RNPS No. 2143
ISSN 2072-6260
Versión Digital

SERIE GRIS

REPORTE TÉCNICO
Minería de Datos

Análisis de redes sociales: una introducción

Airel Pérez Suárez, José E. Medina Pagola,
Raudel Hernández-León y
Andrés Gago-Alonso

RT_020 **octubre 2013**

CENATAV
Centro de Aplicaciones de Tecnologías de Avanzada
MINISTERIO DE LA INDUSTRIA BÁSICA

7ma. No. 21812 e/218 y 222,
Rpto. Siboney, Playa;
La Habana.
Cuba. C.P. 12200
www.cenatav.co.cu

RNPS No. 2143
ISSN 2072-6260
Versión Digital

SERIE GRIS

REPORTE TÉCNICO
Minería de Datos

Análisis de las redes sociales: una introduccion

Airel Pérez Suárez, José E. Medina Pagola,
Raudel Hernández-León y
Andrés Gago-Alonso

RT_020 **octubre 2013**

CENATAV
Centro de Aplicaciones de Tecnologías de Avanzada
MINISTERIO DE LA INDUSTRIA BÁSICA

7ma. No. 21812 e/218 y 222,
Rpto. Siboney, Playa;
La Habana.
Cuba. C.P. 12200
www.cenatav.co.cu

Tabla de contenido

Lista de figuras

Lista de tablas

Análisis de las redes sociales: una introducción

Airel Pérez Suárez, José E. Medina Pagola, Raudel Hernández-León, y Andrés Gago-Alonso

Dpto. Minería de Datos, Centro de Aplicaciones de Tecnologías de Avanzada (CENATAV),
La Habana, Cuba
{asuarez,jmedina,rhernandez,agago}@cenatav.co.cu

RT_020, Serie Gris, CENATAV
Aceptado: 15 de julio de 2013

Resumen. El análisis de las redes sociales es un área de investigación interdisciplinaria con aportaciones de diversas disciplinas, como son las ciencias sociales, la psicología social, matemática, ciencias computacionales, entre otras, permitiendo comprender y visualizar diversas características y, particularmente, evaluar y construir patrones inmersos en sus estructuras. En este trabajo se abordan diferentes concepciones y definiciones asociadas, presentando ejemplos y problemas donde estas redes se aplican. Además, se exponen algunas de las técnicas de minería de datos utilizadas en el análisis de estos tipos de redes.

Palabras clave: red social, red dinámica, análisis de comunidades, análisis de enlaces, red criminal.

Abstract. Social network analysis is an interdisciplinary research area with contributions from different disciplines, as sociology, social psychology, mathematics, computer science, among others, allowing the understanding and visualizing many features and, especially, the evaluation and construction of patterns within their structures. This work deals with different understandings and definitions on this matter, showing examples and problems where they are applied. Besides, we present some data mining techniques used in the analysis of these kinds of networks.

Keywords: social network, dynamic network, community analysis, link analysis, criminal network.

1. Introducción

En los últimos años, debido al intenso flujo y almacenamiento de datos, especialmente sobre escenarios en los que se involucran individuos y actividades sociales, se ha observado un crecimiento significativo de investigaciones y aplicaciones enfocadas a su manipulación y a la extracción de conocimientos. El advenimiento de la Internet y de la *World Wide Web* ha propiciado el creciente interés por tales investigaciones, conformando un campo de trabajo conocido como Análisis de Redes Sociales (*Social Network Analysis*).

El Análisis de Redes Sociales o SNA (por sus siglas en inglés) centra su atención en los problemas asociados con grandes redes, las que no sólo son difíciles de comprender y visualizar, sino que además exigen evaluar y construir patrones inmersos en sus estructuras para un mejor análisis. Por red social se comprende a las redes de relaciones e interacciones entre entidades sociales, tales como individuos, colectivos y organizaciones.

El SNA es aplicable a muchas situaciones prácticas. Muchas de ellas se observan en los problemas surgidos con la WWW; particularmente, donde los vértices representan personas o grupos de individuos. Ejemplo de ellos son las redes de blogs, las redes de intercambio de correos electrónicos, las redes que se

forman en las listas y redes de colaboración, y otras de ámbito global, como las creadas por Facebook y Twitter. Sin embargo, su aplicación puede ir más allá, como las requeridas por las redes de llamadas que se observan en el tráfico de las telecomunicaciones, las redes metabólicas que se presentan en las ciencias biológicas y bioquímicas, etc.

Otro término relacionado, y en cierta forma sinónimo, con SNA es el Análisis de Enlaces (*Link Analysis*) [1,2,3]. Bajo este término se refuerza la relevancia de las interconexiones sociales, tales como las relaciones de amistad, parentesco, sexuales, académicas, intercambios financieros, entre otras.

También relacionado con SNA y el análisis de enlaces se tiene el análisis de Redes Criminales (*Criminal Networks*) [4,5,6,2]. En estas redes se modelan pandillas, redes de tráfico y ventas de drogas, grupos de mafiosos, vándalos, carteristas, terroristas, etc. En estos dominios se han observado estudios criminológicos que justifican la "selección"(en términos sociológicos) de los criminales por tales relaciones, las que pueden ir desde *modus operandi* similares hasta diferentes grados de organizaciones cuyas asociaciones están motivadas por, y reforzadas hacia, perfiles criminales preferidos [7].

En este trabajo se realiza un estudio del estado actual de las investigaciones en el campo de las redes sociales. En la segunda sección se abordan los conceptos básicos, las métricas utilizadas y los tipos y ejemplos de redes sociales. En la tercera sección se aborda la creación y visualización de las redes. En la cuarta sección se analizan algunos de los métodos de Minería de Datos que se utilizan en el análisis de redes sociales. Finalmente, en la sección cinco se indican algunas consideraciones, a modo de conclusión sobre lo abordado en este trabajo.

2. Análisis de redes sociales: definiciones, problemas y aplicaciones

En la presente sección se presentarán aspectos básicos sobre el análisis de las redes sociales, tales como sus fuentes, definiciones y tareas (sección 2.1), conceptos y métricas empleadas en las técnicas de minería de datos (sección 2.2), tipos y ejemplos de redes sociales (sección 2.3), problemas donde se han aplicado el análisis de redes sociales (sección 2.4) y algunas colecciones de prueba utilizadas en las experimentaciones (sección 2.5).

2.1. ¿Qué es SNA?

El Análisis de las Redes Sociales (SNA), más que una teoría formal, es un enfoque o línea de investigación sobre estructuras sociales, denominado en otros círculos como Análisis Estructural (*Structural Analysis*) [8,9].

El SNA parte de la premisa de que las entidades sociales son creadas básicamente por relaciones y por patrones de comportamiento originados por esas relaciones. Formalmente, una red social es usualmente definida como un conjunto de actores sociales, o nodos, cuyos miembros están conectados por uno o varios tipos de relaciones [10]. Los nodos son, mayormente, individuos, grupos u organizaciones, pero el SNA puede considerar redes con otros tipos de nodos, tales como páginas *web*, *blogs*, artículos de revistas, etc. [9].

Entre las fuentes históricas del SNA se tienen los trabajos de los pioneros de la sociología en los finales del siglo XIX, como Émile Durkheim y Ferdinand Tönnies, que plantean que los grupos sociales pueden existir debido a lazos entre los individuos que se forman por compartir valores, creencias o vínculos socio-instrumentales, en donde el fenómeno social se manifiesta por las interacciones entre los individuos y no por las propiedades de los actores.

Sin embargo, es a partir de los años 20 y 30 que se refieren trabajos que cimientan las bases del SNA. Entre estos se citan los del antropólogo Roger Brown y, particularmente, del socio-psicólogo Jacob Levy Moreno, el cual utilizó el término de red, en el sentido que se usa hoy en día, a partir del estudio de pequeños grupos de trabajo [9].

A finales de los 70, el término SNA era internacionalmente reconocido en sociología. Sin embargo, con el surgimiento de la Internet y con la existencia de redes sociales a gran escala, incluso global, es que se observa, desde los años 90, un rápido incremento de las investigaciones donde se modelan y abordan estos problemas por las ciencias computacionales.

La asociación científica que más ha trabajado en este campo es *The International Network for Social Network Analysts* (INSNA), fundada por Barry Wellman en 1977. Esta es una asociación multidisciplinaria, aunque se ha enfocado fundamentalmente en los trabajos sociológicos. En los últimos años se han estado incluyendo estas temáticas en los congresos especializados en ciencias computacionales, destacándose dentro de la minería de datos la *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (ASONAM), celebrada anualmente desde el 2009.

Las redes sociales se asocian con estructuras basadas en grafos. Las aristas de estos grafos permiten modelar diferentes tipos de relaciones sociales como: las de semejanzas, las sociales, las de interacciones y las de flujos [11,12]. Las relaciones de semejanzas se observan cuando dos nodos comparten atributos, donde usualmente las co-membresías a grupos sociales son objetos de análisis. Las relaciones sociales se presentan en redes que modelan roles (amistades, académicas, etc.), afectos y sentimientos (gustos) o relaciones cognitivas. Las de interacciones se refieren a relaciones tales como con quién se habla, a quién se ayuda, se invita, etc. Las de flujos son relaciones asociadas con recursos, informaciones o influencias.

El SNA ha incluido una gran diversidad de tareas, muchas de las cuales consideran métricas específicas. Entre estas tareas se tienen las siguientes [13,5,6,14,15]:

- **Análisis de Centralidad**: Persigue la identificación de los actores de la red más relevantes, los más influyentes o con un determinado poder dentro de la red.
- **Detección de comunidades**: Permite identificar actores que forman grupos o comunidades, usualmente a partir de la estructura y la topología de la red.
- **Análisis de roles o posición**: Buscan los roles que juegan los actores de la red o las posiciones que juegan como enlaces entre grupos.
- **Modelado de redes**: Buscan los tipos que se forman en redes de gran tamaño y complejidad, usualmente con el propósito de su estudio mediante simulaciones.
- **Análisis de la difusión de información**: Analizan cómo se propagan las informaciones en la red, permitiendo la comprensión de la dinámica cultural. En particular, en muchos trabajos sociológicos se ha observado cómo se transmiten las ideas e informaciones de "boca en boca", como sucede en el intercambio de correos y en los *blogs*, entre otros. Este fenómeno también es denominado como Comercialización Viral (*Viral Marketing*), buscándose en muchos casos los actores claves que maximizan la difusión.
- **Clasificación en redes**: Analizan determinados actores para inferir otros con similares comportamientos (por ejemplo, otros terroristas a partir de los identificados).
- **Detección de *outliers***: Al margen de la conveniencia de eliminar los efectos indeseados de los *outliers*, existen situaciones en los que se persigue identificar a estos actores pues en ocasiones los *outliers* pudieran ser los verdaderos líderes de una red ilegal.

Estrechamente vinculado con el SNA, tratada como un subcampo de esta, se tiene el Análisis de Redes Dinámicas (*Dynamic Network Analysis*). Una red dinámica es un tipo especial de red compleja que evoluciona y se modifica en el tiempo. En el sentido de una red social en línea, tales como Facebook

y Twitter, los cambios se producen al incluirse o retirarse comunidades, amigos, o sus conexiones, entre otros. Aunque estos eventos parecerían introducir poco efecto en la red, la dinámica de la red en un período de tiempo prolongado podría producir una transformación significativa de su estructura, representando un reto su procesamiento [16].

Para modelar las redes dinámicas, estas se representan como una serie temporal de vistas de la red conocidas como fotografías o instantáneas (en inglés *snapshots*). Esta forma de representación permite enfrentar otras tareas en el campo de las redes sociales, estrechamente vinculados con el análisis de comunidades dinámicas y la minería de grafos dinámicos, entre las que se tienen las siguientes [17,18,19,15]:

- **Predicción de enlaces**: Se busca predecir el surgimiento de relaciones basado en la historia y dinámica de la red, buscándose en determinados casos no solo su surgimiento, sino cuándo podría surgir en un futuro.
- **Identificación de patrones dinámicos del comportamiento de los actores**: Tipos de patrones temporales y tendencias que exhiben, si estos son cíclicos, si son predecibles, si tienen patrones de comportamiento diferentes.
- **Predicción de cambios de estructura**: Se busca predecir los cambios de roles de los actores. Por ejemplo, la transición de un nodo con alto grado interno hacia uno de alta intermediación. O detectar si la estructura de la red en su conjunto se hace o tiende a una más o menos predecible.
- **Detección de transiciones inusuales o anómalas**: Se evalúa si existen actores en instantes de tiempo con patrones de comportamiento significativamente diferentes basado en su historia.
- **Modelación de la dinámica de meta-comunidades**: La identificación de patrones dinámicos en grupos de entidades latentes. La modelación y análisis de las asociaciones de las entidades y grupos. La identificación de las meta-comunidades y explicar las diferencias en el tiempo entre las comunidades de la meta-comunidad.

En la próxima sección se analizarán algunos de los conceptos y métricas que se aplican en las soluciones de estas tareas.

2.2. Conceptos y métricas

Como se mencionó, las redes sociales se asocian con estructuras basadas en grafos. Sin embargo, aunque muchas de las definiciones y procedimientos se sustentan en conceptos básicos de la teoría de grafos, tales como isomorfismo de subgrafos, subgrafos maximales comunes, distancia de edición en grafos, aristas faltantes, entre otros, estos no son aplicables en los grafos de entidades-relaciones y redes sociales [20].

A continuación se presentan algunos conceptos y métricas asociados con las redes sociales los cuales son empleados en la exposición y solución de las tareas así como en la aplicación de los métodos de minería de datos utilizados en SNA. Estos se presentan en orden alfabético para facilitar su localización.

Autoridad y Centro

Los conceptos de Autoridad y Centro (*Authority and Hub*) son medidas empleadas en el análisis de enlaces. Estas fueron definidas originalmente por Kleinberg [21] para el análisis de los tipos de páginas Web, y extendidas a otros tipos de actores. Estos conceptos se relacionan con la observación de que existen dos tipos de páginas: los Centros, los cuales agrupan enlaces a páginas autorizadas, y las Autoridades, las que son fuentes de información de un tópico dado.

Para descubrir las páginas autoridades y centros se ha propuesto un algoritmo, usualmente conocido como HITS (*hyperlink-induced topic search*), el cual se define a partir de las expresiones siguientes:

$$a_i = \sum_{j \in M_i} h_j \ ,$$
$$h_j = \sum_{i \in L_j} a_i \ ,$$

siendo a_i y h_j las métricas de autoridad y centro de las páginas p_i y q_j respectivamente, M_i el conjunto de las páginas que apuntan a p_i, y L_j el conjunto de las páginas a las que apunta q_j.

Los cálculos de a_i y h_j se realizan a través de un proceso iterativo, partiendo de valores unitarios, hasta que converjan en un número determinado de iteraciones.

Centralidad

La Centralidad es una medida que indica de forma aproximada el poder de un nodo en la conexión de la red. Existen varias medidas de centralidad, ejemplo de ellas son la intermediación, la cercanía y el vector singular.

Cercanía (Centralidad de)

La Cercanía (*Closeness*) es una medida de centralidad que indica el grado en que un nodo está cerca de otros nodos en el grafo (directa o indirectamente). Este refleja la habilidad de acceder a las informaciones a través de otros miembros de la red. En redes inconexas, la centralidad de cercanía debe calcularse por cada componente. Obsérvese que la cercanía es inversa a la suma de las distancias mínimas (también conocida como distancia geodésica) entre el nodo y otros nodos.

Clique

Un clique es un subgrafo en el que existe una arista entre cualquier par de vértice. Este concepto es equivalente a decir que el grafo inducido por el subgrafo es completo. Esta definición es muy restrictiva, por lo que en el SNA se aplican otros conceptos que relajan algunas de sus propiedades. Las propiedades ideales de un clique son definidos por los valores máximos de: la cohesión o familiaridad, el alcance (*Reachability* - relativo al diámetro), la robustez (alta conectividad, siendo difícil la destrucción del grupo por remover sus miembros), y la densidad [22,23].

Coeficiente de Agrupamiento

El Coeficiente de Agrupamiento (*Coeficient Clustering*), definido con la letra C, se define con la siguiente fórmula:

$$C = \frac{3 * \text{número de triángulos del grafo}}{\text{número de tripletas conectadas con los nodos}},$$

Esta medida se relaciona con la transitividad y es precisamente la probabilidad de que dos de los amigos de un actor sean igualmente amigos [24]. Altos valores de esta medida son indicadores de alta completitud de un clique [9]. Una expresión similar asociada a un nodo se ha definido para medir el coeficiente de agrupamiento local.

Cohesión

La Cohesión, o Familiaridad, es la propiedad que indica el grado en que los actores están conectados directamente entre sí. La búsqueda de subgrupos cohesionados es uno de los objetivos de muchas tareas en el SNA, siendo el clique el de máxima cohesión. La cohesión social es utilizada a menudo en trabajos sociológicos, ya que los miembros de subgrupos cohesionados tienden a compartir información, tener ideas similares, compartir ideas, creencias, comportamientos e incluso hábitos y enfermedades [22].

Cohesión Estructural

La Cohesión Estructural (*Structural Cohesion*) mide el menor número de miembros que, si se eliminan de un grupo, pudiera desconectar al grupo [23].

Comunidad

Una comunidad es un conjunto de individuos con fuertes relaciones entre ellos y débiles interacciones externas [25,26]. Conceptos similares encontrados en la literatura en otros contextos son los de grupos, clusters, subgrupos cohesivos y módulos [15]. Pocos trabajos tratan las comunidades como grupos solapados. En Chen *et al*. [27], con el propósito de identificar comunidades solapadas, se propone que una red social puede tener tres tipos de nodos: *Hubs* (nodos con muchas conexiones y que se encuentran en los cubrimientos de comunidades), *Outliers* (nodos con pocas conexiones y que no pertenece a ninguna comunidad) y nodos Normales (nodos con algunas conexiones y que pertenecen a una comunidad).

Comunidad Fuerte

(Ver Conjunto LS)

Conjunto LS

Un Conjunto LS (*LS Set* o Comunidad Fuerte) es un conjunto de nodos donde cualquier subconjunto propio de nodos del mismo tiene más aristas con su complemento dentro del conjunto que las que tiene con los nodos fuera del conjunto [15]. El conjunto LS también se define como el subgrafo cuyo grado interno de cualquier nodo es mayor que su grado externo [25].

Conjunto Lambda

Un Conjunto Lambda (*Lambda Set*) es un subgrafo que requiere más aristas para desconectar cualquier par de nodos internos que para desconectar cualquier nodo interno de los nodos externos a él [15].

Densidad (de aristas)

La Densidad es la propiedad que mide la proporción de aristas de un subgrafo respecto al total de sus posibles aristas [23]. Matemáticamente se define como:

$$Densidad = \frac{m}{\binom{n}{2}},$$

siendo m la cantidad de aristas que existen y n la cantidad de nodos, ambos del subgrafo.

Densidad intra-cluster

La Densidad intra-cluster de un cluster C es la razón de la cantidad de aristas internas a C entre la cantidad de posibles aristas internas [25].

Densidad inter-cluster

La Densidad inter-cluster de un cluster C es la razón de la cantidad de aristas externas a C entre la cantidad de posibles aristas entre clusters del grafo completo [25].

Diámetro

El Diámetro es la mayor entre todas las distancias mínimas (geodésicas) entre cualquier par de nodos.

Equivalencia Estructural

Dos nodos son equivalentes estructuralmente si ambos poseen igual cantidad de enlaces a los mismos vecinos. Dichos nodos no tienen que estar conectados entre sí.

Familiaridad

(Ver Cohesión).

Grado (Centralidad de)

El Grado (*Degree Centrality*), también llamado Valencia, es la medida más simple de centralidad. Esta se define como la cantidad de aristas incidentes a un nodo. En redes dirigidas, existen dos tipos de centralidad de grado: El Grado Interno y el Grado Externo, relacionados con la cantidad de nodos vecinos considerando la dirección de las aristas [28].

Intermediación (Centralidad de)

La Intermediación (*Betweenness*), propuesto por Freeman[1] es una medida de centralidad que mide la cantidad de caminos mínimos que pasan por un nodo de un grafo [24]. Esta medida permite evaluar la conectividad de los vecinos del nodo, considerándose que presentan altos valores a nodos que enlazan a comunidades. En una red de actores, la Intermediación refleja la cantidad de individuos relacionados indirectamente con un actor o, en otros contextos, la influencia de un actor en el flujo de información entre otros individuos, particularmente cuando la información fluye preferentemente a través del camino más corto posible, mostrando los actores más populares, eficientes o poderosos.

Intermediación de Aristas (Centralidad de)

La Intermediación de Aristas (*Edge Betweenness*), propuesto por Girvan y Newman [24], es una medida de centralidad, aplicable a las aristas, definida como la cantidad de caminos mínimos entre pares de vértices que pasan por una arista. De esta forma, las aristas que conectan comunidades tienen valores altos de intermediación de aristas por lo que, eliminando esas aristas, se revelan las estructuras de comunidades del grafo. Según sus autores, en lugar de encontrar los centros de comunidades fuertemente conectados con la Intermediación clásica, esta medida permite detectar las periferias de las comunidades. En algunos trabajos se ha propuesto la intermediación de aristas considerando que los caminos mínimos son menores o iguales a un valor predefinido [29].

Intermediación Dividida (Centralidad de)

La Intermediación Dividida (*Split Betweenness*), propuesto por Gregory en el 2007 [29], provee una forma de decidir (1) cuándo dividir un vértice en lugar de eliminarlo, (2) cuál vértice se divide y (3) cómo dividirlo. Un vértice debe dividirse en dos si estos dos vértices pertenecen a diferentes grupos. Esto se puede verificar contando la cantidad de los caminos mínimos que pudieran pasar entre esos nodos si estos estuvieran unidos por una arista. Entonces, si existieran más caminos mínimos entre esos nodos que con cualquier arista real, el vértice debe dividirse; en caso contrario, el vértice debe ser eliminado.

***k*-Camino sobre una arista (Centralidad de)**

Un *k*-Camino sobre una Arista (o Arista de k-Camino, del inglés *k-path edge*) es una medida de centralidad, propuesto por De Meo *et al*. en el 2013 [30], asociado con una arista *e* y definido como la suma, sobre todos los vértices posibles *s*, de las probabilidades de que un mensaje originado en *s* pase por *e*, asumiendo que el mensaje pasa sólo por *k*-caminos simples.

***k*-Camino Simple (Centralidad de)**

Un *k*-Camino Simple (*simple k-path*) es una medida de centralidad asociada con un camino que contiene cuando más *k* aristas seleccionadas aleatoriamente [30].

***k*-Camino sobre un vértice (Centralidad de)**

Un *k*-Camino sobre un Vértice (o Vértice de k-Camino, del inglés *k-path vertex*) es una medida de centralidad, propuesta por Alahakoon *et al*. en el 2011 [31], asociado con un vértice *v* y definido como la suma, sobre todos los vértices posibles *s*, de las probabilidades de que un mensaje originado en *s* pase por *v*, asumiendo que el mensaje pasa sólo por *k*-caminos simples.

[1] Freeman, L. (1977) Sociometry 40, 35-41.

***k*-Clique**

Un *k*-Clique es un subgrafo cuya distancia mínima (distancia geodésica) entre cualquier par de nodos es menor o igual que *k* [23]. El camino de la distancia que se considera no tiene que estar en el grupo, pudiendo incluso estar este internamente inconexo [15]. Las definiciones dadas por diferentes trabajos asumen que el *k*-clique se refiere al maximal.

***k*-Clan**

Un *k*-Clan (también conocido como Clique Sociométrico) es un *k*-clique cuyo subgrafo inducido tiene un diámetro no mayor que *k*.

***k*-Club**

Un *k*-Club es un subgrafo cuyo grafo inducido tiene un diámetro menor o igual que *k*. Nótese que un *k*-Club exige que los caminos mínimos se encuentren en el subgrafo. Por lo tanto, todo *k*-club es un *k*-clique, pero no todo *k*-clique es un *k*-club.

***k*-Core**

Un *k*-Core es un subgrafo en el que todos sus nodos tienen al menos *k* vecinos en el subgrafo.

***k*-Plex**

Un *k*-Plex es un subgrafo de n_s vértices en el que todos sus nodos tienen al menos $n_s - k$ vecinos en el subgrafo. Un *k*-plex de n_s vértices es un (n_s-*k*)-core [15]. Las definiciones dadas por diferentes trabajos asumen que el *k*-plex se refiere al maximal.

Modularidad

La Modularidad, propuesto por Newman y Girvan en el 2004 y denominada con la letra *Q*, evalúa cuan estructurado en comunidades se encuentra una red [32]. La modularidad se define como la fracción de todas las aristas que se tienen dentro de las comunidades menos el valor esperado de igual valor en un grafo aleatorio, suponiendo que: (1) ambos grafos tienen igual cantidad de nodos, (2) cada nodo del grafo aleatorio tiene igual grado que su par, y (3) las aristas son ubicadas aleatoriamente. Esta definición considera que en los grafos aleatorios no se espera encontrar estructuras de comunidades [33,30]. La aplicación de esta medida y la detección de la estructura de comunidades ha propuesto diversos algoritmos que tratan de lograr óptimos con costos computacionales aceptables, pero aún sigue siendo un problema abierto. La expresión de modularidad más utilizada es la siguiente:

$$Q = \sum \left(e_{ii} - a_i^2 \right) \ ,$$

donde e_{ii} es la fracción de aristas en cada una de las *k* comunidades y a_i la fracción de aristas presentes en los nodos de la comunidad *i*. Esto asume que la probabilidad de que dos nodos se enlacen, bajos las condiciones de aleatoriedad asumidas, está dada por el producto de los grados de ambos nodos.

***R* (Métrica)**

La métrica *R*, propuesta por Chen *et al*. en 2010 [27], se define por la expresión:

$$R(i,j) = \frac{\sum_{x \in N_j} r(x,i) + \sum_{x \in N_i} r(x,j)}{2} \ ,$$

donde N_i son los vecinos de *i*, incluyéndolo, (ídem para N_j) y $r(i,j)$ se define por la expresión:

$$r(i,j) = A_{ij} - \frac{max(k_i, k_j)}{n-1} \ ,$$

siendo A_{ij} el valor 1 si existe la arista (i,j), 0 en caso contrario, k_i el grado del nodo *i* (ídem k_j para el nodo *j*), y *n* la cantidad de nodos en el grafo. La expresión $r(i,j)$ cuantifica la relación entre los nodos

i y *j* considerando las probabilidades de que *i* y *j* estén conectados con cualquier nodo del grafo en un modelo aleatorio. Los autores comentan que la métrica *R* da valores altos a los *outliers* y valores bajos a los *hubs* (ver Comunidad).

Valencia
(ver Grado).

Vector Singular (Centralidad de)
El Vector Singular es una medida de centralidad (conocida como EVC, por sus siglas en inglés *Eigenvector Centrality*) que asigna valores relativos a cada nodo basado en el principio de que las conexiones a nodos con altos valores contribuyen más al valor del nodo en cuestión. También es definido de forma recursiva como una medida de centralidad proporcional a la suma de los valores de centralidad de todos sus nodos vecinos [34]. El EVC es, intencionalmente, similar al PageRank de Google, con la limitante de la complejidad computacional asociado al cálculo de vectores singulares a partir de la matriz de grados de los nodos.

2.3. Tipos y ejemplos de redes sociales

Las redes sociales son estructuras sociales que generalmente son representadas a través de grafos [35]. En estos grafos, los nodos son los individuos del universo de estudio y las aristas entre los nodos están determinadas por las relaciones de interés que existen entre dichos individuos. Naturalmente, en dependencia del tipo de relación que se esté considerando, las aristas entre los nodos pueden ser dirigidas o tener algún tipo de peso. Considerando estas características, el grafo que representa a la red puede ser dirigido y/o pesado.

Supóngase que se tiene un conjunto formado por seis personas: María, Carlos, Juan, Daniela, Pamela y Alex. De este conjunto, se conocen las siguientes relaciones de amistad:

- Alex es amigo de Carlos.
- Carlos es amigo de Juan.
- Alex es amigo de Daniela.
- Alex es amigo de María.
- María es amiga de Pamela.
- María es amiga de Juan.

En la Fig. 1, se muestra la red social que se forma teniendo en cuenta los datos del ejemplo anterior. Como se puede observar en esta figura, las aristas o enlaces entre los nodos no tienen dirección; esto se debe a que la relación de amistad es una relación simétrica (*A es amigo de B* $\Leftrightarrow$ *B es amigo de A*)

Aunque no existe un consenso establecido de cómo clasificar las redes sociales, existen varios criterios que han sido utilizados para clasificar dichas redes [36]. Estos criterios son:

a) **Su tamaño**. Esta clasificación se centra en el *diámetro* de la red. El diámetro de una red social es la distancia del mayor camino que existe entre dos nodos de la red. No existe un valor establecido que determine cuándo una red es grande o pequeña. Por lo general este valor depende de lo que se quiera representar con la información que se tenga. Luego, de acuerdo a este criterio las redes sociales se pueden clasificar en redes a *pequeña escala* o redes *gran escala*:
   - Redes a pequeña escala: Son aquellas que pueden ser analizadas por los sistemas para la visualización de redes; para el análisis de las redes a pequeña escala se utiliza el conjunto de datos completo. Un ejemplo de red a pequeña escala es la formada a partir de la colaboración científica

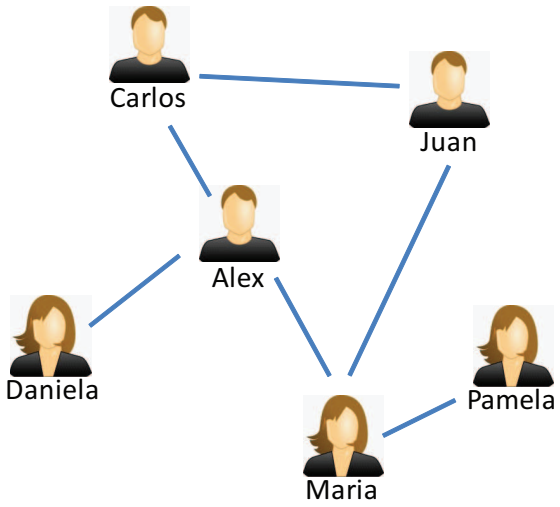

Fig. 1. Ejemplo de red social.

entre investigadores; en esta red, los nodos son los investigadores y las aristas se determinan a partir de los artículos que desarrollan en conjunto los investigadores.
- Redes a gran escala: Son aquellas que no pueden ser analizadas por los sistemas para la visualización de redes; para el análisis de este tipo de red se suele trabajar con un conjunto representativo de la red. Un ejemplo de red a gran escala es la formada por los sistemas de correo electrónico; en esta red los nodos son los distintos usuarios de correo de la web y los enlaces se determinan atendiendo a los contactos de las listas de correo de cada usuario.

b) **Cambios que puede sufrir la red durante el estudio**. Este tipo de redes puede ser de cualquier tamaño y tener diferentes formas. De acuerdo a este criterio, las redes pueden clasificarse en *estáticas* o *dinámicas*:
- Redes estáticas: Son aquellas que no sufren ningún cambio cuando son sujetas a estudio; es decir, mantienen su estructura durante todo el tiempo que dure su estudio.
- Redes dinámicas: Son aquellas que sufren cambios producto de adiciones o eliminaciones de individuos, así como de relaciones entre estos. Un ejemplo clásico de este tipo de redes es la propia Web, la cual está en constante cambio y su tamaño aumenta al pasar el tiempo.

c) **Su origen de datos**. Estas redes dependen de su origen de datos y, en muchos casos, pueden representar comunidades virtuales y/o del mundo real. De acuerdo a este criterio, las redes pueden clasificarse en *off-line* u *on-line*:
- Redes *off-line*: Son aquellas en las cuales las relaciones sociales se establecen sin la intervención de medios electrónicos. En este tipo de redes, la administración y conocimiento de las relaciones recae completamente en el individuo. Un ejemplo de este tipo de redes fue la generada para el caso del supuesto suicidio del científico británico David Kelly, la cual fue creada a partir de documentos que tenía el gobierno sobre el caso en cuestión.
- Redes *on-line*: Son redes que dependen altamente de los medios electrónicos y que se mantienen estrechamente ligadas a los cambios en la tecnología de los sistemas. Ejemplos de este tipo de redes son FaceBook, Twitter, Flirk y Youtube, entre otras. Según un artículo de la empresa CISCO[2], este tipo de redes genera actualmente el 30 % del tráfico total de las redes y se espera que su tráfico aumente en un 500 % para el 2013.

[2] http://newsroom.cisco.com/dlls/2009/prod_102109.html

d) **Su topología**. Esta clasificación se basa en la complejidad de la estructura que representa la red. De acuerdo a este criterio, las redes pueden clasificarse en *simples* o *complejas*:
   - Redes simples: Este tipo de redes son estructuras sencillas y pueden ser fácilmente analizadas con conceptos de teoría de grafos.
   - Redes complejas: Este tipo de redes son estructuras que presentan propiedades no triviales (*e.g.*, coeficiente de agrupamiento, centralidad, etc.) y cuyo estudio se basa en el estudio empírico de redes del mundo real [37]. Algunos ejemplos de este tipo de redes son las *redes aleatorias* [38], las de *mundo pequeño* [39,40], las de *ley de potencia* [41] y las *libres de escala* [42].
     - Redes aleatorias: Este tipo de redes son generadas a partir de un conjunto de datos fijos, agregando aristas entre un par de nodos seleccionados uniforme y aleatoriamente. Este tipo de redes presentan caminos muy cortos entre nodos, un coeficiente de agrupamiento bajo y además, presentan una distribución Binomial o de Poisson.
     - Redes de mundo pequeño: Este tipo de redes tienen un diámetro pequeño y además, un alto índice de agrupamiento. El problema del mundo pequeño fue planteado en 1967 por Stayley Milgram [39] y plantea que cada individuo está conectado a través de a lo sumo cinco personas; o lo que es lo mismo, entre cada par de personas existe un camino de longitud no mayor a 6 que los conecta.
     - Redes ley de potencia: Estas redes cumplen que la probabilidad de que un nodo tenga grado x es proporcional a $x^{-\alpha}$; donde la constante α es llamada coeficiente ley de potencia y es un valor que debe satisfacer $\alpha > 1$. Trabajos de investigación, como el desarrollado en [43], muestran que este tipo de redes siguen una *ley de Pareto*.
     - Redes libres de escala: Son una clase de redes leyes de potencia, en las cuales los nodos con altos grados de conexión tienden a estar conectados con otros nodos de alto grado; es decir, en este tipo de red el número de enlaces está concentrado en un número pequeño de nodos. Este tipo de red presenta una mejor distribución entre sus enlaces en comparación a las redes aleatorias. Muchas redes del mundo real presentan un comportamiento de libre de escala, por ejemplo: la estructura de la red celular [44] y la red de e-mail [45]. Por otra parte, otras redes utilizan las propiedades ofrecidas por este tipo de redes para medir su estructura, por ejemplo: la red formada por las ontologías en la Web semántica [46].

2.4. Problemas de SNA

El análisis de las redes sociales ha emergido como una metodología clave en diversas ciencias sociales, tales como la sociología, la antropología, la psicología social y la economía, entre otras; ganando además un apoyo significativo en la física y la biología. Este análisis produce una visión a la vez alternativa y complementaria a la de los estudios tradicionales en las Ciencias Sociales. En el análisis de redes sociales, los atributos de los individuos son menos importantes que sus relaciones y sus vínculos con otros individuos dentro de la red. Este enfoque ha resultado ser útil para explicar muchos fenómenos del mundo real.

Las redes sociales también se han utilizado para examinar las asociaciones y conexiones entre los empleados de diferentes organizaciones, así como para conocer cómo las organizaciones interactúan unas con otras, caracterizando las múltiples conexiones informales que vinculan a los ejecutivos entre sí. Por ejemplo, a menudo, el poder que tiene un individuo dentro de una organización proviene más del número de relaciones que este tenga con otros individuos dentro de la organización, que de su puesto de trabajo real [10].

Uno de los contextos que ha evidenciado la potencialidad del SNA, es el proceso de elecciones presidenciales de EUA. En este contexto, algunos de los aspectos que resultan de interés y que son investigados

a través del análisis de redes sociales, son la participación de los votantes en las elecciones y sus intenciones de votos. De forma general, una persona no toma las decisiones solamente con base en lo que ve en los medios, sino que recibe información nueva, interpretaciones e influencia de aquellas personas cercanas y con las que comparte actividades en común: amigos, familiares, vecinos, compañeros de trabajo, etc; es decir, integrantes de su red social. Luego, si la mayoría de las personas en una red social tienen intensiones de votar y se lo hacen saber al resto, existe una mayor probabilidad de que el resto de las personas de la red los sigan, ya sea por conveniencia o por competitividad.

Otra de las aplicaciones en las que ha sido utilizado el SNA es en el esclarecimiento de delitos y la desarticulación de redes de criminales. En este sentido, varias herramientas informáticas han sido desarrolladas. Una herramienta bien conocida en el contexto del análisis criminal es Analyst's Notebook (actualmente llamada IBM i2 Analyst's Notebook) [47]. Esta herramienta permite dibujar automáticamente una red social, a través de datos presentes en ficheros textos, e incluye técnicas de análisis de redes sociales que le permiten extraer información acerca de la estructura de la red. Otras herramientas desarrolladas son Sentinel Visualizer [48] y NetReveal [49]. La primera incluye módulos para el enfrentamiento de delitos y análisis de bandas criminales, mientras que la segunda es aplicada en el enfrentamiento del delito, detección de redes del crimen organizado, fraude y lavado de dinero. Otra herramienta es InfiniteInsight [50], de la compañía Kxen. Esta herramienta incluye técnicas de Minería de Datos como la clasificación, regresión, agrupamiento y reglas de asociación, entre otras, así como incluye técnicas de análisis de redes sociales. Esta herramienta puede ser aplicada para la detección de fraudes bancarios, lavado de dinero y enfrentamiento del delito.

Otro ejemplo del uso del SNA se puede encontrar en la MLB (Major Baseball League), la cual utilizó este tipo de análisis para investigar casos de consumo de esteroides entre jugadores. A partir de un reporte donde se mencionaban varios nombres de jugadores que se suponía habían consumido esteroides (*Mitchell Report*), la MLB llevó a cabo una investigaron para conocer qué otros jugadores podían haber consumido también. Por lo general, en la MLB hay frecuentes intercambios de jugadores entre equipos producto de los traspasos de jugadores y de que hay agentes libres (jugadores), que pueden firmar por algún otro equipo de interés si este le paga más. Esta situación hace que los jugadores vayan creando lazos con sus nuevos compañeros de equipo, a la vez que mantienen lazos con sus antiguos compañeros de equipo. Toda esto fue utilizado por la MLB para descubrir información relacionada con este caso. En la Fig. 2 se muestra la red formada por algunos jugadores de los equipos de Baltimore, Los Angeles y New York, que aparecían en el reporte antes mencionado. En esta red, los lazos entre jugadores se establecen si estos pasaron tiempo en común en algún equipo.

Otro ejemplo de aplicación en la que ha sido utilizado el SNA es para descubrir redes terroristas. En este contexto, las métricas utilizadas en el SNA pueden ayudar a responder preguntas tales como: ¿Cuál es el centro de la red?, ¿Cuáles son las personas que tienen mejores condiciones para transmitir informaciones en la red?, ¿Cuál o cuáles elementos hay que eliminar para desarticular la red?. A continuación, se comenta brevemente un caso de estudio, tomado de [51], que permite apreciar la potencialidad del SNA.

A principios del 2000, la CIA fue informada acerca de dos terroristas que se suponía mantenían contactos con al-Qaeda. Nawaf Alhazmi y Khalid Almihdhar fueron fotografiados en una reunión en Malasia, en unión con otros conocidos terroristas. Luego de dicha reunión, retornaron a Los Angeles en donde se habían establecido desde 1999. Ahora, ¿Qué hacer con esta información? ¿Arrestar a los sospechosos o deportarlos? No, lo más correcto sería utilizar esta información para tratar de descubrir a otros miembros de la red. Una vez que los sospechosos han sido identificados, se puede observarlos y escuchar sus llamadas para identificar ¿Quién o quiénes lo llaman o envían emails?, ¿Con quién se reúne, localmente y en otras ciudades? y ¿De dónde proviene su dinero?, entre otras cosas. A partir de esta investigación, se pudo descubrir que ambos sospechosos estaban relacionados con el principal sospechoso de un atentado

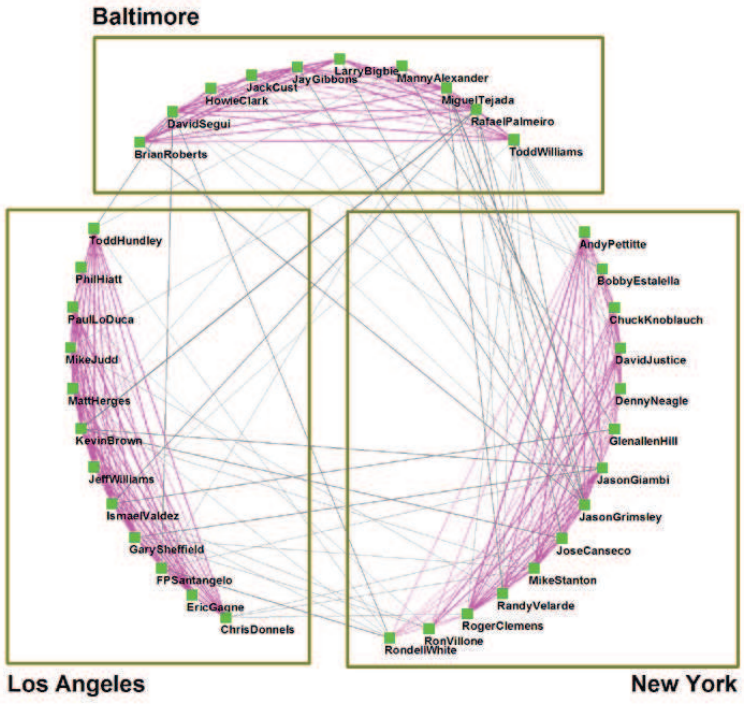

Fig. 2. Ejemplo de red social formada entre algunos jugadores cuyos nombres aparecían en el reporte de la MLB.

de bomba efectuado en EUA en octubre de 2000; una persona que además estuvo presente en la reunion que sostuvieron Alhazmi y Almihdhar en Malasia. La red social que se forma con la información obtenida hasta este momento se muestra en la Fig. 3.

Una vez que se tiene esta información, el próximo paso es buscar las relaciones indirectas de los dos sospechosos; es decir, buscar las relaciones directas de sus relaciones directas. En la Fig. 4, se muestra la red resultado de esta última búsqueda de información.

A partir de la red mostrada en la Fig. 4, se puede observar que los 19 terroristas relacionados con los atentados del 11 de septiembre en EUA (11-S) están, en el caso peor, a dos pasos de los sospechosos originales. Adicionalmente, las métricas de SNA revelan a Mohammed Atta (uno de los 19 terroristas implicados en el 11-S) como el líder local de la red. Utilizando esta información a tiempo se podría haber detenido los sucesos del 11-S. Obviamente, estos datos fueron determinados después del evento. Antes del evento y por la naturalidad de los datos (imprecisos y superfluos), es mucho más difícil descubrir la red.

2.5. Colecciones de prueba utilizadas en problemas de SNA

En la mayoría de los trabajos en los que se ha utilizado SNA, se ha trabajado con datos relacionados con el problema en cuestión; es decir, datos reales del problema analizado. Algunos de los datos provienen de redes sociales on-line como Facebook, Youtube, MySpace o Twitter, entre otras. Por otra parte, otra fuente de datos la constituye la telefonia IP, los emails, el comercio electrónico, etc. No obstante, esta información puede estar sujeta a confidencialidad.

Por lo general, en cuanto a repositorios o colecciones de prueba, se ha seguido dos variantes a la hora de probar algoritmos desarrollados para el SNA. La primera es trabajar con repositorios sintéticos o

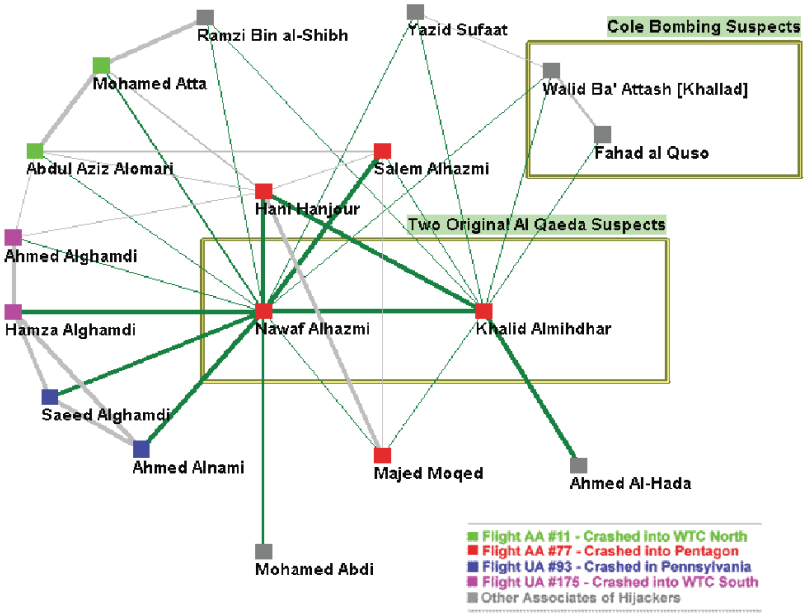

Fig. 3. Red formada con las personas directamente relacionadas con los dos sospechosos.

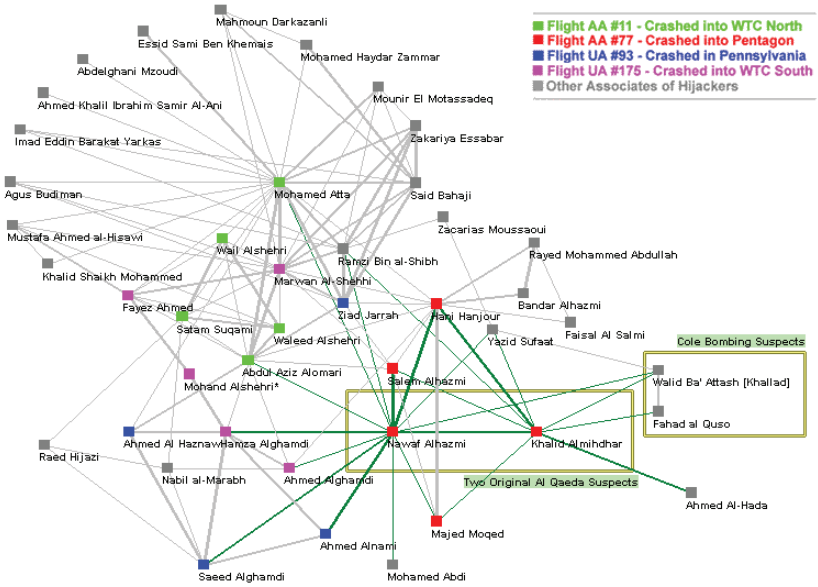

Fig. 4. Red formada con las personas directamente e indirectamente relacionadas con los dos sospechosos.

creados artificialmente. La segunda es trabajar con repositorios estándares, reportados en la literatura. A continuación se describen en detalle ambas alternativas.

Para la generación de repositorios sintéticos o artificiales se utilizan modelos que describen a través de parámetros, configurables por el usuario, el tipo de red social que se desea crear. Existen varios modelos propuestos en la literatura, dependiendo el tipo de red que se desea modelar o crear. Para las redes aleatorias se han propuesto los modelos de Gilbert [52] y el modelo de Erdös-Rényi [53,54,55]; estos modelos son considerados como los primeros modelos de redes sociales. Para las redes de mundo pequeño existe el modelo de Watts-Strogatz [40]; este modelo ha sido utilizado para estudiar el lenguaje humano [56] y la propagación epidemiológica [57]. Para las redes libres de escala existe el modelo Barabási-Albert [58].

Los datos sintéticos generados por las redes aleatorias permiten evaluar de manera controlada los algoritmos desarrollados para el SNA, debido a que la estructura es conocida desde un principio. Sin embargo, siempre es deseable también evaluar los algoritmos de SNA sobre redes del mundo real, las cuales exhiben una estructura diferente. Los repositorios conocidos para este tipo son:

- **Red del club de Karate de Zachary**. Esta red fue creada por el sociólogo Wayne Zachary y representa la interacción social entre 34 miembros de un club de karate [59]. Estos datos han sido utilizados en diferentes trabajos de investigación sobre detección de comunidades [24,32]. Esta red es no dirigida y cuenta con 34 nodos y 78 aristas.
- **Red de libros sobre políticos**. Esta red representa los libros vendidos por www.amazon.com, que hablan sobre políticos estadounidenses. Esta red es no dirigida y cuenta con 105 nodos y 441 aristas. En la misma, los nodos representan a los libros y una arista entre dos nodos representa la frecuencia en que un comprador adquiere los dos libros. Estos datos se encuentran disponibles en http://www.orgnet.com/.
- **Red de delfines**. Esta red representa la interacción entre un conjunto de delfines en Nueva Zelanda. Esta red es no dirigida y cuenta con 62 nodos y 159 aristas. Este conjunto de datos fue presentado y utilizado por primera vez por Lusseau [60].
- **Red de contactos de MySpace**. Esta red representa la interacción entre un conjunto de usuarios de MySpace; los nodos son los usuarios y los enlaces entre los nodos están determinados por las relaciones de los usuarios en MySpace. Esta red es dirigida y cuenta con 100 000 nodos y 6 865 571 aristas. Además, es utilizada en el trabajo de Ahn *et al.* [61].
- **Red de recomendaciones de Amazon**. Esta red representa la interacción entre un conjunto de clientes de www.amazon.com. La red es dirigida y cuenta con 262 111 nodos y 1 234 877 enlaces; donde los nodos representan a los clientes y cada enlace representa la acción de que un cliente recomendó un libro a otro cliente. Esta red es utilizada por primera vez en el trabajo de Leskovec, Adamic y Huberman [62].

Como se puede apreciar de la lista anterior, los repositorios existentes difieren respecto al número de nodos y aristas. Por otra parte, existen redes como la red de políticos y la red de recomendaciones, que son obtenidas del mismo sitio (www.amazon.com) pero con diferentes enfoques. Además de los repositorios anteriores, se encuentran los repositorios SBNS, Ling Spam y PU1. El primero, es de la Escuela de Ciencia de la Computación en la Universidad de Carnegie Mellon (EUA) y puede obtenerse del sitio web http://www.autonlab.org/autonweb/downloads/datasets.html. Los otros dos se pueden obtener del sitio http://www.aueb.gr/users/ion/publications.html.

3. Visualización y obtención de redes sociales

Un método importante para descubrir propiedades de las redes sociales, aunque tiene menos peso teórico en el análisis, es la visualización de las mismas. La visualización tiene la ventaja de alimentar rápidamente la intuición del investigador. Visualizar redes complejas es un gran desafío; por lo general, se busca presentar gran cantidad de información de forma estética. Se busca la claridad y la simpleza, pese a la gran complejidad de los datos, como se ilustra en el ejemplo de la Fig. 5. Hay que considerar que hay muchas potenciales vistas de los datos, que pueden ilustrar propiedades diferentes: centralidad, comunidades, jugadores clave (que, si desaparecen, desconectan la red), etc.

Fig. 5. Visualización de una red social.

Existe una gran variedad de algoritmos para visualizar redes sociales. Cada uno obedece a una idea u objetivo diferente, aunque muchas veces se busca la presentación instantánea. Muchos investigadores [63,64,65,66] han enfocado sus trabajos en cómo representar la información mediante el uso de sistemas. En el caso del análisis de redes sociales existen sistemas (p.ej., Pajek, UCINET, GUESS, PAJEK, etc) que permiten manipular redes sociales de manera gráfica y además permiten realizar cálculos sobre estos datos. En el cuadro 1 se muestran los programas más empleados en los trabajos de investigación [67] sobre la visualización y el análisis de redes sociales. Algunos de los sistemas existentes incorporan análisis más detallados como el caso de SoNIA que permite visualizar redes dinámicas [68,69].

Tabla 1. Programas más utilizados en el análisis de redes sociales.

Programa	Versión	Objetivo	Formato de Entrada	Capacidad
MultiNet	(4.24)	Análisis contextual	dat	5000 nodos
NetDraw	(1.0)	Visualización	mat y dat	10000 nodos
NetMiner	(3.4)	Análisis visual	mat y dat	1000 nodos
Pajek	(1.24)	Análisis y visualización	mat y dat	10000 nodos
StOCNET	(1.8)	Análisis estadístico	mat	5000 nodos
UCINET	(6.05)	Comprensivo	mat y dat	4000 nodos
GUESS	(1.0.3)	Análisis y visualización	gdf	8000 nodos
SIENA	(3.1)	Análisis estadístico	dat	3000 nodos
SoNIA	(1.2)	Análisis dinámico	mat	4000 nodos
GNU R	(2.0)	Análisis estadístico	dat,mat	100000 nodos

Como se observa en el Cuadro 1, los progamas de visualización permiten realizar diferentes tipos de análisis como: análisis contextual, análisis estadístico, análisis visual y análisi dinámico. Por otro lado, las capacidades de nodos (tamaño de la red) que aceptan van desde los 1000 hasta los 100000 nodos.

3.1. Obtención de redes sociales

Para muchos investigadores la información proporcionada por la *Web* representa una fuente de datos muy generosa y en muchos sentidos ilimitada. La necesidad de obtener información de la *Web* para representar problemas en particular, ha permitido que muchos investigadores centren sus estudios sobre cómo obtener dicha información [70,71,72].

En la actualidad los motores de búsqueda (p.ej., Google, Yahoo, Bing, entre otros) son una importante herramienta para hallar contenido en la *Web*, es por eso que diferentes estudios han propuesto sistemas para obtener información usando estas herramientas; tal es el caso de *Flink* [71] y *POLYPHONET* [70] que explotan a los motores de búsqueda para obtener información de la *Web* utilizando los resultados de las búsquedas para estudiar y/o analizar diversos contenidos.

Sin embargo, un proceso de extracción no es tarea sencilla, ya que una mala elección de los datos recolectados puede generar resultados muy diferentes a los esperados. Trabajos de investigación [73] proponen metodologías de extracción para el contenido oculto en la *Web* [74], donde se revela que los resultados proporcionados por los motores de búsqueda no siempre son completos y siempre existe información que se mantiene oculta y que no es presentada en la lista de resultados de la consulta realizada.

Los sistemas de redes sociales en línea permiten obtener una gran cantidad de información debido al elevado número de sistemas y de usuarios que participan en estos sistemas. Mucho del contenido que se encuentra almacenado en estos sistemas es del tipo personal, por lo que el nivel de privacidad de algunos sistemas ha permitido el desarrollo de investigaciones [75] enfocadas a proteger dicha información de ataques de terceros [76].

La forma en cómo se puede representar una red social es generalmente mediante un grafo dirigido; sin embargo, existen redes que se pueden representar como grafos no dirigidos (p.ej., Wikipedia, Messenger, etc). Utilizando las API que proporcionan los sistemas de redes sociales en línea se puede generar un grafo de dicho sistema, donde los actores representan a los usuarios del sistema y las relaciones representan las diferentes interacciones entre usuarios (p.ej., compartición de fotos, envío de mensajes, entre otros).

La forma en cómo interactúan las personas con los sistemas de redes sociales en línea hace de las redes sociales aún más atractivas, ya que la representación de una red puede ser diferente dependiendo del enfoque de estudio. Sin embargo, el proceso de extracción tiene sus inconvenientes como son la pérdida de información al momento de representar la red, según estudios de investigación [77] se muestra que la pérdida de información durante el muestreo puede afectar los resultados de las mediciones.

3.2. Tipos de muestreos aplicados a las redes sociales

En [78] se proponen 3 tipos de métodos para muestreo aplicados en redes sociales, estos son:

- **Muestreo bola de nieve.** Este tipo de muestreo es el más utilizado para obtener redes sociales de un conjunto de información grande y además por su funcionamiento permite representar satisfactoriamente a las redes dirigidas (como el caso de los sistemas de redes sociales en línea). En este tipo de muestreo se empieza con una semilla (un nodo seleccionado aleatoriamente) y posteriormente se enlaza a todos los nodos directamente conectados a este, el proceso de realiza recursivamente para cada

uno de los nodos directamente conectados al del paso anterior. En la Fig. 6(a) se muestra el proceso para este tipo de muestreo.

- **Muestreo de nodo.** Este método consiste en seleccionar un número *n* de nodos de la red original aleatoriamente y posteriormente relacionarlos con los enlaces de la red inicial que existen entre los *n* nodos seleccionados, en la Fig. 6(b) se muestra el proceso para el muestreo de la red original usando este método. Los nodos que son seleccionados y no se enlazan con ningún otro nodo (nodos aislados) son removidos de la nueva red.
- **Muestreo de enlace.** Este método difiere del muestreo de nodo, ya que en vez de seleccionar nodos, seleccionan enlaces de manera aleatoria, para este caso no existen nodos o enlaces aislados. En la Fig. 6(c) se muestra el proceso para el muestreo de enlace.

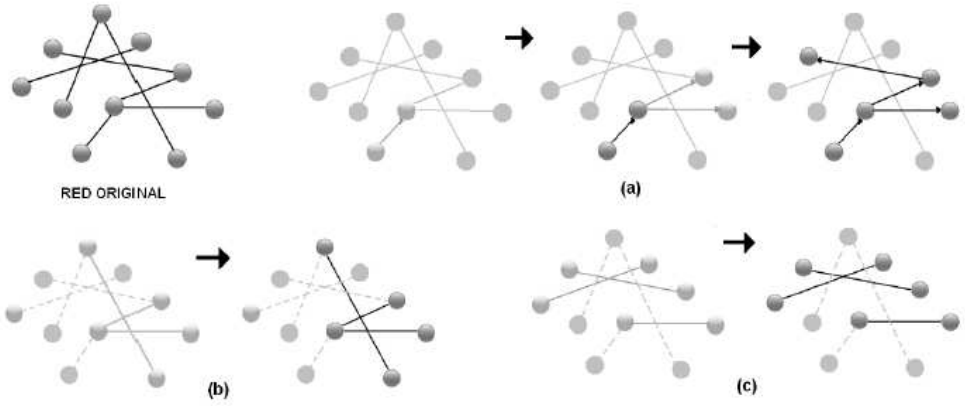

Fig. 6. Tipos de muestreos para una red social. (a) Muestro bola de nieve, (b) Muestreo de nodo y (c) Muestreo de enlace.

3.3. Sistemas para extracción de redes sociales

Como se mencionó anteriormente, existen diferentes técnicas para obtener información de la *Web*, algunos trabajos [70,71] utilizan el poder de los motores de búsqueda para encontrar contenido de la *Web* y poder hacer uso de esta información para diferentes fines. En las siguientes secciones se describen tres formas, dos son sistemas (*Flink* y *POLYPHONET*) que explotan los resultados de las consultas generadas por los motores de búsquedas particularmente Google, y la tercera es el uso de un API de desarrollo de los sistemas de redes sociales en línea.

3.3.1. Flink

Flink [71] es un sistema para extracción, agregación y visualización de redes sociales en línea. *Flink* emplea tecnología semántica para el razonamiento de información personal extraída de fuentes electrónicas, incluyendo páginas *Web*, correos electrónicos, publicaciones y archivos FOAF[3]. En la Fig. 7 se muestra la arquitectura de *Flink* desde la adquisición de metadatos (arriba) hasta la interfaz de usuario (abajo).

[3] Los archivos FOAF (*Friend Of A Friend*) son archivos compartidos por los amigos de un amigo

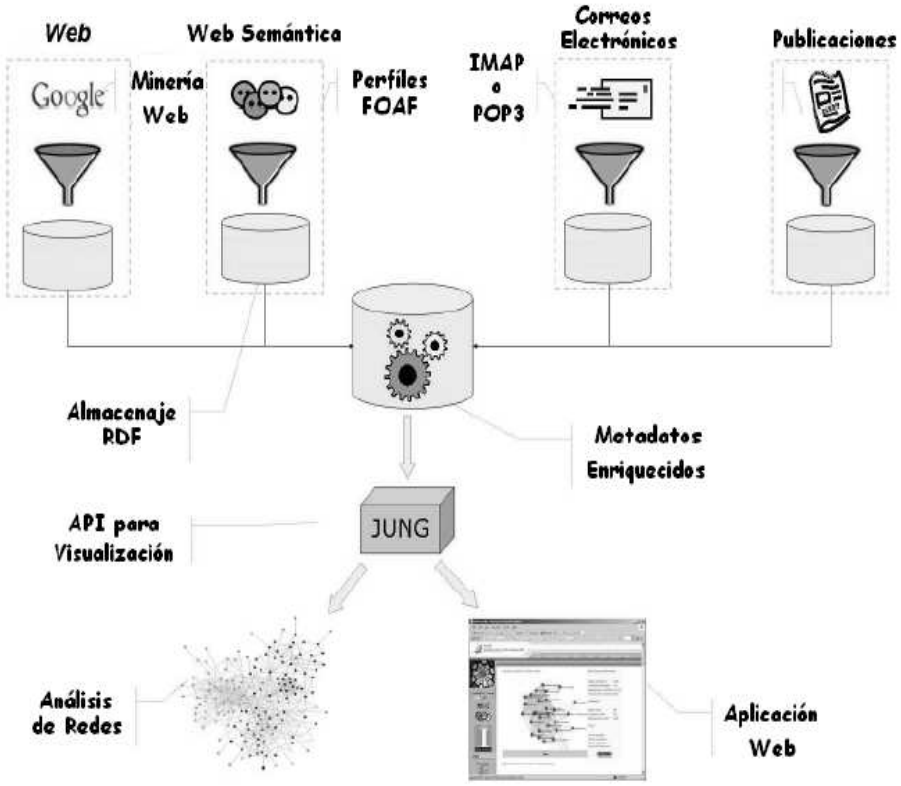

Fig. 7. Arquitectura de Flink. Adquisición de datos (arriba) e Interfaz de usuario (abajo).

Flink utiliza diferentes fuentes de datos para obtener información, después la integra en un base de datos enriquecida y utiliza JUNG[4] para visualizar la información. *Flink* solo utiliza la información con la que cuentan ciertos sistemas y no puede ver información de otros sitios de redes sociales, sin embargo, resulta ser un buen intento para unificar el proceso de extracción de redes sociales.

3.3.2. Polyphonet

En Matsuo *et al.* [70] se propone un conjunto de algoritmos para la extracción de redes sociales de la *Web*, y se integran en un sistema de minería de redes sociales llamado *POLYPHONET*. Este sistema se divide en dos fases, la primera está basada en definir los nodos de la red, y de manera similar que en *Flink* se da una lista de personas (lista de nodos), la segunda fase consiste en encontrar las aristas usando motores de búsqueda especificamente *Google*.

Es importante señalar que el algoritmo para extraer redes sociales utilizado en *POLYPHONET* sólo es aplicado para contenido que se encuentra referenciado por los motores de búsqueda, algo que limita la búsqueda de la información dentro de los sistemas de redes sociales. En la siguiente sección se muestra otro enfoque para la búsqueda de información sobre las redes sociales, basado en el conocimiento de la red social.

[4] *Java Univeral Network Graph* por sus siglas en inglés, es un *framework* hecho en java que facilita la tarea de la visualización de grafos.

3.3.3. OpenSocial y las interfaces de programación de aplicaciones

Sitios como *MySpace*, *Orkut* y *hi5* forman parte de un servicio propuesto por *Google* llamado *OpenSocial*[5], el cual proporciona una colección de funciones y tipos de datos ofrecidos en una biblioteca para el desarrollo de aplicaciones bajo cierto lenguaje de programación, también llamada API[6]. Esto permite a desarrolladores externos crear aplicaciones para los sitios de redes sociales que aceptan estos estándares. Muchos otros son los sitios que proporcionan sus API de manera independiente (p.ej., *Flickr*, *Facebook*, etc) con lo cual se facilita la tarea de búsqueda de contenido para diversos fines.

Este servicio cuenta con alrededor de 35 sistemas (p.ej., *Hi5*, *MySpace*, *Xing*, *LinkedIn*, *Orkut*, etc.) que se han ido incorporado poco a poco, este servicio permite crear aplicaciones en lenguajes como son JavaScript y HTML y proporciona un soporte de versiones de las diferentes API's.

4. Técnicas de Minería de Datos para el análisis de redes sociales

En la presente sección se presentarán las principales técnicas de Minería de Datos empleadas en el análisis de redes sociales. De forma general, son el agrupamiento (sección 4.1), la minería de grafos (sección 4.2) y la clasificación (sección 4.3) las técnicas más utilizadas; siendo otras como el minado de secuencias y el minado de conjuntos frecuentes menos empleadas y por tanto, no abordadas en este reporte.

4.1. Agrupamiento

De forma general, la forma en que los individuos de una red social interactúan determina grupos, de acuerdo a intereses comunes, funciones en las que participan los individuos, objetivos comunes o afiliaciones. En el contexto de las redes sociales, a estos grupos también se les denominan *comunidades*.

En el mundo real, las comunidades no tienen que estar desconectadas o aisladas unas de otras, sino que pueden comunicarse entre sí a través de algunos enlaces entre sus miembros. Por ejemplo: un miembro de la comunidad $M1$ trabaja en la misma empresa que un miembro de la comunidad $M2$. Precisamente, lo anterior es una de las cosas que puede hacer compleja la detección de comunidades en las redes sociales. Cuando se habla de una estructura formal o cuando existe un grupo o comunidad bien definido y un conjunto de personas que dicen pertenecer al mismo, entonces la detección de las comunidades es sencilla; por ejemplo, un país y sus habitantes. Sin embargo, cuando se está tratando con estructuras informales el problema es mucho más complicado. Supóngase que se tiene un grupo de amigos. Este grupo está formado por personas las cuales, en su mayoría, son amigos entre sí; luego, es lógico pensar que este grupo constituye una comunidad. No obstante, cada persona del grupo pudiera tener amigos fuera del grupo, y estos a su vez tener otros amigos. El problema aquí es determinar la frontera del grupo; es decir, ¿cuáles de estas personas pertenecen al grupo y cuáles no?

Otro problema en la detección de comunidades es cómo asumir la pertenencia de los individuos a las comunidades. Existen situaciones las cuales exigen que los individuos sólo pertenezcan a una sola comunidad; por ejemplo, un libro sólo se puede publicar en una editoral. Sin embargo, de forma general los individuos no tienen por qué pertenecer sólo a una comunidad; este tipo de comunidades que pueden compartir miembros entre sí se denominan *traslapadas* o *con traslape*. Por ejemplo, en las redes sociales *on-line*, una persona puede pertenecer a un grupo que comparte preferencias por cierto género de películas

[5] Para más información sobre *OpenSocial* y los sistemas que lo integran visitar http://code.google.com/intl/es-ES/apis/opensocial/

[6] *Application Programming Interface* o Interfaz de Programación de Aplicaciones por su traducción del inglés.

y, a la vez, pudiera pertenecer también a otro grupo que comparte gusto por la cocina. Otro ejemplo, se puede encontrar en redes de proteínas, en las cuales es común detectar proteínas que cumplen más de una función. De forma general, las redes sociales del mundo real presentan comunidades con traslape [36] y en los últimos años, las investigaciones se han centrado en la detección de este tipo de comunidades [79,80]. Por esta razón, esta sección se centra en los algoritmos desarrollados para detectar este tipo de comunidades.

Algunos de los algoritmos mas importantes, reportados para la detección de comunidades con traslape son [79,80,81,82,29,83,84,85,86,87]. Estos algoritmos representan la red social como un grafo (dirigido o no dirigido, pesado o no pesado; dependiendo del problema específico), donde los vértices son los objetos o individuos de estudio y las aristas representan relaciones de interés entre dichos individuos. A continuación se describen brevemente los algoritmos anteriores, exponiendo sus limitaciones.

Uno de los primeros algoritmos propuestos fue introducido por Palla *et al.* en el 2005, para el estudio de redes sociales y biológicas [86]. Este algoritmo trabaja sobre redes binarias (es decir, con aristas no dirigidas y no pesadas) y define como comunidad a la unión de todos los $k-$*cliques* (subgrafos completos de tamaño k) que pueden ser alcanzados de unos a otros, a través de una serie de $k-$*cliques* adyacentes; donde, dos $k-$*cliques* son adyacentes si comparten $k-1$ vértices. Este método primero detecta todos los cliques (subgrafos completos maximales) presentes en la red y, a continuación, detecta cuáles son las comunidades, llevando a cabo una búsqueda estándar de componentes sobre una matrix que contiene la información del traslape entre los cliques detectados [88]. La principal limitación de este algoritmo es su complejidad computacional, la cual puede llegar a ser exponencial; esto limita mucho su aplicación en problemas reales que analicen redes de muchos nodos y enlaces.

Otros algoritmos desarrollados para el descubrimiento de comunidades traslapadas son RaRe-IS [79] y LA-IS^2 [80]. Estos algoritmos representan la red con un grafo no dirigido y no pesado. El algoritmo RaRe-IS [79] consiste en dos etapas: *Inicialización* y *Mejoramiento*. En la etapa de inicialización, RaRe-IS aplica un método de nombre RaRe, el cual construye un conjunto de grupos *semillas*. Para construir este tipo de grupos, RaRe comienza ordenando los vértices del grafo de acuerdo a algún criterio predefinido; por ejemplo, el grado de los vértices o el Page Rank [89]. Posteriormente, se van eliminando iterativamente los vértices con altos valores del criterio de orden utilizado, hasta que el grafo queda descompuesto en componentes conexas que tienen un tamaño indicado por un umbral mínimo y máximo predefinido. Luego, los vértices eliminados son adicionados a las componentes sólo si esta operación mejora de alguna forma alguna la densidad de la componente. Una vez que los grupos semillas son construidos, en la etapa de mejoramiento se aplica un método de nombre IS sobre cada uno de estos grupos. El método IS actualiza iterativamente el grupo, adicionando o eliminando un vértice a la vez, si el valor de alguna métrica sobre el grupo mejora. La complejidad computacional de este método es $O(n^2)$.

La principal limitación de este método es que necesita ajustar valores de al menos cuatro parámetros, cuyos valores dependen de la colección que se desee procesar. Por lo general, los usuarios no tienen conocimiento previo acerca de la colección que desean agrupar; por lo tanto, ajustar diferentes parámetros puede ser una tarea muy complicada. Adicionalmente, como se planteó en [82], el algoritmo RaRe-IS puede producir agrupamientos con mucho traslape. Aunque bien es cierto que formar grupos con traslape es importante y útil para muchas aplicaciones, cuando el nivel de traslape de los grupos es alto se hace muy difícil la obtención de conclusiones útiles y en muchos casos, un alto nivel de traslape puede significar una mala estructuración [29,83].

El algoritmo LA-IS^2 [80] sigue una idea similar a la de RaRe-IS, pero introduce nuevos métodos para las etapas de inicialización y mejoramiento. Para construir los grupos semillas en la etapa de inicialización, LA-IS^2 utiliza un método llamado LA (de *Link Aggregate*). Este método comienza ordenando los vértices de la red en orden decreciente de su Page Rank. A continuación, LA procesa iterativamente los vértices

siguiendo una estrategia similar a la del algoritmo Single Link [90]; es decir, LA adiciona cada vértice a los grupos si esta operación mejora la densidad del grupo; en otro caso, si el vértice no fue adicionado a ningún grupo, entonces dicho vértice constituye por si solo un nuevo grupo. Posteriormente, en la etapa de mejoramiento LA-IS^2 procesa secuencialmente cada grupo semilla, utilizando el método IS^2. Este método es una variación del método IS utilizado por RaRe-IS. En esta variante, los vértices se adicionan a los grupos semillas solamente si estos vértices tienen algún vértice adyacente dentro del grupo semilla. La complejidad computacional del algoritmo LA-IS^2 es $O(n^2)$. La principal limitación de este algoritmo es que realiza asignación irrevocable de los objetos a los grupos; es decir, una vez que un objeto es adicionado a un grupo este no puede ser eliminado del grupo y adicionado a otro, aún cuando esta operación mejore la calidad del agrupamiento formado.

Otro algoritmo de agrupamiento desarrollado para la detección de comunidades con traslape es LA-CIS, propuesto por Goldberg *et al.* [82]. Este algoritmo sigue la idea de LA-IS^2 [80] pero, a diferencia de este, en la etapa de mejoramiento aplica una variación del método IS, llamada CIS. Existen dos diferencias principales entre IS y CIS. La primera es que antes de actualizar un grupo semilla CIS ordena los vértices de acuerdo a su grado. La segunda es que CIS comprueba la conectividad del grupo semilla, cada vez que procesa a todos los vértices; si el grupo consiste en multiples componentes conexas, entonces el grupo es reemplazado por la componente conexa que tenga la mayor densidad. La complejidad computacional de LA-CIS es $O(n^2)$ y, como LA-IS^2, la principal limitación del algoritmo LA-CIS es que realiza asignación irrevocable de los objetos a los grupos. Además, LA-CIS puede producir grupos con alto traslape.

Otros dos algoritmos de agrupamiento desarrollados para descubrir comunidades traslapadas son CONGA [29] y CONGO [83]. Estos algoritmos son variaciones del algoritmo GN [32]. El algoritmo GN construye un conjunto de grupos disjuntos utilizando una estrategia compuesta de cuatro pasos: 1) Calcular el *edge betweenness* (EB) para cada arista del grafo, 2) Buscar la arista con mayor valor de EB y eliminarla, 3) Recalcular el valor de EB de las aristas restantes del grafo, y 4) Repetir los pasos 2 y 3 hasta que no queden aristas.

El algoritmo CONGA [29] extiende el algoritmo GN, permitiendo el traslape entre los grupos. Para este propósito, CONGA introduce el concepto de *split betweenness* (SB) de los vértices del grafo; este concepto provee una forma de decidir cuándo dividir un vértice y qué vértice dividir. Así, CONGA modifica el algoritmo GN de forma tal que en el paso 1, se calcula el EB de cada arista y también el SB de cada vértice. Luego, en el paso 2, CONGA elimina la arista con mayor valor de EB o el vértice con mayor valor de SB. En el tercer paso se recalcula el EB de las aristas y el SB de los vértices; el cuarto paso queda igual que en el algoritmo GN. La complejidad computacional de CONGA es $O(n^6)$ para redes estándares y puede llegar a ser $O(n^3)$ para redes dispersas. La principal limitación de este algoritmo es su complejidad computacional, la cual lo hace poco útil en problemas reales con redes de gran tamaño. Además, CONGA necesita conocer a priori el número de grupos a formar; no obstante, este número es desconocido en la mayoría de los problemas reales.

El algoritmo CONGO [83] extiende el algoritmo CONGA, introduciendo el concepto de *local betweenness*. Este concepto permite, en el paso 3 del algoritmo CONGA, no recalcular el EB de todas las aristas del grafo ni el SB de todos los vértices del grafo. En cambio, sólo se recalculan estas propiedades para los vértices y aristas que pertenecen a una pequeña región alrededor del vértice que se dividió o de la arista que se eliminó. Los pasos 1, 2 y 4 de CONGO son los mismos que en CONGA. La complejidad computacional del algoritmo CONGO es $O(n^4)$ para redes estándares y puede llegar a ser $O(n^2 \cdot \log n^2)$ para redes dispersas, pero haciendo unas asunciones muy duras, como se muestra en [83]. La principal limitación de este algoritmo es su complejidad computacional. Además, CONGO necesita conocer a priori el número de grupos a formar; no obstante, este número es desconocido en la mayoría de los problemas reales.

Otro algoritmo de agrupamiento capaz de formar grupos con traslape es H-FOG [81]. Este algoritmo transforma el problema de agrupamiento de un conjunto de vértices del grafo, en el problema de agrupar el conjunto de aristas del mismo; de esta forma, se permite que los grupos tengan traslape entre sí. H-FOG está basado en probabilidades y utiliza una estrategia similar a la del Single Link [90]. La complejidad computacional de H-FOG es $O(n^6)$. La principal limitación de este algoritmo es su complejidad computacional. Además, necesita conocer a priori el número de grupos a construir.

Otro algoritmo que puede construir comunidades con traslape es RRW [84], diseñado para el descubrimiento de *complexes* y *pathways* en redes de proteínas. Este algoritmo comienza construyendo un vector de afinidad para cada vértice de la red, utilizando la técnica de *recorrido aleatorio* (*random walk technique*). Luego, utiliza el vector de afinidad de cada vértice para construir las componentes fuertemente conexas del grafo; en este contexto, cada componente fuertemente conexa determina un grupo. Posteriormente, ordena los grupos de acuerdo a su significancia estadística y los post-procesa con el objetivo de eliminar aquellos grupos que no satisfagan un umbral de traslape previamente definido; el agrupamiento final está compuesto de los grupos restantes. La complejidad computacional del algoritmo es $O(n^2)$. La principal limitación de RRW es que necesita ajustar los valores de al menos cuatro parámetros.

Otro algoritmo desarrollado para encontrar comunidades con traslape es SSDE-Cluster [85]. Este algoritmo emplea una estrategia compuesta de tres pasos. En el primer paso, SSDE-Cluster utiliza Distancia Muestral y Espectral Embebida (SSDE, por sus siglas en inglés), para embeber un grafo en un número $d << n$ de dimensiones [91]; siendo n el número de vértices del grafo. En el segundo paso, los vértices son agrupados utilizando un Modelo de Mezclas Gaussianas (GMM, por sus siglas en inglés), entrenado utilizando el algoritmo E-M. Luego, el GMM calcula la probabilidades a posteriori y, con base en estas, construye un conjunto de grupos con traslape en el tercer paso. La complejidad del algoritmo es $O(n \cdot k \cdot d)$; siendo k el número predefinido de grupos a formar. La principal limitación de SSDE-Cluster es que necesita ajustar los valores de al menos cuatro parámetros, incluido el número de grupos a formar.

Otro algoritmo que es capaz de construir comunidades traslapadas en redes complejas fue propuesto por Zhang *et al.* [87]. Este algoritmo utiliza *mapeo spectral* sobre la matriz de adyacentes de la red, con el objetivo determinar o formar K vectores de valores propios, que representen la información contenida en la red. Posteriormente, se utiliza el algoritmo Fuzzy C-Means para agrupar estos vectores en k grupos, con $2 \leq k \leq K$. Todas estas particiones difusas son convertidas en agrupamientos con traslape, utilizando un parámetro λ. Finalmente, el agrupamiento con traslape que maximize una variación de la función modular, propuesta en [32], es seleccionado como el agrupamiento final. Una limitación fundamental de este algoritmo es su complejidad computacional que, en el mejor escenario, puede ser $O(n^2 \cdot K \cdot h)$. Otra limitaciones es que necesita ajustar los valores de a varios parámetros, incluido el número de grupos a formar.

Como se puede observar de lo anteriormente expuesto, existen varios algoritmos desarrollados para la búsqueda de comunidades con traslape. Sin embargo, como se describió en esta sección, los algoritmos reportados tienen un conjunto de limitaciones, las cuales pueden reducir su utilidad en problemas reales. Las limitaciones están principalmente relacionadas con: (*a*) la necesidad de ajustar los valores de varios parámetros, cuyos valores dependen de la colección a agrupar, y (*b*) la producción de agrupamientos con alto traslape. Además, existen reportados varios algoritmos que tienen una alta complejidad computacional, por lo que resultan poco útiles en problemas reales.

4.2. Minería de grafos

Generalmente, las redes sociales son representadas como grafos, observándose diversas líneas de investigación en la búsqueda de patrones en grafos obtenidos de los enormes volúmenes de datos disponibles. Una de estas líneas considera la dinámica intrínseca de las redes sociales.

En la actualidad, el estudio del dinamismo en las redes sociales se ha convertido en una necesidad. La mayoría de tales redes cambian en el tiempo; por ejemplo, se añaden/eliminan vértices (actores o entidades), aparecen/desaparecen enlaces entre ellos, etc. Por tal razón, en los grafos de estas aplicaciones se debe tener en cuenta una componente temporal. En tal sentido, se han utilizado los grafos dinámicos.

Formalmente, un grafo dinámico se representa como una sucesión temporal $S = \{F_1, F_2, \ldots, F_t, \ldots\}$, donde cada F_t es un grafo conocido como la fotografía o instantánea del grafo dinámico en el instante t [16] [92] [93].

En los grafos dinámicos que aparecen en aplicaciones reales, particularmente los de las redes sociales, se han realizado estudios que describen la evolución del diámetro en el tiempo [94]. En estos trabajos, se ha concluido que dicha magnitud se va reduciendo hasta estabilizar su valor en el tiempo; este fenómeno es conocido como densificación. Además, se han encontrado relaciones estadísticamente significativas entre el número de vértices y el número de enlaces en el tiempo.

Por otro lado, algunas propiedades, tales como el tamaño y el diámetro de la segunda y tercera componente conexa, han sido caracterizadas estadísticamente en el tiempo [95]. Además, el espectro de la matriz de adyacencia asociada con la mayor componente conexa ha mostrado propiedades interesantes. En el caso de los grafos dinámicos ponderados, se encontraron otros comportamientos en el espectro de la matriz de adyacencia ponderada.

En este epígrafe se presentará la manera en que las técnicas de minería de grafos han sido usadas para caracterizar la evolución en el tiempo de las redes sociales. Esto último consiste en dar respuesta a las siguientes preguntas: ¿Cuáles son las reglas o leyes que determinan los cambios en la red social a través del tiempo? o ¿Cuáles reglas o leyes se observan con frecuencia en las redes sociales a gran escala?

4.2.1. Minería de subgrafos periódicamente recurrentes

Supóngase que se tiene una red social que es representada mediante un grafo dinámico S, donde cada fotografía $F_t = \langle V_t, E_t, \phi_t \rangle$ es un grafo simple, cada V_t es un subconjunto del conjunto único de vértices V llamado universo, E_t es el conjunto de aristas, $\phi_t : E_t \to V_t \times V_t$ es la función de incidencia, y E_t cambia siempre en cada fotografía, o sea $E_{t_1} \cap E_{t_2} = \emptyset$ para cualesquiera sean $t_1 \neq t_2$.

En la minería de subgrafos periódicamente recurrentes (SPR), el dinamismo de la red es capturado para una sub-secuencia $S_t = \{F_1, F_2, \ldots, F_t\} \subset S$, en la serie de instantes de tiempo $T = \{1, 2, \ldots, t\}$ previamente seleccionada por un analista. Asociado con este problema se tienen dos conceptos: Núcleo de dos fotografías y Subgrafo periódico.

El núcleo de dos fotografías F_{t_1} y F_{t_2} se define como el grafo $S[t_1, t_2] = (V, E, \phi)$, donde $V = V_{t_1} \cap V_{t_2}$, y se cumple que para todo par de aristas $e_1 \in E_{t_1}$ y $e_2 \in E_{t_2}$ tal que $\phi_{t_1}(e_1) = \phi_{t_2}(e_2) \subseteq V$ existe una arista $e \in E$ tal que $\phi(e) = \phi_{t_1}(e_1) = \phi_{t_2}(e_2)$. El núcleo de $n > 2$ fotografías F_{t_1}, F_{t_2}, ..., F_{t_n} se define, a partir del concepto anterior, de una manera recursiva como $S[t_1, t_2, \ldots, t_n] = S[S[t_1, t_2, \ldots, t_{n-1}], t_n]$.

Para cada subconjunto I, $I \subseteq T$, se utiliza la notación $S[I]$ para referirse al núcleo de las fotografías en los instantes de I.

Para cada tripleta (t_0, p, s) de enteros $t_0 \geq 1$, $p \geq 1$ y $s \geq \sigma$ tal que $t_0 + p(s-1) \leq t$, siendo σ un umbral previamente definido, se define la secuencia aritmética $A(t_0, p, s) = \{t_0, t_0 + p, \ldots, t_0 + p(s-1)\}$.

Se dice que F es un subgrafo periódico de S_t si para alguna tripleta (t_0, p, s) con las condiciones antes mencionadas se tiene $F = S[A(t_0, p, s)]$ y las fotografías F_{t-p} y F_{t+ps} no contienen a F. En este caso la secuencia $A(t_0, p, s)$ representa el soporte periódico de F.

Tabla 2. Métodos reportados en la literatura para la minería de SPR

Método reportado	Complejidad temporal	Descripción
Apostólico et al. [93]	$O(\|V\| + \frac{t^2}{\sigma}\|\widehat{E}\|)$	Encuentra todas las subestructuras periódicas y propone una forma de garantizar que el soporte sea refinado.
Apostólico et al. [96]	$O((\|V\| + \|E\|)T^2 \ln(\frac{t}{\sigma}))$	Es similar a la de Lahiri & Berger-Wolf [97], pero incluye el filtrado de las subestructuras cuyo soporte no sea refinado.
Lahiri & Berger-Wolf [97]	$O((\|V\| + \|E\|)T^3 \ln t)$	Encuentra subestructuras sin garantizar que el soporte sea refinado.

Un subgrafo periódico F de S_t puede tener varios soportes periódicos. Por las condiciones de maximalidad, si existen soportes periódicos de F con el mismo período p estos obligatoriamente son disjuntos. Si se tienen dos soportes periódicos distintos de F, tales que $A(t_1, p_1, s_1) \subset A(t_2, p_2, s_2)$, entonces decimos que $A(t_2, p_2, s_2)$ incluye a $A(t_1, p_1, s_1)$. Si un soporte periódico no está incluido en ningún otro entonces es llamado refinado. Una subestructura periódica de S_t se define como un par ordenado $(F, A(t_0, p, s))$ donde F es un subgrafo periódico y $A(t_0, p, s)$ es un soporte periódico refinado F.

Dado una sub-secuencia S_t de un grafo dinámico y un umbral $\sigma \geq 2$, la minería de SPR consiste en determinar todas las subestructuras periódicas de S_t.

En el Cuadro 2 se muestra un resumen de los métodos reportados en el estado del arte para este tipo de minería. Esta tabla incluye la complejidad en tiempo de estos métodos para obtener el resultado de la minería.

4.2.2. Reglas de evolución de grafos

Las reglas de evolución de grafos (GERs, por sus siglas en inglés *Graph Evolution Rules*) constituyen una herramienta muy útil para estudiar la evolución de las redes sociales [92].

Supóngase que se tiene una red social que es representada mediante un grafo dinámico S, donde cada fotografía $F_t = \langle V_t, E_t, \phi_t \rangle$ es un grafo simple que va creciendo en el tiempo, o sea $V_t \in V_{t+1}$, $E_t \subseteq E_{t+1}$, y ϕ_t es una restricción de ϕ_{t+1} en E_t, para cualquier t.

Un grafo etiquetado se define como la tupla $\langle V, E, \phi, l \rangle$ donde $\langle V, E, \phi \rangle$ es un grafo, y $l : V \cup E \to L_V \cup L_E$ una función etiquetadora de vértices y aristas donde L_V y L_E son los conjuntos de etiquetas asignables a vértices y aristas respectivamente.

Se dice que dos grafos etiquetados $G_1 = \langle V_1, E_1, \phi_1, l_1 \rangle$ y $G_2 = \langle V_2, E_2, \phi_2, l_2 \rangle$, son l-isomorfos si los grafos $\langle V_1, E_1, \phi_1 \rangle$ y $\langle V_2, E_2, \phi_2 \rangle$ son isomorfos mediante un par de funciones, $f : V_1 \to V_2$ y $g : E_1 \to E_2$, que cumplen, además, $l_1(v) = l_2(f(v))$ para todo $v \in V_1$ y $l_1(e) = l_2(g(e))$ para todo $e \in E_1$. Si existiera un subgrafo de G_2 que es l-isomorfo a G_1 entonces se dice que existe un l-sub-isomorfismo de G_1 a G_2 o, simplemente, que G_2 contiene a G_1.

Un grafo temporizado se define como la tupla $\langle V, E, \phi, \tau \rangle$, donde $\langle V, E, \phi \rangle$ es un grafo, $\tau : E \to \mathbb{N}$ es conocida como función temporizadora y $\mathbb{N}$ es el conjunto de números naturales representando los instantes de tiempo.

Se dice que dos grafos temporales $G_1 = \langle V_1, E_1, \phi_1, \tau_1 \rangle$ y $G_2 = \langle V_2, E_2, \phi_2, \tau_2 \rangle$, son τ-isomorfos si los grafos $\langle V_1, E_1, \phi_1 \rangle$ y $\langle V_2, E_2, \phi_2 \rangle$ son isomorfos mediante un par de funciones, $f : V_1 \to V_2$ y $g : E_1 \to E_2$, que cumple, además, que existe un $\Delta \in \mathbb{N}$, tal que $\tau_2(g(e)) = \tau_1(e) + \Delta$ para todo $e \in E_1$.

En el caso de la red social que se está describiendo en este epígrafe, se asumirá que existe un grafo etiquetado y temporizado, conocido como universo, $G = \langle V, E, \phi, l, \tau \rangle$, tal que $V_t \subseteq V$ y $E_t \subseteq E$ para cualquier t, l es la función etiquetadora, y la temporizadora τ se define como $\tau(e) = \text{mín}\{j \in \mathbb{N} | e \in E_j\}$.

Se dice que $G_1 = \langle V_1, E_1, \phi_1, \tau_1 \rangle$ es un patron contenido en $G_2 = \langle V_2, E_2, \phi_2, \tau_2 \rangle$ si existe un subgrafo de G_2 que es l-isomorfo y τ-isomorfo G_1.

Si P es un patrón en el universo G de la red social, su soporte y frecuencia, $\sigma(P, G)$, se define según los criterios planteados por Bringmann & Nijssen [98].

Una regla de evolución se define como una implicación $P_1 \rightarrow P_2$ donde $P_2 = \langle V_2, E_2, \phi_2, \tau_2 \rangle$ es un patrón del universo G y P_1 es otro patrón definido de manera única a partir de P_2 como $P_1 = \langle V_1, E_1, \phi_1, \tau_1 \rangle$ tal que: $E_1 = \{e \in E_2 | t_2(e) < \text{máx}_{e' \in E_2} t_2(e')\}$, $V_1 = \bigcup_{e \in E_1} \phi_2(e)$ contiene los vértices que forman las aristas de E_1, y las restantes funciones ϕ_1, l_1, y τ_1 se definen como restricciones de ϕ_2, l_2, y τ_2 al nuevo conjunto de vértices y aristas.

Los patrones P_1 y P_2 se denominan antecedente y consecuente de la regla, respectivamente. El soporte de una regla se define como el soporte del consecuente. La confianza de la regla se define como la razón entre el soporte del antecedente y el soporte de la regla.

Algoritmo 1: germ-rec (G, P, δ, R)

Entrada: G - universo de la red social, P - patron frecuente, δ - umbral de frecuencia mínima.
Salida: R - Conjunto de todos los patrones frecuentes

if *isMin(P)* **then return**;
$R \leftarrow R \cup \{P\}$;
$C \leftarrow$ El conjunto de todos los patrones de G que contienen a s;
foreach $Q \in C$ **do**
  if $\sigma(Q, G) \geq \delta$ **then** germ-rec (G, Q, δ, R);
end

Algoritmo 2: germ (G, P, δ, R)

Entrada: G - universo de la red social, δ - umbral de frecuencia mínima.
Salida: R - Conjunto de todos los patrones frecuentes

$R \leftarrow$ Conjunto de todos los grafos de un solo vértice contenidos en G que son frecuentes según δ;
$R_1 \leftarrow$ Conjunto de todos los patrones de una sola arista contenidos en G que son frecuentes según δ;
foreach $P \in R_1$ **do**
  germ-rec (G, P, δ, R);
end

Hasta donde se conoce, el algoritmo GERM [99] es el único algoritmo reportado que permite obtener todos los patrones frecuentes en un universo G (ver Algoritmos 1 y 2). Puede apreciarse que GERM es una adaptación del conocido gSpan [100] para el caso de las redes sociales. En GERM, se ha modificado el código DFS [100] para tener en cuenta la componente temporal de los grafos y patrones que se procesan.

En casi todas las redes sociales, los enlaces entre entidades son dinámicos; o sea, pueden cambian considerablemente en el tiempo. Por ejemplo, en Facebook, existe un sistema que recomienda posibles futuros vínculos de amistad a sus usuarios. De esta forma, la predicción de enlaces ha emergido como una necesidad. En tal sentido, las GERs han sido utilizadas como base para la predicción de enlaces [99].

4.3. Clasificación

Otra importante técnica de Minería de Datos, aplicada al Análisis de Redes Sociales, es la clasificación de grafos; estructuras con las que se representan las redes sociales. El problema de clasificación de grafos se puede enfocar de dos formas: (1) propagación de etiquetas [101,102] y (2) clasificación de grafos [103].

- **Propagación de etiquetas.** Se tiene un subconjunto de nodos etiquetados en un grafo. La tarea consiste en construir un modelo a partir de estos nodos etiquetados y usarlo para etiquetar (clasificar) nuevos nodos no etiquetados.
- **Clasificación de grafos.** Se tiene un subconjunto de grafos etiquetados dentro de una base de datos de grafos. La tarea consiste en construir un modelo a partir de los grafos etiquetados y usarlo para clasificar los grafos no etiquetados.

El primer caso (propagación de etiquetas) surge en el contexto de grafos masivos como las redes sociales, mientras que el segundo caso (clasificación de subgrafos) surge en contextos como la clasificación de compuestos químicos, compuestos biológicos o datos XML.

Varias son las aplicaciones donde se necesitan los algoritmos de propagación de etiquetas [104,105,106]. Por ejemplo, en el análisis de redes sociales se ha utilizado como una herramienta en la comercialización orientada al seguimiento de minoristas que han recibido promociones. Aquellos clientes que respondan a la promoción (realizando una compra) son etiquetados como nodos positivos en el grafo que representa a la red social, y aquellos que no respondan son etiquetados como negativos. El objetivo es enviar promociones a los clientes que tienen mayor probabilidad de responder a las mismas. Todo se reduce a aprender un modelo a partir de los clientes que han recibido promociones y predecir las respuestas de otros potenciales clientes.

El principal desafío de los algoritmos de propagación de etiquetas radica en encontrar una función de distancia que mida la similaridad entre dos nodos del grafo. La función de distancia más utilizada para este problema consiste en contar el número promedio de pasos para llegar de un nodo a otro utilizando la heurística *random walk* [107]. Sin embargo, esta medida tiene una significante limitación: tiene complejidad temporal $O(n^3)$ y complejidad espacial $O(n^2)$. No obstante, muchos grafos asociados a aplicaciones de la vida real son dispersos, lo que reduce la complejidad del cálculo de la distancia [105,108].

4.3.1. Métodos de clasificación de grafos basados en Kernels

Estos métodos utilizan un *kernel* para estimar la similaridad entre dos grafos etiquetados y se basan en la heurística *random walk*. Para cada grafo se enumeran sus caminos y se obtienen las probabilidades para tales caminos. El *Kernel* compara el conjunto de caminos y sus probabilidades entre dos grafos. Un camino aleatorio (representado como una secuencia de etiquetas de nodos y aristas) se genera vía *random walk*: primero, se selecciona de forma aleatoria un nodo del grafo. Durante el próximo y cada uno de los pasos subsecuentes, se termina el procedimiento o se selecciona aleatoriamente un nodo adyacente para continuar el *random walk*. Las elecciones de los nodos están sujetas a una probabilidad de parada y a una probabilidad de transición de nodo. Por medio de la repetición varias veces de esta heurística, se obtiene una tabla de caminos, cada uno de los cuales se asocia con una probabilidad.

Para medir la similaridad entre dos grafos, se necesita medir la similaridad entre nodos, aristas y caminos:

- ***Kernel* Nodo (Arista).** Un ejemplo es el *Kernel* identidad. Si dos nodos (aristas) tienen la misma etiqueta entonces la función núcleo devuelve 1, en caso contrario devuelve 0. Si las etiquetas de los (las) nodos (aristas) toman valores reales, entonces se puede utilizar una función núcleo gausiana.

- ***Kernel* Camino.** Un camino es una secuencia de etiquetas de nodos y de lados. Si dos caminos tienen la misma longitud, el *Kernel* puede construirse como el producto de los *Kernels* Nodo y Arista. Si dos caminos tienen diferente longitud, el *Kernel* devuelve 0.
- ***Kernel* Grafo.** Como cada camino es asociado con una probabilidad, se puede definir ese *Kernel* como la esperanza del *Kernel* Camino sobre todos los posibles caminos en los dos grafos.

La definición anterior de *Kernel* Grafo es directa. Sin embargo, no es factible computacionalmente enumerar todos los caminos. En particular, en grafos cíclicos, la longitud de los caminos es ilimitada, lo cual hace la enumeración imposible. Por tanto, se necesitan estrategias más eficientes para calcular el *Kernel*. Resulta que la definición de *Kernel* puede ser reformulada para mostrar una estructura anidada. En el caso de grafos dirigidos no cíclicos los nodos pueden ordenarse topológicamente tal que no exista camino del nodo j al nodo i si $i < j$, el *Kernel* puede redefinirse como una función recursiva y utilizando programación dinámica, puede resolverse este problema en $O(|X| \cdot |X'|)$, donde X y X' son los conjuntos de nodos de los dos grafos.

4.3.2. Métodos de clasificación de grafos basados en Boosting

Mientras que la clasificación basada en *Kernels* provee una elegante solución a la clasificación de grafos, esta no revela explícitamente qué características de los grafos (subestructuras) son relevantes para la clasificación. Para resolver lo anterior, se introduce una nueva estrategia de clasificación de grafos basada en minado de patrones. La idea es clasificar con base en subestructuras importantes de los grafos.

Se puede crear un vector binario de características basado en la presencia o ausencia de cierta subestructura (subgrafo) y aplicar un clasificador tradicional. Dado que, por lo general, el total de subgrafos es con frecuencia muy grande, se debe centrar en un pequeño grupo de características que sean relevantes. La estrategia más directa para poder encontrar características interesantes es el minado de patrones frecuentes. Sin embargo, los patrones frecuentes no son necesariamente patrones relevantes. Por ejemplo, en grafos químicos patrones obicuos tales como $C-C$ o $C-C-C$ son frecuentes pero casi no tienen significancia en la predicción de características importantes de los compuestos químicos tales como la actividad, la toxicidad, etc. El *Boosting* se usa para seleccionar automaticámente un conjunto relevante de subgrafos como rasgos para la clasificación. El método LPBoost (*Linear Program Boost*) aprende una función discriminante linear para la selección de rasgos. Para obtener una regla interpretable se necestia obtener un vector disperso (*sparse*) de pesos, en el cual sólo pocos pesos sean diferentes de cero. Se muestra en Saigo *et al.* [109] que la clasificación de grafos basada en *Boosting* puede alcanzar mejor eficacia que la basada en *Kernels*.

5. Conclusiones y recomendaciones

El análisis de redes sociales es un área que presenta muchas oportunidades de investigación tanto en las ciencias sociales como en las ciencias computacionales. Sus aplicaciones se observan en muy diversas áreas como las redes de interconexiones sociales (de amistad, laborales, etc.), pasando por las redes criminales y las redes de opiniones, hasta las redes globales como Facebook y Twitter.

Las problemáticas o tareas que se aplican sobre tales redes son muy diversas, requiriendo métricas y estrategias que faciliten y disminuyan las complejidades computacionales de las técnicas asociadas. Muchas de estas métricas ya han sido ampliamente analizadas y aplicadas. No obstante, nuevas métricas y conceptos siguen proponiéndose para problemas específicos y mejoras de técnicas, como la intermediación dividida (*split betweenness*) o la métrica *R*, entre otras.

Entre las técnicas de minería de datos para el análisis de redes sociales se resaltan las asociadas con la detección de comunidades y otras afines, como son el análisis de roles, análisis de difusión de información, análisis de centralidad, en donde las técnicas de agrupamiento presentan un rol central.

Otras de las técnicas de minería de datos que se observan como relevantes son las asociadas con la minería de grafos, particularmente en tarea tales como la identificación de la dinámica del comportamiento, la determinación de perfiles, la predicción de enlaces, entre otras.

En general, a partir del trabajo realizado sobre el análisis de redes sociales, se recomienda continuar los estudios asociados con las técnicas de agrupamiento y de minería de grafos, tales como el agrupamiento conceptual, difuso y evolutivo, la minería de reglas de evolución de grafos y de subgrafos periódicamente recurrentes, para enfrentar aplicaciones en las que se requieran modelar y analizar diferentes aspectos de las comunidades, la explicación de los comportamientos de las redes sociales y, particularmente, de sus dinámicas.

Referencias bibliográficas

1. Françoisse, K., Fouss, F., Saerens, M.: A link-analysis-based discriminant analysis for exploring partially labeled graphs. Pattern Recognition Letters **34**(2) (2013) 146–154
2. Wang, J.H., Lin, C.L.: An association model for implicit crime link analysis. In: Proceedings of the 2010 Pacific Asia conference on Intelligence and Security Informatics. PAISI'10, Berlin, Heidelberg, Springer-Verlag (2010) 15–21
3. Yu, P.S., Han, J., Faloutsos, C.: Link Mining: Models, Algorithms, and Applications. 1st edn. Springer Publishing Company, Incorporated (2010)
4. Ozgul, F., Erdem, Z., Bowerman, C., Bondy, J.: Combined detection model for criminal network detection. In: Proceedings of the 2010 Pacific Asia conference on Intelligence and Security Informatics. PAISI'10, Berlin, Heidelberg, Springer-Verlag (2010) 1–14
5. Qin, J., Xu, J.J., Hu, D., Sageman, M., Chen, H.: Analyzing terrorist networks: A case study of the global salafi jihad network. In Kantor, P., Muresan, G., Roberts, F., Zeng, D., Wang, F.Y., Chen, H., Merkle, R., eds.: Intelligence and Security Informatics. Volume 3495 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2005) 287–304
6. Shaikh, M., Wang, J., Yang, Z., Song, Y.: Graph structural mining in terrorist networks. In Alhajj, R., Gao, H., Li, X., Li, J., Zaïane, O., eds.: Advanced Data Mining and Applications. Volume 4632 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2007) 570–577
7. Canter, D.V.: A partial order scalogram analysis of criminal network structures. Behaviormetrika **31**(2) (July 2004) 131–152
8. Wellman, B., Berkowitz, S.: Social Structures: A Network Approach. Structural Analysis in the Social Sciences, No 2. University Press (1988)
9. Zhang, M.: Social network analysis: History, concepts, and research. In Furht, B., ed.: Handbook of Social Network Technologies and Applications. Springer US (2010) 3–21
10. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge University Press (1994)
11. Borgatti, S.P., Mehra, A., Brass, D.J., Labianca, G.: Network analysis in the social sciences. Science **323**(5916) (2009) 892–895
12. Marin, A., Wellman, B. In: Social Network Analysis: An Introduction, Sage (2010)
13. Kimura, M., Saito, K., Nakano, R., Motoda, H.: Extracting influential nodes on a social network for information diffusion. Data Mining and Knowledge Discovery **20** (2010) 70–97
14. Sun, Y., Han, J., Aggarwal, C.C., Chawla, N.V.: When will it happen?: relationship prediction in heterogeneous information networks. In: Proceedings of the fifth ACM international conference on Web search and data mining. WSDM '12, New York, NY, USA, ACM (2012) 663–672
15. Tang, L., Liu, H.: Graph mining applications to social network analysis. In Aggarwal, C.C., Wang, H., eds.: Managing and Mining Graph Data. Volume 40 of Advances in Database Systems. Springer US (2010) 487–513
16. Nguyen, N., Xuan, Y., Thai, M.: On detection of community structure in dynamic social networks. In Thai, M.T., Pardalos, P.M., eds.: Handbook of Optimization in Complex Networks. Springer Optimization and Its Applications. Springer New York (2012) 307–347
17. Giatsoglou, M., Vakali, A.: Capturing social data evolution using graph clustering. Internet Computing, IEEE **17**(1) (jan.-feb. 2013) 74 –79

18. Rossi, R.A., Neville, J., Gallagher, B., Henderson, K.: Modeling dynamic behavior in large evolving graphs. In: Proceedings of the Sixt ACM Internationa Conference on Web Search and Data Mining, ACM (2013)
19. Takaffoli, M., Sangi, F., Fagnan, J., R.Zaïane, O.: Community evolution mining in dynamic social networks. Procedia - Social and Behavioral Sciences **22**(0) (2011) 49 – 58
20. Khan, A., Li, N., Yan, X., Guan, Z., Chakraborty, S., Tao, S.: Neighborhood based fast graph search in large networks. In: Proceedings of the 2011 ACM SIGMOD International Conference on Management of data. SIGMOD '11, New York, NY, USA, ACM (2011) 901–912
21. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM) **46**(5) (1999) 604–632
22. Balasundaram, B., Butenko, S., Hicks, I.V.: Clique relaxations in social network analysis: The maximum k-plex problem. Operations Research **59**(1) (January/February 2011) 133–142
23. Pattillo, J., Youssef, N., Butenko, S.: Clique relaxation models in social network analysis. In Thai, M.T., Pardalos, P.M., eds.: Handbook of Optimization in Complex Networks. Springer Optimization and Its Applications. Springer New York (2012) 143–162
24. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. In: Proceedings of The National Academy of Sciences of the United States of America. Volume 99. (2002) 7821–7826
25. Fortunato, S.: Community detection in graphs. Physics Reports **486**(3–5) (2010) 75 – 174
26. Zardi, H., Romdhane, L.B.: An o(n2) algorithm for detecting communities of unbalanced sizes in large scale social networks. Knowledge-Based Systems **37**(0) (2013) 19 – 36
27. Chen, J., Zaïane, O., Sander, J., Goebel, R.: Ondocs: Ordering nodes to detect overlapping community structure. In Memon, N., Xu, J.J., Hicks, D.L., Chen, H., eds.: Data Mining for Social Network Data. Volume 12 of Annals of Information Systems. Springer US (2010) 125–148
28. Oliveira, M., Gama, J.: An overview of social network analysis. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **2**(2) (2012) 99–115
29. Gregory, S.: An algorithm to find overlapping community structure in networks. In: Proceedings of the PKDD 2007. (2007) 91–102
30. Meo, P.D., Ferrara, E., Fiumara, G., Provetti, A.: Enhancing community detection using a network weighting strategy. Information Sciences **222**(0) (2013) 648 – 668
31. Alahakoon, T., Tripathi, R., Kourtellis, N., Simha, R., Iamnitchi, A.: K-path centrality: a new centrality measure in social networks. In: Proceedings of the 4th Workshop on Social Network Systems. SNS '11, New York, NY, USA, ACM (2011) 1:1–1:6
32. Newman, M., Girvan, M.: Finding and evaluating community structure in networks. Physical Reviews E **69**(2) (2004)
33. Clauset, A., Newman, M.E.J., Moore, C.: Finding community structure in very large networks. Physical Reviews E (2004)
34. Nanda, S., Kotz, D.: Localized bridging centrality. In Thai, M.T., Pardalos, P.M., eds.: Handbook of Optimization in Complex Networks. Springer Optimization and Its Applications. Springer New York (2012) 197–218
35. Aho, A., Hopcroft, J., Ullman, J.: Data Structures and Algorithms. Addison-Wesley Publishing Company (1983)
36. Mejia-Olivares, C.: Análisis de redes sociales a gran escala. Master's thesis, Centro de Investigación y de Estudios avanzados, Departamento de Computacion, Instituto Politécnico Nacional (IPN) (2010)
37. Zanghi, H., Ambroise, C., Miele, V.: Online and offline social networks: Use of social networking sites by emerging adults. Applied Developmental Psychology **29** (2008) 420–433
38. Newman, M.E.J., Watts, D.J., Strogatz, S.H.: Random graph models of social networks. In: Proceedings of the National Academy of Sciences of the United States of America. (2002) 2566–2572
39. Milgram, S.: The small world problem. Psychology Today **2**(1) (1967) 60–67
40. Watts, D.J., Strogatz, S.H.: Collective dynamics of small-world networks. Nature **393**(6684) (1998) 440–442
41. Barabási, A.L., Ravasz, E., Vicsek, T.: Deterministic scale-free networks. Physica A **299**(3-4) (2001) 559–564
42. Cohen, R., Havlin, S., ben Avraham, D.: Structural properties of scale-free networks. In: Handbook of Graphs and Networks: From the Genome to the Internet. John Wiley & Sons (2003) 85–110
43. Newman, M.E.J.: Power laws, pareto distributions and zipf's law. Contemporary Physics **46**(5) (2005) 323–351
44. Albert, R.: Scale-free networks in cell biology. J Cell Sci **118**(21) (2005) 4947–4957
45. Holger, E., Lutz-Ingo, M., Stefan, B.: Scale-free topology of e-mail networks. Phys. Rev. E **66**(3) (2002)
46. Zhang, H.: The scale-free nature of semantic web ontology. In: Proceedings of the 17th international conference on World Wide Web. (2008) 1047–1048
47. Analyst's Notebook: http://www.i2group.com/us/products/analysis-product-line/ibm-i2-analysts-notebook (2012)
48. FMS: Sentinel Visualizer: http://www.fmsasg.com/ (2012)
49. Deltica NetReveal: http://www.deticanetreveal.com/en/ (2012)
50. InfiniteInsight: http://www.kxen.com/products/social+network+analysis (2012)
51. Klerks, P.: The network paradigm applied to criminal organizations: Theoretical nitpicking or a relevant doctrine for investigators? recent developments in the netherlands. Connections **24**(3) (2001) 53–65

52. Gilbert, E.N.: Random graphs. The Annals of Statistics **30**(30) (1959) 1141–1144
53. Erdos, P., Renyi, A.: On random graphs. Publicationes Mathematicae (Debrecen) **6** (1959) 290–297
54. Erdos, P., Renyi, A.: On the evolution of random graphs. Publ. Math. Inst. Hung. Acad. Sci **5** (1960) 17–61
55. Erdos, P., Renyi, A.: On the strength of connectedness of a random graph. Acta Mathematica Hungarica **12** (1961) 261–267
56. Ferrer, R., Cancho, I., Solé, R.V.: Physical review letters. In: Proceedings of The Royal Society of London. Series B, Biological Sciences. (2001) 2261–2265
57. Kuperman, M., Abramson, G.: Small world effect in an epidemiological model. Physical Review Letters **86**(13) (2001) 2909–2912
58. Barabasi, A.L., Albert, R.: Emergence of scaling in random networks. Science **286**(5439) (1999) 509–512
59. Zachary, W.W.: An information flow model for conflict and fission in small groups. Journal of Anthropological Research **33** (1977) 452–473
60. Lusseau, D.: Evidence for social role in a dolphin social network. Evolutionary Ecology **21**(3) (2007) 357–366
61. Ahn, Y., Han, S., Kwak, H., Moon, S., Jeong, H.: Analysis of topological characteristics of huge online social networking services. In: Proceedings of the 16th international conference on World Wide Web. (2007) 835–844
62. Leskovec, J., Adamic, L.A., Huberman, B.A.: The dynamics of viral marketing. ACM Transactions on the Web (TWEB) **1**(1) (2007)
63. Adar, E.: Guess: a language and interface for graph exploration. (2006) 791–800
64. Jia, Y., Hoberock, J., Garland, M., Hart, J.: On the visualization of social and other scale-free networks. In IEEE Transactions on Visualization and Computer Graphics **14**(6) (2008) 1285–1292
65. Kang, H., Getoor, L., Singh, L.: Visual analysis of dynamic group membership in temporal social networks. In SIGKDD Explor. Newsl. **9**(2) (2007) 13–21
66. Opsahl, T., Panzarasa, P.: Clustering in weighted networks. In Social Networks **31**(2) (2009) 155–163
67. Huisman, M., van Duijn, M.A.J.: Software for social network analysis. (2005) 270–316
68. Berger-Wolf, T.Y., Saia, J.: A framework for analysis of dynamic social networks. (2006) 523–528
69. Demoll, B.S., Mcfarland, D.: The art and science of dynamic network visualization. In journal of Social Structure **7** (2005)
70. Matsuo, Y., Mori, J., Hamasaki, M., Ishida, K., Nishimura, T., Takeda, H., Hasida, K., Ishizuka, M.: Polyphonet: an advanced social network extraction system from the web. (2006) 397–406
71. Mika, P.: Flink: Semantic web technology for the extraction and analysis of social networks. Web Semantics: Science, Services and Agents on the World Wide Web **3**(2-3) (2005) 211–223
72. Mika, P.: Social networks and the semantic web (semantic web and beyond). (2007)
73. Liddle, S.W., Ho, S., Embley, D.W.: On the automatic extraction of data from the hidden web. (2007) 212–226
74. Madhavan, J., Ko, D., Kot, L., Ganapathy, V., Rasmussen, A., Halevy, A.: Google's deep web crawl. 1241–1252
75. Baden, R., Bender, A., Spring, N., Bhattacharjee, B., Starin, D.: Persona: an online social network with user-defined privacy. In SIGCOMM Comput. Commun **39**(4) (2009) 135–146
76. Athanasopoulos, E., Makridakis, A., Antonatos, S., Antoniades, D., Ioannidis, S., Anagnostakis, K.G., Markatos, E.P.: Antisocial networks: Turning a social network into a botnet. (2008) 146–160
77. Kossinets, G.: Effects of missing data in social networks. In Social Networks **28**(3) (2006) 247–268
78. Lee, S., Kim, P., Jeong, H.: Statistical properties of sampled networks. Phys. Rev. **73**(1) (2006)
79. Baumes, J., Goldberg, M., Krishnamoorty, M., Magdon-Ismail, M., Preston, N.: Finding communities by clustering a graph into overlapping subgraphs. In: Proceedings of IADIS Applied Computing. (2005) 97–104
80. Baumes, J., Goldberg, M., Magdon-Ismail, M.: Efficient identification of overlapping communities. In: Proceedings of ISI 2005. (2005) 27–36
81. Davis, G., Carley, K.: Clearing the fog: Fuzzy, overlapping groups for social networks. Social Networks **30**(3) (2008) 201–212
82. Goldberg, M., Kelley, S., Magdon-Ismail, M., Mertsalov, K., Wallace, A.: Finding overlapping communities in social networks. In: Proceedings of SocialCom2010. (2010) 104–113
83. Gregory, S.: A fast algorithm to find overlapping communities in networks. In: Proceedings of the 12th ECML KDD. (2008) 408–423
84. Macropol, K., Can, T., Singh, A.: RRW: repeated random walks on genome-scale protein networks forlocal cluster discovery. BMC Bioinformatics **10**(283) (2009)
85. Magdon-Ismail, M., Purnell, J.: SSDE-Cluster:fast overlapping clustering of networks using sampled spectral distance embedding and gmms. In: Proceedings of SocialCom2011. (2011) 756–759
86. Palla, G., Derényi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature **435**(7043) (2005) 814–824
87. Zhang, S., Wang, R., Zhang, X.: Identification of overlapping community structure in complex networks using fuzzy c-means clustering. Physica A: Statistical Mechanics and its Applications **374**(1) (2007) 483–490
88. Everett, M., Borgatti, S.: Analyzing clique overlap. Connections **21** (1998) 49–61

89. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Stanford Digital Libraries Working Paper (1998)
90. Zhao, Y., Karypis, G.: Criterion functions for document clustering: Experiments and analysis. Technical Report 01-40, Department of Computer Science, University of Minnesota, Minneapolis, MN (2001)
91. Civril, A., Magdon-Ismail, M., Bocek-Rivele, E.: Ssde: Fast graph drawing using sampled spectral distance embedding. In: Graph Drawing. (2007) 30–41
92. Berlingerio, M., Bonchi, F., Bringmann, B., Gionis, A.: Mining graph evolution rules. In: Proceedings of ECML-PKDD. Volume LNCS 5781., Springer-Verlag (2009) 115–130
93. Apostólico, A., Erdös, P.L., Györi, E., Lipták, Z., Pizzi, C.: Efficient algorithms for the periodic subgraphs mining problem. Journal of Discrete Algorithms (2012)
94. Leskovec, J., Kleinberg, J., Faloutsos, C.: Graphs over time: densification laws, shrinking diameters and possible explanations. In: Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, New York, USA, ACM Press (2005) 177–187
95. McGlohon, M., Akoglu, L., Faloutsos, C.: Statistical Properties of Social Nerworks. In: Social Network Data Analytics. Springer (2011) 17–42
96. Apostólico, A., Barbares, M., Pizzi, C.: Speedup for a periodic subgraph miner. Information Processing Letters **111** (2011) 521–523
97. Lahiri, M., Berger-Wolf, T.: Periodic subgraph mining in dynamic nerworks. Knowledge and Information Systems **24** (2010) 467–497
98. Bringmann, B., Nijssen, S.: What is frequent in a single graph? In: Proceedings of 12th PAKDD, ACM Press (2008) 858–863
99. Bringmann, B., Berlingerio, M., Bonchi, F., Gionis, A.: Learning and predicting the evolution of social networks. IEEE Intelligent Systems **25** (2010) 26–35
100. Yan, X., Huan, J.: gspan: Graph-based substructure pattern mining. In: Proceedings International Conference on Data Mining, Maebashi, Japan (2002) 721–724
101. Kashima, H., Tsuda, K., Inokuchi, A.: Marginalized kernels between labeled graphs. In: Proceedings of the ICML. (2003)
102. Kudo, T., Maeda, E., Matsumoto, Y.: An application of boosting to graph classification. In: NIPS Conf. (2004)
103. Zaki, M.J., Aggarwal, C.C.: Xrules: An effective structural classifier for xml data. In: KDD Conf. (2003)
104. Taskar, B., Abbeel, P., Koller, D.: Discriminative probabilistic models for relational data. In: Proceedings of the UAI. (2002) 485–492
105. Zhou, D., Bousquet, O., J.Weston, Sch-olkopf, B.: Learning with local and global consistency. Advances in Neural Information Processing Systems **16** (2004) 321–328
106. Zhou, D., Huang, J., Sch-olkopf, B.: Learning from labeled and unlabeled data on a directed graph. In: ICML Conf. (2005) 1036–1043
107. Kondor, R., Lafferty, J.: Diffusion kernels on graphs and other discrete input spaces. In: ICML Conf. (2002) 315–322
108. Zhu, X., Ghahramani, Z., Lafferty, J.: Semi-supervised learning using gaussian fields and harmonic functions. In: ICML Conf. (2003) 912–919
109. Saigo, H., Nowozin, S., Kadowaki, T., Kudo, T., Tsuda, K.: Gboost: A mathematical programming approach to graph classification and regression. Machine Learning 321–328

RT_020, octubre 2013
Aprobado por el Consejo Científico CENATAV

Editor: Lic. Lucía González Bayona
Diseño de Portada: Di. Alejandro Pérez Abraham
RNPS No. 2143
ISSN 2072-6260
Indicaciones para los Autores:
Seguir la plantilla que aparece en www.cenatav.co.cu
C E N A T A V
7ma. A No. 21406 e/214 y 216, Rpto. Siboney, Playa;
La Habana. Cuba. C.P. 12200
Impreso en Cuba