REDES Revista Hispana para el Análisis de Redes Sociales

Vol.27,#1, (2016), 58-72

http://revistes.uab.es/redes http://dx.doi.org/10.5565/rev/redes.589

Proceso de adquisición de datos del sistema de correo electrónico: Una aplicación a la modelización de una red social

Francisco Rodríguez-Aguilera*
Nieves Arranz Peña
Universidad Nacional a Distancia (UNED)

Juan Carlos Fernández de Arroyabe
University of Essex

RESUMEN

El correo electrónico se ha convertido en un repositorio rico y amplio de la información acerca de las comunicaciones entre los individuos. Desde el punto de vista empresarial las comunicaciones por correo electrónico son el reflejo de las interacciones entre las personas que componen la organización. En este trabajo ideamos las técnicas y las herramientas para la exploración de datos que permiten extraer del correo electrónico la red social que subyace y la información que pueda ser útil para la toma de decisiones en la empresa. Analizando los datos del correo electrónico del aeropuerto de Muenster y Osnabrueck, se muestra la red social conformada a través de dichas relaciones, las propiedades de la red y su evolución a través del tiempo. Nuestra técnica proporciona una aplicación para inferir a través del correo electrónico las conexiones de la red y su caracterización sugiriendo una nueva vía de exploración de datos para estudiar el comportamiento empresarial.

Palabras clave: *Correo electrónico - Análisis de Redes Sociales – Excel.*

ABSTRACT

Electronic mail has become a rich and comprehensive repository of information about communication between individuals. From the business point of view, email communication is a reflection of interaction between the employees of the organization. In this paper we devised the techniques and tools for exploring data, that allow to extract information from email that underlie the social network that may be useful in the company to make a decision. Analyzing data from email communication of the airport Muenster and Osnabrueck shows the properties and the evolution of the network that has been formed by the relation between the employees. Our technique provides an application to infer through email network connections and characterizations, suggesting a new way to explore and study data of business behavior.

Key words: *Email - Social Network Analysis - Excel.*

* *Contacto con los autores: Francisco Rodríguez-Aguilera (frodrigue562@alumno.uned.es), Nieves Arranz (narranz@cee.uned.es), Juan Carlos Fernández (jcfern@essex.ac.uk)*

INTRODUCCIÓN

La teoría de redes proporciona una perspectiva interesante desde la que explicar la estructura y características de la relación entre agentes en los más diversos campos (véase por ejemplo, Newman, 2010; Borgatti y Foster, 2003; Borgatti y Halgin, 2011). Siguiendo a Knoke y Yang (2008), el objetivo del análisis de redes sociales es estudiar las relaciones entre individuos y los atributos de la red que conforman entre ellos, ya que dichos atributos (tales como el tamaño, la densidad, el clustering, etc.) influyen en la difusión de la información dentro de la red (Wasserman y Faust, 1994; Contractor y Monge, 2003). En la actualidad el principal reto que enfrenta el análisis de redes sociales es la obtención de datos fiables y accesibles. Una de las vías más importante de obtención de datos para el análisis es la utilización de archivos de documentos que derivan del uso de las tecnologías de la información y la comunicación. Así, internet se ha convertido en un amplio y rico repositorio de información sobre la interacción entre individuos (Reips 2002; Adamic y Adar, 2003).

Las interacciones creadas entre agentes, a través del sistema de emails, es considerado una red social (Gruzd, 2012), o comunidad virtual (Contractor y Monge, 2003). En este sentido, las interacciones entre agentes crean un sentimiento de pertenencia a una comunidad (Scott, 2009), como consecuencia de que los agentes encuentran un lugar donde comunicarse, sentir y aprender entre ellos, lo cual hace que desde el punto de vista estructural y social, esta sea considerada una red social (Gruzd, 2012). El estudio de las redes de email a través del analisis de redes es un campo de invetigacion relativamente reciente. Así, Yelupula y Ramaswamy (2008) realizaron una clasificación de los emails en el contexto del análisis de las redes sociales, con el objetivo de determinar la relevancia de los agentes en la red. Por su parte Smith (2009), utilizó los datos del correo electrónico para caracterizar una comunidad en forma dinámica y determinar la centralidad de la red, identificándola con el grado de autoridad de los individuos. Rowe et al. (2007) presentaron un algoritmo genético, para extraer datos de Enron emails. Estos Autores, utilizan el grado, densidad y proximidad, para detectar la actividad de los actores en internet. Chapanond et al. (2005) utilizando también Enron email data set, investigan la estructura organizativa de una compañía. Diesner et al. (2005) utilizando igualmente Enron emails, y combinado con análisis de redes, exploran la dinámica y estructura de la propiedades de una red de comunicación, estableciendo una relación con la evolución de la situación económica. En general, podemos observar, que la combinación del análisis de redes sociales con las capacidades de la comunicación digital, es una incipiente línea de investigación la cual puede proporcionar una importante información de la estructura y dinámica de las organizaciones (Trier, 2008). Sin embargo, tratar los mensajes de correo electrónico no es fácil en la actualidad (Alstyne y Zhang, 2003; Trier, 2008), no sólo por su elevada cantidad, y por la volatilidad de los mismos (Trier, 2008), sino porque las leyes que determinan la privacidad de la información hacen que sea muy difícil o prácticamente imposible su acceso (Gross y Acquisti, 2005). Por tanto, siguiendo a Frank et al. (2005) el desafío consiste en encontrar metodologías para la extracción de datos fiables, que sean de bajo coste y accesibles para la aplicación del análisis de redes sociales de modo que permitan estudiar las interacciones y atributos de los integrantes de la red social conformada a través del correo electrónico.

En nuestro caso, hemos desarrollado un procedimiento que permite identificar y analizar la red social de la empresa que gestiona el Aeropuerto de Muenster y Osnabrueck a través del correo electrónico con sus Stakeholders. La primera tarea es desarrollar una metodología que elimine la información no relevante o poco relevante del sistema en la captura de datos de correo electrónico atendiendo las leyes de protección de datos, es decir, sin comprometer los datos personales de los integrantes. Siguiendo los trabajos previos de Cain (2012), y Basso (2011), nos hemos basado en el uso de Microsoft Exchange, que es uno de los sistemas de correo electrónico más populares en las empresas del mundo (Melanchthon, 2010). La segunda tarea es la determinación de la red externa empresarial (Red de Stakeholders) para, posteriormente analizar las propiedades de la red social conformada. La contribución de esta investigación es plantear un método claro y definido para recolectar información a través del correo electrónico, que permita analizar y clasificar la misma con ayuda de herramientas de software que se encuentren en el mercado o que sean de fácil adaptación en la empresa objeto de investigación.

En el siguiente apartado exponemos el marco teórico de las redes sociales y el reto de la captación de los datos para el análisis. En el apartado tres describimos la metodología para la captura de datos del correo electrónico para a partir de éste desarrollar el caso de aplicación y las características de la red social. Finalmente en el apartado de conclusiones se discuten las potenciales aplicaciones de ésta técnica y extraemos las conclusiones generales.

MARCO TEÓRICO

Red Social

La red social es un concepto bien definido en la literatura. Una red es un conjunto de puntos denominados nodos, con conexiones entre ellos denominados vínculos (Borgatti y Kidwell, 2011; Newman, 2010). En el contexto de este estudio, el emisor y el receptor del correo electrónico son los nodos de la red, y el correo electrónico es el vínculo. Como resultado de la multiplicidad de agentes (edad, sexo, etc) que pueden intervenir en una red social, se espera que las conexiones entre ellos sean heterogéneas (Newman, 2010). Dicha heterogeneidad en la distribución de las conexiones es consecuencia de la afinidad y de las relaciones privilegiadas entre los nodos de la red, que resultan de los diferentes roles que éstos adoptan (Wasserman y Faust, 1994; Borgatti, 2009; Borgatti y Foster, 2003). Cuando los nodos comparten similares características interactúan más entre ellos que con aquellos que son diferentes (Coleman, 1986; Haythornthwaite y Wellman, 1998) y por tanto existe un patrón definido que los vincula preferentemente. Dicha afinidad, siguiendo estos autores, tiene como consecuencia la existencia de zonas más densas en la red, derivadas de un mayor nivel de interconexión entre los nodos.

La heterogeneidad de las redes sociales justifica su estudio como un sistema complejo (Newman, 2010). El estudio de la red social como un sistema complejo permite analizar las reglas estocásticas que influyen en la formación de los vínculos, teniendo en cuenta además de la existencia de algún tipo de aleatoriedad; y los rasgos característicos de los nodos (por ejemplo, su grado de conectividad -número de vínculos- o su centralidad) medida en función de la importancia de un nodo que, a su vez, puede verse afectada por sus vínculos con otros nodos (Coleman, 1986; Wasserman y Faust, 1994).

El sistema de correo electrónico: Una comunidad virtual

Los sistemas de correo electrónico, como vínculo entre personas, son un campo de investigación importante (Alstyne y Zhang, 2003; Baym et al., 2004; Gross y Acquisti, 2005; Bird et al., 2006; Kim et al., 2007; Gruzd, 2012). Una primera línea de investigación, ha ido dirigida a la extracción y procesamiento de la información del sistema de correo electrónico. Así, los primeros trabajos en este campo fueron realizados con el sistema "Ahoy!" (Shakes et al., 1997) que desarrolló una metodología de recuperación de la información (DRS), con el objetivo de extractar las direcciones de correo electrónico de páginas de Web. Otros estudios han estado en su mayoría enfocados a la extracción de información de las páginas de internet (Xi, et al., 2002). Borkar et al. (2000) obtuvo una alta precisión en la extracción de datos buscando y usando un número limitado de información (número de casa, calle o carretera, ciudad, estado y código postal). Upstill et al. (2003), por su parte, describen una técnica de recuperación de información que aumenta la extracción de contenidos basándose en las características de las URL. McCallum et al. (2007) chequean los flujos de email con MD5 digest. Shetty y Adibi (2004), y Klimt y Yang (2004), desarrollan una metodología con el objetivo de explorar la distribución de mensajes entre usuarios y tiempo. En general, están proliferando los estudios que tienen como objetivo captar los datos de correo electrónico, sin embargo se observa una baja eficiencia en la recuperación de datos fundamentalmente, como consecuencia de carecer de algoritmos adecuados para ello (Bird et al., 2006; Alstyne y Zhang, 2003) y de la necesidad de proteger la privacidad de los datos (Gross y Acquisti, 2005).

Una segunda línea de investigación se ha ido dirigiendo al estudio del sistema de correo electrónico desde el punto de vista estructural (Núñez y Cárdenas, 2013; Castell, 2001). Desde esta perspectiva se considera que las interacciones creadas por el sistema de correo electrónico entre usuarios conforman una red social (Gruzd y Haythornthwaite, 2013). En este sentido, Contractor y Monge (2003), introducen el concepto de red social de comunicación, la cual refleja el flujo de mensajes a través del espacio y tiempo. La noción de red social ha implicado un primer aspecto que es el sentido de comunidad entre los miembros que la conforman. Freeman et al. (1989) consideran que la comunidad implica además de unas relaciones específicas entre personas, un cierto grado de similaridad entre ellas. En este sentido, internet está creando unas comunidades virtuales, en las cuales la vía de relación son los emails, transformando la interacción física en interacción virtual. Hasta 1970, la comunidad se entendía como un grupo de personas que interaccionaban físicamente (Knoke y Yang, 2009), pero posteriormente se fue introduciendo la clave de distancia, por lo que en la comunidad no era necesaria la interacción física (Wellman y Leighton, 1979). Con la incorporación de las tecnologías de la información, se llega al concepto de comunidad virtual (Guinaliu y Flavian, 2003). En este sentido, tal como señala Gruzd (2012), es importante en estas comunidades virtuales, la creación de un espacio común, en el cual la confianza, seguridad y la colaboración estén presentes, y donde las interacciones desarrollen grupos de identidad.

Jones (1997) y McMillan y Chavis (1986), por su parte, argumentan que para la existencia de una comunidad virtual es necesario que se cumplan cuatro condiciones: interactividad; más de dos agentes; un espacio público de comunicación y, por ultimo un sentimiento de pertenencia. Granovetter (1973) considera que la interactividad supone el intercambio, el dialogo y la dependencia entre los agentes implicados en esta comunicación. Cada agente presta atención al flujo de información, poniendo énfasis en el flujo bidireccional, así como en las conversaciones surgidas entre agentes, creando sensaciones de comunidad (Burbules y Smith, 2005). Tales interacciones construyen una red de información compartida, de aprendizaje y debate, que refleja como las personas interaccionan entre ellas y se implican en esa red social (Gruzd, 2012).

Por tanto, asumiendo que el sistema de correo electrónico, crea una comunidad online, la combinación del análisis de redes sociales con las técnicas de extracción de correo electrónico, proporciona una perspectiva analítica que provee información y visualiza como son las interacciones entre agentes. Desde esta perspectiva, se considera el sistema de correo electrónico, conforma una estructura de interacción entre diversos agentes, lo cual nos permite extraer información sobre aspectos organizativos (Klimt y Yang, 2004; Bekkerman et al., 2004; Berry y Browne, 2005; Priebe et al., 2005; Keila y Skilliconr, 2005). McCallum et al (2007), combinan la red social y email buscando similaridades entre personas en una organización. Smith (2009), analiza la red de correo electrónico de una comunidad, con el objetivo de obtener los niveles de autoridad en la misma, en función de la posición de los agentes en la red de comunicación. Frank (2005), analizan el caso del departamento de Computer Science, con el objetivo de estudiar las interacciones entre los miembros del departamento. Trier (2008), en un estudio longitudinal demuestra como la combinación del análisis de redes y emails, permite no solo el estudio del proceso de comunicación en una organización, sino también como esta estructura reacciona ante eventos externos. Diesner et al. (2005) analizan el caso de Enron, estudiando los comportamientos del sistema de correo electrónico y estableciendo un paralelismo con la evolución de la situación económica. Estos Autores, encuentran, que en situaciones de crisis, la red de correos electrónico es más diversificada, siendo más homogénea cuando nos alejamos de la crisis.

METODOLOGÍA

Proceso de adquisición de datos del sistema de correo electrónico

Causas por las cuales MS-Outlook no es adecuado

Siguiendo la metodología de Cain (2012) y Basso (2011), el sistema de correo usado en esta investigación ha sido Microsoft Exchange 2012 y Microsoft Outlook 2013, que es el más popular sistema de administración de correo electrónico[1]. Este sistema permite la posibilidad de exportación de correos en el formato de la Tabla 1. Sin embargo, observamos que falta un dato importante para el análisis: la fecha de envío del correo electrónico.

Tabla 1

Formato de exportación desde Microsoft Outlook

De	(Nombre)
De	(Dirección)
De	(Tipo)
A	(Nombre)
A	(Dirección)
A	(Tipo)
CC	(Nombre)
CC	(Dirección)
CC	(Tipo)
BCC	(Nombre)
BCC	(Dirección)
BCC	(Tipo)

Otra posibilidad que ofrece Microsoft Outlook para exportar información es utilizar la función de cortar, copiar, pegar de una tabla preparada con las columnas requeridas.

Esta función, que en el primer momento parece práctica, y que puede ser para algunos casos suficiente, sin embargo muestra una debilidad contundente para el análisis exacto de los datos: no se pueden extraer datos de la dirección de correo electrónico en el formato <nombre>@<dominio>.<región>. Solo es posible ver el alias de la dirección y de esta forma no es posible hacer una correlación exacta entre el que manda un correo y el que la recibe. Por ejemplo si en mis contactos tengo guardado un

[1] http://www.peterdehaas.net/ibm/ Computer Profile, Mayo 2014

contacto con el nombre "Carlos Aguilera (DELL)" con dirección c.aguilera@dell.com pero esa persona cuando manda su correo llega con el nombre de "Aguilera, Carlos" <c.aguilera@dell.com>, en MS-Outlook no habrá una correspondencia entre esos dos correos, ya que cuando se manda un correo se envía a "Carlos Aguilera (DELL)" pero se recibe de "Aguilera, Carlos". Esa comodidad en MS-Outlook es un problema para la investigación. Así, es necesario buscar un método diferente para recolectar los correos en una forma que pueda ser analizada.

Método usado para obtener los datos

Buscando alternativas y habiendo probado otros sistemas tales como Outlook Express, Mozilla Thunderbird o Lotus Notes entre los tradicionales y Eudora, KMail, SeaMonkey Mail o Pegasus Mail entre los menos tradicionales, hemos encontrado un medio práctico extracción regular y permanente de correos electrónicos. Se trata de acceder a un servicio de archivo tal como MailStore (http://www.mailstore.com/) donde es posible exportar todos los correos en formato EML. Una vez se ha completado la exportación de correos en formato EML necesitamos un programa que extraiga de todos esos correos la información necesaria para el análisis. El formato que encontramos es del siguiente tipo:

From: "Nombre completo" <correo@dominio.región>
To: "Nombre completo" <correo@dominio.región>
CC: "Nombre completo" <correo@dominio.región>
BCC: "Nombre completo" <correo@dominio.región>
Subject: <Texto legible>
Thread-Topic: <Texto legible>
Thread-Index: <Código índice del mensaje>
Date: <día, fecha, hora y zona horaria>
Message-ID: <Código único del mensaje>
Content-Language: <Código de lenguaje>
X-MS-Has-Attach: <Indicativo si tiene adjuntos>
X-MS-TNEF-Correlator: <Correlator>
Content-Type: <Descripción del contenido>
MIME-Version: <Versión MIME>
X-MailStore-Folder-UTF7: archivo-donde-se-encuentra-el-correo-original
X-MailStore-Message-ID: <Código único del mensaje en el archivo>
X-MailStore-Header-Hash:<Hash de la cabezera>
X-MailStore-Date: <Fecha del archive>
X-MailStore-Flags: <Indicativo interno>

Con ayuda de un programa escrito en Microsoft Visual Basic 2010, se extraen las informaciones necesarias de cada uno de los archivos. Estos son:

From: "Nombre completo" <correo@dominio.región>
To: "Nombre completo" <correo@dominio.región>
CC: "Nombre completo" <correo@dominio.región>
BCC: "Nombre completo" <correo@dominio.región>
Date: <día, fecha, hora y zona horaria>
Message-ID: <Código único del mensaje>

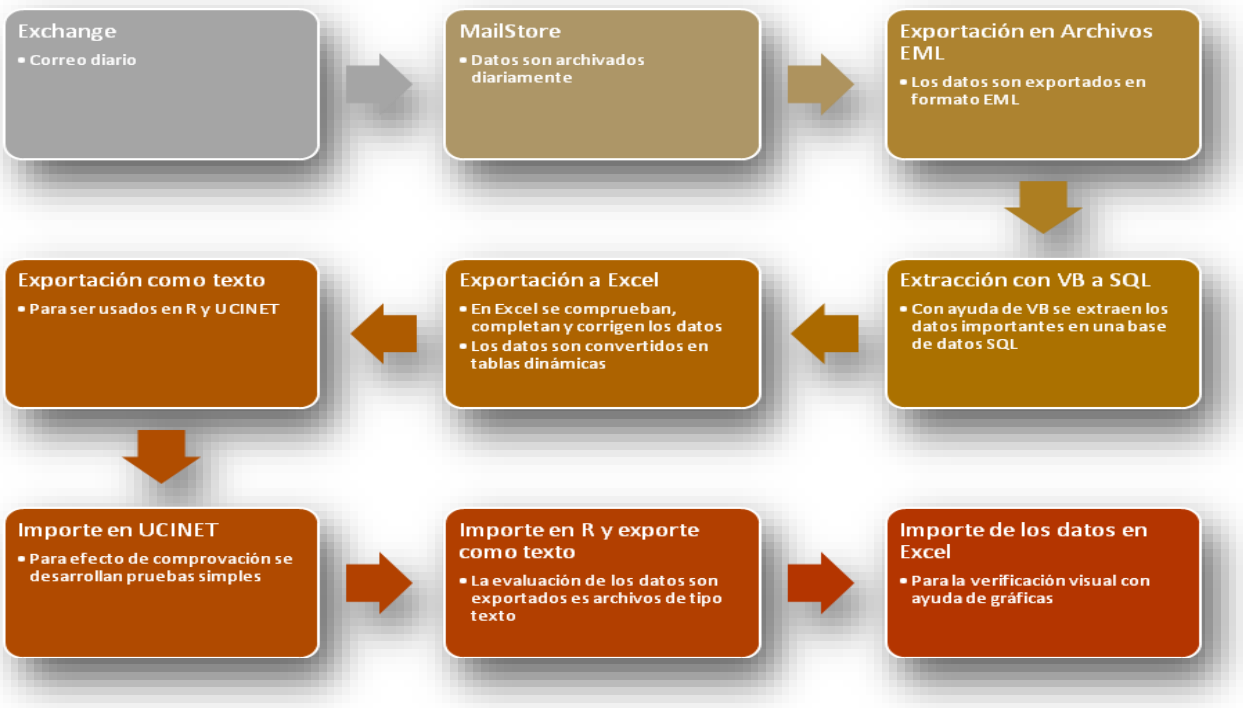

Gráfico 1. Proceso de adquisición de datos.

El resto de la información contenida en el correo no es necesaria para el análisis y como en el caso del texto del mensaje tendría problemas de protección de datos si se usara.

En algunos casos, como es el caso de la compañía objeto de estudio, todo empleado posee al menos dos direcciones que son sinónimos de la misma cuenta y persona que la usa. Además, los miembros de posiciones ejecutivas o de gestión igualmente tienen una tercera dirección de correo electrónico de la función. Por ejemplo:

Correo oficial: francisco.rodriguez@fmo.de

Correo corto: rod@fmo.de

Correo de la función: cio@fmo.de

Por este motivo en el programa de extracción se utiliza una base de datos adyacente que iguala todos estos correos en uno solo. La base de datos en donde se archivan los datos esenciales tiene la forma siguiente:

Tabla 2

Descripción de los registros de la base de datos

Name	Type
SERIALNR	Auto value
EMAILID	Text
EMAILDATE	Date/Time
EMAILFROM	Text
EMAILTO	Text
EMAILCAT	Text
EMAILCLASS	Text

Se observa que no existe un registro especial para CC o BCC. Durante el proceso de diseño de la base de datos se probaron diferentes métodos. El principal problema surge al tratar de diferenciar o dar un valor a la conexión entre los nodos si éstos están conectados en forma directa (TO:), en forma indirecta (CC:) o en forma ciega (BCC:). Es decir, una conexión directa tiene que tener un valor mayor que una indirecta, puesto que en una conversación directa se espera de los actores una actividad más estrecha entre los mismos, en tanto que aquellos que son pasivos en la interacción, tendrán un valor intrínseco menor. Es por eso que se introduce un nuevo registro en la base de datos, es el registro EMAILCAT. En dicho registro se guardan dos valores diferentes: 1 para una comunicación directa (TO:) y 2 para una comunicación indirecta (CC: o BCC:), ya que en el análisis no se hace diferencia alguna entre BCC y CC. También introducimos un nuevo registro llamado EMAILCLASS cuyo objetivo es clasificar los diversos tipos de correos, en función de quién lo ha emitido. Dichas categorías se pueden adaptar al perfil de la empresa que pretendemos analizar. Esta acción tiene que hacerse en forma manual, ya que todo automatismo puede conducir a clasificaciones erróneas.

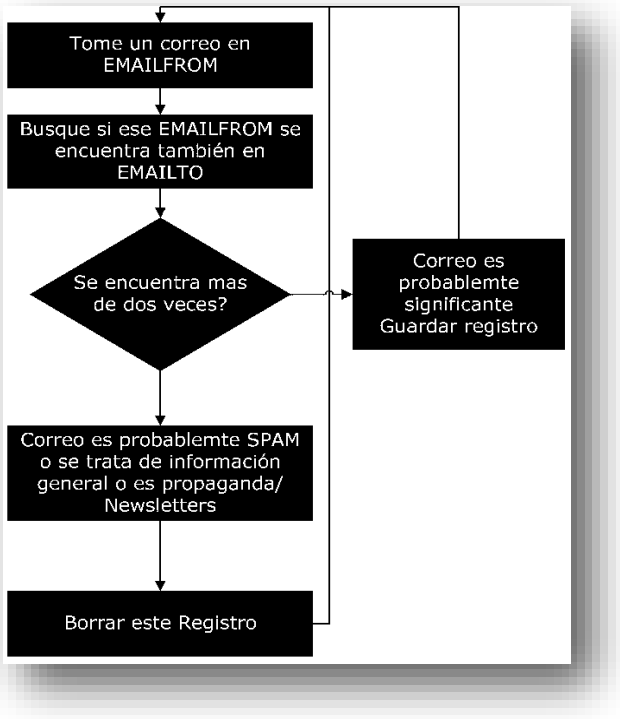

Gráfico 1. *Diagrama de flujo de datos para borrar correos SPAM.*

De la Información contenida en EMAILFROM y EMAILTO se pueden extraer datos importantes para el análisis, tal como el dominio, región o nombre del proveedor. Igualmente haciendo una comparación entre estos dos campos, es posible filtrar toda la información no significante para la detección de una red social tales como el correo spam, newsletters, propaganda, correo a grupos etc[2]. El método recomendado es mostrado en Gráfico 1.

[2] Posibles problemas y recomendaciones para solucionarlos
Los siguientes problemas fueron corregidos con diferentes métodos:

- Mensajes enviados por una persona pero que son recibidos por varios. Para esto se escribió un pequeño macro en Excel que se encarga de generar un registro por cada receptor, tanto para el destinatario como para el originador.
- Las direcciones de una misma persona pueden ser dadas de diferentes formas, con o sin título, con o sin nombre, la dirección de correo electrónico con o sin designación, apellido o dirección incompleta, sobre correo general o de empresa. La corrección de esta anomalía requiere tiempo de análisis tanto visual como funcional y se hace con ayuda de tablas dinámicas.
- El monto de correo llamado spam puede ser significativo, pero debido a su calidad única o a la existencia de un máximo dos (por los filtros anti-spam), eliminando los correos cuyo originador solo aparece 1 ó 2 veces en el espacio analizado, se disminuye el nivel de ellos.
- Por último un correo que aparece muchas veces pero que para el análisis de la red tiene poco significado son los llamados “Newsletters”. Estos son filtrados en una forma sencilla: todo aquel originador que no existe como destinatario es excluido de la base de datos. El diagrama de flujo de datos del Gráfico elimina éstos.

Tabla 3

Ejemplo de Matriz de Adyacencia

	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	N11	N12	N13	N14	N15	N16	N17
N1	*0*	20,5	12,5	3	7,5	22,5	7	8	0	11	22	14	13	12	34	2	1
N2	72	*0*	7	91	12	17	47	24	44	3	33	74	97	99	99	38	42
N3	3	26	*0*	15	82	82	16	30	50	51	13	34	79	18	86	61	27
N4	4	96	92	*0*	68	71	98	91	98	36	89	77	29	55	86	14	4
N5	42	76	51	41	*0*	67	12	8	64	43	36	93	73	54	21	31	47
N6	76	83	6	61	8	*0*	68	92	68	44	17	90	11	30	7	20	17
N7	76	28	52	97	30	24	*0*	75	90	17	61	51	59	63	75	52	17
N8	73	12	14	77	27	40	1	*0*	11	88	72	39	36	98	28	62	88
N9	27	28	33	3	26	51	92	18	*0*	61	46	70	18	12	52	41	73
N10	30	4	62	3	47	33	37	85	10	*0*	57	8	23	15	15	73	87
N11	11	26	36	36	73	19	68	14	42	60	*0*	60	61	48	90	50	48
N12	45	42	50	36	34	7	4	94	55	72	6	*0*	59	64	6	40	44
N13	60	27	90	7	84	96	88	33	37	8	11	95	*0*	73	75	20	39
N14	21	71	91	59	59	40	27	41	49	50	20	49	44	*0*	33	62	84
N15	52	85	89	33	20	35	37	95	26	39	73	70	1	31	*0*	91	44
N16	16	32	7	35	68	23	70	56	19	16	73	25	100	73	21	*0*	24
N17	85	41	39	2	46	61	78	14	32	96	28	59	87	86	92	78	*0*

Creación de la red

Las conexiones entre los actores en la red social se pueden mostrar en un grafo. Este grafo, para nuestra aplicación, presenta las siguientes características. Primero, que es ponderado (Coleman, 1986; Newman, 2010), es decir, tiene como vértices a los nodos de la red, las aristas corresponden a la comunicación o vínculo entre los nodos y la ponderación del link está determinada por la cantidad de correos intercambiados. Dichos correos además, están individualmente ponderados teniendo en cuenta si se establecen de forma directa (TO:) o indirecta (CC: o BCC:). Segundo, que el grafo es conexo (Newman, 2010), es decir, existe un camino para llegar a otro nodo. Tercera, es un grafo dirigido o dígrafo (Coleman, 1986; Newman, 2010), es decir no todos los nodos tienen una conexión bilateral. Esto último se observa especialmente cuando se filtran los datos que no son influyentes en la red, tales como Newsletters, Spam y Publicidad.

A continuación mostramos un ejemplo del grafo correspondiente a dos correos electrónicos:

From: "Nodo77" <Nodo77@dominio2.zona3>
To: <Nodo1@dominio1.zona1>
CC: "Nodo78" <Nodo78@ dominio2.zona3>
From: "Nodo1"<Nodo1@dominio1.zona1>
To: "Nodo77" <Nodo77@dominio2.zona3>
CC: "Nodo78" <Nodo78@ dominio2.zona3>, "Nodo2"Nodo2@dominio1.zona1

El grafo correspondiente se puede visualizar de esta manera:

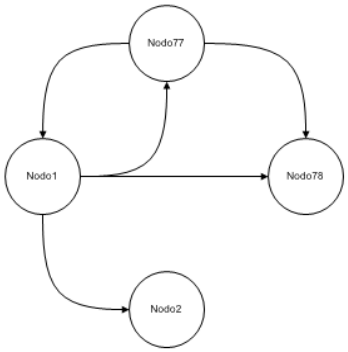

Gráfico 2. Dígrafo ponderado de dos correos.

Como podemos ver tenemos un dígrafo que conforma la red (Coleman, 1986), a partir del cual podemos obtener las matrices de adyacencia y de incidencia.

La matriz de adyacencia:

$$\begin{pmatrix} 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

En esta matriz tanto las columnas como las filas representan los nodos, tomando el valor 1 cuando existe una conexión o 0 si no existe conexión.

La matriz de incidencia:

$$\begin{pmatrix} 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \end{pmatrix}$$

En esta matriz las columnas representan los vínculos y las filas los nodos, dándose el valor de 1 si el nodo tiene ese vínculo y cero en el caso contrario.

Como hemos señalado el grafo es ponderado, por lo que hemos dado el valor a los vínculos en función de la cantidad de correos intercambiados de un nodo a otro. En dicha ponderación también hemos tenido en cuenta la categoría de la conexión, dando el valor 1 a las conexiones directas y 0,5 a las conexiones indirectas. En la tabla 3, se muestra un ejemplo de los datos adquiridos para un mes específico.

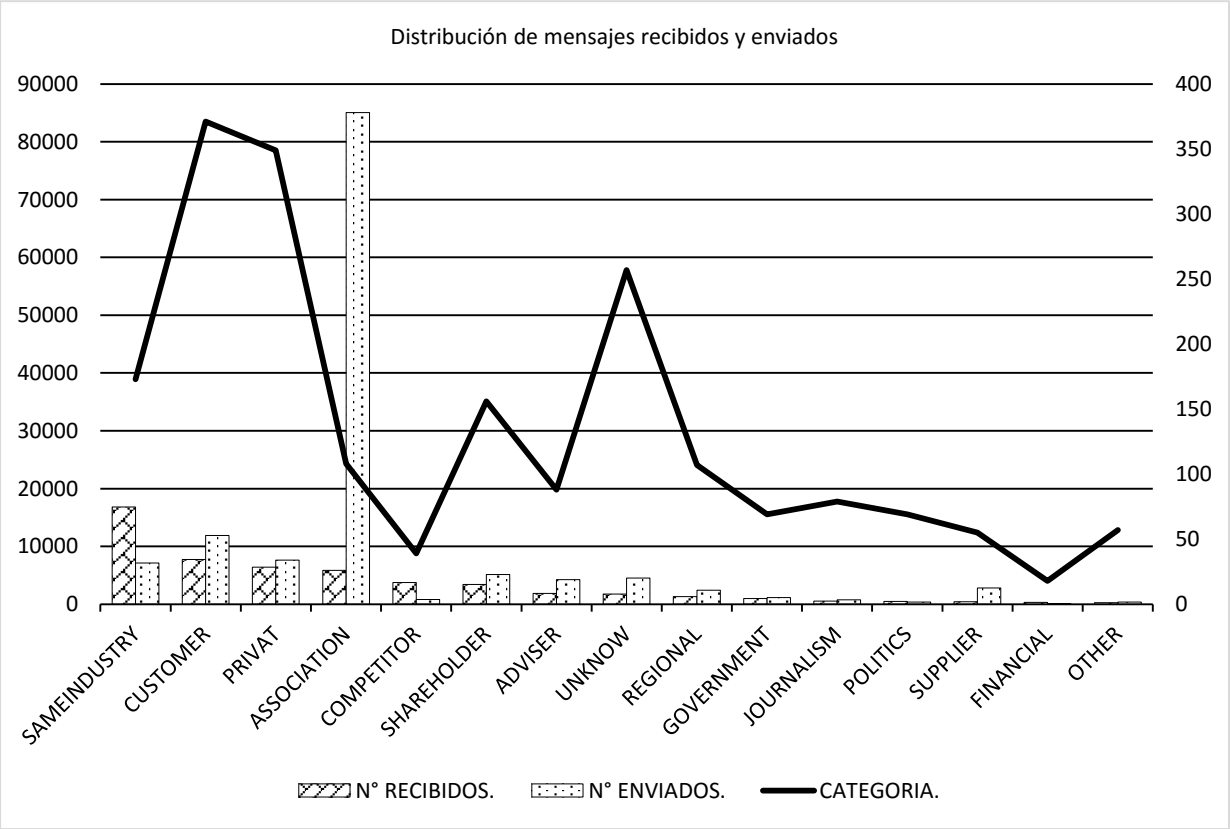

Gráfico 3. Distribución de mensajes recibidos y enviados.

Ejemplo de aplicación: El Aeropuerto de Muenster y Osnabrueck

El Aeropuerto de Muenster y Osnabrueck es una empresa que por su definición tamaño y balance general se puede clasificar como empresa mediana; sin embargo teniendo en cuenta la propiedad de la empresa se puede considerar como empresa gubernamental, ya

que todos los propietarios son empresas públicas.

El Aeropuerto de Muenster y Osnabrueck es una "sociedad con responsabilidad limitada" con una estructura central con aproximadamente 120 empleados y una serie de empresas afiliadas especializadas en diferentes áreas tales como el servicio de rampa (100%), el servicio de carga (33%), el servicio de Check-In, tickets y embarque (33%) y el servicio de control de seguridad (100%).

Es en esta empresa en la que vamos a aplicar la metodología descrita en las secciones anteriores. El periodo de recopilación de datos es del 1 enero del 2002 al 31 de diciembre del 2013, acumulados mensualmente. El objetivo de este caso de aplicación, por una parte es validar la metodología descrita y determinar la red social de correos electrónicos. En segundo lugar, analizar dicha red social y, siguiendo los trabajos de Diesner et al. (2005), y Diesner y Carley (2005), interpretar la evolución de las características estructurales de la red. Hemos considerado los datos extraídos de la red del aeropuerto con sus Stakeholders, lo cual nos permite estudiar la compañía como un sistema abierto (Kadushin, 2012), en el que se producen múltiples interacciones entre ésta y los agentes que conforman su entorno.

Hemos considerado como criterio de análisis el número de agentes que intervienen. Siguiendo un enfoque organizativo y estratégico (Bea y Haas, 2012), clasificamos los siguientes grupos: grupo de consejería (ADVISER) y del sector financiero (FINANCIAL); proveedores (SUPPLIER); otros grupos (OTHER); vínculos de carácter regional (REGIONAL); vínculos con los competidores (COMPETITOR) y con empresas pertenecientes a la misma industria (SAMEINDUSTRY). También se han considerado los grupos de carácter político (POLITICS) y de gobierno (GOVERNMENT); el grupo de propietarios de la empresa (SHAREHOLDER), así como el grupo de periodistas (JOURNALISM). Éste último, por el carácter público de la empresa pasa a tomar un carácter especial y por eso se ha tenido en cuenta por separado. El último grupo incluido es el de las asociaciones (ASSOCIATION) en las cuales la compañía tiene representación o bien en aquéllas que por su carácter y forma pueden influenciar el desarrollo de la empresa sustancialmente, tal es el caso de las organizaciones que reglamentan el tráfico aéreo en una u otra forma (ICAO, IATA, ACI, la Unión Europea, DFS, EUROCONTROL, ADV entre otras).

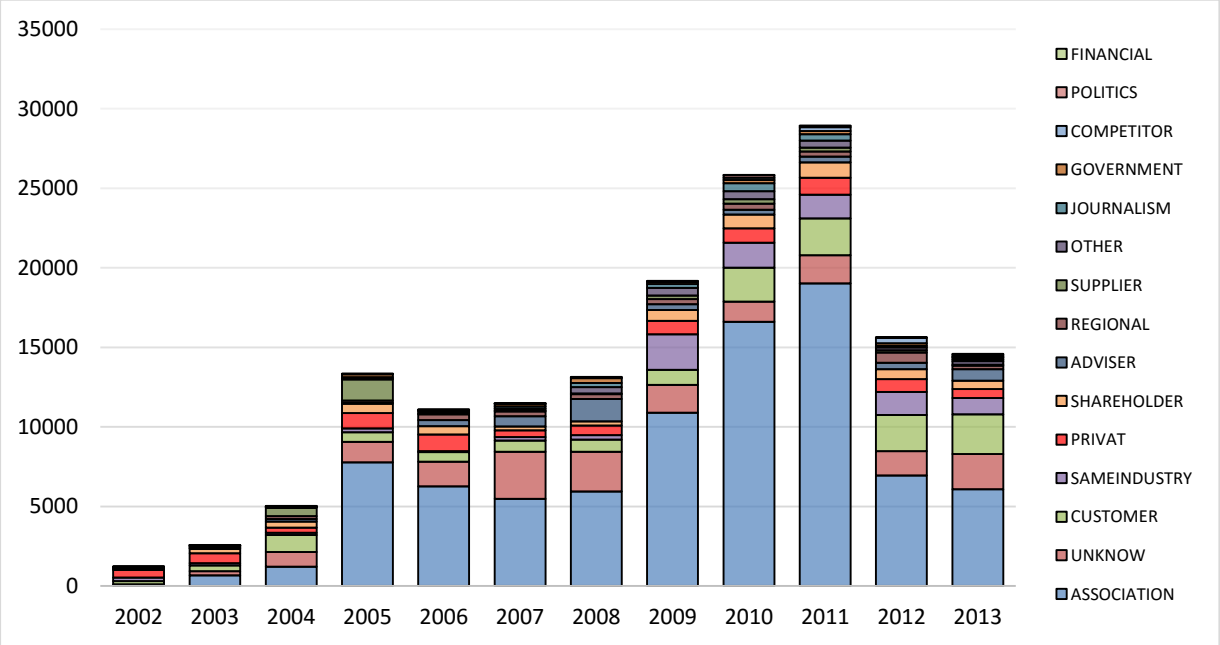

Gráfico 4. Cantidad de mensajes a través del tiempo por grupos y años.

En primer lugar determinamos el alcance de la red. Para ello listamos y clasificamos todos los nodos integrantes de la red social. En el Gráfico ***3*** mostramos la distribución por categoría de los integrantes de la red social, así

como los mensajes enviados y recibidos por grupo de clasificación.

Una vez clasificados los mensajes recibidos entre la empresa y los diferentes grupos que hemos determinado obtuvimos una cantidad de 185.276 mensajes en el período del análisis. De éstos una vez eliminados aquellos mensajes en los cuales el que envía o el que recibe solo es transmisor o receptor, nos deja un volumen de 69.049 mensajes.

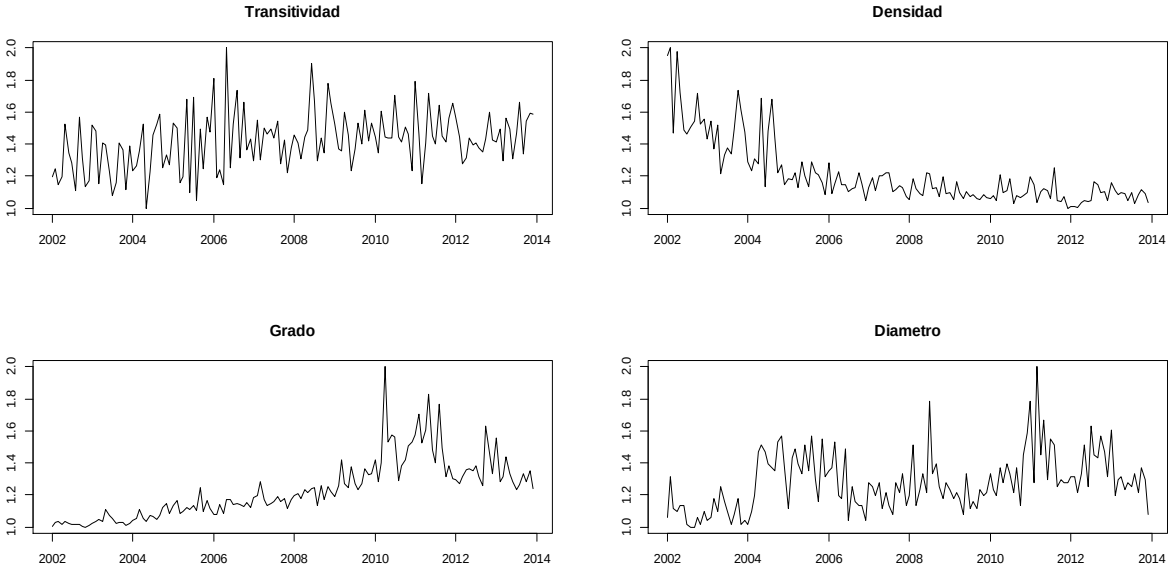

Gráfico 5. Ejemplo de visualización de Transitividad, Densidad, Grado y Diametro de la red social.

Para el análisis de los datos de la red social de Stakeholders del Aeropuerto de Muenster y Osnabrueck hemos utilizado el programa UCINET. Abordamos el estudio exclusivamente a nivel red, determinando las principales propiedades de la red y su evolución en el tiempo.

Siguiendo los trabajos previos de Coleman (1986), Newman (2010), Parkhe et al. (2006) y Borgatti y Halgin (2011), y considerando que la principal características de la red es su heterogeneidad, analizamos las siguientes propiedades: Transitividad, Densidad, Grado, y Diámetro del grafo o red. La primera medida, transitividad or clustering, se define como la probabilidad de que "el socio de mi pareja sea también mi socio", y da una idea de lo que se conoce como la estructura de vecindad de la red (Brass, 1995; Borgatti y Halgin, 2011 y Scott 2000). De acuerdo a Albert y Barabasi (2002) y Borgatti et al. (2009), medimos la transitividad a través del número de triples relaciones dividido por el potencial número de relaciones transitivas que podrían darse en la red. La segunda medida, la densidad, se midió como la suma del número real de enlaces de todos los agentes de la red dividido por la suma de la cantidad máxima posible de lazos entre todos los componentes (Contractor y Monge, 2003; Hagedoorn et al., 2000; Parkhe et al., 2006). La tercera medida, el grado del grafo, siguiendo a Borgatti y Halgin (2011) y Scott (2000), se calculó a través del valor medio del número de conexiones de cada agente. Por último, el diámetro de grafo, que da una idea del tamaño de la red, se midió a través del número de agentes interconectados. En el Gráfico 5 se recogen los cuatro atributos de la red social conformada por los correos electrónicos del Aeropuerto de Muenster y Osnabrueck y sus Stakeholders.

Los resultados muestran que la densidad disminuye a lo largo del tiempo, mientras que el grado de la red presenta un perfil creciente igual que el diametro de la red. Esta evolución podría explicarse por el incremento en el número de contactos como consecuencia de una mayor actividad económica. Sin embargo, como el diámetro de la red no se incrementa sustancialmente, ello hace pensar que se ha producido un desplazamiento de las comunicaciones hacia otros agentes, aunque manteniendo el mismo nivel de comunicación con los antiguos agentes. Por último, los resultados también muestran que el nivel de transitividad se mantiene a lo largo del periodo estudiado lo cual permite señalar la constancia de roles y afinidades en la comunicación entre los agentes. La variación dentro de cada año (mes a mes) responde al componente estacional de este tipo de negocio.

CONCLUSIÓN

El análisis de las redes sociales, se está convirtiendo en una potente herramienta para determinar los fenómenos de relación entre agentes. Más concretamente, y desde el punto de vista económico, las empresas necesitan analizar sus datos con el objetivo de recabar información que pueda ser útil para la toma de decisiones. Una fuente muy importante de información para la empresa pueden ser los correos electrónicos. Hemos demostrado que el estudio del correo electrónico proporciona una visión de la estructura de las redes de contacto de una empresa. No sólo nos revela quién se relaciona con quién, sino que nos proporciona un contexto a través del cual analizar las relaciones internas y externas de la empresa. La obtención de datos del correo electrónico puede ser un proceso costoso y que requiere mucho tiempo para su procesamiento. En este estudio hemos mostrado una forma de aprovechamiento de esta información a través de una metodología sencilla y factible que permite la captación de datos del sistema de correo electrónico de las empresas. Dicha información ha servido para analizar la estructura de la red y describir el comportamiento de la red de relaciones a lo largo del tiempo. También hemos determinado los principales atributos de la red social de contactos de la empresa estudiada, los cuales pueden servir para inferir las características de la red establecida entre la firma y sus principales grupos de interés.

Desde el punto de vista de las implicaciones para la gestión, podemos señalar que esta metodología es factible de aplicación para las empresas, respetando los criterios legales de privacidad. El uso de esta técnica puede permitir a las empresas el acceso a una potencial fuente de datos que, modelizados a través del análisis de redes sociales, proporcione información sobre la evolución de la gestión de la empresa. En este sentido, a través de la red de relaciones de la empresa es factible analizar, las estrategias corporativas, de negocio y funcionales de la compañía. Por ejemplo, una tendencia a diversificar, puede suponer un mayor número de nuevos contactos, o la mayor densidad de la red, una intensificación en la relaciones con nuestros suministradores.

Nuestro trabajo no esta exento de limiatciones. La principal es la necesidad de filtrar los datos de email, con el objetivo de buscar tendencias a largo plazo, y eliminar posibles espurios y ruido a corto plazo. Esta labor de filtrado, puede permitir, establecer patrones de comportamiento con las diferentes variables económicas de las empresas.

Esta contribución abre una línea de investigación que permite comparar la gestión interna de las empresas, en función de las características y evolución de la red social derivada de sus contactos a través del correo electrónico.

REFERENCIAS

Adamic, L. A. y Adar, E. (2003). Friends and neighbors on the Web. *Social Networks*, 25, 211–230. doi: http://dx.doi.org/10.1016/ s0378 -8733(03)00009-1

Albert, R. y Barabási, A.L. (2002). Statistical mechanics of complex Networks. *Reviews of Modern Physics*. http://arxiv.org/pdf/cond-mat/0106096.pdf

Alstyne, M. y Zhang, J. (2003). *EmailNet: A system for automatically mining social networks from organizational email communication.* http://www.researchgate.net/publication/215439725

Basso, M. (27 de Diciembre de 2011). Magic Quadrant for Enterprise Wireless Email Market. http://www.gartner.com/technology/reprints.do?id=1-BOC85F&ct=120809&st=sb

Baym, N., Zhang, Y. y Lin, M. (2004). Social interactions across media: Interpersonal communication on the internet, telephone and face-to-face. *New Media & Society*, 6(3), 299–318. doi: http://dx.doi.org/10.1177/1461444804041438

Bea, F. X. y Haas J. (2012*).* *Strategisches Management*, Ed. 6. UTB, Stuttgart.

Bird, C., Gourley, A., Devanbu, P., Swaminathan, P. y Gertz, M. (2006). Mining email social networks in postgres. En MSR '06: *Proceedings of the International Workshop on Mining Software Repositories*, 225-237. doi: http://dx.doi.org/10.1145/1137983.1138033

Bekkerman, R., McCallum, A. y Huang, G. (2004). Automatic Categorization of Email into Folders: Benchmark Experiments on Enron and SRI Corpora. *Technical Report IR-418*, CIIR, University of Massachusetts.

Berry, M. y Browne, M. (2005). Email Surveillance Using Nonnegative Matrix Factorization. *Computational & Mathematical Organization Theory*, 11, 249–264. doi: http://dx.doi.org/10.1007/s10588-005-5380-5

Borgatti, S. P. y Foster, P. C. (2003). The Network Paradigm in Organizational Research: A Review and Typology. *Journal Management* 29, 991–1013. doi: http://dx.doi.org/10.1016/ s0149-2063_03__00087-4

Borgatti, S. P. y Halgin, D. S. (2011). On Network Theory. *Organization Science, 22*(5), 1168-1181.

Borgatti, S. P., Jones, C. y Everett, M. G. (1998). *Network measures of social capital. Connections, 21*(2), 27-36.

Borgatti, S. P. y Lopez-Kidwell, V. (2011). Network Theorizing. En Carrington, P. y Scott, J. (eds) *The Sage Handbook of Social Network Analysis.* Sage Publications.

Borgatti, S. P., Mehra, A., Brass, D. y Labianca, G. (2009). Network Analysis in the Social Sciences. *Science, 323*(5916), 892–895.

Borkar, V. R., Deshmukh, K. y Sarawagi, S. (2000). Automatically extracting structure from free text addresses. *IEEE Computer Society*, 23(4), 27-32.

Brass, D. (1995). A Social Network Perspective on Human Resources Management. *Research in Personnel and Human Resources Management,* 13, 39-79.

Burbules, N. C. y Smith, R. (2005). "What it makes sense to say": Wittgenstein, rule-following and the nature of education. *Educational Philosophy and Theory*, 37(3), 425-30. doi: http://dx.doi.org/10.1111/j.1469-5812.2005.00130.x

Cain, M. (2012). *Market Scope for Email Systems.* http://www.gartner.com/technology/reprints.do?id=1-1BOC85F&ct=120809&st=sb

Castells, Manuel (2001). *La era de la información. Vol. 1, La Sociedad en Red. Alianza Ed,* Madrid -1ª reimpresión-.

Chapanond, A., Krishnamoorthy, M.S. y Yener, B. (2005). Graph Theoretic and Spectral Analysis of Enron Email Data. *Computer Mathematlical Organization Theory* 11(3),265-281. doi: http://dx.doi.org/10.1007/s10588-005-5381-4

Coleman, J. (1986). Social theory, social research, and a theory of action. *American Journal of Sociology*, 91, 1309-1335. doi: http://dx.doi.org/10.1086/228423

Contractor N. S. y Monge P. (2003). *Theories of communication networks*. Oxford University Press, New York.

Diesner, J. y Carley, K.M. (2005). Revealing and Comparing the Organizational Structure of Covert Networks with Network Text Analysis. *XXV Sunbelt Social Network Conference*, Redondo Beach, CA, February 16-20, 2005.

Diesner, J., Frantz, T. L. y Carley, K. M. (2005). Communication networks from the Enron e-mail corpus it's always about the people Enron is no different. *Computational and Mathematical Organization Theory*, 11 (3) (2005), 201–228. doi: http://dx.doi.org/10.1007/s10588-005-5377-0

Ebel, H., Mielsch, L. y Bornholdt, S. (2002). *Scale-free topology of e-mail networks*. *Physical Review, 66*(3). doi: http://dx.doi.org/10.1103/physreve.66.035103

Frank, O. (2005). Network Sampling and Model Fitting. En *Models and Methods in Social Network Analysis*, J. S. P. Carrington and S. S. Wasserman, eds., 31–56. Cambridge Univ. Press, Cambridge. doi: http://dx.doi.org/10.1017/cbo9780511811395.003

Freemann, L. C., White, D. R. y Romney, A. K. (eds) (1989). *Research Methods in Social Network Analysis*. New Brunswick, NJ: Transaction Books.

Granovetter, M (1973). The strenght of weak ties. *American Journal of Sociology*, 78. 1360-1364. doi: http://dx.doi.org/10.1016/b978-0-12-442450-0.50025-0

Gross, R. y Acquisti, A. (2005). Information Revelation and Privacy in Online Social Networks (The Facebook case). *Procceding ACM Workshop on Privacy in the Electronic Society (WPES)*. AC Workshop. doi: http://dx.doi.org/10.1145/1102199.1102214

Gruzd, A. (2012). Non-Academic and Academic Social Networking Sites for Online Scholarly Communities. En Neal, D.R. (Ed.) *Social Media for Academics: A practical guide*. Chandos Publishing. doi: http://dx.doi.org/10.1016/b978-1-84334-681-4.50002-5

Gruzd, A. y Haythornthwaite, C. A. (2013). Enabling Community Through Social Media. *Journal Mediated Internet Research, 15*(10), e248. doi: http://dx.doi.org/10.2196/jmir.2796

Guinalíu, M. y Flavián C. (2003). La Comunidad Virtual. *Apuntes para asignatura Economía del Comercio Electrónico*, Universidad de Zaragoza. Disponible en: http://www.5campus.org/leccion/comunidadvirtual/comunidadvirtual.doc

Haagedorn, J., Link, A. N. y Vonortas, N. S. (2000). Research partnerships. *Research Policy* 29, 567-586.

Haythornthwaite, C. A., y Wellman, B. (1998). Work, friendship and media use for information exchange in a networked organization. *Journal of the American Society for Information Science*, 49(12), 1101–1114. doi: http://dx.doi.org/10.1002/(sici)1097-4571(1998)49:12%3C1101::aid-asi6%3E3.3.co;2-s

Jones, Q. (1997). Virtual-communities, Virtual settlements & cyber-archaeology: A theoretical outline. *Journal of Computer Mediated Communication*, 3(3), 0. doi: http://dx.doi.org/10.1111/j.1083-6101.1997.tb00075.x

Kadushin, C. (2012). *Understanding Social Networks: Theories, Concepts, and Findings*. Oxford University Press.

Keila, P.S. y Skillicorn, D.B. (2005). *Detecting Unusual and Deceptive Communication in Email*. Queen's University, Ontario, Canada. Disponible en: http://research.cs.queensu.ca/TechReports/Reports/2005-498.pdf

Kim, J., Kim, G., Park, H. y Rice, R. (2007). Configurations of Relationships in Different Media: FtF, Email, Instant Messenger, Mobile Phone, and SMS. *Journal of Computer-Mediated Communication* 12, 1183–1207. doi: http://dx.doi.org/10.1111/j.1083-6101.2007.00369.x

Klimt, B. y Yang, Y. (2004). *The Enron Corpus: A New Dataset for Email Classification Research.* European Conference on Machine Learning.

Knoke, D. y Yang S. (2008). *Social Network Analysis* -2nd ed. Series: Quantitative Applications in the Social Sciences. Sage Publications.

McCallum, A., Wang, X. y Corrada-Emmanuel, A. (2007). Topic and Role Discovery in Social Networks with Experiments on Enron and Academic Email. *Journal of Artificial Intelligence Research* 30, 249-272.

McMillan, D.W. y Chavis, D.M. (1986). Sense of community: A definition and theory. *Journal of Community Psychology*, 14. 6-23. doi: http://dx.doi.org/10.1002/1520-6629(198601)14:1%3C6::aid-jcop2290140103%3E3.0.co;2-i

Melanchthon, D. (2010). *Microsoft Exchange*. Presentación en CEBIT 2010, 2.

Newman M. E. J. (2010). *Networks: An Introduction by M. E. J. Newman, a college-level textbook about the science of networks*. Oxford University Press.

Núñez J. F. y Cárdenas, E. (2013). Identificación de asociaciones y complicidades, vía-email, de un grupo de analistas de redes sociales: ¿Qué intercambian los rederos? *Revista Hispana para el Análisis de Redes Sociales*. *24*(1), 27-52.

Parkhe, A., Wasserman, S. y Ralston, D. A. (2006). New Frontiers in Network Theory Development. *Academy of Management Review*, 31, 560-569. doi: http://dx.doi.org/10.5465/amr.2006.21318917

Priebe, E. C., Conroy, M. J., Marchette, J. D. y Park Y. (2005). Scan statistics on Enron graphs. *Computational and Mathematical Organization Theory*, 11(3), 229–247. doi: http://dx.doi.org/10.1007/s10588-005-5378-z

Reips, U.-D. (2002). Standards for Internet-based experimenting. *Experimental Psychology*, 49(4), 243-256. doi: http://dx.doi.org/10.1026/1618-3169.49.4.243

Rowe, R., Creamer, G., Hershkop, S. y Stolfo, S. J. (2007). Automated social hierarchy detection through email network analysis. *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007*. Workshop on Web mining and social network analysis, 109–117. doi: http://dx.doi.org/10.1145/1348549.1348562

Scott, J. (2000). *Social network Analysis – a handbook*. 2nd Edition. Sage Publications

Shakes, J., Langheinrich, M. y Etzioni, O. (1997). Dynamic reference sifting: A case study in the homepage domain. *Resumen del 6° World Wide Web Conference*. doi: http://dx.doi.org/10.1016/s0169-7552(97)00048-2

Shetty, J. y Adibi, J. (2004). The Enron email dataset database schema and brief statistical report. *Information Sciences Institute Technical Report*, University of Southern California.

Smith, V. (2009). Theory and experiment: What are the questions? *Journal of Economic Behavior & Organization*, 73, 3–15. doi: http://dx.doi.org/10.1016/j.jebo.2009.02.008

Trier, M. (2008). Research Note: Towards Dynamic Visualization for Understanding Evolution of Digital Communication Networks. *Information Systems Research, 19*(3), 350. doi: http://dx.doi.org/10.1287/isre.1080.0191

Upstill, T., Craswell, N. y Hawking, D. (2003). Query independent evidence in home page finding. *ACM Transactions on Information Systems*, 21(3), 286-313. doi: http://dx.doi.org/10.1145/858476.858479

Wasserman, S. y Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge University Press, Cambridge.

Wellman, B. y Leighton, B. (1979). Networks, Neighborhoods and Communities. Approach to the Study of the Community Question. *Urban Affairs Quartely*, 14(3), 363-390. doi: http://dx.doi.org/10.1177/107808747901400305

Xi, W., Fox, E. A., Shu, J. y Tan, R. (2002). "Machine learning approach for homepage finding task". *Resumen del 9° International*

Symposium on String Processing and Information Retrieval, 145–159. doi: http://dx.doi.org/10.1007/3-540-45735-6_14

Yelupula K. y Ramaswamy, S. (2008). Social network analysis for email classication. Proccedings ACM, 2008. *46th Annual Southeast Regional Conference*, 469-474.