CENTRO DE INVESTIGACIÓN Y DE ESTUDIOS AVANZADOS
DEL INSTITUTO POLITÉCNICO NACIONAL
DEPARTAMENTO DE COMPUTACIÓN

Análisis de Redes Sociales a Gran Escala

Tesis que presenta

Cristian Paolo Mejia Olivares

Para obtener el grado de

Maestro en Ciencias en Computación

Directores de Tesis:

Dra. Xiaoou Li Zhang

Dr. Luis E. Rocha Mier

México, D. F. Febrero, 2010

Agradecimentos

A mi padre Ramón Mejia y a mi madre Estela Olivares por brindarme su apoyo siempre y estar en los momentos felices y difíciles de mi vida, por ayudarme a cumplir mis sueños y nunca darme la espalda por más complicado que esto pareciera, porque no existe manera alguna en esta vida en que pueda pagarles todo lo que han hecho por mi.

A mis hermanos, mis sobrinos y mis cuñadas por apoyarme y por brindarme buenos consejos y lecciones muy importantes en mi vida, por haberme dado momentos de alegria y de distracción y sobre todo por ser parte de mi familia.

A Edith Tamaya por ser mi compañera en gran parte de mi vida y por compartir tantos momentos a mi lado, por enseñarme a valorar las cosas, por apoyarme en todo momento y por brindarme un espacio dentro de tu corazón y de tu familia, porque las cosas nunca hubieran sido las mismas sin ti :).

A la Dra. Xiaoou Li y al Dr. Luis Rocha por dedicarme tiempo y orientarme durante mi trabajo de tesis, y por siempre escucharme, brindarme su apoyo y darme consejos cuando yo me acercaba a ustedes.

A mis sinodales, Dr. Pedro Mejia Alvarez y Dr. Amilcar Meneses Viveros por el tiempo dedicado para la revisión de este documento y por sus comentarios para mejorarlo.

Gracias a mis amigos del CINVES, Pam, july, beto, jonhy, don gabo, andres, migue, ray, pau, adri, lupita, lil, christian, madai, paco, bris por dejarme muchos recuerdos felices dentro y fuera del CINVES.

A Sofi por siempre apoyarnos bajo cualquier situación y escucharnos cuando necesitabamos de alguien que nos echara la mano.

A Alan y a Fernanda por brindarme su amistad cuando más la necesitaba.

A la Universidad Rey Juan Carlos de España por permitirme realizar una estancia, y sobre todo a la gente de *LibreSoft* por brindarme su apoyo y ser tan amables durante la estancia.

Al CONACyT y al CINVESTAV por el apoyo económico para la realización de mis estudios de maestría y ofrecerme la oportunidad de conocer muchas cosas nuevas.

Resumen

La mayoría de los estudios sobre la estructura de las redes sociales está basado en redes a pequeña escala [56, 61]. Recientemente, las aplicaciones de los sistemas Web proporcionan una nueva fuente de información para poder estudiar las propiedades de las redes sociales del mundo real. Sistemas como Flickr, Twitter, MySpace y Facebook, han permitido retomar la teoría de redes sociales para poder crear nuevas aplicaciones en base a este tipo de sistemas, ya que permiten modelar redes del mundo real con la gran cantidad de información que poseen.

En la literatura, se pueden encontrar diferentes trabajos sobre redes sociales a pequeña escala y para redes estáticas. Sin embargo, pocos son los trabajos que estudian las propiedades de las redes sociales a gran escala. El análisis de redes sociales (ARS) proporciona los conceptos y las técnicas para el estudio de redes sociales basado en la teoría de grafos y del cálculo matemático.

Nuestra investigación está dividida en tres partes principales, la primera de ellas es un estudio de los diferentes métodos de extracción de información y de las técnicas de muestreo aplicadas a los sistemas de redes sociales en línea. Utilizando un *muestreo de bola de nieve* y haciendo uso de la interacción entre usuarios, se implementó un algoritmo para obtener un conjunto de datos representativo de los sistemas Flickr y Wikipedia.

En la segunda parte de la tesis se estudió la forma de escalar el análisis de redes sociales a gran escala mediante la *detección de comunidades*, la cual se apoya en las diferentes técnias de la *teoría del agrupamiento en grafos*. En base al método de *propagación de etiquetas* [77] para la detección de comunidades disjuntas, se adaptó un algoritmo que permite *detectar comunidades traslapadas* en *redes sociales a gran escala* a partir del cálculo del *coeficiente de agrupamiento* de un nodo en cada comunidad.

Para medir el desempeño del algoritmo se realizaron pruebas con diferentes conjuntos de datos con distintos tamaños y características. En general, los resultados experimentales demuestran que el algoritmo presenta un comportamiento estable y un buen desempeño para conjuntos de datos grandes y permite escalar el estudio de las redes sociales a gran escala por medio del traslapamiento de comunidades dentro de las redes sociales.

Finalmente, presentamos un análisis de la estructura de dos diferentes tipos de redes sociales, Flickr basada en contenido y la Wikipedia basada la colaboración. En general, los resultados muestran que las redes sociales de estos sistemas presentan un modelo de crecimiento como una red libre de escala y poseen una forma del tipo mundo pequeño.

Abstract

Most studies on the structure of social networks is based on networks small scale [56, 61]. Recently, applications of Web systems provide a new source of information to study the properties of social networks in real world. Systems such as Flickr, MySpace and Facebook, allow new applications using social networks theory, in which real-world networks can be modeled with the large amount of information they have.

In literature, much work can be found on studying small-scale social networks and static networks. However, few of them analyze properties of large scale social networks. Social network analysis (SNA) concepts and techniques are generally based on graph theory and mathematical calculation.

Our research is divided into three main parts; the first is a study of different methods of extracting information and sampling techniques applied to Online Social Networks. Using a snowball sampling and using the interaction between users, an algorithm was implemented to obtain a representative data set of systems Flickr and Wikipedia (Between August 2008 to December 2009) were investigated.

In the second part of the thesis we explore the algorithms for detecting social networking communities, and propose a community detection algorithm based on disjoint communities detection using label propagation method of reference . To measure the performance of our algorithm, data sets with different sizes and characteristics were tested. In general, the experimental results show that our algorithm has a stable behavior and a good performance for large-scale networks. Based on disjoint communities detection using label propagation method [77], we adapted an algorithm to detect overlapping communities in large-scale social networks by calculating the clustering coefficient of a node in each community.

For measure the performance of the algorithm this was tested with different data sets and characteristics. In general, experimental results show that the algorithm has a stable behavior and good performance for large data sets and allows you to scale the study of large-scale social networks through the overlap of communities within social networks.

Finally, we did two case studies to demonstrate our approaches described above. Flickr and Wikipedia were selected for their huge scale and popularity as social networks. The results demonstrate that these social networks present a model of growth as a scale-free network and a shape as small world.

Contenido

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

El surguimento de nuevas tecnologías basadas en Web, han permitido que la comunicación e interacción entre usuarios sea cada vez más cercana. Es por eso que hoy en día podemos encontrar diferentes servicios Web que permiten a los usuarios interactuar con personas en diferentes partes del mundo. Este tipo de tecnología ha venido a cambiar la forma en que interactúan los usuarios, ya sea mediante la compartición de archivos, el envío de mensajes, publicaciones en blogs, etc.

La Web 2.0[1] hace referencia a una segunda generación de Web basada en comunidades de usuarios y un conjunto de utilidades como las redes sociales, los blogs, o los wikis, donde la colaboración y el intercambio ágil de información juegan un papel importante[6].

El estudio de las redes sociales a gran escala representa un gran reto para los investigadores del Análisis de Redes Sociales (ARS), ya que estas redes poseen propiedades que no pueden ser obtenidas por simple escalamiento de las redes pequeñas estudiadas por los sociólogos[9].

La tendencia de la gente para formar grupos y crear nuevas relaciones es fundamental en la estructura de la sociedad y la forma en que tales grupos se van desarrollando a través del tiempo y como este crecimiento afecta directamente el comportamiento de la estructura de la sociedad.

1.1 Antecedentes

La teoría de redes tiene sus orígenes e influencias en diferentes corrientes del pensamiento como la antropología, psicología, sociología y las matemáticas. La sociometría se interesó por la estructura de los grupos de amigos aunque sus principales razones fueran terapéuticas [61]. En los últimos años diversas áreas del conocimiento han desarrollado estudios enfocados a la estructura y comportamiento de estos sistemas, y principalmente en la relación

[1]Termino acuñido por Tim O'Reilly en una conferencia en Brainstorming en el año de 2004.

entre usuarios utilizando el ARS como un conjunto de técnicas para el estudio formal de las relaciones entre usuarios y las estructuras sociales que se presenten[84].

Recientemente los servicios de redes sociales en línea[2] han adquirido gran popularidad y se encuentran dentro los sitios más importantes en la Web en cuestión de generar tráfico en la red. Por ejemplo, sitios como MySpace, Facebook, Orkut, Youtube y Flickr poseen una gran cantidad de usuarios que van desde los millones hasta los cientos de millones como en el caso de Facebook y Youtube que en los últimos años han tenido un alto crecimiento y a medida que pasa el tiempo incorporan nuevas tecnologías en base a la aceptación de los usuarios.

Sin embargo, las redes sociales van más allá de solo generar nuevos sistemas y aplicaciones; Distintos grupos de investigadores han visto a las redes sociales en línea como una oportunidad de poder obtener información y poder plantear nuevas problematicas a diversas áreas del conocimiento. Trabajos recientes han propuesto el uso de redes sociales para mitigar el correo SPAM[12], mejorar la eficiencia de los motores de búsqueda[58], para encontrar organizaciones de terroristas y traficantes[88], entre otros.

1.2 Trabajo Relacionado

Existen trabajos que se enfocan en entender la estructura y evolución de las redes sociales a larga escala [60, 4, 42], estos se enfocan en estudiar la estructura de la red y como se comportan entre ellas. Las diferencias que existen entre las distintas redes sociales es notable, podemos ver que existen diferentes tipos de servicios de redes sociales circulando en la Web, donde el surguimiento de nuevas tecnologías propician la generación de nuevos sistemas. Sin embargo, con el surguimiento de estas nuevas tecnologías y la aparición de los servicios de redes sociales en línea, los investigadores se han visto motivados para plantear nuevos problemas sobre su comportamiento, y han retomado trabajos clásicos de la teoría de grafos (p.ej., estudios sobre redes mundo pequeño, redes libres de escala y las estructura de gráfos[84]) para entender la estructura de dichos sistemas.

Trabajos recientes se enfocan en presentar metodologías para poder extraer información de los sistemas en línea[53, 54] para posteriormente obtener una red que represente lo más posible a los individuos y su forma de interactuar entre ellos. Muchos son los problemas que surguen cuando se requiere obtener un conjunto de datos que cumpla con las propiedades de una red social, estos pueden ir desde las restricciones mismas del sistema para proporcionar información hasta el tiempo necesario para procesar la información. Uno de los principales problemas para el muestreo es la arquitectura misma de los servicios de redes sociales en línea, ya que cada sistema posee su propia arquitectura y resulta imposible definir mecanismos que generalicen el proceso de extracción de todos estos sistemas.

[2]Para más detalle sobre redes sociales en línea ver el capítulo 2

Una propiedad interesante de muchas redes es su tendencia a formar grupos. Los sociólogos definen una comunidad en una red social como un conjunto de individuos que comparten en común ciertas características. El problema de la detección de comunidades hace uso de los conceptos del agrupamiento en grafos para explicar de manera matemática sus etudios. Muchos son los trabajos sobre el agrupamiento en grafos [57, 65, 71] que permiten estudiar el problema de la detección de comunidades basados en los conceptos de la teoría de grafos. Trabajos recientes [89, 80, 31, 35] estudian la detección de comunidades utilizando los diferentes tipos de agrupamiento en grafos, proporcionando una poderosa herramienta para medir la estructura de las comunidades dentro de las redes sociales. Pocos son los trabajos de investigación que proporcionan metodologías para la detección de redes sociales a gran escala [46, 47] y sobre todo para comunidades dinámicas [74], es decir, comunidades que se ven afectadas a través del tiempo.

1.3 Motivación

Los sitios de las redes sociales en línea como: Flickr, Facebook, Twitter, Youtube, y MySpace, son utilizados para organizar, almacenar, y compartir contenido entre usuarios. En muchos de estos sitios, el contenido de la información entre usuarios son públicos y pueden ser extraídos automaticamente para generar una red social a gran escala. Lo más importante de este tipo de sistemas es que permiten modelar una red del mundo real, ya que este tipo de información permite analizar las interacciones entre distintos usuarios y permite comprender como se comportará a lo largo del tiempo.

La necesidad de compartir y buscar información entre usuarios se ha convertido en una de las principales características de los sistemas Web actuales, ya que facilitan la forma de compartir información. Sin embargo, el comportamiento de estos sistemas sigue un comportamiento dinámico, el cual es difícil de predecir por simple observación de sus propiedades. Es por eso que la necesidad de realizar un estudio de las diferencias y similitudes de los distintos sistemas de redes sociales en línea es fundamental para el desarrollo de nuevas aplicaciones. Esto también permite entender la forma en como los individuos se relacionan entre sí y generan comunidades dentro de la red.

1.4 Objetivos

Los objetivos de la presente tesis son:

- Basado en técnicas de muestreo existentes (bola de nieve, de nodo y de enlace), extraer representaciones de las redes sociales de diferentes sistemas en línea dentro de la Web, tal como Flickr y Wikipedia.

- Diseñar un algoritmo que permita optimizar el tamaño de los grafos representativos de una red social midiendo la interacción entre individuos utilizando un nivel de amistad, generando redes más representativas y altamente conectadas.
- Análizar la estructura y dinamismo de las redes sociales Flickr(basada en contenido) y Wikipedia(basada en la colaboración) mediante el uso de las técnicas del análisis de redes sociales.
- Basado en métodos sobre la detección de comunidades (p.ej., uso de cliques, caminos aleatorios, intermediación de aristas, funciones de modularidad, etc), desarrollar un algoritmo para la detección de comunidades en redes sociales a gran escala.

1.5 Organización

El resto de la tesis está organizado de la siguiente manera:

En el capítulo 2 se define lo que son las redes sociales, el análisis de redes sociales y se propone una clasificación para las redes sociales en base a diversos factores y se proporciona una lista de los sistemas más importantes dentro de la Web. Se describen algunas de las propiedades más importantes de una red social, así como los modelos clásicos de redes sociales. Finalmente se presentan los diferentes tipos de sistemas para analizar y visualizar redes sociales.

En el capítulo 3 se describe la metodología empleada para el muestreo de datos y la forma en que esta información es adquirida a partir de los sistemas de redes sociales en línea. Se presenta además un algoritmo para obtener redes más representativas mediante la relación de amistad entre usuarios para el caso de redes sociales basadas en contenido.

En el capítulo 4 se proporciona una introducción sobre los conceptos de detección de comunidades y como esta se apoya en el agrupamiento en grafos como la formalidad para el estudio de la detección de comunidades en las redes sociales. Después se estudian los diferentes algoritmos para detección de comunidades, y se presenta un algoritmo para escalar el análisis de redes sociales a gran escala.

En el capítulo 5 se muestran un análisis sobre los conjuntos de datos extraídos de dos redes sociales en línea como son Flickr y Wikipedia. También se muestran los resultados obtenidos sobre el algoritmo de detección de comunidades para redes a gran escala propuesto en esta tesis.

En el capítulo 6 se dan las conclusiones de este trabajo, resaltando la importancia de las redes sociales en diferentes áreas del conocimiento, y como estas estructuras están cambiando la forma social de las personas.

Capítulo 2

Redes Sociales

Las redes sociales se definen como un conjunto finito de actores (individuos, grupos, organizaciones, comunidades, sociedades, etc) vinculados unos a otros a través de una relación o un conjunto de relaciones sociales. Las redes sociales se apoyan en el Análisis de Redes Sociales (ARS) el cual se centra en tomar las relaciones entre actores como el *material* sobre el cual se construye y se organiza el comportamiento social de actores. El punto de análisis deja de ser el individuo (egocéntrica) y pasan a serlo las relaciones[84]; proporcionando un conjunto de métodos y técnicas para el estudio formal de las relaciones entre actores.

En las siguientes secciones se presenta una descripción detallada del concepto de *redes sociales*, las propiedades con las que cuenta, los tipos de modelos de redes, su clasificación en base a su estructura, las técnicas de medición utilizadas y los sistemas que permiten realizar un análisis de estas estructuras.

2.1 Orígen de las Redes Sociales

Una *red social* es una estructura social que está conformada por grupos de individuos, a los cuales se les llama nodos y los cuales estan conectados mediante algún tipo de relación (p.ej., amistad, negocios, parentesco, etc). Este tipo de estructuras son frecuentemente utilizadas para modelar una situación social. En 1930 aparece la sociometría como una manera de formalizar las ciencias sociales, donde se utilizaba la estadística para estudiar poblaciones y a la teoría de grafos servía para modelar la relación entre personas.

En años recientes la teoría basada en la estructura de las redes sociales fue retomada, debido al alto número de sistemas en la Web que permiten generar comunidades virtuales que pueden ser representadas como una estructura de red social.

2.2 Análisis de Redes Sociales

El Análisis de Redes Sociales (ARS[1]) nace en los años 70 con la fundación de la *International Network for Social Network Analysis* (INSNA[2]), la cual es una asociación profesional para investigadores interesados en el análisis de redes sociales aplicada en diferents áreas del conocimiento.

El análisis de redes es una aproximación intelectual amplia para identificar las estructuras sociales que emergen de las diversas formas de relación, pero también es un conjunto específico de métodos y técnicas. El ARS se ha desarrollado como herramienta de medición y análisis de las estructuras sociales que surguen de las relaciones entre actores sociales diversos (individuos, organizaciones, naciones, etc).

2.3 Propiedades de las Redes Sociales

Las redes sociales son representadas mediante grafos, y utiliza técnicas de la teoría de grafos para estudiar la estructura de las redes sociales. Sin embargo, existen diferencias en la forma en como se aplican los distintos conceptos de la teoría de grafos sobre el ARS. Las redes sociales adoptan muchas de las propiedades de las redes sociales, en esta sección estudiamos las diferentes propiedades que presentan las redes sociales en base a su forma, distribución y similaridad entre los conjuntos de nodos y relaciones que existen en la red.

2.3.1 Distancias en las redes

La distancia es una medición de la teoría de grafos que permite definir propiedades más complejas sobre la posición de los individuos y sobre la estructura de la red. Una trayectoría es el numńero de enlaces que existen entre dos nodos en el grafo. La distancia geodésica es el número mínimo de enlaces que llevan de un nodo a otro (trayectoria mínima), esta medida es muy utilizada en el ARS, ya que nos permite obtener el camino más eficiente entre dos actores, sin embargo, para el caso de las redes sociales el camino más corto no siempre es el buscado y dependerá del tipo de red social para determinar el tipo de distancia a utilizar.

El diámetro de un grafo está definido como el máximo geodésico de todos los vértices dentro de un grafo. El diámetro representa el tamaño del grafo y permite saber que tan grande es, se dice que las redes de mundo pequeño (ver sección 2.6.2.2) poseen un diámetro pequeño. En la figura 2.1 se muestra el diámetro la red social del *Club de karate de Zachary*, utilizada en varios trabajos de investigación de redes sociales [34, 63] principalmente para trabajos de detección de comunidades.

[1] *SNA (social network analysis) por sus siglas en inglés*
[2] Para más detalle visitar `http://www.insna.org/`

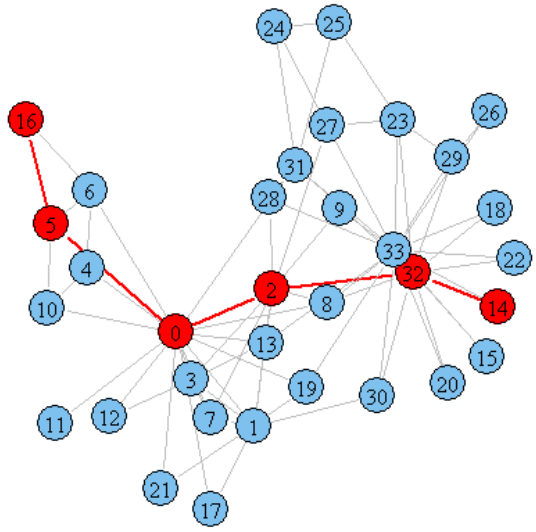


Figura 2.1: Ejemplo del cálculo del diámetro para la red del club de karate de Zachary.

2.3.2 Tipos de interacción

La forma en que interactúan los nodos dentro de un grafo permite establecer un nivel de conexión entre nodos y sus relaciones. A este fenómeno se le llama conectividad y permite reducir las trayectorias entre dos nodos. En un grafo con conectividad alta se puede ir de un nodo a otro con trayectorias más cortas.

La reciprocidad en las redes sociales es un fenómeno presentado en grafos dirigidos, también llamado *simetría de vínculos*. Este fenómeno se presenta cuando en un grafo existe un enlace que va de A hacia B y uno que va de B hacia A. La simetría de vínculos en un grafo reduce el diámetro de la red e incrementa la conectividad de la red.

2.3.3 Coeficiente de Agrupamiento

El coeficiente de agrupamiento es una métrica de similaridad. Para esta tesis se utilizó el coeficiente de agrupamiento local y global para estudiar la similaridad entre nodos de la red. El coeficiente de agrupamiento local mide el nivel de agrupamiento o interconexión de un vértice con sus vecinos, donde el vecindario de un nodo i está formado por el número de nodos que están adyascentes al nodo i, es decir, aquellos nodos con los que se mantiene una relación. El coeficiente de agrupamiento global indica el nivel de agrupamiento de los nodos con respecto a la red total. En [86] Watts y Strogatz proponen el cálculo del coeficiente de agrupamiento para redes de mundo pequeño (ver 2.7.2.1).

2.3.4 Cliques

Un clique en la teoría de grafos es un conjunto de vértices en el cual para cada vértice existe una arista que los conecta. Un clique es un subgrafo en el cual cada vértice está conectado a cada otro vértice del grafo, es decir, el subgrafo puede ser considerado como un grafo completo. En la figura 2.2 se muestra un grafo completo, si es un subgrafo de una red se dice que es un clique de tamaño 6.

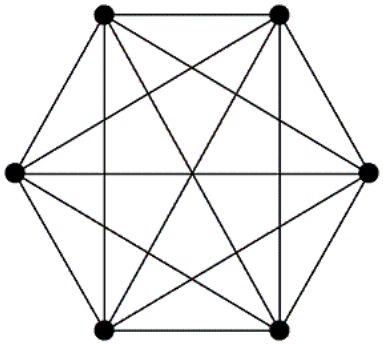


Figura 2.2: Ejemplo de un clique de tamaño 6.

2.4 Clasificación de Redes Sociales

Actualmente no existe una clasificación que permita especificar los tipos de redes sociales, ya que dependiendo el área de investigación, una red social puede ser clasificada según el enfoque del estudio, pero es necesario contar con una clasificación que permita distinguir a las nuevas redes sociales sin caer en ambigüedades. A continuación proponemos una clasificación de los tipos más comunes de redes sociales.

- **Redes basadas en su tamaño.** Este tipo de redes depende del diámetro de la red, es decir, la distancia[3] mayor entre dos actores en la red. No existe una medida exacta para poder determinar cuando una red social es grande o pequeña, por lo que dicho valor dependerá de lo que se quiera representar con la información.

  - **Redes a pequeña escala.** Este tipo de redes pueden ser analizadas por la mayoría de los sistemas para visualización de redes, y generalmente se utiliza al conjunto de datos completo para su análisis. Un ejemplo de este tipo de redes sociales es la red formada a partir de la colaboracción de investigadores, donde los investigadores son los nodos de la red y los enlaces estan formados por los artículos en los que uno o más investigadores participan.

[3]Entiendase como distancia al número de nodos que existen en una trayectoría, donde una trayectoría está formada por los enlaces formados entre los nodos.

  - **Redes a gran escala.** Estas redes no pueden ser analizadas fácilmente y en muchas ocasiones suelen realizarse las mediciones en base a una porción representativa de la red. Aún no existe sistema alguno que permita visualizar este tipo de redes de manera completa. Un ejemplo de estas redes es la formada por los sistemas de correo electrónico, donde las cuentas de usuario son los actores en la red y un enlace es representado por un usuario en la lista de contactos, donde resultaría difícil tratar de representar una red de todas las cuentas de correo que interactúan en la Web.

- **Redes basadas en la evolución.** Este tipo de redes depende de los cambios que sufre la red a través del tiempo. Estas redes pueden ser de cualquier tamaño y tener distintas formas.

  - **Redes Estáticas.** Este tipo de redes no sufre ningún tipo de alteración cuando son sujetas a estudio, por lo que mantienen la misma estructura desde el momento en que son analizadas hasta el final de su estudio.
  - **Redes Dinámicas.** Estas redes sufren cambios en su estructura debido a la incorporación y/o eliminación de nuevos actores y las relaciones entre ellos. Existen muchas redes de este tipo pero la más importante es la propia Web, ya que se encuentra en constante cambio y su tamaño aumenta cada vez más al pasar el tiempo. Diferentes fenómenos se presentan en este tipo de redes y muchos trabajos han sido propuestos como son la predicción de vínculos[69] y los modelos de crecimiento [17].

- **Redes basadas en su origen.** Este tipo de redes depende de su fuente de datos de origen. Muchas de estas redes pueden representar comunidades virtuales y/o del mundo real [92].

  - **Redes fuera de línea (Off-line).** Son aquellas en las que las relaciones sociales son establecidas sin la intervención de un medio electrónico, es decir, la administración y conocimiento de las relaciones recae exclusivamente en el conocimiento del individuo sin ayuda de un sistema software que le permita llevar la gestión de contactos. Un ejemplo de este tipo de redes fué la red generada para el caso del presunto suicidio del científico británico David Kelly [78], dicha red fue generada a partir de documentos del gobierno sobre el caso.
  - **Redes en línea (On-line).** Son redes que dependen altamente de medios electrónicos y se mantienen ligadas a los cambios en la tecnología de los sistemas. Ejemplo de estas redes son Facebook, Twitter, Orkut, entre otras. En la sección 2.4.2 se describe de forma más detallada este tipo de redes.

- **Redes basadas en su topología.** Este tipo de redes depende de la complejidad de la red.

  - **Redes Simples.** Este tipo de redes son estructuras sencillas y pueden ser facilmente analizadas con conceptos básicos de la teoría de grafos.

– **Redes Complejas.** El estudio de este tipo de redes está basada en el estudio empírico de las redes de mundo real, se dice que son complejas por que presentan propiedades no triviales (p.ej., calcular el coeficiente de agrupamiento y la centralidad de la red). Ejemplo de este tipo de redes son las redes aleatorias [68], las de mundo pequeño [56, 86] y las libres de escala [21].

2.4.1 Redes de Mundo Real

Las redes de mundo real son un tipo de redes que representan diferentes situaciones de la vida cotidiana y que a simple vista no parecen tan complejas de entender, en algunos casos estas redes modelan el comportamiento de la sociedad y la forma en como estas interactúan bajo situaciones reales. Ejemplo de este tipo de de redes son las formadas por los alumnos de las escuelas primarias en un determinado estado, o el número de enfermos de cáncer en un hospital, o las comunidades de grupos étnicos en el país, entre otros. Sin embargo, las redes de mundo real no solo se aplican en estructuras sociales, sino que se pueden observar en otro tipo de situaciones como: la propagación de epidemias, las redes de interacción de proteínas en el metabolismo celular, el comportamiento de los animales, las redes de neuronas en los organismos del sistema nervioso, la red de distribución eléctrica, las redes de carreteras, entre otras.

2.4.2 Redes Sociales en línea

Recientemente las redes sociales en línea han adquirido gran popularidad y se encuentran dentro los sitios más importantes en la Web. Sin embargo, este tipo de redes se ha convertido en uno de los principales generadores de tráfico en Internet, donde investigaciones recientes [4] de la empresa CISCO revelan que más del 30% del tráfico en las redes es generado por este tipo de sistemas y se pronostica que para el año 2013 el tráfico generado por estas redes aumentará en un 500%.

La tabla 2.1 muestra la lista de los sistemas de redes sociales en línea más importantes dentro de la Web. Esta tabla fué generada a partir de la información proporcionada por la compañia Alexa[5] en su portal y de datos estadísticos proporcionados por las páginas de los sistemas enlistados.

Según datos estadísticos la cantidad de personas en el mundo es aproximadamente 6.8 mil millones, donde países como China con 1.3 mil millones, India con 1.1 mil millones y Estados Unidos con 295, millones de habitantes son los países con mayor densidad de población. Por otro lado sitios como Facebook, MySpace y Youtube promedian 280 millones de usuarios cada uno. En base a esto podemos decir que las redes sociales en línea tienen

[4]Para leer el artículo completo visitar `http://newsroom.cisco.com/dlls/2009/prod_102109.html`

[5]Alexa es una compañia en la Web que proporciona información sobre el tráfico en Internet y mantiene una lista actualizada de los sistemas más utilizados según su historial de tráfico

condiciones para ser consideradas como redes de mundo real, aunque este tipo de relaciones son inpersonales y poco confiables.

2.5 Representación de las Redes Sociales

En una red social cada nodo, también llamado actor o vértice, puede ser representado como un individuo o grupo de individuos. Una arista, también llamada relación o vínculo, conecta a dos nodos y representa el enlace entre dos individuos en una red social. Las redes pueden tener pocos o muchos actores y uno o más tipos de relaciones entre pares de actores. El ARS usa dos tipos de herramientas de las matemáticas para representar información de las relaciones entre actores sociales, los grafos y las matrices, una razón para usar métodos formales para representar redes sociales es que la representación matemática permite utilizar computadoras para dicho análisis.

El ARS usa un tipo de representación gráfica de estas estructura que consiste de puntos (o nodos) para representar actores y líneas (o aristas) para representar lazos o relaciones. Sin embargo, cuando los sociólogos empezaron con el análisis de redes tomaron esta representación y la llamaron *sociograma*. Matematicamente los sociogramas son conocidos como *grafos dirigidos*. En la figura 2.3 se aprecia una red social como un grafo dirigido o sociograma.

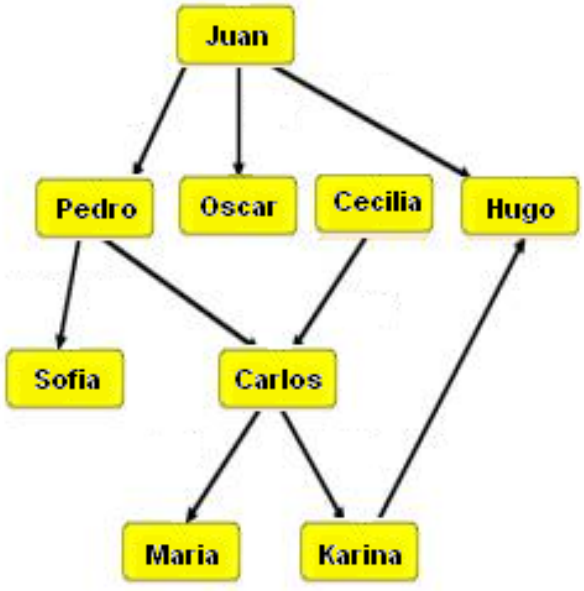


Figura 2.3: Ejemplo de una red social como un grafo dirigido o sociograma.

Red Social	Fecha Inicio	Tipo de Red	Servicios que ofrecen	Usuarios (millones)	Tráfico en Internet *	Mayor Presencia **
Facebook	Mar-97	Personal	Uso de blogs, aplicaciones y chat	300	2	E.U.A(3)
Youtube	Feb-05	Contenido	Vídeos, tagging, etc	285	4	E.U.A(4)
Windows Live	Dic-94	Blogging	Compartir archivos y aplicaciones	120	5	México(2)
Wikipedia	Ene-01	Wiki	Edición de artículos	10	6	E.U.A(5)
QZone	May-95	Personal	Compartir archivos y aplicaciones	200	11	China(2)
Twitter	Ene-00	Microblogging	Notificaciones en tiempo real	44.5	12	E.U.A(12)
MySpace	Feb-96	Personal	Uso de blogs, aplicaciones y chat	264	14	E.U.A(6)
Flickr	Nov-03	FotoBlog	Compartir fotos	32	33	E.U.A(21)
Hi5	Jun-96	Personal	Compartir archivos y aplicaciones	80	44	México(11)
LinkedIn	Nov-02	Profesional	Encontrar u ofrecer trabajo	50	50	E.U.A(19)
LiveJournal	Apr-99	Blogging	Adminsitrar un periódico	23.8	77	Rusia(10)
Friendster	Mar-02	Privada	Compartir archivos y aplicaciones	90	121	Filipinas(5)
Orkut	Dic-02	Privada	Administración de contactos	67	147	India(45)
Fotolog	Abr-02	Contenido	Compartir fotos	20	205	Argentina(11)
Bebo	Jul-03	Personal	Compartir fotos y el uso de blogs	40	285	Inglaterra(31)
Sonico	Mar-00	Personal	Compartir archivos y aplicaciones	17	558	México(70)
* Posición de la red con respecto a la generación de tráfico en Internet. ** El número entre paréntesis indica la posición de la red social dentro de ese país.						

Tabla 2.1: Lista de sitios de redes sociales en Internet.

2.6 Estructura de las Redes Sociales

La estructura de una red social está basada en individuos u organizaciones, que estan conectadas por una o más relaciones, tales como, amistad, contactos profesionales, parentesco, entre otros. Las redes sociales en términos de la teoría de grafos representa una estructura basada en grafos complejos. Las redes sociales pueden ser representadas como grafos dirigidos, no dirigidos, bipartidos y completos.

2.6.1 El componente gigante de las redes sociales

En el año 2000 se publica un artículo [16] sobre las características de la Web, entre sus principales descubrimientos encontraron que un número pequeño de nodos presentaba alto agrupamiento, a este fenomeno le llamaron *Componente Fuertemente Conectado* (CFC) o *Componente Gigante* (CG) y también demostraron que existía la presencia de más de uno. Esto permitió diferenciar ampliamente a las redes aleatorias de las redes de mundo real, ya que las aleatorias presentan una distribución homogénea entre sus nodos, es decir, no hay presencia de CFC's y los nodos presentan una distribución de grado similar.

Un CFC es definido como un conjunto de nodos tal que para cualquier par de nodos u y v en el conjunto hay un camino entre u a v. La figura 2.4 muestra la conectividad de la Web, donde el CFC es el núcleo de la red y el conjunto de nodos en **IN** son los nodos que *entran* en el CFC y el **OUT** es el conjunto de nodos que *salen* del CFC, los *tentáculos* representan a los nodos que conectan a un CFC con otros tipo de formaciones, ya sea otro CFC o islas[6] de nodos.

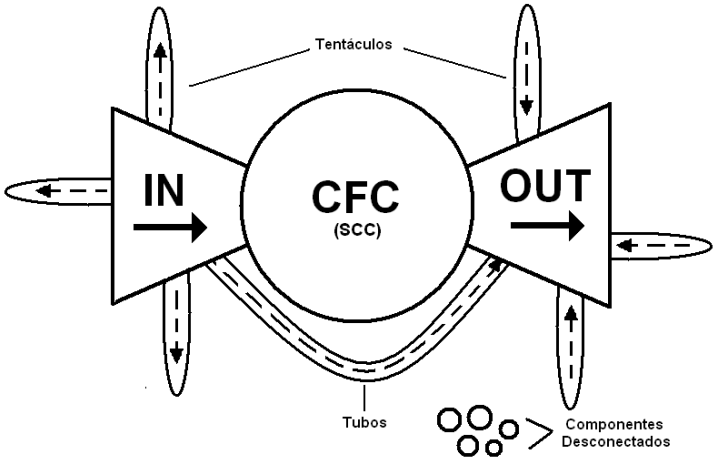


Figura 2.4: Conectividad de la Web, donde el CFC es el núcleo de la red y el conjunto de nodos en **IN** son los nodos que *entran* en el CFC y el conjunto de nodos **OUT** son los que *salen* del CFC. Los *tentáculos* y los *tubos* representan a los nodos externos al núcleo CFC.

[6]Entiendase como isla a un conjunto pequeño de nodos que esta debilmente enlazado a un CFC

2.6.2 Redes Complejas

En trabajos recientes [14, 30] se ha mostrado que las redes de mundo real son más complejas en el sentido que diferentes caracterízticas topológicas son derivadas de la teoría de las redes aleatorias. Un grafo complejo contiene muchos subgrafos diferentes. Muchos sistemas en la naturaleza estan construidos por un alto número de conexiones dinámicas (p.ej., redes neuronales y el Internet) y existen diferentes modelos de redes complejas como las redes aleatorias, mundo pequeño y libres de escala que permiten estudiar el comportamiento de muchas redes del mundo real.

2.6.2.1 Redes aleatorias.

Este tipo de redes son generadas a partir de agregar nodos de manera aleatoria a un conjunto de datos fijos. Este tipo de redes presentan caminos muy cortos entre nodos y un coeficiente de agrupamiento bajo, además presentan una distribución Binomial o de Poisson como en la figura 2.5. Existen diferentes modelos que han sido propuestos para este tipo de redes como el modelo Erdös-Rényi (ER)[67] y el de Gilbert[33], siendo estos los primeros modelos aplicados a las redes sociales. Para más detalle sobre este tipo de modelos ver la sección 2.7.1.

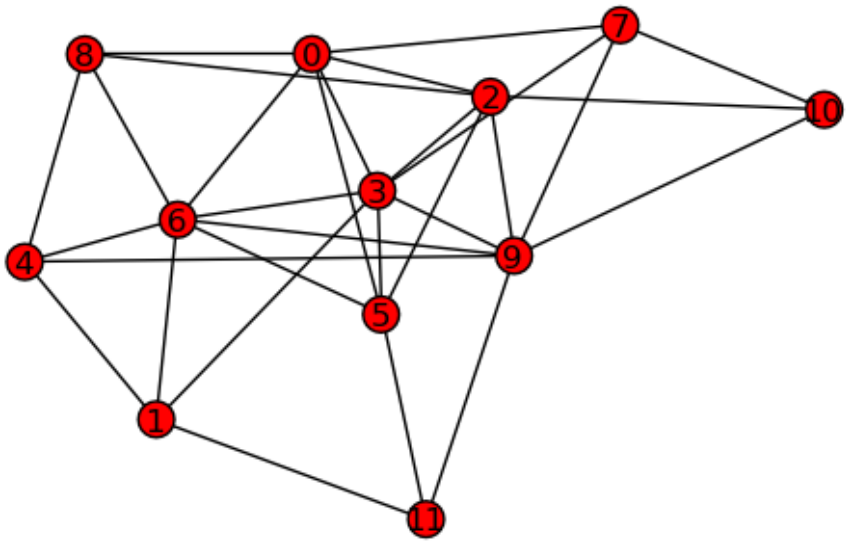


Figura 2.5: Ejemplo de una red aleatoria con 12 vértices y 29 aristas. Las aristas son generadas entre pares de vertices seleccionados uniforme y aleatoriamente. Como resultado se tiene una distribución de grado binomial o de Poisson.

2.6.2.2 Redes de mundo pequeño

Las *redes de mundo pequeño* son redes que tienen un diámetro[7] pequeño y exhibe un alto agrupamiento. El problema del mundo pequeño fue planteado en 1967 por Stanley Milgram[56], donde textualmente se hace estas dos preguntas:

1. *"Comenzando con dos personas cualquiera en el mundo, ¿Cuál es la probabilidad que ellos se conozcan?"*

2. *"Dados dos personas cualquiera en el mundo, persona X y persona Z, ¿Cuantos enlaces intermedios son necesarios antes que X y Z se conecten"*

Estas interrogantes motivaron a Milgram a generar una serie de experimentos, que consistió basicamente en monitorear a un grupo de personas que interactuaban enviando cartas a desconocidos, los resultados mostraron que la red de la sociedad humana presenta una estructura de mundo pequeño y que los individuos estan a *"seis grados de separación"*. Watts [85] describe la teoría de los *seis grados* y plantea que *"el mundo es un pañuelo"*, es decir, cualquiera en el mundo permanece conectado a otra persona a través de otras personas y que el número de personas intermedias es menor a cinco. Watts y Strogatz proponen un modelo basado en el fenómeno de las redes mundo pequeño[86], para más detalle sobre este modelo ver 2.7.2.

Estudios recientes han mostrado que la Web es del tipo mundo pequeño [1], debido al alto índice de agrupamiento y de las distancias tan cortas que exhiben estos sitios. En un estudio sobre redes criminalistas desarrollado por investigadores de la universidad de Arizona [88], emplearon la teroía de las redes de mundo pequeño para determinar que individuos dentro de la red formaban parte o podían en algún momento formar parte de un grupo de terroristas o criminales.

2.6.2.3 Redes ley de potencia

Son redes en donde 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 fijo que debe de satisfacer a $\alpha > 1$. Cuando la probabilidad de alguna variable se distribuye de acuerdo a una ley de potencia, su función de distribución se define como:

$$p(x) = Cx^{-\alpha} \tag{2.1}$$

donde $p(x)$ es la frecuencia (probabilidad) de que la variable tome un valor de x, α es el exponente de la distribución, x es la variable que se requiere analizar y C es una constante que depende del tipo de evento.

[7]Entiendase por diámetro al geodésico (camino más corto entre dos nodos) más grande en la red. El diámetro refleja lo grande que es la red.

Trabajos de investigación [66] muestran que las redes ley de potencia siguen una ley de Pareto, la cual describe la distribución desigual de la riqueza dentro en una red. Una función muy utilizada para encontrar la relación entre una ley de potencia y la ley de pareto es la Función de Distribución Acumulada Complementaria (CCDF[8]) de potencia. Siendo $P(x)$ la distribución acumulada complementaria, entonces:

$$P(x) = \int_x^{\infty} p(x)dx = (\frac{x}{x_{min}})^{-\alpha+1} \tag{2.2}$$

donde x_{min} es el menor valor a partir del cual se satisface la ley de pareto. Mientras más grande sea α, mayor será la frecuencia acumulada $P(x)$ de algún valor x. Un fenómeno que ocurre en este tipo de redes es que los nodos con un alto grado son pocos y los nodos con un bajo grado son muchos, es decir, la concentración de enlaces se da en un número pequeño de nodos de la red. La distribución acumulada complementaria en escala logarítmica está dada por:

$$ln(P(x)) = (-\alpha + 1) \cdot ln(\frac{x}{x_{min}}) \tag{2.3}$$

El exponente α se puede estimar por varios métodos: un primer método es graficar los datos y calcular la pendiente; el otro es utilizar mínimos cuadrados ordinarios, aunque en este caso se recomienda estimar este parámetro usando máxima verosimilitud[9]. La ecuación para estimar el parámetro α por el método de máxima verosimilitud es:

$$\hat{\alpha} = 1 + n(\sum_{i=1}^{n} ln\frac{x}{x_{min}})^{-1} \tag{2.4}$$

donde n es el número de observaciones y x_{min} es el menor nivel de la variable, a partir de donde se cumple la ley de potencia. El error estadístico de este parámetro, estimado por máxima verosimilitud es:

$$\hat{\sigma} = \sqrt{n} \cdot (\sum_{i=1}^{n} ln\frac{x_i}{x_{min}})^{-1} = \frac{\hat{\sigma} - 1}{\sqrt{n}} \tag{2.5}$$

Es necesario estimar un valor de x_{min} que permita hacer un buen ajuste de los datos con la distribución de Pareto, además de una buena estimación del parámetro α. Si se escoge un valor de x_{min} muy bajo es probable que se presente un sesgo en el parámetro estimado. Por otro lado, si se escoge un valor de x_{min} muy alto se excluye información importante.

[8]Complementary Cumulative Distribution Function por su traducción en inglés

[9]El método de máxima verosimilitud es una técnica de estimación para una muestra finita de datos.

Existen varias maneras de estimar el valor de x_{min}. Newman y Clauset [20] recomiendan la prueba de Kolmogorov-Smirnov, en la que se busca minimizar la distancia entre la distribución observada y la distribución con la que se quiere hacer el ajuste, la cual está definida como:

$$D = MAX_{x \geq x_{min}} |S(x) - P(x)| \tag{2.6}$$

donde $S(x)$ es la distribución complementaria observada y $P(x)$ es la distribución complementaria de potencia. En este caso, D es la distancia absoluta entre estas dos distribuciones. La idea fundamental es encontrar el valor de x_{min} que minimice la máxima distancia entre $S(x)$ y $P(x)$. A partir de un muestreo aleatorio de tamaño n sobre los datos observados se estiman los parámetros x_{min} y α.

Algunos investigadores aseguran que muchas redes mundo pequeño son redes ley de potencia como la Web[2], la frecuencia en el uso de palabras del lenguaje humano[94] y el número de citas recibidas por artículos[23]. La figura 2.6 muestra las gráficas de una distribución ley de potencia y una normal.

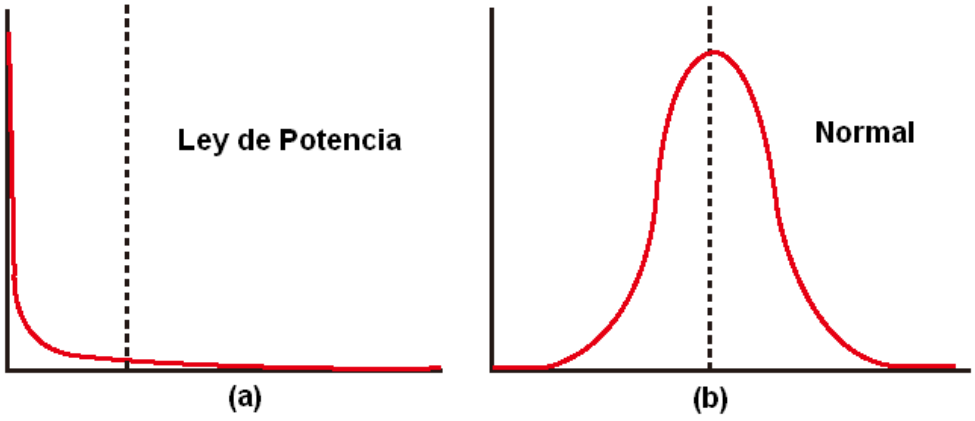


Figura 2.6: Gráfica para una distribución (a) ley de potencia y una (b) Normal o Binomial.

2.6.2.4 Redes libres de escala

Las redes libres de escala son una clase de redes leyes de potencia[11], donde un nodo con un alto grado tiende a ser conectado a otro nodo con un alto grado, es decir, el número de enlaces en la red 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, ya que en este tipo de redes hay más nodos con pocos enlaces que nodos con un gran número de enlaces, con esto se garantiza que el sistema este altamente conectado.

Una de las principales diferencias entre las redes libres de escala y las redes aleatorias es la distribución de grado. No todos los nodos en una red tienen el mismo número de enlaces (grado nodal), una red aleatoria genera sus enlaces aleatoriamente y esto provoca que la mayoría de los nodos tengan el mismo grado dado que siguen una distribución de Poisson (ver. 2.6.2.3). La figura 2.7 muestra la distribución en las redes libres de escala y las redes aleatorias, se puede apreciar que las redes aleatorias presentan una homogeneidad en sus enlaces, mientras que en las redes libres de escala no existe tal homogeneidad.

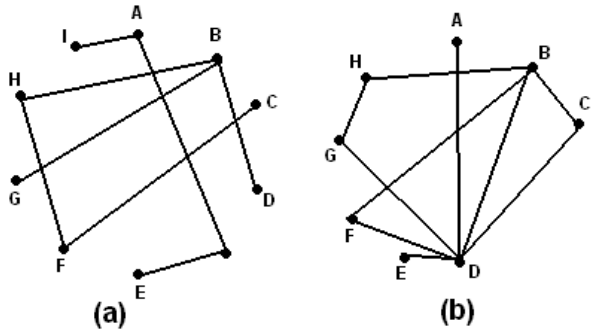


Figura 2.7: Ejemplo de una (a) red aleatoria y una (b) red libres de escala. Nótese que la distribución de enlaces en (a) es homogénea.

La idea principal de este tipo de redes parte del crecimiento de las redes en sistemas complejos con la adición de nuevos nodos y las relaciones formadas en redes existentes y por exhibir un comportamiento de *anexo preferencial*, es decir, existe una alta probabilidad de que un nuevo nodo se agrege a la red formando un enlace con un nodo que posee un gran número de enlaces. Muchas redes del mundo real presentan un comportamiento de libre de escala (p.ej., la estructura de la red celular[5] y la red de e-mail[25]), otras redes utilizan las propiedades ofrecidas por este tipo de redes para medir su estructura (p.ej., la red formada por las ontologías en la Web semántica[93]).

2.6.3 Centralidad en las Redes Sociales

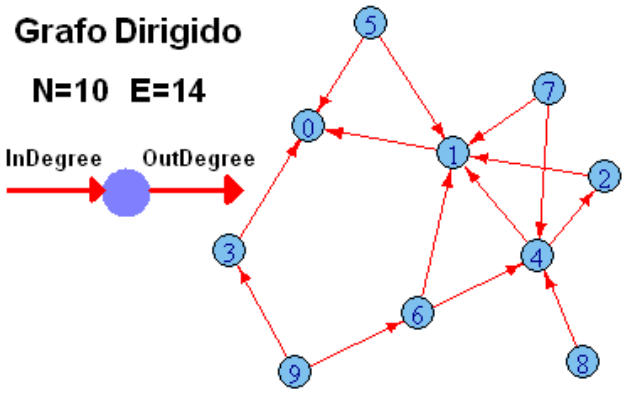
La necesidad por entender el comportamiento de los nodos en una red social es algo que hace al concepto de centralidad muy importante en el análisis de redes sociales. Con la centralidad se puede determinar que tanto un nodo influye en comparación a los demás elementos de la red. La idea de centralidad [32] aplicada en las relaciones humanas fue introducida por Bavelas en 1948, quien dirigió el primer trabajo de investigación sobre centralidad para el *Group Networks Laboratory* del M.I.T. en la decada de los 40. Los cálculos de centralidad se basan en el uso de las matemáticas y los conceptos de la teoría de grafos. En la actualidad existen diversos estudios [36, 22] que aplican las medidas de centralidad para las redes sociales.

El tipo de medición de la centralidad se distingue por el tipo de grafo, en el caso de los grafos no dirigidos el sentido de las relaciones no importa, en los grafos dirigidos si importan. Trabajos anteriores [84] han definido a las relaciones en los grafos dirigidos como *prestigio*, el cual representa y permite formular las medidas de centralidad de distinta forma.

Existen diferentes tipos de medición de centralidad, en esta tesis usamos las siguientes: centralidad de grado (*Degree*), centralidad de cercanía (*Closeness*) y la centralidad de Intermedicación (*Betweenness*), a continuación se detalla de manera más clara este tipo de métricas utilizadas para el análisis de redes.

2.6.3.1 Centralidad de Grado.

Se define al grado de un nodo como el número de enlaces que posee un nodo, es decir, el número de relaciones que tiene el nodo con los otros nodos. Se dice que un nodo tiene un vecindario el cual está formado por los nodos a su alrededor. Para los grafos dirigidos el grado del nodo suele ser dividido en dos formas: el grado de entrada (*in-degree*) y el grado de salida (*out-degree*). El grado de entrada del nodo i en un grafo dirigido es el número de nodos N_v en el vecindario que tienen un enlace que se dirigen hacia i, el grado de salida del nodo es el número de nodos N_v en el vecindario que poseen enlaces que son dirigidos desde i. La figura 2.8 muestra el grado de entrada y el grado de salida para los nodos de un grafo dirigido.



Distribución de Enlaces

Nodo	Grado de Entrada (In-degree)	Grado de Salida (Out-degree)	Grado Nodal (Degree)
0	3	0	3
1	5	1	6
2	1	1	2
3	1	1	2
4	3	2	5
5	0	2	2
6	1	2	3
7	0	2	2
8	0	1	1
9	0	2	2

(b)

Figura 2.8: Ejemplo de los diferentes tipos de grados en un grafo dirigido. La figura (a) muestra un grafo dirigido con 10 nodos y 14 aristas y en la tabla (b) se muestra el valor del grado de entrada (in-degree), grado de salida (out-degree) y el grado (degree) para cada uno de los vértices en la red.

Por lo tanto, el grado k_i de un nodo se calcula por:

$$k_i = \sum_{j}^{n} a_{ij} \tag{2.7}$$

n es el número de enlaces del nodo, a_{ij} es el enlace entre los nodos a_i y a_j.

El cálculo del grado de entrada $K_{i(in)}$ y de salida $K_{i(out)}$ está dado por:

$$k_{i(in)} = \sum_{j=1}^{g_{in}} a_{ij_{in}} \tag{2.8}$$

$$k_{i(out)} = \sum_{j=1}^{g_{out}} a_{ij_{out}} \tag{2.9}$$

donde:

- g_{in} y g_{out} es el número de enlaces totales de entrada y de salida respectivamente.
- $a_{ij_{in}}$ y $a_{ij_{out}}$ representan los enlaces de entrada y de salida entre el vértice i y j respectivamente.

2.6.3.2 Centralidad de Cercanía.

La centralidad de cercanía mide los pasos requeridos para accesar a cada otro vértice de un vértice dado. Esta medida se basa en el uso de la media geodésica entre un nodo y los demás nodos de la red, y se define a la centralidad de un vértice como la inversa de la longitud promedio de los caminos cortos con los demás vértices de la red. Wasserman y Faust [84] define las ecuaciones para el cálculo de la centralidad de cercanía de un vértice para grafos no dirigidos como sigue:

$$C_C(n_i) = \frac{1}{\sum_{j=1}^{g} d(n_i, n_j)} = \frac{1}{MED_{i \neq j}(d(n_i, n_j))} \tag{2.10}$$

donde:

- g es el número de nodos en la red.
- $d(n_i, n_j)$ es el geodésico entre el nodo n_i y el nodo n_j
- *MED* es la media de las distancias.

Para grafos dirigidos se planteó la siguiente ecuación:

$$C_C(n_i) = \frac{g-1}{\sum_{j=1}^{g} d(n_i, n_j)} = \frac{g-1}{MED_{i \neq j}(d(n_i, n_j))} \tag{2.11}$$

2.6.3.3 Centralidad de Intermediación.

Existen dos tipos de centralidad de intermediación, la primera está basada en la frecuencia en la que un *nodo* aparece en el geodésico entre dos nodos, es decir, las veces en que se presenta entre un nodo con trayectoria mínima. La siguiente ecuación calcula la intermediación [10] de v en las trayectorias de z y y

$$C_I(v) = \sum_{v \neq y \neq z} \frac{d_{yz}(v)}{d_{yz}} \tag{2.12}$$

donde:

- d_{yz} es el número de caminos geodésicos que van de y a z.
- $d_{yz}(v)$ representa el número de caminos geodésicos de y a z que cruzan por el vértice v.

El segundo cálculo de centralidad de intermediación está basado en la importancia que tiene una *arista* con respecto a una trayectoria mínima, es decir, las veces en que un enlace se presenta en medio de una trayectoria mínima. La ecuación para calcular la intermediación de aristas es la siguiente:

$$C_I(e) = \sum_{x \neq y} \frac{d_{yz}(e)}{d_{yz}} \tag{2.13}$$

donde:

- d_{yz} es el número de caminos geodésicos que van de y a z.
- $d_{yz}(e)$ representa el número de caminos geodésicos de y a z que cruzan por la arista e.

El cálculo de la intermediación está dado en un tiempo de $O(nm + n\ 2\log \mathrm{n})$ y en espacio $O(n + m)$. En [15] se presenta un algoritmo que realiza el cáclulo de la centralidad de intermediación más eficientemente.

[10] *Betweenness* por su traducción del inglés

2.6.3.4 Ejemplo del cálculo de la centralidad para un grafo.

La figura 2.9 muestra las centralidades de grado, cercanía e intermediación para un grafo dirigido.

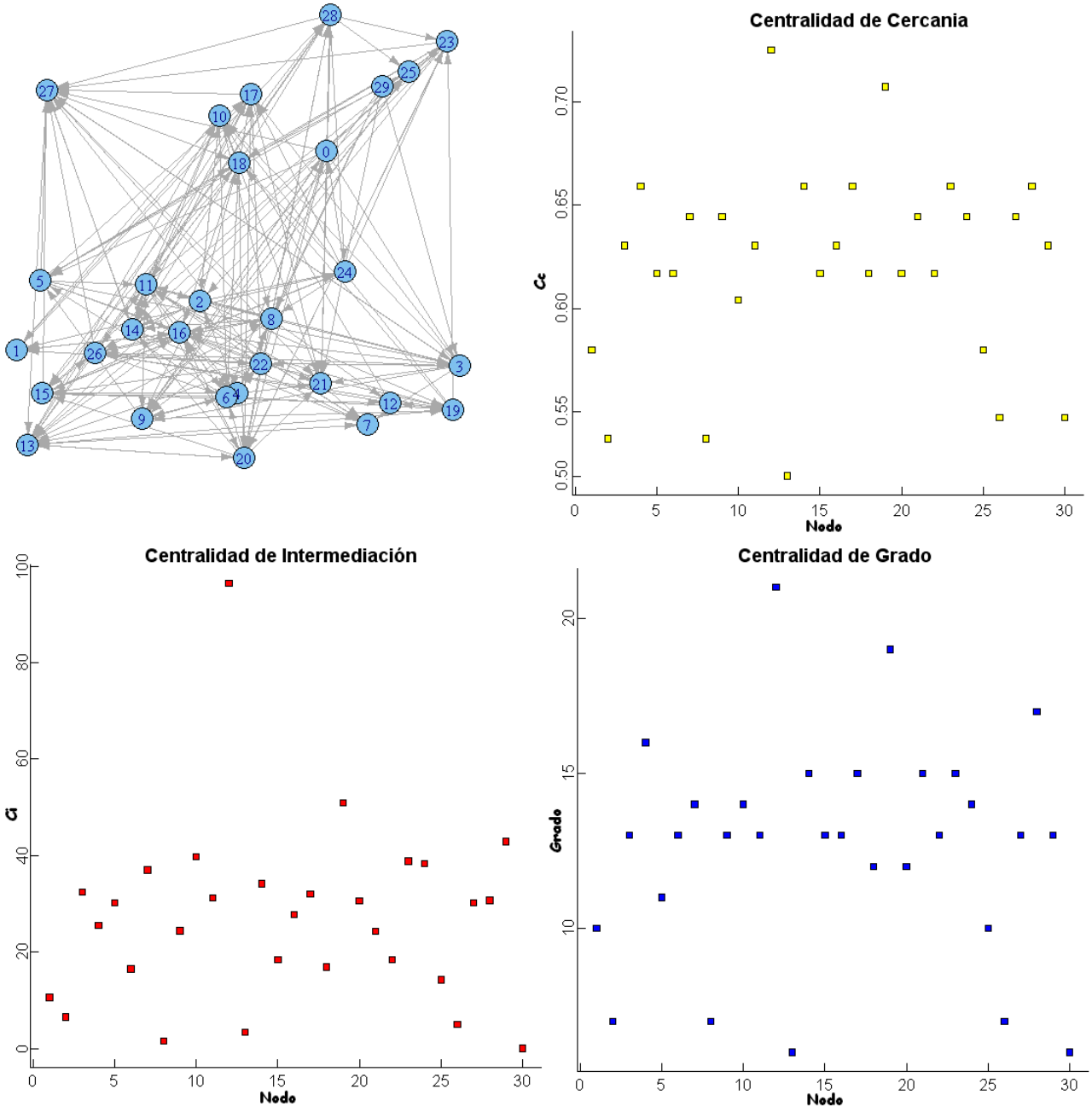


Figura 2.9: Cálculo de las diferentes medidas de centralidad de un grafo dirigido.

2.7 Modelos de redes sociales

El tipo de red y sus propiedades depende de como se lleva a cabo su formación, es por eso que se han propuesto diferentes modelos de crecimiento de redes. Los modelos más antiguos son los modelos de crecimiento de redes aleatorias. Sin embargo, pocas son las redes del mundo real que siguen tal crecimiento. La necesidad de modelar a las redes del mundo real motivo al desarrollo de nuevos modelos para redes libres de escala, donde el *anexo preferencial* permite modelar el fenómeno de *los ricos aún más ricos*. Este tipo de redes son aplicadas para modelar la estructura de las redes sociales.

2.7.1 Modelos de Redes Aleatorias

Los modelos de redes aleatorias son considerados como los primeros modelos de redes sociales. Los modelos más representativos de esta clase son: el modelo de Gilbert y el modelo Erdös-Rényi (ER). Cabe mencionar que ambos modelos poseen características muy similares, pese a que se desarrollaron de manera paralela.

2.7.1.1 Modelo de Gilbert

En el año de 1959 Gilbert [33] introduce un modelo para grafos aleatorios. En este modelo se propone el uso de dos probabilidades P_N y R_N para modelar el comportamiento de la red, donde P_N es la probabilidad global del grafo y R_N es la probabilidad de los enlaces existentes a partir de la construcción de un grafo aleatorio. Donde P_N y R_N son aproximadas mediante las ecuaciones 2.14 y 2.15.

$$P_N \sim 1 - Nq^{N-1} \tag{2.14}$$

$$R_N \sim 1 - 2q^{N-1} \tag{2.15}$$

donde N es el número de nodos en la red y q es el grado del nodo.

2.7.1.2 Modelos Erdös-Rényi

El modelo Erdös-Rényi (ER) llamado así por sus creadores Paul Erdös y Alfréd Rényi, es un modelo que fue presentado en una serie de publicaciones [26, 27, 28] a finales de los años 50 y principio de los 60. El modelo realiza la conexión de los N nodos con una probabilidad p, creando un grafo con aproximadamente $pN(N-1)/2$ aristas distribuidas aleatoriamente. En la figura 2.10 se muestran cuatro diferentes grafos generados[11] utilizando el modelo ER, donde se pueden apreciar diferentes grafos con distintas probabilidades de distribución y número de nodos distintos, las gráficas muestran el grado de distribución para cada grafo,

[11] Estos grafos fueron generados mediante el uso de R (ver sección 2.8.3) y el paquete igraph.

se puede apreciar la forma en que los nodos se distribuyen dentro de la red, presentando una distribución homogénea, es decir, la mayoría de los nodos tienen el mismo número de enlaces. En las gráficas de la figura 2.10 se puede ver como la distribución de cada nodo son parecidos entre sí y presentan una forma binomial.

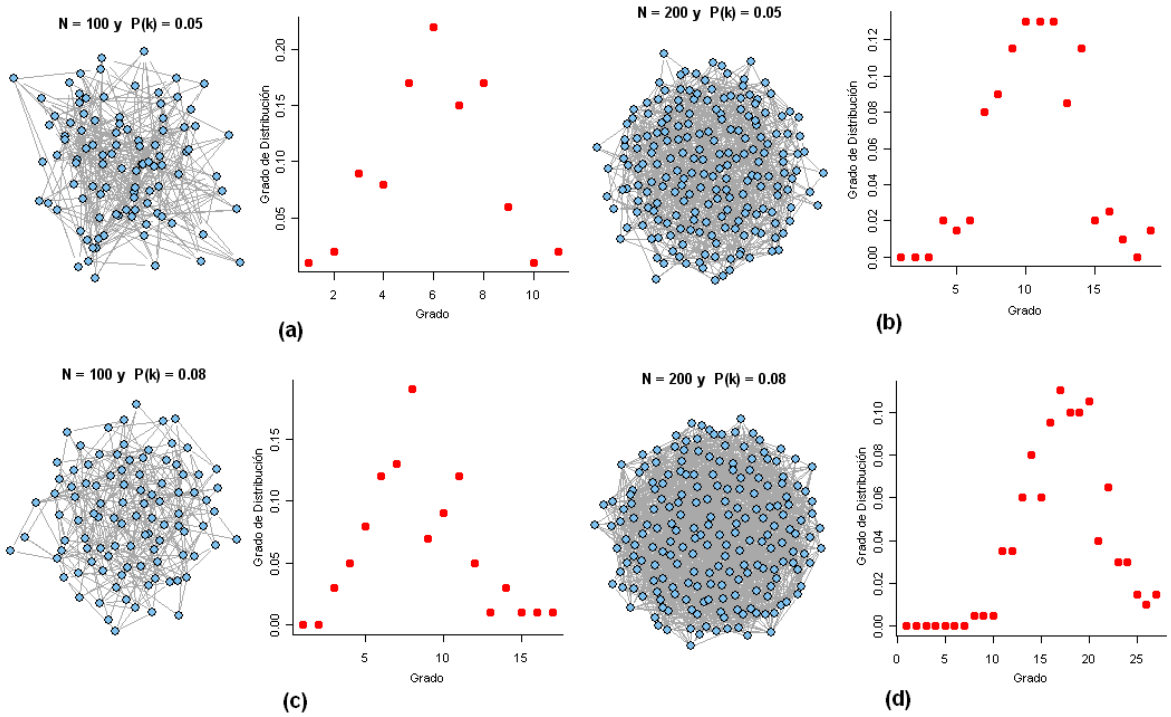


Figura 2.10: Ejemplo de grafos aleatorios generados con el el modelo ER. La gráfica muestra la distribución de grado para cada red con probabilidad y número de nodos diferentes.

2.7.2 Modelo de Redes Mundo Pequeño

Este tipo de modelo parte del fenómeno de mundo pequeño presentado por la mayoría de las redes (ver sección 2.6.2.2). Este tipo de modelos han servido en investigaciones del mundo real como las estudiadas en la propagación epidemiológica [43] y la relación entre el lenguaje humano [30], donde se proponen modelos en base a la estrecha comunicación que existe entre las personas y su manera de relacionarse con los demás.

2.7.2.1 Modelo Watts-Strogatz

El modelo de Watts-Strogatz parte de un proceso de interpolación entre un grafo regular en forma de anillo y un grafo aleatorio. En un grafo con forma de anillo regular con n vértices y k aristas, posteriormente se le aplica un proceso que permite enlazar nuevamente a las aristas

de los vértices con una probabiliad p. Uno de sus principales resultados de este modelo es que para valores intermedios de p bajo $0 < p < 1$, el grafo era una red mundo pequeño que exhibia alto agrupamiento como un grafo regular, pero con longitud de trayectorias cortas, como los grafos aleatorios. En la figura 2.11 se puede apreciar el proceso de interpolación.

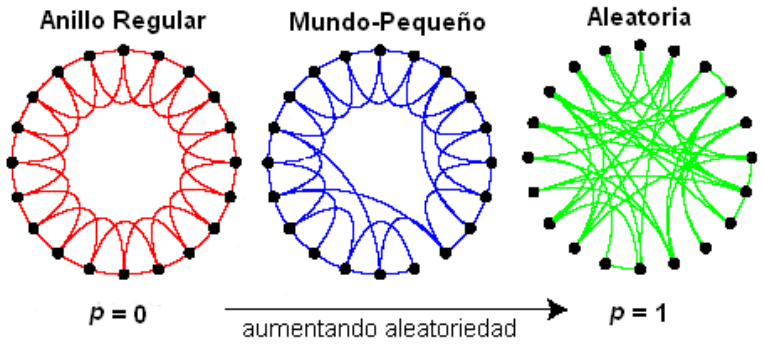


Figura 2.11: Proceso de reescritua aleatorio para la interpolación de un anillo regular a una red aleatoria utilizado en el modelo de Watts-Strogatz

En el modelo de Watts y Strogatz se tiene una longitud L(p) que mide la distancia entre dos nodos en la red y un coeficiente de agrupamiento C(p) que mide los cliques de un vértice con el vecindario. Donde L(p) es pequeño y C(p) es regularmente grande. Para determinar si una red es mundo pequeño basta con calcular la distancia promedio de todos los vertices en la red L_{prom} usando un algoritmo de búsqueda para grafos (p.ej., búsqueda en profundiad o *Depth First Search* en inglés) y determinar el coeficiente de agrupamiento global de la red. En la figura 2.12 se muestra el cálculo para el coeficiente de agrupamiento local de un vértice en la red.

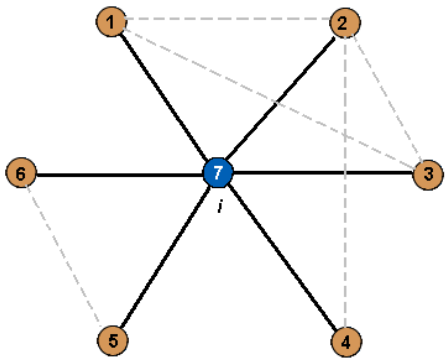


Figura 2.12: Cálculo del coeficiente de agrupamiento local para un vértice. El vertice central i tiene 6 vecinos y en líneas punteadas 5 de los posibles 15 enlaces de los vecinos de i. El coeficiente de agrupamiento para el vertice i es $\frac{5}{15}$.

Watts y Strogatz en [86] proponen la ecuacion 2.16 para calcular el coeficiente de agrupamiento[12] local de un vertice *i*.

$$C_{ws} = \frac{1}{n}\sum_{i=1}^{n} C_i = \frac{1}{n}\sum_{i=1}^{n} \frac{\textit{número de vecinos conectados}}{\frac{1}{2}k_i(k_i-1)} \tag{2.16}$$

donde:

n es el número de nodos en la red, y *k* es el grado del vertice *i*.

Para calcular el coeficiente de agrupamiento global de la red se tiene:

$$C = \frac{\text{3 x (}\textit{número de triangulos en el grafo}\text{)}}{\textit{número de trayectorias de longitud 2}} \tag{2.17}$$

En un principio se pensó que las redes de mundo pequeño generadas por el modelo de Watts-Strongatz tenian una relación casi directa y exclusiva con la distribución de Poisson por la forma en que se planteaba el problema de la interpolación, pero estudios posteriores [21] demostraron que este tipo de redes también se aplican para redes ley de potencia (ver 2.6.2.3) y particularmente en el caso de las redes libres de escala (ver 2.6.2.4).

2.7.3 Modelos de Redes Libres de Escala

En el año de 1999, Faloutsos [29] publica sus resultados trás haber analizado la topología del Internet, donde encontraron que la topología de Internet sigue una distribución ley de potencia (ver 2.6.2.3) y mostraron que las redes ley de potencia presentaban mucha similitud con las redes de mundo real. Este trabajo había sentado las bases para lo que hoy se conoce como *redes libres de escala*.

Barabasí y Albert proponen un modelo para las redes libres de escala en un estudio [10] en el que compara su modelo con los modelos clásicos de redes aleatorias. Se dice que muchas redes que modelan al mundo real se asemejan a las redes de libre escala [5].

2.7.3.1 Modelo de Barabási-Albert

El modelo Barabási-Albert (BA) permite generar redes aleatorias complejas del tipo libres de escala (ver sección 2.6.2.4). El modelo usa las propiedades de *crecimiento* y *anexo preferencial* presentadas en las redes libres de escala, el principio del modelo es el siguiente:

- **Crecimiento.** Dada una red con N_0 de nodos, en determinado tiempo *t* se agrega un nuevo nodo *n* con $j < N_0$ aristas que se unirán a los nodos en la red N_0.

[12] También conocido como *Clustering Coefficient*(CC) por sus siglas en inglés.

- **Anexo preferencial.** El nuevo nodo será agregado a un nodo i con una probabilidad p que dependerá del grado k_i de i, tal que:

$$p(k_i) = \frac{k_i}{\sum_j k_j} \tag{2.18}$$

La distribución de grado para el modelo BA en un tiempo t está dada por:

$$P(k) = \frac{2m^2 t}{k^3} \tag{2.19}$$

Donde m es el número de aristas nuevos, k es el grado del nodo y t es el tiempo en el que un vertice es agregado a la red. En la figura 2.13 se muestran diferentes redes generadas con el modelo BA, así como el grado de distribución del modelo.

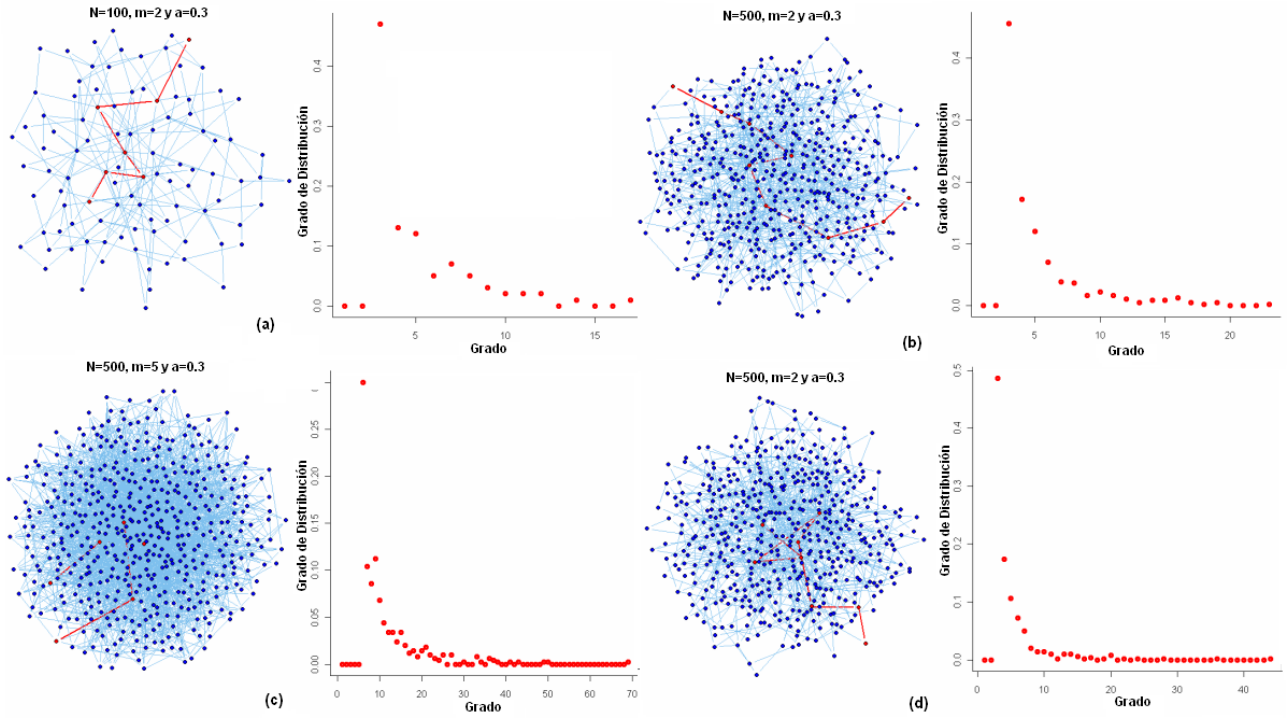


Figura 2.13: Ejemplo de redes libres de escala generadas por el modelo BA con diferentes valores para N, m y α. El camino resaltado representa el diámetro de la red y la gráfica de lado derecho es la distribución de los nodos para cada una de las redes.

2.7.3.2 Comparación de modelos

En la figura 2.14 se muestran los tres tipos de modelos para el análisis de redes sociales, donde se puede apreciar que para la red generada con el modelo ER los nodos están distribuidos de manera homogénea, es decir, existen muchos nodos con grado similar o un grado pequeño. Para la red generada con el modelo BA existen algunos nodos que permiten unir a los demás nodos en la red y para la red de mundo pequeño se puede apreciar un alto índice de agrupamiento y un diámetro pequeño en la red.

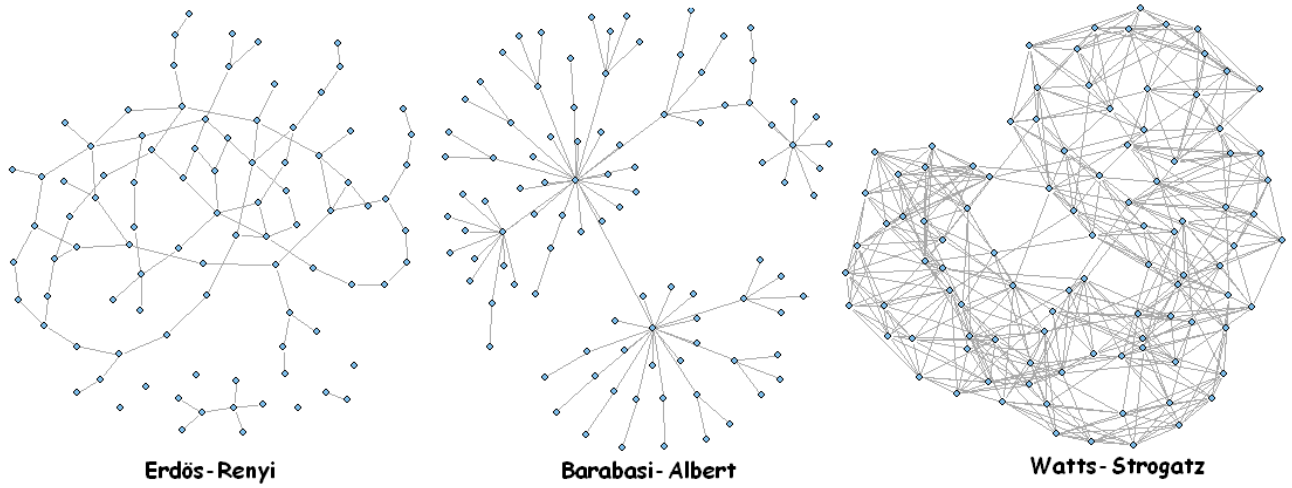


Figura 2.14: Ejemplo de los tres modelos de redes complejas más importantes para el análisis de redes sociales.

La tabla 2.2 muestra la complejidad para el cálculo del diámetro en las tres redes. Así también se presenta el coeficiente de agrupamiento global para cada uno de los modelos.

Tipo de Red	**Modelo**	**Diámetro**	**Coeficiente de Agrupamiento**
Aleatoria	Erdös-Renyi (1960)	$O(logN)$	$CC\alpha N^{-1}$
Libres de Escala	Barabási-Albert (1999)	$O(logN)$ o $O(\frac{\log N}{\log \log N})$	$CC\alpha N^{-0.75}$
Mundo Pequeño	Watts y Strogatz (1998)	$O(N)$ para N pequeño $O(lnN)$ para N larga	$CC(p)\alpha$ $(1-p)^3$
N es el número de nodos y p es la probabilidad de reescritura			

Tabla 2.2: Propiedades de los modelos de redes complejas más comunes.

2.8 Sistemas para el Análisis de Redes Sociales

El análisis de redes sociales posee herramientas de software de diversos tipos, las cuales le permiten realizar mediciones, visualizar los grafos formados y analizar su estructura. Entre los sistemas más comunes para el análisis de redes sociales existen los de exploración de datos, análisis de redes a través del tiempo, análisis estadísticos, entre otros[13].

2.8.1 Visualizadores de redes sociales

La necesidad de agilizar los cálculos y el diseño de algoritmos que permitan visualizar a una red social han motivado a los investigadores [3, 39, 40, 70] enfocar sus trabajos sobre como 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 la tabla 2.3 se muestran los programas más empleados en los trabajos de investigación [38] sobre la visualización y el análisis de redes sociales. Existen diferentes tipos de sistemas para el análisis de redes sociales, algunos incorporan análisis más detallados como el caso de SoNIA que permite visualizar redes dinámicas [24, 13].

Programa	Versión	Objetivo	Formato de Entrada	Capacidad
MultiNet	(4.24)	Análisis contextual	dat	5,000 nodos
NetDraw	(1.0)	Visualización	mat y dat	10,000 nodos
NetMiner	(3.4)	Análisis visual	mat y dat	1,000 nodos
Pajek	(1.24)	Análisis y visualización	mat y dat	10,000 nodos
StOCNET	(1.8)	Análisis estadístico	mat	5,000 nodos
UCINET	(6.05)	Comprensivo	mat y dat	4,000 nodos
GUESS	(1.0.3)	Análisis y visualización	gdf	8,000 nodos
SIENA	(3.1)	Análisis estadístico	dat	3,000 nodos
SoNIA	(1.2)	Análisis Dinámico	mat	4,000 nodos
GNU R	(2.0)	Análisis estadístico	dat,mat	100,000 nodos

Tabla 2.3: Programas más utilizados en el análisis de redes sociales.

2.8.2 Algoritmos para representación gráfica de redes

Dentro de la visualización de redes existen algoritmos para representación gráfica como son: Random, Circle, Sphere, Fruchterman-Reingold, Kamada-Kawai, Spring, entre otros. En la figura 2.15 se muestran diferentes redes generadas con GNU R para el mismo conjunto de datos, pero con diferentes algoritmos de representación gráfica.

[13]Para más información sobre software para redes sociales visitar `http://www.insna.org/software/software_old.html`

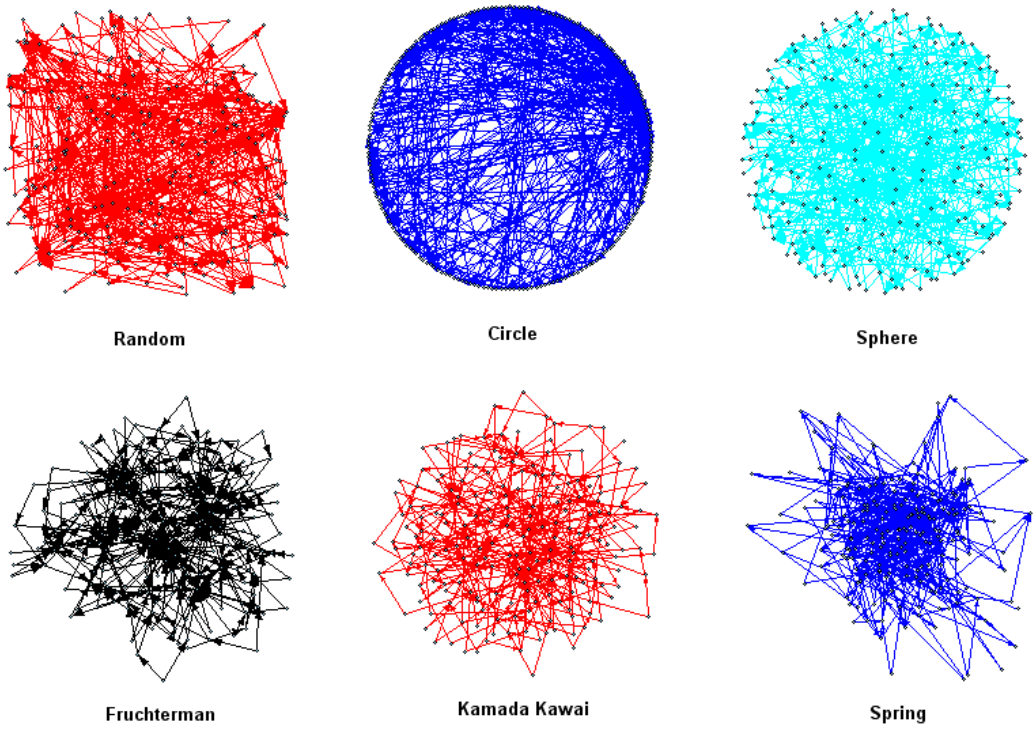


Figura 2.15: Ejemplo de una red con diferentes algoritmos de visualización.

2.8.3 GNU R como sistema para análisis de redes sociales

Dentro de la presente tesis se utilizó *R* como lenguaje de programación para implementar los diferentes algoritmos para el análisis de redes sociales (p.ej., cálculos de centralidad, visualización de grafo, detección de comunidades, generación de redes complejas, etc) mediante el uso de la libreria *igraph*. GNU R es un programa para computación estadística y gráfica. R se distribuye bajo licencia GNU GPL[14] y está disponible para diferentes sistemas operativos.

R es un producto derivado del lenguaje S que fue desarrollado por los Laboratorios Bell en 1993. R proporciona rapidez a la hora de realizar cálculos matriciales y vectoriales. Este sistema es ampliamente utilizado en el campo de la investigación biomédica, bioinformática, y las matemáticas financieras. R está diseñado para interactuar con diversos lenguajes de programación como son C, Python, Java, entre otros. También permite realizar consultas a base de datos (p.ej., MySQL, SQLServer, entre otros). El paquete *igraph* proporciona rutinas para el análisis de redes y grafos simples.

[14]Licencia Pública General de GNU o GNU GPL (General Public License) por sus siglas en inglés, es una licencia orientada a proteger la libre distribución, modificación y uso del software.

Capítulo 3

Muestreo de los sistemas de redes sociales en línea

La gran cantidad de información que está contenida en los sistemas de redes sociales en línea permite a los investigadores realizar diferentes estudios sobre la estructura y comportamiento de estos sistemas en diversas áreas del conocimiento. Sin embargo, la forma en como se encuentra distribuida esta información en ocasiones no resulta ser tan fácil de manipular, y se necesitan mecanismos complejos y de un conocimiento del funcionamiento del sistema para poder obtener la información necesaria para llevar a cabo dicho estudio.

En este capítulo presentamos los diferentes tipos de muestreo aplicados a la extración de información en los sistemas de redes sociales en línea. Se propone un algoritmo para la extracción de información, basado en una de las técnicas de muestreo utilizada por la mayoría de los trabajos de investigación de redes sociales a gran escala, este algoritmo está basado en la relación entre contactos de un usuario y el algoritmo propuesto está basado en la amistad de un usuario y sus contactos. También describimos la estructura de la Wikipedia y en base a esto adaptamos el algoritmo basado en la amistad, de tal manera que nos permita generar un grafo de colaboración de artículos de la Wikipedia para su análisis.

3.1 Extracción 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 [53, 54, 55] sobre como obtener dicha información.

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 [54] y POLYPHONET [53] que explotan a los motores de búsqueda para obtener in-

formación 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 [49] proponen metodologías de extracción para el contenido oculto en la Web [51], 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 contendio 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 [8] enfocadas a proteger dicha información sobre ataques de terceros [7].

La forma en como se puede representar una red social es generalmente mediante un grafo dirigido (o grafo bipartido), 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, envio de mensajes, entre otros).

La forma en como 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 el enfoque de estudio. Sin embargo, el proceso de extracción tiene sus inconvenientes como son la perdida de información al momento de representar la red, según estudios de investigación [41] muestran que la perdida de información durante el muestreo puede afectar los resultados de las mediciones.

3.1.1 Tipos de muestreos aplicados a las Redes Sociales

En [45] 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 figura 3.1(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 figura 3.1(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 figura 3.1(c) se muestra el proceso para el muestreo de enlace.

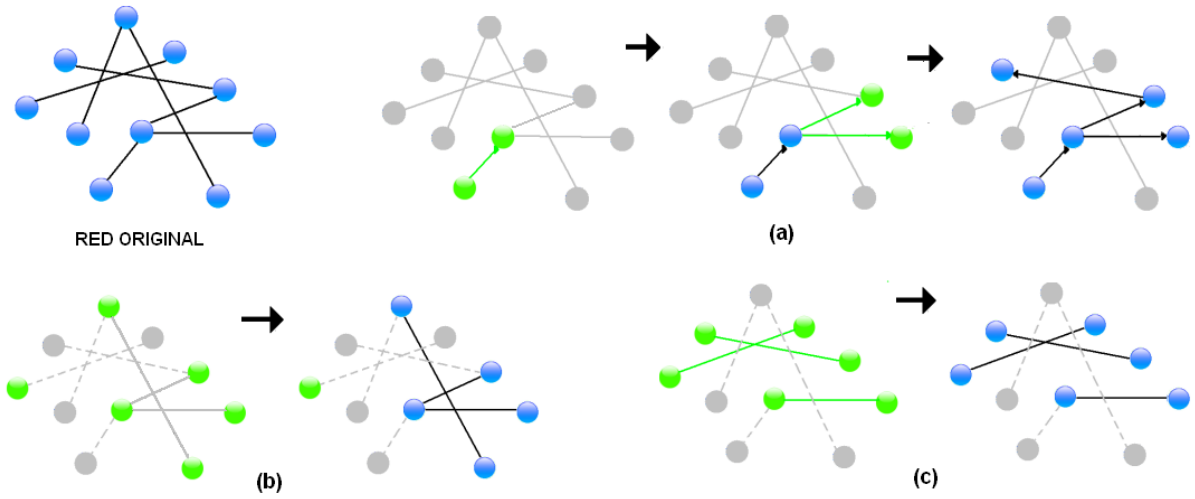


Figura 3.1: Tipos de muestreos para una red social. (a) Muestro bola de nieve, (b) Muestreo de nodo y (c) Muestreo de enlace.

Para el estudio de la tesis se estudiaron a dos redes sociales Flickr y WIkipedia, para lo cual se desarrolló una metodología para el muestreo muy diferente para cada caso, en la sección 3.2 y 3.3 se detalla el proceso de adquisición de datos para cada uno de los sistemas.

3.1.2 Sistemas para extracción de redes sociales

Existen diferentes técnicas para obtener información de la Web, algunos trabajos [54, 53] 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 describiremos tres formas, dos son sistemas (Flink y POLYPHONET) que explotan los resultados de las consultas generadas por los motores de búsquedas particularmente *Google*, y el tercero es el uso de un *API* de desarrollo de los sistemas de redes sociales en línea.

3.1.2.1 Flink sistema para extraer redes sociales

Flink [54] 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 incluyen páginas Web, correos electrónicos, publicaciones y archivos FOAF[1]. En la figura 3.2 se muestra la arquitectura de Flink desde la adquisición de metadatos (arriba) hasta la interfaz de usuario (abajo).

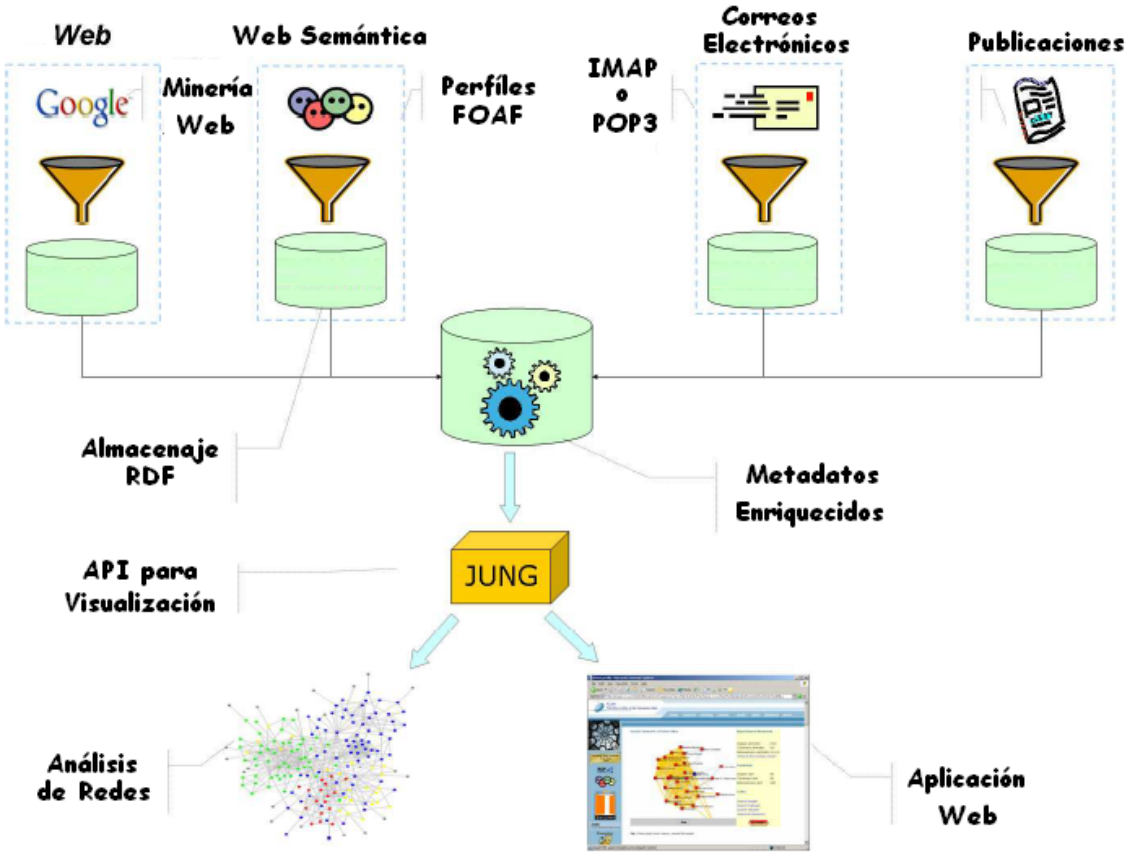


Figura 3.2: Arquitectura de Flink. Adquisición de datos (arriba) e Interfaz de usuario (abajo).

Flink utiliza diferentes fuentes de datos para obtener información, después la integra en un base de datos enriquecida y utiliza JUNG [2] 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.

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

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

3.1.2.2 POLYPHONET sistema para extraer redes sociales

En [53] 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*. El algoritmo es sencillo y se muestra a continuación:

```
Entrada: Dado una lista de personas L, regresar la red social G
Inicio:
for X ∈ L do
| G ⇐ AgregarNodo();
end
for X ∈ L and Y ∈ L do
| // Regresar la ocurrencia entre X y Y r_{X,Y} ⇐ GoogleOcurrencia(X, Y)
end
for X ∈ L and Y ∈ L do
| if r_{X,Y} > umbral then
| | G ⇐ AgregarArista();
| end
end
```

Algoritmo 1: Algoritmo de extracción de redes sociales utilizando POLYPHONET.

Se puede apreciar que el algoritmo para extraer redes sociales utilizado en POLYPHONET solo 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.

3.1.2.3 OpenSocial y las Interfaces de Programación de Aplicaciones

Sitios como MySpace, Orkut y hi5 forman parte de un servicio propuesto por Google llamado *OpenSocial*[3], 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 [4]. 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

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

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

en lenguajes como son JavaScript y HTML y proporciona un soporte de versiones de las diferentes API's. En las siguientes secciones utilizamos este método para extraer información de los sistemas Flickr y Wikipedia en base a las técnicas de muestreo mostradas en la sección 3.1.

3.2 Muestreo de la red de Flickr

Flickr es un sistema de red social en línea basado en la compartición de contenido, específicamente en fotos o imágenes y recientemente en videos. Este sistema empezó a funcionar en 2004 por Ludicop y sus origenes estan en una compañia canadiense de nombre *Game Neverending*. Actualmente Flickr le pertenece a Yahoo quien lo adquirió en marzo del 2005, cuenta con más de 32 millones de usuarios y más de 4 mil millones de fotos para compartir[5]. Sus principales caracterízticas son la búsqueda y etiquetado de fotos, y el uso de licencias *Creative Commons*[6].

En la figura 3.3 se muestra la arquitectura del sistema Flickr[7], donde se pueden ver los diferentes tipos de aplicaciones con las que Flickr trabaja, las desarrolladas por el sistema mismo y las desarrolladas por usuarios externos.

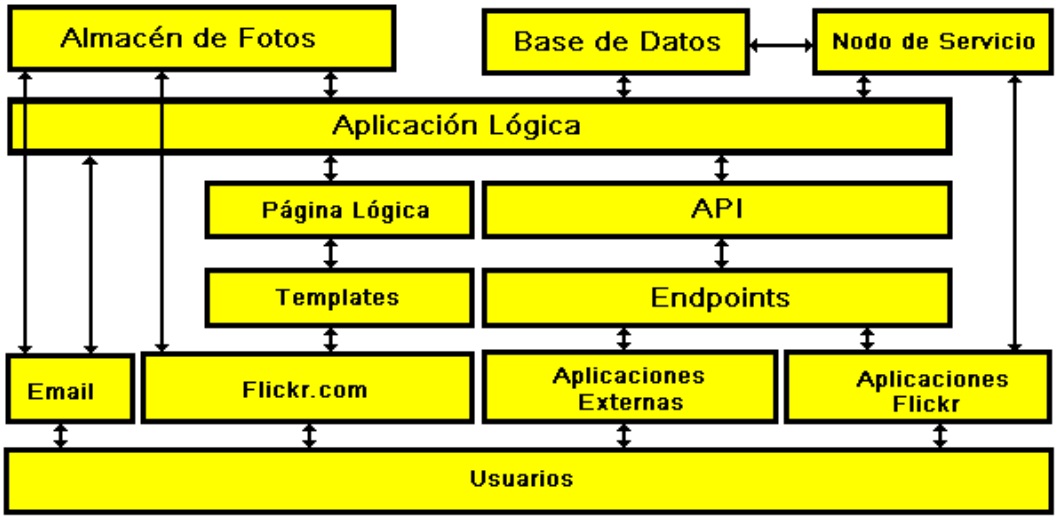


Figura 3.3: Arquitectura del sistema Flickr.

[5]Datos obtenidos de `http://blog.flickr.net/en/2009/10/12/4000000000/`

[6]Flickr maneja diferentes niveles de *creative commons*, los cuales permite compartir fotos entre usuarios.

[7]Esta información fue obtenida de una presentación del director de Flickr *Cal Henderson*, y se puede encontrar una copia disponible en `http://www.niallkennedy.com/blog/uploads/flickr_php.pdf`

3.2.1 El API de Flickr

Flickr proporciona una API de desarrollo para la creación de aplicaciones, las cuales pueden ser integradas al propio sistema o hacer uso de la información del sistema de manera no comercial. Existen distintas API's que se proporcionan para diferentes lenguajes (p.ej., C++, Java, PHP, .NET, Python, entre otros). Trabajos recientes han estudiado a Flickr [59, 18] como una red social, aunque no es considerada como una red social pura, es decir, una red que basa sus relaciones exlusivamente en el conocimiento de otras personas y no en la compartición de contenido y aplicaciones.

Flickr dispone de dos modos para trabajar con su API, ambas orientadas a conexión. La primera está basada en Web, esta permite manipular los datos de Flickr desde una URL externa a Flickr, pero la cual será redireccionada por Flickr. El segundo tipo de aplicaciones son las de escritorio y no necesitan estar en un servidor Web. Sin embargo, las dos maneras deben de pasar por un proceso de autentificación para poder acceder a los datos. La autentificación consiste en registrar la aplicación a desarrollar para obtener dos variables (*clave* y *secreto*) que son asociadas al proyecto y sirven para utilizar el API y llevar una estadística de uso por parte de Flickr.

Los formatos para solicitar información son REST, XML-RPC y SOAP, para la respuesta son REST, XML-RPC, SOAP, JSON y PHP, la figura 3.4 muestra el proceso de una petición utilizando el API de Flickr.

Figura 3.4: Proceso de petición de consulta en Flickr.

3.2.2 Algoritmo de extracción basado en lista de contactos

La forma como se realiza el muestreo dependerá de la forma en como los sistemas proporcionen acceo a la información[8], por tal motivo los algoritmos que permitan generalizar el muestreo en todos los sistemas de redes sociales en línea es algo que aún no es posible debido a su diferencia en su estructura. Flickr dentro de su API de desarrollo proporciona el método **contacts.getPublicList(*id de usuario*)**, el cual recibe como parámetro el identificador del usuario y regresa la lista de contactos asociada a dicho usuario. Esta lista de contactos contiene los identificadores de cada uno de los contactos que el usuario tiene agregado a su lista de contactos.

Un usuario dentro de Flickr posee un identificador de la forma [*Número de 9 dígitos*] @N0 [*Número de 1 dígito*], por ejemplo, 523058314@N09. Partiendo de este hecho la primera etapa del algoritmo empieza por generar n usuarios aleatorios de una posibilidad de 10 mil millones. Sin embargo, Flickr no cuenta con este número de usuarios, para lo que un usuario generado de esta manera puede que no exista en la base de datos de Flickr obligando a generar nuevos usuarios que sean válidos dentro del sistema.

El segundo paso del algoritmo es aplicar un muestreo de nieve (para más información sobre los tipos de muestreos ver la sección 3.1) en donde la primera semilla es el usuario generado aleatoriamente y se empieza buscando su lista de contactos. Posteriormente se busca la lista de contactos de manera recursiva para cada uno de los contactos en la lista inicial. En la figura 3.5 se muestra la consulta de una lista de contactos de un usuario.

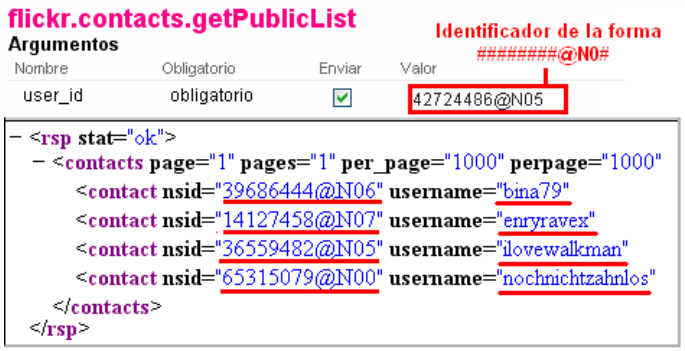


Figura 3.5: Ejemplo de consulta para obtener lista de contactos de un usuario.

El algoritmo para generar la red de contactos de Flickr está inspirado en trabajos de investigación anteriores [60, 18], donde el resultado es un grafo dirigido que representa la interacción de los usuarios de Flickr, el siguiente algoritmo explica el procedimiento para la generación de un grafo a partir de los datos de Flickr.

[8]La información puede ser pública (p.ej., Flickr, MySpace, Wikipedia, etc) o privada (p.ej., Orkut, LinkedIn, Facebook, etc)

```
Entrada: n > 0 usuarios, k > 0 profundidad
Salida: G=(V, E)
Inicio:
i = 0;
m[2...n]
x ⇐ generarUsuario();
while i < n do
    if usuarioValido(x) and noExisteContacto(V_G, x) then
        guardarContacto(x);
        n ⇐ n − 1;
        L[n] ⇐ listaContactos(x);
        m ⇐ numeroContactos(L);
        j = 0
        while m > 0 ∨ j < k do
            y ⇐ obtenerContacto(L[j]);
            if noExisteContacto(V_G, y) then
                V_G ⇐ guardarContacto(y);
            end
            if noExisteRelacion(V_G, x, y) then
                E_G ⇐ guardarRelacion(x, y);
            end
            j ⇐ j + 1;
            m ⇐ m − 1;
        end
        i ⇐ i − 1;
    else
        x ⇐ GenerarUsuario();
    end
end
```

Algoritmo 2: Generación de redes sociales para el sistema Flickr basado en lista de contactos.

El algoritmo da como resultado un grafo dirigido representado por una lista de vértices (usuarios) y una lista de aristas (vínculos) que son las relaciones entre usuarios. La búsqueda finaliza cuando la lista de contactos de un usuario es cero o el nivel de profundiad k [9] cumple $k > \alpha$. El proceso se repite para los n usuarios generados aleatoriamente.

3.2.3 Algoritmo de extracción basado en relaciones por contenido

Una propiedad que tienen los sistemas basados en contenido (p.ej., Flickr, Youtube, Fotolog, etc) es que permiten agregar comentarios a los archivos multimedia que son compartidos, esto permite medir la frecuencia de comunicación entre usuarios y proporcionan información suficiente para poder determinar que tan fuerte puede ser un vínculo entre dos usuarios.

El conjunto de datos obtenidos en el muestreo de redes basadas en las relaciones entre usuarios son utilizadas en la mayoría de las investigaciones, especialmente en Flickr[60, 59,

[9] Entiendase como nivel de profundiad a el número de recursividades realizadas desde el primer elemento hasta cierto nivel.

18]. Sin embargo, el tipo de relaciones que se presentan en Flickr va más allá de solo agregar contactos. Una de las principales funciones de Flickr consiste en compartir fotografías, es por eso que dentro de esta tesis se desarrolló un algoritmo que permita generar un grafo en base a la interacción de contenido entre usuarios.

El API de Flickr nos proporciona dos métodos muy importantes para obtener información sobre las fotos publicadas por los usuarios y los comentarios entre ellos, estas son: $getPublicPhotos(idUser)$ dado un identificador de usuario regresa la lista de identificadores de las fotos asociadadas al usuario dado, y la función $comments.getList(idPhoto)$ que recibe como paramétro un identificador de foto y regresa la lista de usuarios que comentaron la foto. En la figura 3.6 se muestra un ejemplo para consultar los comentarios de una foto de un usuario en Flickr.

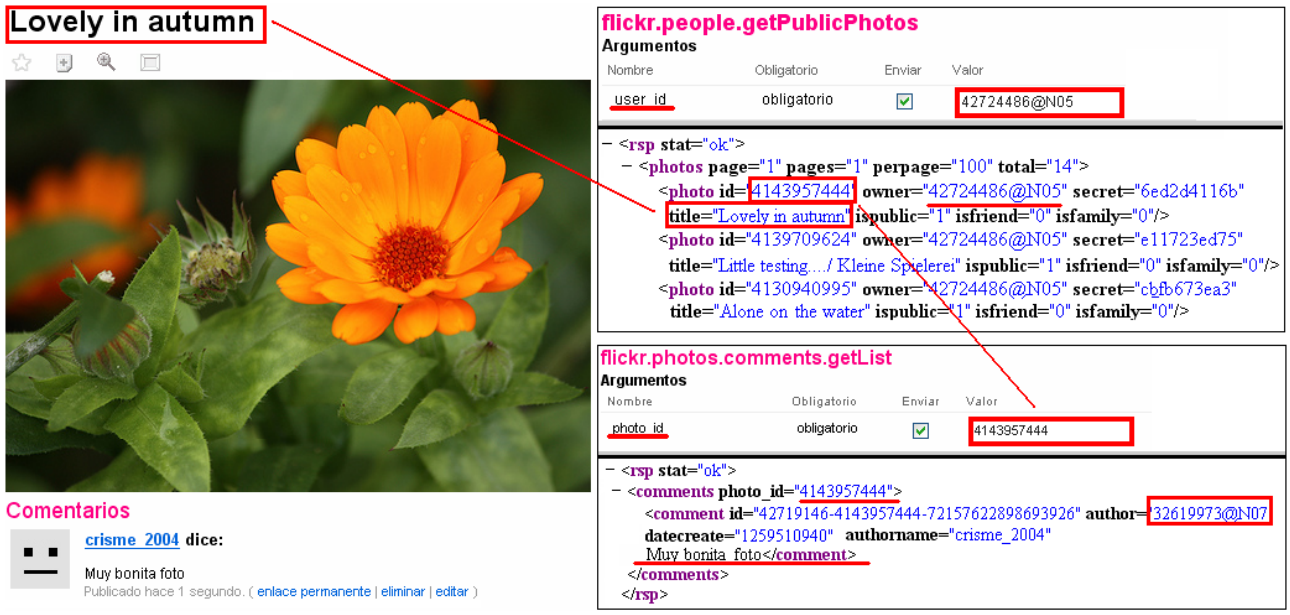


Figura 3.6: Ejemplo de consulta para obtener lista de comentarios de las fotos de un contacto.

El siguiente algoritmo calcula el peso de un vínculo entre dos usuarios con la función *nivelRelacion(x,y)*, donde se calcula el promedio de mensajes entre usuarios y se compara con un valor α que está dado por el promedio de mensajes entre dos usuarios en toda la red.

```
Entrada:  n > 0, k > 0, α > 0
Salida:   G=(V, E)
Inicio:
i = 0;
m[2...n]
x ⇐ generarUsuario();
while i < n do
    if usuarioValido(x) and noExisteContacto(V_G, y) then
        guardarContacto(x);
        n ⇐ n − 1;
        L[n] ⇐ listaContactos(x);
        m ⇐ numeroContactos(L);
        j = 0
        while m > 0 ∨ j < k do
            y ⇐ obtenerContacto(L[j]);
            if noExisteContacto(V_G, y) then
                V_G ⇐ guardarContacto(y);
            end
            if noExisteRelacion(V_G, x, y) and nivelRelacion(x, y) > α then
                E_G ⇐ guardarRelacion(x, y);
            end
            j ⇐ j + 1;
            m ⇐ m − 1;
        end
        i ⇐ i − 1;
    else
        x ⇐ GenerarUsuario();
    end
end
```

Algoritmo 3: Generación de una red social de Flickr con un indice de amistad.

El valor de α representa el índice de amistad entre dos contactos. Existen modelos de asignación de pesos a vínculos como es el caso de *PageRank* [44], el cual propone el uso del modelo de Markov para calcular la probabiliad de cada enlace.

Una primera aproximación para el cálculo del nivel de amistad está dado por:

$$\text{Nivel de amistad} = \frac{\sum_{m(i,j)} + \sum_{m(j,i)}}{\sum \text{Total de mensajes}} \tag{3.1}$$

Donde:

$m(i,j)$ y $m(j,i)$ son los mensajes del nodo *i* al nodo *j* y del nodo *j* al nodo *i* respectivamente.

Una ecuación como la anterior sería aplicada en el caso de que el grado de simetría de la red fuera alto (en el caso de redes no dirigidas), es decir, para una red dirigída no es posible asegurar que un contacto tenga una relación en sentido opuesto o recíproco, además cabe la

posibilidad de que mediante el muestreo se pierdan relaciones. Para nuestra tesis usamos el cálculo del nivel de amistad como:

$$\text{Nivel de amistad} = \frac{\sum_{m(i,j)}}{\sum \text{Total de mensajes de i}} \tag{3.2}$$

Muchos de los servicios de redes sociales en línea permiten a los usuarios interactuar mediante una gran cantidad de aplicaciones (p.ej., blogs, videoblogs, fotoblogs, wikis, etc), esto permite medir la frecuencia en la que un grupo de personas se está comunicando.

3.3 Muestreo de la red de Wikipedia

Wikipedia es una enciclopedia interactiva iniciada en 2001, basa su funcionamiento mediante el autocontenido generado por parte de usuario, es una especie de blog (Wiki) que permite editar contenido de diversos temas. Cuenta con una disponibilidad de más de 13.7 millones de artículos y ocupa el sexto lugar en popularidad dentro de la Web.

En el año de 2004 se crea Wikimedia, un repositorio en el cual se centralizan los diferentes proyectos de la *Fundación Wikimedia*, los cuales son: **Wikipedia**, Wikinoticias, Wikcionario, Wikibooks, Wikiquote, Wikisource, Wikicommons, Wikispecies y Wikiversidad, todas funcionan bajo *MediaWiki* un motor para Wikis[10] que se encuentra bajo licencia GNU.

3.3.1 Estructura de la Wikipedia

La Wikipedia contiene información que puede ser editada por cualquier persona, dentro del sistema existen dos tipos de usuarios, los usuarios registrados que poseen contraseña y un nombre dentro del sistema y los que no. Para el caso de los usuarios que editan contenido y no estan registrados, Wikipedia almacena la dirección IP de la máquina donde se realizó la edición del contenido, esto se hace para llevar cierto control dentro de un historial que maneja las ediciones realizadas a las páginas.

La cantidad de información que contiene Wikipedia y la variedad de idiomas que maneja, permite realizar estudios sobre algún tipo específico de contenido, debido a que Wikipedia permite descargar la Base de Datos[11]. En la figura 3.7 se muestra el proceso de extracción de información de la Wikipedia a una Base de Datos en MySQL.

[10] Entiendase por Wiki a una especie de blog, donde el contenido puede ser editado por diferentes personas que tengan acceso a una publicación.

[11] En `ftp://ftp.rediris.es/mirror/WKP_research/` se pueden descargar diferentes Wikipedias.

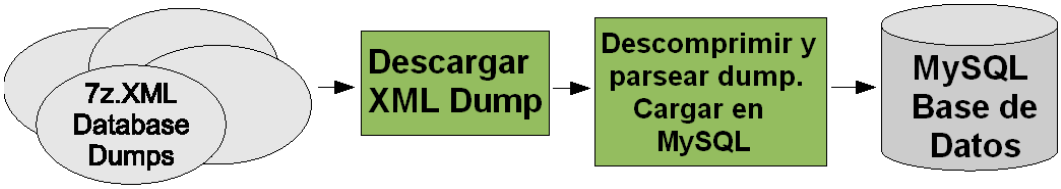


Figura 3.7: Proceso de extracción de datos de la Wikipedia.

La arquitectura general[12] de la Wikipedia se presenta en la figura 3.8.

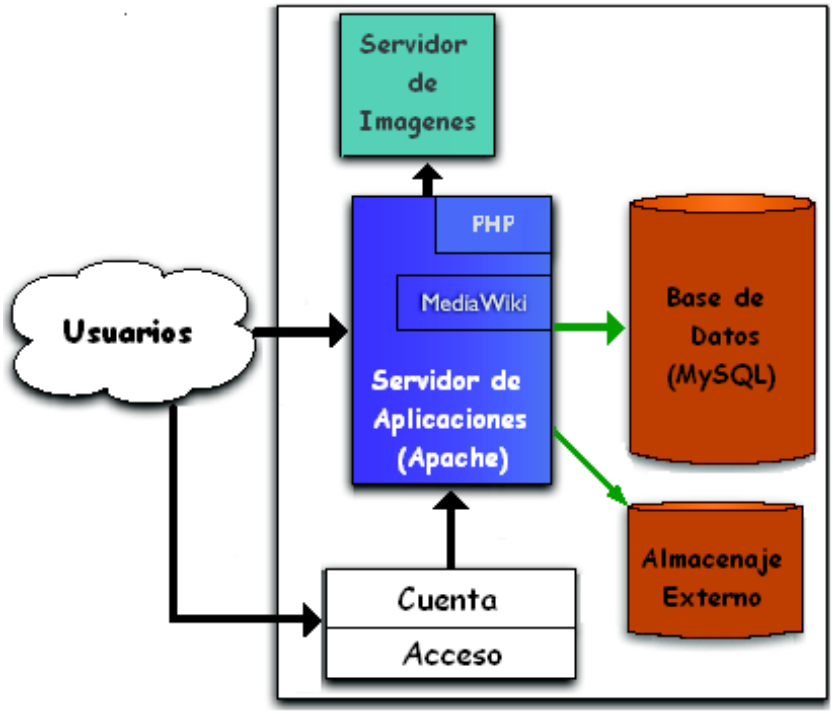


Figura 3.8: Arquitectura general de Wikipedia.

En la figura 3.9 se muestra el diagrama Entidad-Relación de la parte principal de la Base de Datos de Wikipedia [13] utilizada para almacenar la información extraída del sistema de Wikipedia, se puede ver que el núcleo importante de la Wikipedia son las páginas.

[12]Esta información fue obtenida de una presentación de Mark Bergsma de *Wikimedia Foundation Inc.*, una copia de la presentación se puede encontrar en `http://www.nedworks.org/~mark/presentations/san/Wikimediaarchitecture.pdf`

[13]Para ver la Base de Datos completa visitar `http://upload.wikimedia.org/wikipedia/commons/4/41/Mediawiki-database-schema.png`

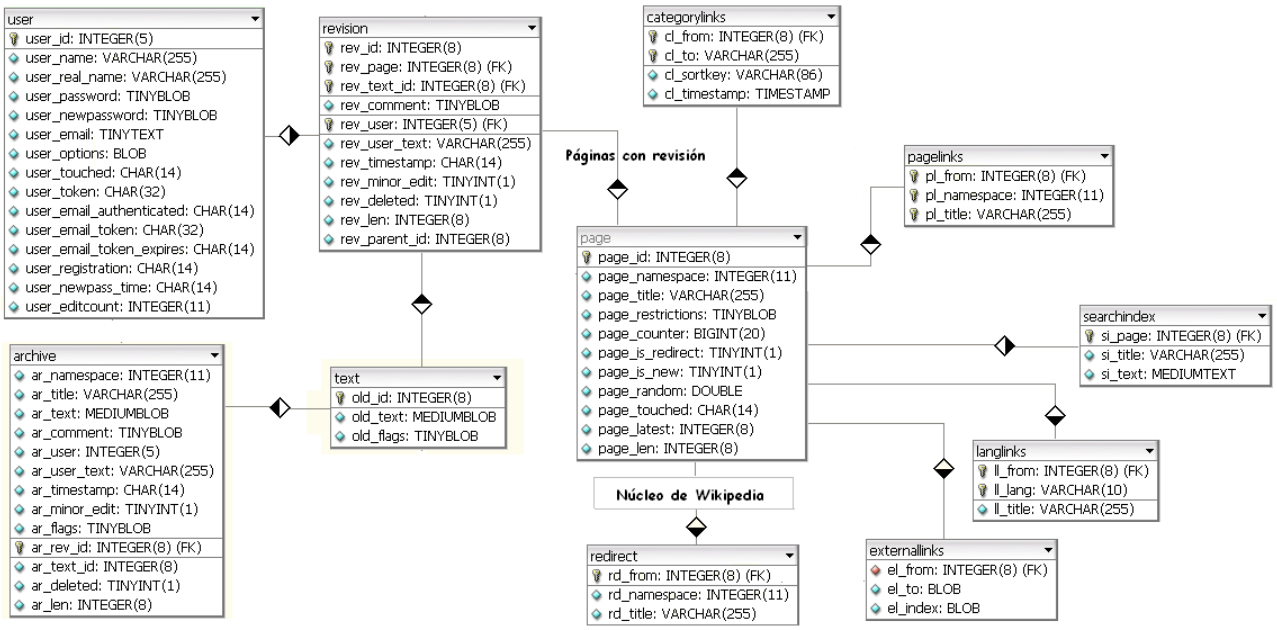


Figura 3.9: Diagrama ER de la Base de Datos de Wikipedia.

3.3.2 Algoritmo para obtener el grafo de Wikipedia

Para realizar un estudio sobre una red social es necesario preparar los datos para que el grafo resultante sea una buena representación de la red social. Para el caso de Wikipedia se propuso tomar a los usuarios como actores de la red y a las páginas como vínculos entre ellos, usando la tabla revisión podemos encontrar la relación de las páginas que han sido editadas por un usuario.

La Wikipedia posee una estructura que puede ser analizadas fuera de línea, pero tiene la particularidad de ofrecer un campo *timestamp* que marca la fecha en que una página es editada y esto permite analizar a la red como una red dinámica. En la figura 3.10 se muestran las tablas utilizadas para almacenar el resultado del algoritmo para obtener la representación de la red social de Wikipedia, se puede apreciar la interacción que existe entre usuario y página mediante el uso de las revisiones.

Para obtener la red social de Wikipedia se propuso una parámetro β que permita eliminar a las relaciones más débiles del sistema, es decir, las relaciones entre usuarios que no interactúan con frecuencia. A continuación presentamos el algoritmo para generar la red social de Wikipedia.

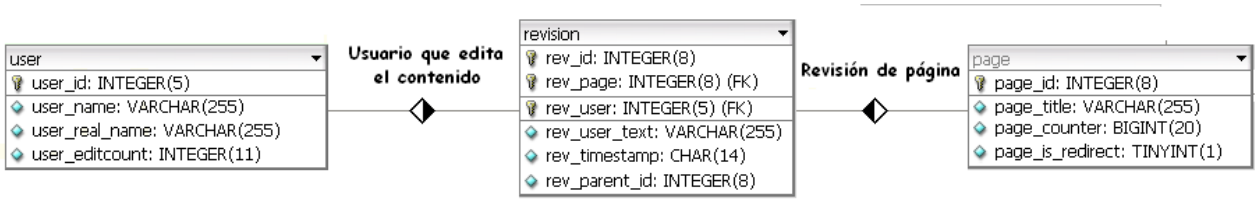


Figura 3.10: Interacción entre un usuario y una página en Wikipedia.

Entrada: $L_{usuario}[1,...,n], L_{pagina}[1,...,m]$
Salida: G=(V, E)
Inicio:
$i = 0;$
while $i < n$ **do**
  // Obtener lista de paginas asociadas a un contacto
  $x \Leftarrow obtenerContacto(L_{usuario}[i]);$
  $guardarContacto(x);$
  $L_{pagUsu} \Leftarrow listaPaginas(x);$
  $j = 0;$
  while $j < m$ **do**
    // Obtener lista de contactos para las páginas de un usuario
    $L_{contacto} \Leftarrow obtenerContactos(L_{pagUsu}[j]);$
    $k = 0;$
    while $k < r$ **do**
      $y \Leftarrow obtenerContacto(L_{contacto}[k]);$
      if $noExisteContacto(V_G, y)$ **then**
        $V_G \Leftarrow guardarContacto(y);$
      end
      if $noExisteRelacion(V_G, x, y)$ **and** $nivelRelacion > \beta$ **then**
        $E_G \Leftarrow guardarRelacion(x, y);$
      end
      $k \Leftarrow k + 1;$
    end
    $j \Leftarrow j + 1;$
  end
  $i \Leftarrow j + 1;$
end

Algoritmo 4: Generación de una red social de colaboración de artículos de Wikipedia.

Este algoritmo utiliza consultas a la Base de Datos de Wikipedia de las tablas mostradas en la figura 3.10, donde las búsquedas para la obtención de listas se realizan sobre la tabla *revision*, la cual posee las relaciones entre páginas y usuarios. El resultado es una *red de colaboración*, para este caso sobre artículos de una red a gran escala. Para el caso de Wikipedia se genera una red no dirigída a diferencia de la red dirigida que se genera para Flickr.

3.4 Discusión

La forma en como los sistemas de redes sociales en línea pueden ser representados para su análisis, depende de un buen muestreo de dichos sistemas. La diferencia entre un buen análisis dependerá de tener una buena representación de la información. Sin embargo, no es muy común realizar un estudio de la estructura completa de las redes sociales en línea, más cuando estas superan los millones de nodos y de relaciones.

Un tema que causa mucha polémica es el que respecta a la seguridad de la información de los sistemas de redes sociales en línea; donde mucha de la información contenida en este tipo de sistemas puede ser fácilmente utilizada para cualquier fin, dejando vulnerable a los usuarios que utilizan estos sistemas para comparitr fotografías, videos, datos personales, entre muchas otras cosas. Sin embargo, muchos son los trabajos que utilizan este tipo de información para representar estadísticas sobre los comportamientos sociales, analizar la estructura de las redes formadas y la evolución de los sistemas, pero pocos trabajos se enfocan en la seguridad de la información de estos sistemas.

Capítulo 4

Detección de Comunidades en Redes Sociales

Una de las principales características de los grafos que representan sistemas reales son las comunidades o aprupamientos. Tales agrupaciones o comunidades pueden considerarse como elementos independientes de un grafo, y formalmente como un subcojnuto del conjunto inicial. Los métodos para el análisis de redes suelen ser muy costos y más cuando se tienen redes a gran escala, el uso de métodos para detectar comunidades permite reducir en diferentes fracciones a la red original, siempre y cuando se cumplan ciertos criterios sobre como generar las particiones.

En este capítulo se presentan los diferentes conceptos sobre comunidades en redes sociales y las técnicas del agrupamiento en grafos como la herramienta matemática que permite detectar estas estructuras dentro de las redes sociales. Finalmente, se propone un algoritmo para la detección de comunidades en base a la estructura de las redes sociales a gran escala.

4.1 Comunidades en Redes Sociales

Una comundiad es un grupo de vértices y aristas que comparten ciertas propiedades en común e influyen de manera similar con la red. La sociedad ofrece una infinidad de posibles comunidades: familiares, círculos de amistad y de trabajo, ciudades, naciones, entre otras. El problema de detección de comunidades es muy importante para tratar diferentes sistemas reales como son: biológicos, computacionales, económicos, políticos, sociales, etc. La formalización matemática al problema de la detección de comunidades se le conoce como *agrupamiento de grafos*, el cual tiene sus inicios desde la decada de los 70 y que detallaremos en las siguientes secciones.

4.2 Mediciones para identificar agrupaciones

En esta sección describimos las dos formas para identicar un buen agrupamiento. La primera, analiza ciertos valores para los vértices del grafo y después clasifica los vértices en agrupaciones basado en el valor obtenido. La segunda forma realiza una medición de aptitud sobre el conjunto de datos de las posibles agrupaciones y después selecciona dentro del conjunto de datos a los candidatos que optimizan la medición a usar.

4.2.1 Similaridad de vértices

Este tipo de análisis se basa en la *similaridad*[1] de los vértices, donde se busca agrupar a los vértices que no estan bien conectados pero que son similares entre ellos. El cálculo de la similaridad no es simple, realizar el cálculo de la similaridad de dos vértices puede ser una tarea más compleja que el mismo agrupamiento del grafo (en algunos casos). A continuación se proponen diferentes mediciones basadas en la *similaridad* de los vértices.

4.2.1.1 Distancia y mediciones de similaridad

Este tipo de medición está definido como: dado un conjunto de datos, una medición de distancia $dist(d_i, d_j)$ debe satisfacer el siguiente criterio;

- La distancia de de un vértice hacia si mísmo es cero: $dist(d_i d_i) = 0$
- Las distancias son *simétricas*: $dist(d_i d_j) = dist(d_j d_i)$
- El *triángulo de desigualdad* está dado por:

$$dist(d_i d_j) \leq dist(d_i d_k) + dist(d_k d_j) \tag{4.1}$$

Para datos en un *n-dimensiones* **Espacio Euclideano**, se representa a la medición de distancia para dos datos $d_i = (d_{i,1}, d_{i,2}, ..., d_{i,n})$ y $d_j = (d_{j,1}, d_{j,2}, ..., d_{j,n})$ incluyendo una *distancia euclideana*

$$dist(d_i d_j) = \sum_{k=1}^{n} \sqrt{(d_{i,k} - d_{j,k})^2} \tag{4.2}$$

Muchas mediciones de similaridad son basadas en el *índice de Jaccard* definido por un conjunto A y un conjunto B tal que:

$$p(A, B) = \frac{|A| \cap |B|}{|A| \cup |B|} \tag{4.3}$$

[1]Entiendase como similaridad a una medida que compara las diferentes estructuras formadas en un grafo.

Esta es fácilmente transformada en una distancia de medición: $dist(A, B) = 1 - p(A, B)$. Esta idea generaliza a un vector binario *n*-dimensional $A = (a_i, a_2, ..., a_n)$ y $B = (b_i, b_2, ..., b_n)$. El *coeficiente de similaridad de Jaccard* para un vector A y B es:

$$p(A, B) = \frac{C_{1,1}}{C_{0,1} + C_{1,0} + C_{1,1}} \tag{4.4}$$

La *distancia de Jaccard* está dada por:

$$dist(A, B) = \frac{C_{1,0} + C_{0,1}}{C_{0,1} + C_{1,0} + C_{1,1}} \tag{4.5}$$

4.2.1.2 Medición basada en la adyacencia

Cuando las aristas inciden en los vértices pueden ser usados para derivar una medición de similaridad para los vértices usando la información de adyacencia directamente. La manera más directa para determinar si dos vértices son similiares dependerá solo de la información de adyacencia que servirá para estudiar el traslapamiento de su vecindario en $G = (V, E)$. Una manera directa para calcular la intersección y la unión de dos conjuntos es:

$$\omega(v, w)\frac{\Gamma(v) \cap \Gamma(w)}{\Gamma(v) \cup \Gamma(w)} \tag{4.6}$$

La medición toma valores $[0, 1]$, cero cuando no hay vecinos en común y uno cuando el vecindario es idéntico.

4.2.1.3 Medición de conectividad

El agrupamiento en grafos puede ser obtenido por la conectividad, es decir, por calcular el número de caminos que existen entre cada par de vértices. Para un buen agrupamiento no es necesario que dos vértices *v* y *u* sean conectados por una arista directa. Un modelo basado en este tipo de mediciones es el *Umbral de la longitud de camino* que requiere que todos los vértices en un agrupamiento tenga una distancia k entre cada nodo. Para ser disponible un conjunto de datos del umbral *k*, uno debe conocer el diámetro del grafo de entrada, esto hace que difícilmente se encuentre un buen valor para el umbral *k*.

4.2.2 Mediciones finas

Este tipo de medición son funciones que se basan en la calidad de un agrupamiento dado. Tales funciones pueden ser utilizadas en la identificación de agrupamientos, seleccionar alternativas entre agrupamientos, y comparar diferentes algoritmos de agrupamiento.

4.2.2.1 Medición de densidad

Este tipo de medición parte de la idea básica de definir a una simple agrupación como un subgrafo que es denso con respecto a una medición de densidad dada. Es decir, dado un grafo no dirigido $G = (V, e)$ una medida de densidad $\delta(.)$ definida sobre el subconjunto de vértices $S \subseteq V$, un entero positivo $k \leq n$, y un número racional $\xi \leq [0, 1]$. Se debe cumplir que el subconjunto $|S| = k$ y la densidad $\delta(S) \geq \xi$

4.3 Agrupamiento de Grafos

La técnica de *agrupamiento* (o clustering) es la clasificación de objetos en diferentes grupos, o la partición de un conjunto de datos en subconjuntos, de modo tal que cada subconjunto comparte propiedades en común. El objetivo de un agrupamiento consiste en dividir el conjunto de datos en agrupaciones tales que los elementos asigandos a un grupo en particular son similares o estan conectados en algún sentdio.

Existen diferentes modelos para la generación de agrupamientos en grafos, un ejemplo, es la generalización del modelo de Gilbert (ver sección 2.7.1.1), el cual fue diseñado especialmente para producir agrupaciones y es llamado *modelo del planteo* λ-particiones, donde un grafo es generado con $n = \lambda k$ vértices que son particionados en λ grupos con k vértices cada uno. Dos probabilidades p y $q < p$ son usadas para construir el conjunto de aristas, cada par de aristas que están en el mismo grupo comparten un arista con la alta probabilidad p.

Existen dos tipos de agrupamientos el global y el local, en el primero se identifican los grupos dentro de la red original de tal manera que los datos en un grupo compartan cierta similitud entre ellos, este tipo de agrupamiento suele ser costoso cuando se maneja mucha información. En el agrupamiento local no es necesario conocer la forma en que se agrupan todos los datos de la red, basta con analizar a un solo elemento y determinar a que grupo pertenece, este tipo de agrupamiento parte de lo microscópico o particular hasta llegar al objetivo. En las siguientes seciones describimos los tipos de agrupamiento de grafos que se utilizan para la detección de comunidades en redes sociales según la clasificación en [82] .

4.3.1 Agrupamiento global de grafos

En este tipo de agrupamiento cada vértice del grafo de entrada es asigando a un agrupamiento para la salida. Una partición global está dada formalmente como: $C_1, ..., C_k$ agrupaciones, tal que $C_i \cap C_j = \emptyset$ cuando $i \neq j$

4.3.1.1 Complejidad del agrupamiento global.

El *problema del minimo k-agrupamiento*. Es un problema de optimización combinatorial donde un conjunto finito de datos *D* es dado junto con una distancia $d : DXD \rightarrow \mathbb{N}$ donde *d* satisface el triángulo de desigualdad de la ecuación 4.1. El objetivo es particionar *D* en *k* agrupa-

ciones $C_1, C_2, ..., C_k$, donde $C_i \cap C_j = \emptyset$ para $i \neq j$, tal que la distancia entre agrupaciones es minimizada (p.ej., la máxima distancia entre dos puntos en una agrupación). Este problema es aproximado con un factor de dos, pero no es aproximado con $(2-\epsilon)$ para cualquier $(\epsilon > 0)$.

El *problema del mínimo k-centro*. Dado un entero *k* y una función de distancia $d : VxV \rightarrow \mathbb{N}$. Encontrar un conjunto $S \subseteq V$, $|\subseteq |S| = k$ tal que la máxima distancia de un vértice al centro más cercano es minimizada. Cuando la función de distancia satisface el triángulo de desigualdad, el problema del mínimo *k*-centro puede ser aproximado con un factor de dos, pero no es aproximado con $(2 - \epsilon)$ para cualquier $(\epsilon > 0)$.

Un algoritmo para agrupar vectores de datos con respecto a su función de distancia es el algoritmo de *k-mediciones*, donde la idea básica de este método consiste en agupar a un conjunto de puntos mediante algúna métrica espacial en *k* agrupaciones, de manera iterativa se mejora el centro de la agrupación y cada grupo de puntos para el agrupamiento con centro cerrado; los centros son selecionados para minimizar la *suma de los cuadrados* de la distancia entre agrupaciones. Desafortunadamente, *k-medias* es NP-difícil siempre para *k*=2.

4.3.1.2 Agrupamiento jerárquico

Un agrupamiento global no necesariamente debe de tener la forma de partición simple o de cobertura, sino que también puede ser representado como una estructura jerárquica, donde los grupos son creados por niveles, el resultado es una jerarquía de grupos. Los métodos de agrupamiento que producen particiones multinivel son llamados *algoritmos de agrupamiento jerárquico*.

Un arupamiento jerárquico es generalmente construido al generar una *partición de secuencias*, donde cada *subagrupación* pertenece a un *superagrupamiento* en esa entidad. Este tipo de agrupamiento se basa en la matriz de distancias y no requiere definir *k* agrupaciones, pero si una condición de paro. Este tipo de agrupamiento genera un árbol al cual se le llama *dendograma* como el de la figura 4.1.

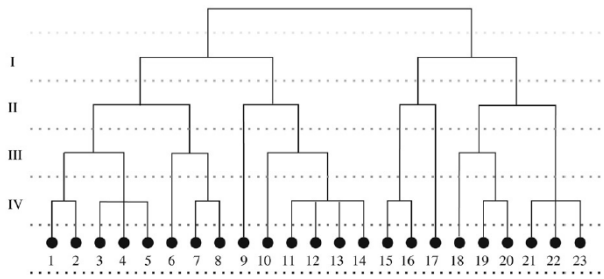


Figura 4.1: Ejemplo de un dendograma que agrupa a 23 elementos en 4 niveles.

Los algoritmos de agrupamiento jerárquico pueden ser divididos en dos clases:

- **Algoritmos de división.** Inicialmente todos los datos forman un solo grupo; los grupos menos coherentes son sucesivamente divididos.
- **Aglomerativos.** Inicialmente cada dato es un grupo; se unen de manera iterativa los grupos más cercanos.

La complejidad en el mejor caso para este tipo de agrupamiento es de $O(n^2)$ cuando son de *enlace simple* y para el peor caso se tiene que $O(n^2 \log n)$ cuando es de *enlace completo o promedio* (ver sección 4.3.1.4). Cuando la definición de la distancia no es tan trivial, la complejidad puede ser muy costosa.

4.3.1.3 Agrupamiento global de división

Este tipo de agrupamiento son de la clase de métodos jerárquicos que operan de arriba hacia abajo, el modelo trabaja recursivamente particionando el grafo en agrupaciones, la divisón de cada iteración es normalmente de dos conjuntos. En las siguientes subsecciones se definen algunos métodos basados en el agrupamiento global de división.

4.3.1.3.1 Método de agrupación por cortes. Una partición de los vértices V de un grafo $G = (V, E)$ en dos conjuntos no vacíos S y $V \setminus S$ es llamado corte y se representa como $(S, V \setminus S)$. Es decir, dados dos subconjuntos A y B de V vértices de un grafo G, existe un corte en G tal que $A \cup B = V$ y $A \cap B = \emptyset$. El *tamaño* de corte es el número de aristas que conectan vértices en S a vértices en $(V \setminus S)$.

Se dice que un corte es *mínimo* cuando el tamaño de dicho corte no es tan grande en comparación a los otros cortes del grafo, en la figura 4.2(a) se muestra un corte mínimo de tamaño 2. Un corte es *máximo* cuando el tamaño del corte no es el más pequeño, en la figura 4.2(b) se muestra un corte máximo igual a $|E| - 1 = 5$, donde $|E|$ es el número de aristas en el grafo.

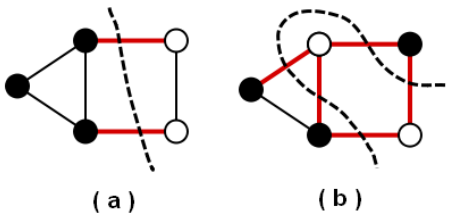


Figura 4.2: Ejemplo de los tipos de cortes en un grafo. (a) Corte mínimo de tamaño 2, y (b) Corte máximo de tamaño $|E| - 1 = 5$

La finalidad de este agrupamiento es buscar los subconjuntos más densos con respecto a la densidad global del grafo, una buena selección del corte separa dos o más agrupaciones, en lugar de romper en dos el conjunto de vértices de cualquier agrupación simple. Este tipo de agrupamiento está basado en dos ideas básicas, el *orden relativo* de los subgrafos separados por el corte y la condición de paro para la partición.

Donde el *orden relativo* de los grafos separados por el corte sucede cuando se realiza un corte y quedan vértices fuera de las agrupaciones y cuando se remueven vértices uno por uno, resultando en una agrupación única y estos no revelan ninguna propiedad estructural interesante. Sin embargo, establecer restricciones sobre como ordenar a los subgrafos, resulta un problema NP-difícil. La *mínima bisección* es el problema de dividir un grafo de $2n$ vértices en dos subgrafos de n vértices de tal manera que el tamaño de corte es minimizado. Las condiciones de paro para la partición del grafo es otro problema que enfrentan este tipo de algoritmos de agrupamiento.

4.3.1.3.2 Métodos de agrupamiento espectral. Este tipo de agrupamiento está basado en la teoría de grafos espectrales [19], donde la información de los grafos puede ser representada por una matriz de adyascencia, la cual es una matirz binaria $A(G)$ tal que $A_{ij} = 1$ solo cuando $(i, j) \in E(G)$. Una matriz relacionada es la matriz *Laplaciana* $L : D - A$, donde *D* es una matriz diagonal tal que $D_{ii} = deg_i$ (el grado del vértice *i*).

Cuando un grafo es formado por una colección de *k* disjuntos cliques, el laplaciano normalizado es una matriz de bloques en diagonal que tienen un eigenvalor cero con multiplicidad *k* y el correspondiente eigenvector sirve como función para indicar a los miembros en los correspondientes cliques. Para este tipo de agrupaciones, un eigenvector o combinación de varios eigenvectores es usada para medir la similaridad de vérticeS para el cálculo de agrupaciones.

Una variante de este método de agrupamiento fue propuesto por Qiu y Hancock en [76], donde utiliza un vector *Fielder*, el cual es el vector propio correspondiente al segundo valor propio más pequeño de la matriz Laplaciana. Para el caso de grafos no dirigidos, los elementos del vector Fielder se clasifican en negativos y positivos para determinar el agrupamiento, el cálculo de este vector puede ser aplicado recursivamente dentro de los grupos determinados de tal manera que se identifique el agrupamiento global del grafo. En el caso de grafos dirigidos el cálculo del vector de Fielder ya no es válido dado que la matriz laplaciana ya no es simétrica y deja de cumplir propiedades de agrupamiento.

4.3.1.4 Agrupamiento global aglomerativo

Este tipo de agrupamiento trabaja de abajo hacia arriba uniendo conjuntos simples de vértices iterativamente dentro de agrupaciones. Normalmente una medida de similaridad es usada para unir vértices dentro de una agrupación. Este tipo de agrupamiento parte desde

n agrupamientos, hasta llegar a obtener un número menor de agrupamientos que el inicial. Existen múltiples métodos para calcular distancias entre agrupaciones. Estos se calsifican como:

- **Métodos de enlace.** Estos métodos determinan las distancias *inter-agrupaciones* usando grafos definidos a partir de las relaciones de distancia entre todas las parejas de puntos compuestas por un miembro de cada agrupación.
  - *Enlace Simple.* La distancia entre dos agrupaciones se define como la mínima distancia entre dos puntos, tal que cada uno de los puntos pertenece a un agrupamiento distinto. Matemáticamente se puede expresar como:

$$D(C_i, C_j) = \min_{i \in C_i j \in C_j} d(i, j) \tag{4.7}$$

    donde:

    $d(i, j)$ es la distancia entre los vertices i y j. C_i y C_j son dos agrupaciones de vertices.
  - *Enlace Completo.* La distancia entre dos agrupaciones se define como la máxima distancia entre dos puntos, tal que cada uno de los puntos pertenece a un agrupamiento distinto. Matematicamente se puede expresar como:

$$D(C_i, C_j) = \max_{i \in C_i j \in C_j} d(i, j) \tag{4.8}$$

    donde:

    $d(i, j)$ es la distancia entre los vertices i y j. C_i y C_j son dos agrupaciones de vertices.
  - *Enlace Promedio.* La distancia entre dos agrupaciones se define como la distancia promedio entre todas las parejas de puntos, que tienen un componente miembro de cada agrupación distinta.
- **Métodos geométricos.** Estos métodos definen un centro para cada agrupación, y lo usan para determinar la distancia entre las agrupaciones.
  - *Centroide.* El centro de una agrupación se define como el centroide geométrico de los puntos miembros del agrupamiento. La distancia entre dos agrupaciones es la distancia euclídeana entre sus centros.
  - *Mediana.* El centro de un agrupamiento se define como el promedio entre los centros de las dos agrupaciones. La distancia entre dos agrupaciones es la distancia euclidiana entre sus centros.

– *Varianza mínima de Ward.* El centro de una agrupación se define como el centroide geométrico de los puntos miembros del agrupamiento. La distancia entre dos agrupaciones es el incremento en la suma de las distancias cuadráticas de cada punto al centro, causada por la aglomeración de ambos.

El algoritmo para este tipo de agrupamiento tiene una complejidad de $O(n^2 \log n)$ y se muestra a continuación:

```
Entrada:  n agrupaciones
Inicio:
1.  Para cada agrupamiento, generar una lista de prioridad con las distancias
entre a cada una de las agrupaciones restantes
2.  Seleccionar las dos agrupaciones más cercanas
3.  Aglomerar las dos agrupaciones seleccionadas
4.  Actualizar las distancias de las listas de prioridad para el nuevo
agrupamiento y los otros
5.  Si quedan dos o más agrupaciones regresar al paso 2
```

Algoritmo 5: Agrupamiento aglomerativo.

4.3.2 Agrupamiento local de grafos

Para grafos grandes, los métodos de agrupamiento global consumen demasiados recursos y tardan mucho en almacenar una gran cantidad de información. Sin embargo, cuando los grafos se encuentran almacenados en un formato que permite conectar subgrafos o listas de adyascencia de vértices cercanos; entonces pueden ser aplicadas ideas similares como el agrupamiento aglomerativo, donde las agrupaciones pueden ser calculadas una a la vez basado en ***vistas parciales*** del grafo general.

Los métodos de búsqueda local son algoritmos heurísticos y/o probabilísticos diseñados para encontrar solución óptima cercana de soluciones candidatas. Estos métodos no exploran el espacio entero de soluciones, pero mediante decisiones probabilisticas establecen una región que contenga a las mejores soluciones. Cada solución candidata es representada por un estado el cual es llamado *espacio de estado*. Una relación de vecindario permite realizar la búsqueda de un estado a otro con muy pocos cálculos. El agrupamiento local de grafos es una área de investigación que en años recientes ha sido retomada para tratar problemas de grafos con un número grande de información, sin embargo, los trabajos [81, 72] sobre este tipo de agrupamiento es mucho menor en relación a los trabajos sobre agrupamiento global.

En la figura 4.3 se muestra un resumen de la clasificación de los tipos de agrupamientos de grafos presentados en este capítulo.

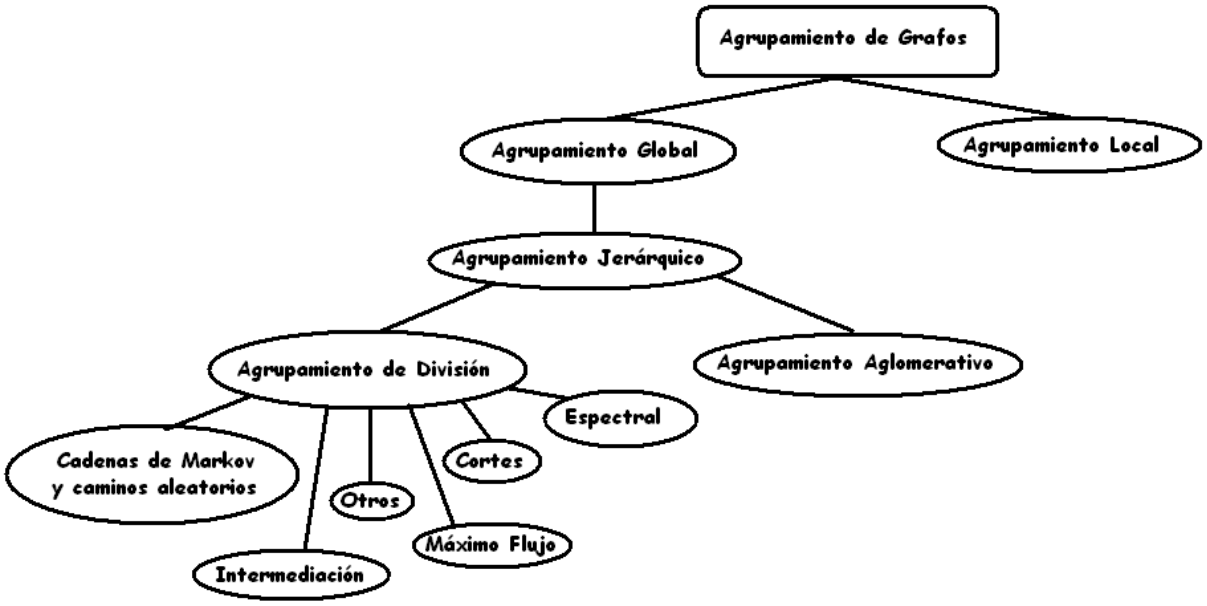


Figura 4.3: Clasificación de los tipos de agrupamiento en grafos

4.4 Detección de comunidades

Existen muchos algoritmos para encontrar la estructura de comunidades en un grafo, muchos de estos realizados en años recientes [35, 77, 79, 80, 89, 90]. Los algoritmos para la detección de comunidades atacan dos problemas principalmente, los cuales son:

- **Detección de comunidades disjuntas.** Para este tipo de algoritmos se generan n comunidades o particiones de individuos, donde cada individuo solo puede pertenecer a una comunidad. Por ejemplo, muchos empleados trabajan para un solo jefe o un departamento, un libro por lo general solo puede ser publicado por una editorial, etc.
- **Detección de comunidades con traslapamiento.** Para este caso un individuo puede pertenecer a más de una comunidad. Un ejemplo de este tipo de detección puede ser visto en los sistemas de redes sociales en línea, donde un individuo puede estar en más de un solo grupo, es decir, una persona puede pertenecer a un grupo que comparte ciertas preferencias sobre música y pertenecer también a un grupo especialista en peliculas y no estar exclusivo a un solo grupo.

Las redes del mundo real presentan comunidades más parecidas a las estudiadas con el problema de traslapamiento. Dentro de la detección de comunidades una solución al problema de traslapamiento consiste en encontrar a los individuos candidatos que compartan comunidades y descomponerlo en n nodos conectados como se muestra en la figura 4.4, donde los n nodos dependerán del número de comunidades a donde pertenece un individuo.

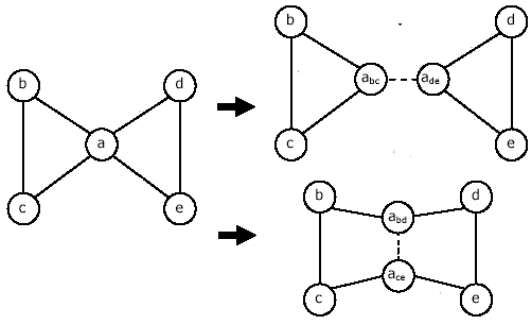


Figura 4.4: Descomposición de un vértice en una red con traslapamiento.

4.4.1 Algoritmo basado en la intermediación

Girvan y Newman[34, 62] proponen un algoritmo para encontrar comunidades disjuntas basado en la centralidad de intermediación de las relaciones (ver sección 2.6.3.3) como se muestra en la figura 4.5, donde en cada paso se calcula la intermediación para cada arista en el grafo y se remueve la arista con mayor valor de intermediación.

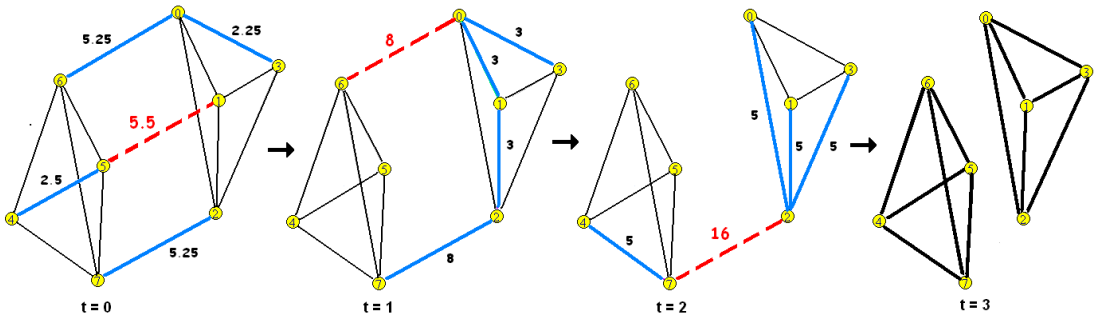


Figura 4.5: Detección de comunidades mediante el cálculo de la intermediación de aristas.

El algoritmo para la detección de comunidades basado en el cálculo de la intermediación de aristas de la red propuesto por Girvan y Newman es el siguiente:

Entrada: Grafo inicial G(E,V)

Inicio:

1. Calcular la intermediación para todas las aristas en la red.
2. Generar un nuevo grafo al remover las aristas con un valor alto de intermediación.
3. Recalcular la intermediación para todas las aristas del nuevo grafo.
4. Repetir el paso 2 hasta que todas las aristas tengan un nivel bajo de intermediación.

Algoritmo 6: Detección de comunidades utilizando la intermediación de aristas.

El valor de intermediación de una arista expresa la importancia de esta con respecto a la red y mide la proporción de trayectorias más cortas entre nodos que pasan por un vínculo (ver ecuación 2.12). Este algoritmo se ejecuta en un tiempo de $O(m^2n)$. Sin embargo, el recalcular la intermediación de cada arista en el grafo es un proceso muy costoso, en el caso de redes que cuentan con millones de aristas este proceso puede resultar muy poco óptimo. En [37] se propone un algoritmo que estima la centralidad de intermediación de una arista basado en el muestreo de pocos pares de nodos.

Newman utiliza el algoritmo de Girvan-Neman para obtener *n* comunidades o grupos de vértices, y propone en [63] un algoritmo que toma estas comunidades, y posteriormente une las comunidades utilizando una función de modularidad, la cual mide que tan buenas son estas divisiones, o que tan separados estan los vértices de una comunidad con respecto a las otras comunidades. La función de modularidad que plantea Newman está dada por:

$$Q = \sum_i (e_{ii} - {a_i}^2) \tag{4.9}$$

donde:

e_{ij} es el número de aristas en la red que conecta vértices del grupo *i* al grupo *j*, y $a_i = \sum_j e_{ij}$.

La función de modularidad mide la calidad de una partición con respecto a las demás particiones. El algoritmo de Neman se ejecuta en un tiempo de $O(mn)$ y tiene una complejidad de $O(mH\log \mathrm{n})$.

4.4.2 Algoritmo basado en cliques

La definición perfecta para una comunidad en una red es un clique(ver sección 2.3.4), ya que está estructura provee de una gran conectividad a los individuos. Este tipo de estructuras han motivado a muchos trabajos sobre la detección de comunidades, ya que toman como elemento principal al clique para poder construir comunidades más complejas a partir de estructuras simples.

Bowen [90] propone un algoritmo para detectar comunidades utilizando este principio básico, primero se obtienen los cliques del grafo y se les considera como comunidades iniciales y en cada paso se van uniendo las comunidades iniciales mediante la función de modularidad propuesta por Newman(ec. 4.9). La figura 4.6 muestra el proceso de detección de comunidades mediante el uso de cliques.

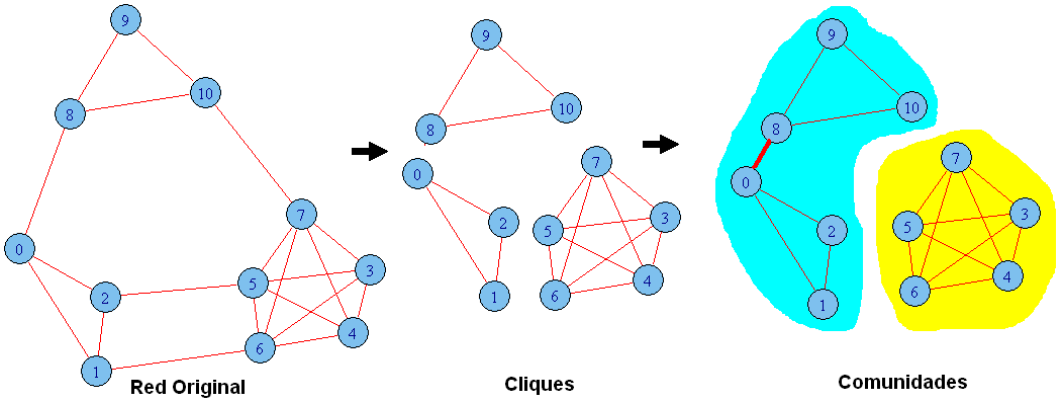


Figura 4.6: Detección de comunidades basado en cliques.

El algoritmo para la detección de comunidades mediante el uso de cliques desarrollado por Bowen es el siguiente:

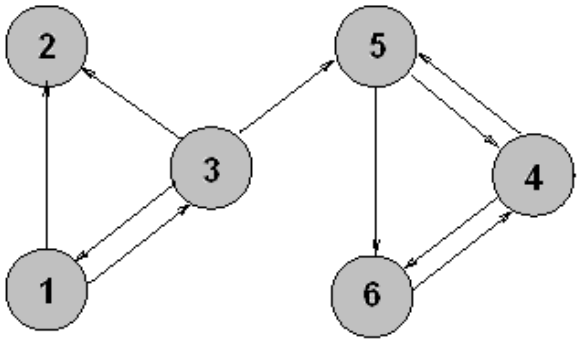
Entrada: Grafo inicial G(E,V)
Inicio:
1. El grafo es dividido en un conjunto separado de cliques, mediante el uso de un algoritmo para encontrar aproximaciones de máximos cliques. Estos cliques son utilizados como las comunidades iniciales de la partición.
2. El número de comunidades se reduce uniendo pares de comunidades. El proceso se repite hasta que se optimiza la función de modularidad de la ecuación 4.9.

Algoritmo 7: Detección de comunidades utilizando cliques.

La complejidad de este algoritmo es de $O(n^2 \log n)$ en el peor caso. La desventaja de este tipo de métodos basados en cliques está en el cálculo del máximo clique, ya que los algoritmos para detectar cliques son NP-completos. Sin embargo, este algoritmo reduce el problema aproximando la búsqueda del máximo clique.

4.4.3 Algoritmo basado en caminos aleatorios

Otros algoritmos para detectar comunidades estan basados en el concepto del camino aleatorio, el cual se describe como: En cada paso un caminante está en un vértice y se mueve a otro vértice en su vecindario de manera aleatoria, la secuencia de visita es una *cadena de Markov*, donde los estados son los vértices del grafo. En cada paso, la probabilidad de transición del vértice *i* al vértice *j* es $P_{ij} = \frac{A_{ij}}{d(i)}$. La probabilidad de ir de *i* a *j* mediante un camino aleatorio de longitud t es P_{ij}^t. La figura 4.7(b) muestra la matriz de transición del camino aleatorio para el grafo de la figura 4.7(a).



$$\mathbf{P}= \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array}$$

(b)

Figura 4.7: Uso de cadenas de Markov para la detección de comunidades.

Un algoritmo basado en camino aleatorio se presenta en [75] y utiliza un algoritmo jerárquico aglomerativo, el cual le permite encontrar comunidades a diferentes escalas. El algoritmo es el siguiente:

Entrada: Grafo inicial G(E,V)
Inicio:
1. Se comienza con una partición P_1 del grafo en *n* comunidades.
2. Se calculan las distancias entre todos los vértices adyascentes.
for $k < n-1$ **do**
  3. Se seleccionan dos comunidades C_1 y C_2 en P_k adyascentes entre sí.
  4. Se unen las dos comunidades en una nueva $C_3 = C_1 \cup C_2$ y se crea una nueva partición: $P_{k+1} = (P_k\{C_1, C_2\}) \cup \{C_3\}$.
  5. Se actualizan las distancias entre comunidades.
end

Algoritmo 8: Detección de comunidades utilizando camino aleatorio.

Para el cálculo de la distancia entre comunidades se utiliza la siguiente ecuación:

$$rC_1C_2 = \sqrt{\sum_{k=1}^{n} \frac{(P^t_{C_1k} - P^t_{C_2k})^2}{d(k)}} \tag{4.10}$$

donde:

- $P^t_{C_1j}$ y $P^t_{C_1j}$ es la probabilidad de ir de una comunidad a un vértice *k* en *t* pasos.
- $d(k)$ es el grado del vértice *k*.

4.4.4 Algoritmo basado en etiquetado

Raghavan [77] propone un algoritmo para encontrar comunidades disjuntas basado en la propagación de etiquetas. Este algoritmo tiene la particularidad de ser el más rápido en encontrar comunidades disjuntas en una red y es uno de los pocos algoritmos que utilizan el etiquetado de vértices para detectar comunidades en una red. El algoritmo es sencillo y no requiere de muchos cálculos matriciales y consume poco tiempo en su ejecucón.

En la figura 4.8 se muestra el proceso de propagación de etiquetado para el algoritmo de Raghavan con una red con tres comunidades.

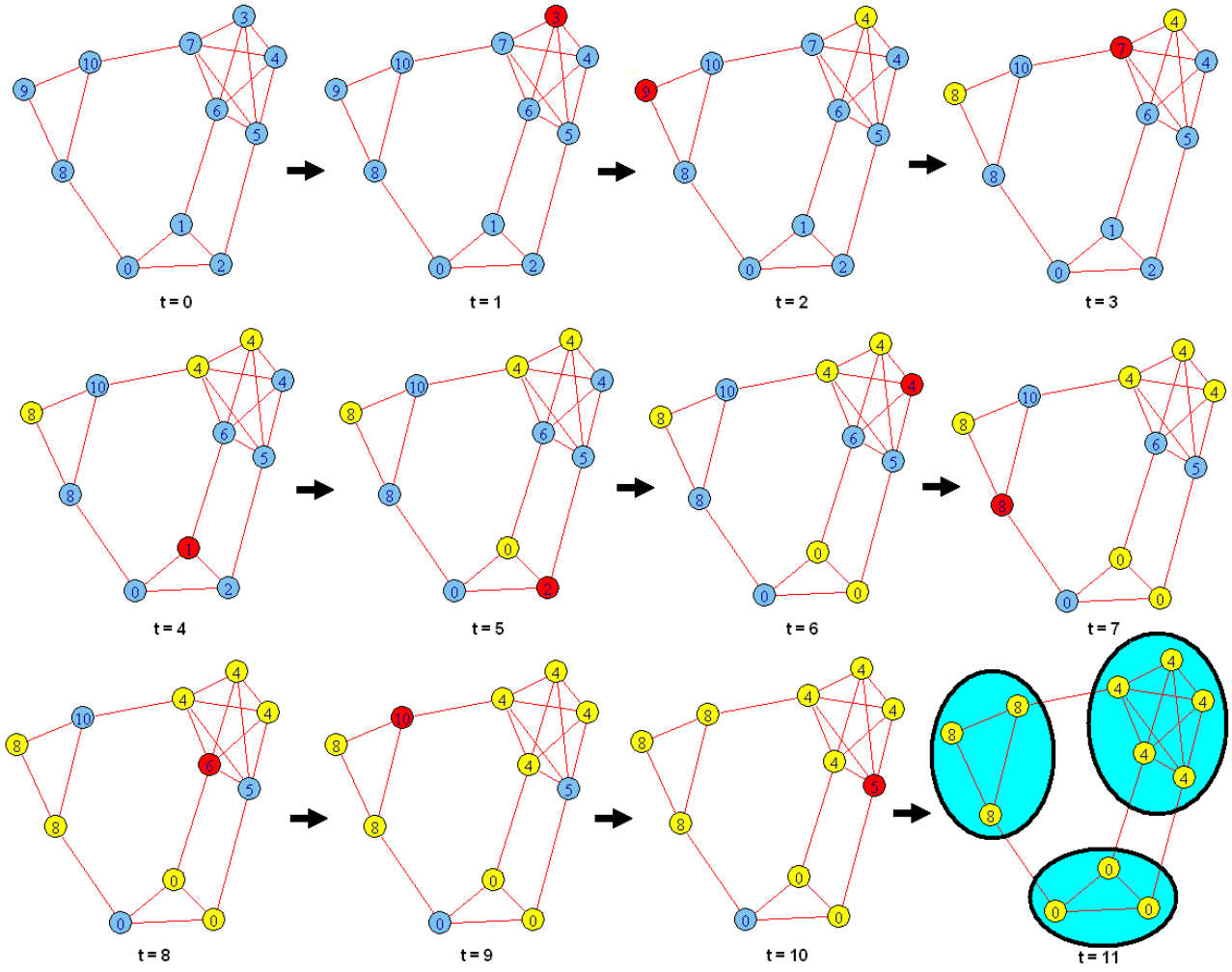


Figura 4.8: Proceso para la detección de comunidades mediante la propagación de etiquetas.

El algoritmo de Raghavan toma un tiempo lineal de $O(m+n)$. Para la parte de la asignación de cada etiquetas para cada vértice se requiere de un tiempo $O(n)$, en cada iteración donde se actualizan las etiquetas se toma un tiempo lineal en el número de aristas $O(m)$, el algoritmo es el siguiente:

```
Entrada: Grafo inicial G(E,V)
Inicio:
1. Se asigna una etiqueta única para cada vértice (generalmente es el número
del vértice).
2. Cada vértice actualiza su etiqueta utilizando la información de sus vecinos,
selecciona la etique más común entre sus vecinos y reemplaza su etiqueta. Si
hay más de una etiqueta entre los vecinos se selecciona aleatoriamente uno de
ellos. Después de varias iteraciones la etiqueta tiende a ser asociada con
todos los miembros de la comunidad
3. Todos los vértices con la misma etiqueta son agrupados en una comunidad
```

Algoritmo 9: Detección de comunidades utilizando propagación de etiquetas.

4.5 Algoritmo propuesto

El algoritmo propuesto tiene dos fases, la primera fase del algoritmo se basa en un algoritmo de búsqueda de comunidades disjuntas, para nuestro caso utilizamos el método de propagación de etiquetas por su rapidez para detectar comunidades. Muchos sistemas de redes sociales en línea (p.ej., Facebook, Twitter, Flickr, etc) poseen más enlaces entre usuarios que usuarios mismos, por lo que métodos de detección de comunidades basados en el análisis de las aristas (p.ej., algoritmo de Girvan-Newman) no resultan ser tan convenientes para redes con millones de aristas. Para la segunda fase se propone un algoritmo para crear comunidades con traslapamiento.

4.5.1 La estrategia

Los algoritmos para la detección de comunidades mediante propagación de etiquetas solo permite encontrar comunidades disjuntas, para el caso de comunidades con traslapamiento se propuso una segunda fase, donde recibe como entrada las comunidades encontradas en la etapa anterior. El problema se reduce en encontrar a los nodos candidatos que seran agregados a otra(s) comunidad(es). Mediante el *coeficiente de agrupamiento local* para un vértice (ver sección 2.3.3) podemos medir la conexión que existe entre un vértice y su vecindario.

El uso del cálculo matricial sobre las redes sociales ofrece una manera muy rápida para realizar operaciones, pero cuando son millones de nodos y vértices estos métodos pueden

resultar muy poco convenientes. Es por eso que para nuestro análisis almacenamos la información en una Base de Datos. La figura 4.9 muestra el proceso de consultas realizadas para la implementación del algoritmo de detección de propagación de etiquetas usado en nuestra primera fase.

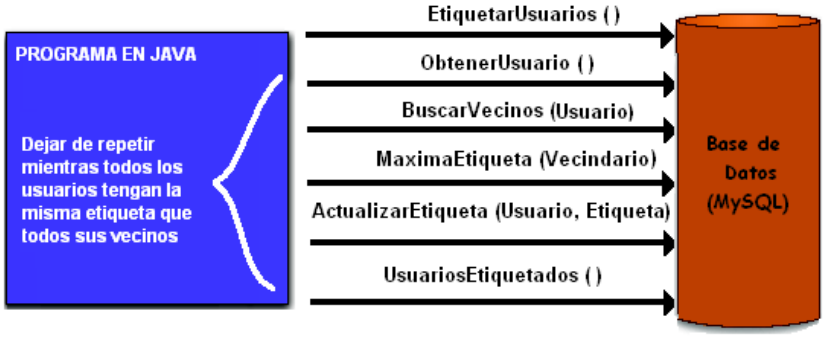


Figura 4.9: Consultas para el algoritmo de detección de comunidades mediante propagación de etiquetas para redes sociales a gran escala.

La segunda fase del algoritmo recibe a las C_n comunidades de la fase uno, a n como el número de comunidades, el grafo inicial $G(E, V)$ con el conjunto de nodos y enlaces iniciales y un parámetro θ, el cual controla el traslapamiento de un nodo en las diferentes comunidades. Los pasos del algoritmo son los siguientes:

- Comparar las n comunidades y obtener los nodos en común, es decir, obtener los nodos que se intersecctan dentro de las distintas comunidades.
- Agregar los nodos y los enlaces obtenidos en la intersección de las comunidades donde potencialmente puedan formar parte.
- Calcular el *coeficiente de agrupamiento local* para cada nodo agregado con respecto a las comunidades tomadas.
- Finalmente, se mide la relación del vértice en las distintas comunidades y se determina mediante un umbral de aceptación θ si el nodo es removido de la comunidad actual o se queda para formar parte de ella.

4.5.2 El algoritmo DCRS

El algoritmo para detectar comunidades con traslapamiento en redes sociales a gran escala propuesto lo llamaremos de aquí en adelante como DCRS. El algoritmo correspondiente para la segunda fase del algoritmo DCRS es la siguiente:

```
Entrada:  C_1, ..., C_n ,  θ,  G(E,V)
Inicio:
while i < n do
    while j < n do
        L_cc ⇐ BuscarNodos(C_i, C_j);
        C_j ⇐ AgregarNodosComunidad(C_j, L_cc)
        // Agregar aristas del grafo original
        C_j ⇐ AgregarEnlaces(E_G[L_cc])
        while k < L_tam do
            // Calcular los coeficientes de agrupamiento de cada vertice
            C_i ⇐ CoeficienteAgrupamiento(C_i, L_cc[k])
            C_j ⇐ CoeficienteAgrupamiento(C_j, L_cc[k])
            C_dif ⇐ C_i − C_j
            if C_dif > θ then
                remover(C_j, k);
            end
        end
    end
end
```

Algoritmo 10: Segunda fase del algoritmo para detección de comunidades con traslapamiento.

Esta segunda fase utiliza las comunidades disjuntas generadas por el algoritmo de propagación de etiquetas y regresa el mismo número de comunidades de entrada, aunque agrega a los nodos que cumplan con el nivel de aceptación θ, el cual debe estar en el rango de $(1, 0]$.

4.5.3 Análisis de la complejidad

La complejidad del algoritmo está determinado por la suma de sus dos fases, para la primera parte tomamos la complejidad mostrada en el algoritmo de detección de comunidades con propagación de eqtiquetas, el cual está dado por $O(m + n)$. Para la parte de la segunda fase tenemos una complejidad de $O(n^2)$ para la etapa en la que se comparan cada uno de las comunidades entre sí. Para el cálculo del coeficiente de agrupamiento local de los nodos tenemos una complejidad de $O(k)$ para los k vecinos del nodo. Por lo tanto, tenemos que la complejidad del algoritmo está dada por $C_{algoritmo} = O(m + n) + O(n^2 + k)$

4.5.4 Pruebas y comparación

En esta sección, se probó el algoritmo DCRS, para detectar comunidades en distintas redes sociales, incluyendo varios conjuntos de datos recolectados de redes de mundo real. Todos los experimentos se realizaron en una PC con procesador 2.6 GHz Xeon y 4 GB de memoria RAM, y los algoritmos utilizados en este cápitulo son implementados con el programa R (para más detalle ver 2.8.3) y otros en Java.

4.5.4.1 Conjuntos de datos de prueba

Para realizar pruebas y comparativas sobre nuestro algoritmo fue necesario utilizar diferentes redes que nos permitieran aplicar y observar la ejecución de los diferentes métodos para detectar comunidades. A continuación se describen los conjuntos de datos seleccionados para llevar a cabo las pruebas sobre los algoritmos vistos en la sección 4.4.

4.5.4.1.1 Conjunto de datos sintéticos. Para realizar las pruebas sobre el rendimiento del algoritmo propuesto y los algoritmos vistos en la sección 4.4, generamos un conjunto de grafos generados mediante una distribución aleatoria mediante el modelo BA (ver sección 2.7.3.1), el procedimiento para generar una red de datos sintéticos se muestra en la figura 4.10. Para probar los distintos algoritmos se generó un grafo con 1,000 vertices, dividido en 5 comunidades de 200 vertices cada una, posteriormente se enlazaron estas comunidades disjuntas al agregar relaciones entre comunidades, y así poder formar la red completa con una estructura conocida, lo cual permite hallar las particiones y comparar resultados de manera sencilla. Usando estos grafos probamos el rendimiento del algoritmo DCRS con respecto a los algoritmos de deteccion de comunidades mostrados en la sección 4.4.

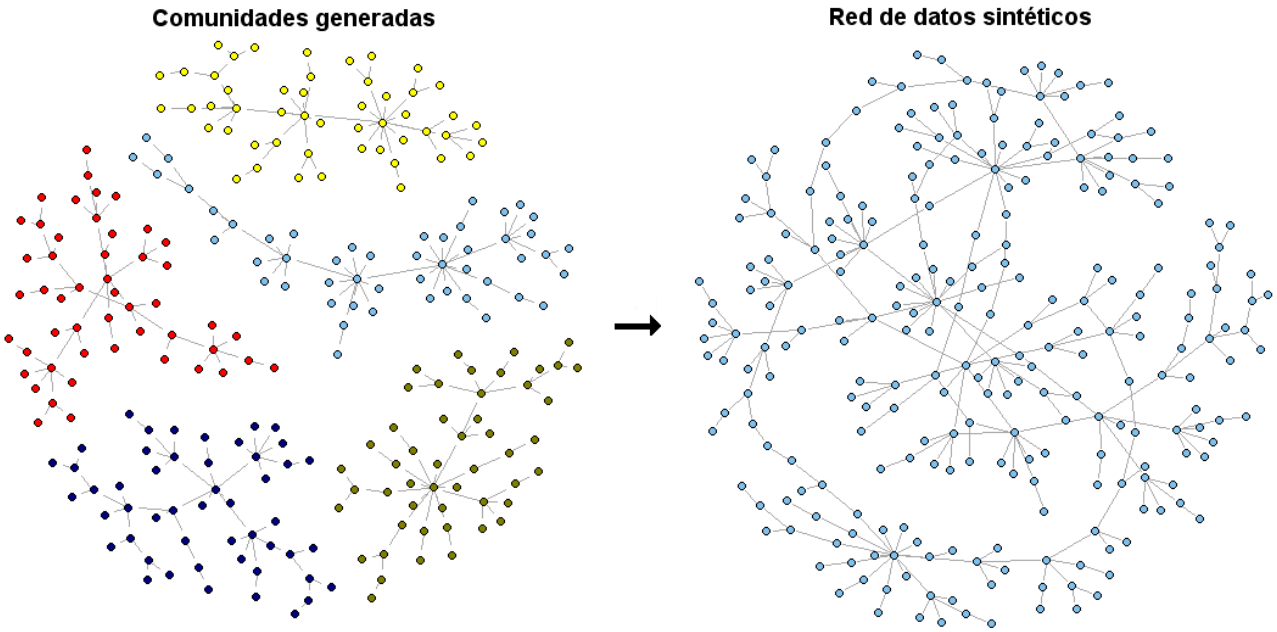


Figura 4.10: Generación de una red de datos sintéticos, mediante el modelo de Barabasi-Albert con 250 vertices en 5 comunidades de 50 vertices cada una.

4.5.4.1.2 Conjunto de datos de redes de mundo real. Los datos sintéticos generados por las redes aleatorias permiten evaluar de manera controlada la detección de comunidades, debido a que la estructura es conocida desde un principio. Sin embargo, es deseable probar y comparar el rendimiento de nuestro algoritmo sobre redes del mundo real, las cuales exhiben una estructura diferente. Para nuestro estudio se tomaron en cuenta 5 conjuntos de datos que representan a redes social de mundo real y son los siguientes:

- **Red del club de Karate de Zachary.** En los años 70's el sociólogo Wayne Zachary estudia por dos años la interacción social entre 34 miembros de un club de karate [91], por medio de esta observación fue capaz de crear una red entre la relación de los miembros del club. Esta red tiene una particular relevancia en el campo de las detección de comunidades debido a que el grupo sufrió una separación en dos pequeños clubs, déspues de una disputa entre el administrador y el instructor del club. Usando esta información se generó una red no dirigida con 34 nodos y 78 enlaces. Estos datos han sido utilizados en diferentes trabajos de investigación sobre detección de comunidades [34, 63]

- **Red de libros sobre políticos.** En esta red los nodos representa a los libros que hablan sobre políticos estadounidenses vendidos por *amazon.com*. Un enlace representan la frecuencia de que un comprador adquiera dos libros, la red formada a partir de estos datos es una red no dirigida con 105 nodos y 441 enlaces. Estos datos se encuentran disponibles en `http://www.orgnet.com/`

- **Red de delfínes.** Esta red representa la interacción entre un conjunto de delfines en Nueva Zelanda, y está representada como un grafo no dirigido con 62 nodos y 159 enlaces, este conjunto de datos fue presentado y utilizado por primera vez en [50]

- **Red de contactos de MySpace.** Los nodos de esta red son representados por usuarios del sistema MySpace y la relacion entre ellos representa los enlaces de la red dirigida que se genera a partir de este conjunto de datos. Este conjunto de datos presenta 100,000 nodos y 6,865,571 enlaces de la red de MySpace en 2006. Esta red es utilizada en el trabajo [4].

- **Red de recomendaciones de Amazon.** Es una red obtenida a partir del sistema Amazon.com, donde los nodos estan representados por los clientes de amazon y los enlaces representan la acción de recomendar a otro cliente un libro. Estos datos generan una red dirigida de 262,111 nodos y 1,234,877 enlaces. Esta red es utilizada por primera vez en [48]

Los algoritmos para detectar comunidades consumen mucho tiempo y utilizan mucho espacio en memoria, por lo que se decidió comparar el algoritmo con conjuntos de datos pequeños y de mediana escala. Se puede apreciar que existen diferentes tipos de redes dependiendo el punto de vista del análisis, como la red de políticos y la red de recomendaciones que son obtenidas de amazon pero con diferentes enfoques.

4.5.4.2 Datos estadísticos para los conjuntos de datos

La tabla 4.1 muestra datos estadísticos calculados para cada conjunto de datos utilizados en nuestra etapa de pruebas, se pueden apreciar las diferentes características ofrecidas para cada conjunto de datos. Para las redes con conjuntos de datos pequeños (Sintética, Karate, Amazon-Pol, Delfínes) se puede apreciar un coeficiente de agrupamiento global es pequeño y para las redes con conjuntos de datos grandes (MySpace y Amazon-Rec) el coeficiente de agrupamiento global es alto, esto es debido a que las redes con más información (especialmente con un número mayor de enlaces) estan altamente conectadas, esto permite tener una amplia conectividad en la red según el modelo de redes de mundo pequeño de Watts-Strogatz (ver sección 2.7.2.1).

Estadisticas	Sintética	Karate	Amazon-Pol	Delfínes	MySpace	Amazon-Rec
Red Dirigida	No	No	No	No	Si	Si
Número de nodos	1,000	34	105	62	100,000	262,111
Número de enlaces	3,413	78	441	159	6,865,571	1,234,877
Número de Cliques	5	5	6	5	500	142
Transitividad (CC)	0.4548	0.2476	0.3484	0.3088	0.7124	0.6240
Diámetro de la red	8	5	7	8	125	29
Trayectoria más corta	4.5500	2.4099	3.0787	3.3570	3.2230	3.5232
Grado promedio	3.4	2.29	4.2	2.5	6.86	4.7
Cercania Global	0.2217	0.4260	0.3296	0.3072	0.2133	0.264
Intermediación de nodo promedio	1773.247	23.2647	108.0952	71.8871	154.21	135.21
Intermediación de enlace promedio	665.9089	17.3333	38.1179	39.9245	102.64	87.21

Tabla 4.1: Estadísticas de los conjuntos de datos utilizados para medir el algoritmo DCRS.

4.5.5 Análisis de resultados

La tabla 4.2 muestra los resultados obtenidos de aplicar los algoritmos para detección de comunidades para cada uno de nuestros conjuntos de datos propuestos. Para nuestro algoritmo se propuso un valor para θ igual a 0.085, tomando en cuenta que el valor de θ es la diferencia del coeficiente de agrupamiento de un nodo en diferentes comunidades, esto hace más difícil que dos comunidades de una red de tamaño pequeño se unan con valores pequeños.

	Comunidades encontradas					**Tiempo de Ejecución (segundos)**				
Conjunto de Datos	**DCRS θ=0.085**	**Newman**	**BK**	**PE**	**CA**	**DCRS**	**Newman**	**BK**	**PE**	**CA**
Sintética	5 n=0	5 Q=0.774	5 Q=0.781	5	5	203.1	302.94	2.12	0.13	0.10
Karate	3 n=0	3 Q=0.363	3 Q=0.416	3	3	0.6	0.05	0.02	0.001	0.010
Delfines	4 n=0	4 Q=0.519	5 Q=0.522	4	4	0.19	0.17	0.08	0.013	0.040
Amazon-Pol	4 n=0	5 0.517	5 Q=0.509	4	4	1.54	1.43	1.09	0.024	0.035
MySpace	8 n=2,451	9 Q=0.426	12 Q=0.412	8	9	8644	8326	7921	1941	3012
Amazon-Rec	10 n=4,115	13 Q=0.322	16 Q=0.314	10	15	9123	7131	6312	4091	4102

BK - Basado en cliques
PE - Propagación de Etiquetas
CA - Camino Aleatorio
n - Número de nodos traslapados
Q - Modularidad de Newman

Tabla 4.2: Estadística de resultados de los diferentes algoritmos para la detección de comunidades sobre diferentes conjuntos de datos.

Capítulo 5

Casos de Estudio

La base principal para un estudio sobre el análisis de redes sociales es la información, dicha información es necesaria para realizar mediciones, comparaciones y pruebas. Donde el tamaño de la información y como esta se genera son factor fundamental para lograr un buen estudio.

Para nuestros casos de estudio seleccionamos dos tipos de sistemas de redes sociales en línea muy distintos entre sí. Primero analizamos a Flickr un sistema basado en la compartición de archivos, ocupa el lugar número 33 [1] en generación de tráfico en Internet y posee aproximadamente 32 millones de usuarios. El segundo caso de estudio es *Wikipedia* pese a no ser un sistema basado en relaciones personales directas, utiliza las relaciones sociales para generar autocontenido, actualmente ocupa el lugar número 6 en generación de tráfico en Internet y cuenta con más de 13.7 millones de artículos.

5.1 El caso Flickr

Estudios recientes [4, 42, 60] han mostrado que los sistemas de redes sociales (p.ej., Facebook, MySpace, Flickr, entre otros) muestran ciertos parecidos en su estructura. Una de las principales características de las redes sociales en línea es su alto dinamismo, es decir, no son redes que permanecen con un tamaño fijo a través del tiempo. Este tipo de redes se ven afectadas por el hecho de agregar o remover actores y/o relaciones en la red. Para este caso de estudio seleccionamos a Flickr (para más detalle ver sección 3.2) por contar con un número alto de usuarios, ser el número uno entre los sistemas de su tipo (compartición de fotos), y por estar dentro de los primeros 20 sitios de redes sociales que generan más tráfico en la red. En la figura 5.1 se muestra las etapas realizadas para la tesis sobre el anásis de la red social Flickr.

[1] Para más información ver la sección 2.4.2

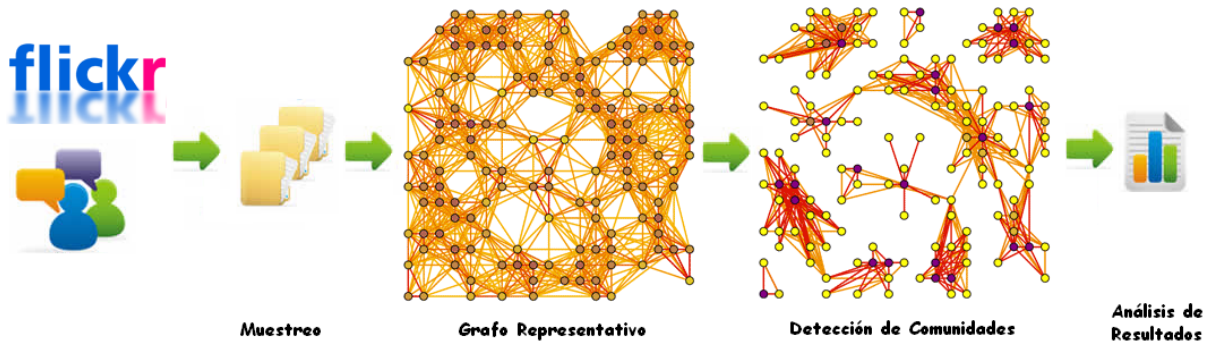


Figura 5.1: Etapas del análisis de la red social Flickr.

Cuando hacemos referencia sobre Flickr, tenemos que hablar sobre usuarios que comparten fotos, y de la tecnología que estos sistemas utilizan para poder permitir a los usuarios interactuar entre sí de diferentes formas. Este tipo de sistemas son del tipo *FotoBlog*, el cual hace uso de etiquetas y permite comentarios sobre las fotos, esto permite a los usuarios relacionarse y ofrece una manera de distinguir a los contactos con mayor interacción, creando un nivel de amistad entre los usuarios. Este tipo de redes sociales en línea difieren de las otras redes por su forma de relacionarse, ya que permiten a los usuarios interactuar mediante el uso de contenido como son: videos, imagenes, aplicaciones, etc.

Es común que en los sistemas de redes sociales personales (p.ej., Facebook, MySpace, Hi5, entre otras) los usuarios que agregan a un usuario generalmente estan representando a personas que conocen en el mundo real, sin embargo, en redes como Flickr y Youtube el interés de tener más contactos está basado por el contenido que pueden ofrecerse. La forma en como las relaciones se dan permiten medir a las redes por un nivel de amistad que va más allá de solo agregar contactos.

5.1.1 Análisis de la red de Flickr

Flickr permite crear una red de contactos mediante el uso de un API de desarrollo, utilizando un proceso de muestreo como el presentado en la sección 3.2, se puede generar un grafo dirigido que represente la interacción de usuarios en la red, donde los usuarios son los nodos de la red y la relación entre usuarios son los enlaces de la red. La figura 5.2 muestra la estructura de la Base de Datos utilizada para almacenar la información durante el proceso de muestreo. En este tipo de sistema un usuario puede tener ninguna o muchas fotos[2] y las fotos pueden tener muchos comentarios de diferentes usuarios, esta característica permite determinar cuales son los enlaces entre usuarios fuertes o débiles dependiendo la interacción que entre ellos exista.

[2]Flickr permite subir un contenido total de hasta 100MB de fotografías al mes, cuentas de paga permiten subir más fotografías.

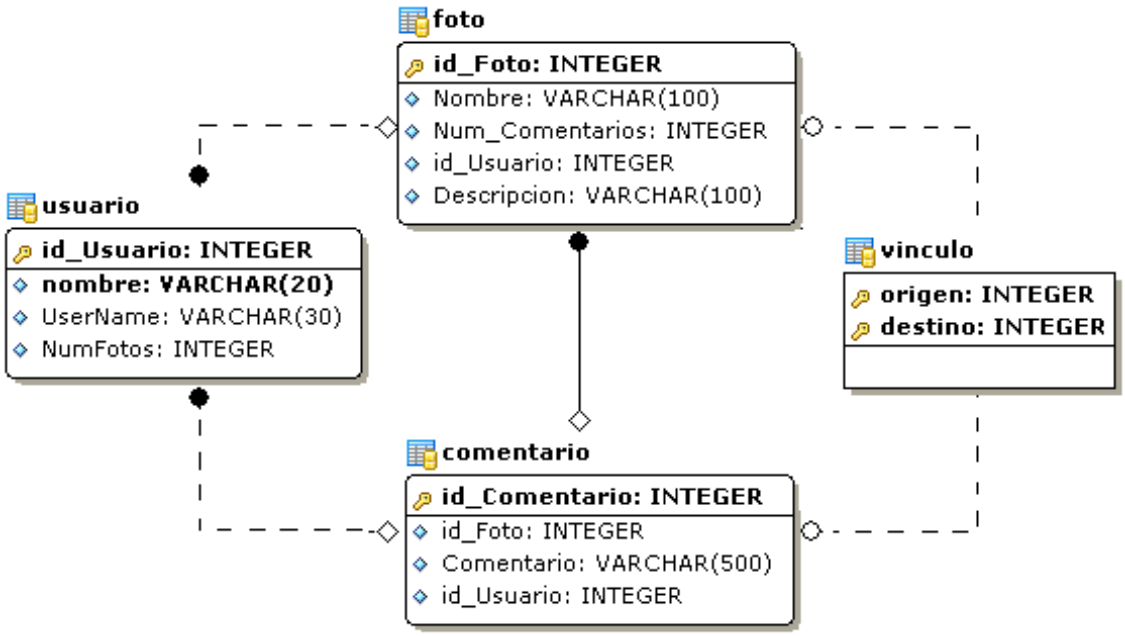


Figura 5.2: Estructura de la Base de Datos utilizada para el muestreo en Flickr.

Para este caso de estudio se hicieron dos análisis principalmente, primero comparamos en base a su estructura las dos redes generadas en el muestreo (red de contactos y de amistad). Para medir la estructura de ambas redes utilizamos los modelos de redes de mundo pequeño y libres de escala (ver sección 2.7.2 y 2.7.3 respectivamente), las medidas de centralidad (de grado, de cercanía y de intermediación) para obtener las características estructurales de ambas redes.

Para nuestro segundo análisis, utilizamos el algoritmo DCRS propuesto en la sección 4.5 para obtener un conjunto de comunidades de la red de amistad estudiada en la primera parte. Se obtuvieron diferentes comunidades con la que pudimos escalar el análisis de la red de Flickr y se realizó un análisis de la evolución de la red de Flickr en el tiempo.

5.1.2 Proceso de muestreo de la red de Flickr

El proceso de muestreo para la presente tesis se realizó en dos etapas, en la primer etapa se realizó un muestreo inicial con un periodo de 3 meses (del 13 de Febrero de 2009 al 22 de Mayo de 2009) utilizando los algoritmos de rastreo descritos en la sección 3.2 aplicados a la red de Flickr[3]. Durante la segunda etapa se realizaron muestreos parciales [4] del grafo inicial

[3]Para ver las estadísticas del número de llamadas a la Base de Datos de Flickr ver el Apéndice A

[4]Entiendase como muestreo parcial a la aplicación de los algoritmos de muestreo en algunos nodos seleccionados de manera aleatoria.

durante 2 meses (del 15 de Junio al 20 de Agosto del 2009). Para hacer el muestreo de Flickr se utilizó una máquina con un procesador 2.0 GHz AMD Sempron y 2 GB de memoria RAM, y la implementación de los algoritmos de muestreo fueron desarrollados utilizando el API *flickrj* bajo el lenguaje de programación de Java. La parte del análisis de las redes sociales se realizó mediante el programa *GNU R* descrito en la sección 2.8.3.

5.1.3 La red de contactos y la red de amistad

Para este caso de uso se generaron dos redes dirigidas a partir del conjunto de datos del sistema Flickr. La primera la llamamos *red de contactos*, la cual está generada a partir de la lista de contactos de los usuarios utilizando el algoritmo de la sección 3.2.2. La segunda es una *red de amistad*, la cual relaciona a los usuarios mediante los comentarios de las fotos que comparten, se implementó el algoritmo propuesto en la sección 3.2.3.

En trabajos relacionados [87] se estudian las *redes de actividad*, las cuales son formadas a partir de la interacción de usuarios mediante el uso de mensajes (en este trabajo utilizaron Facebook para su análisis), tomando en cuenta las relaciones que se presentan en ambos sentidos generando un grafo no dirigido. El *grado de entrada* de las redes de actividad es más bajo que el de la red de contactos según [83]. Para nuestro caso el 33.82% de los usuarios que aparece en la red de contactos, también aparece en la red de amistad, esto refleja el nivel de participación entre los usuarios debido a que en un sistema como Flickr se pueden comentar fotos sin ser contacto de la persona que comparte la foto.

El conjunto de datos obtenidos de la *red de contactos* es de 11,796,457 usuarios y de 332,344,545 relaciones entre usuarios, correspondiente al 36.86% de usuarios totales de Flickr. Para la *red de amistad* se obtuvo un conjunto de datos reducido de 1,028,112 de usuarios (correspondiente al 11.47% de la red de contactos) y de 118,447,596 relaciones entre usuarios, el cual corresponde al 35.64% de la red de contactos, como se muestra en la tabla 5.1.

Tipo de Red	**Usuarios**	**Relaciones**
Red total	32,000,000	1,000,000,000 (*aprox.*)
Red de contactos	11,796,457	332,344,545
Red de amistad	1,028,112	118,447,596

Tabla 5.1: Conjunto de datos de la red de contactos y de la red de amistad de Flickr.

5.1.4 Propiedad del mundo pequeño para Flickr

Según el modelo de mundo pequeño de Watts-Strogatz (ver sección 2.7.2.1) para que una red cumpla con los requerimientos mínimos de estas redes debe cumplir lo siguiente:

- *La trayectoria promedio de todos los vértices en la red debe ser pequeña (según la teoría de los seis grados debe ser menor a 6).*
- *El coeficiente de agrupamiento global de la red debe ser alto.*
- *Un diámetro de red pequeño.*

El **coeficiente de agrupamiento local**[5] mide la relación que un usuario tiene con respecto a sus vecinos. Cuando el vecindario está completamente conectado el valor del coeficiente de agrupamiento global es 1, cuando el coeficiente de agrupamiento global vale 0 significa que el usuario está aislado y no posee ningún enlace con otro usuario.

El **coeficiente de agrupamiento global** por otro lado mide que tan conectada está una red, para nuestras redes obtuvimos los siguientes resultados: para la red de contactos el valor del CC_{global} fue de 0.418 y para la red de amistad el valor de CC_{global} fue de 0.281. Se puede apreciar que la red de amistad está menos conectada que la red de contactos debido a que posee menos enlaces, lo cual reduce la conectividad de la red.

La trayectoria más corta entre dos vértices es el camino con longitud mínima entre ellos, para la red de contactos y de amistad se calculó el promedio de la trayectoria más corta. El valor promedio de la trayectoria más corta para la red de contactos es de 5.15, mientras que para la red de amistad es de 5.31. Este tipo de redes presentan un modelo de *red de mundo pequeño*, esto prueba la famosa teoría de los seis grados de separación propuesta por Milgram[56].

El **diámetro** refleja el tamaño de nuestra red y se cálcula utilizando algoritmos de búsqueda a lo ancho, sin embargo, para redes a gran escala resulta un problema calcular este valor debido al costo computacional que este tipo de algoritmos implican, para la presente tesis utilizamos el algoritmo propuesto en [52] para obtener el valor del diámetro en cada una de nuestras redes. Para la red de contactos se obtuvo un valor de 18 y para la red de amistad con 23, se puede apreciar que el diámetro aumenta para la red de amistad debido a que existen menos enlaces en la red.

En la tabla 5.9 se resumen las propiedades de mundo pequeño encontradas para la red de contactos y de amistad.

5.1.5 Propiedad de libre escala para Flickr

El modelo de crecimiento más utilizado en las redes sociales es el de **Barabasi-Albert** (también llamado como **modelo BA**) [10], en el que la formación de nuevas relaciones está determinado por un **anexo preferencial** (ver sección 2.7.3.1). Para probar que una red sigue un anexo preferencial se debe de mostrar la distribución de grado de los nodos de la red.

[5]Para más detalle sobre como calcular el coeficiente de agrupamiento consultar la sección 2.7.2.1

Red	CC_{Global}	Trayectoría promedio	Diámetro
Flcikr contactos	0.418	5.15	18
Flickr amistad	0.281	5.31	23
LiveJournal [60]	0.330	5.88	20
Youtube [60]	0.136	5.10	21

Tabla 5.2: Propiedades de mundo pequeño para la red de contactos y de amistad para Flickr.

El **grado** de un nodo en nuestra red de contactos y de amistad representa el número de contactos con los que interactúa un usuario, el grado indica que tan importante o activo es el usuario dentro de la red. La **distribución de grado** de una red representa la ocurrencia de los usuarios al agregarse al sistema (*anexo preferencial*). En la figura 5.3 se puede apreciar la **distribución de grado** para la red de contactos y la red amistad obtenida durante el muestreo, en ambas se puede ver que una distribución *libre de escala* (ver sección 2.6.2.4).

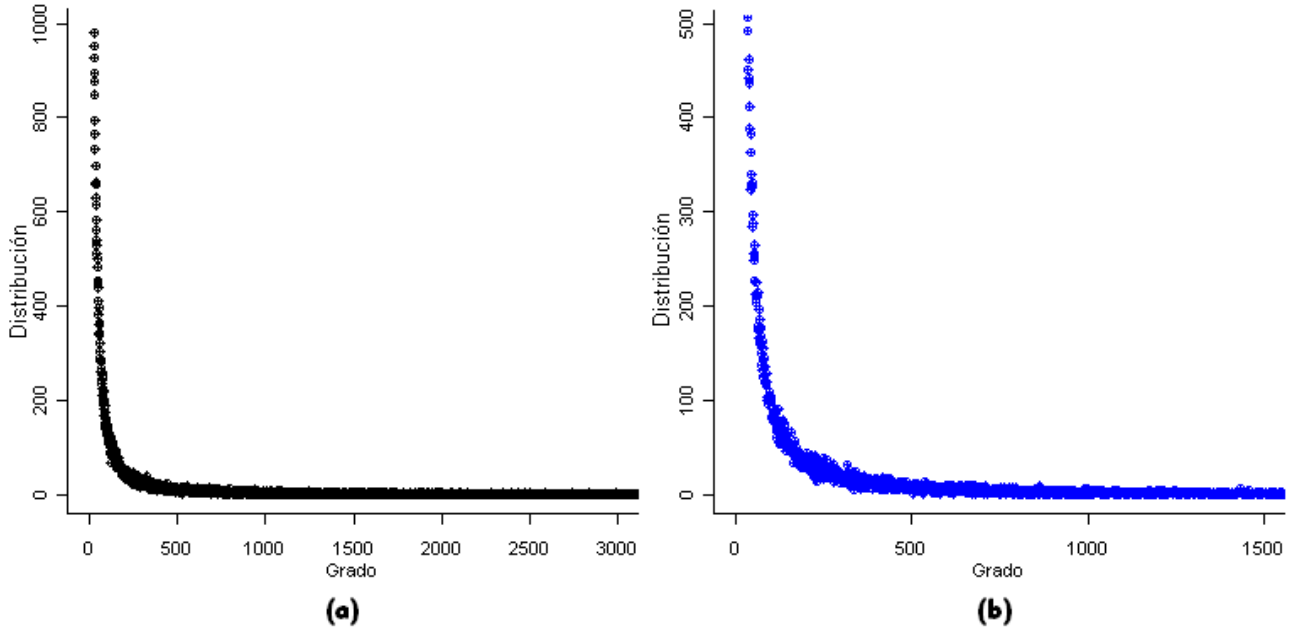


Figura 5.3: Distribución de grado para la red (a) de contactos y (b) de amistad de Flickr.

Sin embargo, trabajos relacionados [17, 59] utilizan los términos de *creación preferencial* y de *recepción preferencial*, para denotar a los mecanismos utilizados para la creación de enlaces en base a su grado de salida y en base a su grado de entrada, donde se analiza la formación de enlaces de entrada y de salida para un usuario en una red dirigida. En la figura 5.4(a) y (b) se muestra la distribución de grado de entrada y de salida respectivamente, y en las figuras 5.4(c) y (d) se muestra la gráfica log-log de la distribución de grado de salida y de entrada. Estos datos muestran que tanto la *creación preferencial* como la *recepción preferencial*

siguen un *anexo preferencial* y un *crecimiento exponencial*.

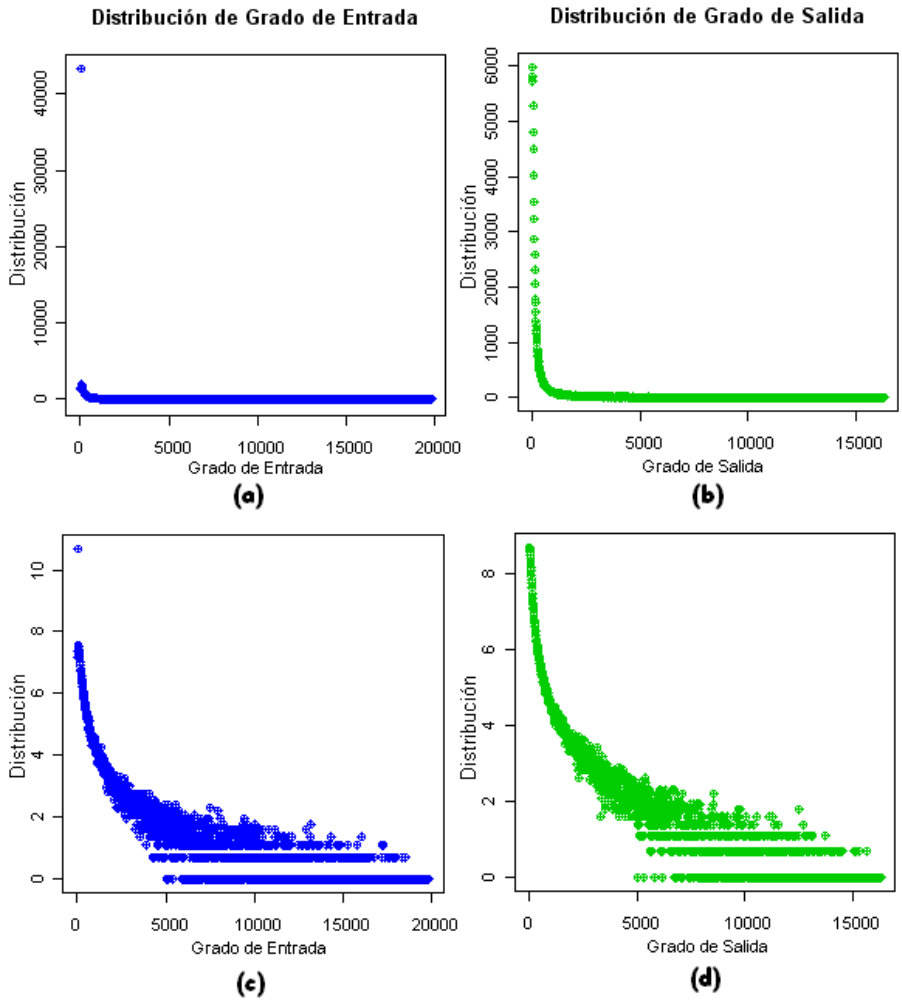


Figura 5.4: Distribución de grado de (a) salida y (b) entrada. Gráfica log-log de la distribución de grado (c) de entrada y (d) de salida

En la sección 2.6.2.4 se mostró que las redes libres de escala siguen una ley de potencia, en la tabla 5.3 se muestra el valor del **coeficiente** α para ambas redes, este valor se calculó mediante el método de máxima verosimilitud presentado en la sección 2.6.2.3, se seleccionó este método debido a que es utilizado en otros trabajos [60].

5.1.6 Características de alto nivel de la red de amistad

Como vimos en secciones anteriores, tanto la red de contactos como la de amistad cumplen con las propiedades de las redes mundo pequeño y libres de escala. La red de amistad posee características que la hacen aproximarse más a un comportamiento social, es decir, posee

Red	Grado de Entrada		Grado de Salida	
	α	D	α	D
Web [16]	2.67	-	2.09	-
Flickr contactos	1.93	0.0392	2.05	0.0034
Flcikr amistad	1.63	0.0592	1.77	0.0249
LiveJournal [60]	1.59	0.0783	1.65	0.1037
Youtube [60]	1.63	0.1314	1.99	0.0094

Tabla 5.3: Estimación para el valor del coeficiente de ley de potencia α y el valor correspondiente de Kolmogorov-Smirnov del método de mámxima verosimilitud para Flickr.

una estructura que permite cuantificar diferentes aspectos entre los usuarios, en la tabla 5.11 se muestran las estadísticas de alto nivel obtenidas para la red de amistad de Flickr durante el periodo de muestreo de los 6 meses (del 13 de Febrero de 2009 al 11 de Agosto del 2009).

Característica	Datos Iniciales	1er. mes	2do. mes	Datos Finales
Usuarios totales	11,468,173	11,512,547	11,593,325	11,796,457
Vínculos totales	276,173,495	295,394,452	311,295,626	332,344,545
Fotos totales	619,489,852	735,596,839	829,596,124	907,210,785
Comentarios Totales	1,796,939,112	1,911,260,085	2,123,253,643	2,408,735,357
Comentarios por fotos	3(2.9)	3(2.59)	3(2.55)	3(2.65)
Fotos por usuario	422	486	521	505
Usuarios con fotos	903,589	1,395,312	1,345,497	1,505,790
Comentarios por usuario	1,224	1,259	1,329	1,338
Usuarios que comentan	369,234	356,332	395,413	425,454

Tabla 5.4: Estadísticas obtenidas de Flickr.

El porcentaje de usuarios que *comparten fotos* es de 83.81% algo esperado tomando en cuenta la finalidad del sistema, el 16.91% de usuarios en la red *no tiene contactos asociados* (usuarios con grado de salida igual a cero), el 37.12% de los usuarios *no interactuan comentando fotos*, el 28.43% de usuarios *comentan una foto* en promedio cada 3 días, y el 38.43% de los usuarios *comparten fotos nuevas* cada 15 días, el 37.92% de las *fotos no tienen comentarios* asociados, el 24.12% de los *amigos de un usuario comentan sus fotos* en los primeros 7 días en que esta fue publicada.

La red de amistad permite analizar el dinamismo de la red con valores reales, cuando un usuario agrega contactos a su lista de contactos, o asocia contenido nuevo a su cuenta (fotos, videos, publicaciones, etc), o interactúa con otro usuario (p.ej., mediante el envío de mensajes usando wikis, blogs, aplicaciones, emails, etc), se le asocia la fecha en que se realizó tal acción. Este tipo de información permite estudiar a la red social a través del tiempo, en la figura 5.5 se aprecia la interacción de los usuarios sobre la red de Flickr por medio de los comentarios sobre las fotos publicadas.

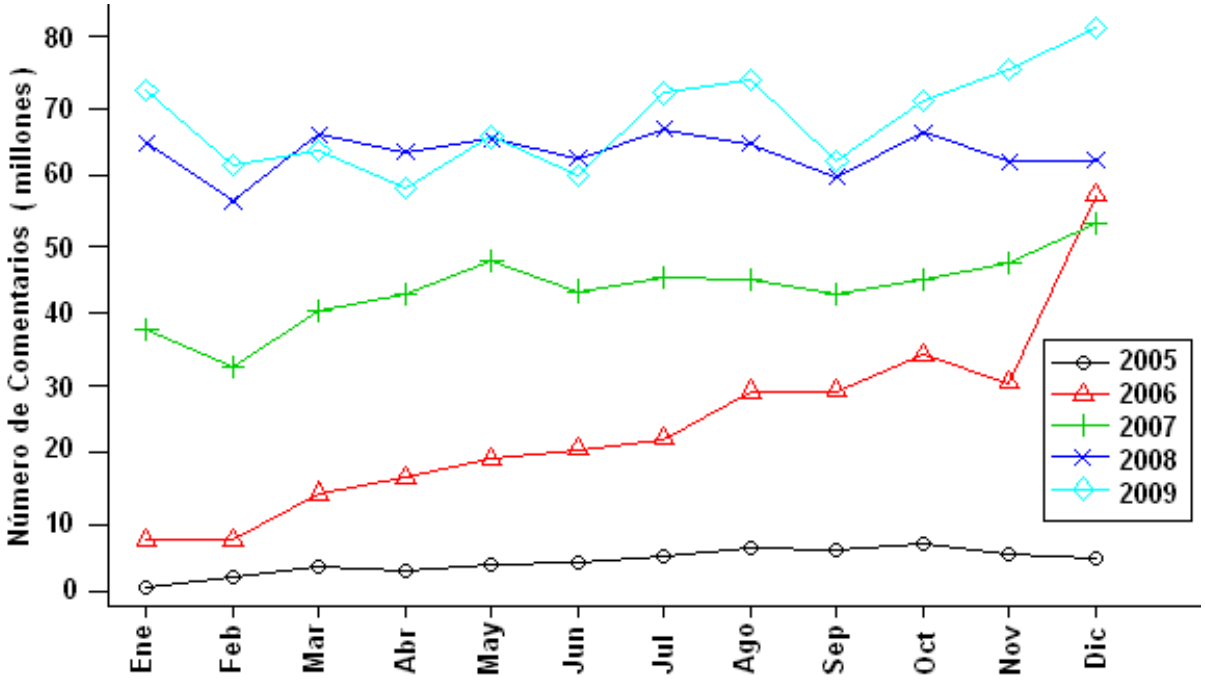


Figura 5.5: Distribución de comentarios para la red de Flickr

5.1.7 Detección de comunidades en la red de Flickr

En el capitulo 4.5 describimos el diseño de una algoritmo para la detección de comunidades para redes sociales a gran escala, para este caso utilizamos la red de amistad descrita en secciones anteriores, para aplicar el algoritmo de detección de comunidades DCRS para la red de Wikipedia se utilizó una máquina con un procesador AMD Sempron a 2.0 GHz y 2 GB de memoria RAM, la implementación del algoritmo se hizo en java y se utilizó MySQL para almacenar la información, el tiempo que se tomó para obtener las comunidades fue de 5 días aproximadamente, se encontaron 12 comunidades para la primera fase del algoritmo, donde el Componente Gigante de la red (ver sección 2.6.1) posee el 33.32% de los nodos totales de la red y el resto se encuentra repartido en las otras comunidades.

En la tabla 5.5 se muestran los resultados obtenidos para diferentes valores de θ de la fase dos del algoritmo DCRS, donde se puede apreciar que para valores de θ altos la red tiende a agregar nodos en las comunidades. Sin embargo, el diámetro de la red disminuye debido a que la red se encuentra más conectada.

Característica	θ=0.07	θ=0.14	θ=0.21	θ=0.28	θ=0.35	θ=0.42
Vínculos perdidos	26.41%	22.8%	18.2%	16.5%	15.4%	12.9%
Diámetro del Componente Gigante	17	16	14	13	12	12
Diámetro promedio sin Componente Gigante	20	19	18	18	17	17
Nodos Compartidos	545	675	721	945	1,273	1,492
Grado promedio	89	94	98	103	107	112
Centralidad promedio	0.462	0.474	0.485	0.493	0.514	0.523
Intermediación de vértice promedio	245	284	321	415	473	492
Intermediación de enlace promedio	143	152	183	201	236	248

Tabla 5.5: Resultados para el algoritmo de detección de comunidades en Flickr. Con diferentes valores de θ

5.2 El caso Wikipedia

El conjunto de datos de Wikipedia está dividido en diferentes idiomas, para la presente tesis analizamos el conjunto de datos de Wikipedia en español, el cual está formado por más de 500,000 artículos y por más de 1 millón de usuarios. El conjunto de datos es proporcionado por Wikipedia y podemos decir entonces que este análisis estudia una red social completa para un caso en particular como es la *Wikipedia en español*.

Actualmente, Wikipedia cuenta con 265 versiones en diferentes idiomas, cada versión trabaja de forma independiente y en total se suman 14 millones de artículos[6]. La versión en inglés es la que cuenta con más artículos (más de 11 millones) y más usuarios (más 19 millones). En la versión en inglés es necesario crear una cuenta para poder editar o crear contenido, para las otras versiones se puede editar contenido como un usuario anónimo.

Para este caso de estudio se realizó una comparativa con las diferentes Wikipedias que existen, posteriormente se realizó un estudio basado en el análisis de redes sociales como el mostrado en el caso de Flickr de la sección 5.1 para el caso de la Wikipedia en español.

5.2.1 Análisis de la red de Wikipedia en Español

En la tabla 5.6 se muestran las estadísticas de las Wikipedias más importantes[7]. La versión en español presenta números bajos tomando en cuenta el número de hispanohablantes.

[6]Para más información consultar `http://es.wikipedia.org/wiki/Wikipedia`

[7]Datos obtenidos de `http://meta.wikimedia.org/wiki/List_of_Wikipedias` actualizados hasa el 22 de septiembre del 2009

Lenguaje	Fecha de Inicio	Número de Páginas	Usuarios	Número de Artículos	Número de Revisiones
Inglés	Enero 2001	3,175,812	19,278,488	11,524,227	363,986,477
Aleman	Mayo 2001	1,017,175	2,905,509	918,870	72,823,066
Frances	Marzo 2001	906,242	3,639,272	758,236	52,502,495
Polaco	Septiembre 2001	670,867	1,181,122	334,581	21,113,293
Itliano	Enero 2002	652,818	2,018,164	472,144	32,923,281
Japones	Septiembre 2002	649,328	1,687,524	385,112	30,788,967
Holandes	Junio de 2001	584,375	1,389,420	290,370	19,877,967
Español	Mayo 2001	556.677	2,232,896	1,351,627	24,769,376
Portugues	Junio 2001	540,229	2,104,977	675,116	18,828,415
Ruso	Mayo 2001	487,647	1,747,484	419,073	22,849,456
Sueco	Mayo 2001	344,758	892,584	154,950	11,238,516
Chino	Octubre 2002	292,349	934,985	778,323	12.318,907

Tabla 5.6: Datos estadísticos generales de las más importantes versiones de Wikipedia.

En la tabla 5.6 la Wikipedia en español se encuentra en la posición 8 en base al número de páginas publicadas, ocupa el cuarto lugar en cuestión del número de revisiones sobre artículos y el quinto lugar con mayor número de usuarios. En la figura5.6 se presenta la evolución de artículos para las distintas versiones de Wikipedia.

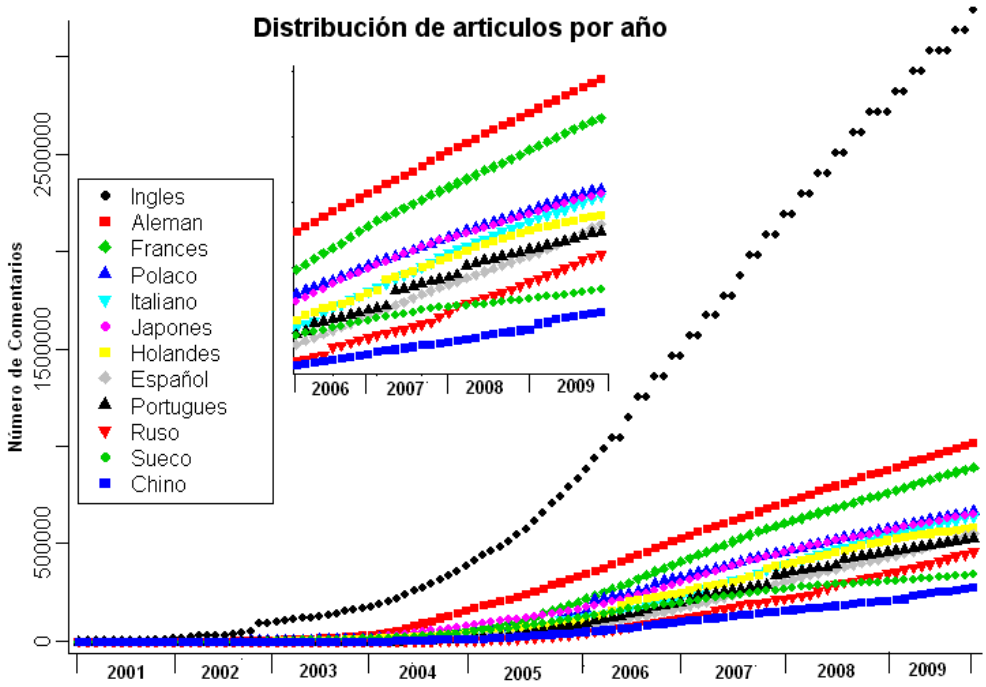


Figura 5.6: Evolución del número de artículos para diferentes idiomas de la Wikipedia.

En la figura 5.7 se muestra la evolución de las revisiones sobre artículos para las Wikipedias más importantes. Se puede apreciar que la versión en inglés está por encima de cualquier otra versión de la Wikipedia y presenta un crecimiento acelerado en comparación a las otras versiones. Sin embargo, estudios recientes [73] hablan sobre una posible freno en el crecimiento en la Wikipedia sobre todo en la versión en inglés, debido a que la información contenida en esta versión está llegando a un máximo, y en comparación a las otras versiones la Wikipedia en inglés ofrece contenido supervisado, debido a que cierta parte de su contenido es editado por personas especializadas y controladas por la propia *Fundación Wikipedia*, algo que supone un número de ediciones y creaciones de artículos limitadas en un futuro.

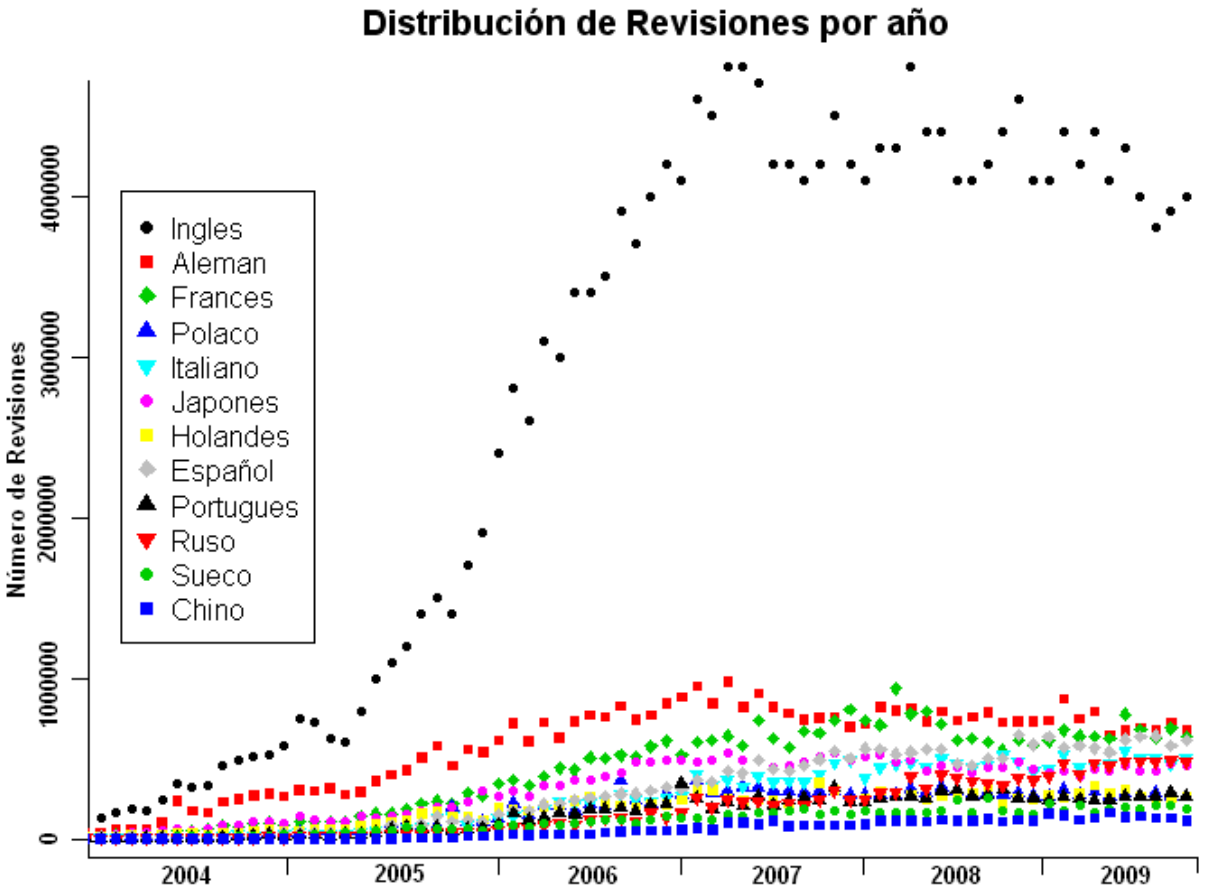


Figura 5.7: Evolución del número de revisiones para diferentes idiomas de la Wikipedia.

5.2.2 Datos de la red de Wikipedia en español

En base a la estructura de Wikipedia (ver sección 3.3) se generó una red no dirigida, como las redes de colaboración a pequeña escala estudiadas en trabajos anteriores [64]. Para el proceso de muestreo se utilizó el algoritmo de la sección 3.3 el cual nos permite optimizar el

tamaño de la Wikipedia. La red completa está formada por 2,232,896 usuarios y 24,769,376 de relaciones entre usuarios, la tabla 5.8 muestra la distribución de la red a través del tiempo.

Característica	**Wikipedia en Español**							
	2002	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**
Nodos	1,342	10,412	193,131	285,012	414,143	482,371	462,221	384,121
Vínculos	9,572	74,735	334,245	1,288,449	3,855,769	6,960,674	8,388,061	3,857,340

Tabla 5.7: Estadísticas de la red general de Wikipedia

Para minimizar el tamaño de la red aplicamos el algoritmo de amistad, donde obtuvimos los datos estadísticos generales del conjunto de datos de la Wikipedia en español reducida a través del tiempo mostrados en la tabla 5.8.

Característica	**Wikipedia en Español**							
	2002	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**
Nodos	26	272	3,258	5,483	8,443	9,223	10,374	10,154
Vínculos	4,572	43,522	113,742	131,132	145,832	362,632	425,314	304,817

Tabla 5.8: Estadísticas de la red optimizada de Wikipedia.

5.2.3 Propiedad del mundo pequeño para Wikipedia en español

Como lo mostramos en el caso de Flickr, seguimos el modelo de Watts-Strogatz (ver sección 2.7.2.1) para poder analizar a la red se como una red de mundo pequeño. En la tabla **??** se resumen las propiedades de mundo pequeño encontradas para la red de Wikipedia en español para el conjunto de datos comleto y el conjunto reducido.

Red	CC_{Global}	Trayectoría promedio	Diámetro
Wikipedia completa	0.363	5.26	24
Wikipedia reducida	0.257	5.79	26
LiveJournal [60]	0.330	5.88	20
Youtube [60]	0.136	5.10	21

Tabla 5.9: Propiedades de mundo pequeño para la red de contactos y de amistad para Wikipedia.

Se puede observar que igual que en el caso de Flickr para la red optimizada la trayectoría promedio aumenta debido a que existen menos *caminos* por los cuales fluya la información y por lo tanto el diámetro de la red es alto. También se puede observar que el coeficiente de agrupamiento en la red de Flickr es más alto que la de Wikipedia, esto es debido a que la red de Flickr posee más información que la red de Wikipedia. Sin embargo, se puede notar que en ambos casos de estudio (Flickr y Wikipedia) la optimización de la red sigue manteniendo la propiedad de mundo pequeño.

5.2.4 Propiedad de libre escala para la Wikipedia

En el caso de uso de la red de Flickr utilizamos el modelo de crecimiento BA, ya que con este modelo podemos medir la forma en como la red se distribuye. Sin embargo, para la red generada para Wikipedia no es necesario medir la distribución de grado de entrada y de salida de la red debido a que la red de Wikipedia es una red no dirigida y el grado de entrada y de salida para la red es igual. En la figura 5.8 se puede apreciar la **distribución de grado** para la red de la Wikipedia en español, se puede apreciar una distribución *libre de escala* (ver sección 2.6.2.4).

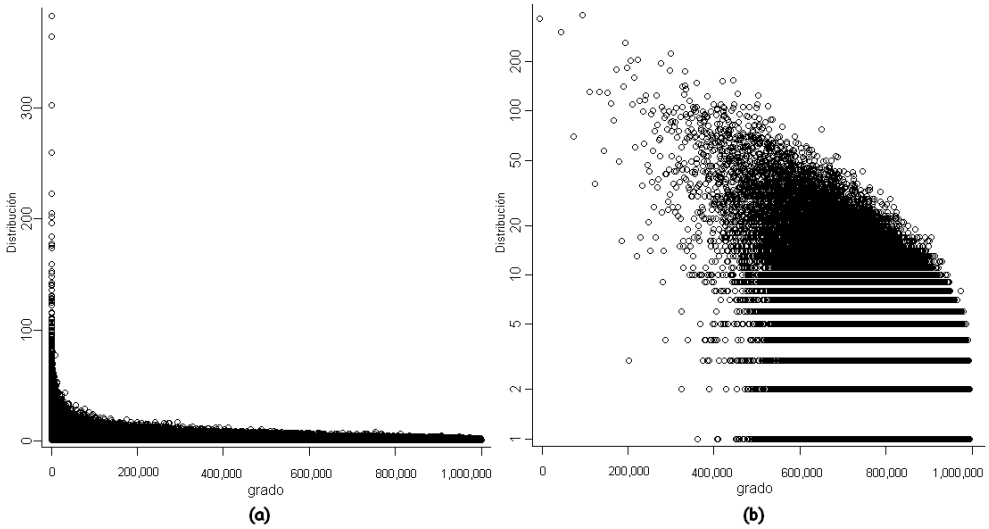


Figura 5.8: (a) Distribución de grado y (b) gráfica log-log de la distribución de grado de la Wikipedia en español.

En la tabla 5.10 se muestra el valor del coeficiente α para ambas redes, al igual que en el caso de Flickr se utilizó el método de máxima verosimilitud presentado en la sección 2.6.2.3 para el cálculo del coeficiente α.

Wikipedia	**Grado**	
	α	D
Completa	2.43	0.1412
Reducida	2.71	0.1567

Tabla 5.10: Estimación para el valor del coeficiente de ley de potencia α y el valor correspondiente de Kolmogorov-Smirnov del método de mámxima verosimilitud para la red completa y optimizada de la Wikipedia en Español.

5.2.5 Características de alto nivel de la red de Wikipedia

Como vimos en secciones anteriores, tanto el conjunto de datos completo y el reducido cumplen con las propiedades de las redes mundo pequeño y libres de escala. En la tabla 5.11 se muestran las estadísticas de alto nivel obtenidas para la red formada por el conjunto de datos reducido de la Wikipedia en español.

Característica	2004	2005	2006	2007	2008	2009
Wikipedistas activos	306	1,144	2,546	3,845	3,934	4,237
Número de artículos	25,000	54,000	347,000	253,000	388,000	497,000
Colaboladores	705	2,340	6,994	19,506	31,925	45,245
Ediciones por artículo	8	15	23	29	36	41

Tabla 5.11: Estadísticas obtenidas de Flickr.

5.2.6 Detección de comunidades en la red de Wikipedia en español

Para aplicar el algoritmo de detección de comunidades DCRS para la red de Wikipedia se utilizó una máquina con un procesador AMD Sempron a 2.0 GHz y 2 GB de memoria RAM, la implementación del algoritmo se hizo en java y se utilizó MySQL para almacenar la información, el tiempo que se tardó en ejecutarse el algoritmo fue de 4 días aproximadamente, se encontraron 9 comunidades para la primera fase del algoritmo, donde el Componente Gigante de la red (ver sección 2.6.1) posee el 35.15% de los nodos totales de la red y el resto se encuentra repartido en las otras comunidades. En la tabla 5.12 se muestran los resultados obtenidos para diferentes valores de θ de la fase dos del algoritmo DCRS en el caso de la Wikipedia en español.

Característica	θ=0.07	θ=0.14	θ=0.21	θ=0.28	θ=0.35	θ=0.42
Vínculos perdidos	32.6%	31.7%	30.4%	28.8%	27.41%	25.9%
Diámetro del Componente Gigante	13	12	12	11	10	10
Diámetro promedio sin Componente Gigante	16	13	11	11	10	10
Nodos compartidos	173	194	224	236	265	281
Grado promedio	36	40	44	49	53	58
Centralidad promedio	0.312	0.351	0.375	0.415	0.456	0.491
Intermediación de vértice	194	211	231	249	289	327
Intermediación de enlace	76	81	95	118	126	157

Tabla 5.12: Resultados para el algoritmo de detección de comunidades en Wikipedia. Con diferentes valores de θ

5.3 Discusión

En esta capítulo se realizó el análisis de dos diferentes sistemas de redes sociales, donde se pudo observar que las dos redes sociales comparten características muy similares entre sí. El algoritmo de detección de comunidades nos permitió obtener comunidades de la red inicial y poder realizar mediciones.

Para ambas redes se puede apreciar un crecimiento siguiendo un **anexo preferencial**, donde los usuarios en el caso de *Flickr* tienden a agregarse con aquellos que comparten más contenido y para el caso de *Wikipedia* la edición de artículos está determinada por un cierto grupo de usuarios que por lo general editan la mayoría del contenido. Sin embargo, el crecimiento de ambas redes es diferente, ya que para el caso de *Wikipedia* llegará un momento en el que la edición de los artículos presentará un crecimiento lento debido a que existirá muy poco contenido el cual modificar. Para el caso de *Flickr* el número de usuarios no parece ser tan importante, algo que si es más importante es la interacción entre usuarios mediante la compartición de fotografías, ya que esto permitirá prolongar el ciclo de uso del sistema.

En *Wikipedia* y *Flickr* se pueden apreciar algunas diferencias en su estructura, la primera es que para el caso de *Wikipedia* se genera una red no dirigida y para *Flickr* se genera una red dirigida, esto permite que la red de *Wikipedia* este mejor conectada que la red de *Flickr*. Sin embargo, la red de Flickr representa mejor el comportamiento de las redes de mundo real, ya que la interacción que se ofrece en este sistema está basada en la comunicación entre usuarios.

Capítulo 6

Resultados, conclusiones, y trabajo a futuro

6.1 Resultados

En este trabajo se abordó el problema del análisis de las redes sociales para conjuntos a gran escala. Se utilizó la teoría de grafos como la herramienta matemática con la cual podemos analizar la estructura de este tipo de redes. Para desarrollar este trabajo de tesis se realizó una investigación de los distintos métodos de la teoría de grafos que son aplicados a las redes sociales.

Se mostraron diferentes sistemas que permiten extraer datos de la Web (*Flink* y *POLYPHONET*), así también se mostraron las técnicas de muestreo aplicadas a la extracción de información de los sistemas de redes sociales en línea. Utilizando un muestreo de **bola de nieve** implementamos un algoritmo que nos permitió extraer información del sistema *Flickr* durante 6 meses (del 13 de febrero al 20 de Agosto del 2009).

Para poder obtener una mejor representación de nuestro conjunto de datos adaptamos el muestreo antes mencionado para que pudieramos medir la interacción entre los usuarios. En base a estos dos muestreos se generó para el caso de *Flickr* una red de contactos y una red de amistad que utiliza la frecuencia de comentarios sobre las fotos compartidas por usuarios del sistema, para estas redes se realizó un estudio comparativo en el cual se pudó observar que ambas redes presentan una estructura muy parecida.

Para el caso de *Wikipedia* no se utilizó un proceso de muestreo debido a que *Wikipedia* presenta una estructura diferente a la de *Flickr*. *Wikipedia* permite obtener una copia del conjunto de datos de diferentes versiones del sistema, esto nos permitió hacer una red de colaboración entre usuarios que editan artículos dentro del sistema. De igual manera que en el caso de *Flickr* implementamos un algoritmo para reducir el conjunto de datos originales en base a la frecuencia de colaboración de los usuarios dentro del sistema, en base a esto se realizó un estudio comparativo de ambos conjuntos de datos.

Uno de los principales problemas del análisis de redes sociales a gran escala es que estas estructuras no pueden ser escaladas fácilmente. En base a las técnicas de *agrupamiento en grafos* se realizó un estudio sobre los diferentes métodos aplicados a las redes sociales, en especial al problema de la *detección de comunidades* en redes sociales. Mediante el uso del algoritmo de detección de comunidades basado en etiquetas propuesto por *Raghavan* [77], se adaptó un algoritmo basado en el problema del escalamiento de redes sociales y haciendo uso del concepto de comunidades traslapadas. Este algoritmo fue utilizado en nuestros casos de estudios y nos permitió escalar a nuestras redes de *Flickr* y de *Wikipedia*, para posteriormente analizar a cada una de las comunidades encontradas.

Como resultado principal de este trabajo de tesis, se presentó un análisis de las redes sociales de *Flickr* y *Wikipedia*, en base a su estructura como red mundo pequeño y red libre de escala, así también se analizó la importancia de ambas redes en base a sus medidas de centralidad (de grado, de cercanía y de intermediación). En general, se pudó apreciar que tanto *Flickr* como *Wikipedia* poseen una estructura muy parecida, pese a que ambas ofrecen distintos servicios y son representadas de diferente manera.

6.2 Discusión

Durante este trabajo de tesis nos enfocamos en estudiar el problema de las **redes sociales a gran escala**, para resolver este problema se estudiaron las diferentes herramientas que provee el **análisis de redes sociales**, las cuales utilizan la *teoría de grafos* y el *cálculo matricial* para analizar la estructura de las redes sociales.

Para nuestro estudio necesitabamos un conjunto de datos que nos permitiera analizar a las redes sociales a gran escala, la solución al problema consistió en aplicar las técnicas de extracción de información en sistemas de redes sociales en línea como *Facebook, Twitter, Wikipedia*, entre otros; los cuales debian proporcionarnos un conjunto de datos con el cual pudieramos trabajar fuera de línea y aplicar las técnicas del *análisis de reds sociales*. En años recientes los trabajos de investigación sobre conjuntos de datos grandes han ido en aumento que permiten analizar la estructura de estas redes, sin embargo, la mayoría de estos conjuntos de datos proporcionan información cuantitativa y para nuestro estudio necesitabamos información que detallara la interacción de los usuarios de la red.

Debido a que necesitabamos estudiar las propiedades de las redes sociales a gran escala, seleccionamos dos sistemas en línea (*Flickr* y *Wikipedia*). Se utilizó *Flickr* debido a que es la red social más importante en la compartición de fotografías y para nuestro caso de estudio buscabamos una red que permitiera medir la interacción entre usuarios. Otro detalle importante sobre *Flickr* es que pertenece a *Yahoo* y todo aquel que tenga una cuenta en este servicio de correos tiene asociada una cuenta de *Flickr*, pese a que no es un sistema muy conocido se espera que *Flickr* en un futuro tome importancia debido a su potencial de usuarios. En

el caso de *Wikipedia* tenemos una red social que está en constante crecimiento y que provee información útil para el análisis de la interacción entre personas, sin embargo, estudios [73] prueban que esta red está próxima de alcanzar su nivel de crecimiento, un fenómeno que se está viendo en diferentes versiones de la *Wikipedia* (p.ej., la versión inglesa, alemana y holandesa).

En el proyecto *OpenSocial* de *Google* se pretende unificar diferentes sistemas de redes sociales en línea con el objetivo de estandarizar la búsqueda de información en los sistemas en línea. Actualmente, este proyecto cuenta con 35 sistemas y provee soporte para redes sociales como son: *CyWorld* (Korea), *YiQi* (China), *Ning*(Estados Unidos), *IDtail*(Korea), *Freebar*(Italia), *MySpace* (en ese momento la red social más importante de la Web y que se incorpora en octubre del 2008), entre otros; muchas de estas redes son regionales, es decir, solo son usadas en ciertas partes del mundo y para nuestro estudio buscabamos algo más global como *Flickr* y *Wikipedia*. En los inicios de nuestro trabajo de tesis (hace un año) *OpenSocial* no incluía tantos sistemas y los pocos con los que contaba carecían de información, redes como Twitter no tenian la presencia e importancia como la tiene en estos momentos.

Debido a que los servicios de redes sociales poseen una gran cantidad de información, es difícil que un proceso de muestreo capture toda la información del sistema. En los sistemas de *redes sociales en línea* existen usuarios que nunca interactúan con otros usuarios del sistema aunque los tengan agregados en su lista de contactos y en otros casos los usuarios ni siquiera usan el sistema para interactuar con los demás. Motivados por este fénomeno se desarrolló un muestreo basado en la interacción entre usuarios, el cual reduce el tamaño del conjunto de datos y ofrece una mejor representación de la red.

Dentro de los principales problemas del *análisis de redes sociales a gran escala* está el estudio de su estructura, esto debido a que el número de datos es muy grande y no es sencillo realizar cálculos sobre este conjunto de datos, esto nos motivó a estudiar la forma de escalar la red y nos encontramos con el problema de la **detección de comunidades** aplicada a las *redes sociales*. Para este problema tuvimos que estudiar los métodos de la teoría del *agrupamiento en grafos*, el cual se encarga de dividir a la red en base a los diferentes conceptos de la *teoría de grafos* (p.ej., cliques, cadenas de markov, centralidad, camino aleatorio, etc), sin embargo, el problema de la *detección de comunidades* utiliza estas técnicas para proponer algoritmos que permitan encontrar comunidades dentro de las redes sociales.

En base a un estudio sobre los métodos para la detección de comunidades se buscó un algoritmo que pudiera ser utilizado con un conjunto de datos grandes. Se implementaron y se probaron diferentes algoritmos para resolver el problema sobre conjuntos de datos reales y sintéticos, entre los métodos estudiados se implementó el método de *Girvan-Newman*, el cual propone un costoso cálculo de intermediación para todas las relaciones de la red en cada proceso iterativo, el método basado en cliques que es aún más restrictivo y costoso que el anterior, ya que utiliza algoritmos con complejidad NP-difícil para obtener los cliques de la red. Un método que nos pareció interesante por sus pocos cálculos sobre los elementos

de la red y por su eficiencia, fué el **método de propagación de etiquetas** de *Raghavan* [77], el cual tiene una complejidad casi lineal y se comporta muy bien para conjuntos de datos grandes.

En nuestro trabajo de tesis pudimos observar que el *algoritmo de propagación de etiquetas* cumple con las condiciones idoneas para detectar comunidades en una red social a gran escala, en el caso que se quiera separar al conjunto de datos en comunidades independientes unas de otras. Sin embargo, las personas tienden a estar en más de una comunidad o grupos sociales, este fenómeno nos motivo a estudiar el problema de *comunidades traslapadas*, el cual consiste en estudiar a los usuarios que forman parte en más de una comunidad. En un estudio que realizamos sobre las diferentes métricas del *análisis de redes sociales* aplicadas al problema de *comunidades traslapadas*, obtuvimos como resultado que el **coeficiente de agrupamiento local** de un nodo en la red permitía agregar a un usuario en más de una comunidad debido a que esta métrica mide la relación de un nodo con respecto a sus vecinos, en base a esto y al *algoritmo de propagación de etiquetas* pudimos resolver el problema de *comunidades traslapadas* juntando ambos conceptos en un algoritmo que nos permitierá distribuir usuarios en diferentes comunidades.

Respecto a los resultados obtenidos en las pruebas realizadas para la red social de *Flickr* y de *Wikipedia*, podemos observar que ambas redes presentan el fenómemo de mundo pequeño y que siguen un comportamiento como libres de escala debido a que presentan un *anexo preferencial* en su formación. Tanto *Flickr* como *Wikipedia* presentan un *componente gigante*, el cual posee un porcentaje alto de los usuarios y las relaciones existentes de la red total.

6.3 Conclusiones

Esta tesis ha presentado un análisis de la estructura de dos diferentes redes sociales *Flickr* y *Wikipedia*, en base a los conceptos del *análisis de redes sociales* y utilizando la *teoría de grafos*. Nuestros resultados muestran que las redes sociales son estructuralmente parecidas, aunque existen algunas diferencias en su forma de ser representadas, ya que estas redes pueden ser vistas como grafos dirigidos y no dirigidos. Dicha representación permite modelar a las redes sociales a gran escala y comparar su estructura con las redes del mundo real, dichas redes pueden ser construidas utilizando diferentes modelos, los cuales fueron analizados en esta tesis.

La forma en como se puede representar una red es muy importante, dentro de la tesis se implementó un algoritmo basado en las técnicas de muestreo y en la interacción entre usuarios aplicado en los sistemas de redes sociales en línea, el cual reduce el tamaño de la red y permite tener una mejor representación de la red. La implementación del algoritmo de extracción de información puede ser aplicado a diferentes sistemas de redes sociales en línea que permitan la interacción entre personas mediante el uso de comentarios en blogs,

compartición de archivos, mensajes privados, etc.

La forma en como se estructuran las redes sociales permite utilizar los algoritmos para detección de comunidades con traslapamiento para escalar el *análisis de redes sociales a gran escala*, en esta tesis se estudiaron los diferentes métodos para la detección de comunidades, destacando el algoritmo de detección de comunidades mediante la *propagación de etiquetas*, el cual permite encontrar comunidades muy rapidamente.

Por último, las aportaciones que se tuvieron con este trabajo de tesis son:

- Un estudio sobre el análisis de redes sociales y su aplicación sobre conjuntos de datos grandes.
- La implementación de un algoritmo basado en técnicas de muestreo y en la interacción entre usuarios, el cual permite obtener un conjunto de datos representativo de la red de *Flickr* y *Wikipedia*.
- Un conjunto de datos para trabajos futuros asociados con el análisis de redes sociales a gran escala sobre la red de *Flickr* y *Wikipedia*, proporcionando información cuantitativa y cualitativa de los individuos dentro de la red.
- La adaptación de un algoritmo que permite escalar el análisis de las redes sociales a gran escala en base al traslapamiento de comunidades y del uso de un algoritmo para detectar comunidades disjuntas basadas en el método de la propagación de etiquetas.
- Un estudio comparativo de la estructura de los servicios de redes sociales en línea *Flickr* y *Wikipedia* basado en las técnicas del *análisis de redes sociales*.

6.4 Trabajo a Futuro

Las redes sociales en línea son estructuras que estan en constante crecimiento y generalmente son muy distintas entre sí, proyectos como *OpenSocial* proponen estándares para unificar diferentes proyectos de redes sociales en línea, trabajos de investigación en un futuro dependerán en gran medida en la forma en como los sistemas de redes sociales en línea interactúen entre ellos.

Como parte del trabajo a futuro proponemos una adaptación del algoritmo para detectar comunidades en redes sociales a gran escala presentado en la tesis, el cual trabaje en línea y que nos permita búscar contenido en los diferentes servicios de redes sociales en línea. Por ejemplo, utilizar redes sociales como *LinkedIn* o *Facebook* para hallar grupos de personas que cumplan con un perfil para un trabajo, o empresas que ofrezcan trabajo a un grupo determinado de personas.

Los servicios de redes sociales en línea proporcionan información necesaria para el estudio de la estructura de las redes sociales y su relación con las redes del mundo real, donde las comunidades virtuales están tomando gran importancia en la vida de las personas y en la forma como se propaga la información, por ejemplo, el caso de *Twitter* ha planteado una nueva forma de difundir información a través de la Web, algo que sin lugar a dudas nos motiva a estudiar en un futuro este tipo de redes.

Un trabajo a futuro iteresante es analizar a las redes sociales que han ido perdiendo popularidad dentro de la Web, sitios como *MySpace* y *Hi5* en años recientes contaban con unas de las redes más prometedoras dentro de la Web, sin embargo, ahora se han visto desplazadas por sistemas como Facebook y Twitter. Tal vez, mucho de este comportamiento tenga que ver con la incorporación de nuevas tecnologías y el diseño de nuevas aplicaciones utilizadas a estos sistemas. Sin embargo, este tipo de servicios en un futuro formaran parte de una nueva generación de la Web, en donde los nuevos sistemas utilicen la información existente para poder crear nuevas aplicaciones con un contenido semántico de los datos.

Finalmente, se pretende desarrollar un par de artículos sobre nuestros casos de estudios, el primero en base al algoritmo para optimizar el tamaño de nuestras redes sociales en relación a la interacción entre usuarios y el segundo basado en los resultados del algoritmo para la detección de comunidades traslapadas utilizado para escalar el análisis de nuestras redes sociales a gran escala.

Bibliografía

[1] Lada A. Adamic. The small world web. In *ECDL '99: Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries*, pages 443–452, London, UK, 1999. Springer-Verlag.

[2] Lada A. Adamic, Bernardo A. Huberman;, A. Barabási, R. Albert, H. Jeong, and G. Bianconi;. Power-law distribution of the world wide web. *Science*, 287(5461):2115a+, March 2000.

[3] Eytan Adar. Guess: a language and interface for graph exploration. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 791–800, New York, NY, USA, 2006. ACM.

[4] Yong-Yeol Ahn, Seungyeop Han, Haewoon Kwak, Sue Moon, and Hawoong Jeong. Analysis of topological characteristics of huge online social networking services. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 835–844, New York, NY, USA, 2007. ACM Press.

[5] Reka Albert. Scale-free networks in cell biology. *J Cell Sci*, 118(21):4947–4957, November 2005.

[6] P. Anderson. What is web 2.0? ideas, technologies and implications for education. Technical report, February 2007.

[7] Elias Athanasopoulos, A. Makridakis, S. Antonatos, D. Antoniades, Sotiris Ioannidis, K. G. Anagnostakis, and Evangelos P. Markatos. Antisocial networks: Turning a social network into a botnet. In *ISC '08: Proceedings of the 11th international conference on Information Security*, pages 146–160, Berlin, Heidelberg, 2008. Springer-Verlag.

[8] Randy Baden, Adam Bender, Neil Spring, Bobby Bhattacharjee, and Daniel Starin. Persona: an online social network with user-defined privacy. *SIGCOMM Comput. Commun. Rev.*, 39(4):135–146, 2009.

[9] A. L. Barabási, H. Jeong, R. Ravasz, Z. Neda, T. Vicsek, and T. Schubert. Evolution of the social network of scientific collaborations. *arXiv:cond-mat*, 0104162, 2001.

[10] Albert-Laszlo Barabasi and Reka Albert. Emergence of scaling in random networks, October 1999.

[11] Albert-Laszlo Barabasi, Erzsebet Ravasz, and Tamas Vicsek. Deterministic scale-free networks, Feb 2002.

[12] Fabricio Benevenuto, Tiago Rodrigues, Virgilio Almeida, Jussara Almeida, Chao Zhang, and Keith Ross. Identifying video spammers in online social networks. In *AIRWeb '08: Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, pages 45–52, New York, NY, USA, 2008. ACM.

[13] Tanya Y. Berger-Wolf and Jared Saia. A framework for analysis of dynamic social networks. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 523–528, New York, NY, USA, 2006. ACM Press.

[14] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424(4-5):175–308, February 2006.

[15] Ulrik Brandes. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25:163–177, 2001.

[16] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure in the web: experiments and models. In *Proceedings of the Ninth International World-Wide Web Conference (WWW9, Amsterdam, May 15-19, 2000-Best Paper)*. Foretec Seminars, Inc. (of CD-ROM), Reston, VA, 2000.

[17] A. Capocci, V. D. P. Servedio, F. Colaiori, L. S. Buriol, D. Donato, S. Leonardi, and G. Caldarelli. Preferential attachment in the growth of social networks: The internet encyclopedia wikipedia. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 74(3), 2006.

[18] Meeyoung Cha, Alan Mislove, Ben Adams, and Krishna P. Gummadi. Characterizing social cascades in flickr. In *WOSP '08: Proceedings of the first workshop on Online social networks*, pages 13–18, New York, NY, USA, 2008. ACM.

[19] Fan R. K. Chung. *Spectral Graph Theory*.

[20] Aaron Clauset, Cosma R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661+, Feb 2009.

[21] R. Cohen, D. Avraham, and S. Havlin. Structural properties of scale-free networks, 2002.

[22] Elizabeth Costenbader and Thomas W. Valente. The stability of centrality measures when networks are sampled. *Social Networks*, 25(4):283–307, October 2003.

[23] Derek J. de Solla Price. Networks of scientific papers. *Science*, 149(3683):510–515, July 1965.

[24] Bender S. Demoll and D. Mcfarland. The art and science of dynamic network visualization. *JoSS: Journal of Social Structure*, Volume 7, 2005.

[25] Holger Ebel, Lutz-Ingo Mielsch, and Stefan Bornholdt. Scale-free topology of e-mail networks, Feb 2002.

[26] P. Erdös and A. Rényi. On random graphs, i. *Publicationes Mathematicae (Debrecen)*, 6:290–297, 1959.

[27] P. Erdös and A. Renyi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5:17–61, 1960.

[28] P. Erdös and A. Renyi. On the strength of connectedness of a random graph. *Acta Mathematica Hungarica*, 12:261–267, 1961.

[29] Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. On power-law relationships of the internet topology. *SIGCOMM Comput. Commun. Rev.*, 29(4):251–262, October 1999.

[30] R. Ferrer I Cancho and R. V. Solé. The small world of human language. *Proc R Soc Lond B Biol Sci*, 268(1482):2261–2265, November 2001.

[31] Santo Fortunato. Community detection in graphs. Jan 2010.

[32] L. Freeman. Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215–239, 1979.

[33] E. N. Gilbert. Random graphs. *The Annals of Statistics*, 30(30):1141–1144, 1959.

[34] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12):7821–7826, June 2002.

[35] Steve Gregory. Finding overlapping communities in networks by label propagation. Oct 2009.

[36] R. Guimerá, S. Mossa, A. Turtschi, and L. A. N. Amaral. The worldwide air transportation network: Anomalous centrality, community structure, and cities global roles. *Proceedings of the National Academy of Sciences of the United States of America*, 102:7794–7799, 2005.

[37] Mark S. Handcock, Adrian E. Raftery, and Jeremy M. Tantrum. Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):301–354, March 2007.

[38] Mark Huisman and Marijtje A.J. van Duijn. Software for social network analysis. In Peter J. Carrington, John Scott, and Stanley Wasserman, editors, *Models and Methods in Social Network Analysis*, pages 270–316. Cambridge University Press, 2005.

[39] Yuntao Jia, Jared Hoberock, Michael Garland, and John Hart. On the visualization of social and other scale-free networks. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1285–1292, 2008.

[40] Hyunmo Kang, Lise Getoor, and Lisa Singh. Visual analysis of dynamic group membership in temporal social networks. *SIGKDD Explor. Newsl.*, 9(2):13–21, December 2007.

[41] G. Kossinets. Effects of missing data in social networks. *Social Networks*, 28(3):247–268, July 2006.

[42] Ravi Kumar, Jasmine Novak, and Andrew Tomkins. Structure and evolution of online social networks. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 611–617, New York, NY, USA, 2006. ACM Press.

[43] Marcelo Kuperman and Guillermo Abramson. Small world effect in an epidemiological model. *Physical Review Letters*, 86(13):2909–2912, Mar 2001.

[44] Amy N. Langville and Carl D. Meyer. Deeper inside pagerank. *Internet Mathematics*, 1(3):335–380, 2003.

[45] Sang H. Lee, Pan-Jun Kim, and Hawoong Jeong. Statistical properties of sampled networks. Nov 2009.

[46] Jure Leskovec, Kevin J. Lang, Anirban Dasgupta, and Michael W. Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. Oct 2008.

[47] Jure Leskovec, Kevin J. Lang, Anirban Dasgupta, and Michael W. Mahoney. Statistical properties of community structure in large social and information networks. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 695–704, New York, NY, USA, 2008. ACM.

[48] Jurij Leskovec, Lada A. Adamic, and Bernardo A. Huberman. The dynamics of viral marketing, Sep 2005.

[49] Stephen W. Liddle, Sai Ho Yau, and David W. Embley. On the automatic extraction of data from the hidden web. In *Revised Papers from the HUMACS, DASWIS, ECOMO, and DAMA on ER 2001 Workshops*, pages 212–226, London, UK, 2002. Springer-Verlag.

[50] David Lusseau. Evidence for social role in a dolphin social network. *Evolutionary Ecology*, 21(3):357–366, May 2007.

[51] Jayant Madhavan, David Ko, Lucja Kot, Vignesh Ganapathy, Alex Rasmussen, and Alon Halevy. Google's deep web crawl. *Proc. VLDB Endow.*, 1(2):1241–1252, 2008.

[52] Clémence Magnien, Matthieu Latapy, and Michel Habib. Fast computation of empirically tight bounds for the diameter of massive graphs. *ACM Journal of Experimental Algorithmics*, 13, 2008.

[53] Yutaka Matsuo, Junichiro Mori, Masahiro Hamasaki, Keisuke Ishida, Takuichi Nishimura, Hideaki Takeda, Koiti Hasida, and Mitsuru Ishizuka. Polyphonet: an advanced social network extraction system from the web. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 397–406, New York, NY, USA, 2006. ACM Press.

[54] Peter Mika. 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):211–223, October 2005.

[55] Peter Mika. *Social Networks and the Semantic Web (Semantic Web and Beyond)*. Springer, September 2007.

[56] S. Milgram. The small world problem. *Psychology Today*, 2(1):60–67, 1967.

[57] Nina Mishra, Robert Schreiber, Isabelle Stanton, and Robert E. Tarjan. Clustering social networks, Nov 2007.

[58] Alan Mislove, Krishna P. Gummadi, and Peter Druschel. Exploiting social networks for internet search. In *In Proceedings of the 5th Workshop on Hot Topics in Networks (HotNets-V*, 2006.

[59] Alan Mislove, Hema S. Koppula, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. Growth of the flickr social network. In *WOSP '08: Proceedings of the first workshop on Online social networks*, pages 25–30, New York, NY, USA, 2008. ACM.

[60] Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks. In *IMC '07: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42, New York, NY, USA, 2007. ACM.

[61] J.L. Moreno. *Who shall survive? : a new approach to the problem of Human Interrelations*, volume 58 of *Nervous and mental disease monograph series*. Nervous and Mental Disease Publ., Washington, 1934.

[62] M. E. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 69(2 Pt 2), Feb 2004.

[63] M. E. J. Newman. Fast algorithm for detecting community structure in networks. Sep 2003.

[64] M. E. J. Newman. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5200–5205, April 2004.

[65] M. E. J. Newman. Random graphs with clustering. Mar 2009.

[66] Mark E. J. Newman. Power laws, pareto distributions and zipf's law, May 2006.

[67] Mark E. J. Newman, Albert L. Barabási, and Duncan J. Watts, editors. *The Structure and Dynamics of Networks*. Princeton University Press, 2006.

[68] Mark E. J. Newman, D. J. Watts, and S. H. Strogatz. Random graph models of social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(Suppl 1):2566–2572, February 2002.

[69] David L. Nowell and Jon Kleinberg. The link prediction problem for social networks. In *CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management*, pages 556–559, New York, NY, USA, 2003. ACM.

[70] J. O'Madadhain, D. Fisher, P. Smyth, S. White, and Y. B. Boey. Analysis and visualization of network data using jung. *Journal of Statistical Software*, 2005.

[71] Tore Opsahl and Pietro Panzarasa. Clustering in weighted networks. *Social Networks*, 31(2):155–163, May 2009.

[72] Pekka Orponen and Satu Elisa Schaeffer. Local clustering of large graphs by approximate fiedler vectors. In *Proceedings of the Fourth International Workshop on Efficient and Experimental Algorithms (WEAŠ05), volume 3505 of Lecture Notes in Computer Science*, pages 524–533. Springer-Verlag GmbH, 2005.

[73] Felipe Ortega. *Wikipedia: A Quantiative Analysis*. PhD thesis, Universidad Rey Juan Carlos, Madrid, Spain, 2009. http://libresoft.es/Members/jfelipe/phd-thesis.

[74] Gergely Palla, Albert-Laszlo Barabasi, and Tamas Vicsek. Quantifying social group evolution. *Nature*, 446(7136):664–667, April 2007.

[75] Pascal Pons and Matthieu Latapy. Computing communities in large networks using random walks (long version). Dec 2005.

[76] Huaijun Qiu and Edwin R. Hancock. Graph matching and clustering using spectral partitions. *Pattern Recogn.*, 39(1):22–34, 2006.

[77] Usha N. Raghavan, Reka Albert, and Soundar Kumara. Near linear time algorithm to detect community structures in large-scale networks, Sep 2007.

[78] Seth Richards. A social network analysis into the david kelly tragedy. *Connections*, 26:25–32, 2005.

[79] Barna Saha and Pabitra Mitra. Dynamic algorithm for graph clustering using minimum cut tree. In *ICDMW '06: Proceedings of the Sixth IEEE International Conference on Data Mining - Workshops*, pages 667–671, Washington, DC, USA, 2006. IEEE Computer Society.

[80] Venu Satuluri and Srinivasan Parthasarathy. Scalable graph clustering using stochastic flows: applications to community discovery. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 737–746, New York, NY, USA, 2009. ACM.

[81] Satu Elisa Schaeffer. Stochastic local clustering for massive graphs. In T. B. Ho, D. Cheung, and H. Liu, editors, *Proceedings of the Ninth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-05)*, volume 3518 of *Lecture Notes in Computer Science*, pages 354–360. Springer-Verlag GmbH, 2005.

[82] Satu Elisa Schaeffer. Graph clustering. *Computer Science Review*, 1(1):27–64, 2007.

[83] Bimal Viswanath, Alan Mislove, Meeyoung Cha, and Krishna P. Gummadi. On the evolution of user interaction in facebook. In *WOSN '09: Proceedings of the 2nd ACM workshop on Online social networks*, pages 37–42, New York, NY, USA, 2009. ACM.

[84] Stanley Wasserman, Katherine Faust, and Dawn Iacobucci. *Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences)*. Cambridge University Press, November 1994.

[85] Duncan J. Watts. *Six degrees: The science of a connected age*. WW Norton & Company, 2003.

[86] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, June 1998.

[87] Christo Wilson, Bryce Boe, Alessandra Sala, Krishna P. Puttaswamy, and Ben Y. Zhao. User interactions in social networks and their implications. In *Proceedings of the Fourth ACM European conference on Computer Systems (EuroSys)*, pages 205–218, New York, NY, USA, 2009. ACM Press.

[88] Jennifer Xu and Hsinchun Chen. The topology of dark networks. *Commun. ACM*, 51(10):58–65, 2008.

[89] Bowen Yan and Steve Gregory. Detecting communities in networks by merging cliques. Technical report, 2009.

[90] Bowen Yan and Steve Gregory. Detecting communities in networks by merging cliques. In *2009 IEEE International Conference on Intelligent Computing and Intelligent Systems (ICIS 2009)*, pages 832–836. IEEE, November 2009.

[91] W. W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33:452–473, 1977.

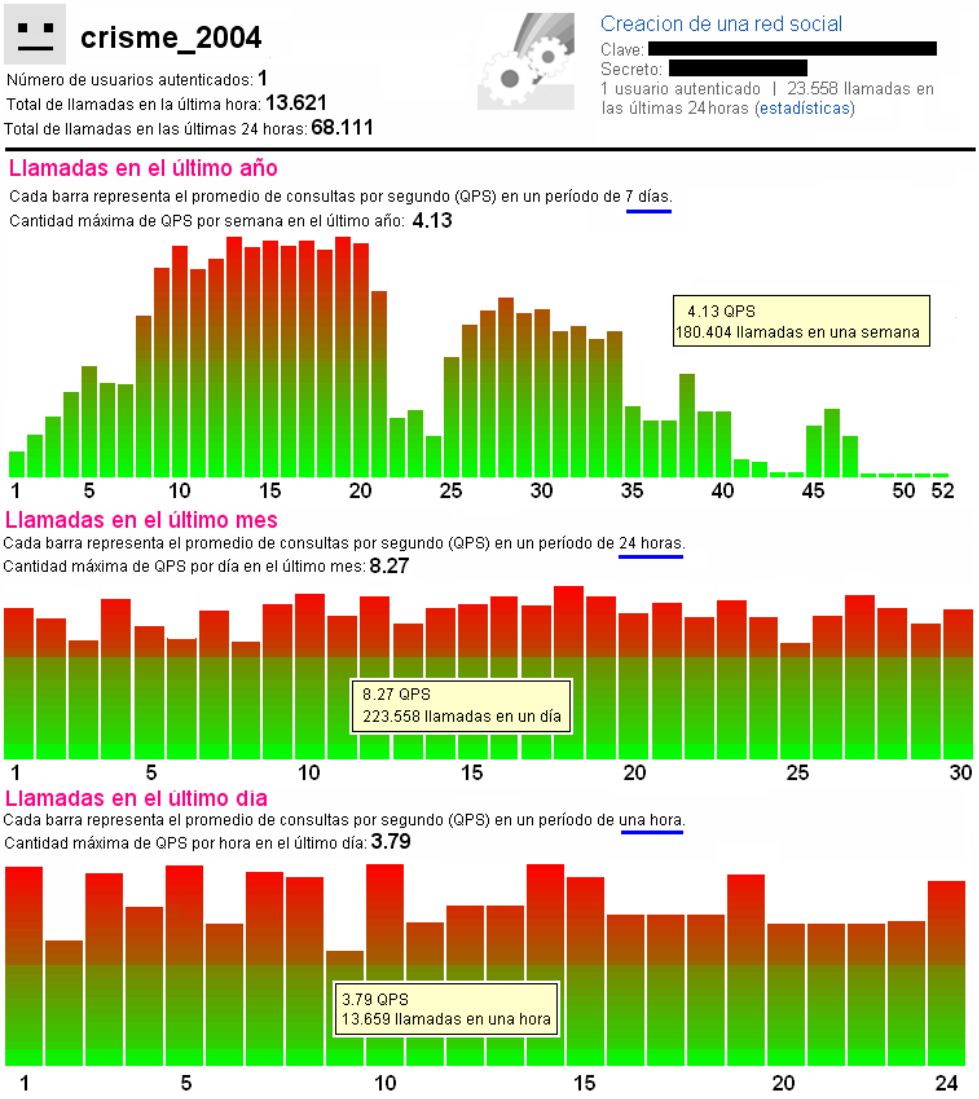
[92] H. Zanghi, C. Ambroise, and V. Miele. Online and offline social networks: Use of social networking sites by emerging adults. *Applied Developmental Psychology*, 29:420–433, August 2008.

[93] Hongyu Zhang. The scale-free nature of semantic web ontology. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 1047–1048, New York, NY, USA, 2008. ACM.

[94] George K. Zipf. *Human Behaviour and the Principle of Least Effort*. Addison Wesley, Cambridge MA, 1949.

Apéndice A

Estadísticas de llamadas a la BD de Flickr



Apéndice B

Recursos Electrónicos

B.1 Fuente de Datos

- **Base de Datos.** Wikipedia.
  http://download.wikimedia.org/eswiki/
- **Dataset.** MySpace de [4]
  http://an.kaist.ac.kr/traces/WWW2007.html

B.2 Programas para el análisis de redes sociales

- **MultiNet 4.24.** Paquete para el análisis de redes sociales.
  http://www.sfu.ca/personal/archives/richards/Multinet/Pages/multinet.htm
- **NetDraw 1.0.** Visualización de redes sociales.
  http://www.analytictech.com/Netdraw/netdraw.htm
- **NetMiner 3.4.** Análisis exploratorio y visualización de datos.
  http://www.netminer.com/NetMiner/home_01.jsp
- **Pajek 1.24.** Análisis y visualizaión de redes grandes.
  http://vlado.fmf.uni-lj.si/pub/networks/pajek/
- **StOCNET 1.8.** Análisis de redes sociales estadístico.
  http://stat.gamma.rug.nl/stocnet/
- **UCINET 6.05.** Paquete para el análisis de redes sociales.
  http://www.analytictech.com/ucinet/
- **GUESS 1.0.3.** The Graph Exploration System.
  http://graphexploration.cond.org/documentation.html

- **SIENA 3.1.** Simulation Investigation for Empirical Network Analysis.
  `http://stat.gamma.rug.nl/siena.html`
- **SoNIA 1.2.** Visualizador dinámico de redes.
  `http://www.stanford.edu/group/sonia/`
- **JUNG 2.0.** Java Universal Network/Graph Framework.
  `http://jung.sourceforge.net/`
- **MySQl 5.1.24.** Manejador de Base de Datos.
  `http://dev.mysql.com/downloads/mysql/5.1.html`
- **Connector/ODBC 5.1.4.** Fuente de datos ODBC para MySQL.
  `http://dev.mysql.com/downloads/connector/odbc/5.1.html`
- **R-project 2.7.0.** Proyecto para estadística computacional.
  `http://cran.r-project.org/bin/`
- **RODBC 1.2.3.** Paquete de R para conexión con ODBC.
  `http://cran.r-project.org/web/packages/RODBC/index.html`