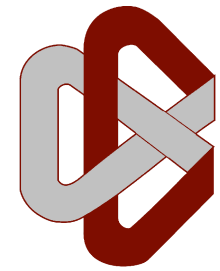

CIMAT

Centro de Investigación en Matemáticas, A.C.

Revisión sistemática de herramientas y métodos para las diferentes fases del análisis de redes sociales en línea.

REPORTE TÉCNICO

Que para obtener el grado de

Maestro en Ingeniería de Software

P r e s e n t a

Julio Alonso García González

Director(a) de Reporte Técnico

Dra. Alejandra García Hernández

Zacatecas, Zacatecas., 06 de Noviembre de 2013

Agradecimientos

A todos los maestros
por la gran paciencia
que han tenido hacia conmigo.

Y un agradecimiento especial
a la *Doctora Alejandra García H.*
por su constante ayuda
en la finalización de este reporte.

Índice de Tablas

Índice de Ilustraciones

Resumen

Las Redes Sociales en Línea (OSN-por sus siglas en inglés), se están convirtiendo en una parte activa e indispensable del que hacer humano. Los datos acumulados en estas redes sociales son cada vez más usados en diferentes estudios, y cada vez surgen nuevas interrogantes.

Durante este trabajo de investigación identificamos que existen muchas herramientas y métodos que permiten hacer análisis de redes sociales en línea. Sin embargo, aún no está claro cuáles utilizar al momento de querer abordar una investigación o estudio. Por lo tanto, el objetivo general de este trabajo de investigación es generar un compendio o esquema que muestre las diferentes etapas, artefactos y algoritmos que faciliten el desarrollo del análisis de redes en línea. El esquema que se presenta muestra algunos de los métodos y herramientas más utilizados durante la etapa de extracción de datos y durante la etapa de análisis de datos en estudios de redes sociales. De esta forma en la etapa de extracción de datos, se realizó una comparativa de las fortalezas y debilidades de los diferentes métodos de extracción de datos de redes sociales en línea, enfocándonos en el uso de las respectivas API que cada plataforma de red social ofrece. Posteriormente, en la etapa de análisis de datos, nos enfocamos en el análisis de desambiguación de perfiles, por ser un análisis de reciente interés en la comunidad científica; y en el análisis de influencia, ya que es uno de los análisis que genera más interés en el público y aplicaciones comerciales. También incluimos una descripción de lo que es el Social network analysis software, explicando algunas de las principales herramientas, y describimos las métricas más utilizadas actualmente en los diferentes estudios de análisis de redes sociales.

Finalmente, durante la última fase de este trabajo de investigación se realizó un experimento que abarca la fase de extracción de datos y análisis de datos. La extracción de la información se realizó en la red social de Facebook y Twitter con la finalidad de comprobar los algoritmos de análisis de datos en un caso de estudio real. Para realizar el caso de estudio se seleccionaron 7 candidatos a la Presidencia de la Ciudad de Zacatecas durante los comicios del 2013, se extrajo información de ellos en las dos redes sociales seleccionadas durante un periodo de 30 días antes del día de las votaciones, y se probaron los algoritmos Similar Text y la Distancia de Levenshtein, para probar su efectividad en la desambiguación de perfiles (identificar perfiles de una misma persona en diferentes redes sociales). Para el análisis de influencia proponemos una nueva fórmula que calcula un factor de influencia (FI) de cada uno de los candidatos en las redes sociales, y para probar su efectividad comparamos la fórmula propuesta con fórmulas de influencia comunes en el análisis de redes, y realizamos un análisis de correlación de los resultados utilizando las diferentes fórmulas de influencia, con los resultados reales de la votación. Los resultados del experimento muestran que en la fase de desambiguación de perfiles, se logró obtener 810 perfiles de Twitter con una probabilidad de más del 90% de encontrar una coincidencia de perfil en Facebook. En el análisis de influencia encontramos un comportamiento similar de la influencia de los candidatos en las redes sociales con los resultados finales de la votación. Sin embargo es importante mencionar que la fórmula de influencia que se propone mostró una mayor correlación R^2=0.46867 que la métrica de retweets (influencia) la cual obtuvo una

correlación de R^2=0.40465, lo que siguiere una mayor precisión de medición de la influencia real de los actores o eventos sociales.

Por último anexamos compendios de herramientas de análisis de redes sociales que pueden servir de referencia para nuevos investigadores de las redes sociales en línea.

Introducción

Mucho se ha hablado de cómo las redes sociales basadas en internet están cambiando la forma de interactuar y comunicarse en nuestra época, cada vez más las personas ven como algo natural el contar con un perfil en una de las distintas redes sociales basadas en internet como Twitter, Facebook y LinkedIn. Ya no se trata solo de una moda, muchas de las compañías destinan campañas con altos presupuestos para tener presencia y promocionarse a través de estos medios.

Otro suceso, es que cambios en la tecnología hacen que estas redes nos acompañen en los distintos quehaceres de la vida cotidiana, y no solo a través de las computadoras, un sin fin de dispositivos como celulares, smartphones y tabletas están haciendo que estas redes se masifiquen gracias a la movilidad que ofrecen, y sean naturales para las nuevas generaciones.

Las redes sociales acumulan información de la mayoría de las personas en el mundo, están siendo usadas para perfilar cada vez más a las personas, además de permitir conocer sus opiniones y sus relaciones o su comunidad.

Junto con el boom de las redes sociales, llegó el boom de la liberación de API[1]. Las API en las redes sociales (y en cualquier entorno de programación) son esos métodos a los que se les puede llamar para recoger datos sobre las mismas. Vamos, es la puerta que nos abren los grandes servicios sociales para acceder a parte de su información. Gracias a ellos se pueden llevar a cabo desarrollos por parte de terceros.

En la presente investigación trataremos de mostrar un panorama sobre cuáles herramientas y métodos, son los más usados actualmente para la minería de datos en las redes sociales de Facebook, LinkedIn y Twitter, mencionaremos también que pros y contras hemos encontrado, y cómo se pudieran aprovechar las redes sociales basadas en Internet, para generar herramientas o métodos propios de extracción de datos, así como algoritmos de análisis que nos ayuden a resolver nuevos cuestionamientos, de esta forma la investigación se divide en dos partes:

La primera parte está dedicada a el Estado del Arte el cual está organizado de la siguiente manera: en el Capítulo 1 se da una breve introducción del trabajo de investigación y se definen los principales conceptos que se estarán utilizando a lo largo de este trabajo. En el Capítulo 2 se presenta un análisis de las herramientas o librerías proporcionadas por las propias redes sociales, y su integración con lenguajes de programación que permiten la explotación de la información recolectada y generada en estos medios. Posteriormente, en el Capítulo 3 se presenta un análisis de las herramientas o librerías de terceros también llamadas "Social network analysis software (SNA software)", generando un índice de las diferentes iniciativas. Durante el Capítulo 3 se analizan también las métricas más comunes usadas en el Análisis de Redes Sociales para el análisis de datos extraídos; y se realiza también una investigación sobre los algoritmos más usados para identificar perfiles de usuarios en diferentes redes sociales en línea, así como nuevas propuestas para analizar las tendencias de influencia de los actores de estas redes.

[1] Interfaz de programación de aplicaciones ó API Application Programming Interface.

La segunda parte está dedicada a la parte experimental del trabajo de investigación. En el Capítulo 4 incluimos un experimento analizando los datos de los candidatos a la alcaldía de Zacatecas en las elecciones de Julio de 2013, en la primera parte del experimento se utilizaron algoritmos que permiten hacer una desambiguación de perfiles de usuarios de los seguidores de estos candidatos, entre las redes sociales en línea, Facebook y Twitter. En la segunda parte del experimento se realizó un análisis de tendencias de influencia, se propuso una nueva fórmula para obtener un factor de influencia (FI) de cada uno de los candidatos en las redes sociales en línea durante las campañas, y se hizo un análisis de correlación entre las tendencias obtenidas en las redes sociales en línea con las métricas tradicionales y la nueva métrica (FI), versus el resultado real de las votaciones de la elección electoral, lo que nos permitió corroborar la efectividad de las métricas utilizadas en el análisis de redes sociales en línea para el caso presentado en este trabajo de investigación. Cabe señalar que las fases del experimento son independientes una de la otra y los objetivos de cada una es demostrar la efectividad de los algoritmos utilizados para cada análisis.

Por último en el Capítulo 5 se mencionan las conclusiones a las que se han llegado después de ver los resultados tanto de la revisión literaria que se realizó, así como de los resultados de los experimentos ejecutados.

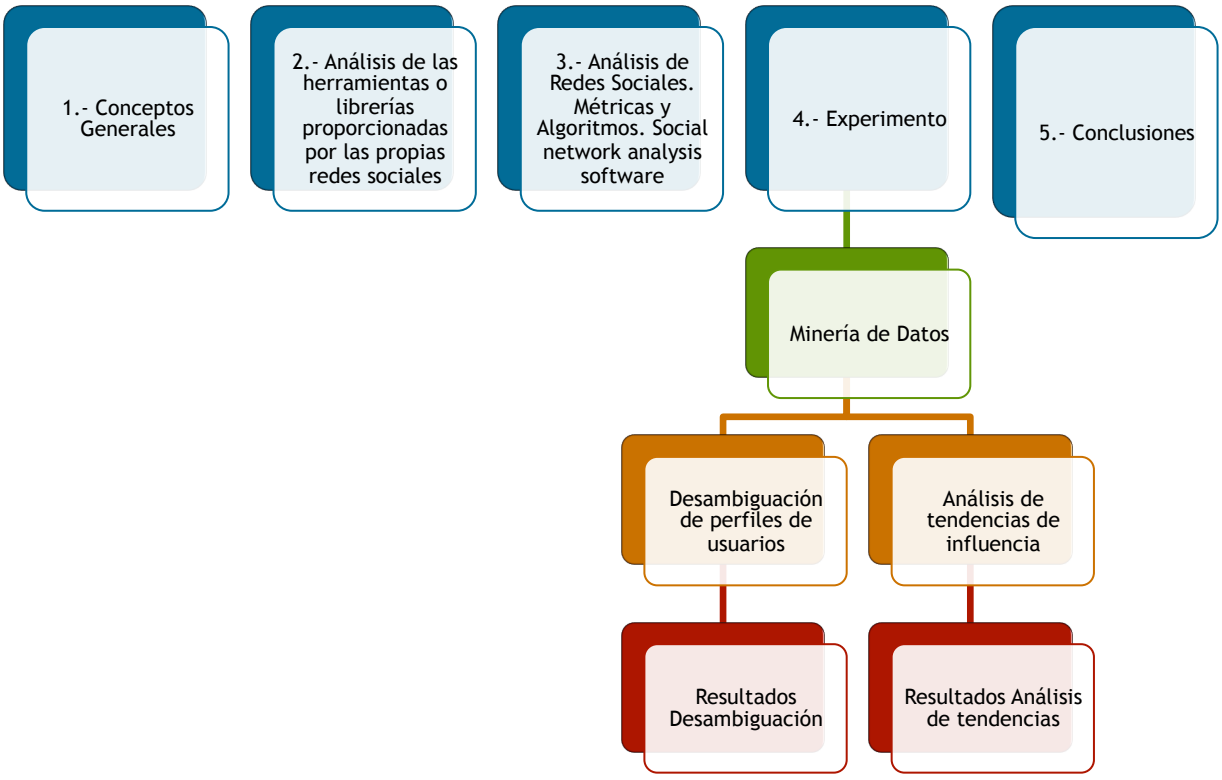

Ilustración 1 Estructura de la "Revisión sistemática de herramientas y métodos para las diferentes fases del análisis de redes sociales en línea".

1 Conceptos Generales

1.1 ¿Qué es una Red social?

Una red social es una estructura social compuesta de individuos que se encuentran relacionados entre sí. Las relaciones pueden ser de distinto tipo, como intercambios financieros, amistad, entre otros.

A partir del 2003 aparecen en internet espacios en los que se genera interacción social a través de plataformas residentes en internet como MySpace, Hi5, Facebook y Twitter, estas herramientas aprovechan el uso de diferentes componentes electrónicos como chats, foros, galerías fotográficas y blogs para lograr esta interacción social. Actualmente la interacción social a través de internet se ha extendido en gran medida con el uso de dispositivos móviles electrónicos.

En España, el Instituto Nacional de Tecnologías de la Comunicación (INTECO) en su "Estudio sobre la privacidad de los datos y la seguridad de la información en las redes sociales online", del año 2009, define a las redes sociales en línea como "los servicios prestados a través de Internet que permiten a los usuarios generar un perfil público, en el que plasmar datos personales e información de uno mismo, disponiendo de herramientas que permiten interactuar con el resto de usuarios afines" (Urueña & Ferrari 2011).

Wikipedia, uno de los medios de comunicación más consultados por los internautas, las define como: "estructuras sociales compuestas de grupos de personas, las cuales están conectadas por uno o varios tipos de relaciones, tales como amistad, parentesco, intereses comunes o que comparten conocimientos"[2].

Las funcionalidades de una red social en internet varían en algunos casos considerablemente. Algunas permiten alojar fotografías, vídeos, pueden tener mensajería instantánea o permiten el envío y la recepción de mensajes privados de forma similar al correo. Muchas, en la actualidad, se apoyan en la telefonía móvil y están segmentadas por los más variados intereses: hacer amigos, buscar pareja, hacer negocios, compartir música y un largo etcétera. Merece especial mención la apuesta que muchas redes sociales están realizando, por la integración del comercio electrónico a través del desarrollo del comercio social ("social commerce"), mediante la incorporación de tiendas en línea a través de las páginas creadas por empresas en este tipo de redes sociales.

El atractivo de las redes sociales para la publicidad, radica en la potencial capacidad de poder enviar mensajes a una gran cantidad de usuarios (potenciales consumidores), en muy poco tiempo, a través de un soporte que resulta mucho más económico que los medios tradicionales y que cuenta con la gran ventaja de una elevada capacidad de segmentación.

Según la Asociación Mexicana de Internet AMIPCI en su estudio "Hábitos de los usuarios de internet en México" del 2012, y el "Estudio de consumo de medios entre internautas mexicanos" de enero del 2013 efectuado por la agencia de mercadotecnia IAB México, las

[2] Wikipedia http://es.wikipedia.org/wiki/Red_social -- Red social, consultado en Julio de 2012.

redes sociales más populares son Facebook, Twitter y YouTube (AMIPC 2012; IAB-México 2013)

1.2 La importancia del análisis de los datos generados por las redes sociales

Al día de hoy existen varias iniciativas que estudian los datos acumulados en las distintas redes sociales, tratando de comprender el cómo las redes sociales están afectando o transformando el comportamiento de esta nueva sociedad. Estas iniciativas responden a preguntas como:

- ¿Quién conoce a quién y qué amigos se tienen en común?
- ¿Con qué frecuencia ciertas personas se comunican con otras?
- ¿Qué tan simétrica es la comunicación entre las personas?
- ¿Qué personas son las más influyentes o populares en las redes sociales?
- ¿De qué hablan las personas y cuáles con sus intereses?

Las respuestas a este tipo de preguntas generalmente conectan a un grupo de personas, y apuntan hacia un contexto que indica por qué existen las conexiones o cómo se forma la red social (Russell 2011).

Desafortunadamente no todas las redes son abiertas a ofrecer su contenido públicamente, muchas como Facebook y LinkedIn, entre otras, tienen fuertes condiciones de privacidad, lo que dificulta la minería de los datos generados en estas.

Afortunadamente existen redes sociales como Twitter que por filosofía son públicas, lo que permite acceder a un universo gigantesco de datos de diferente índole, y proporciona herramientas y métodos para la extracción y análisis de estos (Interfaz de programación de aplicaciones o API Application Programming Interface).

Si bien existen muchos programas de software, librerías y herramientas que permiten hacer un análisis de los datos generados por las redes sociales, aún queda mucho por hacer en lo referente a la extracción de los datos de redes, cada vez más las diferentes redes sociales integran nuevos servicios que recolectan datos, ya sean datos textuales, de lenguaje natural, de multimedia, gráficos, y es común que estos datos sean clasificados, perfilando a cada usuario; nuevas interrogantes pueden ser contestadas a partir de esta información y cada vez más las diferentes organizaciones privadas, gubernamentales, ONG, voltean hacia estos instrumentos tratando de encontrar respuestas a diferentes problemas que van desde comerciales, predicción de tendencias y hasta de bienestar social.

Las redes sociales se están convirtiendo en el mayor universo de datos de las personas en el mundo, y es por ello que es de mucha importancia el que se conozcan los métodos de extracción y análisis de ésta información, que nos ayuden a obtener la respuesta adecuada a las preguntas que se formulen en diferentes investigaciones.

En cinco años las redes sociales serán el segundo canal más importante de atención al cliente, según The Global CEO Study de IBM (IBM et al. 2012). En línea con el punto anterior,

Gartner dice que, para 2014, negarse a atender a los clientes a través de las redes sociales será tan perjudicial como no hacerlo hoy por teléfono.

Las empresas que añadan las comunidades en línea a sus servicios de atención al cliente tendrán un ahorro de entre un 10 y un 50% en costos (IBM et al. 2012). Además, para 2015 se estima que los temas dominantes relacionados con la atención al cliente serán procesos colaborativos en servicio al cliente (co-working), migración de aplicaciones a la nube (Cloud) y soporte a consumidores en dispositivos móviles, que empiezan a ser una de las grandes tendencias ahora mismo (IBM et al. 2012).

2 Minería de datos de las redes sociales

2.1.1 Antecedentes de la minería de datos de redes sociales

Podemos clasificar las técnicas para recolectar datos de una plataforma de Online Social Media OSN en dos grandes categorías: la primera categoría se basa en el uso de las API ad-hoc, que por lo general son proporcionadas por la plataforma web social en sí, y una segunda categoría que consiste básicamente en examinar y extraer los datos desde el HTML de páginas web.

En cuanto a la primera categoría, señalamos que en la actualidad las plataformas de redes sociales proporcionan API potentes (a menudo disponibles para su consumo en varios lenguajes de programación), que permiten recuperar de forma fácil y rápida una amplia gama de información de la propia plataforma. Esta información, en particular, se refiere no sólo a conexiones sociales de los miembros de las plataformas, sino también al contenido publicado y a los usuarios (Ferrara et al. 2012).

La técnica más fácil de acceder a la información de las plataformas de Redes Sociales en Línea (Online Social Networks - OSN) es a través de sus API, sin embargo, debido a las limitantes de privacidad de estas plataformas con el fin de proteger la información privada de los usuarios, se han explorado diferentes técnicas de extracción de datos.

La principal técnica de extracción de datos, consiste en utilizar el HTML desplegado por estas plataformas para consumir el Modelo de Objetos del Documento (DOM) del HTML, limpiarlo y acceder a atributos, que las API tienen restringidas.

El Modelo de Objetos del Documento (DOM) es una interfaz de programación de aplicaciones (API) para documentos HTML y XML. Define la estructura lógica de los documentos y el modo en que se accede y manipula un documento. En la especificación del DOM, el término "documento" se utiliza en un sentido amplio. XML se utiliza cada vez más como un medio para representar muchas clases diferentes de información que puede ser almacenada en diversos sistemas, y mucha de esta información se vería en términos tradicionales, más como datos que como documentos. Sin embargo, XML presenta estos datos como documentos, y se puede usar el DOM para manipular estos datos. En pocas palabras se trata de interpretar el HTML como si fuera XML.

Con el Modelo de Objetos del Documento DOM, los programadores pueden construir documentos, navegar por su estructura, y añadir, modificar o eliminar elementos y contenido. Se puede acceder a cualquier cosa que se encuentre en un documento HTML o XML, salvo algunas excepciones. En particular, aún no se han especificado las interfaces DOM para los subconjuntos internos y externos de XML.

En el artículo "Web Data Extraction, Applications and Techniques: A Survey" (Ferrara et al. 2012), se menciona el método de extracción utilizando el Modelo de Objetos del Documento (DOM) que permite interpretar los HTML como si fueran documentos de XML, con una estructura similar a la de un árbol, de esta forma es posible utilizar algoritmos de recorrido de árboles, para la extracción de los atributos.

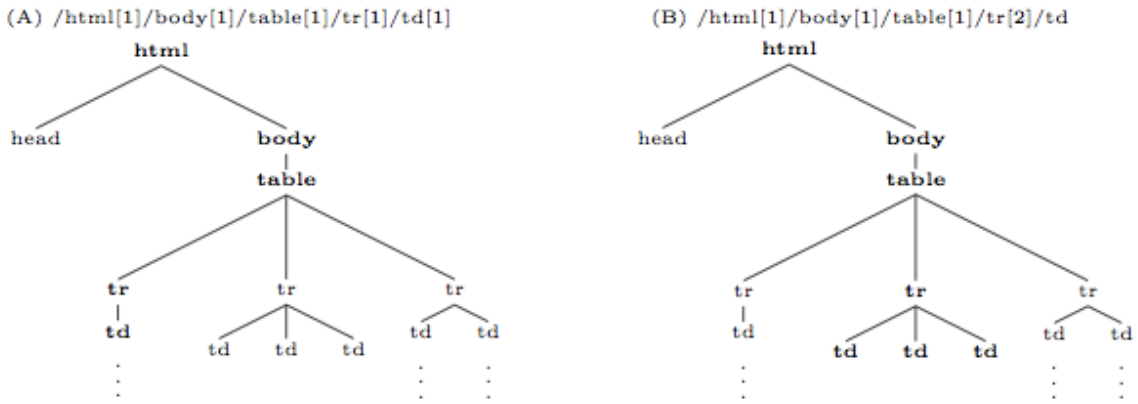

Ilustración 2 Ejemplo de acceso vía DOM con XML a una tabla de HTML (Ferrara et al. 2012)

En la ilustración 1 se muestra cómo se accede a los registros de una tabla utilizando el DOM. En el inciso A se representa la ruta de acceso para acceder al contenido de la columna uno del primer registro de tabla en el HTML, en el inciso B se representa cómo acceder a la columna uno del segundo registro de la tabla, ejemplificando como solo es necesario modificar los subíndices del apuntador del objeto para variar el acceso a los distintos objetos del HTML.

El uso de estos algoritmos permite construir sistemas automatizados que pueden recorrer grandes cantidades de información, en tiempos relativamente rápidos. Google utiliza estos algoritmos -también llamados de araña- en sus robots de indexación de datos de la WEB.

Estas técnicas fueron probadas en la Universidad de Indiana (Ferrara et al. 2012), con redes sociales carentes de API como MySpace, y arrojaron buenos resultados al obtener las conexiones o subscripciones de sus páginas o perfiles de MySpace. Esto supondría que pudiera aplicarse a redes sociales como Facebook, donde el API está restringida para consultar las conexiones de un usuario por ejemplo. Una posible solución para obtener las conexiones de un usuario, sería construir una aplicación que analizara el HTML que Facebook presenta, analizándolo mediante el recorrido de su estructura utilizando el DOM y obtener los diferentes Id de usuarios encontrados en los links del HTML mostrados en las listas de amigos, para después extraer más datos de perfil de estos Id a través del API de Facebook. Esta es en teoría una solución factible para la recuperación de datos, sin embargo aparecen nuevas limitantes, como el que cada vez más, las plataformas usan técnicas de criptografía de links en los HTML, lo que hace casi imposible descifrar los valores arrojados por el DOM. Aunque Facebook aun presenta los ID de usuarios en los links, es posible que en un futuro Facebook deje de hacerlo.

Otra limitante de estas técnicas es el uso constante de Ajax, una técnica asíncrona que impide que la estructura del HTML sea fija y completa, ya que está continuamente enviando eventos en las plataformas y que hacen que se esté actualizando el contenido desplegado por el HTML; por ejemplo, el muro de Facebook está actualizándose continuamente. Aun así, hay secciones que permanecen estáticas como las listas de amigos.

También tenemos que tomar en cuenta que este tipo de extracción de datos se basa completamente en llamadas de HTML, así que la información mostrada por la plataforma corresponderá solo a usuarios autentificados o en sesión del navegador. Aunque es posible

acceder a más información en la plataforma, el llegar a la información que realmente es interesante para el análisis, pudiera depender de acciones que el usuario tendría que ejecutar manualmente, por ejemplo el buscar a un contacto no amigo del usuario en sesión.

Por último, tenemos que recordar que las diferentes plataformas tampoco muestran todo sobre los usuarios que no están en la red del usuario en sesión, la información mostrada depende de los permisos o perfiles públicos especificados por cada usuario u objeto (páginas de fans, eventos, lugares, etc.), de esta forma el análisis del HTML estará también limitado por esos permisos.

La última limitante es que se requiere de una infraestructura capaz de soportar el continuo procesamiento de los algoritmos de búsquedas y extracción de los datos en línea, lo que aumenta el costo monetario de operación, ya que estos algoritmos son complejos de implementar y mantener, debido a que la estructura de los HTML en las plataformas está en constante cambio.

2.1.2 Minería de datos en Facebook

En la actualidad -finales del año 2012-, vemos integradas las redes sociales en línea prácticamente en todas partes, Facebook tiene más de 1 billón (mil millones) de usuarios, la media de amigos o conexiones por usuarios pasaron de 160 en el 2009 a 330 en 2012: Facebook tiene más de 600,000 usuarios móviles y se ha convertido en la red social más popular en el mundo. México es uno de los 5 países con mayor crecimiento en número de usuarios (Facebook-News 2012).

Facebook ha evolucionado mucho, y las herramientas de conexión para terceros proporcionadas por la compañía también lo han hecho, actualmente Facebook proporciona una API llamada Graph API (Facebook-Developers 2012). Esta API permite a los desarrolladores acceder al núcleo de Facebook, contiene métodos que permiten extracción de datos de diferentes objetos. Cada objeto tiene un identificador único, por ejemplo el identificador de la página oficial de desarrolladores de Facebook es 19292868552, y mediante este identificador se puede acceder a sus propiedades, mediante llamadas de tipo REST[3], como el siguiente ejemplo: https://graph.facebook.com/19292868552 y los resultados son presentados en una colección de datos con formato JSON[4].

Alternativamente también se pueden usar los nombres de usuario como identificador para acceder a los diferentes objetos, por ejemplo: https://graph.facebook.com/usuario/picture, el resultado será devuelto en formato JSON de cualquier manera. Todos los objetos pueden ser

3 La Transferencia de Estado Representacional O REST (por sus siglas en inglés -Representational State Transfer) es una interfaz web simple que utiliza XML y HTTP, sin las abstracciones adicionales de los protocolos basados en patrones de intercambio de mensajes como el protocolo de servicios web SOAP (http://es.wikipedia.org/wiki/Representational_State_Transfer).

4 JSON, acrónimo de JavaScript Object Notation, es un formato ligero para el intercambio de datos. JSON es un subconjunto de la notación literal de objetos de JavaScript que no requiere el uso de XML.

utilizados de la misma manera y los métodos disponibles están clasificados de la siguiente manera:

- **User (Usuarios).** Permite acceder a los atributos y métodos para compartir contenido en el muro y perfil del objeto usuario de la plataforma de Facebook.
- **Pages (Páginas de Fans).** Este objeto contiene la colección de atributos y métodos para compartir contenido en las "Fan Pages" de Facebook.
- **Events (Eventos).** Permite acceder a la colección de atributos del objeto Eventos que los usuarios o Fan Pages publican**.**
- **Groups (Grupos).** Permite acceder a los atributos y métodos para compartir del objeto Grupos de Facebook, aunque para acceder a las listas de usuarios está suprimida actualmente.
- **Applications (Aplicaciones).** Permite acceder a los atributos del perfil de las aplicaciones de terceros colocadas en la plataforma de Facebook, este objeto solo contiene atributos de solo lectura.
- **Status messages (Mensajes de Estatus o Muro).** Permite hacer búsquedas directamente en los post públicos de toda la plataforma, aun cuando los usuarios no hayan otorgado permisos de acceso a la aplicación, para ellos usa palabras clave de búsqueda.
- **Photos (Fotografías).** Permite acceder a los atributos de las colecciones de fotos y álbumes de los diferentes objetos en la plataforma de Facebook.
- **Checkins Places (Lugares o visitas).** Permite acceder a los atributos del objeto Places de Facebook, este objeto contiene la colecciones de posicionamientos efectuados por usuarios de la plataforma, los atributos corresponde al lugar que visita el usuario.

Todos los objetos en Facebook están conectados entre sí a través de relaciones llamadas "conexiones". Estas conexiones se pueden examinar a través de la siguiente estructura de URL: `https://graph.facebook.com/ID/CONNECTION_TYPE,` pudiendo acceder a las diferentes conexiones entre los objetos y que pueden ser: Friends, Profile feed (Wall ó Muro), Likes, Movies, Music, Books, Notes, Permissions, Photo Tags, Photo Albums, Video Tags, Video Uploads, Events, Groups, Checkins Places, Objects with Location, lamentablemente algunas de estas conexiones están limitadas por los acuerdos de privacidad, y para poder acceder a ellas requeriríamos de permisos del usuario u objeto que se quisiera examinar.

Los atributos que las llamadas del API regresa dependen del objeto examinado, y son muy variados, en el Anexo 1 se muestran los atributos del objeto User, los permisos necesarios para obtener el atributo y el valor que regresa cada atributo. Se pueden seleccionar los atributos que se requieren en cada búsqueda, usando el parámetro "fields" de la siguiente manera: `https://graph.facebook.com/me?fields=id,name`

Además de los atributos podemos obtener también las conexiones de objeto. En el Anexo 2 se muestran las conexiones, permisos necesarios para la consulta y valor de retorno de la conexión.

El API Graph de Facebook permite también hacer búsquedas a través de todos los objetos públicos de la red social usando la forma `https://graph.facebook.com/search?q=QUERY,` con

este método podemos encontrar casi cualquier objeto en Facebook, por lo que se convierte en una herramienta muy útil en la estrategia de extracción de datos de esta red social.

Otra de las herramientas que el API de Facebook ofrece, es Facebook Query Language (FQL), que permite ejecutar consultas de estilo SQL para consultar los datos expuestos por el API Graph. FQL prevé algunas características avanzadas que no están disponibles en la API Graph, incluyendo el procesamiento por lotes: varias consultas en una sola llamada.

Las consultas son de la forma SELECT [campos] FROM [tabla] WHERE [condiciones]. A diferencia de SQL, la cláusula FROM FQL sólo puede contener una sola tabla. Puede utilizar la palabra clave IN SELECT o hacer sub-consultas, pero las sub-consultas no pueden hacer referencia a las variables de ámbito de la consulta externa.

Al utilizar una tecnología de tipo REST, el API permite ser consumida prácticamente por cualquier lenguaje de programación.

2.1.3 Minería de datos en Twitter

Twitter es la red social que tiene menos limitantes o restricciones de privacidad sobre la información que se publica, además está siendo adoptada por nuevos usuarios cada vez con mayor frecuencia, situándose en el año 2012 en el segundo lugar de popularidad en México y el Mundo (AMIPC 2012; IAB-México 2013).

Twitter ofrece tres API: Streaming API, REST API y Search API, aplicables a diferentes necesidades (Twitter 2012). Estas API pueden ser usadas en conjunto con diversos lenguajes de programación tales como: Java, Python, lenguajes .Net, PHP. Los resultados se presentan en forma de colecciones de datos en formatos XML[5], JSON, RSS[6] y Atom[7].

El **Streaming API** proporciona un sub-conjunto de tweets en casi tiempo real. Se establece una conexión permanente por usuario con los servidores de Twitter y mediante una petición http se recibe un flujo continuo de tweets en formato JSON. Se puede obtener una muestra aleatoria, un filtrado (estados/filtro) por palabras claves o por usuarios. Sin embargo, cuando hay mucha actividad de los usuarios en la red social de Twitter, es difícil obtener todo el flujo de tweets, o los tweets que tienen enlaces (estado/links), o los tweets con retweets (estados/retweets), lo anterior se debe a la sobrecarga de búsquedas de los usuarios de Twitter

5 XML, siglas en inglés de eXtensible Markup Language ('lenguaje de marcas extensible'), es un lenguaje de marcas desarrollado por el World Wide Web Consortium (W3C). Usado principalmente para el intercambio de datos.

6 RSS son las siglas de Really Simple Syndication, un formato XML para indicar o compartir contenido en la web. Se utiliza para difundir información actualizada frecuentemente a usuarios que se han suscrito a la fuente de contenidos.

7 Atom fue desarrollado como una alternativa a RSS, y que pretende reducir la incompatibilidad entre diferentes consumidores del estándar de intercambio de datos.

El **REST API** ofrece a los desarrolladores el acceso al "core[8]" de los datos de Twitter. Todas las operaciones que se pueden hacer vía web es posible realizarlas desde el API. Dependiendo de la operación, requiere o no autenticación con el mismo criterio que en el acceso web.

El **Search API** suministra los tweets con una profundidad en el tiempo de 7 días que se ajustan a la Búsqueda (Query) solicitada. Es posible filtrar por cliente, lenguaje y localización. No requiere autenticación y los tweets se obtienen en formato JSON o ATOM. También el **Search API** ofrece información limitada del tweet, en concreto solo indica el Id del autor, el screen_name y la url de su avatar. Los otros dos API (REST y Streaming) si ofrecen el perfil completo del autor en el momento de la escritura del tweet.

En el libro "Mining the Social Web" (Russell 2011), se encontraron varios métodos de extracción de datos, usando el API de Twitter y el lenguaje de programación Python.

Python es un lenguaje de programación de alto nivel cuya filosofía hace hincapié en una sintaxis muy limpia y que favorece un código legible. Se trata de un lenguaje de programación multi-paradigma ya que soporta orientación a objetos, programación imperativa, y en menor medida programación funcional. Es un lenguaje interpretado y multiplataforma.

Russell nos ejemplifica cómo poder extraer datos a través del REST API de Twitter, parte desde lo más sencillo como hacer llamadas a las funciones básicas del API, para luego ir incrementado en la complejidad de las búsquedas. Los resultados de las búsquedas son arrojados en formato json, los cuales son recorridos para transformarlos en archivos de datos simples en formato de texto ASCII, que pueden ser usados en conjunto con otras herramientas de análisis de datos como "Protovis," una librería para graficar datos y redes, basada en HTML 5 y JavaScript permitiendo tener una noción gráfica de los resultados.

También se ejemplifica cómo hacer uso de otras librerías en conjunto con la API de Twitter, para hacer las búsquedas más específicas, tal es el caso de "The Natural Languaje Toolkit (NLTK)" que es una suite de librerías y programas para Python que permite el procesamiento estadístico y simbólico del lenguaje natural, lo que permite hacer análisis de lo que se está publicando en el texto de cada tweet arrojado por la búsqueda. Para hacer este análisis se requiere de definir una serie de tokens que corresponden a expresiones de lenguaje natural o de uso común, como por ejemplo: "@justinbieber", "U For", "U2", "yeahhh."; es decir modismos que nos permiten clasificar la información que obtenemos. Este procesamiento de lenguaje natural se efectúa en tiempo real al realizar las búsquedas, por lo que no es necesario un reprocesamiento posterior para la reclasificación de los datos, aunque este se puede efectuar para tomar distintos criterios de clasificación diferentes al del análisis de lenguaje natural.

El análisis de las relaciones entre los tweets es otro de los aspectos contemplados en los artículos de "Mining the Social Web" (Russell 2011), y se trata de conocer los enlaces o links de los usuarios de Twitter y el mundo real. Usando expresiones regulares podemos encontrar diferentes tipos de relaciones como:

[8] Núcleo, Corazón, Actividad principal ó mas importante.

- Mentioned by a username
- RT followed by a username
- Via followed by a username

Estas relaciones se convierten en aristas que unen a los diferentes nodos, los cuales pueden ser graficados en las distintas herramientas de análisis gráfico.

Complementando los algoritmos de minería de datos de twitter, Russell nos presenta diferentes técnicas para responder a preguntas como:

- ¿Qué tantos amigos / seguidores (followers) tengo?
- ¿De quién soy seguidor, pero sin que este usuario me siga?
- ¿Quién me sigue sin que yo lo siga?
- ¿Quién es el más amistoso y menos amistoso de las personas en mi red?
- ¿Quiénes son mis amigos mutuos (personas en mi red que están es sus propias redes conectados)?
- ¿Cuál es mi influencia potencial a través del análisis de retweets entre mis seguidores y sus seguidores?
- ¿Cuáles son las principales entidades (temas, hastags, usuarios, palabras) que yo y mi red usamos?

Los principales datos que se han recuperado en los ejercicios de extracción de datos mediante el API de Twitter son (Twitter Dev 2012):

Campo	Tipo de dato	Descripción
Coordinates	Coordinates	Representa las coordenadas geográficas de donde se ha originado el tweet.
created_at	String	Tiempo UTC de cuando el tweet fue creado.
current_user_retweet	Object	Objeto con una colección de datos de los usuarios que han usado el tweet resultante en la búsqueda.
Entities	Entities	Colección de diversos parámetros complementarios devueltos en el tweet.
Favorited	Boolean	Indica si el tweet ha sido marcado como favorito por el usuario autentificado.
Id	Int64	Representa un identificador único para este tweet.
id_str	String	Representa también un identificador único del tweet.
in_reply_to_screen_name	String	Si el tweet ha sido contestado, este parámetro indica el usuario original del tweet.
in_reply_to_status_id	Int64	Representa el id numérico único del texto original del tweet.
in_reply_to_status_id_str	String	Representa el id alfanumérico único del texto original del tweet.
in_reply_to_user_id	Int64	Representa el id numérico único del usuario original del tweet.
in_reply_to_user_id_str	String	Representa el id alfanumérico único del usuario original del tweet.
Campo	Tipo de dato	Descripción
place	Places	Representa que el tweet está asociado con un lugar, aunque no necesariamente haya sido el lugar de origen.

retweet_count	Int	Representa el número de veces que el tweet ha sido retwiteado.
retweeted	Boolean	Indica Verdadero (True) si el tweet ha sido retwiteado, y Falso (False) en caso contrario.
text	String	Representa los 140 caracteres del tweet.
truncated	Boolean	Indica si el tweet fue cortado en su logitud o no por exceder de los 140 caracteres.
user	Users	Colección de datos sobre el usuario que hizo la publicación (Post) del tweet.
withheld_copyright	Boolean	Indica si el tweet contiene material restringido por derechos de autor.
withheld_in_countries	String	Código en letras del país de origen del tweet.

Tabla 1 Campos más comunes resultantes de las búsquedas usando el API de Twitter. *datos tomados de (Twitter blog 2012)

La lista anterior solo muestra los campos más comunes que arroja la extracción de datos, sin embargo es importante decir que las búsquedas usando el API pueden adaptarse para que devuelva otros conjuntos de datos, los cuales pueden verse en la documentación del foro de desarrollo del API de Twitter (Twitter Dev 2012); cabe mencionar que estos datos están en continuo cambio dependiendo de la evolución del API y los métodos de consulta ofrecidos.

2.1.4 Minería de datos en LinkedIn.

LinkedIn es una red social en línea orientada a negocios, fue fundada en diciembre de 2002 y lanzada en mayo de 2003 principalmente como una red profesional[9] (Linkedin-About 2012). En septiembre de 2012 LinkedIn alcanzó los 187 millones de usuarios en más de 200 países, convirtiéndose en la red de profesionistas más grande del mundo. Sesenta y tres por ciento de los miembros de LinkedIn se encuentran fuera de los Estados Unidos. Los usuarios de LinkedIn hicieron casi 4,2 mil millones de búsquedas en la plataforma en 2011 y están en camino de superar a 5,3 mil millones en 2012.

En octubre de 2008, LinkedIn reveló sus planes de abrir su red social de 30 millones de profesionales a nivel mundial, como una muestra de su potencial para la investigación de negocio a negocio. Se está poniendo a prueba el potencial de redes sociales de ingresar al modelo de investigación, que para algunos parece más prometedora que el simple uso de publicidad que se ha estado dando actualmente.

Al igual que Facebook, LinkedIn ofrece a terceros la posibilidad de usar un API para interactuar con el core de la plataforma. A diferencia de otras API de redes sociales, en LinkedIn es requisito que la aplicación que hará uso del API esté registrada en la plataforma, para poder obtener los tokens o llaves para usar el API.

[9] Una red profesional (o, en un contexto de Internet, un servicio de red profesional) es un tipo de servicio de red social que se enfoca en la interacción y relacionamiento de naturaleza comercial y profesional, en vez de las relaciones personales.

El API de LinkedIn ofrece 2 formas de hacer uso o implementar el API, vía REST o JavaScript, aunque ambas formas ofrecen los mismos métodos de interacción con el core de LinkedIn. A continuación se listan las categorías de los métodos que el API de LinkedIn ofrece:

- **People (Personas)**. Ofrece la posibilidad de acceder a los datos de los usuarios registrados, y a sus conexiones.
- **Share and Social Stream (Compartir y Muro Social)**. Permite a los usuarios de la aplicación que hace uso del API consumir y distribuir contenidos.
- **Groups (Grupos)**. Permite acceder a los contenidos compartidos en los grupos, así como a los datos de los perfiles de usuario que integran el grupo.
- **Comunications (Comunicaciones)**. Permite a los miembros establecer y ampliar sus redes con las invitaciones y mensajes a conexiones directamente dentro de la aplicación que hace uso del API.
- **Companies (Compañías)**. Permite acceder a los perfiles de empresa así como a sus actualizaciones de estado, además permite también hacer actualizaciones de estado de las compañías directamente al muro de seguidores de la empresa.
- **Jobs (Trabajos)**. Permite acceder a los trabajos de los miembros de la plataforma, hacer búsquedas de posiciones de trabajo en compañías, así como ofertas de trabajo publicados y hacer búsquedas de trabajos sugeridos de acuerdo al perfil del usuario.

Las llamadas al API usando REST y los métodos del API son llamados vía URL, algo muy similar a como se hace en la API de Facebook. Por ejemplo: http://api.linkedin.com/v1/people/~:(id)?format=json, los datos resultantes de la llamada son devueltos en colecciones de XML por default, pero como se muestra anteriormente el formato se puede cambiar para que se devuelvan en formato de JSON.

También se pueden especificar los atributos del método que se quieren recuperar mediante un selector de campos, que se envía al método como un parámetro más de la siguiente manera: http://api.linkedin.com/v1/people/~/connections:(id,first-name,last-name,industry). En el Anexo 3 se muestra una tabla con los atributos completos para usuarios y conexiones que pueden ser regresados por el API People de LinkedIn.

2.2 Limitaciones de la extracción de datos de la Redes Sociales en Línea

2.2.1 Limitaciones en la extracción de datos de Facebook

El propósito del API Graph es proporcionar a los desarrolladores externos un marco de trabajo o framework, que les permita crear aplicaciones para interactuar con el core de Facebook, no está diseñada para hacer explotaciones masivas de información, por lo que los resultados arrojados en las llamadas del API, comprenden rangos de información muy limitados, apenas uno o dos días en el caso de los feeds de comentarios o publicaciones (muro de usuarios). Para las búsquedas de objetos de usuarios, páginas y eventos, el API prácticamente proporciona todo el universo de la base de datos actual, la única restricción son los permisos de seguridad que cada objeto tenga configurados.

Para hacer uso del API Graph se debe estar autentificado, Facebook utiliza el protocolo de autenticación llamado OAuth[10], éste sistema genera las llaves (tokens) de permisos que hacen posible la ejecución de los métodos del API.

Como se explicó anteriormente algunas conexiones de objetos requieren de permisos para poder ser consultadas, de esta forma si se quisiera obtener el listado de amigos de un usuario requeriríamos generar un token con la autorización del usuario para hacerlo, es por eso que las aplicaciones actuales que requieren interactuar con Facebook, solicitan a sus usuarios el que se autentifiquen en Facebook y otorguen estos permisos a la aplicación.

El API Graph está en constante evolución al igual que la plataforma de Facebook, lamentablemente las condiciones de privacidad son muy cambiantes también, lo que a veces hace que métodos nuevos dejen de funcionar o bien que cambien en estructura, esto impactará en el diseño de cualquier herramienta de extracción de datos que se desarrolle. En el artículo de "Crawling Facebook for Social Network Analysis Purposes" (Catanese et al. 2011), publicado en mayo del 2011 hacen mención de 2 métodos de extracción de datos de Facebook. "Breadth-first-search sampling" que consiste básicamente en una búsqueda recursiva a través de los listados de amigos de cierto objeto en particular, avanzando en diferentes niveles hasta llegar al nivel indicado para formar la red del objeto; y "Uniform sampling" que requiere la generación de una pila de IDs aleatorios de usuario que se solicitarán a Facebook, si el usuario existe o se tiene acceso a él, el usuario y su lista de amigos se extraen, de lo contrario el usuario se elimina y se procede a analizar el siguiente. Las ventajas de este último enfoque se basan en la independencia de la distribución de IDs de usuarios con respecto a la distribución de amistades en la plataforma de Facebook. Lamentablemente con las nuevas políticas de seguridad, los dos algoritmos anteriores no pueden ser ya implementados, este es un claro ejemplo del impacto de estas restricciones.

En cuanto el número de ejecuciones, Facebook no tiene ningún límite explícito, por lo que es factible lanzar un sin fin de ejecuciones por un mismo token.

2.2.2 Limitaciones en la extracción de datos de Twitter

En el **Streaming API** el flujo es continuo y la velocidad de recepción de tweets dependerán del ancho de banda de los dos extremos de la conexión y la sobrecarga de los servidores de Twitter (Twitter blog 2012).

En el **Search API** y en el **REST API** existe una limitación de 150 peticiones a la hora por usuario o por IP si la llamada no está autenticada, si la llamada esta autenticada el número de llamadas se incrementa a 300 peticiones. También es necesario configurar el tamaño de las páginas para poder aprovechar al máximo las llamadas efectuadas (Twitter blog 2012).

[10] OAuth (Open Authorization) es un protocolo abierto que permite autorización segura de un API de modo estándar y simple para aplicaciones de escritorio, móviles, y web. OAuth permite a un usuario del sitio A compartir su información en el sitio A (proveedor de servicio) con el sitio B (llamado consumidor) sin compartir toda su identidad.

API	Petición	Max. Tamaño Pagina	Max. Total
Search	search	200 tweets	1500 tweets-
REST	status	200 tweets	3200 tweets
REST	friends/ids	5.000 id users	Todos los que exiten
REST	followers/ids	5.000 id users	Todos los que exiten

Tabla 2 Límites de páginas por API de Twitter
***datos tomados de (Twitter blog 2012).**

Es importante saber que aunque todos los tweets residen en la base de datos de Twitter hay una limitación de tiempo para obtenerlos.

API	Limitación temporal	Limitación tamaño
Streaming	Solo tiempo real	No aplica por ser streaming en tiempo real.
Search	Menos de 7 días	1500 últimos tweets
REST	NO	3200 últimos tweets

Tabla 3 Persistencia de datos de las peticiones por API de Twitter
***datos tomados de (Twitter blog 2012)**

2.2.3 Limitaciones en la extracción de datos en LinkedIn.

La principal diferencia del API de LinkedIn con otras redes como la de Facebook y Twitter, es que para hacer uso de ésta, la aplicación tiene que estar registrada en la plataforma.

Al igual que Facebook, LinkedIn usa el protocolo de autenticación abierto OAuth, que genera los tokens para que la aplicación pueda usar el API, pero con la diferencia de que quien se autentifica es la aplicación, no el usuario.

Las aplicaciones de LinkedIn solo permiten 5000 llamadas o ejecuciones del API de manera gratuita, después de esto, si se requieren más búsquedas se tiene que contratar un plan de llamadas del API, que varía en precio de acuerdo al volumen y frecuencia de llamadas. Es posible generar un nuevo registro de la aplicación y obtener otros 5000 accesos, sin embargo eso implica actualizar el ID de la aplicación, haciéndola pasar como si fuera otra.

En cuanto a los datos que el API regresa, dependen directamente de los permisos que cada usuario, compañía o grupo definan en su perfil; aunque para cada objeto existe un conjunto básico de atributos que son de perfil público, y no se requieren permisos explícitos del objeto para poder ser accedidos. A continuación mostramos una tabla con los atributos públicos del perfil People y que pueden ser accedidos sin requerir permisos del objeto:

Field	Parent Node	Description
Id	Person	A unique identifier token for this member
first-name	Person	The member's first name
last-name	Person	The member's last name
maiden-name	Person	The member's maiden name
formatted-name	Person	The member's name formatted based on language
Headline	Person	The member's headline (often "Job Title at Company")
location:(name)	Person	Generic name of the location of the LinkedIn member, (ex: "San Francisco Bay Area")
location:(country:(code))	Person	Country code for the LinkedIn member
Industry	Person	The industry the LinkedIn member has indicated their profile belongs to (http://developer.linkedin.com/node/1011)
Distance	Person	The degree distance of the fetched profile from the member who fetched the profile
relation-to-viewer:(distance)	Person	The degree distance of the fetched profile from the member who fetched the profile
current-share	Person	The member's current share, if set
num-connections	Person	The # of connections the member has
num-connections-capped	Person	*true* if the value of num-connections has been capped at 500. *false* otherwise.
Summary	Person	A long-form text area where the member describes their professional profile
Specialties	Person	A short-form text area where the member enumerates their specialties
Positions	Person	A collection of positions a member has had, the total indicated by a *total* attribute
picture-url	Person	A URL to the profile picture, if the member has associated one with their profile and it is visible to the requestor
site-standard-profile-request	Person	The URL to the member's authenticated profile on LinkedIn (requires a login to be viewed, unlike *public profiles*)
api-standard-profile-request:(headers)	Person	A collection of fields that can be re-used as HTTP headers to request an out of network profile programmatically
public-profile-url	Person	A URL to the member's public profile, if enabled

Tabla 4 Atributos del Objeto People del API LinkedIn (linkedin-Developer 2012)

En caso de requerir acceso a la lista de atributos completa será necesario solicitar permisos de acceso al objeto, que actualmente solo es posible con el objeto Persona y se hace a través de una autentificación en la aplicación.

LinkedIn reserva el derecho de desactivar los perfiles públicos de cada objeto de la plataforma, y también se reserva el poder cambiar, restringir o adicionar atributos de estas colecciones.

Al igual que Facebook, LinkedIn limita la búsqueda de conexiones de un usuario, aunque es posible llegar al listado de conexiones de primer grado cuando un usuario autoriza a la aplicación para acceder a su perfil completo. En búsqueda de publicación no se detectaron límites preestablecidos, aunque en las pruebas del API realizadas se pudo apreciar que solo mostraba actualizaciones del muro de los últimos 15 días.

Al tener versiones para llamadas del API vía REST y JavaScript, puede usarse cualquier lenguaje actual de programación para su implementación. En la documentación de LinkedIn (linkedin-Developer 2012) dan ejemplos de cómo hacer las implementaciones en Python, PHP y Ruby.

3 Análisis de las redes sociales

Según Menéndez "En el análisis de redes se describen y estudian las estructuras relacionales que surgen cuando diferentes organizaciones o individuos interaccionan, se comunican, coinciden, colaboran etc., a través de diversos procesos o acuerdos, que pueden ser bilaterales o multilaterales; de este modo la estructura que emerge de la interrelación se traduce en la existencia de una red social." (Menéndez 2003)

Algunos de los análisis de redes sociales en línea más comunes son:

- Influencia.
- Epidemia viral (Alcance).
- Detección de opiniones positivas, negativas y neutras.
- Información cuantitativa y cualitativa sobre el texto completo escrito por cada persona.
- Evolución de tendencias y pronósticos.

A continuación expondremos sobre dos tipos de análisis, que en la presente investigación se consideraron importantes de acuerdo al experimento que se explicará en el Capítulo 4.

3.1 Análisis de Influencia

Con la rápida adopción de las redes sociales por las personas, como un medio de obtener información y participar en la opinión pública de sucesos, productos, eventos, etc., para las empresas y otras entidades públicas, la identificación de los autores influyentes en los medios sociales es fundamental, ya que las opiniones que expresan pueden extenderse rápidamente por todas partes.

Christiankis y Fowler en su libro "Conectados" hacen una reflexión sobre la influencia de las redes sociales: "Las redes sociales difunden felicidad, generosidad y amor. Siempre están ahí, ejerciendo una influencia sutil y al mismo tiempo determinante en nuestras elecciones, acciones, pensamientos, sentimientos y también en nuestros deseos. Además, esas conexiones no terminan en las personas que conocemos; más allá de nuestros horizontes sociales, los amigos de los amigos de nuestros amigos pueden impulsar reacciones en cadena que acaben por alcanzarnos" (Christakis & Fowler 2009).

La influencia social determina que una idea, un comportamiento o un producto se difundan a través de las redes sociales como una epidemia. Es la influencia que otros ejercen sobre nosotros para actuar de manera similar, ya sea en la moda, en la adopción de una tecnología, en una acción de Marketing Viral, en el boca a boca, etc. Son muy pocos los autores que han expresado de una manera tan clara esta serie de conceptos, uno de ellos es Malcolm Gladwell en su libro "The Tipping Point", en el que realiza un análisis sobre el funcionamiento de las "epidemias sociales". Para Gladwell, "la respuesta radica en que el éxito (alta influencia) de una epidemia social depende enormemente de la participación en ella de un cierto tipo de persona, dotada de unos rasgos especiales y poco habituales" (Gladwell 2000)

Gladwell hace una clasificación de la clase de personas que desempeñan un papel crucial en la propagación de cualquier información en un contexto particular (Gladwell 2000):

- **Conectores**: Son personas muy hábiles para unir al mundo. Conoce a muchas personas que creen en él, a la gente que hace falta, gente influyente que pertenece a diversos grupos. Es muy sociable y le gusta estar en el centro de los acontecimientos.
- **Enterados**: Son especialistas en información. Son el tipo de personas al que recurrimos cuando tenemos un problema. Los "enterados" están constantemente buscando y compartiendo información, les encanta dar consejos sin esperar nada a cambio, asumen seriamente el desafío de dar buenos consejos y se sienten motivados por ayudar y enseñar.
- **Vendedores**: Son los convencedores del mundo.
- **El poder del contexto**: Toda epidemia social está sujeta a las condiciones y circunstancias del momento y del lugar en el que ocurre.

Tanto conectores como enterados y vendedores son necesarios para desatar una epidemia (efecto de influencia). Con eso, nos aproximamos a los factores que debemos tener en cuenta a la hora de medir la influencia de determinadas personas en las redes sociales. Factores, que van mucho más allá de un número X de seguidores en Twitter y Facebook. No deberíamos preocuparnos tanto por el número de seguidores sino más bien por quiénes son aquéllos que nos siguen y cómo se encargan de propagar nuestros mensajes.

Los medios sociales han revolucionado muchas cosas, pero uno de sus mayores impactos ha sido el potenciar la influencia y el papel que determinadas personas juegan sobre otras.
Es necesario comprender una serie de ideas fundamentales (Gladwell 2000):

- La influencia sin contexto es irrelevante.
- La popularidad no es lo mismo que la influencia.
- Conectar y conversar no tiene nada que ver con la actividad y la gestión de las distintas herramientas y plataformas de la web social.

3.2 Desambiguación de perfiles (profile matching) en diferentes Redes sociales en Línea.

Con la creciente popularidad y el uso de las redes sociales en línea, se ha observado que los medios de comunicación y las personas ahora tienen cuentas (algunos varias veces) en múltiples y diversos servicios como Facebook, LinkedIn, Twitter y YouTube. Veremos una nueva generación de aplicaciones de multiposting que se integran en los medios sociales y un mayor uso de las API de redes sociales profesionales.

La información públicamente disponible se puede utilizar para crear un rastro digital de cualquier usuario utilizando las redes sociales en línea. La generación de este rastro digital puede ser muy útil para la identificación de perfiles en diferentes redes sociales, es decir desambiguar los perfiles que pertenecen al mismo usuario en diferentes redes sociales, puede ser útil también para gestionar estos perfiles, o para detectar comportamientos maliciosos de usuarios.

La búsqueda semántica es muy útil para empresas cuyo *core* sea la información misma: medios de comunicación, policía, abogados, etc.; y puede aplicarse a muchos otros ámbitos,

como para administrar una base de datos de comentarios, o la administración de currículo (en el caso de una empresa que se dedique al headhunting). A nivel personal, cuando buscamos información de cualquier tipo en internet seguramente nos facilitará más las búsquedas y perderemos menos tiempo desechando links que no nos sirvan.

Actualmente la desambiguación de perfiles (profile matching) y la búsqueda semántica va a ser un tema de debate y se convertirá en una obligación para bolsas de empleo y los sistemas de ATS[11], así como para los diversos estudios o análisis en las redes sociales en línea.

Una aplicación muy importante de análisis de rastros digitales de los usuarios, es la de proteger a los usuarios de posibles riesgos de privacidad y seguridad que surgen de la enorme información que tiene el usuario a disposición del público. También puede ser usada en la identificación de promotores u organizadores de eventos, organizados a través de las redes sociales.

3.3 ¿Qué es el Social network analysis software?

Los Software de Análisis de Redes Sociales (SNA software) facilitan el análisis cuantitativo o cualitativo de las redes sociales, lo hacen mediante la descripción de las características de una red, ya sea a través de la representación numérica o visual. [12]

Existen una gran variedad de este tipo de software, en el Anexo 4 podemos encontrar una lista de ellos (en inglés) y sus principales características, sin embargo aquí ahondaremos en un par de ellos tratando de explicar cómo están siendo usados para hacer minería de datos de las redes sociales y sus funcionalidades.

3.3.1 NodeXL

NodeXL es una plantilla gratuita (Add-in) y de código abierto para Excel 2007/2010, desarrollado en .Net con el lenguaje C# que permite el análisis y visualización gráfica de redes, calcula un conjunto básico de indicadores de la red. Soporta la extracción de datos desde correo electrónico, también de redes sociales como Twitter, YouTube, Facebook, hyperlinks de WWW y Flicker. NodeXL genera a partir de los datos importados, listas de nodos y vértices que representa gráficamente. Permite una fácil manipulación y filtrado de los datos subyacentes en formato de hoja de cálculo. Permite también tener múltiples planos de la red de visualización. NodeXL lee y escribe archivos de UCINET[13] así como de GraphML, que consiste en un archivo XML que contienen los elementos de los nodos y vértices en una lista ordenada (GraphML 2012); además permite la importación de datos desde archivos tipo .CVS (Archivos de texto, separados por comas o caracteres especiales).

---

[11] ATS (Applicant Tracking System).

[12] Vista en http://en.wikipedia.org/wiki/Social_network_analysis_software#cite_note-7

13 UCINET es un paquete de software para el análisis de los datos de las redes sociales. Fue desarrollado por Lin Freeman, Everett Martin y Steve Borgatti. Viene con la herramienta de visualización NetDraw UCINET (2012). "UCINET Software." visto en Julio de 2012, en https://sites.google.com/site/ucinetsoftware/home.

Como se mencionó anteriormente NodeXL tiene módulos que facilitan la importación de diferentes redes sociales, sin embargo para objeto de esta investigación se revisaron los módulos de importación de Twitter que a continuación se reseñan (Vladimir Barash & Golder + 2011):

NodeXL tiene 3 métodos principales de extracción de datos de Twitter:

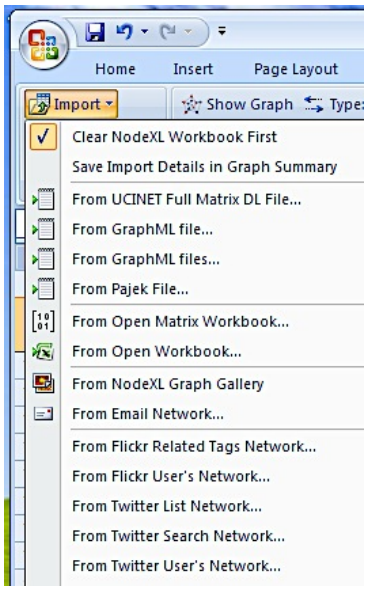

Ilustración 3 Menú de Importación de datos de NodeXL

- *Import from Twitter List Network.* Se refiere a que puede generar un conjunto de resultados a partir de una lista de usuarios definida en alguna cuenta de Twitter o bien una lista de nombres de usuarios manualmente definidos.
- *Import from Twitter Search Network.* Esta opción obtiene la red de personas en cuyas publicaciones (de tweets) aparece una o varias palabras en específico.
- *Import from Twitter User´s Network*. Esta opción obtiene la red de personas que son seguidores de un usuario de Twitter o las personas a las que este usuario sigue; es posible también definir ambas.

Una vez que se selecciona el método, las opciones son muy simples de usar, solo hay que especificar la lista de usuarios, palabras clave o bien el usuario específico a buscar. A continuación se especifican los parámetros.

En el caso de "*Import from Twitter User´s Network*" se debe especificar si se deberán buscar los usuarios que siguen al usuario, o bien los que el usuario sigue o ambas opciones, el análisis de la red esta versionada en 3 tipos, nivel 1.0 la cual forma una red de los usuarios conectados en primer grado, el nivel 1.5 que forma una red de usuarios conectados en primer grado y las relaciones existentes entre los nodos encontrados, y el nivel 2.0 que forma una red de hasta segundo grado y las relaciones entre los diferentes nodos de la red.

Representación gráfica de los niveles de búsqueda:

Una vez que se ha seleccionado el tipo y parámetros de la búsqueda, es necesario especificar el tipo de cuenta que se usará para invocar el API de Twitter, se recomienda usar una cuenta autentificada de Twitter, ya que éstas están menos restringidas en el uso o ejecución de múltiples consultas; también es posible hacer las búsquedas sin una cuenta de Twitter pero esto incrementa el tiempo de extracción de datos y no es recomendable.

Una vez que NodeXL termina de ejecutar las búsquedas, genera un workbook de Excel que permite ver el listado de nodos y vértices que integran nuestra red, según los parámetros y métodos seleccionados, este workbook nos muestra una lista con los siguientes campos:

Campo	Tipo de dato	Descripción
Usuario	String	Usuario encontrado en la red según los parámetros utilizados
imagen file	String	Ruta URL a la imagen del avatar del usuario
Followed	Integer	Número de personas a las que el usuario sigue
Followers	Integer	Número de personas que siguen al usuario
Favorited	Boolean	Indica si el Twitter ha sido marcado como favorito por el usuario autentificado.
Tweets	Int64	Número de tweets hechos por el usuario
Favorities	Integer	Número de tweets marcados como favoritos
Description	String	Campo de descripción del usuario.
Location	String	Coordenadas geográficas del origen del tweet
Web	String	Página web del usuario
Time Zone	String	Uso horario del usuario
Joined Twitter Date (UTC)	String	Indica desde cuando el usuario usa Twitter.
Latest Tweet	String	Texto del último tweet del usuario
URLs in Latest Tweet	String	Url usados en los últimos tweets o el último.
Hashtags in Latest Tweet	String	Hastags usados en el último tweet del usuario.
Latest Tweet Date (UTC)	Date	Fecha del último tweet hecho por el usuario

Tabla 5 Lista de campos estándar generada por NodeXL (Twitter Dev 2012).

NodeXL también permite desarrollar módulos para extender la funcionalidad de los diferentes módulos de importación, al ser un proyecto de código abierto pueden usarse scripts de C# o bien Visual Basic, para ejecutar llamadas al API de Twitter desde un framework conjunto de librerías de NodeXL en conjunción con XML, esto permitiría poder importar a NodeXL diferentes conjuntos de datos, diferentes columnas y agrupaciones de datos, a los que nativamente la herramienta ofrece (Derek L. Hansen et al. 2011a).

3.3.2 Gephi

Gephi es una plataforma de visualización interactiva y de exploración de todo tipo de redes y sistemas complejos, dinámicos y gráficos jerárquicos. Es una herramienta para las personas que tienen que explorar y comprender los gráficos. El usuario interactúa con la representación, permite manipular las estructuras, formas y colores para revelar las propiedades ocultas. Utiliza un motor de render 3D para mostrar las grandes redes en tiempo real y acelerar la exploración. Gephi tiene una arquitectura flexible y multi-tarea, ofrece nuevas posibilidades para trabajar con conjuntos de datos complejos y produce valiosos resultados visuales.

Gephi permite la importación de datos en los formatos GraphViz(.dot), Graphlet(.gml), GUESS(.gdf), LEDA(.gml), NetworkX(.graphml, .net), NodeXL(.graphml, .net), Pajek(.net, .gml), Sonivis(.graphml), Tulip(.tlp, .dot), UCINET(.dl), yEd(.gml), Gephi (.gexf), Edge list(.csv), databases. Es posible construir plugins (utilizando el lenguaje de programación Java) que ayuden con la importación de los datos desde diferentes redes sociales, haciendo uso de las diferentes APIs de extracción de datos de las redes sociales; aunque Gephi es de código abierto existen pocas referencias a las librerías que expone el core de la aplicación, lo que dificulta el desarrollo de estos plugins. En la página web de este software se pueden encontrar algunos ejemplos de cómo se hace una importación de datos desde Twitter, sin embargo se hace a través de un formato de intercambio de datos de UCINET(.dl), el ejercicio muestra cómo construir el archivo con el estándar, para de ahí importarlo a Gephi y hacer el análisis gráfico de la red.

Una ventaja que tiene Gephi en el análisis gráfico, es que el motor de render de las gráficas está basado en OpenGL[14], lo que permite aprovechar el poder de la aceleración gráfica por hardware, haciendo al software más veloz y permitiéndole mostrar gráficas de gran tamaño, en número de nodos, aristas y vectores.

3.4 Métricas obtenidas en Social network analysis software

Actualmente el análisis de las redes sociales se ha convertido en una metodología clave en las modernas ciencias sociales, entre las que se incluyen la sociología, la antropología, la psicología social, la economía, la geografía, las ciencias políticas, los estudios de comunicación, estudios organizacionales y la mercadotecnia. Las instituciones públicas y privadas cada vez con mayor frecuencia, están dedicando equipos de trabajo en el análisis y recolección de datos de las redes sociales.

Las métricas dependerán en gran medida de las preguntas a resolver con el análisis de los datos arrojados por las diferentes búsquedas, sin embargo existen ya una serie de medidas estándar que nos dan información de cómo está integrada una red (Derek L. Hansen et al. 2011b).

[14] OpenGL (Open Graphics Library) es una especificación estándar que define una API multilenguaje y multiplataforma para escribir aplicaciones que produzcan gráficos 2D y 3D.

Algunas métricas pudieran ser subjetivas como la influencia y la epidemia (viralidad) de los tweets, ya que dependen de múltiples factores en los que influyen desde quién hace los retweets, y en este caso puede suceder que el tweet sea atribuido a otro usuario aunque éste no haya sido el origen. También hay que tomar en cuenta que en Twitter existen cuentas llamadas bots, que son cuentas automáticas y que no tienen a una persona detrás de ellos, lo que pudiera afectar el nivel real de la influencia que un usuario pudiera generar en la red.

Algunas de las métricas comunes en el análisis de redes sociales que podemos mencionar son (Russell 2011; Derek L. Hansen et al. 2011b): [15]

Intermediación

La medida refleja el número de personas que una persona conecta indirectamente a través de sus vínculos directos. Es la medida en que un nodo se encuentra entre los demás nodos en una red. Esta medida toma en cuenta la conectividad de los vecinos del nodo, dando un mayor valor a los nodos que conectan a grupos.

Conector

Un lazo puede ser llamado conector si su eliminación causa que los puntos que conecta se transformen en componentes distintos de un grafo.

Centralidad

"Intermediación", "Cercanía", y "Grado" son medidas de centralidad. Esta medida da una idea aproximada del poder social de un nodo basándose en lo bien que se "conecte" éste a la red.

Centralización

La diferencia entre el número de enlaces para cada nodo, dividido entre la cantidad máxima posible de diferencias. Una red centralizada tendrá muchos de sus vínculos dispersos alrededor de uno o unos cuantos puntos nodales, mientras que una red descentralizada es aquella en la que hay poca variación entre el número de enlaces que cada nodo posee.

Cercanía

Es el grado en que un nodo está cerca de todos los demás (nodos) en una red (directa o indirectamente). La cercanía es la inversa de la suma de las distancias más cortas entre cada nodo y cada una de los otros nodos en la red.

Coeficiente de agrupamiento

Una medida de la probabilidad de que dos personas (nodo) vinculadas a un nodo se asocien a sí mismos. Un coeficiente de agrupación más alto indica un mayor «exclusivismo».

[15] Visto en http://es.wikipedia.org/wiki/Red_social

Cohesión
El grado en que los actores se conectan directamente entre sí por vínculos cohesivos.

Grado
El recuento del número de vínculos con otros actores en la red. Véase también grado (teoría de grafos):

Densidad (nivel individual)
El grado de relaciones de conocerse unos a otros / proporción de lazos entre las menciones de un individuo. La densidad de la red, o densidad global, es la proporción de vínculos en una red en relación con el total de vínculos posibles.

Flujo de centralidad de intermediación
El grado en que un nodo contribuye a la suma del flujo máximo entre todos los pares de nodos excluyendo ese nodo.

Centralidad de Eigen-vector (Autovector)
Una medida de la importancia de un nodo en una red. Asigna puntuaciones relativas a todos los nodos de la red basadas en el principio de que las conexiones a los nodos que tienen una puntuación más alta, contribuyen más a la puntuación del nodo en cuestión.

Estas son las métricas más comunes que podemos encontrar sin embargo, no siempre se hace uso de ellas. Dependerá del tipo de análisis que se requiera y de los datos de la red social.

4 Experimento de Análisis de Redes Sociales en Línea

4.1 Definición del Experimento

Una vez que hemos revisado las múltiples técnicas de recolección de datos, así como mencionado algunos de los tipos de análisis que pueden hacerse sobre los datos obtenidos de la redes sociales en línea, surge la necesidad de comprobar y definir un método simple pero efectivo, que permita hacer análisis de los datos generados en las redes sociales, permitiéndonos identificar tendencias y poner a prueba diferentes métodos de análisis comúnmente usados por instituciones, empresas y aplicaciones comerciales.

Para realizar el experimento se seleccionaron las dos redes sociales con mayor impacto en la actualidad, Facebook y Twitter. El experimento que se determinó hacer, se refiere al análisis de influencia en Facebook y Twitter de los candidatos a la alcaldía de la ciudad de Zacatecas en los comicios del año 2013, después de identificar la influencia de los candidatos en las redes sociales, se realizó un análisis de correlación entre la influencia de cada uno de los candidatos (posts hechos en Facebook y Twitter) y los resultados de la votación. Se decidió extraer los datos de Facebook y Twitter, por ser las más populares o usadas en México (AMIPC 2012; IAB-México 2013).

Además del análisis de la influencia se analizaron también algunos métodos de desambiguación de perfiles, con el objetivo de identificar seguidores de los candidatos que tuvieran perfiles en ambas redes.

Desambiguar múltiples identidades en línea de los usuarios tiene muchas ventajas, por ejemplo: gestión de perfiles, la gestión de la creación y la construcción de un perfil en una red social global, ayudar a supervisar usuarios y controlar la fuga de información personal, la portabilidad de perfil de usuario, y la personalización. Además, al unir múltiples perfiles en línea de los usuarios, puede facilitar el análisis a través de diferentes redes sociales lo cual puede ayudar a detectar y proteger a los usuarios de diversas amenazas de privacidad de seguridad que surgen debido a la vasta cantidad de información que el usuario pone a disposición del público (Malhotra et al. 2012). La detección de múltiples identidades se realizó también con la intención de unificar el universo de influencia de cada candidato, un puente entre las dos redes sociales (Facebook y Twitter). A partir de los datos obtenidos concluimos que es posible considerar los datos como si se tratasen de una sola red de influencia aunque hablemos de dos redes diferentes (Facebook y Twitter).

4.2 Extracción de datos de la Redes Sociales en Línea

En esta sección describiremos cómo se formaron los conjuntos de datos usados en los diferentes tipos de análisis del experimento. Inicialmente se procedió a identificar los perfiles de los candidatos, tanto en Facebook como en Twitter. A continuación se muestra una lista con los nombre de los candidatos:

Rogelio Cárdenas Hernández
Salvador Llamas
Fernando Bueno
Carlos Peña Badillo
Rogelio Lara
Xerardo Ramirez Muñoz
Martin Uvario Gaspar

Tabla 6 Lista de candidatos a la alcaldía de Zacatecas, México

Lamentablemente, pudimos comprobar que el uso de redes sociales no está extendido uniformemente, de hecho el candidato Martin Uvario Gaspar se tuvo que eliminar de la lista de Twitter ya que no contaba con un perfil, y en el caso de Facebook se encontraron Fan Pages[16] de cada uno de los candidatos. Se encontró que algunos de los candidatos tenían 2 o más perfiles, o que hacían suponer que les pertenecía, sin embargo se optó por tomar aquellas Fan Pages con más actividad; se consideraron como perfiles fallidos a los que no mostraban actividad en comparación de los perfiles paralelos, o como sospechosos de ser bots[17], esto debe ser considerado ya que al tratarse de un tema político es fácil que existan campañas negras en contra de los candidatos.

La extracción de los datos se hizo a través de las APIs proporcionadas por cada una de las redes, utilizando llamadas de tipo REST, y utilizando el lenguaje de programación PHP 5.3. Los datos fueron recolectados del 1 de Junio de 2013 al 5 de julio de 2013, ya que las elecciones se efectuaron el día 7 de julio de 2013. Los datos de los votos obtenidos por cada candidato fueron obtenidos de la página del IEEZ[18] el día 9 de julio de 2013.

Inicialmente se pretendió formar estructuras de redes sociales por cada uno de los candidatos hasta un segundo grado, lamentablemente las limitaciones impuestas por las políticas de seguridad de Facebook impidieron hacerlo para ésta red social. En Facebook tampoco se pudo acceder a la lista de followers de cada "fan page" directamente, por lo que la recolección de datos se hizo a través de los "likes" hechos por los followers a los post del candidato en dicha página. Se obtuvieron 9,633 registros de seguidores de las Fan Pages de los candidatos, de un total de 27,937 usuarios que le dieron like a la fan page al día 5 de

16 Son páginas públicas. A diferencia de los perfiles personales, las Fan Pages pueden ser de acceso público. Es decir, no necesariamente tienes que ser "amigo" de la persona ni tener una cuenta en Facebook para acceder a ellas.

17 Los bots o robots son programas automatizados que generan mensajes de Twitter y cumplen diversas funciones. Algunos permiten programar mensajes para su difusión a determinada hora, replican tweets con palabras claves o emiten spam (correo basura).

18 Instituto Electoral del Estado de Zacatecas.

julio. Se obtuvieron también 970 post que fue de donde se extrajeron los datos de los seguidores.

En el caso de Twitter si se logró hacer la extracción de datos directamente y recursivamente, aunque con la limitante de 5000 followers en primer nivel y 5000 followers de segundo nivel por cada follower de primer nivel.

El total acumulado de seguidores de primer y segundo grado en Twitter de los 7 candidatos fue de 2,589,636 registros, se trató de obtener para cada uno de ellos el Id de Twitter, Nombre Completo, el Screen_name o Nick, la Localidad y Fecha de Inclusión, así como Número de Seguidores. Se almacenó la información de Twitter en una base de datos, ésta base de datos cuenta con un total de Id de Twitter de 2,589,636, de los cuales sólo se obtuvieron los datos completos de 821,659; esto debido a que existe una restricción de privacidad que impide la extracción de los datos de algunos de los seguidores, así como en el número de peticiones (consultas) que se pueden hacer por hora al API de Twitter, el límite es de 100 llamadas por hora, lo que repercutió bastante en que se pudieran extraer el total de los datos para el número de seguidores recolectados, aun así se consideró como una muestra suficiente para los efectos del experimento.[19]

A continuación se muestra una tabla con el resumen de los datos recolectados:

	Candidatos							
	Rogelio Cárdenas Hernández	Salvador Llamas	Fernando Bueno	Carlos Peña Badillo	Xerardo Ramírez Muñoz	Rogelio Lara	Martin Uvario Gaspar	Total
Votos	***2,447***	***1,904***	***7,554***	***18,442***	***16,431***	***920***	***1,761***	***49,459***
FACEBOOK								
FB Followers	1,505	4,142	7,447	9,585	4,192	828	238	27,937
FB followers en DB	613	390	2,354	2,708	3,250	203	115	9,633
FB Post	127	138	162	159	174	163	47	970
TWITTER								
Followers de 1er grado	*191*	*285*	*1,219*	*2,011*	*1,747*	*732*	*NA*	*6,185*
Twitters	*210*	*249*	*272*	*576*	*311*	*645*	*NA*	*2,263*
Followers de 2do grado	*97,297*	*207,750*	*733,405*	*495,571*	*619,547*	*429,881*	*NA*	*2,583,451*
E1	*52*	*42*	*133*	*532*	*105*	*23*	*NA*	*887*
E2	*1,004*	*126*	*764*	*2,364*	*397*	*46*	*NA*	*4,701*
Retwets posibles	*40,110*	*70,965*	*331,568*	*1,158,336*	*543,317*	*472,140*	*NA*	*2,616,436*

E1 = Followers de 1er grado que alguna vez han retweteado un tweet de un candidato.
E2= Total de veces que se ha retweteado un tweet de un candidato por los usuarios de 1er grado.

Tabla 7 Estadísticas de la extracción de datos de las redes sociales para cada candidato.

En la tabla 7 encontramos los datos obtenidos por cada uno de los candidatos en cada una de la redes sociales (Facebook y Twitter). Para el caso de Facebook al ser extraídos los datos desde los post públicos de las fan pages de cada candidato, se pudieron obtener registros de

[19] Se considera una muestra suficiente para el experimento ya que los datos recolectados de Twitter representan mas de un 98% de usuarios de primer grado de followers de Twitter por candidato. Para el caso de Facebook la muestra variaron entre el 10% de followers para el candidato con menor recuperación, hasta un 77% para el candidato con mayor recuperación de registros de followers.

contactos de primer grado, también se contabilizaron el número de post publicados en el periodo de tiempo. El primer dato es el de followers y es el total de followers que el candidato registraba en el contador de la fan page al día 5 de julio, el segundo dato son los followers que se pudieron recuperar del total de followers, el tercer dato representa el número de post que permitió obtener la información de los followers totalizados en el segundo dato. Así por ejemplo tenemos que para Rogelio Cárdenas se obtuvieron 613 registros de followers extraídos de 127 post, pero el total de followers (usuarios que le dieron "me gusta" a la fan page del candidato) al día 5 de julio era de 1505. La diferencia pudo ser originada debido a que no todos los followers le dieron like a alguno de los 127 post. Como ya se mencionó anteriormente, lamentablemente por cuestiones de privacidad no se pudo acceder a la lista de followers de la fan page directamente, por eso se optó por hacerlo a través de los post, que al ser públicos convierten en pública la información relacionada con el post, de esta manera se hace público quien le da like al post también.

En el caso de Twitter, se obtuvo el número de followers totales al día 5 de julio por candidato, dato que contrastó con el total de registros con información completa en la base de datos, ya que algunos de los followers no permitieron la extracción de sus datos; ya sea por restricciones de privacidad o bien por que la cuenta ya no existía aunque los Id de los followers si permanecían asociados al candidato. De esta manera podemos ver como en el caso de Xerardo Ramírez solo se obtuvieron los datos completos de 1747, cuando el total de followers indicaba que eran 1856. Los demás datos son algo simples, como el número de tweets hechos por el candidato Xerardo, 311 por ejemplo, el número de followers de segundo grado, que básicamente son los followers de los followers del candidato 619,547 de 1747 followers de primer grado, tenemos también followers de 1er grado que alguna vez han retweteado un tweet de un candidato 105, otro dato es el total de veces que se ha retweteado un tweet de un candidato por los usuarios de 1er grado 397 y el último dato representa el número de retweets que hipotéticamente pudieron ser posibles si todos los usuarios de primer grado hubieran hecho retweet a cada uno de los tweets de los candidatos, 543,317 para el caso de Xerardo Ramírez.

4.3 Definición de algoritmos para encontrar Desambiguación de perfiles ó coincidencia de perfiles en diferentes Redes Sociales en Línea (Matching)

Uno de los objetivos iniciales de la investigación fue el de encontrar mecanismos o algoritmos fiables para encontrar coincidencias de perfiles o desambiguación de perfiles en diferentes redes sociales, es decir usar información de un usuario en Twitter para encontrar su identidad en Facebook.

La identidad de un usuario en una red social incluye un conjunto de atributos del perfil, los cuales proporcionan información básica sobre el usuario, como nombre de usuario, nombre completo, ubicación, género, descripción, etc. Es posible que los usuarios mantengan estos atributos en la creación de sus perfiles en diferentes redes sociales. Esto pudiera permitir hacer uso de estos tributos para identificar al usuario entre las diferentes redes sociales. Twitter tiene un número limitado de atributos a disposición del público, mientras que Facebook tiene un conjunto de atributos mayor, sin embargo son privados. Aun así el username y el nombre se consideraron los atributos más discriminantes para desambiguar perfiles de usuario (Malhotra et al. 2012). Es por ello que los campos que se usaron para la comparación fueron el username de twitter y el screen_name de Facebook que son los campos que se usan en las diferentes redes para el apodo del usuario, el otro campo comparado fue el del nombre completo, que se normalizó para que incluyera los dos apellidos de los usuarios.

Para hacer esta comparación de atributos, se utilizaron un par de algoritmos simples, pero potentes para hacer comparaciones entre cadenas de texto, estos algoritmos son: Similar Text y la Distancia de Levenshtein. Estos algoritmos han sido utilizados ampliamente en diferentes comparativas de algoritmos para identificar cadenas de texto duplicadas, y han mostrado ser muy eficientes y fáciles de implementar (Dănăilă et al. 2012), por esta razón fueron seleccionados para ver si son efectivos detectando usuarios comunes en las diferentes redes sociales.

La implementación del método de Similar Text se realizó en PHP, el método arroja como resultado una ponderación porcentual de la distancia, o mejor dicho del porcentaje de igualdad entre los campos comparados (Username vs. Screen_name, y nombre completo de ambas redes). En el caso de la Distancia de Levenshtein se tuvo que desarrollar una fórmula para obtener el indicador en porcentaje de semejanza.

El primer intento para ejecutar esta fase del experimento, fue hacer una comparación de todos los registros de usuarios de Facebook versus los registros de usuarios obtenidos en Twiter de todos los candidatos, sin embargo esto fue descartado debido a que no se le vio mucho valor el hacer comparaciones de username y nombres muy dispares; además de que el número de comparaciones se disparaba a cerca de 3,435,700,818. También se intentó hacer una clasificación previa de los datos a comparar, utilizando el algoritmo Soundex[20] ,

[20] Soundex es un algoritmo fonético, un algoritmo para indexar nombre por su sonido, al ser pronunciados en Inglés. El objetivo básico de este algoritmo es codificar de la misma forma los nombres con la misma pronunciación. Soundex es el algoritmo fonético conocido más ampliamente y es usada en ocasiones (de forma incorrecta) para describir el "algoritmo fonético".

lamentablemente la inclusión de caracteres especiales (principalmente en los username de las cuentas de Twitter) impidieron que ésta clasificación pudiera ser utilizada. Finalmente se optó por hacer una clasificación alfabética simple; de esta forma los usuarios de Twitter cuyo username iniciara con A solo serían comparados con los usuarios de Facebook con screen_name iniciado con A, así sucesivamente hasta llegar a la Z. Se excluyeron todos los username iniciados con números, ya que en Facebook no hubo ni un solo registro de usuario cuyo username iniciara con números. También fueron descartados todos los username que iniciaran con los códigos ASCII[21] 35, 39 registros en Twitter y 216 registros en Facebook.

Letra	Codascii	Registros Tw	Registros FB	Total comparaciones
A	65	81278	1146	93144588
E	69	41008	552	22636416
R	82	40983	472	19343976
S	83	47584	398	18938432
D	68	39786	340	13527240
P	80	37962	346	13134852
G	71	31700	361	11443700
F	70	28298	337	9536426
B	66	28166	216	6083856
I	73	24581	229	5629049
K	75	21320	250	5330000
N	78	22068	234	5163912
V	86	19312	238	4596256
T	84	28155	130	3660150
H	72	17180	146	2508280
O	79	14602	162	2365524
Y	89	13272	166	2203152
Z	90	5806	78	452868
W	87	8081	33	266673
U	85	5563	26	144638
X	88	3737	32	119584
Q	81	2858	14	40012
Total de Comparaciones				240,269,584

Tabla 8 Agrupamiento y número de comparaciones del análisis de desambiguación.

[21] ASCII (acrónimo inglés de American Standard Code for Information Interchange — Código Estándar Estadounidense para el Intercambio de Información), pronunciado generalmente [áski] o [ásci] , es un código de caracteres basado en el alfabeto latino, tal como se usa en inglés moderno y en otras lenguas occidentales.

4.4 Definición de algoritmo y obtención de métricas de análisis de influencia

Muchos de los estudios actuales de las redes sociales así como de las métricas de análisis de redes que existen, están enfocados en medir la influencia de los usuarios o actores relevantes de una comunidad o grupo en particular. Es por esto que en esta parte del experimento nos enfocamos en identificar la influencia generada por cada uno de los candidatos a la alcaldía de Zacatecas en Facebook y Twitter y su correlación con los resultados de la elección.

El número de seguidores ha sido tradicionalmente la métrica más usada por muchas de las soluciones comerciales, así como por agencias de mercadotecnia para determinar la influencia de un actor en una red social, sin embargo, a diferencia de un sitio web, blog o post al que le dan like, no hay forma de trazar la acción de si alguien realmente leyó el Tweet o simplemente lo dejó pasar. En Facebook el propietario de una fan page puede tener miles de seguidores, sin embargo puede ser que sólo algunos le den like a sus post, y hay una alta posibilidad de que siempre sean las mismas personas haciendo esto. Es por eso que se decidió cruzar varios indicadores de influencia con el fin de hacer una aproximación más real a la verdadera influencia de los candidatos, y ver la correlación entre cada factor de influencia y los resultados finales de la votación.

Para llevar a cabo ésta fase, se obtuvieron los registros de los usuarios que dieron al menos un retweet o like en alguno de los post del candidato en las diferentes redes sociales. En el caso de Facebook como los registros se extrajeron a partir de los likes de los post en las fan pages, prácticamente fueron todos los registros de cada candidato, sin embargo para el caso de Twitter sólo se seleccionaron los followers de 1er nivel que hicieron algún retweet a cualquier tweet del candidato del 1 de junio al 5 de julio de 2013.

La metodología para medir la influencia consistió en agrupar solo los usuarios que hicieron algún retweet del candidato, este se multiplicó por el número de followers del usuario y se dividió entre la sumatoria de los followers de primer y segundo nivel de todos los candidatos.

Para hacer más normal la muestra a analizar, se decidió eliminar los valores atípicos[22], por ésta razón se eliminaron los usuarios de comités nacionales de los partidos de cada uno de los candidatos, así como cuentas de periódicos o medios de comunicación, lo anterior con la finalidad de dejar en el universo de datos sólo a personas que de cierta manera se consideraron como posibles votantes.

[22] En estadística, un valor atípico es una observación que es numéricamente distante del resto de los datos. Las estadísticas derivadas de los conjuntos de datos que incluyen valores atípicos serán frecuentemente engañosas. Los valores atípicos pueden ser indicativos de datos que pertenecen a una población diferente del resto de la muestra establecida.

La fórmula para calcular la influencia fue:

$$FI = \Sigma\,(nrFC*(1+fllwrFFC))/\,\Sigma\,fllwTC)$$

Donde:
FI= Factor de influencia.
nrFC: Representa el número de retweets efectuados por cada follower del candidato, de alguno de los tweets (publicaciones) hechos por el candidato.
fllwrFFC: Representa el número de followers que tiene el follower del candidato.
Σ fllwTC: Representa el total de followers a los que llegaron los retweets de todos los candidatos (éste factor nos permitió tener un universo común para la comparación de influencia entre todos los candidatos).

4.5 Resultados del Experimento.

4.5.1 Análisis de desambiguación (matching)

El primer análisis que se realizó fue el de desambiguación de perfiles. De las 240,269,584 comparaciones se obtuvo un conjunto de 810 registros con un porcentaje de desambiguación superior al 90% en las comparaciones del atributo de nombre completo, la comparación del atributo de username versus scree_name fue menos consistente, a continuación se muestra una gráfica de tendencias de estos registros, según los algoritmos y comparaciones usadas para obtener las desambiguaciones:

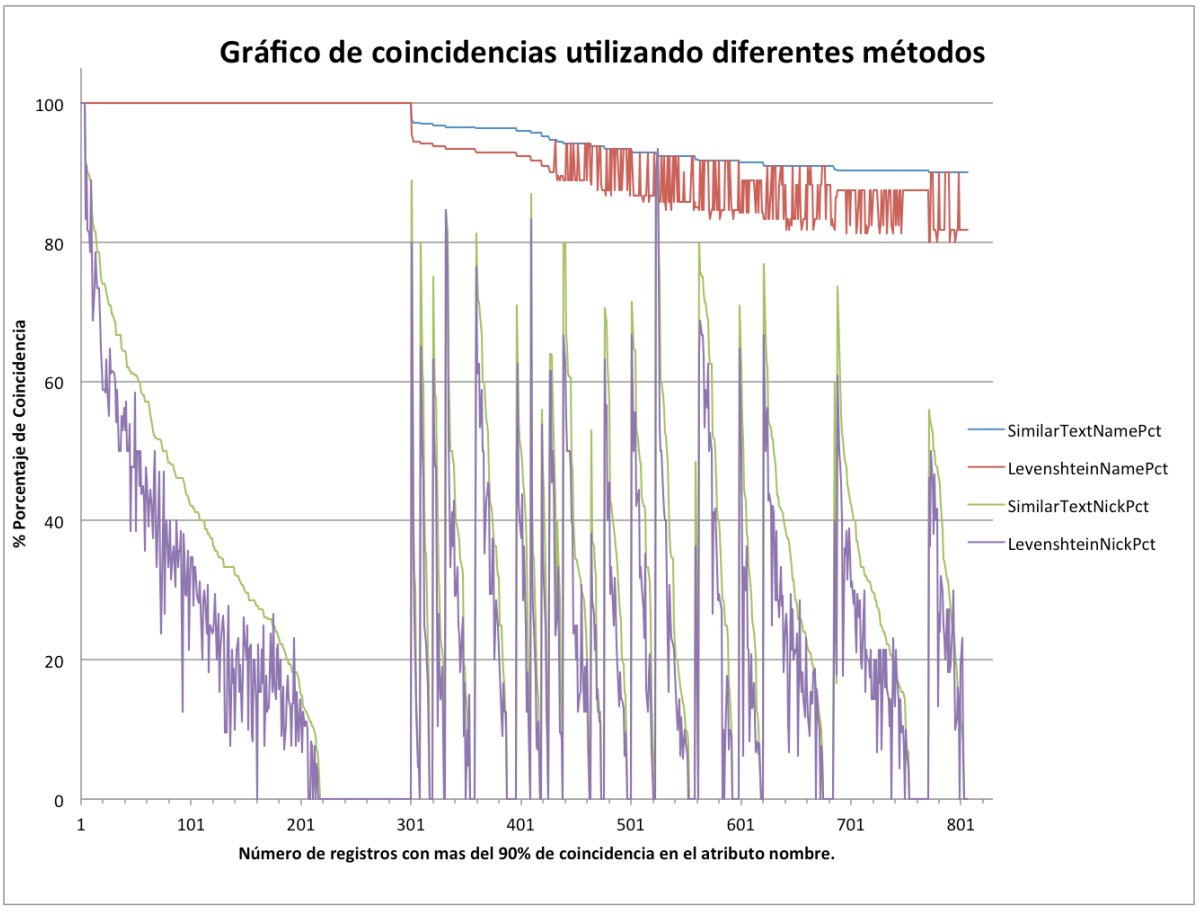

Ilustración 4 Gráfica de tendencias de resultados en el experimento de desambiguación

Como podemos observar en la ilustración 3, se obtuvieron resultados del 100% de coincidencia en 301 registros para el atributo de nombre completo usando los algoritmos de Similar-Text y de distancia de Levenshtein, sin embargo podemos apreciar también que sólo una pequeña facción tuvieron una coincidencia del 100% en la comparativa de username con solo 4 registros de los cuales se tiene una certeza del 100% que se trata de la misma persona en ambas redes sociales.

También podemos apreciar como existe una variación en los resultados de los diferentes algoritmos usados, y aunque las tendencias de los resultados son muy parecidas, el algoritmo de Similar-Text mostró ser más consistente al tener valores más regulares, el algoritmo de Levenshtein mostró más picos en sus resultados, que en algunos casos llegaron a ser de cerca de 20 puntos porcentuales de diferencia contra el valor obtenido con el algoritmo de Similar-Text; sin embargo en algunos casos también obtuvimos resultados con un mayor porcentaje usando el algoritmo de Levenshtein aunque nunca rebasó los 14 puntos porcentuales en diferencia con el algoritmo de Similar-Text.

Los algoritmos mostraron ser útiles en la identificación de posibles usuarios comunes en las redes sociales, pueden ser usados para agilizar la identificación. Cuando los atributos coinciden plenamente no deja dudas de que se trata de la misma persona, sin embargo se puede apreciar que es difícil que los usuarios usen los mismos atributos en la diferentes redes sociales, por lo que deberá implementarse un análisis manual de aquellos registros en los que los porcentajes no lleguen al 100% y poder descartar que se traten de personas homónimas.

4.5.2 Análisis de Influencia

Los resultados del análisis de influencia de los candidatos a la alcaldía de Zacatecas se muestran en la siguiente Tabla 9. El Factor de Influencia está basado en la fórmula descrita en la sección 4.4 del presente documento:

Candidato	***Retweets***	***Followers impacto de retweets***	***Factor de Influencia FI= Σ (nrFC*(1+fllwrFFC))/ Σ fllwTC)***	***Votos***
1231494354 - Rogelio Cárdenas	1004	25584	**1.082725346**	2447
1176876972 - Salvador Llamas	126	63034	**0.623335915**	1904
1024904191 - Fernando Glez Bueno	764	90683	**1.134648988**	7554
439520630 - Carlos Peña	2237	64221	**1.235325731**	18442
143068598 - Rogelio Lara	46	15718	**0.117922574**	920
84466594 - Xerardo Ramírez	396	73826	**1.117673374**	16431
Total	**4573**	**333066**		

Tabla 9 Factor de influencia en Twitter de los candidatos a la alcaldía de Zacatecas 2013

Los factores de influencia calculados para cada candidato nos permitieron hacer una comparativa con los resultados de la votación, y calcular la correlación que se dio entre la influencia en Twitter de cada candidato y el número de votos reales que obtuvieron. A continuación mostramos las gráficas obtenidas:

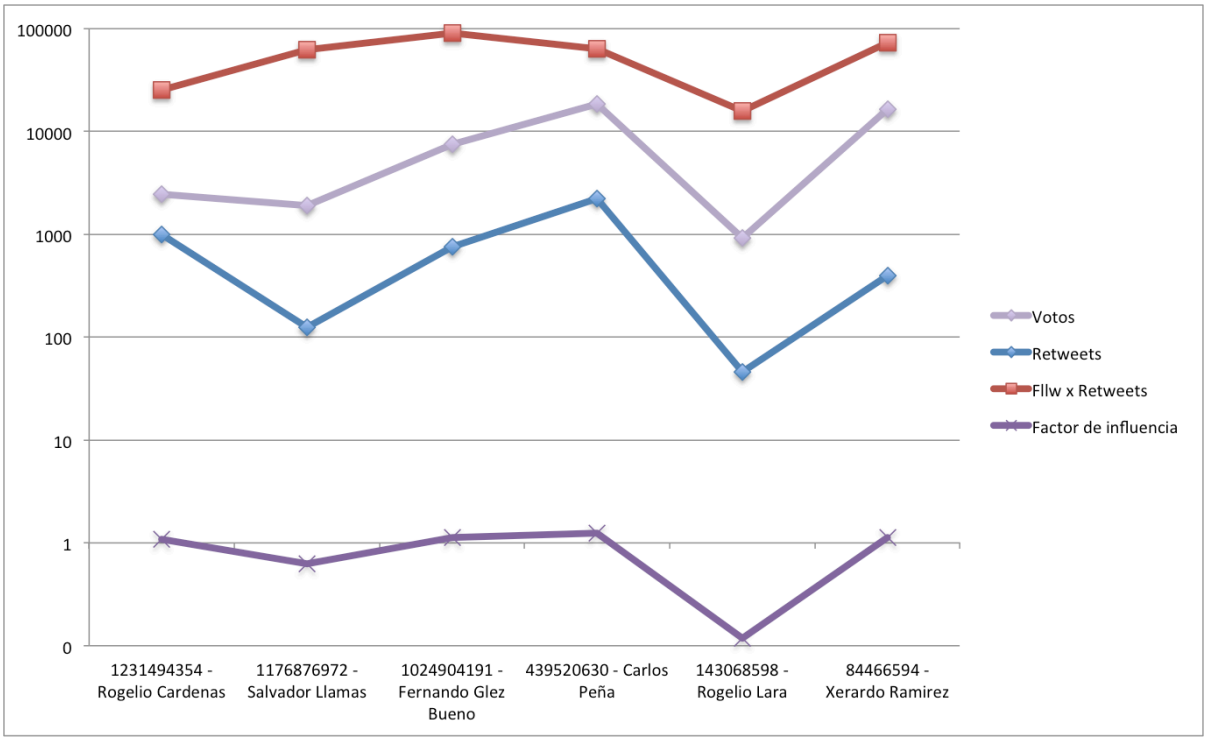

Ilustración 5 Gráfica de tendencias de factores usados para el cálculo de la influencia de los candidatos.

En la Ilustración 4 podemos observar las tendencia de los factores que se usaron para calcular la influencia de los candidatos, la primer impresión que causa es que la simple métrica de retweets se asemeja mucho a la votación real que obtuvo cada candidato, sin embargo el valor del factor de influencia es muy pequeño en escala en comparación con los valores de los otros factores, por lo que las gráficas se separaron para tratar de mejorar las comparativas.

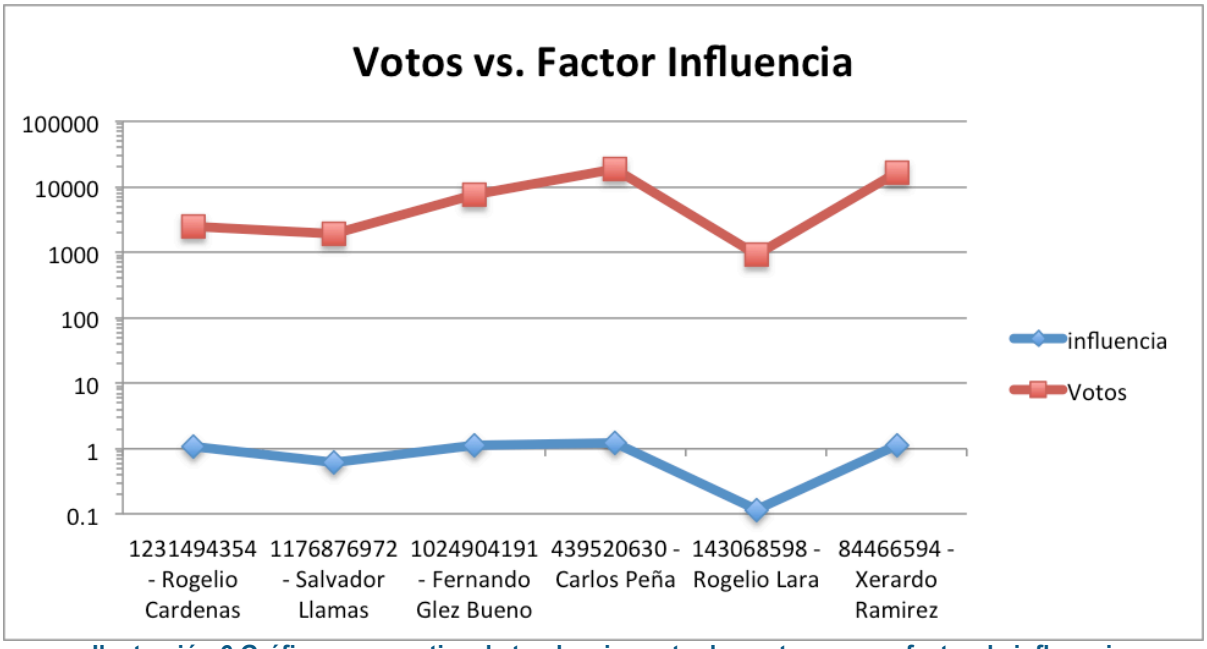

Ilustración 6 Gráfica comparativa de tendencias entre los votos versus factor de influencia.

La Ilustración 5 nos permite apreciar cómo la gráfica de tendencia entre los votos obtenidos por cada uno de los candidatos es muy similar a la del factor de influencia, se puede apreciar como el factor de influencia indica que Carlos Peña es el candidato más influyente en las redes sociales, seguido de Fernando González Bueno y Xerardo Ramirez, a diferencia de la gráfica de tendencia de los retweets donde tenemos en primer lugar a Carlos peña, seguido de Rogelio Cárdenas y Xerardo Ramírez. A continuación mostramos las tablas de predicciones utilizando las diferentes métricas:

Predicción de Posición	Retweets	Posición Real	Votos
Carlos Peña	2237	Carlos Peña	18442
Rogelio Cárdenas	1004	Xerardo Ramírez	16431
Fernando Glez Bueno	764	Fernando Glez Bueno	7554
Xerardo Ramírez	396	Rogelio Cárdenas	2447
Salvador Llamas	126	Salvador Llamas	1904
Rogelio Lara	46	Rogelio Lara	920

Tabla 10 Predicción de resultados usando la métrica de retweets versus votos reales.

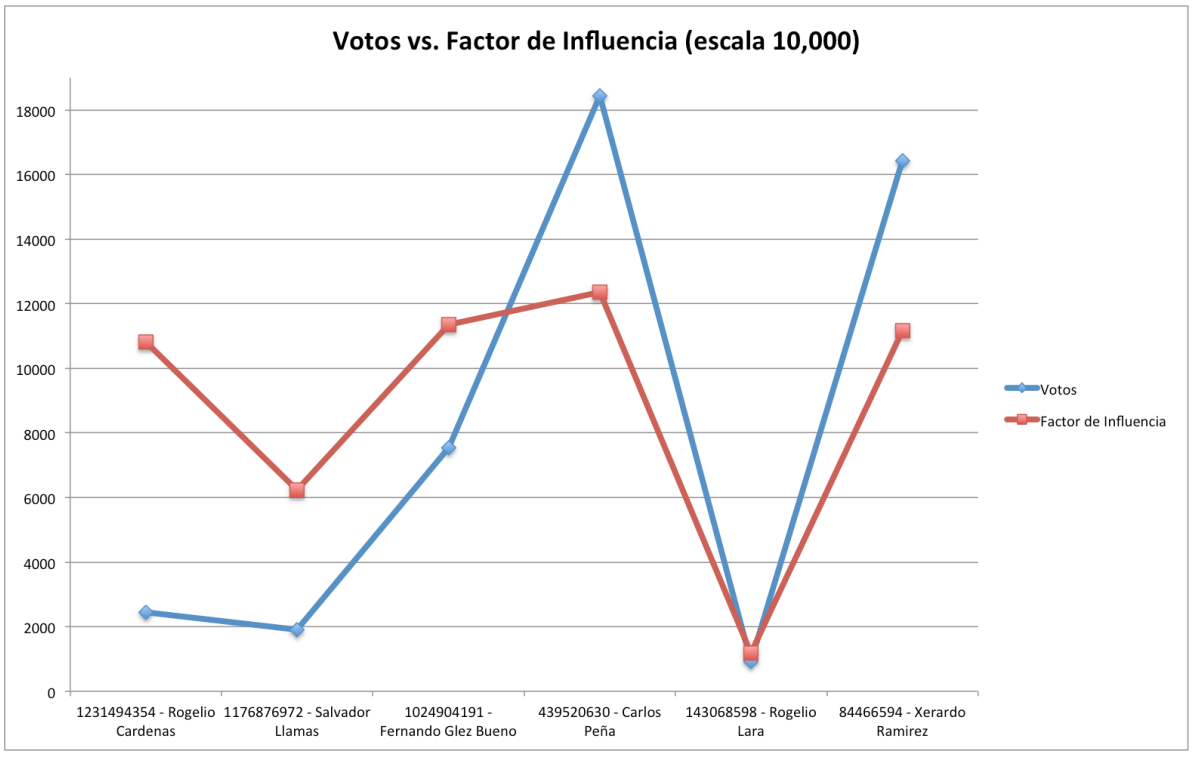

Ilustración 7 Escala 10000:1 de la gráfica de Votos versus Factor de influencia.

Predicción de Posición	Factor de Influencia	Posición Real	Votos
Carlos Peña	1.235325731	Carlos Peña	18442
Fernando Glez Bueno	1.134648988	Xerardo Ramírez	16431
Xerardo Ramírez	1.117673374	Fernando Glez Bueno	7554
Rogelio Cárdenas	1.082725346	Rogelio Cárdenas	2447
Salvador Llamas	0.623335915	Salvador Llamas	1904
Rogelio Lara	0.117922574	Rogelio Lara	920

Tabla 11 Predicción de resultados usando el factor de influencia calculado versus votos reales.

La Ilustración 6 nos permite observar claramente las posiciones que el factor de influencia calculó para cada uno de los candidatos, así como la comparación contra el resultado final de la votación, podemos observar como a diferencia de la métrica de retweets Rogelio Cárdenas se desplaza hasta una 4 posición y subiendo a Fernando González Bueno hasta una segunda posición, esto debido en parte a que este candidato fue quien más seguidores tenía con un total de 90,683 entre su primer y segundo nivel. Aun así ambas métricas mostraron diferencia versus el resultado final de la votación, por lo que se procedió a hacer una validación calculando la correlación lineal entre estas métricas y el resultado final de la votación. El resultado se puede apreciar en las siguientes ilustraciones:

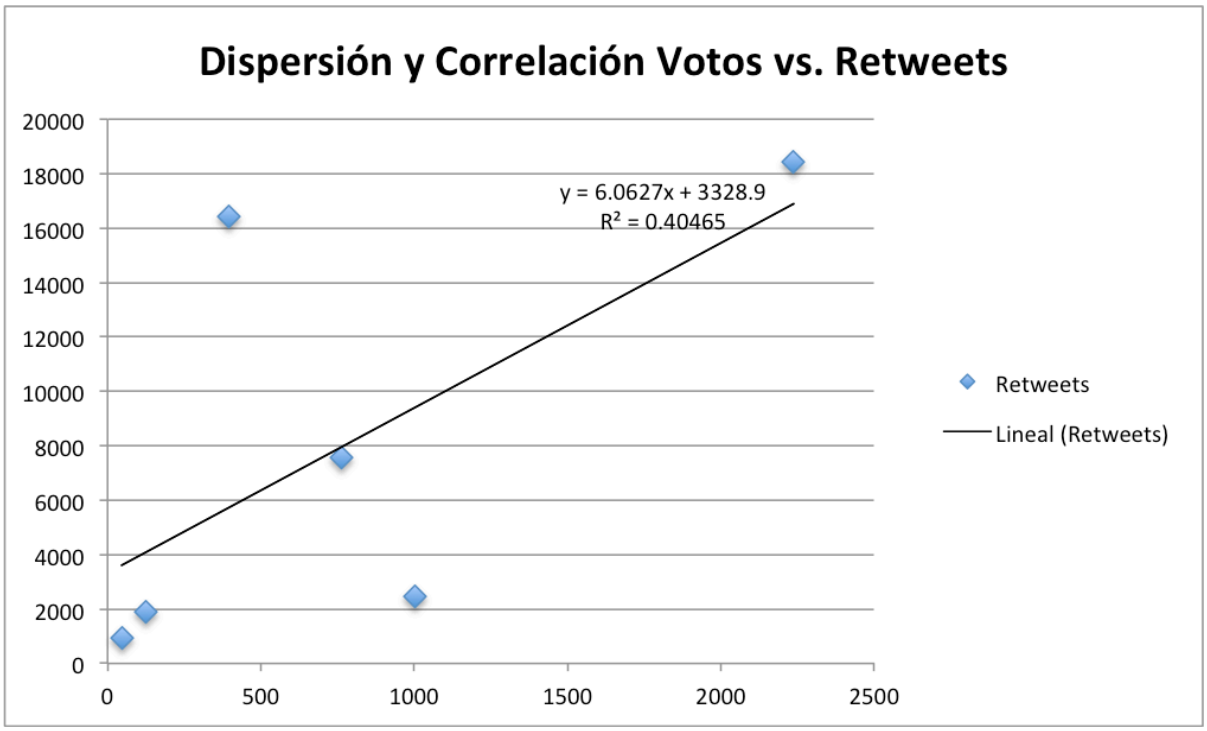

Ilustración 8 Dispersión y correlación entre retweets y votos finales.

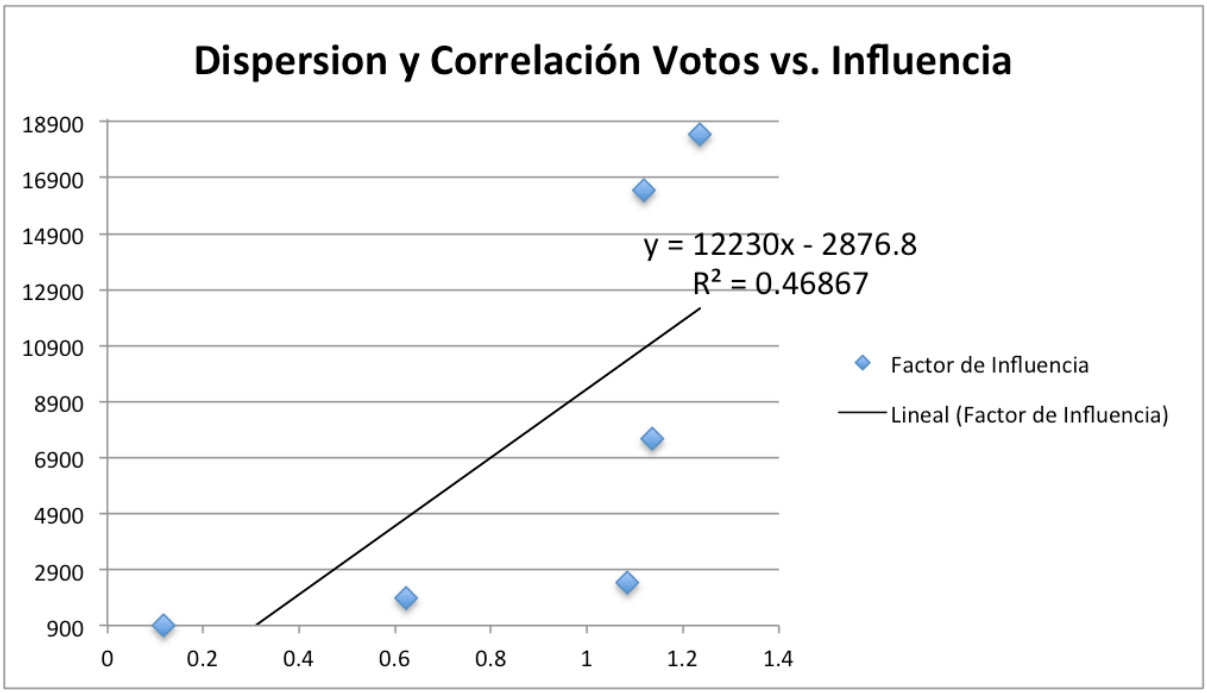

Ilustración 9 Dispersión y Correlación entre el factor de influencia y los votos finales.

Como resultado final podemos afirmar que el método que proponemos para calcular la influencia de usuarios o actores de redes sociales es más acertado, que el simple uso de métricas como el número de seguidores, twitter(post) o retweets por si solas, una correlación R^2 en retweets de 0.40465 versus una correlación R^2 de 0.46867 de nuestro factor de influencia. Nuestra metodología incluye como factor el número de seguidores impactados por cada retweet y establece un universo común de influencia, lo que da certeza de hacer una comparación justa entre los actores analizados.

5 Conclusiones

En la presente investigación recorrimos el mundo del análisis de redes sociales en línea, desde el estado del arte que es muy cambiante dada la dinámica de actualizaciones en las plataformas de las redes sociales, en las herramientas y las métricas usadas para diferentes análisis, hasta cómo usar algoritmos que nos permitieron efectuar experimentos de análisis que resultaron exitosos y con un gran potencial de reutilización en otros ejercicios del análisis científico de estas redes sociales.

Pudimos conocer las diferentes API de las principales redes sociales en línea (Online Social Networks), Facebook, LinkedIn y Twitter, las cuales exponen diferentes métodos para ejecutar las búsquedas, así como diferentes métodos de devolución de los datos resultantes de estas búsquedas, y sus principales limitantes. Estas API están diseñadas para poder ser usadas con casi cualquier lenguaje de programación actual, y están en constante renovación ya que las redes sociales están agregando servicios con bastante frecuencia, lo que también impacta en el tipo de datos que pueden o no estar expuestos por el API. Se conocieron algunos métodos y algoritmos alternos a las API para tratar de complementar las búsquedas y obtener datos que las API tienen restringidas. Pudimos observar algunas adaptaciones del uso de las API con lenguajes de programación tipo PHP y C#, además de cómo se hace uso de estas API en diferentes software de análisis de redes sociales (**SNA software**). También se pudo encontrar la semejanza en los métodos de devolución de datos de las diferentes API y se encontró que prácticamente todas usan JSON y XML para esto.

Durante el trabajo de investigación pudimos observar cómo las herramientas que antes solo se dedicaban al análisis de redes en general,-herramientas que van desde redes de distribución, de comunicaciones, etcétera-, están incluyendo apartados específicos o especializados para el análisis de redes sociales con plataformas basadas en internet, algunas como el caso de Gephi hacen un excelente uso de las capacidades gráficas de las actuales computadoras, permitiendo el análisis gráfico de redes de gran tamaño, que en el caso de NodeXL está limitado a el número de registros que Excel puede contener, sin embargo Ghepi es un software de graficado de redes en general, no es exclusivo para el análisis o graficación de datos de redes sociales con plataformas en internet, por lo que carece de plug-ins que faciliten la importación de datos de estas plataformas, teniendo que ser construidos estos manualmente.

La lista de herramientas y librerías para la extracción y análisis de datos de las redes sociales identificada, es bastante grande, son más de 80 herramientas y librerías (véase Anexo 4) y cada vez se van agregando más, si bien tenemos ya algunos años usando redes sociales, estamos apenas comenzando en su análisis, por lo que en un futuro inmediato ésta área requerirá de tener un compendio claro y actualizado de las iniciativas globales que están siendo usadas para responder a las preguntas que las diferentes instituciones, investigadores y empresas se están planteando.

En la parte del experimento podemos concluir que la parte más demandante fue la extracción de datos (minería de datos) de los seguidores de los candidatos a la alcaldía de Zacatecas, en Facebook y Twitter, ya que las restricciones de ejecución de las API hicieron que este proceso fuera lento, llevándonos un mes en la recolección de los datos. La base de datos creció bastante, llegando a un tamaño de cerca de 450 MB de espacio en disco duro, y

aunque no parece ser mucho el servidor web sobre el que hicimos la extracción, fue insuficiente para realizar las consultas y comparaciones que los experimentos requirieron, por lo que se tuvo que trasladar la información a un servidor que no tuviera sus recursos compartidos. Un buen diseño de la base de datos contenedora de los datos recolectados, es clave fundamental a la hora de hacer las consultas de estos datos para efectuar las comparaciones u operaciones de los experimentos de análisis que se requieran, también pudimos darnos cuenta que es muy importante la indexación de los datos para acelerar estas operaciones.

El experimento de desambiguación efectuado nos permitió definir un método simple para efectuar análisis en redes sociales a través de múltiples redes sociales en este caso Facebook y Twitter, y poder identificar el perfil de los usuarios o actores de un movimiento social. Sin embargo la técnica es fácilmente replicable a cualquier otra red social, lo único que se requiere es hacer una normalización, es decir, igualar o tener atributos comunes o tener datos que sean comparables en las diferentes redes sociales. Los algoritmos de Similar Text y la Distancia de Levensthein demostraron que pueden ser usados para rastrear similitudes en los diferentes perfiles de las redes sociales, y además pudimos observar que se complementan bastante bien. Los atributos de los usuarios que usamos fueron solamente el username y nombre completo de los usuarios en las diferentes redes sociales, bastándonos para procesar una base de datos con 240,269,584 comparaciones, encontrando coincidencias en solo 810 registros de usuarios, es decir con posibilidades del más del 90% de ser el mismo usuario en las dos redes sociales analizadas (Facebook y Twitter), lo que reduce significativamente el trabajo de verificación de los 2,589,636 perfiles recuperados en la extracción de datos.

Es importante señalar que los algoritmos de Similar Text y la Distancia de Levensthein funcionaron con los atributos username y nombre completo de los usuarios, ***sin embargo es necesario implementar mejoras en este tipo de análisis para poder incluir más atributos que puedan ser recolectados de la base de datos de las diferentes redes sociales en línea, como son la localidad o ubicación del usuario, o incluir algoritmos de lenguaje natural que realicen comparativas de las publicaciones efectuadas por un mismo usuario en diferentes redes sociales***; lo anterior con la finalidad de mejorar la efectividad de la desambiguación y descartar personas que puedan ser homónimas o que tengan cuentas falsas del tipo bots.

El análisis de tendencias nos permitió definir una fórmula sencilla de influencia, que incluyera un universo común para todos los actores participantes en el experimento. Las métricas más usadas actualmente por las herramientas de análisis de influencia, son el número de seguidores, likes o retweets, las cuales tienden a ser imprecisas, ya que no delimitan el universo de influencia, sino que toman como base el total del tamaño de la red social; de esta forma podría decirse que la influencia representada por estas métricas es a nivel mundial, sin embargo la fórmula que proponemos calcula el factor de influencia (FI), reduciendo el universo de datos a partir de la sumatoria de seguidores de los actores involucrados en el análisis, de esta forma se delimita el universo y se hace más acorde a la realidad del tamaño de la muestra para la cual se quiere conocer el nivel de influencia, lo que nos da una mayor posibilidad de obtener datos más precisos y acotados a la realidad de impacto de influencia en una zona (universo definido y común) de los actores analizados. La fórmula que proponemos mostró una mayor correlación con los resultados reales de la

votación, en comparación con la fórmula que considera como influencia el simple conteo de retweets.

Si bien las correlaciones mostradas tanto por los retweets R^2 =0.404065, como por nuestro factor de Influencia R^2 =0.46867 no es muy alta y está por debajo de la media de .5, muestran claramente un apego a la tendencia que se evidenció en los resultados finales. **Aun así creemos que haría falta realizar más experimentos de este tipo para asegurar que la fórmula sea totalmente confiable y verificar si el factor de correlación aumenta con un universo más estable de los datos.** Cabe mencionar que aunque el conjunto de datos de cada candidato fueron limpiados de "valores atípicos", lamentablemente por tratarse de un evento político se detectó la existencia de cuentas propagandistas que no necesariamente incluyen seguidores de impacto en el evento social, además de cuentas de tipo bots usadas para librar guerras sucias, que alteran la realidad de la influencia de los actores.

Aun así podemos concluir satisfactoriamente que el método para calcular la influencia propuesto en el experimento puede ser utilizado en la medición efectiva de influencia para la predicción de resultados de eventos sociales en general, y en el análisis de tendencias de actores o publicaciones en redes sociales en línea, entre otros casos de estudio.

Finalmente, como futuras líneas de investigación trataremos de incorporar mejoras en los algoritmos usados en el experimento de investigación. Para la parte de desambiguación consideramos que es pertinente incluir el análisis con algoritmos de lenguaje natural, y de ubicación geo referenciada lo que en teoría nos daría más certeza en la detección de perfiles únicos de usuario, o bien la detección de personas homónimas.

En la parte de análisis de influencia deberá optimizarse el algoritmo de cálculo del factor de influencia (FI) para lograr una correlación más cercana al 1.0 versus los resultados reales de los eventos sociales. Orientar los esfuerzos en encontrar metodologías de definición del universo de influencia, así como la manera de calcular pesos de influencia entre los diferentes grados de seguidores, extendiéndolo hasta n grados usando algoritmos recursivos, ya que la metodología actual solo contempló el uso del primer y segundo grado de seguidores.

Se deberá incluir algoritmos para la detención de orígenes de movimientos sociales, es decir, ¿quiénes o qué están disparando los eventos sociales, o movimientos en las redes sociales en línea?, ¿qué influencia ejerce este origen en las comunidades, mercados o seguidores del evento?, ¿cómo predecir eventos importantes antes de que impacten socialmente?.

Y por último la inclusión de extracción de datos de nuevas redes sociales, mejorando las metodologías de extracción de datos de ellas, que aunque estas están muy limitadas por las políticas de seguridad y privacidad de cada red social, deberá establecerse un marco de trabajo (framework) que facilite la extracción y almacenamiento de estos datos, que representan la base de todo análisis que se pretenda realizar en las redes sociales en línea.

Referencias

AMIPC, 2012. *Hábitos de los usuarios de Internet en México*, Available at: http://www.amipci.org.mx/?P=esthabitos.

Catanese, S.A. et al., 2011. Crawling Facebook for Social Network Analysis Purposes. , pp.0–7.

Christakis, N.A. & Fowler, J.H., 2009. *Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives*,

Dănăilă, I. et al., 2012. String Distances for Near-duplicate Detection. *Polibits*. Available at: http://www.scielo.org.mx/scielo.php?pid=S1870-90442012000100004&script=sci_arttext&tlng=pt [Accessed September 22, 2013].

Derek L. Hansen, Ben Shneiderman & Smith, M.A., 2011a. NodeXL for Programmers. In *Analyzing Social Media Networks with Nodexl*.

Derek L. Hansen, Ben Shneiderman & Smith, M.A., 2011b. Social Network Analysis - Measuring, Mapping, and Modeling Collections of Connections. *Analyzing Social Media Networks with Nodexl*.

Facebook-Developers, 2012. The Graph API. , 2012, p.The Graph API is the core of Facebook Platform, en. Available at: https://developers.facebook.com/docs/.

Facebook-News, 2012. Facebook reached 1 billion monthly active users. Available at: http://newsroom.fb.com/download-media/4227.

Ferrara, E. et al., 2012. Web Data Extraction , Applications and Techniques : A Survey. , V(July), pp.1–54.

Gladwell, M., 2000. *The Tipping Point How Little Things Can Make a Big Difference*, Oxford University Press.

GraphML, 2012. The GraphML File Format. , 2012(Julio de 2012), p.webpage de http://graphml.graphdrawing.org/. Available at: http://graphml.graphdrawing.org/.

IAB-México, I.A.B., 2013. *Estudio de consumo de medios entre internautas mexicanos*, Available at: http://www.iabmexico.com.mx/.

IBM, C-suite & Studies, 2012. Liderar en un mundo hiperconectado Conclusiones de "The Global Chief Executive Officer Study." *IBM Institute for Business Value*, p.68.

Linkedin-About, 2012. Pagina Oficial de Noticas de Linkedin. , 2012. Available at: http://press.linkedin.com/About.

linkedin-Developer, 2012. LinkedIn Developer Page. , 2012. Available at: http://developer.linkedin.com/.

Malhotra, A., Totti, L. & Jr, W.M., 2012. Studying user footprints in different online social networks. *… in Social Networks …*. Available at: http://dl.acm.org/citation.cfm?id=2456933 [Accessed September 20, 2013].

Menéndez, L.S., 2003. Análisis de redes sociales : o cómo representar las estructuras sociales subyacentes.

Russell, M.A., 2011. *Mining The Social Web*,

Twitter, inc., 2012. Twitter developers API Documentation and Programer reference. , 2012(julio 2012), p.web page dor develors of twitter applications. Available at: https://dev.twitter.com/.

Twitter blog, inc., 2012. About Twitter Limits (Update, API, DM, and Following). , 2012(07/01/2012), p.Pagina de blog para desarrolladores de twitter, si. Available at: https://support.twitter.com/articles/15364-about-twitter-limits-update-api-dm-and-following.

Twitter Dev, inc., 2012. A field guide to Twitter Platform objects. , 2012(Julio 2012), p.Descripción de datos resultado de las busquedas po. Available at: https://dev.twitter.com/docs/platform-objects/users.

Urueña, A. & Ferrari, A., 2011. Las Redes Sociales en Internet.

Vladimir Barash & Golder +, S., 2011. Twitter - Conversation, Entertainment, and Information, All in One Network! In *Analyzing Social Media Networks with Nodexl*. pp. 143–164.

Apéndices y Anexos

Anexo 1 Lista de atributos del objeto User del Facebook Graph API.

Name	Description	Permissions	Returns
id	The user's Facebook ID	No access_token required	string
name	The user's full name	No access_token required	string
first_name	The user's first name	No access_token required	string
middle_name	The user's middle name	No access_token required	string
last_name	The user's last name	No access_token required	string
gender	The user's gender: female or male	No access_token required	string
locale	The user's locale	No access_token required	string containing the ISO Language Code and ISO Country Code
languages	The user's languages	user_likes	array of objects containing language id and name
link	The URL of the profile for the user on Facebook	No access_token required	string containing a valid URL
username	The user's Facebook username	No access_token required	string
third_party_id	An anonymous, but unique identifier for the user; only returned if specifically requested via the fields URL parameter	Requires access_token	string
installed	Specifies whether the user has installed the application associated with the app access token that is used to make the request; only returned if specifically requested via the fields URL parameter	Requires app access_token	object containing type(this is always "user"), id(the ID of the user), and optional installed field (always true if returned); The installed field is only returned if the user has installed the

			application, otherwise it is not part of the returned object
timezone	The user's timezone offset from UTC	Available only for the current user	number
updated_time	The last time the user's profile was updated; changes to the languages, link,timezone, verified,interested_in,favorite_athletes,favorite_teams, andvideo_upload_limits are not not reflected in this value	Requires access_token	string containing an ISO-8601 datetime
verified	The user's account verification status, eithertrue or false (see below)	Requires access_token	boolean
bio	The user's biography	user_about_me orfriends_about_me	string
birthday	The user's birthday	user_birthday orfriends_birthday	Date string inMM/DD/YYYY format
cover	The user's cover photo (must be explicitly requested usingfields=cover parameter)	Requires access_token	array of fields id, source, and offset_y
currency	The user's currency settings (must be explicitly requested using a fields=currencyURL parameter)	Requires access_token	object with fieldscurrency (detailed here),id
devices	A list of the user's devices beyond desktop	User access_token required; only available for friends of the current user	array of objects containing os which may be a value of 'iOS' or 'Android', along with an additional field hardwarewhich may be a value of 'iPad' or 'iPhone' if present, however may not be returned if we are unable to determine the hardware model - Note: this is a non-default field and must be explicitly specified as shown below
education	A list of the user's education history	user_education_history orfriends_education_history	array of objects containing year and typefields, and school object

			(name, id, type, and optional year, degree,concentration array,classes array, and witharray)
email	The proxied or contact email address granted by the user	Email	string containing a valid RFC822 email address
hometown	The user's hometown	user_hometown orfriends_hometown	object containing name andid
interested_in	The genders the user is interested in	user_relationship_details orfriends_relationship_details	array containing strings
location	The user's current city	user_location orfriends_location	object containing name andid
political	The user's political view	user_religion_politics or friends_religion_politics	string
payment_pricepoints	The payment price-points available for that user	User access_token	array of objects containing user_price,credits andlocal_currency fields
favorite_athletes	The user's favorite athletes; this field is deprecated and will be removed in the near future	user_likes or friends_likes	array of objects containing id and namefields
favorite_teams	The user's favorite teams; this field is deprecated and will be removed in the near future	user_likes or friends_likes	array of objects containing id and namefields
picture	The URL of the user's profile pic (only returned if you explicitly specify a 'fields=picture' param)	access_token required for pages with whitelisting/targeting restrictions, otherwise noaccess_token required	string; If the "October 2012 Breaking Changes" migration setting is enabled for your app, this field will be an object with the url andis_silhouette fields;is_silhouette is true if the user has not uploaded a profile picture
quotes	The user's favorite quotes	user_about_me or friends_about_me	string
relationship_status	The user's relationship status: Single, In a relationship, Engaged,Married, It's complicated,In an open	user_relationships orfriends_relationships	string

	relationship,Widowed, Separated,Divorced, In a civil union, In a domestic partnership		
religion	The user's religion	user_religion_politics orfriends_religion_politics	string
security_settings	Information about security settings enabled on the user's account (must be explicitly requested using afields=security_settingsURL parameter)	Available only for the current user	object containingsecure_browsing (an object with a single field,enabled, which indicates whether Secure Browsing is enabled)
significant_other	The user's significant other	user_relationships orfriends_relationships	object containing name andid
video_upload_limits	The size of the video file and the length of the video that a user can upload; only returned if specifically requested via the fields URL parameter	Requires access_token	object containing lengthand size of video
website	The URL of the user's personal website	user_website orfriends_website	string containing a valid URL
work	A list of the user's work history	user_work_history or friends_work_history	array of objects containing employer,location, position,start_date and end_datefields

Anexo 2 Lista de Conexiones del objeto User del Facebook API Graph.

Name	Description	Permissions	Returns
accounts	The Facebook apps and pages owned by the current user.	manage_pages yieldsaccess_tokens that can be used to query the Graph API on behalf of the app/page	array of objects containing account name,access_token, category, id
achievements	The achievements for the user.	user_games_activity orfriends_games_activity.	array of achievement(instance) objects
activities	The activities listed on the user's profile.	user_activities orfriends_activities.	array of objects containing activity id, name,category and create_time fields.
albums	The photo albums this user has created.	user_photos orfriends_photos.	array of Album objects.
apprequests	The user's outstanding requests from an app.	Requires app access_token.	array of app requests for the user within that app.
books	The books listed on the user's profile.	user_likes orfriends_likes.	array of objects containing book id, name,category and create_time fields.
checkins	The places that the user has checked-into.	user_checkins orfriends_checkins.	array of Checkin objects
events	The events this user is attending.	user_events orfriends_events.	array of objects containing event id, name,start_time, end_time, location andrsvp_status defaulting to the past two weeks.
family	The user's family relationships	user_relationships.	array of objects containing id, name, andrelationship fields.
feed	The user's wall.	read_stream	array of Post objects containing (up to) the last 25 posts.
friendlists	The user's friend lists.	read_friendlists.	array of objects

			containing id and name fields of the friendlist.
friendrequests	The user's incoming friend requests.	user_requests.	array of objects containing to, from,message, created_time and unread fields of the friend request
friends	The user's friends.	Any valid access_token of the current session user.	array of objects containing friend id and namefields.
games	Games the user has added to the Arts and Entertainment section of their profile.	user_likes	array of objects containing id, name,category, and created_time
groups	The Groups that the user belongs to.	user_groups orfriends_groups.	An array of objects containing theversion(old-0 or new Group-1), name, id,administrator (if user is the administrator of the Group) and bookmark_order(at what place in the list of group bookmarks on the homepage, the group shows up for the user).
home	The user's news feed.	read_stream.	array of Post objects containing (up to) the last 25 posts.
inbox	The Threads in this user's inbox.	read_mailbox.	array of thread objects
interests	The interests listed on the user's profile.	user_interests orfriends_interests.	array of objects containing interest id, name,category and create_time fields.
likes	All the pages this user has liked.	user_likes orfriends_likes.	array of objects containing like id, name,category and create_time fields.
links	The user's posted links.	read_stream.	array of Link objects.

locations	Posts, statuses, and photos in which the user has been tagged at a location, or where the user has authored content (i.e. this excludes objects with no location information, and objects in which the user is not tagged). See documentation of the location_posttable for more detailed information on permissions.	user_photos,friend_photos,user_status,friends_status,user_checkins, orfriends_checkins.	array of objects containing the id, type,place, created_time, and optionalapplication and tags fields.
movies	The movies listed on the user's profile.	user_likes orfriends_likes.	array of objects containing movie id, name,category and create_time fields.
music	The music listed on the user's profile.	user_likes orfriends_likes.	array of objects containing music id, name,category and create_time fields.
mutualfriends	The mutual friends between two users.	Any valid access_token of the current session user.	array of objects containing friend id and namefields.
notes	The user's notes.	read_stream.	array of Note objects.
notifications	App notifications for the user.	Any valid access_token of the current session user.	array of objects containing template and href.
outbox	The messages in this user's outbox.	read_mailbox.	array of messages
payments	The Facebook Credits orders the user placed with an application. See theCredits API for more information.	app access_token	array of order objects.
permissions	The permissions that user has granted the application.	None.	array containing a single object which has the keys as the

			permission names and the values as the permission values (1/0) – Permissions with value 0 are omitted from the object by default; also includes a `type` field which is always `permissions` if the query param`metadata=1` is passed.
photos	Photos the user (or friend) is tagged in.	`user_photo_video_tags` or `friends_photo_video_tags` .	`array` of `Photo` objects.
photos/uploaded	All of the updates photos of a user. Cursor based pagination.	`user_photos`	`array` of `Photo` objects containing all of the photos a user has uploaded in order of upload time.
picture	The user's profile picture.	No `access_token` required.	HTTP 302 redirect to URL of the user's profile picture (use `?type=square` | `small` | `normal` |`large` to request a different photo). If you specify `?redirect=false`, this connection will return the URL of the profile picture without a 302 redirect. Additionally, you can specify`width` and `height` URL parameters to request a picture of a specific size. This will return an available profile picture closest to the requested size and requested aspect ratio. If only `width` or `height` is specified, we will return a picture whose width or height is closest to the requested size,

			respectively; if`width=height`, we will always return a square picture. If the "October 2012 Breaking Changes" migration setting is enabled for your app, this connection will return a JSON object with `url`, `width`, `height`, and `is_silhouette`fields, where the width and height specify the actual dimensions of the returned picture;`is_silhouette` is a boolean which specifies whether the profile picture is the default picture (i.e. the user has not uploaded a profile picture).
pokes	The user's pokes.	`read_mailbox`.	an `array` of objects containing `to`, `from`,`created_time` and `type` fields.
posts	The user's own posts.	Any valid `access_token` or`read_stream` to see non-public posts.	`array` of `Post` objects.
questions	The user's questions.	`user_questions`	`array` of `Question` objects.
scores	The current scoresfor the user in games.	`user_games_activity` or`friends_games_activity`.	`array` of objects containing `user`,`application`, `score` and `type`.
sharedposts	Returns shares of the object. Cursor based pagination.	`read_stream`	`array` of `Post` objects.
statuses	The user's status updates.	`read_stream`.	An `array` of `Status message` objects.
subscribedto	People you're subscribed to.	Any valid `access_token`	`array` of objects containing user `id` and `name`fields.

subscribers	The user's subscribers.	Any valid access_token	array of objects containing user id and namefields.
tagged	Posts the user is tagged in.	read_stream	array of objects containing id, from, to, picture, link, name, caption, description,properties, icon, actions, type,application, created_time, andupdated_time
television	The television listed on the user's profile.	user_likes orfriends_likes.	array of objects containing television id,name, category and create_time fields.
updates	The updates in this user's inbox.	read_mailbox.	array of messages
videos	The videos this user has been tagged in.	user_videos orfriends_videos.	array of Video objects.

Anexo 3 Colección de atributos del LinkedIn API.

Field	Parent Node	Description
last-modified-timestamp	person	The timestamp, in milliseconds, when the member's profile was last edited
proposal-comments	person	A short-form text area describing how the member approaches proposals
associations	person	A short-form text area enumerating the Associations a member has
honors	person	A short-form text area describing what Honors the member may have
interests	person	A short-form text area describing the member's interests
publications	person	A collection of publications authored by this member
patents	person	A collection of patents or patent applications held by this member
languages	person	A collection of languages and the level of the member's proficiency for each
skills	person	A collection of skills held by this member
certifications	person	A collection of certifications earned by this member
educations	person	A collection of education institutions a member has attended, the total indicated by a *total* attribute
courses	person	A collection of courses a member has taken, the total indicated by a *total* attribute
volunteer	person	A collection of volunteering experiences a member has participated in, including organizations and causes, the totals indicated by a *total* attribute
three-current-positions	person	A collection of positions a member currently holds, limited to three and indicated by a *total* attribute. You can use the <positions> collection to get the full set or use this collection to limit the return to just the first three positions.
three-past-positions	person	A collection of positions a member formerly held, limited to the three most recent and indicated by a*total* attribute. You can use the <positions> collection to get the full set or use this collection to limit the return to just the

Field	Parent Node	Description
		first three positions.
num-recommenders	person	The number of recommendations the member has
recommendations-received	person	A collection of recommendations a member has received.
mfeed-rss-url	person	a URL for the member's multiple feeds
following	person	a collection of people, company, and industries that the member is following
job-bookmarks	person	a collection of jobs that the member is following
suggestions	person	a collection of people, company, and industries suggested for the member to follow
date-of-birth	person	member's birth date
member-url-resources	person	A collection of URLs the member has chosen to share on their LinkedIn profile
member-url-resources:(url)	person/member-url-resources	The fully-qualified URL being shared
member-url-resources:(name)	person/member-url-resources	The label given to the URL by the member
related-profile-views	person	A collection of related profiles that were viewed before or after the member's profile

Anexo 4 Colección de herramientas de análisis de redes sociales y librerías.

A continuación se muestra una tabla con una lista de diferentes herramientas de análisis de redes sociales y librerías que actualmente están en desarrollo y siendo usadas en diferentes partes del mundo (http://en.wikipedia.org/ 2012):

Tabla 12 Colección de herramientas de análisis de redes sociales y librerías.

Product	Main Functionality	Input Format	Output Format	Platform	License and cost	Notes
AllegroGraph [3]	Graph Database.RDF with Gruff visualization tool	RDF	RDF	Linux, Mac, Windows	Free and Commercial	AllegroGraph is a graph database. It is disk-based, fully transactional OLTP database that stores data structured in graphs rather than in tables. AllegroGraph includes a Social Networking Analytics library. Gruff is a freely downloadable triple-store browser that displays visual graphs of subsets of a store's resources and their links. By selecting particular resources and predicates, you can build a visual graph that displays a variety of the relationships in a triple-store. Gruff can also display tables of all properties of selected resources or generate tables with SPARQL queries, and resources in the tables can be added to the visual graph.
AutoMap [4]	Network Text Analysis	.txt	DyNetML [5],. csv	Any (it's in Java)	Freeware for non-commercial use	Text mining tool that supports the extraction of relational data from texts. Distills three types of information: content analysis, semantic networks, ontologically coded networks. In order to do this, a variety of Natural Language Processing/ Information Extraction routines is provided (e.g. Stemming, Parts of Speech Tagging, Named-Entity Recognition, usage of user-defined ontologies, reduction and

						normalization, Anaphora Resolution, email data analysis, feature identification, entropy computation, reading and writing from and to default or user-specified database).
Centrifuge Visual Network Analytics [6]	Visual Network Analytics	Any data source that supports connection through JDBC	Web Browser, CSV data, PNG images, published visualizations for collaboration, URL images which include visualizations	Windows and Linux	Free Evaluations, commercial and government editions, enterprise licensing and OEM licensing	Centrifuge Visual Network Analytics (VNA) help organizations discover insights, patterns and relationships hidden in public, cloud, social network and enterprise data. Centrifuge® Systems delivers a unique approach to interactive data visualization – Combining agile data integration, dynamic relationship mapping, and interactive visual analytics to reveal insights in big data. Using Centrifuge visualizations and link intelligence, analysts discover, measure and communicate risk and fraud.
						Centrifuge solves challenging visual network, relationship mapping and data analysis problems in the areas of fraud and money laundering, organized retail crime, pharma risk analysis, intelligence analysis, cyber security and other domains. Centrifuge actively partners with software companies, information providers and others looking to enhance their solutions by embedding Centrifuge visualizations
CFinder [7]	Finding and visualizing communities	.txt	.txt, .pfd, .ps, .svg, .svg, .emf, .gif, .raw, .ppm, .bmp, .jpg,.png, .wbmp	Linux, Mac OS X, Windows, Solaris	Freeware for non-commercial use	A software for finding and visualizing overlapping dense communities in networks, based on theclique percolation method. It enables customizable visualization and allows easy strolling over the found communities. The package contains a command line version of the program as well, suitable for scripting.

C-IKNOW [8]	Survey design, data collection, visualization, recommendation.	.DL, .txt GraphML	.DL, .txt GraphML	Mac, Windows, Linux	Free and Commercial	C-IKNOW is a powerful web-based software tool for social network analysis investigation. It has been designed around real-world problems, and it can store and analyze virtually any type of network data. The documentation provides a basic step-by-step walkthrough of how to get started on a C-IKNOW project as well as more advanced support, including the C-IKNOW Question-Type Primer. C-IKNOW's visualization and analytics suite allows both administrators and users to access visualizations, recommendation tools, and analytical measures for their networks.
Commetrix [9]	Dynamic network visualization & analysis	Commetrix-Files, direct import from data sources/DB's, (standard DB and File Specs upcoming)	CSV Tables for SNA Metrics over time,(Graph Videos per Screencast), Keywords, Graphs, etc. in GUI	Any system supporting java (developed for Windows Platform)	Free trial, commercial licenses, free research collaboration (in beta-user group),	Commetrix is a Software Framework and Tool for Dynamic Network Analysis and Visualization. It provides easy exploratory access to network graphs and has been applied to study co-authorship, Instant Messaging, manual SNA surveys, e-mail, newsgroups, etc. Each node and each linking event can have properties, e.g. types of messages or rank of nodes, but also types, topics, or time stamps. This allows animations of network growth, structural change, and topic diffusion. A short introductory video is available on the website.
CoSBiLab Graph [10]	Network visualization, analysis and manipulation	.dot, .txt, .dl(UCINet), .spec(Beta WB), .txt (MRMC)	.dot, .txt, .dl(UCINet), .txt (MRMC), .pm(PRISM), .png	Windows (.NET 3.5 required)	Freeware for non-commercial use	CoSBiLab Graph is an application for visualization analysis and manipulation of networks. It provides a high customizable graphical representation of networks based on local properties. Nodes can be aggregated and arranged on the space manually or by choosing from a list of predefined layouts. A set of indices is provided for measuring the positional importance of nodes in the network and

						they can be combined together defining new mathematical expressions. The manual and a set of examples are available on the website.
Cuttlefish [11]	Dynamic network visualization and simulations using different layouts	cxf, pajek, graphml, MySQL, PostgreSQL,	tikz, jpeg, cxf, MySQL, PostgreSQL, Commetrix CSV,	Any system supporting Java	GNU General Public License	Cuttlefish is a network workbench application that visualizes the networks with some of the best known layout algorithms. It allows detailed visualizations of the network data, interactive manipulation of the layout, graph edition and process visualization as well as different input methods and outputs in tex using Tikz and PSTricks. It is developed by the Chair of Systems Design of ETH Zürich, a research group that applies a complex system approach to investigate economic and social networks.
Cytoscape [12]	General complex network data integration, analysis, and visualization.	SIF (Simple Interaction Format, GraphML, XGMML, GML, KGML, SBML, BioPAX, Excel, and text tables (including csv, tab delimited tables)	SIF, XGMML, GML, GraphML, Cytoscape Session(.cys), vector/bitmap images including jpg, png, pdf, ps.	Any system supporting Java	Open source (LGPL)	An open source platform for complex network data integration, analysis, and visualization. Originally Cytoscape was developed for bioinformatics research and now it is a problem domain independent platform. Many plugins are available for users and developers can expand its functionality by writing them.
Deep Email Miner [13]	Social Network Analysis and text mining of an Email corpus	MySQL database	pajek and MySQL database	Any system supporting Java	GPL V2	A software solution for the multistaged analysis of an Email Corpus. Social network analysis and text mining techniques are connected to enable an in depth view into the underlying information.
Detica NetReveal [14]	Social Network Analysis for insurance or banking fraud, crime detection, intelligence, tax evasion,	csv, txt, XML and databases	csv, txt, XML and native Oracle database	Any system supporting Java	Commercial	A platform that can process billions (often at national scale) of multi-format data sources and builds social networks. In doing so, a single view of entity (customer, business, telephone, bank account, vehicle, address, citizen,

	border control and network risk based targeting					etc.) can be generated across multiple, poor quality data sources. Social networks and entities can be scored using a range of powerful analytics and a full free text entity centric search is available across all records. The platform includes network visualization tools, workflow and real time rules engine to score incoming events in real time.
DEX [15]	Graph database for query processing and network analysis.	csv, jdbc	csv, graphml, graphviz	Linux, Windows & Mac OS	Free evaluation version (up to 1 Million nodes, no restriction on edges, 1 concurrent user). For larger graphs or commercial ask for licenses quotation.	DEX is a high-performance graph database written in Java and C++ . One of its main characteristics is its performance storage and retrieval for large graphs, in the order of billions of nodes, edges and attributes, allowing the analysis of large scale networks.
EpiFast [16]	EpiFast is a model that simulates the spread of an infectious disease across a social network - see Stephen Eubank and Keith Bisset at Virginia Tech.	(add)	(add)	(add)	(add)	See the journal article from the SC '08 Proceedings of the 2008 ACM/IEEE conference on Supercomputing - [17].
Discourse Network Analyzer[18]	Extract networks from structured text data	Text via copy&paste, .DNA files (a simple XML format)	DL, GraphML, CSV, Commetrix SQL, SON (Sonia)	Any system supporting Java 1.6	Freeware	Discourse Network Analyzer serves two purposes: manually coding text data for statements of actors in a QDA-like fashion, and exporting one- or two-mode networks from these

						structured data. Dynamic algorithms for the longitudinal analysis of discourses are available.
DyNet [19]	Data analysis	*.agf(proprietary), *.net(pajek), *.txt	*.agf(proprietary), *.net(pajek), *.txt	?	Proprietary(starting from $3000/user)	DyNet SE (Standard Edition) is an innovative software tool to analyse pools of complex data unveiling relations and interconnections via graphical and verbose outputs. DyNet SE is based on social network theory therefore relational data is visualised in terms of networks.
EgoNet Active Development orExplanation	Ego-centric network analysis	Conducts interviews or takes any valid XML file	Output to CSV and convertible to almost any other format	Any system supporting Java	Open Source, seeking contributors	Egonet is a program for the collection and analysis of egocentric network data. Egonet contains facilities to assist in creating the questionnaire, collecting the data and providing general global network measures and data matrixes that can be used in further analysis by other software programs.
EveSim [20]	EvESimulator	XML, SimCase	XML	Any system supporting Java	Open Source	The EvESimulator provides a simulation framework for biologically inspired P2P systems - the EvE as a part of the DBE. Although its focus is on the EvE, the EvESimulator simulates a DBE. Besides from that, the EvESimulator constitutes a collaborative platform for interdisciplinary research acting as a framework for understanding, visualising and presenting the DBE concepts to contributors.
FirmNet Online [21]	Social Network Analysis survey, visualization & reports.	Survey data collected via online questionnaire	png, jpg, svg network images	Any browser	Commercial. Academic research supported	FirmNet Online (FNO) is a fully web-based Organizational Network Analysis tool for consultants. Online ONA survey, network visualization and reports integrated into one process based platform. Consultants can start using FNO after a certification training.

Future Point Systems [22]	Visual analytics platform called Starlight for all-source analysis, including social network analysis (SNA)	Virtually any format, including MSFT Office, PDF, XML, .txt, database, HTML, web services, POP or IMAP mail servers, RSS, ESRI SHP	XML, CSV, ESRI SHP, KML, copy to clipboard, web reports, PDF, .jpg, .bmp, .png	Windows	Government GSA pricing and commercial pricing	Starlight is a visual analytics platform that transforms data into actionable intelligence. SNA capabilities include centrality, path-finding and metrics support.
FNA [23]	Online platform for network analysis of financial transaction, trade or link data.	Arc list (.csv, .txt), Matrix (.txt), Pajek (.net), Graphml (.graphml) - from files or any JDBC database	Arc list (.csv, .txt), Matrix (.txt), Pajek (.net), Graphml (.graphml) - to files or any JDBC database	Linux, Windows, IE9-/Chrome/Firefox 3.6-	Proprietary or Web Service	FNA is an analytics platform that helps financial institutions and regulators manage and understand financial data with network analysis and visualization. Its particularly suited for the analysis of large transaction, trade or link databases in finance and for monitoring continuous data via dashboards. You can use FNA for free online at http://www.fna.fi.
Gephi [24]	Graph exploration and manipulation software	GraphViz(.dot), Graphlet(.gml), GUESS(.gdf), LEDA(.gml), NetworkX(.graphml, .net), NodeXL(.graphml, .net), Pajek(.net, .gml), Sonivis(.graphml), Tulip(.tlp, .dot), UCINET(.dl), yEd(.gml), Gephi (.gexf), Edge list(.csv), databases	GUESS(.gdf), Gephi(.gexf), .svg, .png	Any system supporting Java 1.6 and OpenGL	Open Source (GPL3), seeking contributors	Gephi is an interactive visualization and exploration platform for all kinds of networks and complex systems, dynamic and hierarchical graphs. It is a tool for people that have to explore and understand graphs. The user interacts with the representation, manipulate the structures, shapes and colors to reveal hidden properties. It uses a 3D render engine to display large networks in real-time and to speed up the exploration. A flexible and multi-task architecture brings new possibilities to work with complex data sets and produce valuable visual results.

GraphStream [25]	Dynamic Graph Library	GraphStream(.dgs), GraphViz(.dot), Graphlet(.gml), edge list	GraphStream(.dgs), GraphViz(.dot), Graphlet(.gml), image sequence	Any system supporting Java	Open Source	With GraphStream you deal with graphs. Static and Dynamic. You create them from scratch, from a file or any source. You display and render them.
graph-tool [26]	Python module for efficient analysis and visualization of graphs.	GraphViz(.dot),GraphML	GraphViz(.dot), GraphML, .bmp, .canon, .cmap, .eps, .fig, .gd, .gd2, .gif, .gtk, .ico, .imap, .cmapx, .ismap, .jpeg, .pdf, .plain, .png, .ps, .ps2, .svg, .svgz, .tif, .vml, .vmlz, .vrml, .wbmp, .xlib	GNU/Linux, Mac	Free Software (GPL3)	graph-tool is a python module for efficient analysis of graphs. Its core data structures and algorithms are implemented in C++, with heavy use of Template metaprogramming, based on the Boost Graph Library. It contains a comprehensive list of algorithms.
Graphviz	Graph vizualisation software	GraphViz(.dot)	.bmp, .canon, .cmap, .eps, .fig, .gd, .gd2, .gif, .gtk, .ico, .imap, .cmapx, .ismap, .jpeg, .pdf, .plain, .png, .ps, .ps2, .svg, .svgz, .tif, .vml, .vmlz, .vrml, .wbmp, .xlib	Linux, Mac, Windows	Open Source (CPL)	Graphviz is open source graph visualization framework. It has several main graph layout programs suitable for social network visualization.

iDETECT [27]	Visual Network Analytics	Any structured and unstructured data sources, Database-agnostic	Web Browser, CSV data, published visualizations for collaboration	Any platform supporting Java (back-end), Web-browser (front-end)	Commercial	iDETECT provides next-generation investigative monitoring and analysis platform. Its technology empowers financial institutions, private sector businesses and government agencies to monitor, detect, investigate, and defeat the most sophisticated forms of crimes - often those which have never been found before. iDETECT is redefining the ways of extracting actionable intelligence from structured and unstructured data through innovative analytics processes.
Idiro SNA Plus [28]	Highly scalable Social Network Analysis for Telecoms	All databases e.g. Oracle, DB2, Teradata & flat file	All databases & flat text Files	Linux	Commercial	Idiro SNA Plus is the market leading SNA platform for telecoms with a particular focus on churn prediction, viral marketing, acquisition and family unit identification. Idiro SNA Plus takes social network analysis from academia and into the realm of business where the focus is on deriving real value from the application of SNA to real-world problems. With Idiro SNA Plus users can do the following: 1. Churn - Predict churners and quantify the damage a person would cause if they were to churn 2. Viral marketing - Identify key influencers for viral marketing3. Family units - Identify family units for marketing purposes4. Acquisition - Identify targets for member-get-member campaigns 5. Rotational churn - Identify rotational churners
igraph [29]	Analysis and visualization of very large networks	.txt (edge list), .graphml, .gml, .ncol, .lgl, .net	.txt (edge list), .graphml, .dot. .gml, .ncol, .lgl, .net	Windows, Linux, Mac OS X	Open source (GNU GPL)	igraph is a C library for the analysis of large networks. It includes fast implementations for classic graph theory problems and recent network analysis methods like community structure search, cohesive blocking, structural holes, dyad and triad census and motif count estimation.

						Higher level interfaces are available for R, Python, and Ruby.
iPoint [30]	Analysis and visualization of social networks trends, geo-location, age, gender and sentiment	Take any valid XML	XML, Flex	Windows, Linux, Mac OS X	Commercial	iPoint monitors and analyzes Consumer Generated Media, the full privacy of the author is maintained and its reporting dashboard reads from iMediaStreams web services. The analysis is easily viewed and managed from the worldwide, to the state, to the hyperlocal neighborhood level.
InFlow [31]	Interactive network mapping and network metrics in one integrated application for social and organizational network analysis.	Easy data import from Microsoft Office[PC/MAC] and CSV files	Export graphics to Microsoft Office [PC/MAC] -- Powerpoint, Word, Visio—and network files to interactive Java applet for WWW	Windows 2000, XP, Vista	Commercial, Site, and Academic licenses available. Training & Mentoring in social network analysis, data gathering, and software application, is also available.	InFlow is intended for business users, and is designed for ease-of-use, multiple networks per node set, and what-if capabilities. Network data can be entered via 1) CSV files, from data bases and spreadsheets, 2) automated survey tools such as NetworkGenie, Optimice, etc. 3) data entry screens with paper surveys, or 4) drawn by hand with mouse, using node & link tools in graphics window. Most popular network metrics included: Density, Geodesics, Freeman Centralities, Watts-Strogatz Small World, Structural Equivalence, Cluster Analysis,Krackhardt E/I Ratio, and Krebs Reach & Weighted Average Path Length. Metrics are executed based on current network view—you measure what is mapped. Many network layouts are possible using automated algorithms and geometric layouts[arcs, lines, etc.] resulting in an unlimited number of custom views. Different actions can be taken on selected nodes vs. unselected nodes.

Java Universal Network/Graph (JUNG) Framework	network and graph manipulation, analysis, and visualization	built-in support for GraphML, Pajek, and some text formats; user can create parsers for any desired format	built-in support for GraphML, Pajek, and some text formats; user can create exporters for any desired format	Any platform supporting Java	Open source (BSD license)	JUNG is a Java API and library that provides a common and extensible language for the modeling, analysis, and visualization of relational data. It supports a variety of graph types (including hypergraphs), supports graph elements of any type and with any properties, enables customizable visualizations, and includes algorithms from graph theory, data mining, and social network analysis .
Jerarca [32]	Social network analysis, community structure, hierarchical clustering of networks.	.txt (List of links)	Text, output to MEGA [33], output to Cytoscape[34], hierarchical tree in Newick format	Linux, Windows	Open Source (GNU GPLv3)	Jerarca is a suite of hierarchical clustering algorithms that provides a simple and easy way to analyze complex networks. It is designed to efficiently convert unweighted undirected graphs into hierarchical trees by means of iterative hierarchical clustering. Moreover, Jerarca detects and returns the community structure of the network.
Keyhubs [35]	Social Network Analysis			Web-based		Keyhubs provides software and services for workplace social analytics: www.keyhubs.com.
KrackPlot [36]	Network visualization	UCINET, Mathematica	UCINET, Mathematica	?	?	KrackPlot is a program for network visualization designed for social network analysts.
KXEN Social Network (KSN)[37]	Powerful Social Network Analysis	All databases, Text Files, other Input formats	All databases, Text Files, other Output formats, graph structure export : dot format used by GraphViz, link structure.	Unix, Linux, Windows	Commercial	KSN is a Social Network Analysis module designed for extracting many Social Networks from CDRs, extracting many attributes from a Social Network, integrating Social Network attributes into the customers database and exploiting Social Network attributes to build predictive models
libSNA [38]	Basic network statistics	Csv	Csv	Any platform supporting python	Open source(LGPL)	libSNA is a widely-used open source library for conducting SNA research. Written in the object oriented programming language Python, libSNA provides a simple

						programming interface for applying SNA to large scale networks. libSNA is built on top of the open source libraryNetworkX; without NetworkX, libSNA would not be possible.
Meerkat [39]	Static and dynamic networks: community mining, visualization , exploration, and filtering software.	Pajek(.net), Edge list, Meerkat (.meerkat), GraphML (.graphml)	Meerkat (.meerkat), GraphML (.graphml), .jpg, .pdf, .cvs	Windows, Ubuntu and OSX, Java 1.6+	Free to use Lite version, pay to use Full to be released .	Meerkat allows interactive visualization of networks, and provides facilities and algorithms for community mining, filtering on edge and node properties, network statistics, and node metrics. In particular, it provides dynamic network community mining, or community evolution event analysis, which allows abstraction and better understanding of changes to communities across timeframes for dynamic networks.
MetaSight [40]	Email / communication network visualization and analysis	MS Exchange and Lotus email servers	Interactive user interface	Windows Server	Commercial	MetaSight is an enterprise social software application which uses data from routine e-mail to infer and map business expertise and relationships. Applications include expertise location and external relationship management.
NEO4J [41]	Graph Database with several modules such as rdf or visualization	GraphML, rdf, csv, other	?	?	AGPL and commercial	Neo4j is a graph database. It is an embedded, disk-based, fully transactional Java persistence engine that stores data structured in graphs rather than in tables.
Network Overview Discovery Exploration for Excel (NodeXL) [42]	Network overview, discovery and exploration	email, .csv (text), .txt, .xls (Excel), .xslt (Excel 2007), .net (Pajek), .dl (UCINet), GraphML	.csv (text), .txt, .xls (Excel), .xslt (Excel 2007), .dl (UCINet), GraphML	Windows XP/Vista/7	Free (Ms-PL)	NodeXL is a free and open Excel 2007/2010 Add-in and C#/.Net library for network analysis and visualization. It integrates into Excel 2007 and 2010 and adds directed graph as a chart type to the spreadsheet and calculates a core set of network metrics and scores. Supports extracting email, Twitter, YouTube, Facebook, WWW and flickr social networks. Accepts edge lists and matrix representations of graphs. Allows for easy manipulation and filtering of underlying data in

						spreadsheet format. Multiple network visualization layouts. Reads and writes UCINet and GraphML files.
NetMiner 4 [43]	All-in-one Software for Network Analysis and Visualization	.xls(Excel),.xlsx (Excel 2007), .csv(text), .dl(UCINET), .net(Pajek), .dat(StOCNET), .gml; NMF(proprietary)	.xls(Excel),.xlsx (Excel 2007), .csv(text), .dl(UCINET), .net(Pajek), .dat(StOCNET), NMF(proprietary)	Microsoft Windows	Commercial with free trial	NetMiner is a software tool for exploratory analysis and visualization of large network data. NetMiner 4 embed internal Python-based script engine which equipped with the automatic Script Generator for unskilled users. Then the users can operate NetMiner 4 with existing GUI or programmable script language.Main features include : analysis of large networks(+10,000,000 nodes), comprehensive network measures and models, both exploratory & confirmatory analysis, interactive visual analytics, what-if network analysis, built-in statistical
Network Genie [44]	Social Network Survey Data collection	Online survey and project design environment	Output to CSV, InFLow [45], NEGOPY [46], MultiNet [47], Pajek [48], Siena[49], and UCINET [50]	Any social web browser	Payment assessed for completed surveys at $.50 per survey. Data collection is free. Payment is required prior to data download.	Network Genie is used to: (1) Design complete, egocentric, and hybrid social network surveys using a wide variety of survey question formats; (2) Manage social network projects, including manage a collaborative team who have privileges defined by a project coordinator; (3) Collect social network data using online forms; and (4) Download and export data to the social network analysis program of your choice. Registration is free.
Network Workbench[51][52]	Modeling, Analysis and Visualization of Large Scale Networks	.net, .mat, .graphml, .nwb, .csv, xgmml	.net, .mat, .graphml, .nwb, .csv, xgmml, .eps, .pdf	Linux, Mac OS X, Windows, Solaris	Open Source (Apache 2.0)	Contains a variety of algorithms and features useful for analyzing networks, including Page Rank, Pathfinder Network Scaling, Small World network generation, and Blondel

						Community Detection to name a few. The underlying OSGi plugin model allows users to expand on Network Workbench's core functionality.
NetworkX	Python package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks.	GML, Graph6/Sparse6, GraphML, GraphViz (.dot), NetworkX (.yaml, adjacency lists, and edge lists), Pajek (.net), LEDA	GML, Gnome Dia, Graph6/Sparse6, GraphML, GraphViz (.dot), NetworkX (.yaml, adjacency lists, and edge lists), Pajek (.net), LEDA, and assorted image formats (.jpg, .png, .ps, .svg, et al.)	Open source (GPL and similar)	Free	NetworkX (NX) is a toolset for graph creation, manipulation, analysis, and visualization. User interface is through scripting/command-line provided by Python. NX includes a several algorithms, metrics and graph generators. Visualization is provided through pylab and graphviz.NX is an open-source project, in active development since 2004 with an open bug-tracking site, and user forums. Development is sponsored by Los Alamos National Lab.
Nevada [53]	Dynamic network visualization & analysis	XML-based GAML (similar to GraphML), Pajek files	SVG, PNG, and GraphML	Any system supporting Java	Open Source, GNU General Public License	Nevada is a tool for interactive visualization of dynamic networks. Unlike other dynamic network visualization tools it's focussing on visualizations preserving the user's mental-map. Import of Pajek files is supported.
[NGCE][54]	Graph Generator, Graph Analyzer	txt	txt, net – Pajek compatible	Any system supporting Java	University of Illinois/NCSA Open Source License	NGCE is graph generator and statistical analyzer which can create reproducible graphs for simulations or other scientific experiments. In particular it is capable of creating: Homogeneous Graphs, Random Graphs, Scale-Free Graphs, Random Graphs with Fixed Connectivity and Custom Graphs.

ONA Surveys [55]	Social Network Survey Data collection	Online survey and project design environment	Output to Microsoft Excel, Inflow [56]and d Netdraw [57]	Microsoft Internet Explorer	Subscription based. Unlimited number of surveys. Pricing for commercial and academic use.	ONA Surveys is a tool aimed primarily at ONA/SNA practitioners to help collect data about relationships. Free registration provides full functionality, but export is limited to first 5 nodes. Paid subscription provides unlimited surveys with unlimited respondents. Supports multiple languages.
ORA [58]	Social Network Analysis, Network Visualization, Meta-Network Analysis, Trail Analysis, Geospatial Network Analysis, Network Generation	DyNetML [59], .csv	DyNetML, .csv	Windows	Freeware for non-commercial use	*ORA is a dynamic meta-network assessment and analysis tool containing hundreds of social network, dynamic network metrics, trail metrics, procedures for grouping nodes, identifying local patterns, comparing and contrasting networks, groups, and individuals from a dynamic meta-network perspective. *ORA has been used to examine how networks change through space and time, contains procedures for moving back and forth between trail data (e.g. who was where when) and network data (who is connected to whom, who is connected to where …), and has a variety of geo-spatial network metrics, and change detection techniques.
Pajek [60] [61]	Analysis and Visualization of Large Scale Networks	.net, .paj, .dat(UCINET), .ged, .bs, .mac, .mol	.net, .paj, .dat(UCINET), .xml(graphML), .bs	Windows, Linux, Mac OS X	Freeware for non-commercial use	A widely used program for drawing networks, Pajek also has analytical capabilities, and can be used to calculate most centrality measures, identify structural holes, block-model, and so on. Macros can be recorded to perform repetitive tasks. Data can be sent directly to R, to calculate additional statistics.

R	Social network analysis within the versatile and popular R environment	R will read in almost any format data file	R has write capability for most data formats	Windows, Linux, Mac	Open source	R contains several packages relevant for social network analysis: *igraph* is a generic network analysis package; *sna* performs sociometric analysis of networks; *network* manipulates and displays network objects; *tnet* performs analysis of weighted networks, two-mode networks, and longitudinal networks; *ergm* is a set of tools to analyze and simulate networks based on exponential random graph models exponential random graph models; *Bergm* provides tools for Bayesian analysis for exponential random graph models, *hergm* implements hierarchical exponential random graph models; *latentnet* has functions for network latent position and cluster models; *degreenet* provides tools for statistical modeling of network degree distributions; and *networksis* provides tools for simulating bipartite networks with fixed marginals.
Sentinel Visualizer [62]	Comprehensive network analysis and visualization	Structured XML, and databases such as SQL Server, Oracle, and Access. Also, Excel, Text and HTML formats.	Open database architecture in SQL Server; Structured XML and unstructured documents in Word, PDF, Excel, Text and HTML formats.	Microsoft Windows	Commercial (starting from $2000/user) 45-day free trial available	Sentinel Visualizer is a Windows-based program that provides data visualization, analysis and knowledgebase management within one product. Sentinel Visualizer produces interactive dynamic link charts, timeline and geospatial views, and provides a variety of analysis tools including Social Network Analysis, temporal analysis and entity and relationship weighting. Sentinel Visualizer includes a multi-user knowledgebase for efficiently and economically storing analysis data.

SNA-Forte [63]	A Scalable Social Network Analysis Solution	Any data source that can be imported into SAS (DB connections, text files, SAS files)	Any data source that can be exported from SAS (DB connections, text files, SAS files)	Any platform supporting SAS	Free software; is provided as part of professional services.	SNA-Forte is a social network analysis solution using raw telecommunications CDR data as its input and automatically identifying communities of customers, as well as segments of these communities and roles of individuals in each community, based on selected parameters and weights. The algorithm is implemented as an open-source solution in SAS and is already in use in commercial environment.
SNA-Network [64]	A toolkit for Social Network Analysis	.gdf (Guess), .net (Pajek)	.gdf (Guess), .net (Pajek)	Any platform supporting Perl 5	Free software; may be redistributed/modified under the same terms as Perl itself.	SNA-Network is a bundle of modules for network algorithms, specifically designed for the needs of Social Network Analysis (SNA), but can be used for any other graph algorithms. It represents a standard directed and weighted network, which can also be used as an undirected and/or unweighted network. Data structures have been designed for SNA-typical sparse network operations, and consist of Node and Edge objects, linked via references to each other.
Social Networks Visualizer[65]	Social Networks Visualization and Analysis Tool	.xml (GraphML), .net (pajek), .dot (GraphViz), .sm/.net (Sociomatrix), .net (UCINET)	.xml (GraphML), .net (pajek), .dot (GraphViz), .sm/.net (Sociomatrix)	Linux, Windows, Mac (Qt toolkit needed)	Free Software (GPL3)	SocNetV (Social Networks Visualizer) is an open-source graphical application, developed in C++ language and the cross-platform Qt toolkit. The user interface is friendly and simple, allowing the researcher to draw social networks or plain graphs by clicking on a canvas. SocNetV computes basic network properties (i.e. density, diameter, shortest path lengths), as well as more advanced statistics, such as centralities (i.e. closeness, betweeness, graph), clustering coefficient, etc. Various layout algorithms are supported. For instance, nodes can be automatically

						positioned on circles or levels according to their betweeness centralities. Random networks and small world creation is also supported. SocNetV can handle any number of nodes, although with a speed penalty when nodes are more than 3000 or the graph is quite dense (many edges).
Socilyzer [66]	Easy-to-use organizational and social network analysis tool for managers and consultants. Complete with guidebook and study templates.	Data is collected with online surveys. Pro users can also copy-paste raw matrix data into a data editor.	Export network visualizations and copy statistics tables to PowerPoint, Word, print etc. Pro users can also generate and export data in VNA-format.	All platforms (web-based)	Commercial, free 30-day trial available.	An all-in-one social network analysis analysis tool with built-in questionnaire design, data collection, data visualization and statistics. Find the guidebook and study templates at:https://socilyzer.com
SocioMetrica [67]	EgoNet, LinkAlyzer, and VisuaLyzer applications	DyNetML, Excel, DL, text, UCINET	DyNetML, Excel, DL, text, UCINET, SPSS	Windows	Shareware	A set of applications for interview-based gathering of egocentric data (EgoNet), linking of data records through matching of node attributes (LinkAlyzer), and visualization (VisuaLyzer). VisuaLyzer also provides prototype functionality for analysis using a relational algebra model. A relational programming language, RAlog, derives and analyzes representations in this relation algebra.
SocProg [68]	Analyses movements of individuals, social and population structure. Prepares data for population size			Cross-platform (requiresMATLAB and Statistics Toolbox) or Windows (stand-alone)	Freeware	

analysis.

SONAMINE [69]	Scalable network scoring and analysis up to hundreds of millions nodes and billion edges	any comma separated text file	comma separate text file	Windows, Linux	Commercial, free eval. Enterprise software license or hosted.	SONAMINE graph scoring engine is software for analysts. It distributes work over multiple servers, is fault tolerant and horizontally scalable. It is used for node scoring and data mining. SONAMINE graph query server is a real time high performance graph query engine.
SONIVIS [70]	Network visualisation and analysis, especially Wiki-based information spaces	.xml(graphML)	.xml(graphML)	Windows, Linux	open-source (GPL)	SONIVIS:Tool is a Java-based, open-source application, which is based on the Eclipse Rich Client Platform (RCP). The user interface is organized into three main perspectives: Analysis, Manipulation, and Statistics. Besides various Wiki and network analysis metrics, the tool provides predefined and user-definable graphical analyses. It offers a quick overview on current Wiki states or developments
statnet [71]	Social network analysis within the versatile and popular R environment	R will read in almost any format data file	R has write capability for most data formats	Windows, Linux, Mac	Open source (GPL)	A suite of R packages for social network analysis: sna performs sociometric analysis of networks; network manipulates and displays network objects; ergm implements exponential random graph models for networks; latentnet has functions for network latent position and cluster models; degreenet provides tools for statistical modeling of network degree distributions; and networksis provides tools for simulating bipartite networks with fixed marginals; the statnet meta-package allows for package management.

StOCNET [72]	Software package for the advanced statistical analysis of social networks	Text (.dat, .txt)	Text	Windows	Freeware/Open source	StOCNET is a software system for the advanced statistical analysis of social networks, focusing on probabilistic (stochastic) models. The program consists of several statistical models for network analysis. In the present version, six modules are implemented: BLOCKS (stochastic blockmodeling of relational data), p2 (analysis of binary network data with actor and/or dyadic covariates), PACNET (constructing a partial algebraic model for observed multiple complete networks using a statistical approach), SIENA (analysis of repeated measures on social networks and MCMC-estimation of exponential random graphs), ULTRAS (analysis of binary undirected network data using ultrametric measurement models), and ZO (simulation and/or enumeration of graphs with given degrees).
tnet [73]	Social network analysis of weighted, two-mode, and longitudinal networks in R	Edgelist	R has write capability for most data formats	Windows, Linux, Mac	Open source (GPL)	A packages for social network analysis of weighted, two-mode, and longitudinal networks. Possible extensions are discussed here [74]
Tulip	Social Network Analysis tool	Tulip format (.tlp), GraphViz (.dot), GML, txt, adjacency matrix	.tlp, .gml	Windows Vista, XP, 7/ Linux / Mac OS	LGPL	Tulip is an information visualization framework dedicated to the analysis and visualization of relational data. Tulip aims to provide the developer with a complete library, supporting the design of interactive information visualization applications for relational data that can be tailored to the problems he or she is addressing.

UCINET [75]	Social Network Analysis tool	Excel, DL, text, Pajek .net, Krackplot, Negopy, proprietary (##.d & ##.h)	Excel, DL, text, Pajek .net, Krackplot, Mage, Metis, proprietary (##.d & ##.h)	Windows	Shareware	A comprehensive package for the analysis of social network data as well as other 1-mode and 2-mode data. Can handle a maximum of 32,767 nodes (with some exceptions) although practically speaking many procedures get too slow around 5,000 - 10,000 nodes. Social network analysis methods include centrality measures, subgroup identification, role analysis, elementary graph theory, and permutation-based statistical analysis. In addition, the package has strong matrix analysis routines, such as matrix algebra and multivariate statistics.
UNISoN [76]	Download usenet messages and save SNA output files	Reads from free NNTP servers	Creates CSV files and Pajek .net files	Any system supporting Java	Freeware	A java application that can download Usenet messages from free NNTP servers, show the saved messages, then allow filtering of data to save to a Pajek network file or CSV file. It creates networks using the author of each post. If someone replies to a post, there is a unidirectional link created from the author of the post to the author of the message they are replying to. There is also a preview panel that shows the network visually.
UrlNet [77]	Generation of social network analysis program input files	World Wide Web pages, online search engine result sets, and Internet Web Service APIs	currently generates Pajek projects[78] and GUESS .gdf files [79]	Any (it's written in Python, requires v2.5 or higher [80]; source code is included)	Freeware for non-commercial use	UrlNet is a Python class library for generating networks based on Internet linkages. In the simplest case, UrlNet creates a tree by harvesting the outlink URLs from the page referenced by a root URL (level zero); retrieving each of those pages (level 1), harvesting their outlink URLs; retrieving those pages (level 2), harvesting their outlink URLs; et cetera to a caller-specified depth. UrlNet can also create "forests", the union of multiple tree networks. Specialized classes are provided for generation of

						networks from search engine result sets (6 search engines are currently supported).
VennMaker [81]	Egocentric network analysis and interview tool	Copy&Paste edges from other programs	JPEG, PNG, SVG, CSV	Java 1.6 (Windows, Linux, MacOS)	Commercial (free demo version)	VennMaker is a software program that allows data to be jointly gathered both qualitatively and quantitatively at the same time during an interview. In addition, the interview is recorded on an audio track. While the respondent is visualising and describing the personal network in conversation, qualitative data can be collected visually and/or by way of audio recording and adding comments. Quantitative data can be directly established on the digital network map or with an optional electronic questionnaire. The interviewer is able to configure the digital questionnaire prior to the survey, making adjustments according to the focus in question. For non-standardised surveys, e.g. narrative interviews, it is possible to adjust all set-tings of the network map during the course of the interview as well.
visone [82]	Interactive analysis and visualization of social networks	GraphML, UCINET (.dl), Pajek (.net), Excel (.csv), Matrices, and Edge lists	Same as input and images as PNG, PDF, JPEG	Java (Windows, Linux, MacOS)	Freeware for non-commercial use	Interactive graphical tool for manipulating, analyzing, and visualizing social networks. Analysis methods include centrality indices, clustering, cliques, components, and centralization. Generic graph layout algorithms and tailored network visualizations are available. visone supports many graphical properties and generates high-quality images in PNG, PDF, etc.

VisuaLyzer [83]	Network visualization	Edgelist/Edgearray, Excel or GraphML formats	Edgelist/Edgearray, Excel or GraphML formats	?	Commercial	Interactive tool for entering, visualizing and analyzing social network data.Create nodes and links directly in VisuaLyzer, or import data from Edgelist/Edgearray, Excel or GraphML formats.
WAND [84]	Ecological network analysis	Scor files	Scor files	Microsoft Windows	Open source(GPL)	
Xanalys Link Explorer [85]	Visual analytics combining link analysis with temporal and spatial analysis	ODBC databases, flat files, direct data entry	*.wf, *.xas (Proprietary formats), Excel	Microsoft Windows	Commercial, Free trial	Interactive visual analytics tool combining data acquisition and querying with link analysis, temporal analysis and spatial analysis (GIS) techniques. Integrates with other desktop applications and services such as Excel and Bing mapping
Plotonic [86]	social network analysis	keywords/phrases	graphics, text	PHP	Commercial	Interactive web service for tracking brands on social networks. Compiles social network data into psychographic, demographic, and geographic charts and maps.
Jungle Torch [87]	social network analysis, seo report, inbound marketing	keywords/phrases	Graphics, Excel, Text	any web browser	Free Trial, Commercial	Jungle Torch gives an SEO view of websites. Jungle Torch provides users with a social network analysis and can determine who is saying what about a particular company good or bad.

Fin del documento.

"Al final, todo está conectado"
- Charles Eamnes