}{\partial M_S} \neq 0$ *and* $\frac{\partial O}{\partial M_A} \neq 0$ **do**

$$M_{S_{new}} = M_{S_{old}} \cdot \frac{\left[\frac{\partial O}{\partial M_S}\right]_-}{\left[\frac{\partial O}{\partial M_S}\right]_+};$$

$$M_{A_{new}} = M_{A_{old}} \cdot \frac{\left[\frac{\partial O}{\partial M_A}\right]_-}{\left[\frac{\partial O}{\partial M_A}\right]_+};$$

**end**

$C_A \longleftarrow M_A$;

$C_S \longleftarrow M_S$;

**return** $C_S$ and $C_A$;

**Algorithm 1:** Estimación de comunidades con RMOCA

Para estimar $C_S$ y $C_A$ primero se inicializa con valores aleatorios las matrices $M_S$ y $M_A$ como se muestra en la Figura 4.12(a) para el ejemplo de la Figura 4.2; posteriormente se comienzan las iteraciones según el ciclo del Algoritmo 1 donde se remplazan las matrices anteriores $M_{S_{old}}$ y $M_{A_{old}}$ por los nuevos calculados $M_{S_{new}}$ y $M_{A_{new}}$ con las Ecuaciones 4.12 y 4.14 a partir de los gradientes dados por Ecuación 4.10 y 4.11. Inicialmente las comunidades son asignadas de manera aleatoria como en la Figura 4.12(a). La Figura 4.12(b) muestra el resultado de la primera iteración suponiendo que se tengan los valores binarizados ya que, como se mencionó en la Sección 4.3, los valores estimados están en un rango de $M'_{v_i v_j} = [0, 1]$ y $X'_{v_i a_k} = [0, 1]$. Se observa en la Figura 4.12(c) que en la segunda iteración las comunidades detectadas abarcan más nodos en el caso de $M_S$ y se reducen en el caso de $M_A$, aproximándose cada vez más a $M$ y $X$, como podemos observar en los pasos que siguió el ejemplo de prueba mostrado. Cabe destacar que el algoritmo está implementado en el lenguaje de programación C.

#### 4.3.4.1 Complejidad

Considerando $m = |E|$ como el número de aristas en el sociograma, $n = |N|$ el número de nodos y $k = |A|$ el número de atributos, la estimación de $C_S$ toma $O(m + nk)$ y la de $C_A$ toma $O(nk)$ para cada comunidad $c$, de tal forma que la complejidad de RMOCA es $O(m+nk)+O(nk)$ que se reduce a $O(m+nk)$ por cada comunidad $c$ y cada iteración, por lo que el incremento del número de comunidades a detectar podría aumentar la complejidad del algoritmo.

## 4.4 Experimentos

Haciendo uso de redes sociales reales, se analizó la detección encontrada por RMOCA. En esta sección se muestra primero un análisis cualitativo de las comunidades detectadas añadiendo atributos, posteriormente se muestra una comparación con otros algoritmos del estado del arte evaluando primero las comunidades detectadas y luego comparándolas con las comunidades reales; para esto se hace uso de las métricas expuestas en Capítulo 3.

### 4.4.1 Experimento 1: demostración de comunidades detectadas añadiendo atributos

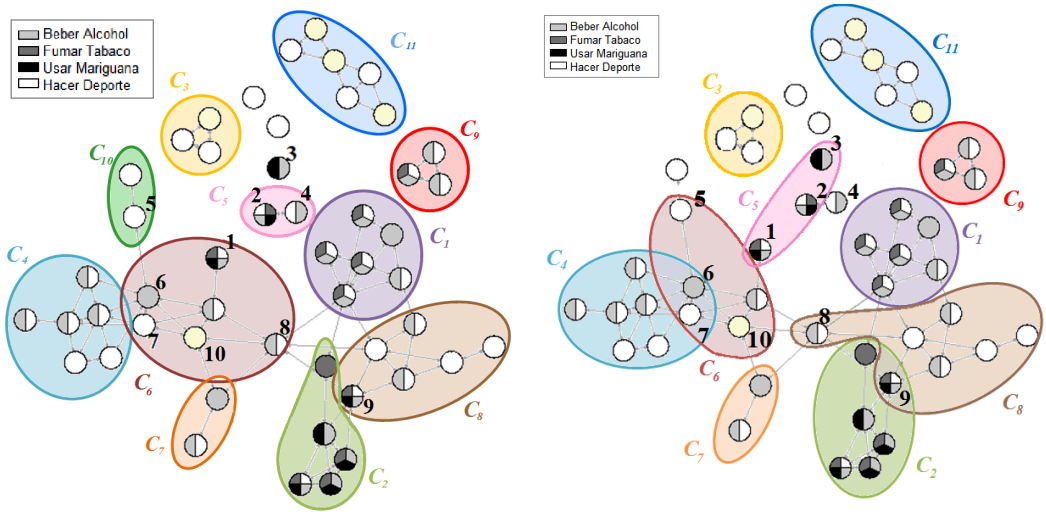
En este experimento se evalúa de manera cualitativa las comunidades detectadas en una red social pequeña para poder visualizarla y analizarla. Se comparan las comunidades detectadas usando el método de RMOCA y el algoritmo de Girvan-Newman (GN) [76] explicado en la Sección 3.1.3.

#### 4.4.1.1 Descripción de red social

La **red social de amigas adolescentes y su estilo de vida** estudiada por Michell, L. y A. Amos [68] tiene 50 adolescentes con 4 atributos que representan sus hábitos (hacer deporte, fumar tabaco, beber alcohol y consumir mariguana) los cuales son mostrados en escala de grises en la Figura 4.13, mientras que las aristas representan la amistad que existe entre ellas según lo que declararon.

#### 4.4.1.2 Resultado

Por el tamaño de esta red podemos estudiar las comunidades detectadas por RMOCA. En la Figura 4.13(a) tenemos las comunidades detectadas por el algoritmo de Girvan-Newman [76] y en la Figura 4.13(b) están las comunidades obtenidas por RMOCA. El algoritmo GN detectó once comunidades por lo que esa cantidad de comunidades se solicitó a RMOCA para tener un mejor parámetro de comparación.




(a) Comunidades detectadas con algoritmo GN (b) Comunidades detectadas con RMOCA

Figura 4.13: Comunidades detectadas en red social donde los nodos representan adolecentes, las aristas representan amistades y el relleno de los nodos en escala de grises los atributos.

#### 4.4.1.3 Análisis de comunidades detectadas

Como se puede observar en la Figura 4.13, las comunidades $c_1, c_3, c_7, c_9$ y $c_{11}$ en colores morado, amarillo, naranja, rojo, y azul rey son iguales en ambos algoritmos. A simple vista se observa que las comunidades de GN son más naturales considerando la estructura, sin embargo podemos observar que no todos los elementos que componen una misma comunidad comparten los mismos atributos. A continuación se detallan las diferencias entre GN y RMOCA con respecto a las comunidades detectadas, haciendo énfasis de casos particulares.

**Comunidad 6 de GN con comunidad 6 de RMOCA**

La comunidad $c_6$ (color guinda en Figura 4.13(a)) detectada por GN tiene 6 nodos con características muy diferentes, de hecho uno de ellos (nodo 10) no tiene ningún atributo y otro de ellos tiene todos los atributos (nodo 1); el resto tiene los atributos de hacer deporte y/o beber alcohol. En contraste la misma comunidad detectada por RMOCA descarta el nodo que consume drogas y fuma (nodo 1) dado que es el único que tiene esos atributos y considera un nodo más que hace deporte.

Para evaluar la estructura contaremos el número de aristas inter-comunidad, en el caso de la $c_6$ de GN tenemos 11 aristas que salen de ese comunidad: una hacia $c_{10}$, cinco a $c_4$, dos a $c_7$, dos a $c_8$ y una a $c_1$. Mientras que en $c_6$ de RMOCA tenemos 10 aristas inter-comunidades: cinco hacia nodos en $c_6$, una a $c_7$, tres a $c_8$, una a $c_5$ y otra a un nodo aislado ($c_{10}$ en GN), de tal forma que RMOCA tiene menos cortes para separarse del resto de la red que GN.

**Comunidad 8 de GN con Comunidad 8 de RMOCA**

La $c_8$ de GN y la $c_8$ de RMOCA difieren sólo en el nodo 8. Este nodo tiene tres aristas hacia la comunidad en que lo detecta GN ($c_6$ en guinda) y también tiene tres aristas que lo vinculan con la comunidad en la que lo detecta RMOCA ($c_8$ en café). Sin embargo, se observa que este nodo es igual a solo otro nodo en $c_6$ y a dos nodos en $c_8$, por lo que RMOCA lo clasifica en $c_8$.

**Comunidad 2 de GN con Comunidad 2 de RMOCA**

En $c_8$ de GN y de RMOCA (marcada en color café) todos los nodos hacen deporte y la mitad de ellos bebe alcohol con regularidad, sin embargo existe un nodo que es el único que consume drogas y fuma (nodo 9) por lo que comparte más características con los nodos de la comunidad $c_2$, de tal forma que este nodo es detectado por RMOCA en ambas comunidades, generando un traslape entre comunidades.

**Comunidad 5 de GN con Comunidad 5 de RMOCA**

La comunidad $c_5$ mostrada en color rosa en la Figura 4.13a de GN abarca dos nodos (nodos 2 y 4) conectados entre sí y aislados del resto de las comunidades, ya que se basa en la detección por estructura. Como RMOCA considera también atributos, agrupa los nodos 1, 2 y 3, donde 1 y 2 son iguales y 3 comparte uno de los atributos que está muy poco presente en el resto de los nodos: consumo de marihuana. Note que estos tres nodos no están conectados entre sí, sin embargo, no los agrupa con la comunidad $c_2$ (verde claro) que tiene nodos parecidos porque RMOCA también considera la estructura y los nodos en $c_2$ comparten muchas aristas entre ellos, lo que hace que deje fuera a los nodos en $c_5$.

**Comunidad 4 de GN con Comunidades 4 y 6 de RMOCA**

Otro traslape de comunidades se observa en $c_4$, la cual contiene nodos que hacen deporte y que consumen alcohol por lo que, a diferencia de GN, RMOCA tiene más nodos con estas características. Note que $c_4$ de GN tiene 5 aristas inter-comunidades y $c_4$ de RMOCA también tiene 5, por lo que estructuralmente no existe ninguna repercusión. Los traslapes dados por RMOCA son suaves por lo que no generan riesgo de tener dos comunidades casi iguales como suele suceder en algoritmos como GAMer.

### 4.4.2 Experimento 2: evaluación de la calidad de las comunidades.

Se evalúan las comunidades detectadas por RMOCA con respecto a las detectadas por métodos del estado del arte que sólo consideran estructura y que consideran estructura y atributos. Los experimentos se realizaron sobre cuatro redes sociales que serán descritas en esta sección. Para evaluar la calidad de las comunidades detectadas, se evalúa la estructura de las comunidades y se evalúan los atributos de los nodos en las comunidades .

#### 4.4.2.1 Medidas y algoritmos base

Para evaluar la calidad de las comunidades detectadas usamos dos medidas que permiten saber qué tan parecidos son los nodos en una comunidad (entropía) y que tantas conexiones hay en una comunidad (densidad). La entropía fue definida en la Ecuación 3.24 y la densidad en la Ecuación 3.2. Los valores de entropía y densidad fueron normalizados, por lo que se encuentran en un rango entre 0 y 1, donde 1 se refiere a un grafo completo para el caso de densidad y a que todos los nodos son

iguales (con los mismos atributos) para el caso de la entropía.

Se evalúan las comunidades detectadas por RMOCA con el algoritmo de **GN** (ver Sección 3.1.3) que sólo considera estructura y con **CESNA** (ver Sección 3.3.1) que considera tanto estructura como atributos.

#### 4.4.2.2 Conjuntos de datos

Para evaluar RMOCA se usaron dos redes sociales con pocos nodos y dos de mayor tamaño, las cuales se muestran en la Tabla 4.1. La primera de las pequeñas es la de los adolescentes usada en la Sección 4.4.1 y la segunda es una red de políticos mexicanos (*Pol.Mex.*) descrita en la Sección 2.1, la cual cuenta con dos comunidades reales, cuyos atributos representan años en que tomaron su primer cargo, así como sus profesiones.

Se realizaron experimentos en dos redes sociales reales de mayor tamaño. La primera de éstas es una red de jugadores de Football [40][1] de la división de IA en la temporada regular de Otoño 2000. Ésta cuenta con 115 nodos y 613 aristas no dirigidas. Los atributos de cada nodo indican en que conferencias han jugado. La segunda red social es la producida por *Blogs* de Política (*Blogs Pol.*) [1], la cual muestra interacciones entre *blogueros* en 2004, los atributos corresponden a los *Blogs* en los que comentaron y existen dos comunidades etiquetadas manualmente por los autores basados en las aristas que salen y las que entran.

Tabla 4.1: Redes sociales con atributos usadas para la evaluación de RMOCA

| | $\|N\|$ | $\|E\|$ | $\|A\|$ | $\|C\|$ |
|---|---|---|---|---|
| Adolescentes | 50 | 113 | 4 | - |
| Pol.Mex. | 35 | 117 | 11 | 2 |
| Football | 115 | 613 | 12 | 12 |
| Blogs Pol. | 1490 | 19091 | 9 | 2 |

[1]http://www-personal.umich.edu/ mejn/netdata/

#### 4.4.2.3 Resultados

Se evalúa la densidad y la entropía de las comunidades detectadas por los tres métodos en las cuatro redes sociales. Dado que se busca que se tenga en conjunto una buena densidad y entropía, se hace una tercera evaluación con ambas medidas juntas.

### Densidad

En la Tabla 4.2 se ve que la densidad obtenida por RMOCA es de las mejores. En caso de no ser la mejor se encuentra muy cercana a la que fue mejor. Esto nos indica que la matriz $M$ obtuvo una muy buena aproximación derivada de la función objetivo (Ecuación 4.8) en la que interviene las comunidades de nodos y atributos ($C_S$ y $C_A$).

Tabla 4.2: Comparación de densidad

| | CESNA | GN | RMOCA |
|---|---|---|---|
| Adolescentes | 0.02851524 | 0.04827031 | **0.06637168** |
| PolMex | **0.20940171** | 0.16025641 | 0.2008547 |
| Football | 0.0560087 | 0.05872757 | **0.08939641** |
| Blogs Pol. | 0.33713595 | **0.45662917** | 0.44377563 |

### Entropía

Dado que RMOCA considera atributos en la aproximación de la matriz $X$, la entropía es mejor que los algoritmos del estado del arte. Sin embargo, cuando se tienen pocos atributos en relación con el número de nodos, el desempeño de RMOCA no es el mejor, como es el caso de la red de Blogs de Política, como se observa en la Tabla 4.3.

Tabla 4.3: Comparación de entropía

| | CESNA | GN | RMOCA |
|---|---|---|---|
| Adolescentes | 0.20052160 | 0.25784601 | **0.264839995** |
| Pol.Mex. | 0.0848392 | 0.13713341 | **0.18625127** |
| Football | 0.0105135 | 0.04958609 | **0.21907635** |
| Blogs Pol. | 0.0712984 | **0.14770272** | 0.1165593 |

### Densidad y Entropía

Se busca que las comunidades detectadas tengan una buena entropía (se parezcan los nodos) y buena densidad (que existan muchas relaciones dentro de la comunidad). En la Figura 4.14 vemos que RMOCA obtiene mejoras significativas que las comunidades detectadas por GN que sólo considera estructura y que CESNA, que considera tanto estructura como atributos. En esta figura se aprecia que RMOCA no supera a GN en el conjunto de Blog de Políticos. Como se mencionó, este conjunto tiene muy pocos atributos comparado con la cantidad de nodos de esta red, por lo que no hay una mejor entropía, aún cuando la densidad es similar.

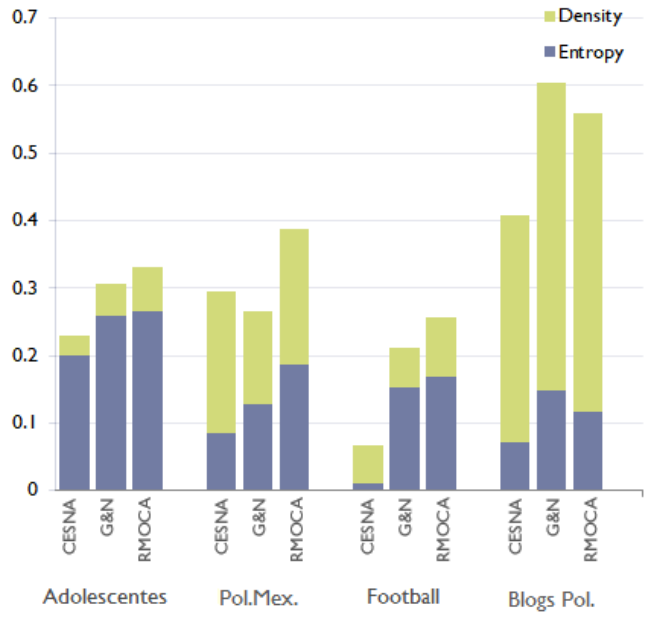



Figura 4.14: Comparación de estructura y atributos en las comunidades detectadas

### 4.4.3 Experimento 3: evaluación de comunidades detectadas con respecto a las reales

Algunos conjuntos de datos cuentan con comunidades reales con los que podemos comparar el desempeño de las comunidades detectadas. En este caso se evalúan las comunidades detectadas de RMOCA y algoritmos del estado del arte en estos conjuntos de datos.

#### 4.4.3.1 Algoritmos base y conjunto de datos

Las redes sociales de políticos mexicanos, de Football y de *Blogs* de política descritos en la Sección 4.4.2 cuentan con comunidades reales (*ground truth communities*) por lo que se evaluó qué tan parecidas fueron las comunidades detectadas a la reales.

#### 4.4.3.2 Métricas

Para evaluar qué tan parecidas son las comunidades reales a las detectadas se usaron las medidas de pureza de la Ecuación 3.7 y la medida $F_1$ descrita en la Ecuación 3.8. Ambas hacen uso de la precisión y recuperación de la información con respecto a las comunidades reales. La pureza nos permite ver qué tan buena precisión tienen las comunidades detectadas y la medida $F_1$ es un balance entre la precisión y la recuperación. Dado que CESNA y RMOCA detectan comunidades sobrepuestas, también se hace uso del índice omega de la Ecuación 3.12.

#### 4.4.3.3 Resultados

A continuación, se analizan los resultados obtenidos para cada una de las medidas en los métodos de CESNA, GN y RMOCA para los conjuntos de datos de Pol.Mex.,

Football americano escolar y red de Blogs Pol. de política.

### Pureza

En la Tabla 4.4 tenemos los resultados de pureza, la cual nos muestra que RMOCA obtiene muy buena precisión con respecto a los conjuntos esperados. Se observa que a pesar de que la entropía y densidad obtenida en el conjunto de Blogs Pol. no mejoró a otros algoritmos; la pureza de las comunidades obtenidas en RMOCA supera substancialmente a los del estado del arte. Por lo tanto, RMOCA detecta muy pocos falsos positivos.

Tabla 4.4: Comparación de la medida de pureza de las comunidades detectadas

| | CESNA | GN | RMOCA |
|---|---|---|---|
| Pol.Mex. | 0.317029 | 0.394928 | **0.562500** |
| Football | 0.925000 | 0.927143 | **0.988888** |
| Blogs Pol. | 0.297110 | 0.650374 | **0.935839** |

### Medida $F_1$

La medida de $F_1$ es la media armónica de la precisión y la recuperación, en la Tabla 4.5 tenemos los resultados obtenidos del promedio de los $F_1$ obtenidos de las comunidades detectadas con respecto a las reales y de las reales con respecto a las detectadas. A pesar de que RMOCA tiene una gran precisión, la recuperación de los datos no es tan buena por lo que no supera en este ámbito a otros algoritmos. Esto significa que deja fuera algunos elementos que pertenecen a determinadas comunidades, aún cuando agrupa nodos disjuntos.

Tabla 4.5: Comparación de la Medida de $F1$ de las Comunidades Detectadas

| | CESNA | GN | RMOCA |
|---|---|---|---|
| Pol.Mex. | 0.473856 | 0.533946 | **0.5726495** |
| Football | **0.9141005** | 0.766452 | 0.6632065 |
| Blogs Pol. | 0.423554 | **0.671318** | 0.608575 |

### Índice Omega

Una de las medidas para evaluar la detección de comunidades sobrepuestas es el índice omega que busca la similitud entre comunidades evaluando por pares de nodos. Como se puede observar en la Tabla 4.6 los algoritmos de CESNA y RMOCA tienen mejor desempeño que GN. A pesar de que GN no detecta comunidades sobrepuestas, el índice omega es alto dado que se sigue comparando con las comunidades esperadas y, al detectar comunidades naturales, su desempeño se ve favorecido. RMOCA obtuvo mejores resultados que CESNA en redes con menor densidad.

Tabla 4.6: Comparación del índice omega

| | CESNA | GN | RMOCA |
|---|---|---|---|
| PolMex | 0.2975661 | 0.53942059 | **0.61058314** |
| Football | **0.99062989** | 0.95790914 | 0.47563757 |
| Blogs Pol. | 0.54702466 | 0.53097584 | **0.61799048** |

## 4.5 Conclusiones

La detección de comunidades basada en modelo evita la definición de una medida que podría no contener todos los parámetros. La propuesta de RMOCA está basada en un modelo de regresiones con estimaciones estadísticas, que sigue un principio básico de multiplicación de matrices. La ventaja del modelo RMOCA está en la alternancia del cálculo de la matriz de adyacencia y la matriz de atributos que permitió la definición de la función objetivo.

La optimización por el método del gradiente conjugado nos da valores aproximados

con la primera derivada, pero la convergencia de este método es lenta a pesar de que el uso de los parámetros de $\beta$ permite moderar el error y propiciar una convergencia más rápida. Por otro lado, podrían usarse otros métodos de optimización, sin embargo el aporta está en la definición del modelo RMOCA.

Como se mencionó, la estimación de $C_S$ esta vinculada con $C_A$ por lo que $C_A$ podría ayudar para el etiquetado de las comunidades detectadas. Algo que debe ser evaluado con mayor detalle.

# Capítulo 5

# Uso de atributos para medir calidad de comunidades

Existen múltiples algoritmos para detectar comunidades, pero ¿cómo evaluar la calidad de las comunidades cuando no se conocen las comunidades reales? Esta pregunta se complica ya que no existe una definición global para comunidades. Es aceptado que las comunidades deben tener una densidad alta entre los individuos que la forman, tal que la mayoría de los enlaces de un nodo están dentro de su comunidad y muy pocos hacia nodos que pertenecen a otras comunidades, sin embargo, no existe una definición general ni cualitativa.

Determinar qué comunidad es mejor que otra depende de las necesidades que se deseen cubrir. La mayoría de las medidas se enfocan a la estructura de la red, sin embargo se ha demostrado que el uso de los atributos mejora las comunidades. Las medidas mixtas integran atributos y estructura de la red, las investigaciones revelan que debería existir un balance entre la estructura y los atributos tal que se tenga una medida como la mostrada en la Ecuación 3.26, que suma una medida de estructura con una de atributos. Existen muchas medidas para estructura y muy pocas para

atributos vinculados a la estructura de la red.

Existen dos problemas principales con las medidas de atributos del estado del arte. El primero es que todos los atributos son considerados igual de importantes, aún cuando son pocos alrededor de los cuales las entidades se agrupan, aunado a esto, los atributos tienen un impacto diferente en todo el grafo comparado con el impacto que tienen dentro de una comunidad, por lo que se deberían considerar diversos pesos para el mismo atributo. El segundo se relaciona con el uso de distancia lineales cuando generalmente se cuenta con una gran cantidad de atributos.

Se propone una medida de calidad de comunidades considerando atributos $Q_A$ que aborde los problemas antes expuestos. Posteriormente, esta medida es integrada con un Balance entre Atributos y Estructura que llamaremos **BAS** (por sus siglas en inglés, *Balanced Attribute and Structure*) que hace uso de una medida exclusiva de la estructura de la red.

En los procesos de detección de comunidades, las medidas de calidad de comunidades se integran a través de procesos de optimización de la medida, ya sea por maximización o minimización, como guía de caminos aleatorios (*random walk*) o con procesos de expansión que buscan el incremento de la calidad de la comunidad, como se propone en este trabajo.

En este capítulo se describe $Q_A$ y su integración a BAS-*measure*, posteriormente es usada para mejorar las comunidades y así detectar comunidades sobrepuestas. Finalmente, se evalúa el comportamiento de la medida en métodos tradicionales como Girvan-Newman (GN), mientras que la integración de BAS al modelo de detección de comunidades propuesto (RMOCA) se describirá en el Capítulo 6 con el método mixto.

## 5.1 Importacia de atributos para las comunidades

La medida de calidad de comunidades basada en atributos $Q_A(c)$ de una comunidad $c$ considera el grado de los nodos, el grado de atributos, la densidad, la conductividad y la asortatividad. La asortatividad es la preferencia de los nodos de una red para estar cerca de otros similares. Algunos métodos lo interpretan en términos del grado de un nodo [77], en este caso, se considera no sólo el grado de nodos sino también el grado de atributos.

Para definir la medida de calidad de comunidades basada en atributos ($Q_A(c)$) se explican los términos que se han introducido a esta medida como lo son: la importancia global de un atributo que considera la influencia de un atributo para crear comunidades, la importancia local de un atributo $W_a(c)$ que evalúa la probabilidad de que un nodo en una comunidad $c$ tenga el atributo $a$, de acuerdo a su importancia global, y la fracción de atributos en el conjunto de comunidades $H_a(C)$.

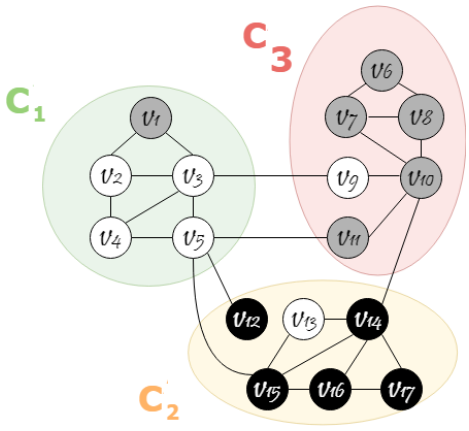



Figura 5.1: Ejemplo de red con tres atributos mostrados en blanco, gris y negro.

### 5.1.1 Importancia Global de Atributos

Definiremos como importancia global de un atributo $a$ a la relevancia que tiene un atributo en todo el grafo $G$ para formar comunidades por lo que se considera, principalmente, la propiedad de homofília.

**Definición 5.1** ***Importancia Global de un Atributo*** $W_a(G)$. *Dado un grafo* $G'(V, E, A)$ *con un conjunto de vértices* $V$, *aristas* $E$, *atributos* $A$ *y una relación* $\Lambda$ *entre atributos y vértices, la importancia global de un atributo* $a \in A$ *en un grafo* $G$ *está dada por el número de aristas entre nodos* $|(i, j) \in E : a \in \Delta(i), a \in \Delta(j)|$ *y el número de aristas que tienen los nodos con a, tal que* $|(i, j) \in E : a \in \Delta(i), a \in \Delta(j) \vee a \notin \Delta(j)|$ *y* $i, j \in V$.

De tal forma que la importancia global de un atributo $W_a(G)$ está dada por la Ecuación5.1:

$$W_a(G) = \frac{|(i, j) \in E : a \in \Delta(v), a \in \Delta(j)|}{|(i, j) \in E : a \in \Delta(i), a \in \Delta(j) \vee a \notin \Delta(j)|} \tag{5.1}$$

La propuesta, importancia global de un atributo, considera que un atributo es importante si:

1. los nodos que lo contienen están conectados,
2. si los nodos que no lo tienen están conectados, por lo que sería un grupo contra este atributo,
3. si es importante en más de un grupo disconexo y
4. no si es un máximo local.

De tal forma que el peso global de un atributo $W_a(G)$ evalúa la proporción de aristas entre nodos que comparten un atributo con respecto a aquellos que son inter-comunidades.

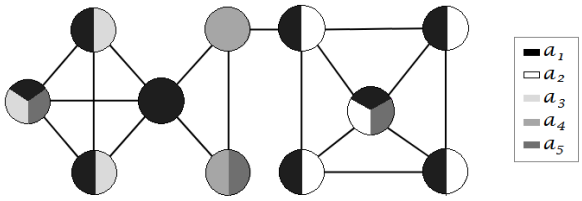
En la Figura 5.1 la importancia del atributo *gris* está dada por las seis aristas entre los nodos con ese atributo $\{(v_6, v_7), (v_6, v_8), (v_7, v_8), (v_7, v_{10}), (v_8, v_{10}), (v_{10}, v_{11})\}$ y las once aristas que enlazan al menos un nodo con *gris*, las cuales son: $\{(v_6, v_7), (v_6, v_8), (v_7, v_8), (v_7, v_{10}), (v_8, v_{10}), (v_{10}, v_{11}), (v_9, v_{10}), (v_1, v_2), (v_1, v_4), (v_5, v_{11}), (v_{10}, v_{14})\}$ definidas en la Ecuación 5.1 como $|(v_i, v_j) \in E : a_k \in \Delta(v_i), a_k \in \Delta(v_j) \vee a_k \notin \Delta(v_j)|$. Por lo tanto, el peso global del atributo *gris* es $W_{gris}(G) = \frac{6}{11} = 0.5454$ ya que existen fuertes conexiones entre los nodos con ese atributo. El atributo menos importante es *blanco* con un peso global de $W_{blanco}(G) = \frac{5}{14} = 0.3571$ ya que se encuentra en todas las comunidades.

#### 5.1.1.1 *Ranking* de atributos

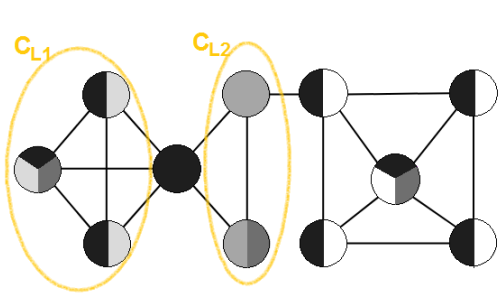
La importancia global de un atributo también es usada para seleccionar los atributos más importantes, para ello se hace un *ranking* y se seleccionan aquellos con el mayor impacto para formar comunidades.

Supongamos el ejemplo de la Figura 5.2(a), en el que tenemos once vértices y cinco atributos $A = \{a_1, a_2, a_3, a_4, a_5\}$. En la Tabla 5.1 podemos observar el orden de importancia que se obtuvo con la importancia global de un atributo $W_a(G)$ comparada con la importancia *Global Weighting* de la Ecuación 3.30 en el Capítulo 3.

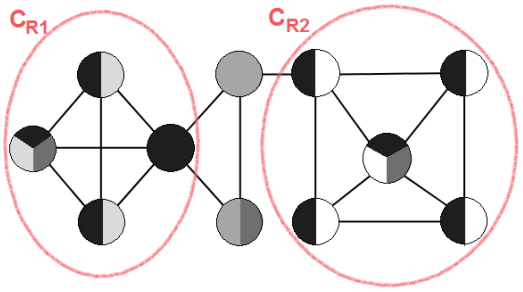
Se observa que para *Global Weighting* de Linkrec los mejores dos atributos son $a_3$ y $a_4$, si hiciéramos comunidades considerando estos dos atributos, tendríamos las comunidades $C = \{c_{L1}, c_{L2}\}$ como las mostradas en Figura 5.2(b), de tal forma que $c_{L1}$ agrupa a los nodos con el atributo $a_3$, mientras que $c_{L2}$ agrupa a los nodos con $a_4$. Se observa que estructuralmente esta separación de grupos hace cortes en seis aristas




(a) Grafo con atributos




(b) Comunidades dadas por los mejores dos atributos con *Global Weighting*




(c) Comunidades dadas por los mejores dos atributos con $W_a(G)$

Figura 5.2: Comparación de comunidades dadas por los dos atributos más importantes definidos por importancia global $W_a(G)$ y por *Global Weighting*

Tabla 5.1: *Ranking* de atributos dado por importancia global $W_a(G)$ y por *Global Weighting* en la red social de la Figura 5.2

| *Ranking* | *Global Weighting* | $W_a(G)$ |
|---|---|---|
| 1 | $a_3$ | $a_2$ |
| 2 | $a_4$ | $a_1$ |
| 3 | $a_2$ | $a_3$ |
| 4 | $a_1$ | $a_4$ |
| 5 | $a_5$ | $a_5$ |

ya que $c_{L1}$ corta el grafo completo que se forma por los cuatro nodos de la izquierda.

En la Tabla 5.1 se puede ver que los dos primeros lugares de importancia que otorga importancia global $W_a(G)$ son para el atributo $a_2$ y $a_1$. Nuevamente, si construyéramos las comunidades agrupando los nodos con estos atributos, las comunidades que se detectarían serían $C = \{C_{R1}, C_{R2}\}$ mostradas en la Figura 5.2(c).

Vemos que $C_{R1}$ detecta un subgrafo completo y la comunidad $C_{R2}$ se separa del resto del grafo con tan sólo el corte de una arista.

Ambos algoritmos consideran la estructura de la red para obtener la importancia de los atributos. Sin embargo *Global Weighting* de LINKREC cae en máximos locales porque busca grafos completos sin minimizar los cortes como lo hace nuestra mediad de importancia global $W_a(G)$.

### Ejemplo en la red social Facebook

Ahora comparemos el *ranking* generado por nuestra importancia global $W_a(G)$ y el *Global Weighting* de LINKREC. Usaremos una ego-red de Facebook que contiene 57 nodos, 42 atributos y 292 aristas.

El conjunto completo de atributos de la red social de Facebook lo podemos ver en la Figura 5.3, en donde las etiquetas de los atributos con el tamaño de letra más grande muestran a los atributos que fueron evaluados como más importantes. En la Figura 5.3(a) vemos los atributos evaluados por *Global Weighting* para el procesos de Linkrec y en la Figura 5.3(b) para la evaluación realizada por importancia global $W_a(G)$.

La evaluación de los atributos con importancia global $W_a(G)$ nos permite obtener los $Top - n$ atributos, es decir, los mejores $n$ atributos. Notamos que los atributos en Facebook considerados más importantes se relacionan con el lugar y la educación. En el ejemplo que se muestra en la Figura 5.3, el género aparece como un atributo importante para segmentar grupos, aparentemente es de un estudiante de tecnología en donde por lo regular no hay el mismo número de alumnos en ambos géneros y, dado que se evalúan los enlaces con otros nodos y hay pocos nodos de un género, predomina el otro.


(a) *Global Weighting* (b) Importancia Global $W_a(G)$

Figura 5.3: Bolsa de palabras de los atributos. El tamaño del nombre de los atributos representa la relevancia de cada atributo.

### 5.1.2 Importancia Local de un Atributo

**Definición 5.2** *La* ***Importancia Local de un Atributo*** $W_a(c)$. *Dado un grafo* $G'(V, E, A)$ *con un conjunto de vértices* $V$, *aristas* $E$, *atributos* $A$ *y una relación* $\Lambda$ *entre atributos y vértices, la importancia local de un atributo* $a$, *en la comunidad* $c$, *está dada por la relación del número de nodos con el atributo* $a$ *en la comunidad* $c$ *y el número total de atributos en la comunidad* $c$, *ponderada por la importancia global del atributo* $a$ *en* $G$ *dada por* $W_a(G)$, *la cual está dada en un rango de* $0 \leq W_a(c) \leq 1$

De tal forma que $W_a(c)$ se expresa como se muestra en la Ecuación 5.2 para $i \in V$:

$$W_a(c) = W_a(G) \cdot \frac{|i : a \in \Lambda(i), i \in c|}{|i : i \in c|} \tag{5.2}$$

Esta medida considera, además, que cada atributo tiene una importancia global diferente, esto es necesario ya que la importancia de una comunidad será mayor si tiene atributos más importantes. Si un nodo en la comunidad $c$ no tiene un atributo la importancia local del atributo $a$ es $W_a(c) = 0$, pero si la comunidad $c$ tiene nodos con el atributo $a$, la importancia local depende de la importancia global y de la probabilidad de que los nodos en esa comunidad tengan esa propiedad.

### 5.1.3 Densidad de un Atributo

**Definición 5.3** ***Densidad de un Atributo en una comunidad*** $H_a(c)$. *Dado un grafo* $G'(V, E, A)$ *con un conjunto de vértices* $V$, *aristas* $E$, *atributos* $A$ *y una relación* $\Lambda$ *entre atributos y vértices, la densidad de un atributo* $a$ *en una comunidad* $c$ *está dada por la relación del número de nodos con el atributo* $a$ *en la comunidad* $c$ *y el número total de nodos con el atributo* $a$ *en el grafo* $G$.

De tal forma que $H_a(c)$ se expresa en la Ecuación 5.3:

$$H_a(c) = \frac{|i : a \in \Lambda(i), i \in c|}{|i : a \in \Lambda(i)|} \tag{5.3}$$

En la Figura 6.2, la probabilidad de que el atributo *negro* esté en la comunidad $c_2$ es $H_{negro}(c_2) = 5/6$ según la Ecuación5.3 porque existen cinco nodos con negro de los seis nodos en la comunidad $c_2$.

## 5.2 BAS: calidad de comunidad basado en atributos y estructura

Dadas la importancia global de un atributo $W_a(G)$, la importancia local de un atributo $W_a(c)$ y la densidad de un atributo $H_a(c)$, se define la calidad de la comunidad $Q_A(c)$ basado en la premisa de que los atributos son diferente e independientes entre ellos, por lo que la medida toma el principio de la similitud coseno.

Tomando en cuenta que la densidad de un atributo $a$ en $c$ dada por $H_a(c)$ mide la importancia de los atributos según su densidad y la importancia local de un atributo $a$ dada por $W_a(c)$ es la importancia de los atributos, según los enlaces, se considera que la calidad de la comunidad es mejor si estos dos indicadores son similares por

lo que se calcula la distancia en todas las $k$ dimensiones, donde $k = |A|$ equivale al número de atributos. También, se considera la importancia global del atributo, lo que evita mínimos locales.

De tal forma que la **calidad de una comunidad basada en atributos** es definida como se muestra en la Ecuación5.4:

$$Q_A(c) = \frac{\sum_{i=1}^{k} H_i(c) \cdot W_i(c)}{\sqrt{\sum_{i=1}^{k} H_i(c)^2 \cdot \sum_{i=1}^{k} W_i(G)^2}} \tag{5.4}$$

La calidad de la comunidad según los atributos de la comunidad $c_1$ en la Figura 6.2(a) está dada por $Q_A(c_1) = \frac{\frac{1}{6}\cdot\left(\frac{6}{11}\cdot\frac{1}{5}\right)+\frac{4}{6}\cdot\left(\frac{5}{14}\cdot\frac{4}{5}\right)+\frac{0}{5}\cdot\left(\frac{5}{10}\cdot\frac{0}{5}\right)}{\sqrt{\left(\frac{1}{6}^2+\frac{4}{6}^2+\frac{0}{5}^2\right)\cdot\left(\frac{6}{11}^2+\frac{5}{14}^2+\frac{1}{2}^2\right)}} = \frac{0.2086}{0.6805} = 0.3065$ en donde se consideran los atributos *blanco* y *gris* porque el atributo *negro* no está en la comunidad $c_1$.

Para el **balance entre atributos y estructura (BAS)** se hace uso de la calidad de comunidades basada en atributos de la sección anterior y de la conductividad. Como se ha mencionado, la conductividad $\phi(c)$ es el número de aristas entre grupos, dadas por cortes, divididas por las aristas internas dentro del conjunto. Para ser una buena comunidad el conjunto de nodos debería tener una conductividad baja. Es por esto que algunos algoritmos [63] usan una de aproximación para minimizar la conductividad en el problema de cortes. Cabe destacar que la conductividad tiende a ser mejor cuando se tienen pocos grupos. Para la calidad de comunidades dadas por estructuras ($Q_S(c)$), se usó la conductividad ($\phi$), ya que ésta corresponde a la premisa de que una comunidad es un conjunto de nodos que están más conectados internamente que con el exterior. La Ecuación 3.6 muestra la conductividad de una comunidad $c$.

Haciendo uso de la conductividad $\phi(c)$ y la calidad de comunidades basada en atributos $Q_A(c)$ de la Ecuación 5.4, se hace un balance de calidad de comunidades

sugerido por Ecuación 3.26. BAS es la medida de calidad de una comunidad $c$ dada por la Ecuación 5.5:

$$Q(c) = \alpha(1 - \phi(c)) + \gamma(Q_A(c)) \tag{5.5}$$

donde $\gamma = \alpha - 1$ y $\alpha + \gamma = 1$, de tal forma que para $\alpha = 0$ la estructura no es considerada y para $\gamma = 0$ los atributos no son considerados. En los experimentos nos referiremos al uso de la Ecuación 5.5 como BAS.

## 5.3 Experimentos

Para evaluar la mejora de las comunidades con la medida $Q_A(C)$ se pueden tomar como comunidades iniciales aquellas que vengan de un proceso de detección basado en estructura, en atributos o en ambos. Para los experimentos de esta sección, se usó como comunidades de entrada aquella derivadas del algoritmo Girvan-Newman (GN) [40], ya que éste encuentra comunidades con divisiones naturales y no requiere que se le especifiquen el número de comunidades ni restricciones de tamaño. Cabe recalcar que este algoritmo se basa en la estructura del grafo.

Los experimentos del proceso de mejora de la calidad de las comunidades se dividen en dos: el primero considera la medida de calidad de comunidades basada en atributos $Q_A(C)$ y el segundo la calidad de comunidades que Balances Atributos y estructura (BAS) dado por $Q(C)$:

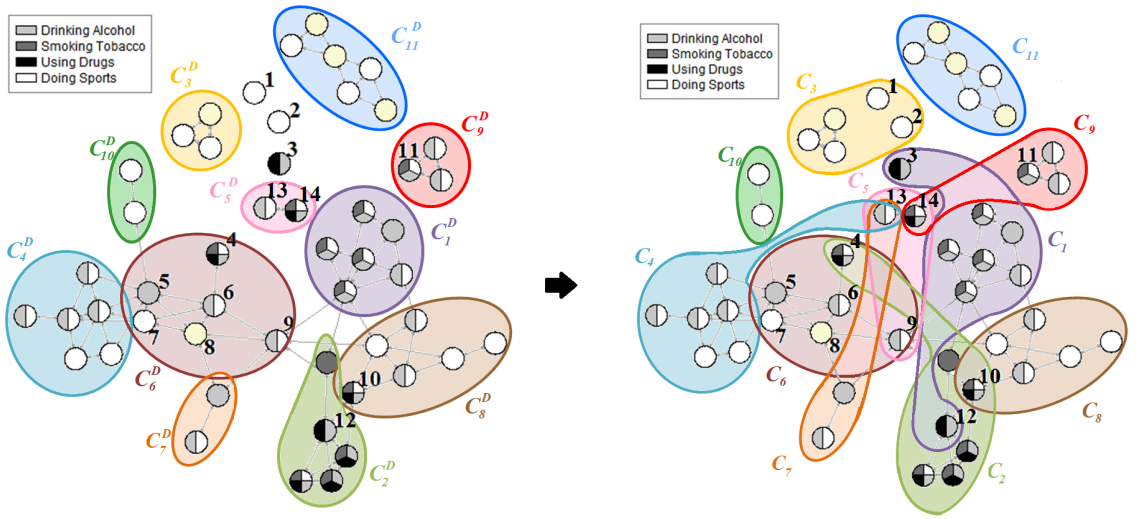
- **GN + $Q_A$:** Se usan las comunidades dadas por el algoritmo Girvan-Newman (GN) que considera estructura y se mejoran haciendo uso de la medida de calidad según atributos $Q_A$ de la Ecuación 5.4.

- **GN + $BAS$:** Se usan las comunidades dadas por el algoritmo Girvan-Newman (GN) que considera estructura y se mejoran usando como medida la calidad

según atributos $BAS$ de la Ecuación 5.5.

### 5.3.1 Experimento 1: variación de las comunidades usando atributos con la medidad $Q_A(C)$

Dado que se propone una nueva medida para evaluar los atributos, se compara de manera cualitativa y cuantitativa la mejora de comunidades usando $Q_A(C)$.




(a) Comunidades detectadas por GN

(b) Comunidades considerando atributos (GN+$Q_A$)

Figura 5.4: Comparación de comunidades detectadas en red social de cincuenta adolescentes, donde el color de los nodos representa cuatro atributos (beber alcohol, fumar tabaco, consumir mariguana y hacer deporte)

A través del estudio de la red de adolescentes [68] se comparan dos conjuntos de comunidades evaluados con la medida $Q_A$ que considera la similitud de atributos según su estructura y de su densidad. La red de adolescentes (*Teenage Friends and Lifestyle Network*), como ya se ha explicado, cuenta con 50 nodos y 4 atributos que representan hábitos de adolescentes. Por el tamaño de ésta se puede observar detalladamente la comparación de las comunidades.

El algoritmo de Girvan-Newman detecta once comunidades representadas en la Figura 5.4(a) que, para diferenciarlas de las comunidades usando atributos, se representan como $C^D$. Los grupos aislados son detectados como comunidades, las cuales son: $c_3^D$, $c_5^D$, $c_9^D$ y $c_{11}^D$. Los nodos $1, 2$ y $3$ no pertenecen a ninguna comunidad, ya que no tienen aristas. La comunidad $c_6^D$ en color guinda tiene seis nodos que no se parecen, pero que están altamente conectados; de hecho, existe un sub-grafo completo entre los nodos $5, 6, 7$ y $8$. El nodo 10 comparte dos atributos con los nodos en $c_8^D$ y al menos tres atributos con los nodos en $c_2^D$, sin embargo pertenece a $c_8^D$.

**Importancia global**

La importancia global de los atributos de esta red social, dada por la Ecuación 5.1, determinó que el atributo más importante para crear grupos es el alcohol con $W_{alcohol}(G) = 0.64$ y la actividad menos importante para formar grupos es fumar con $W_{tabaco}(G) = 0.33$. Esto significa que beber alcohol es una actividad mucho más "sociable", seguida por la práctica de deportes, mientras que fumar tabaco es la actividad más "solitaria". Específicamente, observemos que en esta red social existen 12 nodos que fuman tabaco (atributo en gris obscuro), una cuarta parte de ellos no tienen amigos que también fumen los cuales están marcados en la Figura 5.4 como los nodos $1, 2$ y $3$; mientras que otra cuarta parte (nodos $4, 5$ y $6$) sólo tienen un amigo que fumen. Por ello, fumar tabaco no es considerado como un atributo importante para formar grupos. En cambio, observemos que los nodos que beben alcohol tienen amigos que también consumen alcohol, excepto por el nodo 7, que es el único de veintinueve nodos que consumen alcohol y no tienen amigos con este atributo; por lo que el consumo de alcohol es una actividad que serviría para formar grupos.

### Importancia local

Para analizar la importancia local de los atributos, observe que la comunidad $c_1^D$ de color morado en 5.4(a) tiene seis nodos que no usan drogas, por lo que ese atributo se convierte en el menos importante dado por la Ecuación 5.2. En la comunidad $c_2^D$ en verde olivo, el atributo menos importante es hacer *deporte*, aún cuando globalmente era el segundo más importante, ya que sólo un nodo tiene este atributo. En la comunidad $c_2^D$ hay cinco nodos que fuman, cinco que beben alcohol y cinco que consumen drogas. Debido a que la importancia local considera la importancia global, el atributo más importante continúa siendo el *alcohol*.

En la Tabla 5.2 se observa la comparación de los pesos locales entre las comunidades disjuntas de GN y las comunidades sobrepuestas mostradas en la Figura 5.4(b), donde $W_{a_i}(C)$ es el promedio de los pesos locales de cada atributo $a_i$ en todas las $c_j$ comunidades, tal que $W_{a_i}(C) = (\sum_{j=1}^{k} W_{a_i}(c_j))/k$. A excepción de *tabaco*, las comunidades que consideran atributos ($GN+Q_A$) incrementan el promedio de importancia local de los atributos; la excepción se debe a que *tabaco* tiene la importancia global más baja.

Tabla 5.2: Comparación de peso local de atributos $W_{a_1}(C)$ en la red de adolescentes

| | GN | GN + $Q_A$ |
|---|---|---|
| $W_{alcohol}(C)$ | 0.3859 | **0.3920** |
| $W_{deporte}(C)$ | 0.4411 | **0.4530** |
| $W_{droga}(C)$ | 0.0846 | **0.1048** |
| $W_{tabaco}(C)$ | **0.0790** | 0.0757 |

### Peso de un nodo con atributo ($H_{a_i}(C)$)

La probabilidad de que un nodo con el atributo $a_i$ pertenezca a una comunidad está expresada en la Ecuación 5.3. Si comparamos esta probabilidad para cada atributo en cada comunidad, se ve que se tiene un incremento cuando se consideran atributos

y si los nodos comparten atributos. La Tabla 5.3 compara GN con $GN + Q_A$ en el promedio de la probabilidad dado por $H_{a_i}(C) = (\sum_{j=1}^{k} H_{a_i}(c_j))/k$.

Tabla 5.3: Comparación de la probabilidad de un atributo $H_{a_1}(C)$ en la red de adolescentes

| | GN | GN + $Q_A$ |
|---|---|---|
| $H_{alcohol}(C)$ | 0.0877 | **0.1128** |
| $H_{deporte}(C)$ | 0.0883 | **0.1085** |
| $H_{droga}(C)$ | 0.0795 | **0.1363** |
| $H_{tabaco}(C)$ | 0.0909 | **0.1136** |

### Calidad de la comunidad

Finalmente, la Tabla 5.4 muestra una comparación de la calidad obtenida para cada comunidad en la red de adolescentes en ambos algoritmos. El mayor incremento se tiene en la $c_2$, marcada en verde olivo, ya que la comunidad al considerar atributos considera un nodo 10, el cual consume mariguana, al igual que casi todos los miembros de esa comunidad, y este atributo era el segundo más importante a nivel global, mientras que localmente es en la comunidad en la que tiene la mayor importancia.

Tabla 5.4: Comparación de la calidad según atributos, $Q_A(c_i)$, aplicando GN y GN+$Q_A$ en la red de adolescentes

| | $Q_A(c_1)$ | $Q_A(c_2)$ | $Q_A(c_3)$ | $Q_A(c_4)$ | $Q_A(c_5)$ | $Q_A(c_6)$ | $Q_A(c_7)$ | $Q_A(c_8)$ | $Q_A(c_9)$ | $Q_A(c_{10})$ | $Q_A(c_{11})$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| GN | 0.3155 | 0.4998 | 0.033 | 0.1989 | 0.147 | 0.1374 | 0.0574 | 0.1909 | 0.1446 | 0.0495 | 0.0371 |
| GN + $Q_A$ | **0.3564** | **0.8135** | **0.0792** | **0.2415** | **0.1708** | 0.1374 | **0.0981** | 0.1909 | **0.229** | 0.0495 | 0.0371 |
| Incremento | 0.0409 | 0.3137 | 0.0462 | 0.0426 | 0.0238 | 0.0000 | 0.0407 | 0.0000 | 0.0844 | 0.0000 | 0.0000 |

Se evalúa la mejora de las comunidades disjuntas obtenidas por el método de GN a las comunidades sobrepuestas resultantes usando los atributos $A$. Para ello, se usan las medidas de densidad y entropía, ya que se buscan comunidades con nodos fuertemente conectados y parecidos.

En la Tabla 5.5 se observa la densidad de cada una de las comunidades de la Figura 5.4. Como se ve, existe un pequeño incremento en la densidad aún cuando se añaden nodos disjuntos. Como se ha dicho, el atributo menos importante es *tabaco*,

y es el único atributo en el que no se incrementa la entropía, como se puede ver en la Tabla 5.6. Con $Q_A$ se esperaba tener un incremento en la entropía y no afectar tanto la densidad, sin embargo también se tuvo un ligero incremento en la densidad.

Tabla 5.5: Comparación de la densidad entre GN y GN + $Q_A$ de cada atributo en la red de adolescentes

| Community | $c_1$ | $c_2$ | $c_3$ | $c_4$ | $c_5$ | $c_6$ | $c_7$ | $c_8$ | $c_9$ | $c_{10}$ | $c_{11}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| GN | 0.0973 | 0.0796 | 0.0176 | 0.0973 | 0.0000 | 0.0707 | 0.0000 | 0.0796 | 0.0176 | 0.0000 | 0.0880 |
| GN + $Q_A$ | **0.1238** | **0.0973** | 0.0176 | **0.1150** | 0.0000 | **0.1061** | **0.0176** | 0.0796 | **0.0530** | 0.0000 | 0.0880 |

Tabla 5.6: Comparación de la entropía entre GN y GN + $Q_A$ de cada atributo en la red de adolescentes

| Atributo | *tabaco* | *droga* | *alcohol* | *deporte* |
|---|---|---|---|---|
| GN | **0.1139** | 0.1321 | 0.3451 | 0.4400 |
| GN + $Q_A$ | 0.0849 | **0.1571** | **0.3979** | **0.5587** |

Como era esperado, el uso de atributos incremente la similitud dentro de las comunidades dada por la entropía, por lo que los nodos seleccionadas para integrarse a las comunidades fueron seleccionados apropiadamente por la medida $Q_A$.

### 5.3.2 Experimento 2: mejora de la calidad de las comunidades usando atributos

Se compara la calidad de las comunidades detectadas en términos de densidad (ver Ecuación 3.2) y entropía (ver Ecuación 3.24) con el fin de evaluar qué tan conectados están los nodos y qué tan parecidos son los nodos entre ellos. En esta sección se muestran las diferencias entre el algoritmo GN y la mejora de las comunidades haciendo uso de atributos con la medida $Q_A$. Cabe destacar que la comparación con el estado del arte será descrita en el Capítulo 6.

Tabla 5.7: Redes sociales con atributos usadas en los experimentos de mejora de comunidades

| | $\|N\|$ | $\|E\|$ | $\|A\|$ | $\|C\|$ |
|---|---|---|---|---|
| Adolescentes | 50 | 113 | 4 | - |
| Football | 115 | 613 | 12 | 12 |
| Twitter | 145 | 7642 | 1185 | 17 |
| Facebook | 347 | 5038 | 224 | 24 |
| Pol.Blogs | 1490 | 19091 | 9 | 2 |

#### 5.3.2.1 Conjuntos de redes sociales

Las redes sociales que son usadas en este experimento incluyen la red de adolescentes y redes más grandes con comunidades reales para comparar su efectividad. Las características de estos conjuntos de datos los podemos observar en la Tabla 5.7 y su descripción se muestra a continuación:

- **Red de adolescentes** (*Teenage Friends and Lifestyle Network*) como ya se ha explicado, esta red cuenta con 50 nodos y 4 atributos que representan hábitos de adolescentes. Por el tamaño de ésta, se puede observar detalladamente el comportamiento de los algoritmos a evaluar.

- **Red de football [40]**[1]. Es una red de futbol americano de los juegos en la división colegial IA durante la temporada regular en Otoño del 2000. Cuenta con 115 nodos y 613 aristas no dirigidas. Los atributos de cada nodo indican en qué conferencia han jugado. Es una red social clásica del estado del arte.

- **Twitter [64]**[2]. Ego-red e Twitter de datos públicos, los nodos representan a usuarios, los enlaces a la opción de 'seguir' o 'ser seguido' por lo que su representación estructural es un grafo dirigido y sus atributos representan a tópicos relacionados a temas de tendencia (*hashtags*). En esta sección se usa

[1]http://www-personal.umich.edu/ mejn/netdata/

[2]http://snap.stanford.edu/

un conjunto de datos con 145 nodos con una elevada cantidad de atributos que supera los mil.

- **Blogs de Política [1]**. La red de *Blogs* de política muestra la interacción en 2004 de usuarios de *blogs* donde se discutía política. Los autores agruparon manualmente esta red según, sus enlaces, en dos grupos: conservadores y liberales. Los atributos representan los nueve *blogs* en los que podían comentar los usuarios.

- **Facebook. [64]**[3]. Ego-red de Facebook anonimizada, donde los nodos representan a los usuarios, las aristas la amistad y los atributos la información del perfil. En esta sección se usa un conjunto de datos con 1045 nodos y 224 atributos.

#### 5.3.2.2 Resultados

Se comparan las redes sociales de: Adolescentes, Football, Twitter, Facebook y Blog Pol. descritas previamente con el algoritmo de detección de comunidades disjuntas GN y GN + $Q_A$ que considera los atributos.

En la Tabla 5.8, se observa la mejora dada por GN + $Q_A$. Se tiene la mejor entropía en todos los casos al considerar atributos con la medida de $Q_A$. La mejora de GN + $Q_A$ sobresale cuando se tienen muchos atributos como en Facebook y Twitter. Y la menor mejoría se presenta en las redes de Adolescentes y Blog Pol. con únicamente 4 y 2 atributos respectivamente.

La medida de calidad de comunidades $Q_A$ fue diseñada para mejorar la similitud de los nodos dentro de una comunidad, sin embargo ésta también tiene elementos de la estructura como el hecho de que la importancia de los atributos considere el

[3]http://snap.stanford.edu/

Tabla 5.8: Mejora de la entropía en las comunidades detectadas

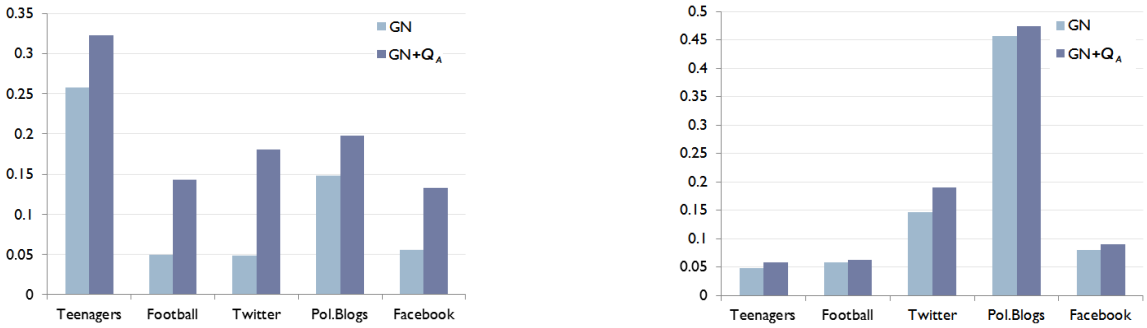
| | GN | GN + $Q_A$ |
|---|---|---|
| Adolescentes | 0.2578 | **0.3225** |
| Football | 0.0495 | **0.1432** |
| Twitter | 0.0482 | **0.1807** |
| Blogs Pol. | 0.1477 | **0.1981** |
| Facebook | 0.0555 | **0.1332** |

número de aristas y no sólo el número de nodos con ese atributo. Por ello, se esperaba que la densidad no fuese afectada significativamente; sin embargo encontramos que la densidad incluso mejora comparada con la del algoritmo base GN. En la Tabla 5.9 se muestran los resultados obtenidos del promedio de las densidades de las comunidades. Como se puede observar, la mejora de las comunidades con $Q_A$ tiene mejor densidad que las comunidades dadas originalmente por $GN$.

Tabla 5.9: Mejora de la densidad en las comunidades detectadas

| | GN | GN + $Q_A$ |
|---|---|---|
| Adolescentes | 0 .0482 | **0.0587** |
| Football | 0.0587 | **0.0619** |
| Twitter | 0.1471 | **0.1907** |
| Blogs Pol. | 0.4566 | **0.4743** |
| Facebook | 0.0802 | **0.0895** |

La Figura 5.5 muestra con mayor claridad la diferencia en entropía y densidad entre los algoritmos. Se observa que el incremento de entropía con respecto al algoritmo base de GN es alto y la densidad es equiparable a la original. La mejora con el uso de atributos incrementa la similitud entre nodos dentro de la comunidad. A pesar de que se pueden agregar nodos desconectados de la red la densidad no se ve afectada como se observa en la Figura 5.5(b).




(a) Entropía (b) Densidad

Figura 5.5: Comparación de la calidad de las comunidades detectadas

### 5.3.3 Experimento 3: comparación de comunidades reales con comunidades detectadas usando atributos

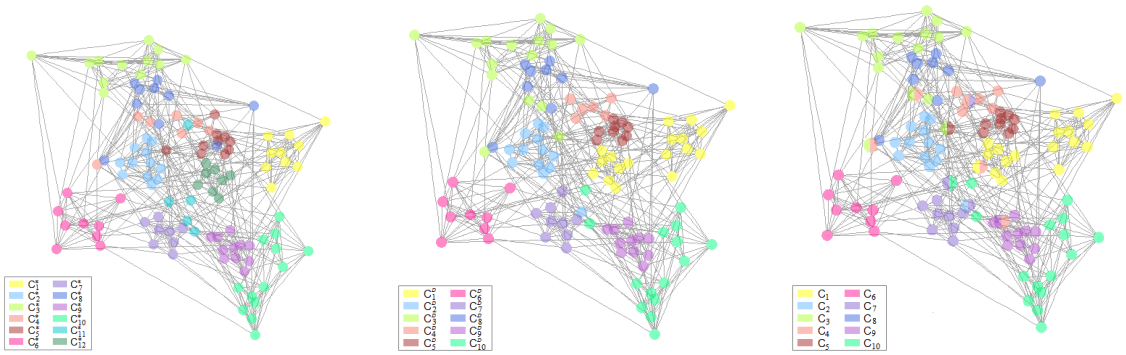
Se comparan las comunidades reales con las detectadas por GN + $Q_A$ y GN en los conjuntos de datos de: Football, Twitter, Facebook y Blog Pol. descritas en el experimento anterior. Para ello se hace uso de la medida $F_1$ (ver Ecuación 3.8) y de pureza (ver Ecuación 3.7).

La pureza evalúa la precisión de un algoritmo al compararlo con las comunidades esperadas. Los nodos aumentan su probabilidad de pertenecer a la comunidad correcta cuando se tienen comunidades sobrepuestas. En nuestro caso, la mejora con atributos genera comunidades sobrepuestas por lo que se esperaba un incremento en la pureza, ya que disminuyen los falsos positivos como se obtuvo en los experimentos que se muestran en la Tabla 5.10. Se observa que el mayor incremento es en las redes sociales de Facebook y Twitter que tienen más atributos.

A diferencia de la pureza, la medida $F_1$ considera tanto la precisión como la recuperación de las comunidades reales esperadas, de tal forma que nos da un balance entre la cantidad de nodos correctos y la cantidad de nodos esperados en esa comunidad. Como se puede observar en la Tabla 5.11, a diferencia de la pureza vemos que la

Tabla 5.10: Evaluación de pureza en las comunidades detectadas

| | GN | GN + $Q_A$ |
|---|---|---|
| Football | 0.9271 | **0.9514** |
| Twitter | 0.4000 | **0.7726** |
| Blogs Pol. | 0.5124 | **0.6503** |
| Facebook | 0.4224 | **0.9494** |



(a) Comunidades reales ($C^*$) (b) Comunidades disjuntas detectadas por GN ($C^D$) (c) Comunidades sobrepuestas de GN +$Q_A$ ($C$)

Figura 5.6: Comparación de comunidades reales en la red de football con las detectadas por GN y GN + $Q_A$

diferencia es muy poca, aunque favorece en la mayoría de los casos a las comunidades que usan atributos.

Tabla 5.11: Comparación de la medida $F_1$ en comunidades detectadas

| | GN | GN + $Q_A$ |
|---|---|---|
| Football | **0.766452** | 0.7537345 |
| Twitter | 0.042051 | **0.11328** |
| Blogs Pol. | 0.671318 | **0.718672** |
| Facebook | 0.153404 | **0.159691** |

Para al caso de la red de football podemos ver en la Figura 5.6 las comunidades reales y las obtenidas por GN y por GN + $Q_A$, donde los nodos que tienen un traslape se muestran en un mayor tamaño. Vemos que las comunidades reales $C^*_{11}$ and $c^*_{12}$ no fueron detectadas por GN y la mejora con atributos mezcla $c^D_1$ con $c^D_{12}$ en $c_1$, además de hacer más grade las comunidades más pequeñas como $c_4$. La precisión

Tabla 5.12: Redes sociales con atributos usadas en los experimentos de mejora de comunidades

| | $\|N\|$ | $\|E\|$ | $\|A\|$ | $\|C\|$ |
|---|---|---|---|---|
| Adolescentes | 50 | 113 | 4 | - |
| Pol.Mex. | 35 | 117 | 11 | 2 |
| Football | 115 | 613 | 12 | 12 |
| Blogs Pol. | 1490 | 19091 | 9 | 2 |

y recuperación permiten evaluar la comunidades obtenidas, la precisión fue muy alta tal que el promedio fue de 0.927, mostrando la correcta clasificación de los nodos. Sin embargo, la recuperación no fue tan alta cuyo promedio fue de 0.745 porque se ve afectada por la mezcla de dos de las comunidades.

### 5.3.4 Experimento 4: observaciones del balance entre atributos y estructura

En este experimento, a diferencia de los Experimentos 1, 2 y 3, haremos uso de la medida de calidad de comunidades $Q(C)$ que balancea atributos y estructura a la cual nos referimos como BAS.

#### 5.3.4.1 Conjuntos de redes sociales

Las redes sociales que son usadas en este experimento incluyen dos conjuntos de datos pequeños: red de adolescentes y la red de políticos mexicanos; y redes más grandes con comunidades reales para comparar su efectividad que serán la de *football* americano y la de *blogs* de política. Las características de estos conjuntos de datos los podemos observar en la Tabla 5.12 y su descripción se muestra a continuación:

La primera de las pequeñas es la de los adolescentes usada en la Sección 4.4.1 y la segunda es una red de políticos mexicanos descrita en la Sección 2.1, la cual cuenta con dos comunidades reales y cuyos atributos representan años en que tomaron su

primer cargo, así como sus profesiones.

Se realizaron experimentos en dos redes sociales reales de mayor tamaño. La primera de éstas es una red de jugadores de *football* [40][4] de la división de IA en la temporada regular de Otoño 2000. Ésta cuenta con 115 nodos y 613 aristas no dirigidas. Los atributos de cada nodo indican en que conferencias han jugado. La segunda red social es la producida por *blogs* de política [1], la cual muestra interacciones entre *blogueros* en 2004, los atributos corresponden a los *blogs* en los que comentaron y existen dos comunidades etiquetadas manualmente por los autores, basadas en las aristas que salen y las que entran.

#### 5.3.4.2 Resultados

En esta sección comparamos las comunidades detectadas por GN con la mejora dadas por $Q_A(C)$ y por BAS. Para evaluar las comunidades detectadas se usa densidad (ver Ecuación 3.2) y entropía (ver Ecuación 3.24).

Haciendo uso de las redes sociales de: Adolescentes, Políticos Mexicanos, Football y Blog Pol. se compararon las comunidades detectadas por GN + $BAS$ para $\alpha = \gamma = 0.5$. Como ya ha sido explicado, la densidad evalúa si las comunidades son densas o dispersas, entre más alto sea el valor existe una mayor conexión entre los miembros de esa comunidad. Por otro lado, la entropía muestra que tan similares son los nodos que pertenecen a una misma comunidad.

En la Figura 5.7 se observa la densidad y la entropía de las comunidades detectadas. Las que tienen contorno negro hacen referencia a las mejoras a comunidades considerando atributos ya sea con $Q_A(C)$ o $Q(C)$. Como se vio en los experimentos anteriores, la mejora con $Q_A$ daba un incremento a la entropía (similitud entre atributos), por lo que con BAS se busca una mejora en la densidad sin afectar la entropía

[4]http://www-personal.umich.edu/ mejn/netdata/

que da por sí sola $Q_A$. Se observa que hay un ligero aumento en la densidad y que la entropía se mantiene casi igual.

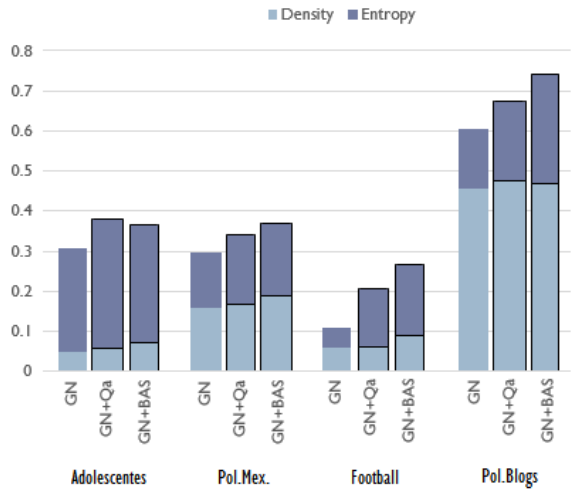



Figura 5.7: Comparación del índice omega

Como se observa en las Tablas 5.13 y 5.14, el uso de BAS comparado con $Q_A$ mejora tanto la densidad como la entropía (valores con letra inclinada), sin embargo aún no supera en algunos casos a los obtenidos en el estado del arte como se verá más adelante en esta tesis. Sin embargo, si acumulamos la densidad y la entopía como se observa en la Figura 5.7 la mejora de comunidades con atributos da mejores resultados en casi todos los casos.

Tabla 5.13: Comparación de la entropía de las comunidades detectadas

| | GN | GN + $Q_A$ | GN+BAS |
|---|---|---|---|
| Adolescentes | 0.257846 | **0.322523** | 0.293179 |
| Pol.Mex. | 0.137133 | 0.175264 | **0.183234** |
| Football | 0.049586 | 0.143297 | **0.175353** |
| Blogs Pol. | 0.147702 | 0.198172 | **0.273969** |

Tabla 5.14: Comparación de la densidad de las comunidades detectadas

| | GN | GN + $Q_A$ | GN+BAS |
|---|---|---|---|
| Adolescentes | 0.048270 | 0.057924 | **0.072405** |
| PolMex | 0.160256 | 0.166666 | **0.188034** |
| Football | 0.058727 | 0.061990 | **0.090701** |
| Blogs Pol. | 0.456629 | **0.474371** | 0.467801 |

### 5.3.5 Experimento 5: comparación de comunidades reales con comunidades usando atributos y estructura

Se compara GN + BAS con los algoritmos de comunidades sobrepuestas: Infomap, SLPA y CESNA, además como referencia se muestran también los resultados de GN y GN + $Q_A$. Las redes con comunidades reales evaluadas son Pol.Mex., Football y Blog Pol. descritas en el experimento anterior. Y las medidas usadas son $F_1$ (ver Ecuación 3.8) y de pureza (ver Ecuación 3.7) explicadas en la Sección 3.1.6).

La pureza evalúa la precisión en la recuperación de las comunidades detectadas con respecto a las comunidades reales o esperadas. A pesar de que con BAS se tiene un incremento sobre GN, no siempre se supera la mejora del uso único de atributos con $Q_A$ como se observa en la Tabla 5.15, sin embargo se observa un incremento significativo en la red de *blogs* de política (Blogs Pol.), por lo que las interacciones en esta red muestran ser más importantes que los atributos.

Tabla 5.15: Comparación de la pureza de las comunidades detectadas

| | GN | GN + $Q_A$ | G&N+BAS |
|---|---|---|---|
| Pol.Mex. | 0.394928 | **0.438406** | 0.427536 |
| Football | 0.927143 | 0.951429 | **0.972619** |
| Blogs Pol. | 0.650374 | 0.650374 | **0.922194** |

$F_1$ nos permite evaluar que tan buena ha sido la recuperación de las comunidades con respecto a las originales. Como se observa en la Tabla 5.16, cuando se hace uso de la mejora usando atributos ya sea con $Q_A$ o con BAS se tienen los mejores resultados. Cuando se hace uso del balance dado por BAS se obtiene comunidades más parecidas

a las originales para redes más grandes.

Tabla 5.16: Comparación de la medida F1 de las comunidades detectadas

| | GN | GN + $Q_A$ | GN+BAS |
|---|---|---|---|
| Pol.Mex | 0.533946 | **0.550843** | 0.545248 |
| Football | 0.766452 | 0.753734 | **0.757618** |
| Blogs Pol. | 0.671318 | 0.718672 | **0.757618** |

## 5.4 Conclusión

La medida de calidad propuesta considera propiedades importantes como: la importancia local y global de los atributos según la estructura, el grado de un nodo, el grado de un atributo y la asortatividad. Los atributos nos permiten evaluar la similitud estructural de los atributos con $Q_A$ y la similitud entre nodos, dando un balance a los atributos y estructura con BAS.

La mejora de comunidades considerando atributos se enfoca en conjuntos con comunidades previamente detectadas que sólo consideraban estructura y que se desea sean enriquecidas con las información de los atributos de los nodos. Por lo que se consideraron los atributos en adición a métodos basados en estructura. Esta metodología también es usada en la recomendación de enlaces, por lo que podría evaluarse la factibilidad de implementarlo para ese efecto. Cabe destacar que si algún nodo estaba mal clasificado puede ser detectado en la comunidad correcta, sin embargo no es eliminado de la primera, por lo que se plantea explotar esta opción en el futuro.

Uno de los problemas es el proceso estocástico para la selección del nodo que será agregado a la comunidad. La mejor opción sería seleccionar el nodo que aporte la mayor mejora, sin embargo esto aumenta la complejidad, por lo que se ha planteado un algoritmo que reduce este problema aunque no lo resuelve en su totalidad.

# Capítulo 6

# Método mixto: basado en modelo y distancia

En virtud de las hipótesis planteadas, se optó por un modelo mixto para la detección de comunidades, que permitiera obtener la similitud a comunidades reales dada por la detección de comunidades basada en modelos y la inmersión de una distancia. Esta distancia es usada como una expansión para la generación de comunidades sobrepuestas, en lugar de hacer directamente la detección de comunidades basada únicamente en distancia. La mezcla de los dos genera un modelo de dos fases evaluado en la Sección 6.2.

Por otro lado, no todos los atributos son importantes en el procesos de detección de comunidades. Por ello se plantea hacer una selección de los atributos más importantes como un pre-proceso al proceso de selección de comunidades. En la Sección 6.3 se aplica la selección de atributos, dando origen a las tres fases de la detección de comunidades. La selección de los atributos fue evaluada en nuestro modelo, así como en otro algoritmo que también detecta comunidades sobrepuestas, usando estructura y atributos.

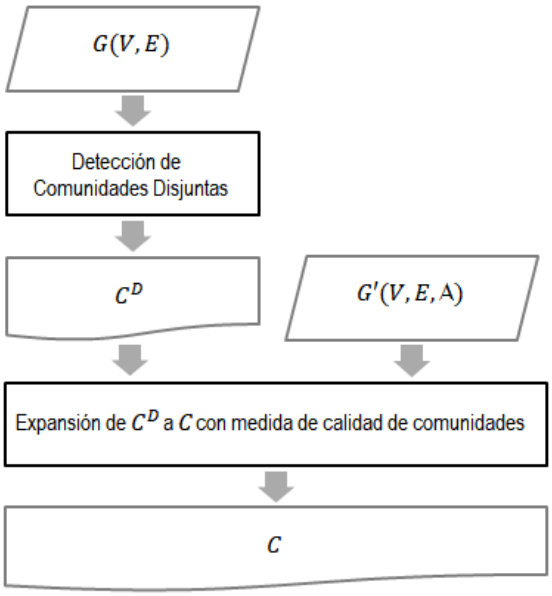



Figura 6.1: Proceso general para la generación de comunidades sobrepuestas a partir de comunidades disjuntas

## 6.1 Estrategia general

A partir de comunidades dadas por cualquier método se generan comunidades sobrepuestas. Las comunidades de las que se parten pueden o no haber considerado atributos. Para generar la sobreposición expandimos las comunidades con una medida de calidad $M_Q$ que podría ser la calidad de comunidades basada en atributos $Q_A(c)$ (dada por la Ecuación 5.4) o la calidad de comunidad con balance $Q(c)$ (dada por la Ecuación 5.5) que hemos llamado BAS.

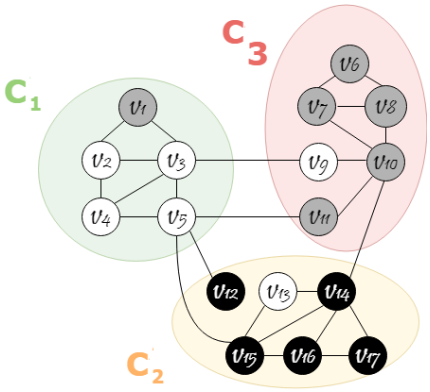
En la Figura 6.1 podemos ver que a una red social $G$ se le aplica cualquier método de detección de comunidades basado en estructura, lo cual generará comunidades disjuntas o sobrepuestas bajas $C^D$. La expansión de comunidades toma las comunidades previamente detectadas $C^D$ y considera el grafo con atributos $G'$. Con ambos se generan las comunidades sobrepuestas $C$ con un proceso de expansión que involucra alguna medida de calidad.

La generación de comunidades sobrepuestas expande las comunidades iniciales agregando aquellos nodos que mejoren la calidad de la comunidad según la medida de calidad de comunidades $M_Q$ usada. De tal forma que, si la calidad de una comunidad $c$ con un nuevo nodo $v$ incrementa la calidad original, entonces $c$ agrega el nodo $v$, como se describe en el Algoritmo 2.

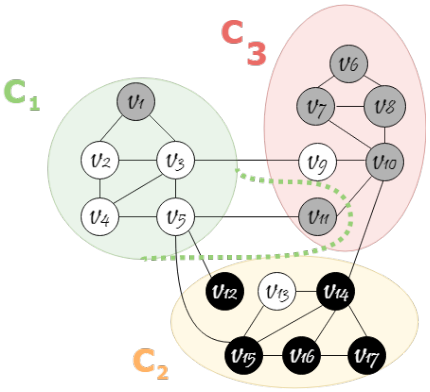
**Input**: $C^D$
**Output**: $C$
$C \leftarrow C^D$;
**for** $c \in C$ **do**
  $q \leftarrow M_Q(c)$;
  **for** $v \in V$ *tal que* $v \notin c$ **do**
    **if** $M_Q(c \cup v) > q$ **then**
      $c \leftarrow c \cup v$;
      $q \leftarrow M_Q(c \cup v)$;
    **end**
  **end**
**end**
**return** $C$;

**Algorithm 2:** Proceso de expansión de comunidades con una medida $M_Q(c)$

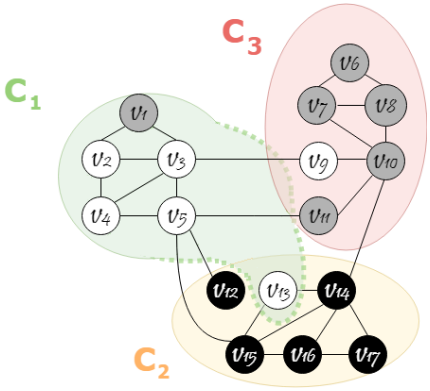
Supongamos que $M_Q$ está dada por la medida de calidad de comunidades basada en atributos $Q_A(C)$. Para el ejemplo de la Figura 6.2(a) la comunidad $c_1 = \{v_1, v_2, v_3, v_4, v_5\}$ busca integrar nodos externos. Supongamos que se intenta agregar al nodo $v_{11}$, la calidad sería $Q_A(c_1 \cup v_{11}) = 0.3043$ con la Ecuación 5.4, lo que decrementa la calidad de la comunidad original $Q_A(c_1)$, por lo que este nodo no es añadido a $c_1$ como se muestra la 6.2(b). Ahora, evaluemos la adición del vértice $v_{12}$ donde $Q_A(c_1 \cup v_{12}) = 0.2686$, lo cual es mucho menor que la original por lo que también se descarta. Cuando se evalúa el nodo $v_{13}$ (ver Figura 6.2(c)) la calidad es $Q_A(c_1 \cup v_{13}) = 0.3116$, lo cual incrementa el original de $Q_A(c_1) = 0.3065$ por lo que ahora $c_1 = \{v_1, v_2, v_3, v_4, v_5, v_{13}\}$. Se continúa el procedimiento con el resto de los nodos, así como para todas las comunidades, de tal forma que la expansión final de las comunidades se muestra en la Fig. 6.2(d) donde $c_1 = \{v_1, v_2, v_3, v_4, v_5, v_9, v_{13}\}$ con $Q_A(c_1) = 0.3161$ y $c_3 = \{v_1, v_6, v_7, v_8, v_9, v_{10}, v_{11}\}$



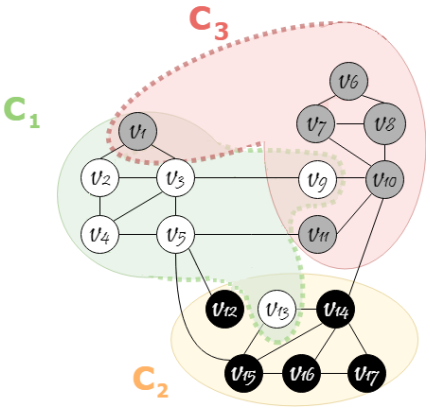

(a) Evaluación de comunidades donde $Q_A(c_1) = 0.3065$




(b) Evaluación de la expansión de $c_1$ con el vértice $v_{11}$ donde $Q_A(c_1^D \cup v_{11}) = 0.3043$




(c) Evaluación de la expansión de $c_1$ con el vértice $v_{13}$ donde $Q_A(c_1^D \cup v_{13}) = 0.3116$




(d) Evaluación de la expansión de $c_3$ donde $Q_A(c_3^D) = 0.5516$ mejora a $Q_A(c_3) = 0.6607$

Figura 6.2: Expansión con $Q_A$

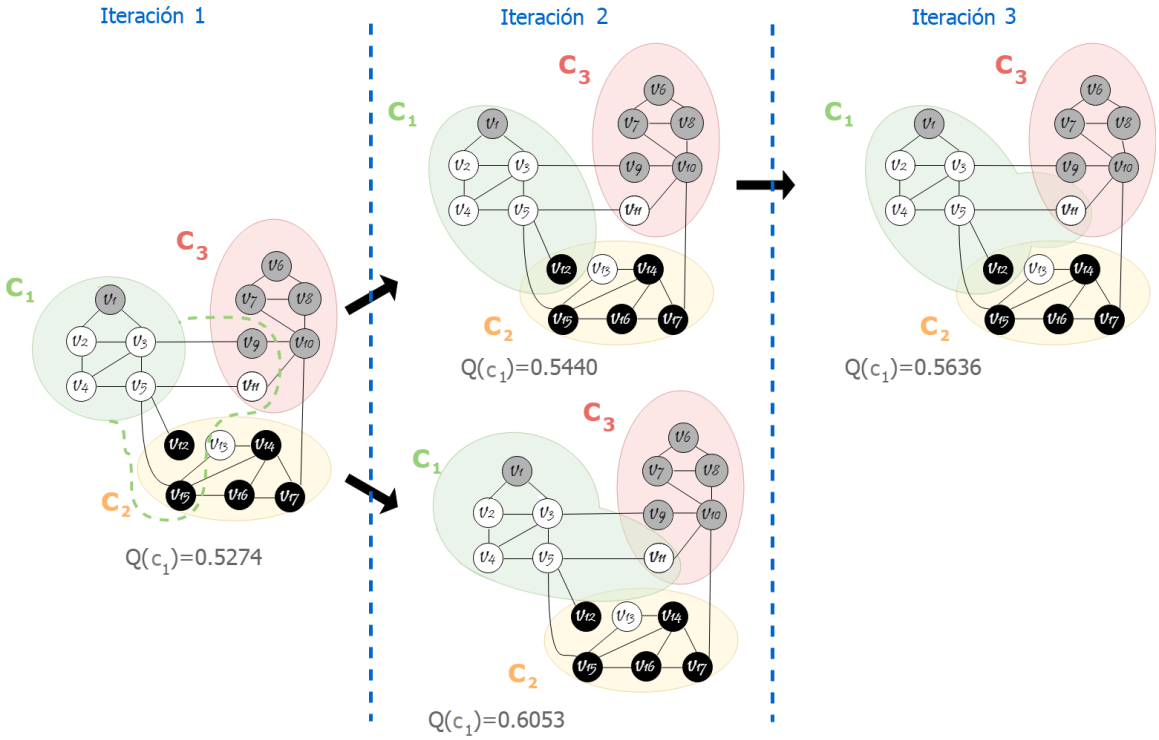



Figura 6.3: Ejemplo de afectación del proceso estocástico de expansión

con $Q_A(c_3) = 0.6607$.

La complejidad para la medida de calidad según atributos $Q_A(C)$ es $O(A(E+V))$. Ésta calcula una vez $W_{a_k}$ por cada atributo. Para cada comunidad calcula una vez $H_{a_k}$ con $O(AC)$ operaciones. Considerando que $W_{a_k}(G)$ y $H_{a_k}$ son constantes, nos permite reducir la complejidad cuadrática del coeficiente coseno, de tal forma que para $Q_A(C)$ la complejidad es $O(A(E+V))$ por cada comunidad $c_i$.

## Afectación del orden de selección de atributos

En el grafo de la Fig. 6.3 se tiene diecisiete nodos con tres diferentes atributos $A = \{blanco, gris, negro\}$ representados por esos tres colores. Supongamos que se cuentan con tres comunidades $C = \{c_1, c_2, c_3\}$ mostradas en la Etapa 1. Para expandir

$c_1$ con $M_Q(c_n) = Q(c_n)$ se buscan las comunidades que comparten atributos en $c_1$, en este caso, $c_2$ comparte los atributos *gris* y *blanco* y $c_3$ comparte los atributos *blanco* y *gris*, por lo tanto se evaluarían los nodos en esas comunidades. Para disminuir el proceso estocástico, se buscan los nodos más cercanos a la comunidad $c_1$ que serían los nodos $v_9, v_{11}, v_{12}, v_{13}$ como se ve en la Etapa 1 de la Figura 6.3. Sin embargo, aún con esta restricción la selección entre estos cuatro nodos sigue siendo aleatoria. En la Etapa 2 de la Figura 6.3 se presentan dos posibilidades: la primera evalúa primero los nodos en $c_2$ y la segunda evalúa primero los nodos en $c_3$.

La calidad de la comunidad de $c_1$ en la Etapa 1 es $Q(c_1) = 0.5274$ para $\alpha = \gamma = 0.5$ con $\phi(c_1) = \frac{4}{min(11,19)}$ y $Q_A(c_1) = 0.4185$. De las dos posibilidades en la Etapa 2, en el primer caso se calcula la calidad de la comunidad $c_1$ incluyendo el nodo $v_{12}$, la conductividad $\phi(c_1 \cup F) = 0.2727$ decrementa mostrando una mejor separación estructural entre grupos, ya que este nodo está más relacionado con $c_1$ que con $c_2$, mientras que la calidad dada por los atributos $Q_A(c_1 \cup F) = 0.3619$ empeora con respecto al original porque $v_{12}$ tiene un atributo que no fue considerado en $c_1$. Si damos la misma importancia a la estructura que a los atributos $\alpha = \gamma = 0.5$ entonces la calidad $Q(c_1 \cup F) = 0.5446$ mejorará la original de $c_1$ por lo que incluirá al nodo tal que $c_1 = c_1 \cup v_{12}$. Con $v_{12}$ incluido ahora evaluamos el nodo $v_{15}$ que no mejora la calidad de la comunidad. Después se evalúa el nodo $v_{11}$, el cual da mejor calidad por sus atributos $Q_A(c_1 \cup v_{11}) = 0.3773$ y la misma conductividad, por lo que la calidad de la comunidad es mejor $Q(c_1) = 0.5636$. Posteriormente evaluamos el nodo $v_9$ y, al no mejorar la calidad de la comunidad, se procede a buscar a los nodos vecinos de los nodos añadidas y se repite el procedimiento, de tal forma que $c_1 = \{v_1, v_2, v_3, v_4, v_5, v_{11}, v_{12}\}$.

En el segundo caso de las dos posibilidades mostradas en la Figura 6.3 de la Etapa 2, se calcula la calidad de la comunidad $c_1$ incluyendo $v_{11}$, donde la conductividad $\phi(c_1 \cup v_{11}) = 0.33$ tiene un decremento que marca una mejor separación de esta

comunidad, además tiene un incremento en la calidad dada por atributos ($Q_A(c_1 \cup N) = 0.4329$) porque incluye a nodos con *blanco*. Al considerar de igual importancia atributos y estructura, la calidad es $Q(c_1 \cup v_1 1) = 0.6053$. Posteriormente se evalúan los nodos restantes, pero no se tiene incremento en la calidad. Como se observa, al agregar un solo nodo ($v_{11}$); incrementó más la calidad que al agregar dos nodos; esta variación dada por el proceso estocástico podría ser reducida.

## 6.2 Detección de comunidades con método mixto

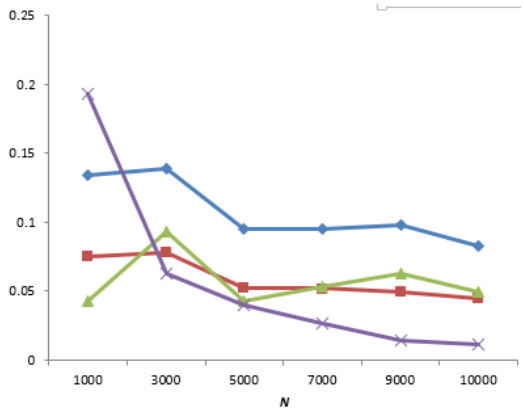
Se integra la detección de comunidades basada en modelo descrita en el Capítulo 4, llamada RMOCA, con la distancia que hace un balance entre la estructura y los atributos planteada en el Capítulo 5 con el nombre de BAS según la estrategía general del método mixto descrita en la sección anterior.

Primero se evaluará el desempeño de RMOCA+BAS en redes sintéticas dadas por *LFR benchmark* que nos permitirá evaluar la calidad de las comunidades sobrepuestas cuando no se presentan atributos. Posteriormente se hará una evaluación y comparación en redes sociales pequeñas para detallar algunos aspectos del modelo mixto propuesto. Finalmente se hará una comparación en redes sociales de mayor tamaño con algunos algoritmos del estado del arte.

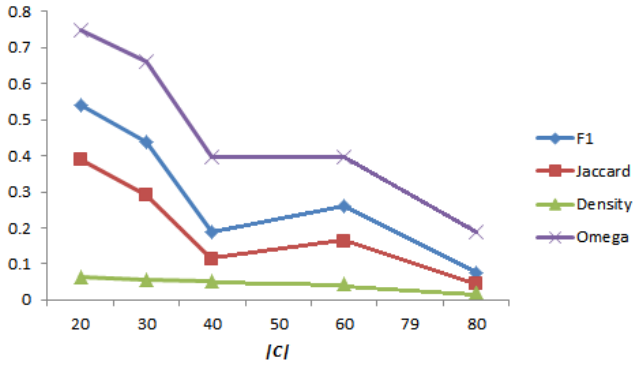
### 6.2.1 Experimento 1: desempeño de RMOCA+BAS en redes sintéticas

Para evaluar el desempeño de RMOCA+BAS se realizaron experimentos en redes sintéticas aleatorias del LFR benchmark[1][62]. Éste permite el análisis de redes artifi-

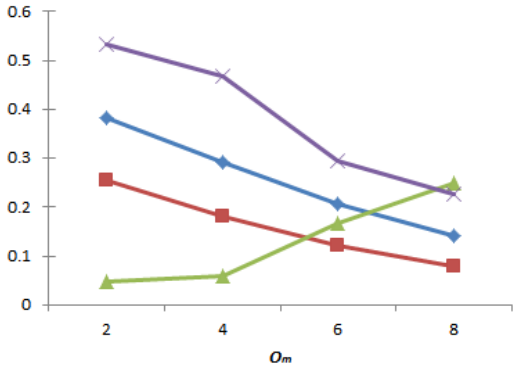
[1]https://sites.google.com/site/andrealancichinetti/files



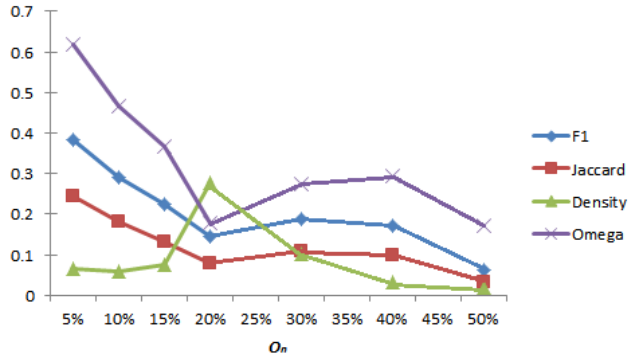

(a) Número de nodos



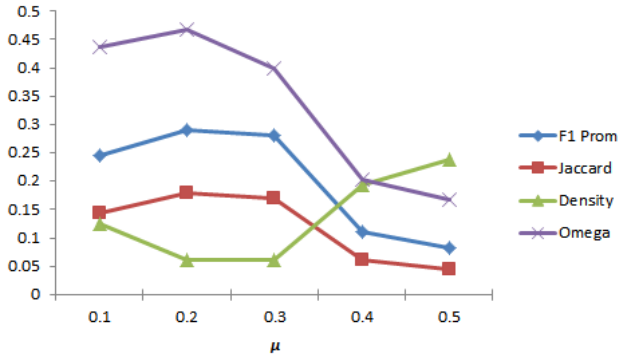

(b) Número de Comunidades




(c) Pertenecia de los nodos en traslape




(d) Fracción de los nodos en traslape




(e) Parámetro de mezcla

Figura 6.4: Comportamiento de RMOCA+BAS en el proceso de detección de comunidades sobrepuestas sin atributos en la redes de LFR variando $N$, $|C|$, $0_m$, $O_n$ y $\mu$

ciales construidas en comunidades sobrepuestas. LFR cuenta con heterogeneidad en la distribución de grados de nodo y de tamaños de comunidad.

#### 6.2.1.1 Configuración de *LFR Benchmark*

Los valores constantes en las redes de LFR fueron $n = 5000$ para el tamaño de la red con un grado promedio (*average degree*) de $\bar{k} = 20$. Para obtener el grado promedio deseado se fijó una máxima distribución de $k_{max} = 100$. El parámetro *mixing* $\mu$ es la fracción de aristas que están ente las comunidades tal que el rango varía de $0 \leq \mu \leq 1$, y fue fijado en 0.2. El tamaño de las comunidades era de 50 a 100 nodos. El grado de traslape es controlado por los parámetros $O_m$ y $O_n$; el primero es la pertencia de los nodos que están en un traslape $O_m = 4$; el segundo se refiere al número de nodos en traslape para lo que se usó un 10% de los nodos.

#### 6.2.1.2 Métricas de Evaluación

Usamos la medida F1, la similitud Jaccad y el Índice Omega para comparar las comunidades detectadas en relación con las del *benchmark* F1 y Jaccard permiten evaluar la relación de precisión y recuperación, el índice omega nos ayuda para evaluar por pares el traslape y se usa la densidad para evaluar la cantidad de aristas dentro de las comunidades. Cabe destacar que las redes generadas por LFR no tiene atributos, por lo que se añade un atributo a todos los nodos de tal forma que no influya en el proceso de detección de comunidades sobrepuestas. Para evaluar RMOCA+BAS se varía el número de nodos y comunidades, así como los parámetros de traslape como se verá en la Fig. 6.4.

#### 6.2.1.3 Resultados de la Evaluación

El número de nodos se varió de 1000 a 9000. La medida F1 y Jaccard tienen un decremento al aumentar el número de nodos. La densidad en general es muy baja dado que se tuvieron hasta 1264 comunidades. El índice omega también muestra un decremento al aumentar el tamaño de la red como se observa en la Figura 6.4(a). Con $n = 1000$, se variaron los parámetros del mínimo y máximo número de nodos por comunidad para evaluar el desempeño al variar el número de comunidades. Se observó que RMOCA+BAS tiene mejores resultados cuando hay menos comunidades para el mismo número de nodos (Figura 6.4(b)) porque al incrementar las dimensiones de las matrices $M$ y $X$ el modelo de regresiones decrementa su precisión en el proceso de optimización.

Los parámetros de traslape $O_m$ y $O_n$ se variaron de 2 a 8 y de 5% a 50%, respectivamente. En ambos casos se obtuvo un promedio de 70.18 comunidades. Cuando se decrementa el número de comunidades se tiene una mejor densidad aunque la precisión decrementa (ver Figura 6.4(c)). Cuando el traslape es bajo (5% de los nodos) la precisión aumenta pero disminuye la densidad ya que el parámetro de mezcla permanece en 0.2 (pocas aristas entre comunidades) pero el número de nodos en traslape aumenta de tal forma que LFR balancea estos elementos y se tiene un decremento en la densidad, se observa que cerca del 20% se tiene un balance entre la precisión y la densidad (Figura 6.4(d)). Cuando el parámetro de mezcla es alto, es difícil obtener la comunidades ya que no existe una clara separación de las comunidades por lo que la precisión es muy baja (vea F1, Jaccard and Omega de la Figura 6.4(e)) pero la densidad aumenta porque aumentan el número de aristas.

### 6.2.2 Experimento 2: análisis de comunidades detectadas en redes sociales reales

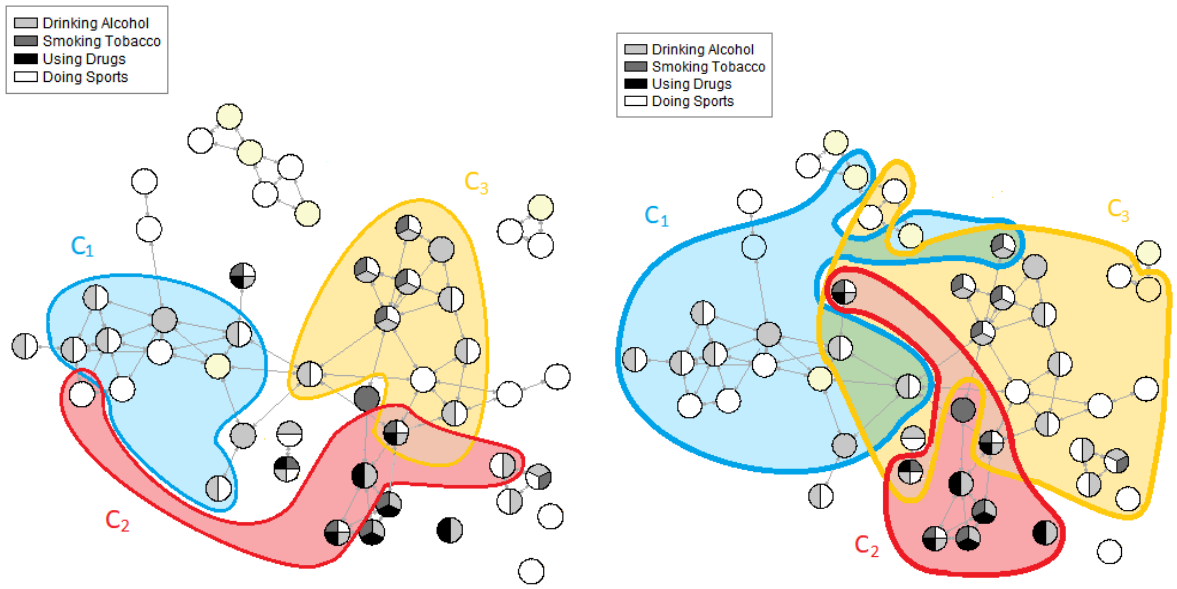
A continuación se comparan las comunidades detectadas por RMOCA+BAS con el algoritmo de CESNA. Ambos consideran la estructura de la red dada por el grafo, así como los atributos. En ambos casos se obtienen comunidades sobrepuestas. Para evaluar visualmente las comunidades se analizarán las comunidades de dos redes: la red de adolescentes y la red de blogs de política:

- **Red de Adolescentes** (*Teenage Friends and Lifestyle Network*) como ya se ha explicado esta red cuenta con 50 nodos y 4 atributos que representan hábitos de adolescentes. Por el tamaño de ésta se puede observar detalladamente el comportamiento de los algoritmos a evaluar.
- **Political Blogs** [**1**]. La red de *Blogs* de política muestra la interacción en 2004 de usuarios de *blogs* (atributos) donde se discutía política, cuenta con dos grupos: liberales y conservadores.

#### 6.2.2.1 Red de Adolescentes

La Figura 6.5 muestra la red social de adolescentes en donde se solicitaron tres comunidades a los algoritmos de CESNA y de RMOCA+BAS. Como se observa RMOCA+BAS tiene una mayor cobertura de los nodos, además de que encuentra más traslapes.

Se observa que en la $c_2$ mostrada en rojo los atributos no solo se encuentran cerca estructuralmente sino que también comparten los mismos atributos, especialmente el del consumo de mariguana que muy pocos nodos poseen. También se observa que algunos nodos disjuntos de la red fueron agrupados en alguna de las comunidades ya que la expansión no solo es estructural.



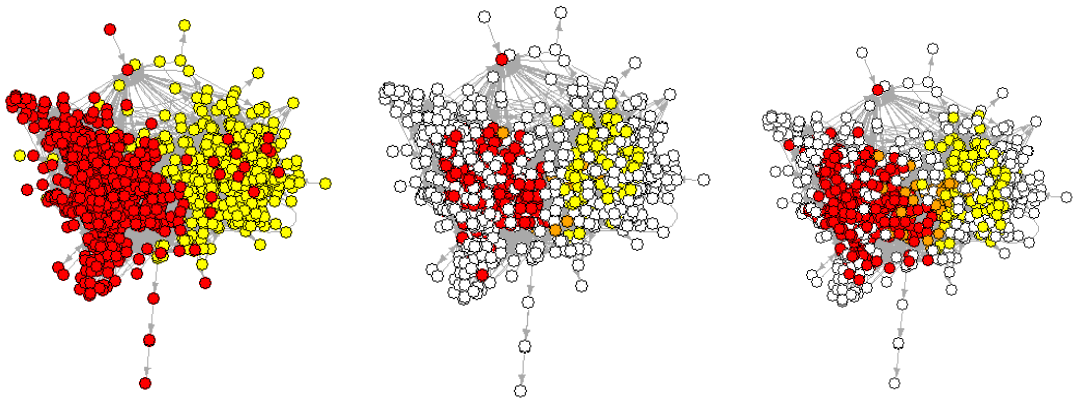

(a) CESNA (b) RMOCA+BAS

Figura 6.5: Red de adolescentes donde se solicitaron 3 comunidades que se visualizan en rojo, amarillo y azul.

#### 6.2.2.2 Blogs de Política

En la Figura 6.6 se observa las comunidades esperadas, las comunidades representan a las personas liberales y conservadores que opinaron en *blogs* de política en el 2004. Nuevamente, se aprecia que RMOCA+BAS tiene una mayor cobertura de los nodos. En naranja podemos observar aquellos nodos que pertenecen a ambas comunidades según CESNA y RMOCA+BAS, es decir aquellos nodos que se encuentran en un traslape (nodos en naranja).

En la Tabla 6.1 se evaluó la entropía y densidad de las comunidades de la Figura 6.6. Previa a la expansión dada por BAS, las comunidades de QMUCA están disjuntas en su mayoría por lo que se observa la entopía y densidad de las comunidades con y sin traslape de tal forma que nos permite evaluar el incremento dado por BAS. Si de $c_1$ y $c_2$ (comunidades en rojo y amarillo) se quitan los nodos del traslape tanto la



(a) Comunidades por auto-etiquetado (b) Comunidades detectadas con CESNA (c) Comunidades detectadas con RMOCA+BAS

Figura 6.6: Las comunidades representan a conservadores y liberales en el *blog* de política estudiado

entropía como la densidad disminuyen. Además la comunidad $c_2$ fue la que tuvo una mayor entropía, mientras que la comunidad $c_1$ fue la de mejor densidad.

Tabla 6.1: Entropía y densidad en RMOCA

| $c$ | Entropía | Densidad |
|---|---|---|
| $c_1 \cup c_2$ | 0.1171 | 0.4447 |
| $c_1 \cup c_2 - (c_1 \cap c_2)$ | 0.0989 | 0.3595 |
| $c_1$ | 0.2313 | 0.4714 |
| $c_1 - (c_1 \cap c_2)$ | 0.1965 | 0.3595 |
| $c_2$ | 0.2559 | 0.418 |
| $c_2 - (c_1 \cap c_2)$ | 0.2308 | 0.3091 |
| $c_1 \cap c_2$ | 0.1252 | 0.0197 |

### 6.2.3 Experimento 3: comparación del uso de BAS en GN y en RMOCA

Primero se muestra una comparación entre las propuestas de esta tesis que involucran la integración de la medida BAS a métodos tradicionales así como la integración al modelo propuesto de RMOCA.

En las Tablas 6.2 y 6.3 se muestra la entropía y densidad obtenida. Se comparan el tradicional GN, GN con la medida $Q_A$, GN con la medida $BAS$, el modelo RMOCA y la conjunción del modelo mixto RMOCA+BAS en dos redes pequeñas (Adolescentes y Pol.Mex.) y dos redes grandes típicas del estado del arte (Footbal y Blogs Pol.).

Como se observa en la Tabla 6.2 el modelo mixto de RMOCA+BAS, que hace uso de los atributos desde la primera fase, permite una mejor entropía, es decir encuentra comunidades con nodos más parecidos.

Tabla 6.2: Comparación de la entropía entre las fases propuestas

| | GN | GN+$Q_A$ | GN+BAS | RMOCA | RMOCA+BAS |
|---|---|---|---|---|---|
| Adolescentes | 0.25784601 | 0.32252361 | 0.29317913 | 0.26591868 | **0.4840043**8 |
| PolMex | 0.13713341 | 0.17526483 | 0.18323481 | 0.18625127 | **0.19502254** |
| Football | 0.04958609 | 0.14329726 | 0.17535376 | 0.21907635 | **0.22638684** |
| Blogs Pol. | 0.14770272 | 0.19817217 | 0.27396904 | 0.1165593 | **0.46225625** |

En la Tabla 6.3 el modelo mixto obtiene buenos resultados, sin embargo afecta la densidad cuando se tiene redes grandes. Es por ello que en la siguiente sección se hace una evaluación del modelo mixto RMOCA+BAS con otros algoritmos del estado del arte en redes reales de mayor tamaño.

Tabla 6.3: Comparación de la densidad entre las fases propuestas

| | GN | GN+$Q_A$ | GN+BAS | RMOCA | BAS+RMOCA |
|---|---|---|---|---|---|
| Adolescentes | 0.04827031 | 0.05792438 | 0.07240547 | 0.25663717 | **0.28097345** |
| Pol.Mex. | 0.16025641 | 0.16666667 | 0.18803419 | 0.2008547 | **0.23717949** |
| Football | 0.05872757 | 0.06199021 | **0.09070147** | 0.08939641 | 0.08115824 |
| Blogs Pol. | 0.45662917 | **0.47437178** | 0.46780044 | 0.44377563 | 0.13670574 |

### 6.2.4 Experimento 4: comparación con algoritmos del estado del arte

En este experimento se compara el modelo mixto BAS+RMOCA con algoritmos del estado del arte en redes sociales reales como Facebook, Google Plus y Twitter.

#### 6.2.4.1 Algoritmos base

Se evalúa el desempeño de RMOCA+BAS comparándolo con otros algoritmos de detección de comunidades sobrepuestas. Infomap [89] y SLPA [116] consideran la estructura del grafo y CESNA [122] usa tanto la estructura como los atributos. Se evalúa la entropía y la densidad para evaluar las comunidades detectadas.

#### 6.2.4.2 Conjuntos de redes sociales

Además de las redes con atributos Football y Blogs de Política, se evaluaron redes sociales en línea, de las que se obtienen ego-redes de Facebook, Twitter, y Google Plus[2] cuyas características se encuentran en la Tabla 6.4 y Figura 6.7 y que se describen a continuación:

- **Ego-redes de Facebook.** Comprende usuarios anonimizados, con aristas no dirigidas que representan la amistad y cuyos atributos son los perfiles. Se hace uso de ocho ego-redes con un total de 4042 nodos, 2191 atributos y 170342 aristas.
- **Ego-redes de Twitter.** Círculos de Twitter de datos públicos donde el usuario es un nodo, *seguir* representa las aristas dirigidas y los temas en que participan los usuarios son los atributos. Los temas están dados por los temas de tendencia (*hashtags*) y menciones. Se valúan siete ego-redes con un total de 1109 nodos, 6373 atributos y 43494 aristas.
- **Ego-redes de Google Plus.** Los datos de Google Plus son de usuarios con aristas que representan *seguir* a otro usuario y los atributos son elementos en el perfil. Se usaron nueve ego-redes con un total de 11035 nodos, 6263 atributos, and 608376 aristas.

[2]http://snap.stanford.edu/

Tabla 6.4: Características de las redes sociales con atributos de Facebook, Twitter y Google Plus

Facebook

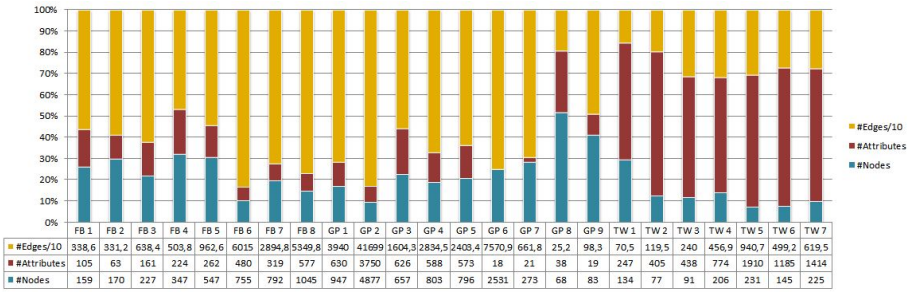
| | Nodos | Atributos | Aristas |
|---|---|---|---|
| FB 1 | 159 | 105 | 3386 |
| FB 2 | 170 | 63 | 3312 |
| FB 3 | 227 | 161 | 6384 |
| FB 4 | 347 | 224 | 5038 |
| FB 5 | 547 | 262 | 9626 |
| FB 6 | 755 | 480 | 60150 |
| FB 7 | 792 | 319 | 28948 |
| FB 8 | 1045 | 577 | 53498 |

Twitter

| | Nodos | Atributos | Aristas |
|---|---|---|---|
| TW 1 | 134 | 247 | 990 |
| TW 2 | 77 | 405 | 1868 |
| TW 3 | 91 | 438 | 3234 |
| TW 4 | 206 | 774 | 7024 |
| TW 5 | 231 | 1910 | 14672 |
| TW 6 | 145 | 1185 | 7642 |
| TW 7 | 225 | 1414 | 8064 |

Google Plus

| | Nodos | Atributos | Aristas |
|---|---|---|---|
| GP 1 | 947 | 630 | 39400 |
| GP 2 | 4877 | 3750 | 416992 |
| GP 3 | 657 | 626 | 16043 |
| GP 4 | 803 | 588 | 28345 |
| GP 5 | 796 | 573 | 24034 |
| GP 6 | 2531 | 18 | 75709 |
| GP 7 | 273 | 21 | 6618 |
| GP 8 | 68 | 38 | 252 |
| GP 9 | 83 | 19 | 983 |




| | FB 1 | FB 2 | FB 3 | FB 4 | FB 5 | FB 6 | FB 7 | FB 8 | GP 1 | GP 2 | GP 3 | GP 4 | GP 5 | GP 6 | GP 7 | GP 8 | GP 9 | TW 1 | TW 2 | TW 3 | TW 4 | TW 5 | TW 6 | TW 7 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| #Edges/10 | 338,6 | 331,2 | 638,4 | 503,8 | 962,6 | 6015 | 2894,8 | 5349,8 | 3940 | 41699 | 1604,3 | 2834,5 | 2403,4 | 7570,9 | 661,8 | 25,2 | 98,3 | 70,5 | 119,5 | 240 | 456,9 | 940,7 | 499,2 | 619,5 |
| #Attributes | 105 | 63 | 161 | 224 | 262 | 480 | 319 | 577 | 630 | 3750 | 626 | 588 | 573 | 18 | 21 | 38 | 19 | 247 | 405 | 438 | 774 | 1910 | 1185 | 1414 |
| #Nodes | 159 | 170 | 227 | 347 | 547 | 755 | 792 | 1045 | 947 | 4877 | 657 | 803 | 796 | 2531 | 273 | 68 | 83 | 134 | 77 | 91 | 206 | 231 | 145 | 225 |

Figura 6.7: Proporción de Nodos, Atributos y Aristas de las redes de Facebook, Google Plus y Twitter

### 6.2.4.3 Resultados de densidad y entropía

En la Figuraa 6.8 se muestra la densidad y entropía de las comunidades detectadas por diversos métodos. Se muestran juntas dado que se espera que el resultado sean comunidades con muchas conexiones internas y con nodos similares. Como se observa RMOCA+BAS supera fuertemente a los algoritmos del estado del arte, incluso a las comunidades reales. En el caso de la red de Políticos la entropía es mucho mejor que otros sin embargo la densidad se vio afectada como se veía en la Tabla 6.3, sin embargo al considerar los dos aspectos juntos supera a otros métodos.

La entropía de RMOCA+BAS es mejor en todos los conjuntos de Facebook y Twitter evaluados como se ve en la Tabla 6.5. Por otro lado, la densidad es muy buena comparado con otros algoritmos, Infomap que solo considera estructura obtiene mejores resultados en algunos casos pero nuestro modelo tiene valores próximos a éste.

RMOCA+BAS no es superior en los conjuntos de Google+. La entropía fue superior con el algoritmo de Infomap, mientras que la densidad favoreció a SLPA. Al observar detalladamente estos conjuntos de datos en la Tabla 6.5, Infomap obtiene mejores de entropía en conjuntos con muchos atributos en proporción al número de nodos, de hecho se tienen casi el mismo número de nodos que de atributos, dando mejor entropía sobre todo en conjuntos como GP7, GP8 y GP9. Por otro lado, SLPA tiene mejores resultados en la densidad en grafos poco densos como GP3, GP4 y GP5, mientras que RMOCA+BAS tiene mejores resultados en conjuntos como GP1, GP2, GP7 y GP9 donde se tiene más aristas ya que la matriz $M$ del modelo RMOCA tiene más elementos con los cuales aproximar.

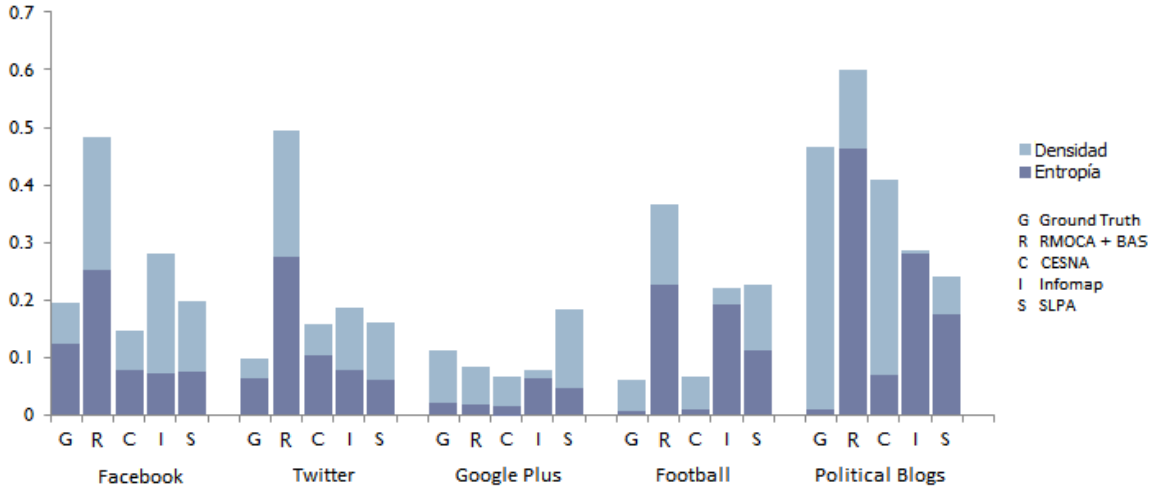



Figura 6.8: Densidad y Entropía de RMOCA+BAS comparado con otros algoritmos del estado del arte

#### 6.2.4.4 Resultados de $F_1$ y Jaccad

Dado que esos conjuntos de datos cuentan con comunidades reales se evalúa que tan parecidas son las comunidades a las reales, para ello se ocupó la medida de $F_1$ y la similitud Jaccard descritas en la Sección 3.1.6. RMOCA+BAS obtiene resultados similares a CESNA en la evaluación de los conjuntos individuales de Facebook,

Tabla 6.5: Características de las redes sociales con atributos de Facebook, Twitter y Google Plus

| | Entropía | | | | Densidad | | | |
|---|---|---|---|---|---|---|---|---|
| | RMOCA+BAS | CESNA | Infomap | SLPA | RMOCA+BAS | CESNA | Infomap | SLPA |
| FB 1 | **0.170178** | 0.107964 | 0.125121 | 0.124906 | **0.273243** | 0.121340 | 0.113039 | 0.242912 |
| FB 2 | **0.338055** | 0.150958 | 0.142607 | 0.139012 | **0.238325** | 0.054693 | 0.114357 | 0.196014 |
| FB 3 | **0.294334** | 0.119450 | 0.115325 | 0.106411 | 0.216035 | 0.059680 | **0.332185** | 0.196053 |
| FB 4 | **0.535065** | 0.073496 | 0.000032 | 0.064817 | **0.459618** | 0.040376 | 0.046069 | 0.052711 |
| FB 5 | **0.358014** | 0.055524 | 0.062031 | 0.050442 | 0.098243 | 0.032230 | **0.195387** | 0.062594 |
| FB 6 | **0.119859** | 0.041083 | 0.049969 | 0.041457 | 0.240977 | 0.090964 | **0.333211** | 0.087681 |
| FB 7 | **0.146843** | 0.046068 | 0.050221 | 0.044143 | 0.135486 | 0.054306 | **0.199729** | 0.073725 |
| FB 8 | **0.067166** | 0.037328 | 0.045727 | 0.038299 | 0.168874 | 0.101465 | **0.330654** | 0.059633 |
| TW 1 | **0.121588** | 0.041451 | 0.066641 | 0.002180 | 0.136716 | 0.054337 | 0.039778 | **0.247518** |
| TW 2 | **0.309971** | 0.124176 | 0.115349 | 0.114719 | 0.218577 | 0.050590 | **0.243515** | 0.000000 |
| TW 3 | **0.387624** | 0.195753 | 0.117638 | 0.116330 | **0.303969** | 0.087286 | 0.184577 | 0.000000 |
| TW 4 | **0.303019** | 0.107275 | 0.076988 | 0.072791 | **0.27360** | 0.052497 | 0.109822 | 0.249781 |
| TW 5 | **0.253714** | 0.076028 | 0.030117 | 0.000000 | **0.190984** | 0.033579 | 0.061815 | 0.000000 |
| TW 6 | **0.390674** | 0.115549 | 0.106206 | 0.084650 | **0.191807** | 0.029469 | 0.06675 | 0.000000 |
| TW 7 | **0.167961** | 0.069340 | 0.047352 | 0.046925 | **0.220314** | 0.074539 | 0.036333 | 0.190024 |
| G1 | 0.041077 | 0.047456 | **0.053317** | 0.023747 | **0.044495** | 0.037997 | 0.007093 | 0.000000 |
| G2 | 0.002698 | 0.008716 | 0.001716 | **0.007393** | **0.065168** | 0.001938 | 0.000909 | 0.037947 |
| G3 | 0.008178 | 0.018781 | **0.065891** | 0.029926 | 0.081406 | 0.172832 | 0.004898 | **0.296017** |
| G4 | 0.006160 | 0.009733 | 0.031770 | **0.026647** | 0.065073 | 0.019768 | 0.009173 | **0.249912** |
| G5 | 0.019571 | 0.017502 | 0.086247 | 0.027264 | 0.049162 | 0.017242 | 0.002686 | **0.310158** |
| G6 | 0.005255 | **0.008059** | 0.007899 | 0.006238 | 0.026769 | **0.012031** | 0.011698 | 0.111752 |
| G7 | 0.010500 | 0.008407 | **0.091185** | 0.049666 | 0.082049 | **0.097664** | 0.005628 | 0.000000 |
| G8 | 0.029209 | 0.023951 | 0.103141 | **0.121997** | 0.107143 | 0.064048 | 0.040584 | **0.242063** |
| G9 | 0.043292 | 0.004746 | **0.135164** | 0.127386 | **0.067141** | 0.007694 | 0.044881 | 0.000000 |

Twitter y Google Plus como se puede observar en la Tabla 6.6. Como se observa Infomap y SLPA que tuvieron buena entropía y densidad en Google+, no tienen buena aproximación a los conjuntos reales comparados con CESNA y RMOCA+BAS ya que ambos están basados en modelos y en el caso de RMOCA+BAS también está basado en distancia lo cual mejora la entropía y densidad que pudiese CESNA. El conjunto de Facebook donde ninguno de estos dos algoritmos es el mejor es el conjunto FB7, donde SLPA tiene mejor resultado ya que este conjunto tiene más aristas en proporción al número de nodos y atributos.

La Figura 6.9 muestra el promedio de $F_1$ y Jaccard para los conjuntos de datos de Facebook, Twitter y Google Plus. El modelo de CESNA considera un traslape denso dado un estudio previo de estas redes sociales por lo que tiene resultados superiores a otros algoritmos del estado del arte en estas medidas que evalúan la similitud a grupos reales. Como se observa en la Figura 6.9, RMOCA+BAS tiene muy buenos resultados, incluso superiores a los de CESNA a excepción de la red de Football donde

Tabla 6.6: Comparación de la medida $F_1$ y Jaccard para evaluar la similitud de las comunidades detectadas con las reales

| | F1-measure | | | | Jaccard | | | |
|---|---|---|---|---|---|---|---|---|
| | RMOCA+BAS | CESNA | Infomap | SLPA | RMOCA+BAS | CESNA | Infomap | SLPA |
| FB 1 | **0.603238** | 0.602061 | 0.183101 | 0.457807 | **0.516128** | 0.482167 | 0.152067 | 0.361757 |
| FB 2 | **0.535000** | 0.326727 | 0.398709 | 0.366133 | **0.395000** | 0.205922 | 0.287551 | 0.281541 |
| FB 3 | 0.485020 | **0.518575** | 0.27737 | 0.322906 | 0.358259 | **0.378369** | 0.233528 | 0.263523 |
| FB 4 | **0.316537** | 0.283223 | 0.242863 | 0.282346 | **0.211698** | 0.180781 | 0.173627 | 0.183267 |
| FB 5 | 0.142939 | **0.202643** | 0.056767 | 0.099332 | 0.080438 | **0.118073** | 0.032851 | 0.057266 |
| FB 6 | **0.329376** | 0.275107 | 0.071474 | 0.201282 | **0.242596** | 0.188172 | 0.047635 | 0.154115 |
| FB 7 | 0.381275 | 0.423544 | 0.167974 | **0.496409** | 0.276346 | 0.319984 | 0.127174 | **0.403935** |
| FB 8 | **0.355512** | 0.328399 | 0.207683 | 0.294081 | **0.246712** | 0.234862 | 0.160407 | 0.225973 |
| TW 1 | **0.345544** | 0.335478 | 0.264311 | 0.158691 | 0.227698 | **0.249292** | 0.191721 | 0.102881 |
| TW 2 | **0.469554** | 0.374043 | 0.120463 | 0.156939 | **0.345506** | 0.258068 | 0.116621 | 0.104973 |
| TW 3 | 0.368119 | **0.379735** | 0.286135 | 0.123737 | 0.247594 | **0.347369** | 0.193151 | 0.074481 |
| TW 4 | **0.291601** | 0.164888 | 0.042999 | 0.071204 | **0.178725** | 0.174442 | 0.024322 | 0.042308 |
| TW 5 | 0.297017 | **0.319782** | 0.291277 | 0.000000 | 0.191242 | **0.200521** | 0.191202 | 0.000000 |
| TW 6 | **0.470067** | 0.33269 | 0.293236 | 0.077666 | **0.324094** | 0.209629 | 0.186908 | 0.047716 |
| TW 7 | 0.226981 | **0.304882** | 0.187531 | 0.129605 | 0.143295 | **0.207984** | 0.115611 | 0.082798 |
| G1 | **0.376456** | 0.31030 | 0.118572 | 0.181365 | **0.272502** | 0.192699 | 0.075047 | 0.121553 |
| G2 | **0.223090** | 0.16321 | 0.011715 | 0.027368 | **0.134493** | 0.091610 | 0.000000 | 0.015810 |
| G3 | 0.128205 | 0.10638 | 0.025878 | **0.265837** | 0.068493 | 0.056180 | 0.015129 | **0.195085** |
| G4 | **0.552384** | 0.09344 | 0.247834 | 0.027368 | **0.381593** | 0.049010 | 0.140762 | 0.015810 |
| G5 | **0.260514** | 0.19280 | 0.032400 | 0.130294 | 0.163297 | **0.173332** | 0.047914 | 0.074627 |
| G6 | 0.292584 | **0.30902** | 0.217376 | 0.186102 | 0.196466 | **0.210200** | 0.154540 | 0.152048 |
| G7 | **0.721739** | 0.54737 | 0.052293 | 0.691170 | 0.564626 | 0.376812 | 0.028687 | **0.535448** |
| G8 | **0.388889** | 0.21500 | 0.121109 | 0.203333 | **0.241379** | 0.128182 | 0.072456 | 0.120790 |
| G9 | **0.308083** | 0.14286 | 0.083916 | 0.023810 | 0.029911 | **0.076923** | 0.055128 | 0.012048 |

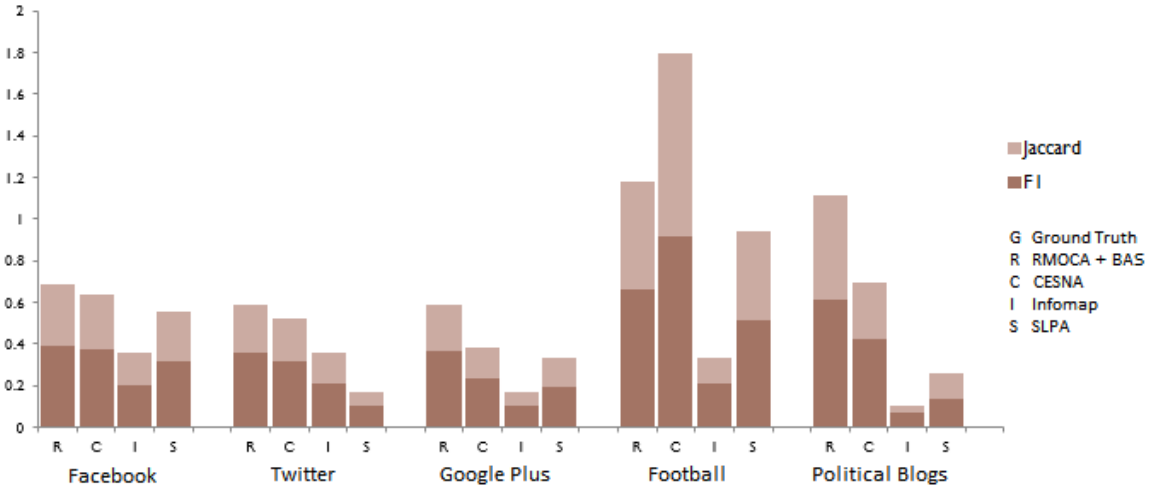



Figura 6.9: $F_1$-measure y Jaccard de RMOCA+BAS comparado con otros algoritmos del estado del arte

la recuperación de las comunidades estructuralmente no había sido buena.

De tal forma que tanto CESNA como RMOCA+BAS obtienen comunidades más

parecidas a las reales, sin embargo las comunidades detectadas por RMOCA+BAS contienen nodos más parecidos y con más conexiones que los detectados por CESNA.

## 6.3 Pre-Selección de atributos integradas al método mixto

La dimensionalidad de las redes sociales eleva la complejidad en el proceso de detección de comunidades, es por ello que siguiendo una de las hipótesis que dio origen a este trabajo de tesis se hace una selección de los atributos más relevantes, a través del *ranking* propuesto por la Ecuación 5.1, que lleva por nombre importancia global de atributos $W_a(G)$. Este proceso de selección de atributos es planteado como una etapa previa en la que se transformará un grafo inicial $G(V, E, A)$ a un grafo con un subconjunto de atributos $G^R(V, E, A^R)$ tal que $A^R \subseteq A$ y $A^R$ representa los atributos relevantes.

Seleccionar los mejores atributos para el procesos de detección de comunidades puede traer varias ventajas como la reducción de la gran cantidad de variables, quitar los elementos irrelevantes, redundantes, inesperados, erróneos o sospechosos. Esta selección puede tener otros usos como encontrar los intereses más relevantes, dirigir publicidad según el perfil de la persona, analizar el comportamiento de los usuarios con respecto a un tema y la segmentación de mercado.

Integrando la selección de atributos al método mixto, primero obtenemos el ranking de los mejores atributos del conjunto original de redes sociales con la medida $W_a(G)$, después seleccionamos los mejores atributos y reconstruimos la red, finalmente aplicamos un algoritmo de detección de comunidades sobrepuestas para encontrar las comunidades, que en este caso es RMOCA+BAS.

La selección del los atributos disminuye la complejidad de cualquier algoritmo de

detección de comunidades. El cálculo de la medida $W_a(G)$ toma $O(N+AE)$ donde $N$ es el número de nodos, $A$ el numero de atributos y $E$ el número de aristas. RMOCA toma $O(E+NA)$ para la estimación de $C_S$ y $O(NA)$ para la estimación de $C_A$ para cada $c_i \in C$. BAS toma $O(A(E+N))$. De tal forma que RMOCA+BAS tiene una complejidad de $O(A(E+N))$ para cada $c_i$ y se ve reducida a $O(\alpha(E+N)+A)$ tal que $\alpha$ es una constante menor a $A$ que va de 15 a 25 aproximadamente.

Primero, se evaluará el desempeño de $W_a(G)$+RMOCA+BAS comparando al tiempo empleado en la detección de comunidades con todos los atributos y con solo los más relevantes. Posteriormente, se hace uso de las medida F1 y Jaccard para medir el parecido de las comunidades detectadas a las esperadas en función al número de atributos. Finalmente, se compara con el algoritmo de CESNA que también hace detección de comunidades sobrepuestas.

### 6.3.1 Experimento 1: recursos en la detección de comunidades sobrepuetas

Como es de suponerse el decremento de los atributos implica una disminución significativa del tiempo que requiere el procesamiento. Se seleccionaron los mejores veinte atributos de los conjuntos de datos dados por las redes de Facebook, Twitter y Google Plus. Posteriormente se reconstruye el grafo y finalmente se integra al proceso de detección de comunidades sobrepuestas. Se considera el tiempo que consume la selección de atributos, así como el de la detección de comunidades. Además, para hacer una comparación justa se aplicó la selección tanto a CESNA como a RMOCA+BAS.

En la Figura 6.10 se muestra el tiempo empleado por ambos algoritmos con y sin la medida $W_a(G)$ en los conjuntos de datos de Facebook. Como se observa el decremento es muy significativo con una reducción de 56.85% promedio del tiempo,

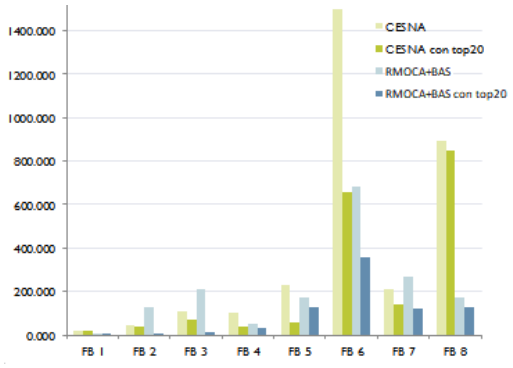



Figura 6.10: Tiempo empleado en el proceso de detección de comunidades con y sin la selección de los mejores atributos

teniendo las mayores reducciones en los conjuntos de datos de Twitter ya que éstos contaban con más atributos llegando a ocupar hasta solo 13.23% del tiempo original.

### 6.3.2 Experimento 2: evaluación de la cantidad de atributos relevantes

El *ranking* dado por la medida $W_a(G)$ nos devuelve el orden de importancia de los atributos, sin embargo se requiere seleccionar los $k$ mejores atributos. Determinar el valor apropiado de $k$ no es una tarea fácil. Se determinó que sin importar el número de atributos del conjunto inicial, los mejores resultados en cuanto a la calidad de la comunidad y el tiempo es de aproximadamente 20 atributos principalmente para aquellos conjuntos que exceden de los 250 atributos. A continuación se evalúa la cantidad de nodos a seleccionar en relación a las comunidades esperadas.

Para evaluar las comunidades detectadas se hicieron pruebas en conjuntos de datos de las redes sociales en línea de Facebook, Google Plus y Twitter. Se consideraron los tres conjuntos de datos con mayor número de atributos de cada red social, de tal forma que se evaluaron los conjuntos de datos: FB6, FB7, FB8, GP1, GP2, GP3, TW5, TW6 y TW7 que van de 319 a 3750 atributos, la descripción de los conjuntos

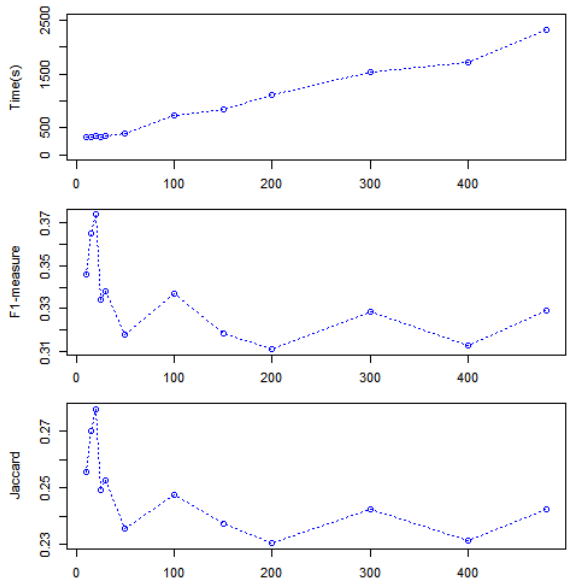



Figura 6.11: Variación de $F_1$ y Jaccard según el número de atributos seleccionados.

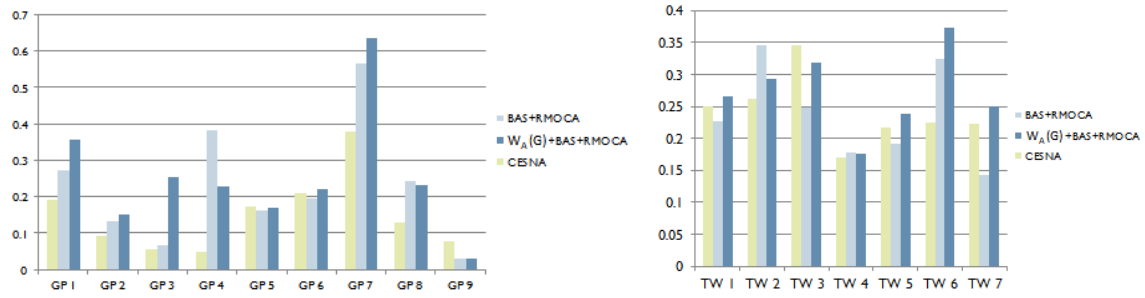
de datos se encuentra en la Tabla 6.4.

Para la pruebas se seleccionaron los $k$ mejores atributos, donde $k$ varió de 5 en 5 para los primeros 40 atributos, de 50 en 50 hasta los 200 atributos y de 100 en 100 hasta el número máximo de atributos, en este caso 3700 atributos. Se observó que no se presentan mejoras significativas en la similitud de las comunidades detectadas a las esperadas en la mayoría de los casos ya que la mejora es poca, sin embargo existen picos que suelen darse en algunos conjuntos de datos ente la selección de 15 a 25 atributos como se puede ver en la Figura 6.11 para el conjunto de datos de Facebook que tiene 480 atributos. En ella se puede ver la relación del incremento en tiempo que implica usar todos los atributos y como la calidad de la comunidad no cambia significativamente. Sin embargo existe un pico cuando el número de atributos e alrededor de 20, donde no solo se reduce el tiempo sino que las comunidades son mejores según las medidas de $F_1$ y Jaccard que se observan en esa imagen. Cabe destacar que cuando se seleccionaron menos de 5 atributos la calidad de las comunidades se veía afectada.

### 6.3.3 Experimento 3: comparación de comunidades sobrepuestas detectadas

Como era de esperarse se obtuvo una reducción en el tiempo, sin embargo, la reducción de la cantidad de los atributos podría comprometer los resultados obtenidos en la calidad de las comunidades detectadas. Seleccionando los mejores 20 atributos en los conjuntos de Facebook y Google Plus, y los mejores 25 para los conjuntos de Twitter, se extendieron los experimentos de la sección anterior a todos los conjuntos de la Tabla 6.4.

La diferencia de las comunidades detectadas fue medida usando F1 y Jaccard en la Figura 6.12 se pueden comparar los resultados obtenidos en CESNA, en RMOCA+BAS y en RMOCA+BAS con la selección de los atributos. En el caso de Twitter casi siempre se tiene una mejora cuando se hace la selección de los atributos, este incremento suele incluso superar a CESNA en conjuntos de datos en los que este algoritmo había obtenido mejores resultados que RMOCA+BAS como en TW1, TW5 y TW7.




(a) Google Plus (b) Twitter

Figura 6.12: Evaluación con Jaccard de las comunidades detectadas con y sin selección de los mejores atributos.

Dado que el ranking de la medida $W_a(G)$ puede ser aplicado a cualquier algoritmo, se aplicó al algoritmo de CESNA que detecta comunidades sobrepuestas considerando atributos. El incremento en CESNA fue en promedio de 9% en $F_1$ y cerca del 11% con

Jaccard. Dado que los primeros conjuntos de Facebook son más pequeños que de las otras redes sociales, aplicar el ranking en nuestro algoritmo afectó las comunidades detectadas, no siendo así para las de CESNA como se observa en la Figura 6.13, sin embargo, cuando el tamaño del conjunto aumenta no afecta la calidad de la comunidad.

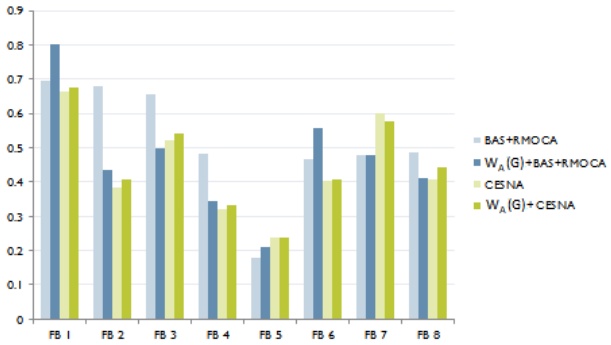



Figura 6.13: Comparación de $F_1$ de las comunidades detectadas con y sin selección de los mejores atributos en RMOCA+BAS y en CESNA.

# Capítulo 7

# Conclusiones

La detección de comunidades es un campo de estudio que ha sido abordado desde hace muchos años; sin embargo, aplicar estos algoritmos a la redes sociales en línea no es apropiado porque originalmente estaban dirigidos a grafos aleatorios y no a grafos con características propias de las redes sociales, como la de seis grados de separación. La integración de atributos al grafo, dada por las relaciones en redes sociales, permite la obtención de mejores resultados, lo cual ha sido comprobado por varios trabajos en el área; sin embargo, este tipo de integración a sido abordado en menor medida dada la complejidad que pueden llegar a tener la representación del grafo.

El principal problema en la detección de comunidades con atributos es el uso de dos tipos de información que podrían arrojar dos conjuntos de comunidades diferentes; conjuntar la estructura del grafo y los atributos para obtener un solo conjunto de comunidades sigue siendo un reto.

Los métodos existentes de detección de comunidades en redes sociales con atributos suelen obtener comunidades similares a las esperadas o comunidades con nodos muy conectados y parecidos. El problema resuelto en esta tesis fue lograr obtener ambos tipos de comunidades conjuntando la estructura del grafo y los atributos.

## 7.1 Conclusión

En esta tesis se presenta un método mixto que integra el método basada en modelo (RMOCA) y una medida que balancea atributos y estructura (BAS), la cual hace uso de la medida propuesta de calidad de atributos basada en estructura ($Q_A$). El método mixto de detección de comunidades propuesto considera los atributos a través de la mezcla de propiedades de los algoritmos basados en modelo y los algoritmos basados en distancia. RMOCA, al estar basado en modelo, permite obtener comunidades parecidas a las comunidades reales, mientras de la medida BAS ayuda a obtener mejor calidad de las comunidades en términos de entropía y densidad.

El modelo propuesto, RMOCA, no sólo permite obtener comunidades con nodos similares (nodos que comparten atributos) también encuentra nodos con múltiples enlaces al interior de las comunidades. Y a diferencia de otros algoritmos que también lo logran, nuestro modelo obtiene comunidades muy parecidas a las originales al considerar atributos. Sin embargo, la optimización usada para minimizar la función definida por el modelo es tardada y en ocasiones no logra converger, por lo que se usó un número máximo de iteraciones. A pesar de ello, se logran mejoras con respecto a algoritmos del estado del arte.

La medida de calidad, BAS, presenta una integración de la valoración de las relaciones (aristas) así como de los atributos en los nodos, lo cual mejora las comunidades previamente obtenidas por otros algoritmos. BAS balancea la estructura y los atributos, para ello hace uso de la conductividad y la nueva medida de calidad de atributos $Q_A$; esta última considera importancia global de un atributo, importancia local de un atributo y densidad de atributos. Esta medida podría ser optimizada para detectar comunidades de manera directa, sin embargo requeriría de seleccionar buenas semillas (nodos iniciales) y conocer el número de comunidades a detectar.

El método mixto para la detección de comunidades toma como base esta medida de

calidad para generar comunidades sobrepuestas, a partir de la mezcla de comunidades. El proceso de expansión de comunidades logra mejorar la precisión de las comunidades detectadas y permite considerar los atributos en comunidades previamente detectadas. Sin embargo, si las comunidades iniciales contienen nodos erróneos, éstos no son eliminados aunque podrían ser detectados como un traslape en la comunidad correcta.

Un elemento que ayuda a reducir esta complejidad es reducir la dimensionalidad, como lo es la selección de atributos. La selección de atributos propuesta permite obtener aquellos que son más importantes considerando si alrededor de los nodos que lo poseen, o en contra, se forma una o varias comunidades. Adicionalmente, esto permite eliminar aquellos atributos que son erróneos o irrelevantes. El preproceso de selección de atributos relevantes ayuda a reducir el tiempo y espacio en la detección de comunidades sobrepuestas, sin repercusiones negativas en la calidad de las comunidades, incluso en algunos casos se obtuvo una mejora en las comunidades detectadas, esto se debe a que se eliminan atributos innecesarios o erróneos.

## 7.2 Trabajo a futuro

A diferencia de otras ciencias, las ciencias sociales son constituidas a través de significado, motivos y definiciones, de tal forma que involucran un proceso de interpretación; el modelo desarrollado permite obtener conjuntos de atributos por lo que falta interpretar la relación de éstos con las comunidades de nodos encontradas en este mismo procedimiento.

La medida fue propuesta para un proceso de expansión, sin embargo también podría usarse para eliminar aquellos nodos mal clasificados en el proceso inicial, mejorando la pureza de las comunidades detectadas.

Actualmente, el proceso de expansión sigue siendo afectado por el proceso es-

tocástico de selección de atributos. La eficiencia de las comunidades se podría beneficiar si el nodo que se agrega es el mejor candidato, sin embargo podría elevar la complejidad por lo que es un estudio que se debe realizar.

BAS podría ser usada en un proceso de optimización para la detección de comunidades con las adecuaciones apropiadas, de tal forma, que se obtenga un modelo basado únicamente en distancia; o bien podría ser integrada a cualquier algoritmo tradicional de agrupamiento que haga uso de una función de distancia.

El balance entre los atributos y la estructura continúa siendo un problema abierto por lo que es necesario ampliar el análisis sobre algunas otras propiedades de las redes para obtener conclusiones precisas.

La selección de atributos dada por la importancia de atributos $W_A(G)$ permite eliminar atributos irrelevantes y erróneos, sin embargo otro de los problemas que se tiene en redes sociales es la duplicidad, por lo que debería expandirse para detectar aquellos atributos redundantes, ya sea para eliminarlos o para integrarlos en uno solo.

# Referencias


[1] Adamic, L. A., & Glance, N. (2005, August). The political blogosphere and the 2004 US election: divided they blog. In Proceedings of the 3rd international workshop on Link discovery (pp. 36-43). ACM.

[2] Akoglu, L., Tong, H., Meeder, B., & Faloutsos, C. (2012, April). PICS: Parameter-free identification of cohesive subgroups in large attributed graphs. *Proceedings of the 2012 SIAM international conference on data mining* (pp. 439-450). Society for Industrial and Applied Mathematics.

[3] Alsaleh, S., Nayak, R., & Xu, Y. (2011, July). Finding and matching communities in social networks using data mining. *International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2011* (pp. 389-393). IEEE.

[4] Baumes, J., Goldberg, M. K., Krishnamoorthy, M. S., Magdon-Ismail, M., & Preston, N. (2005). Finding communities by clustering a graph into overlapping subgraphs. *IADIS AC*, 5, 97-104.

[5] Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *science*, 286(5439), 509-512.

[6] Biswas, A., & Biswas, B. (2017). Defining quality metrics for graph clustering evaluation. *Expert Systems with Applications*, 71, 1-17.

[7] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993-1022.

[8] Boden, B., Ester, M., & Seidl, T. (2014, September). Density-based subspace clustering in heterogeneous networks. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 149-164). Springer, Berlin, Heidelberg.

[9] Bonacich, P. (1987). Power and centrality: A family of measures. American journal of sociology, 92(5), 1170-1182.

[10] Brockmann, D., & Helbing, D. (2013). The hidden geometry of complex, network-driven contagion phenomena. science, 342(6164), 1337-1342.

[11] Brandes, U., Delling, D., Gaertler, M., Görke, R., Hoefer, M., Nikoloski, Z., & Wagner, D. (2007). On finding graph clusterings with maximum modularity. *Graph-Theoretic Concepts in Computer Science* (pp. 121-132). Springer Berlin/Heidelberg.

[12] R. Breiger (1974), "The duality of persons and groups," Social Forces, 53(2): 181–190.

[13] Campo, D. N., Stegmayer, G., & Milone, D. H. (2016). A new index for clustering validation with overlapped clusters. Expert Systems with Applications, 64, 549-556.

[14] Chakraborty, T. (2015). Leveraging disjoint communities for detecting overlapping community structure. *Journal of Statistical Mechanics: Theory and Experiment*, 2015(5), P05017.

[15] Chakraborty, T., Dalmia, A., Mukherjee, A., & Ganguly, N. (2017). Metrics for community analysis: A survey. *ACM Computing Surveys (CSUR)*, 50(4), 54.

[16] Chen, J., Zaïane, O., & Goebel, R. (2009, July). Local community identification in social networks. *International Conference on Advances in Social Network Analysis and Mining, 2009. ASONAM'09.* (pp. 237-242). IEEE.

[17] Chen, W., Liu, Z., Sun, X., & Wang, Y. (2010). A game-theoretic framework to identify overlapping communities in social networks. *Data Mining and Knowledge Discovery,* 21(2), 224-240.

[18] Cheng, H., Zhou, Y., & Yu, J. X. (2011). Clustering large attributed graphs: A balance between structural and attribute similarities. *ACM Transactions on Knowledge Discovery from Data (TKDD),* 5(2), 12.

[19] Combe, D., Largeron, C., Egyed-Zsigmond, E., & Géry, M. (2012, August). Combining relations and text in scientific network clustering. *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)* (pp. 1248-1253). IEEE Computer Society.

[20] Cullum, J. K., & Willoughby, R. A. (2002). Lanczos Algorithms for Large Symmetric Eigenvalue Computations: Vol. 1: Theory (Vol. 41). Siam.

[21] Cruz, J. D., Bothorel, C., & Poulet, F. (2013). Community detection and visualization in social networks: Integrating structural and semantic information. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(1), 11.

[22] Davis, H. T. H. T. (1941). The theory of econometrics (No. 04; HB139, D38.).

[23] Dang, T. A., & Viennet, E. (2012, January). Community detection based on structural and attribute similarities. *In International conference on digital society (icds)* (pp. 7-14).

[24] Deng, X., Zhai, J., Lv, T., & Yin, L. (2017). Efficient vector influence clustering coefficient based directed community detection method. *IEEE Access*, 5, 17106-17116..

[25] Dhillon, I. S., Guan, Y., & Kulis, B. (2007). Weighted graph cuts without eigenvectors a multilevel approach. *IEEE transactions on pattern analysis and machine intelligence*, 29(11).

[26] Ding, Y. (2011). Community detection: Topological vs. topical. *Journal of Informetrics*, 5(4), 498-514.

[27] Dodds, P. S., Muhamad, R., & Watts, D. J. (2003). An experimental study of search in global social networks. *science*, 301(5634), 827-829.

[28] Donath, W. E., & Hoffman, A. J. (1973). Lower bounds for the partitioning of graphs. *IBM Journal of Research and Development,* 17(5), 420-425.

[29] Elhadi, H., & Agam, G. (2013, August). Structure and attributes community detection benchmark and a novel selection method. *International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM* (pp. 1474-1476). IEEE.

[30] Emmons, S., Kobourov, S., Gallant, M., & Börner, K. (2016). *Analysis of network clustering algorithms and cluster quality metrics at scale*. PloS one, 11(7), e0159161.

[31] Erdös, P., & Rényi, A. (1959). On random graphs, I. Publicationes Mathematicae (Debrecen), 6, 290-297.

[32] Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. *In Kdd* (Vol. 96, No. 34, pp. 226-231).

[33] Evans, T. S., & Lambiotte, R. (2009). Line graphs, link partitions, and overlapping communities. Physical Review E, 80(1), 016105.

[34] Evans, T. S. (2010). Clique graphs and overlapping communities. Journal of Statistical Mechanics: Theory and Experiment, 2010(12), P12037.

[35] Faloutsos, M., Faloutsos, P., & Faloutsos, C. (1999, August). On power-law relationships of the internet topology. *ACM SIGCOMM computer communication review* (Vol. 29, No. 4, pp. 251-262). ACM.

[36] Fiedler, M. (1973). Algebraic connectivity of graphs. Czechoslovak mathematical journal, 23(2), 298-305.

[37] Fortunato, S. (2010). Community detection in graphs. Physics reports, 486(3-5), 75-174.

[38] Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social networks*, 1(3), 215-239.

[39] Gil-Mendieta, J., & Schmidt, S. (1996). The political network in Mexico. *Social Networks*, 18(4), 355-381

[40] Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12), 7821-7826.

[41] Goh, K. I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., & Barabási, A. L. (2007). The human disease network. *Proceedings of the National Academy of Sciences*, 104(21), 8685-8690.

[42] Gregory, S. (2010). Finding overlapping communities in networks by label propagation. *New Journal of Physics*, 12(10), 103018.

[43] Gregory, S. (2007). An algorithm to find overlapping community structure in networks. *Knowledge discovery in databases: PKDD* 2007, 91-102.

[44] Gunnemann, S., Farber, I., Boden, B., & Seidl, T. (2010, December). Subspace clustering meets dense subgraph mining: A synthesis of two paradigms. *IEEE 10th International Conference on Data Mining (ICDM), 2010* (pp. 845-850). IEEE.

[45] Günnemann, S., Boden, B., & Seidl, T. (2012). Finding density-based subspace clusters in graphs with feature vectors. *Data mining and knowledge discovery,* 1-27.

[46] Günnemann, S., Boden, B., Färber, I., & Seidl, T. (2013, April). Efficient mining of combined subspace and subgraph clusters in graphs with feature vectors. *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 261-275). Springer, Berlin, Heidelberg.

[47] Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

[48] Hoang, T. A., & Lim, E. P. (2014, November). On joint modeling of topical communities and personal interest in microblogs. *International Conference on Social Informatics* (pp. 1-16). Springer, Cham.

[49] Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2(1), 193-218.

[50] Jeong, H., Mason, S. P., Barabási, A. L., & Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, 411(6833), 41-42.

[51] Jin, H., Wang, S., & Li, C. (2013). Community detection in complex networks by density-based clustering. *Physica A: Statistical Mechanics and its Applications*, 392(19), 4606-4618.

[52] Jin, D., Gabrys, B., & Dang, J. (2015). Combined node and link partitions method for finding overlapping communities in complex networks. *Scientific reports*, 5.

[53] Kannan, R., Vempala, S., & Vetta, A. (2004). On clusterings: Good, bad and spectral. Journal of the ACM (JACM), 51(3), 497-515.

[54] Kaufman, L., & Rousseeuw, P. J. (1990). Partitioning around medoids (program pam). *Finding groups in data: an introduction to cluster analysis*, 68-125.

[55] Karypis, G., & Kumar, V. (1998, November). Multilevel algorithms for multi-constraint graph partitioning. *Proceedings of the 1998 ACM/IEEE conference on Supercomputing* (pp. 1-13). IEEE Computer Society.

[56] Kelley, S., Goldberg, M., Magdon-Ismail, M., Mertsalov, K., & Wallace, A. (2012). Defining and discovering communities in social networks. I*n Handbook of Optimization in Complex Networks* (pp. 139-168). Springer US.

[57] Kim, Y., & Jeong, H. (2011). Map equation for link communities. *Physical Review E,* 84(2), 026110.

[58] Körner, J. (1973). Coding of an information source having ambiguous alphabet and the entropy of graphs. *In 6th Prague conference on information theory* (pp. 411-425).

[59] Kulis, B., & Guan, Y. (2010). Graclus—Efficient graph clustering software for normalized cut and ratio association on undirected graphs, 2008.

[60] Kumpula, J. M., Kivelä, M., Kaski, K., & Saramäki, J. (2008). Sequential algorithm for fast clique percolation. *Physical Review E,* 78(2), 026109.

[61] Lancichinetti, A., Fortunato, S., & Kertész, J. (2009). Detecting the overlapping and hierarchical community structure in complex networks. New Journal of Physics, 11(3), 033015.

[62] Lancichinetti, A., & Fortunato, S. (2011). Limits of modularity maximization in community detection. *Physical review E,* 84(6), 066122.

[63] Leskovec, J., & Horvitz, E. (2008, April). Planetary-scale views on a large instant-messaging network. *Proceedings of the 17th international conference on World Wide Web* (pp. 915-924). ACM.

[64] Leskovec, J., & Krevl, A. (2015). SNAP Datasets:Stanford Large Network Dataset Collection. Recuperado de http://snap.stanford.edu/data.

[65] Li, H., Nie, Z., Lee, W. C., Giles, L., & Wen, J. R. (2008, October). Scalable community discovery on textual data with relations. *Proceedings of the 17th ACM conference on Information and knowledge management* (pp. 1203-1212). ACM.

[66] Liu, Y., Niculescu-Mizil, A., & Gryc, W. (2009, June). Topic-link LDA: joint models of topic and author community. *Proceedings of the 26th annual international conference on machine learning* (pp. 665-672). ACM.

[67] MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, No. 14, pp. 281-297).

[68] Michell, L., & Amos, A. (1997). Girls, pecking order and smoking. *Social science & medicine,* 44(12), 1861-1869.

[69] McDaid, A., & Hurley, N. (2010, August). Detecting highly overlapping communities with model-based overlapping seed expansion. *International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2010* (pp. 112-119). IEEE.

[70] Milgram, S. (1967). The small world problem. *Psychology Today* 1(May):61-67

[71] Moser, F., Colak, R., Rafiey, A., & Ester, M. (2009, April). Mining cohesive patterns from graphs with feature vectors. *Proceedings of the 2009 SIAM International Conference on Data Mining* (pp. 593-604). Society for Industrial and Applied Mathematics.

[72] Muller, E., Sánchez, P. I., Mulle, Y., & Bohm, K. (2013, April). Ranking outlier nodes in subspaces of attributed graphs. *IEEE 29th International Conference on Data Engineering Workshops (ICDEW), 2013* (pp. 216-222). IEEE.

[73] Nguyen, N. P., Dinh, T. N., Xuan, Y., & Thai, M. T. (2011, April). Adaptive algorithms for detecting community structure in dynamic social networks. *INFOCOM, 2011 Proceedings IEEE* (pp. 2282-2290). IEEE.

[74] Neville, J., Adler, M., & Jensen, D. (2003, August). Clustering relational data using attribute and link information. *Proceedings of the text mining and link analysis workshop, 18th international joint conference on artificial intelligence* (pp. 9-15). San Francisco, CA: Morgan Kaufmann Publishers.

[75] Neville, J., Adler, M., & Jensen, D. (2004). Spectral clustering with links and attributes. MASSACHUSETTS UNIV AMHERST DEPT OF COMPUTER SCIENCE.

[76] Newman, M. E., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E,* 69(2), 026113.

[77] Newman, M. E. (2004). Detecting community structure in networks. *The European Physical Journal B-Condensed Matter and Complex Systems,* 38(2), 321-330.

[78] Nicosia, V., Mangioni, G., Carchiolo, V., & Malgeri, M. (2009). Extending the definition of modularity to directed graphs with overlapping communities. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(03), P03024.

[79] Ng, A. Y., Jordan, M. I., & Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems* (pp. 849-856).

[80] Palla, G., Derényi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043), 814-818.

[81] Parthasarathy, S., Ruan, Y., & Satuluri, V. (2011). Community discovery in

social networks: Applications, methods and emerging trends. *Social network data analytics* (pp. 79-113). Springer US.

[82] Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. *Stanford InfoLab*.

[83] Parimala, M., & Lopez, D. (2015, March). Graph clustering based on structural attribute neighborhood similarity (SANS). *IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, 2015 (pp. 1-4). IEEE.

[84] Plantié, M., & Crampes, M. (2013). Survey on social community detection. *Social media retrieval* (pp. 65-85). Springer London.

[85] Rehman, S. U., Khan, A. U., & Fong, S. (2012, August). *Graph mining: A survey of graph mining techniques*. In Digital Information Management (ICDIM), 2012 Seventh International Conference on (pp. 88-92). IEEE.

[86] Reid, F., McDaid, A., & Hurley, N. (2013). Partitioning breaks communities. *Mining Social Networks and Security Informatics* (pp. 79-105). Springer Netherlands.

[87] Rehman, S. U., Khan, A. U., & Fong, S. (2012, August). Graph mining: A survey of graph mining techniques. *2012 Seventh International Conference on Digital Information Management (ICDIM)* (pp. 88-92). IEEE.

[88] Revelle, M., Domeniconi, C., Sweeney, M., & Johri, A. (2015, September). Finding Community Topics and Membership in Graphs. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 625-640). Springer, Cham.

[89] Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex

networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4), 1118-1123.

[90] Roy, S. B., Eliassi-Rad, T., & Papadimitriou, S. (2015). *IEEE Transactions on Knowledge and Data Engineering Fast best-effort search on graphs with multiple attributes*, 27(3), 755-768.

[91] Ruan, Y., Fuhry, D., & Parthasarathy, S. (2013, May). Efficient community detection in large networks using content and links. *Proceedings of the 22nd international conference on World Wide Web* (pp. 1089-1098). ACM.

[92] Sánchez, P. I., Müller, E., Irmler, O., & Böhm, K. (2014, June). Local context selection for outlier ranking in graphs with multiple numeric node attributes. *Proceedings of the 26th International Conference on Scientific and Statistical Database Management* (p. 16). ACM.

[93] Schaeffer, S. E. (2007). Graph clustering. *Computer science review*, 1(1), 27-64.

[94] Scholz M.(2013) Network science. The research of complex networks and systems. Recuperado de: www.network-science.org

[95] Scott, J. (2011). Social network analysis. *Sage*. LondonUk,

[96] Shen, H., Cheng, X., Cai, K., & Hu, M. B. (2009). Detect overlapping and hierarchical community structure in networks. Physica A: Statistical Mechanics and its Applications, 388(8), 1706-1712.

[97] Shi, C., Li, Y., Zhang, J., Sun, Y., & Philip, S. Y. (2017). A survey of heterogeneous information network analysis. *IEEE Transactions on Knowledge and Data Engineering*, 29(1), 17-37.

[98] Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. IEEE Transactions on pattern analysis and machine intelligence, 22(8), 888-905.

[99] Silva, A., Meira Jr, W., & Zaki, M. J. (2012). Mining attribute-structure correlated patterns in large attributed graphs. *Proceedings of the VLDB Endowment,* 5(5), 466-477.

[100] Steinhaeuser, K., & Chawla, N. V. (2010). Identifying and evaluating community structure in complex networks. *Pattern Recognition Letters*, 31(5), 413-421.

[101] Smith, L. M., Zhu, L., Lerman, K., & Percus, A. G. (2016). Partitioning Networks with Node Attributes by Compressing Information Flow. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 11(2), 15.

[102] Sun, Y., Aggarwal, C. C., & Han, J. (2012). Relation strength-aware clustering of heterogeneous information networks with incomplete attributes. *Proceedings of the VLDB Endowment,* 5(5), 394-405.

[103] Sun, P. G., & Sun, X. (2017). Complete graph model for community detection. *Physica A: Statistical Mechanics and its Applications*, 471, 88-97.

[104] Thakur, G. S., Tiwari, R., Thai, M. T., Chen, S. S., & Dress, A. W. M. (2009). Detection of local community structures in complex dynamic networks with random walks. *IET systems biology,* 3(4), 266-278.

[105] Tang, J., Chang, Y.,& Liu, H. (2014). Mining social media with social theories: a survey. *ACM SIGKDD Explorations Newsletter*, 15(2), 20-29.

[106] Tang, L., & Liu, H. (2010). Graph mining applications to social network analysis. *Managing and Mining Graph Data* (pp. 487-513). Springer US.

[107] Tong, H., Faloutsos, C., Gallagher, B., & Eliassi-Rad, T. (2007, August). Fast best-effort pattern matching in large attributed graphs. *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 737-746). ACM.

[108] Wang, X. F., & Chen, G. (2003). Complex networks: small-world, scale-free and beyond. *IEEE circuits and systems magazine,* 3(1), 6-20.

[109] Whang, J. J., Gleich, D. F., & Dhillon, I. S. (2013, October). Overlapping community detection using seed set expansion. *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management* (pp. 2099-2108). ACM.

[110] Wang, Z., Zhou, X., Zhang, D., Yang, D., & Yu, Z. (2014). Cross-domain community detection in heterogeneous social networks. *Personal and ubiquitous computing*, 18(2), 369-383.

[111] Wang, T. C., Phoa, F. K. H., & Hsu, T. C. (2015). Power-law distributions of attributes in community detection. *Social Network Analysis and Mining*, 5(1), 45.

[112] Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *nature*, 393(6684), 440-442.

[113] Watts, D. J. (2004). Six degrees: The science of a connected age. *WW Norton & Company*. 2003.

[114] Wasserman, S., & Faust, K. (1994). Social network analysis in the social and behavioural sciences. *Social network analysis: Methods and Applications*, 1994, 1-27.

[115] Wu, Z., Lin, Y., Wan, H., & Tian, S. (2010, November). A fast and reasonable method for community detection with adjustable extent of overlapping. In Intelligent Systems and Knowledge Engineering (ISKE), 2010 International Conference on (pp. 376-379). IEEE.

[116] Xie, J., Szymanski, B. K., & Liu, X. (2011, December). Slpa: Uncovering overlapping communities in social networks via a speaker-listener interaction dy-

namic process. *IEEE 11th International Conference on Data Mining Workshops (ICDMW), 2011* (pp. 344-349). IEEE.

[117] Xie, J., Kelley, S., & Szymanski, B. K. (2013). Overlapping community detection in networks: The state-of-the-art and comparative study. *Acm computing surveys (csur)*, 45(4), 43.

[118] Xu, Z., Ke, Y., Wang, Y., Cheng, H., & Cheng, J. (2012, May). A model-based approach to attributed graph clustering. *In Proceedings of the 2012 ACM SIGMOD international conference on management of data* (pp. 505-516). ACM.

[119] Xu, Z., Ke, Y., Wang, Y., Cheng, H., & Cheng, J. (2014). GBAGC: a general bayesian framework for attributed graph clustering. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 9(1), 5.

[120] Yang, B., Cheung, W., & Liu, J. (2007). Community mining from signed social networks. *IEEE transactions on knowledge and data engineering*, 19(10).

[121] Yang, J., & Leskovec, J. (2015). Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems*, 42(1), 181-213.

[122] Yang, J., & Leskovec, J. (2013, February). Overlapping community detection at scale: a nonnegative matrix factorization approach. *Proceedings of the sixth ACM international conference on Web search and data mining* (pp. 587-596). ACM.

[123] Yang, T., Chi, Y., Zhu, S., Gong, Y., & Jin, R. (2011). Detecting communities and their evolutions in dynamic social networks—a Bayesian approach. *Machine learning*, 82(2), 157-189.

[124] Yang, J., McAuley, J., & Leskovec, J. (2013, December). Community detection in networks with node attributes. *IEEE 13th international conference on Data Mining (ICDM), 2013* (pp. 1151-1156). IEEE.

[125] Yin, Z., Gupta, M., Weninger, T., & Han, J. (2010, April). Linkrec: a unified framework for link recommendation with user attributes and graph structure. P*roceedings of the 19th international conference on World wide web* (pp. 1211-1212). ACM.

[126] Yin, Z., Gupta, M., Weninger, T., & Han, J. (2010, August). A unified framework for link recommendation using random walks. *International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2010* (pp. 152-159). IEEE.

[127] Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. Journal of anthropological research, 33(4), 452-473.

[128] Zhang, S., Wang, R. S., & Zhang, X. S. (2007). Identification of overlapping community structure in complex networks using fuzzy c-means clustering. Physica A: Statistical Mechanics and its Applications, 374(1), 483-490.

[129] Zhang, Y., Levina, E., & Zhu, J. (2016). Community detection in networks with node features. *Electronic Journal of Statistics*, 10(2), 3153-3178.

[130] Zhou, H., 2003a, *Phys. Rev. E* 67(6), 061901.

[131] Zhou, Y., Cheng, H., & Yu, J. X. (2009). Graph clustering based on structural/attribute similarities. *Proceedings of the VLDB Endowment*, 2(1), 718-729.

[132] Zhou, Y., Cheng, H., & Yu, J. X. (2010, December). Clustering large attributed graphs: An efficient incremental approach. *IEEE 10th International Conference on Data Mining (ICDM), 2010* (pp. 689-698). IEEE.

[133] Tomada el 5 de diciembre de 2017 de: http://scienceblogs.com/goodmath/wp-content/blogs.dir/476/files/2012/04/i-24b2db50d71be775f98de9a464113aca-maximal-cliques.jpg