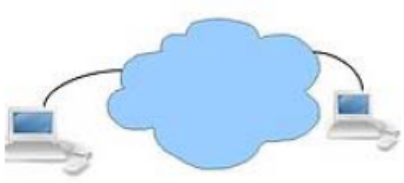
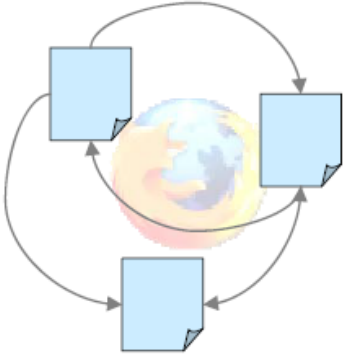
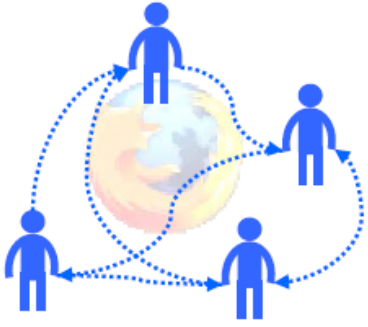
Minería de Medios Sociales
Máster en Ciencia de Datos e Ingeniería de Computadores

Tema 0: Minería de Medios Sociales

Oscar Cordón García
Dpto. Ciencias de la Computación e Inteligencia Artificial
ocordon@decsai.ugr.es

WEB 2.0 Y MEDIOS SOCIALES

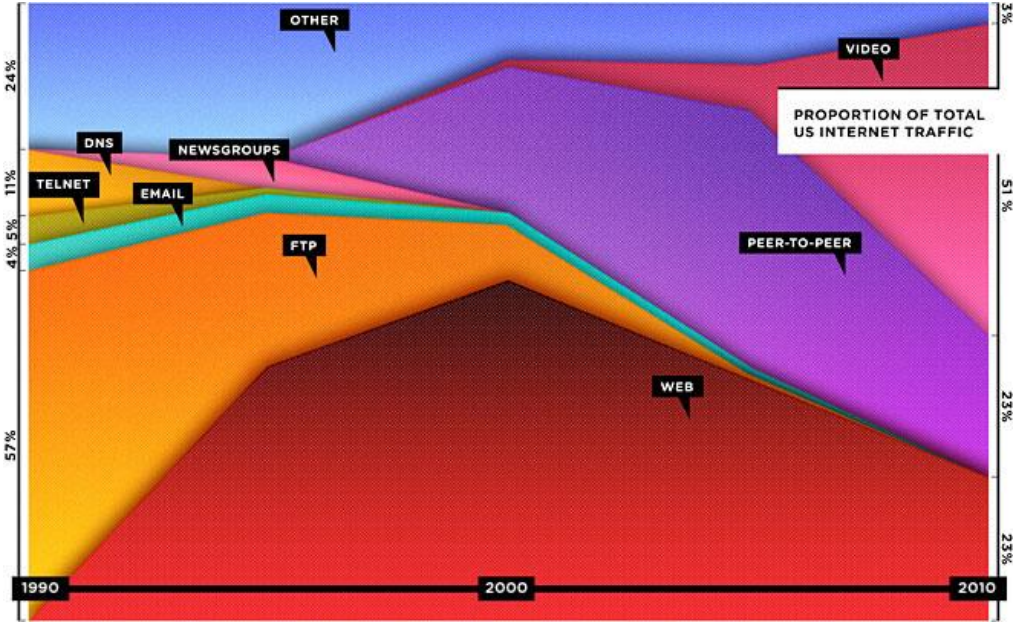
EVOLUCIÓN DE LA WEB (1)

Internet	World Wide Web	Web 2.0, The Social Web
		
connecting computers	connecting documents	connecting people
1960's	1990's	2004-present
	Google	flickr facebook
	Y!	twitter

EVOLUCIÓN DE LA WEB (2)

Chris Anderson, Wired, “The Web is dead. Long Live the Internet”:

http://www.wired.com/2010/08/ff_webrip/all/



Two decades after its birth, the World Wide Web is in decline, as simpler, sleeker services — think apps — are less about the searching and more about the getting

MEDIOS SOCIALES (*SOCIAL MEDIA*): Definición

Los ***Medios Sociales*** comprenden el uso de herramientas electrónicas y de Internet para compartir y comentar información y experiencias entre seres humanos de forma eficiente

Los ***Medios Sociales*** son un “conjunto de aplicaciones basadas en Internet que constituyen los fundamentos ideológicos y tecnológicos de la **Web 2.0**, y que permiten la creación y el intercambio de contenidos generados por los usuarios”

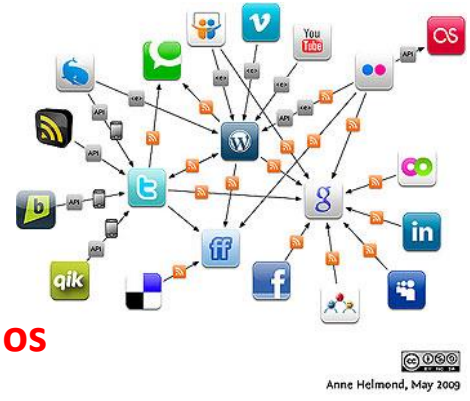
MEDIOS SOCIALES (2)

Vivimos en la era del *big data*. Con cientos de millones de personas invirtiendo horas y horas en los medios sociales para compartir, comunicar, conectar, interactuar y crear contenidos generados por los usuarios con un volumen sin precedentes, los medios sociales se han convertido en una fuente única de *big data*. Dicha fuente muestra un gran potencial para la investigación y el desarrollo. Desafortunadamente, tener más datos no necesariamente implica generar mejor información, sólo los datos de buena calidad lo permiten. Así, es necesario disponer de nuevos métodos computacionales para hacer minería sobre esos datos. Los datos de medios sociales son ruidosos, sin formato, de longitud variable y multimedia. Además, las relaciones sociales entre las entidades, llamadas redes sociales, forman una parte inseparable de los datos de medios sociales. Por tanto, es importante que se combine el uso de las teorías sociales y los métodos de investigación con los métodos estadísticos y de minería de datos. Es un tiempo propicio para la minería de medios sociales.

This chapter is from *Social Media Mining: An Introduction*.
By Reza Zafarani, Mohammad Ali Abbasi, and Huan Liu.
Cambridge University Press, 2014. Draft version: April 20, 2014.
Complete Draft and Slides Available at: http://dmml.asu.edu/smm

MEDIOS SOCIALES (3): Características Diferenciales

- Contenido generado por los usuarios: masivo, dinámico, amplio, instantáneo y ruidoso
- **Rico en interacciones entre los usuarios: datos con relaciones/enlaces (*linked data*)**
- Entorno colaborativo, sabiduría de las masas
- Muchos grupos pequeños (fenómeno de cola larga)
- La atención es cara



- **Comunicación muchos a muchos: la información llega a los usuarios**

  … vía la influencia personal en sus redes sociales

  … a través de la transmisión de los medios de masas

- Los medios sociales están diseñados para ser diseminados mediante **interacción social**
  - ¿Cómo interacciona la información transmitida con la influencia personal de las redes sociales?
  - Tensión entre los efectos globales de los medios de masas y los efectos locales provocados por la estructura social

MEDIOS SOCIALES (4): Entorno

MEDIOS SOCIALES (5):

- Redes sociales on-line (Facebook, LinkedIn)
- *Microblogging* (Twitter)
- Compartición de fotos (Flickr, Picasa)
- Agregadores de noticias (Google reader)
- Compartición de video (YouTube)
- *Livecasting* (Justin.TV)
- Mundos virtuales (Kaneva)
- Juegos on-line (Warcraft)
- Búsqueda "social" (Google, Bing)
- Mensajería instantánea (Google Talk, Skype)
- …

Categorías

MEDIOS SOCIALES (6): Ejemplos de Datos

- Un artículo de una wiki
- Un post en un blog
- Una opinión/puntuación en una web de restaurantes sobre una hamburguesería famosa de tu ciudad
  - Ejs: Yelp.com, Tripadvisor.es
- Los datos de un usuario en una red social on-line de contactos profesionales
  - Ejs: LinkedIn.com, Facebook.com
- Los datos de una app que indica la localización de las cámaras en una calle
- Los datos de una recomendación de Amazon

MINERÍA DE MEDIOS SOCIALES

MINERÍA DE MEDIOS SOCIALES (1): Definición

*La **Minería de Medios Sociales** es el proceso de representar, analizar y extraer patrones con significado a partir de datos de medios sociales*

La ***Minería de Medios Sociales*** es un área multidisciplinar formada por:

- Redes sociales:
  - interacciones y comunidades
- Medios sociales:
  - Gran cantidad de datos (masivos, dinámicos, amplios, instantáneos y ruidosos)
  - Doble naturaleza: información de contenido y relaciones
- Minería de datos:
  - Modelado, Aprendizaje y Predicción

MINERÍA DE MEDIOS SOCIALES (2): Una Nueva Área Multidisciplinar

- Aparición de nuevos fenómenos observables de las **interacciones** entre personas en los medios sociales

- Oportunidades sin precedentes para **estudiar el comportamiento humano** mediante investigación interdisciplinar y colaborativa entre las ciencias sociales y de la computación:
  - Teorías sociales derivadas de años de investigación
  - Tecnologías computacionales y algoritmos escalables para Minería de Datos/Big Data

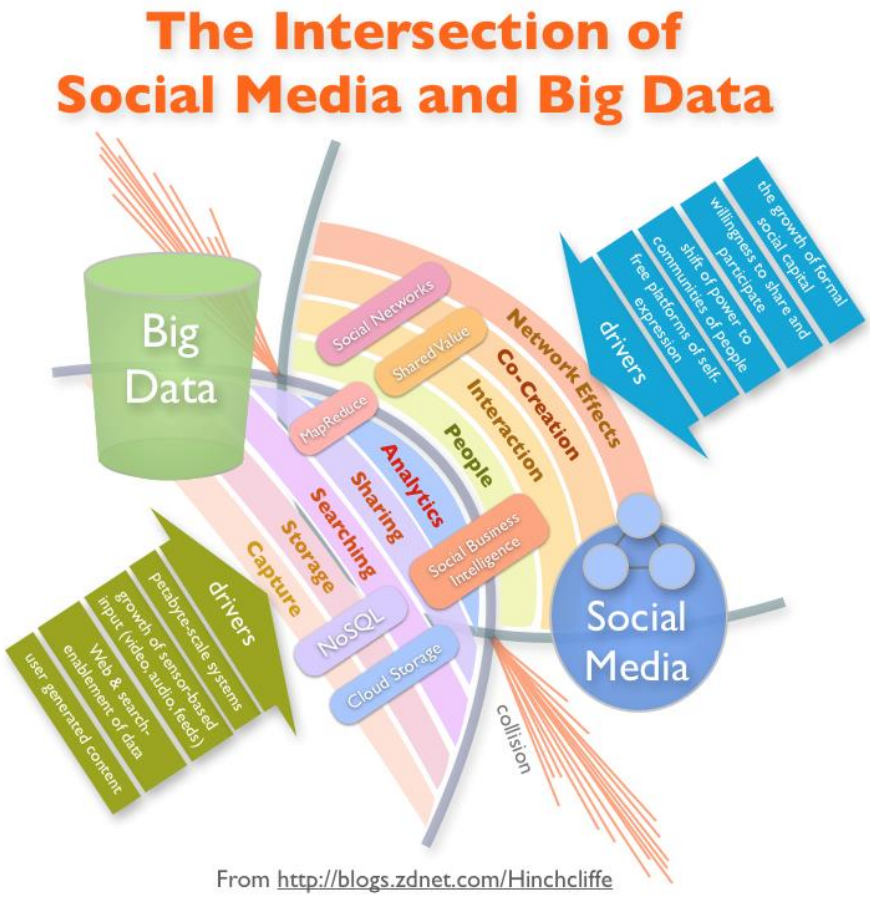
MIT Laboratory for Social Machines:
http://socialmachines.media.mit.edu/

mit media lab

MINERÍA DE MEDIOS SOCIALES (3): ¿¿Big Data??

- 1060 millones de páginas web a principios de 2014. Billones de búsquedas en Google cada día
- Un billón de usuarios de Facebook. Más de un billón de páginas de Facebook. En Marzo de 2010 Facebook implicó más del 7% del tráfico de Internet en EEUU, superando por primera vez a *Google*:

  http://money.cnn.com/2010/03/16/technology/facebook_most_visited
- Cientos de millones de cuentas de Twitter. Cientos de millones de Tweets cada día
- Información multimedia (texto, imágenes, videos, …). Comportamientos, preferencias, tendencias, …
- Facilidad de acceso: APIs, Web spiders, Conjuntos de datos existentes, etc.



MINERÍA DE MEDIOS SOCIALES (4): Aplicaciones (1)

Análisis de imagen de marca:

- ¿Qué está opinando la gente de nuestra marca? (**análisis de sentimientos/minería de opiniones**, **análisis de confianza**)

Campañas de marketing:

- Compañías tratando de posicionar sus productos: publicidad "ganada" en Redes Sociales on-line (***earned media***, **boca a boca, procesos de difusión en redes, marketing viral**)

Recomendaciones personalizadas:

- **Sistemas de recomendaciones, filtrado colaborativo, marketing viral**

Informes de productos:

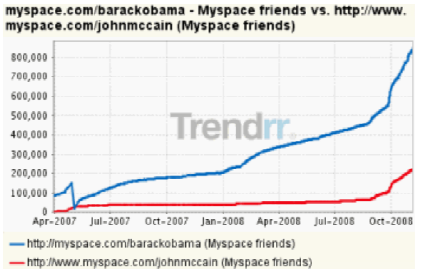
- Minería de revisiones de productos en foros para obtener información sobre la percepción de sus características, identificar nuevas necesidades, etc. (**análisis de sentimientos**, **análisis de confianza**, **minería de textos**)

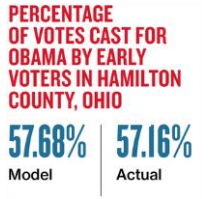
MINERÍA DE MEDIOS SOCIALES (5): Aplicaciones (2)

Campañas políticas:

- ¿Por qué los ciudadanos apoyan a un candidato? Predicción de/ Influencia en procesos electorales (**boca a boca, procesos de difusión en redes, análisis de influencia, minería de blogs**)

  http://www.technologyreview.com/featuredstory/509026/how-obamas-team-used-big-data-to-rally-voters/



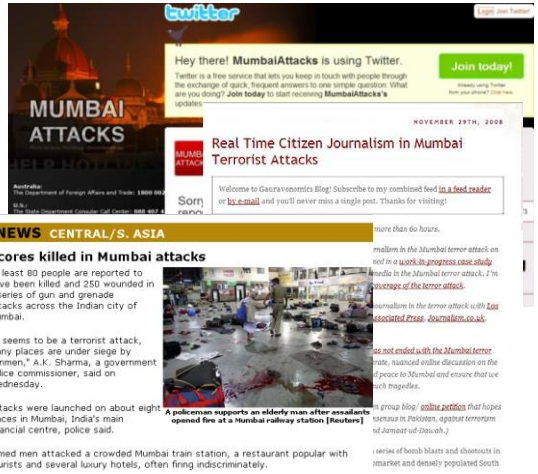


Cuerpos de Seguridad del Estado:

- Predicción de actividades tales como disturbios de bandas y manifestaciones no autorizadas (**minería de *tweets***)

  NYT: Sending the Police Before There's a Crime:

  http://www.nytimes.com/2011/08/16/us/16police.html?_r=1



Periodismo ciudadano:

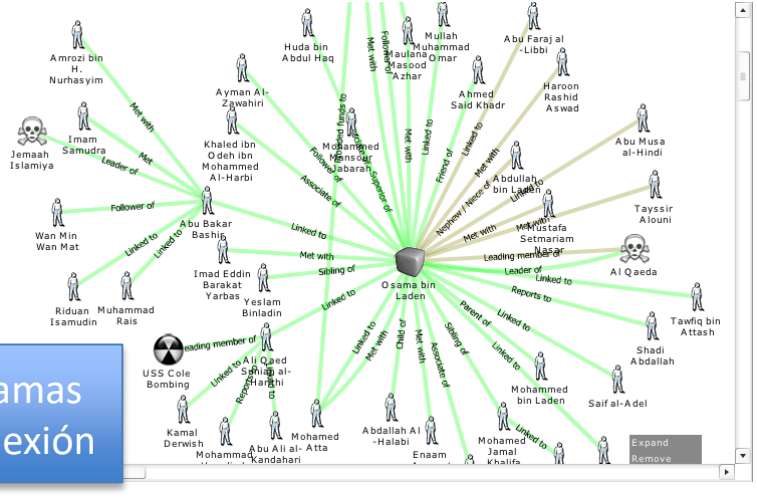
- Mayor valor que los teletipos. Problema de tener que "bucear" en muchos *posts* para localizar información útil
- Creación de resúmenes en tiempo real (**análisis de influencia, minería de textos**)

MINERÍA DE MEDIOS SOCIALES (6): Aplicaciones (3)

Predicción de comportamiento humano:

- Procesado de datos de medios sociales para obtener herramientas de análisis para:
  - Identificar redes sociales: miembros, grupos (**descubrimiento de comunidades**)
  - Identificar tópicos y sentimientos (**minería de sentimientos**)
  - Monitorización de tendencias culturales (**minería de *tweets***)
  - Salud 2.0: Predicción de epidemias (**procesos de difusión en redes complejas**)

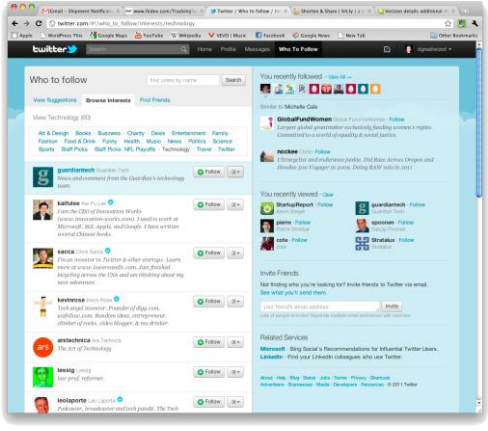
Modelado predictivo



Diagramas de conexión

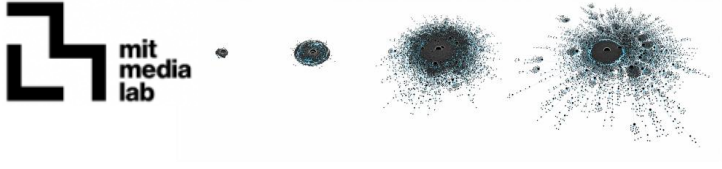
APLICACIONES (1): Noticias Sociales en Twitter

- Los usuarios tweetean mensajes cortos:
  - Se retweetean mensajes de otros
  - Los tweets pueden contener URLs de contenido on-line, noticias

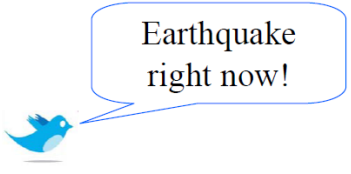


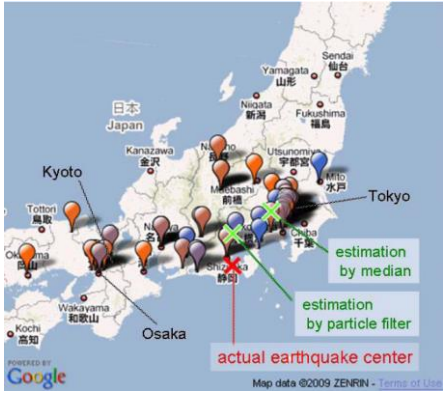
- **Redes sociales:** Los usuarios siguen a los amigos (*friends*) para ver tweets y retweets de amigos

- **Contagio social (cascadas):**
  - ¿se retwiteará un tweet? ¿hasta donde llegará?
  - La mecánica de los *memes* es similar a la de los virus biológicos e informáticos



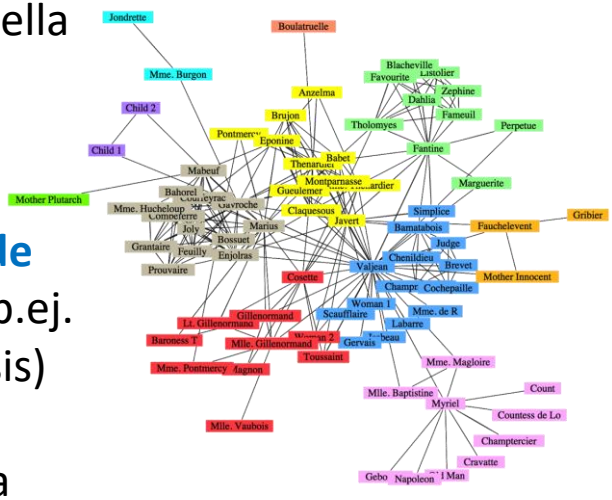
- **Detección de eventos en tiempo real:** epidemias, terremotos, movimientos ciudadanos, etc.





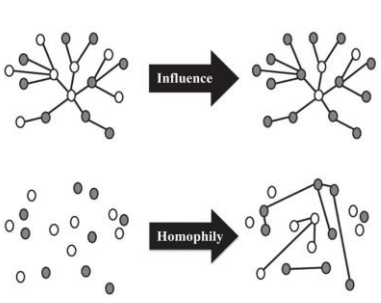
APLICACIONES (2): Influencia en la Red y Detección de Comunidades

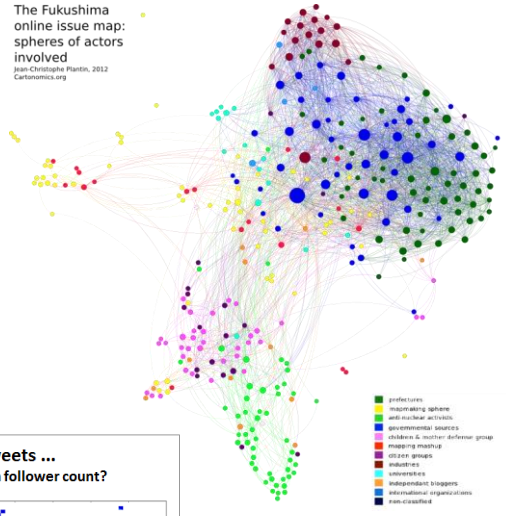
- Las redes sociales on-line y los servicios de microblogging se configuran en una **estructura de red** (compleja) **social**
- La estructura de la red determina los procesos dinámicos de transmisión de información que se producen sobre ella (**análisis de redes sociales**)
- Es posible determinar la **influencia de un usuario** a partir de la red social (p.ej. a quién seguir en Twitter en una crisis)
- Igualmente, es posible determinar la **estructura de comunidades** subyacente

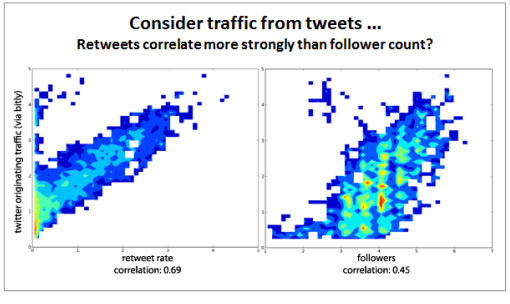


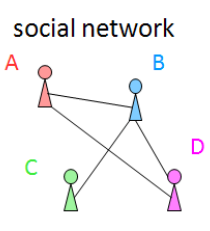
APLICACIONES (3): Influencia y Predicción de Enlaces

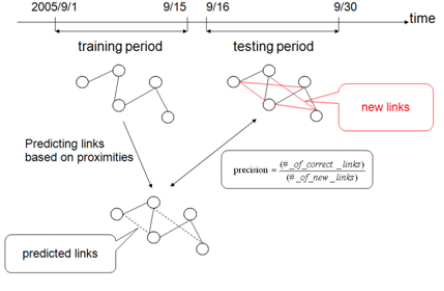
- Las teorías de relación social (**asortatividad**) son trasladables a los medios sociales:
- **Influencia:** personas relacionadas tienden a tener intereses similares. **Homofilia:** personas con intereses similares tienden a relacionarse
- En Twitter se modelan con cuatro relaciones: *co-post*, *co-following*, *co-followed*, *following*
- Se usan medidas como el grado de entrada, el número de menciones o el de retweets para medir **influencia de usuarios**
- En redes sociales on-line, se pueden usar para **recomendar amigos/contactos** (listas ordenadas de pares de nodos proclives a conectarse)
- Son también muy útiles en sistemas de recomendaciones colaborativos









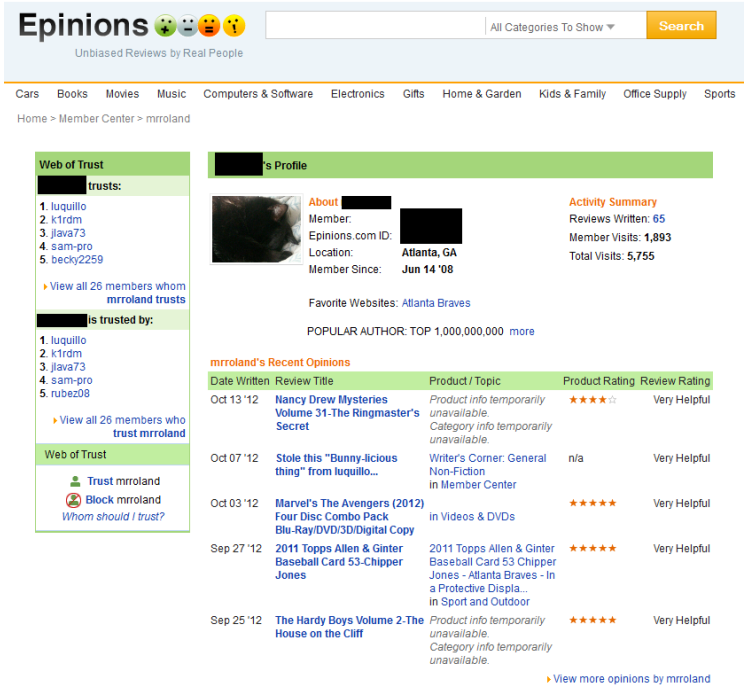


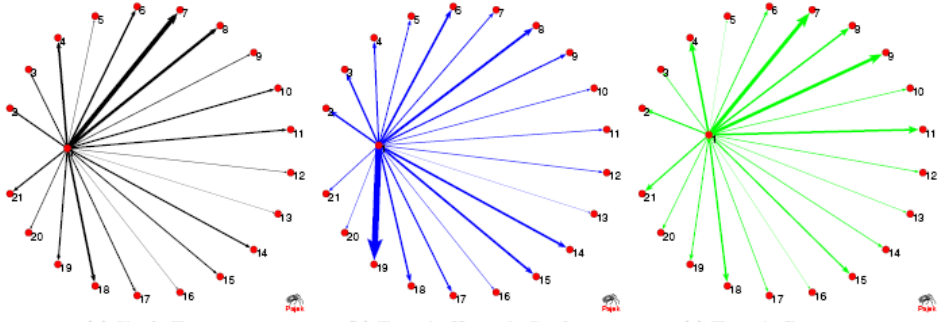
APLICACIONES (4): Determinación de Confianza en Usuarios

- El papel de la confianza es crítico en algunas comunidades on-line tales como los **sitios web de revisión de productos** y **de comercio electrónico**

- Los usuarios de sitios de este tipo (ej. el antiguo ***Epinions***) proporcionan tanto revisiones de productos como **redes de confianza**

- Estas redes se explotan para predecir la calidad de la revisión y para mejorar la precisión de los sistemas de recomendaciones

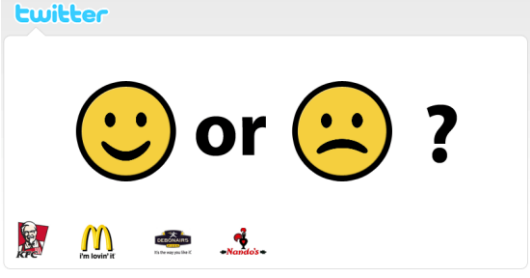


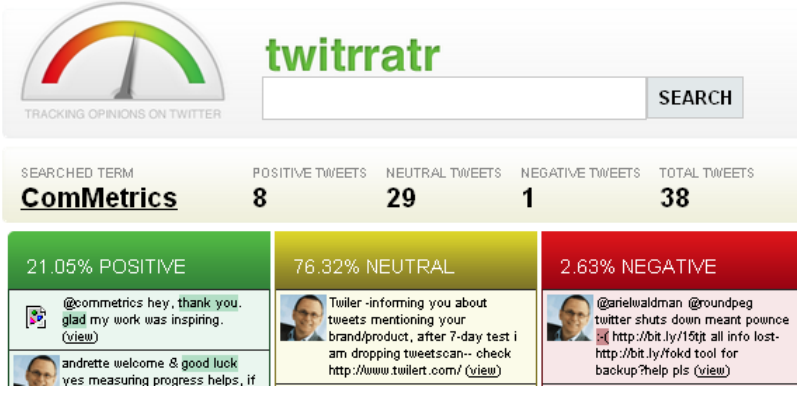


(a) Single Trust (b) Trust in Home & Garden (c) Trust in Restaurants

APLICACIONES (5): Análisis de Sentimientos/Minería de Opiniones

- El objetivo del análisis de sentimientos es clasificar un fragmento de texto con una **valoración emocional positiva o negativa**
- La agregación de los sentimientos puede dar una idea de la percepción de la gente sobre una empresa, un producto o un tema
- Es un área de aplicación de la **minería de textos** y la **clasificación automática**. Se puede considerar también la influencia de las relaciones sociales
- Se realiza sobre tweets, entradas en un blog, opiniones en un portal, informes de productos en una web, etc.





APLICACIONES (6): Marketing Viral

- Los usuarios de foros on-line promocionan las marcas, un 79.2% ayuda a amigos en sus decisiones de compra (frente a un 47.6% de no usuarios). Un 65% aconseja sobre lo leído on-line (frente a un 35%)

  http://www.socialmediaexaminer.com/new-studies-show-value-of-social-media

postrelease

In thinking about the following activities, please indicate which, if any, you're involved in.

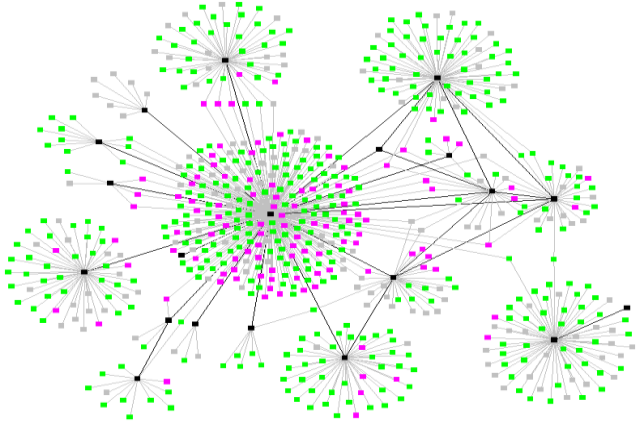
Responses (in order of frequency among overall population)	Those who contribute to online forums	Those who do NOT contribute to forums
Helping a friend or family member make a decision about a product purchase	79.2%	47.6%
Sharing advice (offline and in person) based on information that I've read online	65.0%	35.0%
Posting online ratings/reviews of products/services	66.0%	16.8%
Proactively recommending that someone make a particular purchase	57.7%	16.9%
Sharing links to articles about new products or with reviews of products	43.6%	12.0%
Attending an offline event or meet up where people with similar interests or who share the same hobby connect	35.6%	13.8%
Publishing a blog	20.6%	2.1%
Taking an active role in organizing an offline event or meet up for a group that met originally online	18.8%	2.4%
None of the above	0.0%	39.4%

Source: PostRelease.com/Synovate

- El contagio social permite modelar procesos de publicidad no convencional para campañas de marketing: **boca a boca**

negro: líderes de opinión; rojo: influenciados
verde: no influenciados; gris: indecisos

- Analizando la influencia de los usuarios de la red se puede determinar las ***semillas*** a las que es necesario recompensar (cupones, promociones, etc.) para que hablen bien de un producto



- Este proceso se hace para provocar una **cascada** en la red con el menor presupuesto de campaña posible

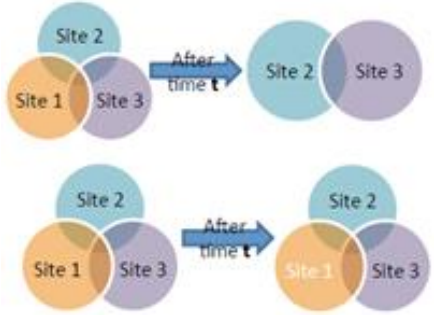
APLICACIONES (7): Predicción de Migración de Usuarios (*Churn*)

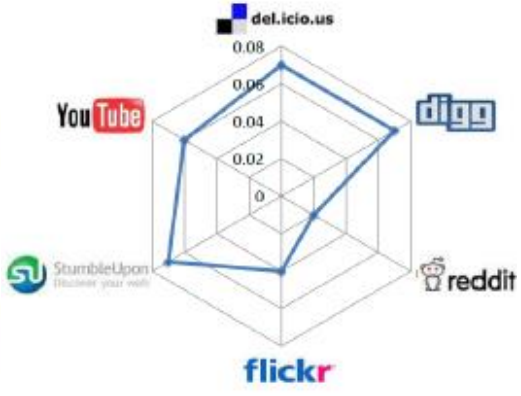
- Toda empresa (ej. Telecom.) sufre de la **migración voluntaria de consumidores** a la competencia

- La prevención de este abandono es un factor clave para su supervivencia. Se aplican técnicas de **clasificación automática**

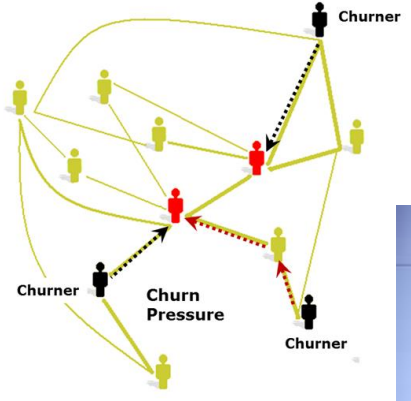
- Las relaciones sociales son muy influyentes. Considerar la **estructura de red social/comunidades** y la **influencia** de los usuarios mejora la detección

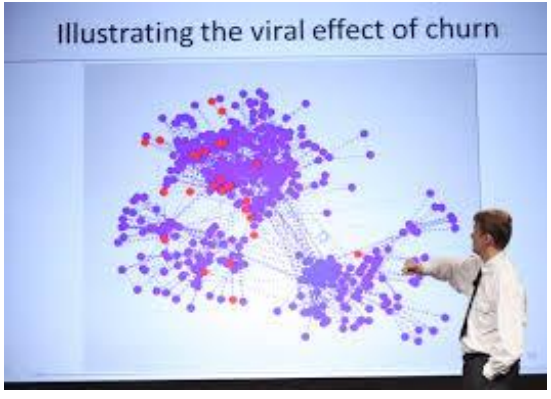
- Se modela el *churn* como **procesos de contagio complejo en redes sociales**





(f) Twitter





MINERÍA DE MEDIOS SOCIALES (7): Problemática

1. Paradoja del Big Data:
   - Los datos de medios sociales son “Big Data” pero no están distribuidos uniformemente
   - Es habitual que cuando nos centramos en un individuo concreto nos falten datos y haya que obtenerlos por agregación

2. Obtención de muestras representativas:
   - La información se obtiene de de las APIs de las aplicaciones. ¿Cómo podemos saber si nuestra muestra de datos es representativa del conjunto completo?

3. Falacia de la eliminación de ruido:
   - El preprocesamiento habitual en minería de datos no es adecuado: eliminar muchos datos ruidosos provoca la pérdida de información útil
   - La eliminación del ruido es compleja y dependiente del problema

4. Dilema de la evaluación:
   - Evaluar los resultados es complejo al no tener un ground truth en muchos casos

Referencias y Agradecimientos

Para diseñar los materiales de este tema, he hecho uso de material desarrollado por expertos en el área disponible en Internet:

- P. Gundecha, H. Liu. “Mining Social Media: A Brief Introduction”.
  Tutorial INFORMS 2012: http://www.public.asu.edu/~pgundech/book_chapter/smmslides.pdf
- H. Liu. “Some Computational Challenges in Mining Social Media”.
  Tutorial ASONAM 2013: http://www.public.asu.edu/~huanliu/papers/ASONAM13.pdf

- K. Lerman. “Social Media. A Responsible User’s Guide”
  University of South California: http://www.isi.edu/integration/people/lerman/talks.html

- J. Leskovec. “Social Media Analytics”. Tutorial ACM SIGKDD 2011
  Stanford University: http://snap.stanford.edu/proj/socmedia-kdd