1. MÉTODOS DE INVESTIGACIÓN EN LA RED

1. TRANSFORMACIONES SOCIALES RESULTADO DE LA RED

Internet y las nuevas tecnologías de comunicación en su conjunto, junto con otros procesos en marcha, han generado una inmensa transformación social. Dicha transformación se ha dado en múltiples facetas de la sociedad. Por poner algunos ejemplos, en aspectos económicos favoreciendo transacciones financieras y comerciales en tiempo real en extremos opuestos del globo (Ocampo y Stiglitz, 2008), a nivel de la sociedad civil promoviendo una mayor transnacionalización de algunos movimientos sociales (Tarrow y Della Porta, 2005) y al reducir los costes de coordinación en el interior de las organizaciones en especial organizaciones dispersas geográficamente (Benkler, 2006; Bimber, Stohl, y Flanagin, 2008; Skocpol, 2004), a nivel electoral en la forma como los candidatos políticos movilizan sus bases y obtienen financiación para sus campañas (Castells, 2009; Gibson, 2009), en términos de movilización ciudadana al lograr sortear la posible censura al permitir la difusión desde múltiples fuentes de mensajes a miles de posibles simpatizantes y con un coste marginal cercano a cero (Donk, 2004; Earl, 2010; Passy, 2003); entre otros múltiples fenómenos sociales.

Internet, junto con el creciente uso de ordenadores, parece fortalecer la emergencia de la denominada aldea global (McLuhan, 1996), un mundo cada vez más interconectado, lo que favorece aspectos tales como una difusión casi en tiempo real de la información sobre desastres naturales, crisis económicas o escándalos políticos. Desde una perspectiva positiva, el alto nivel de interconexión parece haber favorecido un sentimiento de solidaridad o repudio más allá de las fronteras nacionales ante hechos que no llegaban a conocerse o pasaba mucho tiempo hasta que se supiesen. Es el caso de movilizaciones a nivel transfronterizo, como los movimientos en contra del G8 durante los noventa, el movimiento ecologista, y en particular contra el calentamiento global en el mismo periodo, y recientemente el movimiento de los "indignados" en el 2011, entre otros.

No obstante, desde una perspectiva negativa, el constante y creciente flujo de datos se traduce en un aumento de la incertidumbre en cuanto en pocos segundos, en el otro extremo del globo, miles de personas pueden verse afectadas por hechos aparentemente ajenos a su contexto. Es el caso de decisiones financieras automatizadas mediante algoritmos que, ante el menor síntoma de pérdida de rentabilidad o ante un nivel de riesgo mayor al programado, o simplemente siguiendo el comportamiento de otros agentes en el mercado, deciden de forma masiva retirar fondos, decisión con importantes y a veces nefastas consecuencias humanas. Un ejemplo es la compra o venta masiva de títulos de deuda de un país, lo que puede llevar a su quiebra, o que el precio de bienes esenciales ante la amenaza de una plaga o una guerra o por simple espe-

culación al usarse como biocombustible hace que inmediatamente su precio se dispare, lo que desencadena una hambruna entre los países más pobres que dependen del consumo de este producto, entre otras decisiones, donde el constante flujo y el proceso automatizado de análisis de datos tiene importantes repercusiones humanas (Bauman, 2011; Sánchez, 2013; Pardo, 2010).

Es pertinente recordar que los algoritmos antes mencionados no se programan solos, su diseño y programación depende de científicos y, en particular, de programadores que han sido sus creadores vendiendo su conocimiento al mejor postor. Sin perder de vista el factor ético que deben tener los científicos, en este contexto en el que es posible procesar masivas cantidades de datos, los científicos y científicas sociales a través de aproximaciones como el Big Data y análisis de datos en Internet, que expondremos a continuación, pueden tener un rol más protagónico en el tipo de sociedad que deseamos construir.

El presente documento busca establecer un panorama de las múltiples posibilidades que nos ofrece Internet para la investigación social en general y la investigación sobre y con la población joven en particular. Primero exponemos ciertas consideraciones a tener en cuenta dentro de la investigación social y las tecnologías de la comunicación e información TIC; para, en segundo lugar, exponer los debates teóricos más actuales sobre la investigación de fenómenos sociales en la Web, lo que se conoce como Big Data (Boyd y Crawford, 2012; Manovich, 2011; Zikopoulos y Eaton, 2011) y el debate entre los métodos virtuales (Hine, 2000) y los denominados métodos digitales (Rogers, 2013; Marres y Rogers, 2008). En esta parte se expone un resumen del estado del arte sobre estos temas. Posteriormente se realiza un análisis crítico de estas aproximaciones, en el sentido de resaltar no sólo sus ventajas, sino también los límites y desafíos a los que se debe hacer frente. Posteriormente hay una sección dedicada al estudio de la Red como el nuevo espacio de discusión en el cual coexisten e interaccionan múltiples canales en los que cada uno presenta características propias. Para finalizar se presenta una reflexión de lo que se conoce como investigación de minería de datos en la Red, donde mencionamos las técnicas más comunes para intentar comprender este espacio multicanal. En particular, cómo estudiar la información que se publica en estos canales por parte de la población y qué tipo de información y análisis de distintos fenómenos sociales podemos realizar, sin olvidar la importancia de los aspectos metodológicos y los éticos, de gran relevancia y que nunca han de perderse de vista. Por último se realiza una reflexión del rol del investigador/a social en relación a la investigación a través de la Red y en la Red.

2. LAS TECNOLOGÍAS DE LA INFORMACIÓN Y SUS CONSECUENCIAS EN LAS CIENCIAS SOCIALES

Las tecnologías de comunicación e información (TIC) pueden ser vistas tanto como un campo de investigación —la investigación sobre las TIC y la sociedad— como un canal para llevar a cabo la investigación —las TIC como artefactos metodológicos—. La investigación siempre ha dependido de la matriz de los medios de comunicación dominantes (Johns, Chen, y Hall, 2004), y las TIC representan la última etapa de este desarrollo. Aunque las TIC evolucionan rápidamente y aún están en una fase embrionaria, algunas tendencias principales pueden ser identificadas.

Hine (2000) destaca dos sentimientos que acompañan la adopción de las TIC para la investigación: el entusiasmo y la ansiedad. El entusiasmo ante el potencial innovador. Pero Hine también es consciente de que este aspecto innovador representa una fuente de ansiedad. La innovación implica necesariamente romper viejos, confiables y establecidos modos de investigación, dejando un campo de parámetros experimentales y métodos no probados. La ansiedad en la investigación de las TIC con mayor frecuencia surge de la idea de que "nada puede darse por sentado" sobre todo porque la "netiqueta", en general en las TIC y la ética de la investigación en línea —como una nueva forma de interacción social, tanto para los investigadores como para los investigados— parece ser un tema primordial ante nuevas preguntas y retos en los que la ansiedad juega un papel (Hine, 2000).

Como Rutter y Smith exponen en *Ethnographic Presence in a Nebulous Setting*, "la definición del marco de la investigación no se convierte en un punto de partida, sino en una cuestión de investigación primaria que requiere un examen cuidadoso y continuo por el investigador a lo largo del trabajo de campo." (Rutter y Smith, 2005: 85).

Cuatro diferentes tipos de relaciones entre los objetivos de la investigación y las TIC

Como paso previo, en esta sección vamos a reflexionar sobre la relación entre las TIC y la investigación. Distinguimos cuatro tipos de relación. Como se mencionó anteriormente, el uso de las TIC para la investigación y la investigación sobre las implicaciones socio-político-culturales de las TIC son dos cosas diferentes. Es decir, el método de la investigación y el objeto no necesariamente tienen que ir de la mano. Sin embargo, es común para la investigación sobre las TIC y la sociedad desarrollar algún tipo de etnografía en línea, y el trabajo con frecuencia adopta otros métodos de TIC. El uso de las TIC para la investigación de las implicaciones de las TIC en la sociedad o en el uso de las TIC por los distintos actores sociales es el primer tipo de relación entre las TIC y la investigación. Este es por ejemplo el caso de la investigación de cómo los movimientos sociales utilizan Facebook mediante la realización de etnografías virtuales en Facebook.

El segundo tipo es lo opuesto al primero, es decir, el uso de métodos de TIC para investigar cuestiones ajenas a las preguntas de TIC. Sin embargo, el uso de métodos de TIC para la investigación de cuestiones que no están vinculadas a las TIC es menos frecuente y es relativamente nuevo (Rogers, 2009). Por ejemplo, cuando el objetivo de la investigación es mediante el análisis de datos en la Web, identificar el perfil sociológico de un participante en una manifestación. En este caso, el uso de métodos de TIC sigue siendo relativamente poco frecuente. Quizás por desconocimiento y desconfianza de los investigadores se continúa privilegiando métodos tradicionales de investigación tales como encuestas o grupos de discusión. Una de las razones principales de esta situación era las bajas tasas de penetración de ciertas TIC y en particular de Internet dentro de la población, lo que dificultaba la posible extrapolación de los resultados y cuestionaba la representatividad de la muestra.

Un tercer tipo es la investigación que utiliza las TIC como indicadores. Es por ejemplo la adopción de hipervínculos o hiperenlaces como indicadores de las conexiones entre las organizaciones (lo explicaremos con mayor detalle en el apartado de minería de datos). En este caso, la di-

mensión en línea de los actores no es el foco principal de la obra, pero se utiliza como un indicador o *proxi* de la existencia de una relación entre las organizaciones o agentes bajo estudio (Thelwall, 2009).

Por último, tal vez el caso más común, es que los investigadores utilizan las TIC para acercarse a un objeto de la investigación, como visitar el sitio web de un actor para obtener una primera impresión u obtener contactos, o para comunicarse con sus informantes con el fin de ponerse de acuerdo sobre el desarrollo de métodos de investigación "fuera de línea".

Antes de continuar, conviene aclarar que las categorías de línea y fuera de línea (o, de acuerdo a los términos utilizados en este trabajo, TIC para la investigación frente a la investigación independiente de las TIC) deben ser utilizados con precaución. En línea generalmente se refiere a cualquier interacción mediada por un ordenador, mientras que fuera de línea se considera como algo que no encaja en la definición de la línea e implica la interacción física. Sin embargo, hay una zona ambigua entre las dos. Una entrevista puede llevarse a cabo en un teléfono móvil, por ejemplo, y no está claro si esto se debe considerar en línea o fuera de línea. Además, es difícil encontrar situaciones puramente fuera de línea o en línea. En conclusión, consideramos que estas categorías (en línea y fuera de línea – *online* y *offline)* parten de una transición histórica en la adopción de TIC que bien puede perder sentido con relativa rapidez.

3. ENFOQUES DE LA INVESTIGACIÓN EN RED: EL TRÁNSITO DE LOS MÉTODOS VIRTUALES A LOS DIGITALES Y EL BIG DATA

En los siguientes apartados presentaremos los diversos enfoques que han ido apareciendo en el tiempo en torno al uso de la Red como método de investigación.

Big Data es un concepto que ha ganado mucha predominancia en los últimos años, que en un sentido amplio se utiliza como un paraguas para referirse a todo aquello que tiene que ver con el uso y la relación de grandes bases de datos y/o datos que provienen de la Red (que por su naturaleza también suelen ser masivos), cuya recogida y elaboración requieren de un trabajo computacional para recopilarlos y/o para "limpiarlos" o prepararlos para su ulterior análisis.

Más allá del concepto de Big Data, que es aplicable también a datos que no provienen de Internet, han emergido otros dos enfoques que ponen su acento en el uso de la Red como método de investigación.

Por una parte, y en un primer momento, aparecieron en los estudios ciberculturales los llamados métodos virtuales, que se caracterizan por la adaptación de los métodos "tradicionales" de investigación empírica a la Red (Hine, 2000). De tal manera, la etnografía pasaría a ser etnografía virtual, la entrevista a e-entrevista, las encuestas a e-encuestas. Este enfoque tiene presente cómo el uso del medio en red puede afectar o hacer variar la validez y el funcionamiento de los métodos tradicionales.

A este primer enfoque, le siguió otro —métodos digitales— (Rogers, 2009) en el que el acercamiento se caracteriza, no por adoptar los métodos al medio, sino por extraer los métodos del medio.

Métodos virtuales

Fueron los primeros en aparecer. Como indicábamos anteriormente, se refieren a la adaptación de métodos tradicionales al entorno virtual. Seguidamente expondremos los casos de la e-encuesta y de la e-entrevista, y el análisis codificado de organizaciones, ejemplificando algunas de sus aplicaciones.

La e-encuesta

La e-encuesta o encuestas en línea se refieren a encuestas o cuestionarios puestos a disposición a través de una plataforma o sitio web. Se invita a los potenciales encuestados a ir a la página web y rellenar el cuestionario (por ejemplo, enlazando una página en una invitación por *e-mail,* o directamente el enlace a la encuesta es puesto en la página web de la organización que la realiza). Un ejemplo es la encuesta a participantes en línea llevada a cabo en el Foro Social Europeo (FSE) (celebrado en Atenas en octubre de 2006) por el grupo de investigación política de la Universidad de Antwerp. Esta experiencia señala las ventajas y limitaciones de la utilización de encuestas en línea.

Para difundir la encuesta entre la población objeto de estudio (activistas en Europa), durante el FSE de Atenas se distribuyeron versiones impresas del cuestionario en línea (600), pero sólo 68 fueron recibidos inmediatamente. Con el fin de involucrar a más personas, a través de la página web se entregaron folletos en el FSE con mensajes cortos y la dirección URL del cuestionario en línea. También se envió un *e-mail* a las listas de correo del FSE para invitar a la gente a participar (a unos 700 abonados) y se envió un correo electrónico a cerca de 1.500 direcciones (de los y las participantes del FSE que se inscribieron a través de la página web del FSE). El resultado fue un archivo de datos interesante que contiene 510 encuestados[1].

> «Una ventaja del uso de la e-encuesta en línea es que las 510 respuestas ya están insertadas en un formato de base de datos y los investigadores no tienen que insertar las respuestas correspondientes a cada pregunta. Los costos no son muy altos tampoco porque la comunicación por Internet no es muy costosa. En contrapartida, el porcentaje de respuesta es bajo, pues el número de activistas alcanzados es relativamente bajo teniendo en cuenta el número de activistas contactado. Otros de los problemas identificados durante esta experiencia fue la pérdida de control en la representatividad de la e-encuesta, en este caso, las respuestas tienen un problema de sobre-representación de algunos colectivos (en este caso los que recibieron directamente la invitación por correo electrónico) y el posible sesgo de contacto con los activistas más conectados en línea. Otro problema adicional fue la recurrencia de cuestionarios incompletos, casi la mitad de los cuestionarios (200) estabán incompletos, posiblemente debido a que el cuestionario requería demasiado tiempo respecto a la percepción de tiempo en línea que invita a una mayor inmediatez.»

1. Más información del grupo en http://www.m2p.be

Otras experiencias de e-encuesta similares las encontramos en "15M: Retrato de un clima" de Juan Linares, Óscar Marín, Yolanda Quintana y Ariadna Fernández y en encuesta "occupy" que se llevó a cabo entre los activistas del movimiento occupy por el colectivo Occupy Research (del que también dan cuenta Pablo Rey y Alfonso Sánchez).

Entrevistas online

Las entrevistas telefónicas se han utilizado tradicionalmente en las ocasiones en que, por diversas razones, no pudiera llevarse a cabo las entrevistas fuera de línea. Hoy en día, el sustituto de la entrevista física también puede incluir el correo electrónico, la videoconferencia o la entrevista vía *chat*. En otros casos, las entrevistas en línea no son un sustituto, sino la primera opción para los investigadores y las investigadoras por una variedad de razones.

El formato en que se desarrolla una entrevista influye en el contenido extraído de la comunicación. Por ejemplo, las entrevistas por correo electrónico pueden ser diferentes en estilo, dimensiones temporales y el sentido de la intimidad y la confianza establecida en el proceso en comparación con una entrevista cara a cara (Kivits, 2005 en Hine, 2000). Pero aunque las peculiaridades de la comunicación textual se tienen que tener en cuenta, la comunicación textual de una entrevista por *e-mail* puede ser tan rica como una entrevista cara a cara.

Sin embargo, la obtención de entrevistas en línea no siempre es fácil. La tasa de respuesta a las entrevistas por correo electrónico tiene una gran variación dependiendo de colectivo al que se dirija y la confianza que el investigador tenga con éste. Siguiendo las reflexiones de Van Laer anteriores (J. van Laer, entrevista por correo electrónico, 25 de marzo de 2007), depende del perfil que el investigador desea alcanzar. En la investigación de fenómenos sociales altamente relacionados con Internet (como el movimiento por el *software* libre o el movimiento de los derechos de comunicación) la comunicación a través de *e-mail* es muy frecuente, incluso cuando los activistas se encuentran en el mismo espacio físico.

El uso frecuente de la comunicación por correo electrónico entre los activistas de estos movimientos podría explicar por qué las respuestas a los métodos en línea entre este tipo de activistas son más positivas. Pero el alto uso de las TIC entre la población objeto de estudio no es garantía de éxito. Los investigadores de las comunidades en línea, una población familiarizada con la comunicación mediada por ordenador, mencionan que solicitar entrevistas en línea para involucrar a los participantes en las comunidades en línea generalmente da como resultado tasas de respuesta pobres (Reagle, 2005). Fuster Morell (2011), apunta que en su investigación de comunidades en linea que el procedimiento más eficaz para asegurar entrevistas en línea fue asistir a reuniones fuera de línea. En su opinión, las mayores tasas de respuesta de los informantes en las reuniones fuera de línea se relacionan principalmente con ganarse la confianza y atraer su atención.

Análisis web: el análisis estadístico de las características de los sitios web de organizaciones

Este enfoque se basa en los análisis estadísticos comparativos de las características de los sitios web de las organizaciones juveniles. Este enfoque parte de la literatura sobre calidad democrática

(Berg-Schlosser, 2004; Bollen, 1990; Bollen y Paxton, 2000; Diamond y Morlino, 2004; Morlino, 2004; Munck y Verkuile,n 2002). La investigación empírica utilizando este enfoque fue desarrollado por primera vez para el análisis de las web de los partidos políticos (Davis, 1999; De Landtsheer, Krasnoboka y Neuner, 2001; Gibson, Nixon y Ward, 2003; Norris, 2003; Römmele, 2003; Trechsel *et al,* 2003). A continuación, se pasó a examinar los actores políticos no convencionales, tales como las Organizaciones No Gubernamentales (ONG) (Vedres, Bruszt y Stark, 2005a, 2005b), las organizaciones de los movimientos sociales (Della Porta y Mosca, 2006, 2009; Sudulich, 2006; Van Aelst y Walgrave, 2004) y los *blogs* políticos (Navarria, 2007). Un aspecto común de estas investigaciones es que los investigadores no "tratan de deducir los efectos sociales de las propiedades de las tecnologías" (Vedres, Bruszt y Stark, 2005b).

De acuerdo con este cuerpo de literatura, los actores sociales modelan su uso de Internet en función de sus propios estilos, estrategias organizativas y lógicas (Vedres, Bruszt y Stark, 2005b). En *Buscando por la Red: las cualidades democráticas de Internet,* Della Porta y Mosca (2006, 2009) analizan estadísticamente los sitios web de los movimientos sociales, considerando dimensiones tales como el suministro de información, la construcción de la identidad, la rendición de cuentas externas, la movilización y la reducción de la desigualdades (brecha digital). Los distintos estilos de sitios web reflejan diferentes modelos de democracia (y de comunicación democrática) presentes en las organizaciones de los movimientos sociales (Della Porta y Mosca, 2006). El punto importante de esta investigación es que no todas las dimensiones están correlacionadas: esto confirma que las organizaciones optan por la maximización de algunas, pero no todas, las dimensiones del potencial democrático en la Red.

Para la obtención de datos, los investigadores Della Porta y Mosca (2006, 2009) visitaron 261 webs vinculadas a organizaciones de movimiento social, verificando la posible presencia o ausencia de las listas de elementos que se consideran indicadores de la calidad democrática. Durante la visita a los sitios web, los investigadores también informaron sobre aspectos particulares relacionados con los sitios. Este trabajo pone de manifiesto las dificultades de la codificación de una gran variedad de sitios web, señalando la importancia de tener una noción clara de los indicadores, aún así dejando espacio para los comentarios sobre cuestiones no consideradas inicialmente, y la necesidad de la verificación de la codificación cuidadosamente antes de comenzar la recolección de datos.

Ventajas y retos del uso de métodos virtuales

Seguidamente pasaremos a exponer algunas de las ventajas y retos de los métodos virtuales. Algunos son particularmente referidos a los mismos, mientras que otros, como veremos más adelante, también están presentes en los enfoques de Big Data y los métodos digitales.

Entre las ventajas del uso de los métodos virtuales se encuentra: 1) **tener una variedad más amplia de opciones**, aunque la mayoría de la comunicación de las TIC es textual y asíncrona, tenemos un margen más amplio, teniendo en cuenta todo tipo de comunicación —tanto textuales como visuales, síncronos y asíncronos—. 2) Se presenta una **reducción de los costes potenciales,** no hay necesidad de colocar a las personas en un lugar físico. Las TIC son generalmente menos costosas

que otras comunicaciones tecnológicas (como el teléfono) y encuentros físicos. 3) **Agilizan la recombinación de los datos.** La digitalización de los datos y el acceso a máquinas de gran alcance permite que los datos se vuelvan a combinar de varias maneras. 4) Por último, se trata de **investigación sin fronteras.** No hay limitaciones geográficas o de distancia. Por ejemplo, si se están utilizando métodos en línea para analizar la comunidad de Christiania - Copenhague está tan cerca como el centro social Can Mas Deu en Barcelona. 5) Por último los métodos virtuales, aunque dependen de las TIC, requieren un conocimiento técnico menor que los enfoques de Big Data y métodos digitales. Requieren adaptar los métodos tradicionales al entorno virtual, pero un científico social que conozca los métodos tradicionales requerirá una formación específica mucho menor que con Big Data o métodos digitales, como veremos seguidamente.

Pero la comunicación mediada por ordenadores también plantea otros retos importantes:

1. Un primer reto se refiere al establecimiento de la **confianza y la identificación de los actores a través de las TIC.** En un medio en el que el investigador suele permanecer distante y sin rostro, en comparación con el cara a cara de la comunicación física, y donde la mayoría de la comunicación no verbal se pierde, constituye un gran desafío. Para la transparencia de investigación y el fomento de la confianza una posibilidad relativamente común es la construcción de un espacio de referencia en línea, tal como un sitio web en el que se presenta previamente la investigación. Otra opción es llegar a ser útil para el sujeto de investigación, por ejemplo mediante el intercambio de datos sistematizados.

   La comunicación en línea también es difícil en a cuanto la identificación del informante para el investigador, ya que es posible que el investigador sólo tenga un nombre o alias y una dirección de correo electrónico o una URL como puntos de referencia de su informante.

2. Otro reto es la **"trampa del tecno-entusiasmo",** que se refiere a la necesidad de adaptar la tecnología a la meta de la investigación. Una herramienta de investigación de las TIC podría funcionar técnicamente; sin embargo, esto no garantiza que se adapte a la audiencia esperada. Por lo tanto, es necesario adaptar la tecnología del método al comportamiento de las TIC. Un aspecto a considerar que podría afectar el uso y la respuesta al método en línea se refleja en el enfoque del informante a las TIC. Otro aspecto a considerar es el nivel de educación tecnológica de la población a la que el investigador se dirige, y también para reproducir la personalidad virtual de la persona. Hay gente muy rápida en la interacción a través de las TIC, y hay gente a la que le toma más tiempo.

3. Otro reto se refiere a la **representatividad de la muestra,** las nuevas fuentes de sesgo y la brecha digital. La participación en un método disponible en línea (como una encuesta) podría limitarse a determinadas personas o podría ser abierto a cualquier persona. En la segunda opción, la política de difusión es muy importante, porque guiará a las personas que llegan al sitio. También hay que tener presente el sesgo que pueda crear la brecha digital. El riesgo es perder el control sobre la representatividad de la muestra.

4. Otro problema que se puede plantear, no necesariamente menor, está relacionado con **posibles sorpresas con los tiempos de la investigación.** Cuando el investigador depende de la interactividad de los informantes, los métodos de TIC no necesariamente reducen el tiempo requerido para la recogida de datos. Especialmente cuando se utilizan métodos asincrónicos,

una gran cantidad de tiempo se dedica a ser el *"time-keeper"*, es decir, tener que recordar a los informantes que envíen la información. Por ejemplo, si le pedimos una entrevista por correo electrónico, un elemento digno a tener en cuenta es que, como podría ocurrir con el investigador, el informante puede tener el correo colapsado de correos electrónicos. Incluso puede ocurrir que la persona piense que la petición es *spam*. Una posible manera de "bloquear el tiempo" para los métodos en línea que requieren la interacción del informante, como cuestionarios o entrevistas en línea de correo electrónico, es la fijación de una "cita" cuando el investigador también está presente a través de Skype o de otro sistema de mensajería instantánea. De esta manera, el investigador puede proporcionar apoyo al informante mientras él o ella responde a las preguntas si tuviera alguna dura.

5. Otro potencial problema es la **sobrecarga de datos y la selección de los datos relevantes.** Podemos reconocer que con los métodos de las TIC el investigador es capaz de obtener grandes cantidades de notas de campo, porque una mayor cantidad de información está registrada y disponible. Sin embargo, esto plantea un problema muy conocido por los investigadores en línea, a saber, el de la sobrecarga de información. El investigador en línea se enfrenta a un problema en la selección de la información relevante. En este sentido, es importante tener un esquema claro y disciplinado de los datos concretos que se requieren.

El Big Data y los métodos digitales

En la actualidad, el creciente volumen de comunicaciones de distinto tipo que se viven en la Red, de la mano del *e-mail* y la Web o más recientemente del auge de las denominadas redes sociales tales como Facebook, YouTube, Twitter, WhastApp entre otros, hacen de la Red un inmenso e inagotable laboratorio para observar diferentes dinámicas sociales, dinámicas que anteriormente eran muy difíciles de estudiar con un amplio soporte empírico.

Es el caso de los estudios sobre la difusión de la información a lo largo del tiempo entre colectivos sociales: por ejemplo en campañas políticas o de concienciación social (Bakshy *et al.,* 2012; Bruns, 2012; Earl, 2010) o en la transmisión de la innovación, de ideas y de rumores (Rogers 2010; Weng *et al.,* 2012) o la identificación de masas críticas para lograr la acción colectiva (Xie *et al.,* 2011). Para este tipo de estudios, tradicionalmente, las únicas formas de recrear la red de actores era entrevistando a líderes activistas o encuestando a personas que hubiesen participado en la movilización o mediante análisis históricos de contenido en prensa como la investigación pionera de McAdam (1983). Estos métodos de obtención de información presentan diferentes sesgos y pérdida de información. Años antes del desarrollo de las redes sociales y cuando la Web no presentaba las tasas de penetración de uso actuales[2], autores especializados en estudios de protesta como Oliver y Myers (1998) ya señalaban las dificultades metodológicas de acceder a datos precisos, y en particular datos que tuviesen información del tipo de relación entre los individuos que se movilizaban, la dirección que presentaba el flujo de información y en qué momento exacto se había presentado ese punto de quiebre o masa crítica necesaria para favorecer un amplio proceso de difusión (Cristancho y Salcedo, 2013).

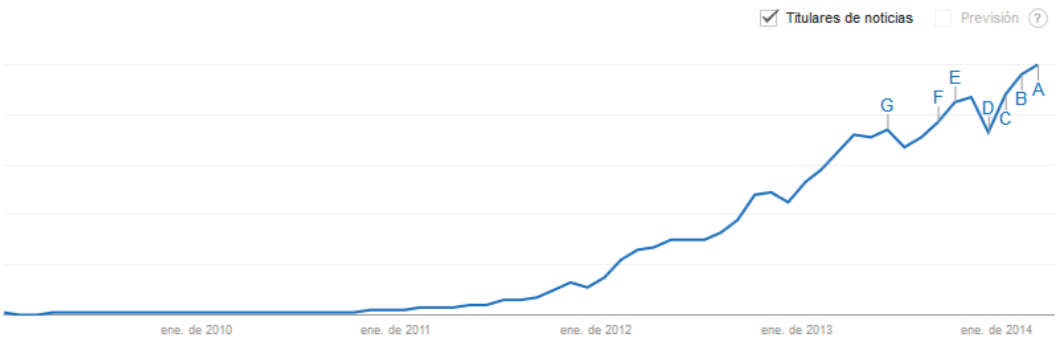
2. Fundamentalmente en países de la OCDE, la tasa de penetración de Internet supera el 50% de la población (Internet World Stats 2011).

En la presente sección presentamos dos aproximaciones metodológicas centrales de la investigación de fenómenos sociales con base en el análisis de datos en la Web. La primera aproximación es lo que se conoce como Big Data y la segunda es la denominada investigación a través de los llamados métodos digitales.

Big Data

En primera estancia, consideramos importante señalar que el concepto de Big Data es un concepto relativamente nuevo, si nos guiamos usando como indicador el volumen de entradas en noticias sobre el término, en la Figura 1 podremos observar que el término comienza a aparecer a partir del 2009 y presenta su máxima visibilidad en los últimos dos años.

FIGURA 1
INTERÉS A LO LARGO DEL TIEMPO DEL TÉRMINO BIG DATA EN NOTICIAS

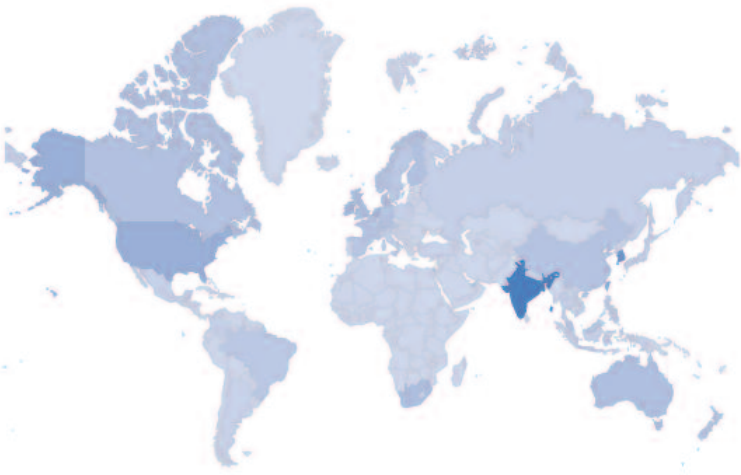


Fuente: Google Trends. Datos normalizados, búsqueda realizada 16/06/2013

El concepto de Big Data está asociado con el surgimiento de las organizaciones más emblemáticas en la actualidad en la Red (Google, Yahoo, Facebook, Wikipedia, Twitter…) y con la tecnología, tanto *hardware* y en especial *software* que las ha hecho posible. Tecnologías que permiten la indexación de millones de registros estilo Nutch en el 2003, o la computación distribuida o en redes que en sus orígenes permitió realizar proyectos como el SETI para mapear señales de radio-frecuencia provenientes del universo en la búsqueda de vida inteligente, o proyectos en la búsqueda del genoma humano, sólo por citar algunos de los más conocidos, que utilizan el poder computacional que tiene la Red al poder interconectar millones de ordenadores y servidores, creando de esta manera el superordenador más grande, por lo menos hasta la fecha.

En términos geográficos, el interés acerca del Big Data se concentra especialmente en Estados Unidos, India, Reino Unido y Francia. En la Figura 2 una mayor intensidad del color en determinado país refleja una mayor frecuencia de apariciones de noticias sobre Big Data. En España, comparativamente, el interés en el tema sigue siendo bajo.

FIGURA 2
INTERÉS POR REGIÓN EN EL TÉRMINO DE BIG DATA



Fuente: Google Trends. Datos normalizados, búsqueda realizada 16/06/2013.

El término Big Data tiende a asociarse principalmente con un gran **volumen** de datos, aspecto que es importante, pero no es el único. Primero, porque depende de la tecnología disponible; lo que hoy consideramos como Big Data, en un futuro cercano, ante la velocidad de los avances tecnológicos (Ley de Moore[3]) puede considerarse como un tamaño mediano o incluso pequeño de datos. Segundo, además del volumen debemos considerar la **velocidad** de producción de los datos en el mundo actual y el carácter **relacional** de los mismos que nos permite identificar y analizar diversidad de fenómenos sociales.

En relación a la velocidad debemos considerar el volumen de datos que minuto a minuto son producidos por diversos medios sociales y la velocidad en la que los podemos procesar. Es tal el volumen de datos y la producción de los mismos que parece superar la velocidad de obtención, procesamiento y análisis de los mismos. En este sentido, sólo llegamos a observar una parte de todos los datos que se producen. También en términos de poder interpretar y comprender los datos, no nos interesa un mapa cuyo tamaño sea igual al territorio; es necesario identificar dentro del mar de datos qué muestra nos permite hacernos una idea del conjunto.

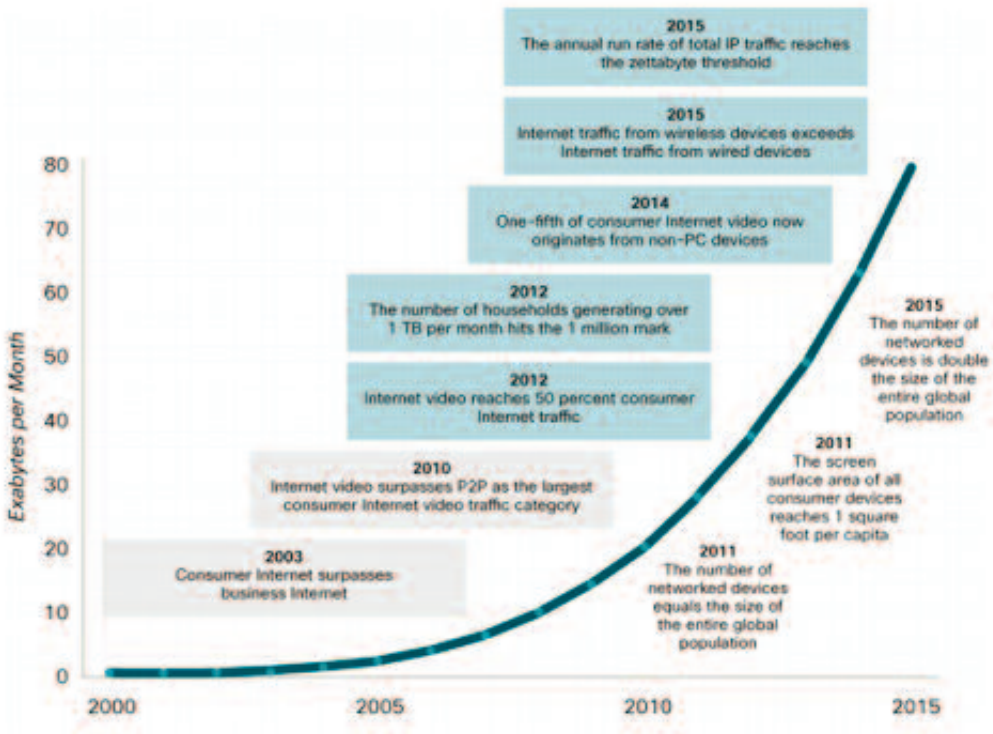
Para hacernos una idea de las dimensiones de velocidad y volumen de datos que se producen, para el año 2007, declaraciones de Google mencionaban el procesar 20 petabytes de información diaria (Dean y Ghemawat, 2008). Para hacerse una idea de los volúmenes de información

3. Cada año y medio el tamaño de los chips se reduce a la mitad y se duplica su capacidad de procesamiento.

de los que estamos hablando, una tesis de 400 páginas en formato .doc con gráficos aproximadamente tiene unos 10 megas, 1024 megas es lo que se conoce como un gigabyte (hoy un *laptop* promedio tiene 100 gigas de memoria), 1024 gigabytes es lo que se conoce como un terabyte (10 exp 12 bytes), lo cual es espacio suficiente para almacenar aproximadamente 800 películas de dos horas de duración con la calidad de un DVD convencional. Empresas como Twitter generan al día más de 7 terabytes de información, Facebook más de 10 terabytes. En la actualidad hablamos de peta y exabytes. Mientras un terabyte son 10^{12} bytes, un petabyte son 10^{15} bytes; hoy ya se habla de cientos de exabytes, que son 10^{18} bytes, para calcular las dimensiones de la Web. En la actualidad, fuentes como IBM Research declaran que el 90% de la información que circula en la Web ha sido creada en los últimos dos años (Zikopoulos y Eaton, 2011). Masivas cantidades de datos se producen segundo a segundo alrededor del mundo.

Otra característica del Big Data es el carácter relacional de los datos y la posibilidad de cruzar los mismos para su interpretación. Por ejemplo cruzar bases de datos del Registro Civil con perfiles de cuenta de Twitter, asociados a sus *post* y preferencias en Facebook, ofrece una gran variedad de posibilidad de identificar perfiles personales, con un gran nivel de detalle en términos de las preferencias, gustos e intereses que puedan tener estos individuos.

FIGURA 3
NIVEL DE PRODUCCIÓN Y TRÁFICO DE DATOS, PROYECCIÓN 2015

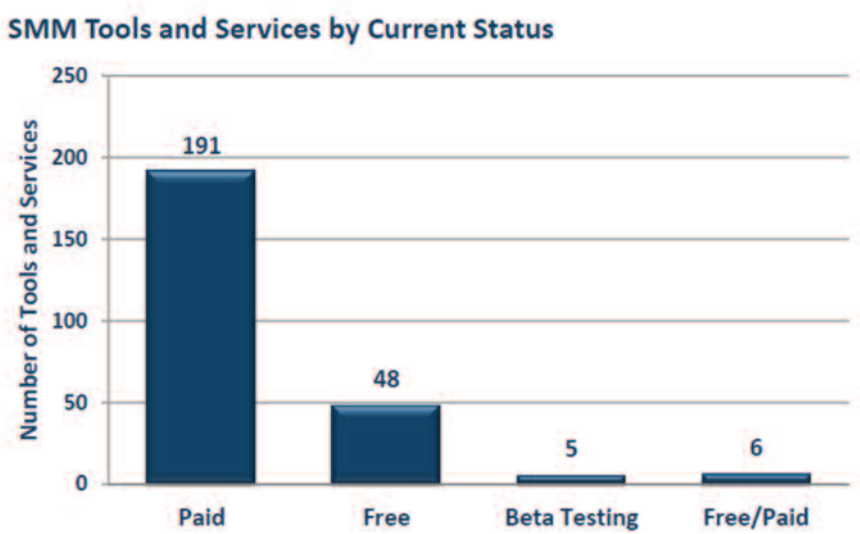


Fuente: http://blogcmt.com/wp-content/uploads/2011/06/exabyte1.png, CISCO VNI 2011 (20/04/2012)

Además de las características mencionadas (volumen, velocidad, carácter relacional de los datos), existe una amplia **diversidad de fuentes** para obtener datos tanto de individuos como de organizaciones. En ciertas ocasiones se asume que Big Data se refiere sólo a datos obtenidos a través de redes sociales. Esta es sólo una de estas fuentes, pueden relacionarse datos de fuentes censales, historias clínicas de hospitales, bases de datos de repositorios públicos, de centros de opinión, departamentos nacionales de estadística e información de empresas de diferentes tipos (desde agencias de viaje que solicitan nuestros datos cuando compramos un billete, hasta portales de venta de artículos como Amazon o E-bay que pueden a partir de las compras que realizan sus usuarios, personalizar el tipo de publicidad que les puede interesar, o empresas como Google que mediante el registro de búsquedas también personalizan la publicidad que envían al usuario de sus servicios). En este sentido, las fuentes pueden ser tan diversas como el tipo de fenómenos que estamos estudiando.

De acuerdo al volumen de datos que estamos manejando han sido creadas múltiples **herramientas** que facilitan el almacenamiento y procesamiento de datos distribuidos, lo que implica menos tiempo para obtener el resultado de los cómputos que se estén realizando en comparación al mismo proceso en un ordenador o servidor tradicional. Sólo por citar algunos, es el caso de servicios como Google Big Query, IBM infosphere, Microsoft Server y Power Pivot entre otros. El inconveniente o posible límite es que muchos de estos servicios o herramientas son de pago. Por ejemplo, en la Figura 4 y de acuerdo al estudio de la consultoría Herramientas de Seguimiento de Social Media podemos observar la proporción entre herramientas, de pago, gratuitas y versiones de prueba para análisis de Big Data.

FIGURA 4
NÚMERO DE HERRAMIENTAS PARA ANÁLISIS DE DATOS DE MEDIOS SOCIALES



Fuente: Social media monitoring tools 2012

Una alternativa a los servicios corporativos comparativamente minoritaria, y usualmente con una interfaz menos sencilla es recurrir al uso en red de lenguajes abiertos como Haddop, R o Phyton entre los más conocidos, algunos de ellos son las base de la arquitectura de las aplicaciones antes mencionadas. No obstante, nuestro poder de cómputo también estará en relación a la red de ordenadores de la que dispongamos y en particular de servidores. También existe la posibilidad de contratar servicios de procesamiento y reducción de datos en la nube tales como los que ofrecen empresas como Amazon o Google, entre las más conocidas.

Sin embargo, ante estas características, da la impresión que el Big Data está sólo al alcance de aquellos investigadores que dispongan de las infraestructuras, recursos y conocimientos para acceder a la tecnología que permita procesar grandes volúmenes de información. Sin desconocer esta situación, es posible hacer investigación de frontera desde nuestro ordenador personal *(desk research),* recurriendo a algoritmos desarrollados en el contexto de Big Data y haciendo uso de la información disponible en la Red de acuerdo a nuestros intereses de investigación. Este marco es lo que autores como Rogers han denominado "métodos digitales" (Rogers, 2013).

Métodos digitales

Cuando se hace referencia a métodos digitales, retomamos el concepto de *online groundedness* (Rogers, 2009) que implica que lo que sucede en la Red es reflejo de lo que sucede en otros espacios de la realidad. De acuerdo a autores como Bennet y Segerberg (2011) es anacrónico hablar de lo *online-offline* como si la realidad fuera esquizofrénica. Analizando lo que sucede en la Red, y más en sociedades con un alto nivel de penetración de Internet como es el caso español con tasas cercanas al 70% de la población y entre la población joven por encima del 90% (Fundación Telefónica, 2013; ONTSI, 2013), podemos explicar fenómenos sociales cuyos efectos van más allá de su manifestación en la Red. Un ejemplo tradicional dentro de la literatura sobre el tema es la posibilidad de identificar zonas posibles de contagio del virus de la gripe (con base al análisis de los resultados en motores de búsqueda (Ginsberg *et al.,* 2009)

Otro ejemplo es ajustar automáticamente la duración de los semáforos, o predecir rutas de desplazamiento para mejorar rutas de tráfico, de acuerdo a los datos de geolocalización (GPS) de los diferentes coches obtenidos a lo largo del tiempo y la información obtenida a través de cámaras distribuidas a lo ancho y largo de una ciudad, lo que permitiría optimizar los periodos de desplazamiento, con el correspondiente ahorro de combustible y el menor impacto ambiental. O predecir si una persona está enferma o tiene alguien cercano enfermo o se encuentra embarazada por cierto datos, lo que abre múltiples cuestiones éticas que comentaremos en otra sección.

Los métodos digitales van más allá de ser aplicaciones de métodos de investigaciones tradicionales pero aplicadas a través de la Web, como es el caso de los métodos virtuales (Hine, 2000). En los métodos digitales, las técnicas de obtención y análisis de datos surgen del entorno mismo de la Red. El análisis se concentra en aquellos objetos digitales que se crean en el uso por parte de la población de la Web y de sus múltiples canales: por ejemplo, al realizar análisis de búsquedas en Google, de los hipervínculos entre páginas web, de las palabras clave de búsqueda, o del intercambio de *mails* entre listas de usuarios; también a nivel de las redes sociales todas las

marcas sociales tales como *tags,* "me gustas", comentarios, contenidos compartidos, *retweets,* menciones entre otros. En general todo lo que se comenta con respecto a cierto tema en redes sociales como Facebook o de *microblogging* como Twitter o lo que se difunde en plataformas de vídeo como YouTube o en redes de mensajería como WhatsApp, o cada vez que entramos a una página y tenemos autorizado la instalación de *cookies* o registros de navegación. Los objetos digitales son todos los datos personales que voluntariamente (consciente e inconscientemente) las personas suministran al usar la Web, lo que hace de ésta un laboratorio para que los científicos sociales puedan estudiar fenómenos que anteriormente era casi imposible, bien por la dificultad de obtención de los datos, bien porque el comportamiento de la población bajo estudio podía verse afectada por la presencia física del investigador.

4. DEBATES ABIERTOS: LÍMITES Y DESAFÍOS DEL BIG DATA Y LOS MÉTODOS DIGITALES

Novedades que nos ofrece esta aproximación

Una aproximación clásica a Internet es asumirlo como un medio de comunicación y tratarlo como tal, enfoque que en la presente investigación buscamos superar y por ello dedicamos todo un apartado a la comprensión de Internet más allá de una simple herramienta de comunicación. Desde el enfoque de Internet como un medio, algunos de los análisis de datos en la Web están muy influenciados por los estudios de medición de audiencias.

Una de las primeras formas de medir la audiencia en medios fue mediante encuestas. Cuatro veces al año se entrevistaba a familias, preguntándoles qué programas de radio y televisión veían. A pesar de que se siguen utilizando, son técnicas que presentan múltiples sesgos de respuesta y han comenzado a quedar rezagadas con la aparición de métodos más automáticos que evitan la injerencia del entrevistador y la posible falta de sinceridad del encuestado. En 1940 se comienzan a utilizar los primeros audiómetros, que registraban qué frecuencia de radio se estaba escuchando y por cuánto tiempo, lo que permitía identificar el nivel de audiencia y fidelidad a un programa. En 1950 estos *ratings* comienzan a aplicarse en televisión que, a través del codificador de la tele permitía enviar datos vía telefónica sobre qué canal se estaba sintonizando, a qué hora y en qué momento se daban los mayores picos de audiencia, información que era utilizada para determinar el precio que debía pagar la publicidad de acuerdo al horario de mayor *rating* (Wimmer y Dominick, 2010).

En este sentido, la medición automática de audiencias y de consumo de información no es algo novedoso que surja con Internet, ya presenta una extensa tradición en otros medios como radio y televisión. Sin embargo, se presenta un salto cualitativo con Internet al permitir una identificación mucho más detallada del perfil de quién consume cierto tipo de información. Más allá de un papel pasivo en el que sólo se recibe o consume información, la interactividad característica de la Red permite conocer no sólo si se presenta exposición a determinado tipo de información sino qué se opina sobre la misma y qué tipo de contenidos y temas produce la persona titular de distintos perfiles en redes sociales.

En este sentido sería anacrónico conformarnos con una información pasiva de una unidad de análisis agregada como el "hogar", cuando podemos tener como unidad de análisis directamente al individuo, no sólo en su rol de receptor sino también como emisor y creador de contenidos que interactúa mediante las redes sociales con otros individuos y organizaciones. Interacciones que quedan registradas, tanto su contenido cómo la frecuencia de éstas.

Esto permite hacer un análisis en mayor detalle de los intereses y preferencias del individuo, como identificar con quién comparte ciertos contenidos y con qué frecuencia o intensidad lo hace.

Así, el análisis de datos sociales en la Web va más allá de una simple medición de audiencias. Es lo que autores como Rogers (2013:35) denominan investigación post-demográfica, nombre que hace referencia a que mientras los anteriores estudios demográficos recurrían a encuestas para conocer edad, fecha de cumpleaños, si se estaba soltero o casado, además de preferencias y demás datos personales, en la actualidad mediante el estudio de los diferentes objetos digitales que los individuos dejan al utilizar la Web es posible obtener esta información e incluso con mayor detalle, sin el riesgo de un posible sesgo por parte del entrevistador o falta de honestidad por parte del entrevistado al sentirse intimidado por la entrevista o encuesta.

A través de objetos o marcas digitales (su dirección IP, los hiperenlaces entre páginas web de la organización o el *blog* del individuo, los "me gusta" —*likes*— a una causa en su perfil de Facebook, las imágenes o vídeos que sube y consulta en la Web, los grupos a los que pertenece en redes sociales, el *tweetear* o *retweetear* sobre cierto tema, así como identificar con quién y con qué frecuencia e intensidad comparte este tipo de información. . .) es posible crear un perfil demográfico del individuo o de una organización sin que sean conscientes de estar siendo observados y analizados..

Las aplicaciones son múltiples, a nivel comercial para explicar preferencias de compra, opiniones sobre una marca, interés en un programa; en temas de campañas políticas para observar la preferencia sobre un candidato, sobre lo que de él se comenta y entre qué tipo de colectivos puede tener una mayor aceptación social o rechazo; en temas de movilización social y política, el poder identificar a los distintos líderes de opinión, los canales o esferas que privilegian e incluso la interacción entre los múltiples canales para ganar adeptos a una causa (Costanza-Chock, 2011). En términos de políticas públicas, para mejorar aspectos como la regulación del tráfico o el suministro de bienes y servicios en ciertos lugares de una ciudad en horas punta concentrando recursos como servicios de transporte o de seguridad, entre otros.

En investigaciones sobre juventud, poder analizar sobre qué temas conversan los jóvenes, qué les interesa, qué les preocupa sin tener la presencia física del observador y el sesgo que esto puede generar. La información se obtendría analizando los contenidos que los jóvenes vuelcan en las redes sociales, con quién interactúan en la misma, qué temas comparten y prefieren e incluso si llegan a usar la Red con fines políticos y de movilización social y, de ser el caso, cómo la utilizan. Además de poder identificar dentro de las redes de discusión de jóvenes, los líderes de la discusión, los actores más centrales y, en muchos casos, los más influyentes.

5. LA COEXISTENCIA DE MÚLTIPLES CANALES EN LA RED

En la mayoría de investigaciones sobre Internet y sus efectos en la acción colectiva, usualmente la Red se asume como si fuera un todo indivisible y homogéneo "la Web" o "Internet" o se analiza sólo un canal específico, sean *blogs,* redes sociales, o de *microblogging* para luego extrapolar las observaciones acerca de un canal como si fueran los efectos de Internet en la acción colectiva (Chadwick y Howard, 2008)[4]. Una investigación novedosa a partir de la aproximación de métodos digitales es lo que se conoce como "investigación transmedia" (Costanza-Chock, 2011), que implica analizar las prácticas que movimientos sociales y en nuestro particular interés las movilizaciones juveniles hacen de los canales o esferas de la Web. Canales tales como redes de *microblogging* (Twitter), redes sociales (Facebook, Link, Pininterest), motores de búsqueda (Google, Yahoo, Bing), redes de vídeo (Vimeo, YouTube), agregadores de noticias (Meneame, Digg), entre otros.

Un tipo de estudio transmedia implica analizar el uso diferencial que los movimientos sociales y en general diversos tipos de actores políticos (candidatos, partidos, grupos de interés) hacen de los distintos canales o esferas que se encuentran en Internet para hacer visible su mensaje. Desde el uso de los medios tradicionales que tienen su versión *online* (prensa, radio, televisión) hasta lo que llamamos medios nativos (buscadores, redes sociales y de *microbloging,* bitácoras, etc.). Una movilización transmedia implica que los mensajes son difundidos a través de múltiples medios de comunicación. "En la actualidad, las historias y mensajes más importantes tienden a fluir a través de múltiples plataformas de medios" (Jenkins, 2004) con el propósito de aumentar su posible alcance, pero siempre buscando crear un mensaje coherente y coordinado. (Costanza-Chock, 2012).

Para hacer un análisis transmedia como parte de la aplicación de métodos digitales es necesario identificar las características que presentan los diferentes canales que encontramos en la Red y así poder adaptar mejor nuestro análisis. Tal como se mencionó, la interactividad es una característica central de canales inherentes a la Web tales como las redes sociales (Chadwick, 2008), en las que los usuarios son tanto difusores como creadores de los contenidos más allá de ser simples receptores.

Dentro de la amplia diversidad de canales que encontramos en la Web podemos clasificar dos grandes grupos: los denominados medios tradicionales y los medios de auto-comunicación de masas MCM (Castells, 2009). Los primeros, aunque tienen presencia en la Web, su origen en la mayoría de los casos antecede a la Web y en la construcción, difusión y edición de contenidos

4. Cuando hablamos de canal hacemos referencia al espacio que surge como resultado de la interacción de los usuarios en determinado servicio (ejemplo: motores de búsqueda, redes sociales, blogs, etc..) con las características inherentes del servicio. Esto genera que cada canal presente una configuración personalizada para cada usuario. Un ejemplo son los resultados de los motores de búsqueda, los resultados que cada usuario obtiene son consecuencia tanto del conjunto de búsquedas realizadas por el usuario como de los algoritmos que organizan los resultados de otros usuarios y el marco legal por el cual se rige cada servicio. De forma similar sucede con redes sociales donde este tipo de canal es resultado de los contenidos que vuelcan voluntariamente los usuarios (consciente e inconscientemente) y las especificidades técnicas y legales del servicio que organizan o indexan la información.

es un equipo humano o un editor que por los criterios económicos y políticos del medio al que representan decide privilegiar una información sobre otra y cómo se presenta esta información.

En el caso de lo que llamamos MCM, son canales de distribución y producción de contenidos que en su mayoría han nacido con la Web, pero donde los usuarios juegan un rol protagónico en la producción de los contenidos (suministrando sus datos personales, fotografías, vídeos u opiniones sobre diversos temas...) aunque es la plataforma o canal bajo criterios de un programador o equipos de programadores que diseñan los algoritmos que han de organizar los contenidos y definir los derechos de quién podrá acceder a ellos o modificarlos.

Los dueños de la plataforma o canal definen el marco legal que el usuario debe aceptar si decide hacer uso del canal, indistintamente de la nacionalidad a la que éste pertenezca. Por ejemplo, los usuarios de Facebook o Twitter o cualquier otra plataforma en el momento de crear una cuenta aceptan los términos y condiciones que determina la plataforma indistintamente de la legislación que como nacional de un país el individuo esté obligado a cumplir La información que se difunda en el caso de los MCM ya no sólo depende del criterio humano o del editor, sino de la interacción entre los usuarios que producen los contenidos junto con las características técnicas y el marco legal del canal o plataforma que se esté utilizando.

El conjunto de canales presentes en la Web configura el espacio público en red, público en cuanto espacio de encuentro entre diversidad de personas que usan estos canales para expresar e intercambiar sus ideas y mensajes, un espacio que permite la comunicación y la creación de diversas formas institucionales. Tal como lo plantea Benkler (2006) este espacio provee diversas alternativas comunicativas a los ciudadanos, en la que puntos de vista minoritarios y con difícil acceso a medios tradicionales logran hacer difusión de sus puntos de vistas y contactar a sus pares.

Pero cada MCM presenta particularidades que las podemos clasificar bajo diversas dimensiones tal como es posible leer en la Tabla 1.

En un tipo de investigación transmedia, y tal como observamos en la Tabla 1, los MCM los podemos clasificar para su análisis bajo seis grandes dimensiones: la primera hace referencia al número de individuos que pueden asumir tanto el rol de creadores y difusores de contenido como el tamaño de la población objetivo a la que el canal permite que el contenido llegue. De acuerdo a esta dimensión, un movimiento social o un individuo privilegiaría un canal sobre otro, cuando se privilegia un tipo de comunicación más personalizada y focalizada en un individuo muy seguramente se privilegiarán canales como el *e-mail* o *post* dirigidos específicamente para un destinario, en el caso de buscar un mayor alcance se privilegiarán redes sociales y mensajes en muros *(wall)* o a través de redes de *microblogging*.

La segunda dimensión considera las diferencias entre una comunicación en tiempo real a una en diferido o asincrónica de acuerdo al canal que se privilegie y el objetivo que se busque. También de acuerdo al tamaño o extensión de los contenidos que desean transmitirse o su grado de elaboración, un canal será más apropiado que otro. En la creación de contenidos complejos y de un alto nivel de elaboración, plataformas como *wikis* o sistemas de documentos públicos compartidos en la nube serían los más apropiados.

TABLA 1
DIMENSIONES DE LOS MEDIOS DE AUTOCOMUNICACIÓN DE MASAS
Web 2.0. La distribución de la producción en las manos de muchos (Nick Carr)

DIMENSIONES	EJEMPLOS	CARACTERÍSTICAS
1. N° de emisores y tamaño de la población objetivo Medios sociales soportan diferentes escalas de producción y consumo de objetos digitales	E-mail	Una a una - una persona a muchas
	Wikis	Un grupo pequeño a un gran grupo
	Redes sociales y foros	Una a una - una persona a muchas
	Nasa-genoma	Masas
2. Fase de interacción Momento en el que se realiza el intercambio de información	Mail, foros, wikis, post	Asincrónica
	Chat, mensajería instantánea, videoconferencia, telefonía IP, videojuegos en línea	Sincrónica
	Redes sociales	Ambas
3. Tamaño del contenido	Twitter	140 caracteres, compartir links
	E-mail	Pocos párrafos, no grandes discursos
	Webs, blogs	Contenidos más largos. Todo el contenido que se desee y el servidor permita, convergencia de formatos y canales
4. Formato del contenido Granularidad tiempo y esfuerzo para obtener un bien público (critical mass and tipping point)	Youtube	Vídeos, comentarios, categorías de vídeos
	Flickr	Fotos, comentarios
	Mi tele	TV shows
	Facebook	Confluencia formatos (fotos, contenidos, mensajes, vídeo), preferencias, gustos
	Twitter	Tiende a confluencia de formatos (micro-mensajes, fotos, vídeos)
5. Tipo de vínculo o conexión	Páginas web y blogs	Links, tags
	Twitter	Seguidores, retweet, favorito, menciones
	Flickr, Pininterest, Instagram	Identificar (tagging) fotos-imágenes
	Facebook	Múltiples tipos (amistad o conocidos que comparten intereses o creencias, o momentos de la vida)
6. Propiedad y derechos de uso de contenido A quién pertenece la información, quién ostenta su titularidad, qué derechos tienen los usuarios sobre los contenidos	Google, Facebook, Yahoo, Microsoft	Corporaciones, derechos reservados sobre control contenidos que usuarios vuelcan en sus servicios, seguimiento constante y análisis de los contenidos con fines comerciales e investigación.
	Duck, N1:Lorea, Telegram, Wikipedia, etc.	Licencias abiertas, mayores derechos y poder de decisión de usuarios sobre los contenidos que vuelcan en estos servicios.

Fuente: Elaboración propia

La dimensión o tamaño de los contenidos y su nivel de dificultad suelen estar relacionados con la fase de interacción que el medio permite. Los medios de comunicación que se clasifican como sincrónicos facilitan la comunicación en tiempo real a través de mensajes usualmente cortos *(chats,* Twitter, WhatsApp). Este tipo de medios facilita la rápida movilización de potenciales simpatizantes y miembros de una organización ante eventos inesperados o que por sus características son momentos apropiados para la movilización (escándalos, desastres, atentados. . .).

Entre los medios asíncronos, su diseño facilita una tarea más reflexiva y de colaboración *(wikis,* documentos compartidos, foros, páginas web, *blogs,* etc.). Estos sistemas tienen la ventaja de permitir a cada participante programar su tiempo para decidir participar, sin que tenga que coincidir con otras personas. Por ejemplo, en el caso de las *wikis* o foros, son herramientas útiles que permiten diferentes niveles de compromiso o disponibilidad para realizar una tarea determinada (Hansen, Shneiderman y Smith, 2010), lo que usualmente permite la obtención de contenidos más extensos y estructurados para su análisis.

Relacionado con la extensión del contenido está el formato del mismo; podrá ser un mensaje de texto, un vídeo en diferido o en directo *(streaming)* o una imagen. El tipo de formato que se utilice también está en relación con los límites que imponga el canal junto con el ancho de banda del que dispongamos en términos de la extensión del mensaje o del tamaño de la imagen o del vídeo. Todos los contenidos por sí mismos no tendrán sentido si no se difunden, pero es importante precisar a través de qué vínculo o enlace en determinados canales se transmite un mensaje, cuál es el tipo de vínculo característico del MCM que permite que se difunda un mensaje y que se deberá considerar para su posterior análisis. El contenido se comunica a través de un *link* o un *post* en la Red a conocidos del movimiento, o se transmite a nuestros seguidores en una red de *microblogging* o por el contrario se enlaza en una página web o se publica en un *blog.* Los vínculos característicos de cada uno de estos canales influirán directamente en el alcance del mensaje y finalmente a quién potencialmente le llegue. De igual forma, debe tenerse presente en el momento que se estén analizando los datos de qué canales se obtienen éstos para poder saber el significado de la relación que se está analizando: la relación puede ser desde la emisión de un mensaje a un hiperenlace entre la diversidad de tipos de marcadores digitales que es posible indentificar ("me gustas", categorías temáticas, favoritos . . .).

La última dimensión a considerar en un análisis transmedia es que los MCM más populares en su mayoría son propiedad de corporaciones privadas. Es crucial conocer las condiciones legales que el usuario asume al utilizar estos canales y lo que se permite en cuanto al análisis de datos; por ejemplo qué tipo de uso pueden dar los dueños de los canales a los contenidos suministrados por los usuarios (consciente e inconscientemente), quiénes son los dueños de los contenidos albergados en la plataformas, los individuos que consigan su información personal o las empresas que suministran un servicio a cambio de que el individuo accede a que su información se utilice con fines publicitarios o son contenidos públicos. A pesar de que estas plataformas puedan ser libres de pago, no son gratis y quizás el precio a pagar puede llegar a ser muy alto. Es crucial considerar el nivel de confidencialidad y privacidad que ofrecen las plataformas, en qué casos los dueños del canal le suministra a gobiernos, autoridades o terceros la información consignada en las mismas.

En todo análisis transmedia debemos tener en cuenta las seis dimensiones antes mencionadas. Además no debemos desconocer que es el nivel de conectividad de la red de comunicación y la estructura de la misma lo que finalmente favorecerá directamente el alcance que logre un mensaje (Gonzalez-Bailon, 2013), lo que implica reconocer la existencia de jerarquías en el interior de las redes, donde no todos los nodos o actores presentan igual nivel de importancia, en cuanto a la mayor influencia que pueden llegar a tener ciertos nodos en la difusión de un mensaje resultado de su posición y nivel de interconectividad que presenta en la Red.

6. INVESTIGACIÓN DE MINERÍA DE DATOS EN LA RED

Usualmente no es posible hablar de una sola técnica o método para realizar la obtención y el análisis de datos digitales. Depende también de cuál sea la pregunta de investigación que se pretenda resolver y de qué fuentes se estén obteniendo los datos. También es totalmente anacrónica la diferencia de enfoques cualitativos y cuantitativos; en la investigación a través de métodos digitales la triangulación es fundamental (Boyd y Crawford, 2012). Tal cómo se ha mencionado, la Red es una fuente inagotable de información, pero su acceso no en todos los casos es igual de sencillo, existen consideraciones técnicas, éticas e incluso legales de lo que deseamos obtener. Nos centraremos en aquellas fuentes y procedimiento de carácter legal, que aunque requieren ciertos tecnicismos pueden desarrollarse desde investigación de escritorio.

El propósito de este apartado no es ir al detalle de cómo hacer el proceso de obtención y análisis, es exponer un panorama de las fuentes disponibles, de las técnicas más comunes, pero no las únicas y de las consideraciones metodológicas y éticas a tener en cuenta. Para el detalle recomendamos dirigirse a manuales especializados y páginas web de soporte para cada una de las técnicas y fuentes aquí mencionadas.

Fuentes

Los datos en la Web pueden presentarse tanto de una forma estructurada como desestructurada. En el primer caso, los datos están organizados y el proceso de clasificación, limpieza y análisis es relativamente sencillo. En datos desestructurados debe crearse la base de datos de acuerdo a los criterios que determine el investigador, lo que implica nuevos desafíos técnicos y, en algunos casos, computacionales para organizarlos, limpiarlos y permitir su posterior lectura y análisis. Entre las fuentes de datos estructurados podemos encontrar:

APIs (Application Programming Interfaces)

Las APIs que ofrecen servicios como Twitter, Huffington Post, Change.org o Facebook, entre las más conocidos, tienen la ventaja de que al obtener la información claramente son identificables las diferentes dimensiones que puede presentar la misma. En el caso de Twitter, por ejemplo, en la información que el API ofrece es posible obtener el titular que envió el *tweet*, el momento en que lo hizo, el número de *retweets* y menciones que recibió, desde dónde lo hizo, si el usuario tiene activada la geo-localización en el dispositivo desde donde lo envió, la descripción del perfil de

usuario que lo envió, el número de seguidores, la antigüedad con la que lleva usando Twitter, entre otras variables que el servicio ofrece. No obstante, a través de los API sólo es posible obtener una muestra del conjunto de datos bajo análisis. Twitter sólo permite obtener hasta un 1% del total de mensajes que en determinado momento se están discutiendo en el servicio. Si hay un alto nivel de tráfico y demanda de datos, el servicio pude ser interrumpido unilateralmente por la empresa, si la frecuencia y volumen de generación de datos *(tweets)* ante cierto evento es muy alta, el ancho de banda del que se disponga (conexión a Internet) no puede ser suficiente para descargar el volumen de datos.

Esto genera problemas en la calidad de la muestra de datos que finalmente logremos obtener. También dentro del mismo servicio pueden existir diversos tipos de API: algunos nos permiten obtener datos de manera casi ininterrumpida durante el periodo de tiempo que programemos, es el caso de la *stream* API de Twitter, otros buscan los términos específicos que definamos —el Search API—, o simplemente el API nos permite graficar nuestro conjunto de amigos y *post* entre ellos, como ofrece el *graph* API de Facebook.

La mayoría de APIs exigen el registro y autenticación del usuario que lo demanda; en algunos casos son restringidos sólo para ciertos usuarios como en el caso de APIs para datos de Facebook. En todo caso para intentar mejorar la calidad de la muestra se pueden crear diferentes cuentas para hacer múltiples requerimientos de información y utilizar, por ejemplo, conexiones a través de redes virtuales y así reducir la posibilidad de ser bloqueado por el servicio.

Otro inconveniente del uso de APIs, y en particular para el caso de Twitter, es que sólo podemos analizar eventos con una antigüedad no mayor a dos semanas, o que estén sucediendo durante el momento del análisis. Está situación dificulta poder analizar en su conjunto toda la dinámica que presenta un evento en la Red, cuando muchas veces sólo comenzamos a recopilar datos cuando el evento ya ha empezado y tiene un volumen suficiente que permita que nos enteremos del mismo, perdiendo así todo el inicio.

Compra de datos

Otra opción que soluciona el problema de muestreo y ofrece la posibilidad de acceder a datos históricos es la compra de los datos. Diferentes empresas ofrecen este servicio y cobran en función del volumen de los datos, el tiempo de utilizar la plataforma que ellos ofrecen y, en algunos casos, según la complejidad de la búsqueda que se esté solicitando. Cuantos más filtros o más específico sea el conjunto de datos que se solicita, más aumenta el precio. Las empresas más conocidas son TOPSY, recientemente adquirida por Apple, GNIP y Datasift. A pesar de que ofrecen descuentos para uso académico, el precio más económico está en los 500€ mensuales y en plataformas que ofrecen análisis de texto y sentimiento incluyendo datos, se acerca a los 2000€ mensuales.

Además de ser un negocio, estas empresas buscan recuperar la inversión que debieron pagar a empresas como Twitter por el hecho de poder utilizar lo que se conoce como el *firehose* o su base datos completa Aunque no se conoce la cifra oficial, comentan que en el 2009 Google pagó aproximadamente 20 millones de dólares USA por tener acceso al *firehose* de Twitter (Luis

Fer Mtz, 2010). En la actualidad Twitter no ha renovado el acuerdo con Google, con el propósito de que las búsquedas que tengan que ver con temas difundidos en Twitter se hagan a través de su motor de búsqueda (TOPSY, GNIP, Datasift).

Extracción de datos (scrapping)

Si financieramente no es posible acceder a datos históricos, o por el tipo de investigación y fuentes de los mismos éstos no están disponibles a través de un API, es posible aplicar lo que se conoce como un *scrapper,* que son pequeños programas informáticos para extraer información específica de páginas web y ponerla en un formato que facilite su análisis. Este procedimiento puede ser dirigido prácticamente a cualquier contenido que se encuentre visible en la Web (excepto PDF). A través del procedimiento se busca identificar marcadores en el contenido (texto, imágenes, tablas) que se desea obtener y, a partir de los marcadores, descargar los datos en un formato por ejemplo tipo tabla, CVS (separado por comas y tabuladores) o de imagen (JPG) que permita su posterior análisis, en cualquier hoja de cálculo o programa de análisis de contenido, todo depende de la información que obtengamos y de lo que estemos buscando resolver.

Es una técnica de obtención de datos dirigida en particular a contenidos no estructurados. Puede ser utilizada, por ejemplo, en análisis de páginas web. Un caso particular puede ser el análisis de páginas web de medios de comunicación para el análisis de agenda e incluso de *timelines* en redes sociales cuando no es posible acceder al API o se buscan datos históricos y no se pueden comprar los datos. El inconveniente de este último proceso es saber qué tan estadísticamente representativos son el conjunto de los *tweets* históricos que se obtienen a través del *timeline* o del buscador específico para el caso de Twitter.

Siempre, al utilizar esta técnica, es pertinente ser prudente con el volumen de demanda de datos al servidor que aloja el sitio web al que se le estén solicitando los datos. Una demanda excesiva o reiterativa puede generar el bloqueo de nuestra IP al interpretar el servidor que está siendo atacado. La demanda de datos debe darse en un sentido amigable y en intervalos de tiempo que no saturen el tráfico del servidor que aloja el sitio web. Tal como se mencionaba, es posible utilizar redes virtuales para evitar que nuestra IP sea reconocida, no obstante es fundamental respetar el sitio que nos ofrece la información que estamos buscando.

Hiperenlaces entre páginas web

Otro tipo de datos a obtener son los hiperenlaces o vínculos que presenta una página web. Los hiperenlaces tienen un rol en la visibilidad que puede llegar a alcanzar un sitio web. No es lo mismo que enlace a una página muy popular que a muchas páginas que nadie visita.

De igual forma, aunque existen enlaces de carácter estructural, que por defecto la página enlaza a sus cuentas sociales o al servidor que la alberga, una vez realicemos un ejercicio de limpieza, los enlaces también nos reflejan una intencionalidad: con quién deciden compartir información los titulares de la página web o hacía qué sitios desean los titulares de la página web que sus lectores probablemente se dirijan. Usualmente hay una afinidad entre los enlaces que se ponen

en la página web y los sitios a los cuales éstos se dirigen. En este sentido es posible, a través del análisis de hiperenlaces, establecer la red temática que hay alrededor de un tema (Rogers, 2008; Thelwall, 2009).

Entre los límites de este tipo de análisis, el primero es el alcance del rastreador o araña *(crawler* o *spider)* del cual dispongamos. De qué tamaño es la red temática de páginas o sitios web que deseamos obtener. Es determinante el momento en el cual se realiza la búsqueda, porque los hiperenlaces pueden cambiar en el tiempo y, de acuerdo al momento de la controversia o asunto que sigamos, éstos pueden variar. También es fundamental un trabajo documental previo para establecer el conjunto de sitios web del cual parta el análisis, los puntos de partida a partir de los cuales el rastreador que se utilice comenzará a explorar la Red (el conjunto de hiperenlaces) entre los sitios web alrededor de un asunto. Ante la variación en el tiempo de los hiperenlaces y de los sitios web pude ser necesario otro tipo de análisis.

Acceso a páginas web históricas

Otro tipo de información que puede ser útil y se mencionó en parte en la sección de métodos virtuales es el análisis de páginas web. No obstante, hay páginas web que han sido ya descatalogadas. Si lo que buscamos es hacer un análisis cross-temporal o incluso longitudinal de cómo ha cambiado una página web a lo largo del tiempo, un excelente recurso es acceder a repositorios digitales tales como el WayBack Machine (http://archive.org/web/). En su página también encontramos referencias a otros archivos digitales. Este tipo de recursos nos permiten analizar, por ejemplo, cómo páginas web de organizaciones han cambiado a lo largo del tiempo y con la incorporación de las nuevas tecnologías de información y comunicación, el caso de las redes sociales. Este archivo puntualmente tiene registros desde 1996 y cuenta con más de 368 billones de páginas web archivadas.

Técnicas en los métodos digitales

Ante la amplia diversidad de información que es posible encontrar en la Red, son múltiples las técnicas a utilizar para el análisis de los datos digitales. Sin embargo las dos más utilizadas parten tanto del análisis de redes como del análisis de contenido.

Análisis de redes

El análisis de redes es una de las técnicas que más se utilizan para el análisis de datos en la red, en cuanto privilegia en su análisis la relación que se presenta entre los objetos que estamos analizando. Sean hipervínculos entre páginas web, intercambio de mensajes y, en general, intercambio de contenidos entre miembros de redes sociales o de *microbloging*.

A diferencia de las técnicas de estadística tradicional, no es necesario cumplir el supuesto de independencia entre las unidades de observación (Hanneman y Riddle, 2005), precisamente el análisis de redes se concentra en las relaciones entre estas unidades, en estudiar el tipo de relación y, de acuerdo a las preguntas de la investigación, la fuerza e intensidad de las relaciones (Granovetter, 1983). En un primer nivel interesa observar las características que puede presentar

la Red en su conjunto, por ejemplo la densidad de la misma, en cuanto a nivel de interacciones, los grados de separación que ésta presenta, si es una red aleatoria o una red de pequeño mundo (Watts, 2003) en la que existen interconexiones entre los nodos que reducen la distancia entre los nodos de la red, lo que favorece la velocidad de difusión de la información.

También es relevante identificar cómo se distribuyen las interacciones entre los miembros de la red, si es una red de libre escala (Barabási, 2003) en las que unos pocos nodos (cuentas de redes sociales, páginas web, organizaciones…) concentran la mayoría de interacciones, mientras para la mayoría de nodos su nivel de interconexiones es mínimo, o presenta lo que se conoce como una distribución de larga cola. Entre otras implicaciones, el poder identificar los nodos más interconectados permitirá desarrollar mejores estrategias de difusión de información.

También puede interesar analizar cómo cambia la red en el tiempo. Tal como se ha mencionado, las interacciones son dinámicas (intercambio de mensajes, de *post,* de hiperenlaces…). El momento en el que la red se analiza, influye en la configuración de la misma; por ejemplo, al seguir un debate de una política o un proceso de movilización no es lo mismo hacer la medida al principio del debate que en el auge del mismo (Andrews y Biggs, 2006; Baños *et al.,* 2013, entre otros).

Además de analizar la Red como un todo, también puede ser de interés analizar nodos específicos de la Red (cuentas de redes sociales, páginas web, organizaciones, etc.). Por ejemplo, analizar qué tan central es un nodo (la cuenta de una persona o su perfil), en términos del número de interacciones que recibe (ejemplo: *links, retweets*, comentarios) lo que incidirá directamente en su nivel de visibilidad. También el rol que puede tener un intermediario importante *(betweenness)* entre grupos de nodos que encontramos en la red bajo análisis, lo que le da un alto nivel de prestigio dentro de la Red (Wasserman y Faust, 1994). También hay nodos cuyo nivel de influencia aumenta (cercanía-*closness)* por el simple hecho de estar cerca de nodos muy populares.

Identificar este tipo de indicadores permite analizar cómo se difunde la información a través de la red temática. Si se desea que un mensaje llegue a tener una mayor repercusión y alcance es necesario poder identificar los nodos más visibles y concentrar en ellos la atención. Poder identificar qué actores son los más influyentes o los más centrales o juegan un rol clave como intermediarios en la difusión de información. De acuerdo a la pregunta de investigación que se tenga, el interés es identificar los nodos más activos, los que presentan el mayor envío de enlaces o interconexiones (mensajes, hipervínculos, etc.) dentro de la red bajo análisis.

Análisis de contenido

Otra técnica de uso común, es el análisis de contenido. En gran medida entra a complementar el análisis de redes. También es posible realizar análisis de contenido recurriendo a indicadores utilizados para análisis de redes; en este caso, el propósito es analizar los términos más visibles en la red temática y con qué términos se asocian, lo que nos permitirá tener una idea de qué es lo que se comenta en la red temática, qué temas son los recurrentes, qué temas son casi invisibles o prácticamente no se mencionan. En nuestro caso de interés, qué temas comentan los jóvenes con respecto a cierto asunto o en determinada organización juvenil o a través de los medios sociales o un conjunto de páginas web, o analizar cuánto aparecen reflejadas las preocupaciones de los jóvenes en los medios tradicionales en un periodo electoral o el que se considere relevante analizar.

Dentro del análisis de contenido usualmente se recurre a lo que se denomina análisis automático de texto, dentro del cual es posible hacer análisis de sentimientos y, en un nivel de sofisticación mayor, realizar lo que se conoce como aprendizaje de máquina. En el análisis de texto, los contenidos que se obtengan en la web, se analizan buscando identificar los términos más recurrentes y qué temas son los que más se comentan dentro del conjunto de datos bajo análisis (por ejemplo: *tweets* o páginas de medios de comunicación o de organizaciones o comunicaciones de las mismas).

También mediante un ejercicio de contrastación con conjuntos de palabras o *corpus* de palabras con cierta valoración, se busca contrastar en qué sentido se trata un tema, si al tema se la da una connotación positiva, negativa o neutra. El conjunto de palabras o *corpus* que utilicemos para contrastar lo definimos nosotros. Ya existen en la Web muchos *corpus* asociados a sentimientos y el sentido positivo o negativo que puede tener un término, incluso existen *corpus* de palabras especializados de acuerdo al tema que estemos analizado. No obstante, sigue siendo un desafío lograr identificar de forma automática aspectos como la ironía o el sarcasmo con que se utiliza una palabra en determinado contexto.

El *machine learning* o aprendizaje de máquina dentro de la minería de datos es un conjunto de procedimientos mediante el cual se crean algoritmos o programas que de acuerdo a los criterios que definamos permiten automatizar procesos de clasificación y categorización de un gran volumen de datos; por ejemplo millones de *tweets,* miles de páginas web o documentos cuya clasificación manual requeriría mucho tiempo. Dentro del *machine learning* existen diferentes tipos de procedimientos por medio de los cuales se busca automatizar procesos para identificar patrones dentro del conjunto de datos —clusterización, aprendizaje asociativo, redes neuronales, máquinas de soporte vectorial entre otros— (Romero *et al.,* 2011; Witten y Frank, 2005). Por ejemplo, entre estos procedimientos, la clusterización parte de criterios definidos por el programador para crear grupos de contenido que compartan criterios comunes; otros procedimientos a partir de un modelo de categorización previamente definido, permiten al ordenador asignar probabilidades para replicar las categorías en el conjunto de datos que se asigne, es lo que se conoce como análisis vectorial de aprendizaje de máquina (Pang, Lee y Vaithyanathan, 2002; Witten y Frank, 2005). Diferentes servicios de la Web, tales como traductores automáticos y buscadores utilizan este tipo de tecnologías. A medida que más usuarios realizan búsquedas o traducciones, el sistema va mejorando sus resultados.

Consideraciones metodológicas

A pesar de las posibles ventajas que puedan presentar el análisis y cruce de masivas cantidades de datos, es importante no olvidar que como toda aproximación técnica presenta ciertos supuestos y limitaciones.

Apophenia

El término significa ver o identificar patrones donde no existe nada. Simplemente por las masivas cantidades de datos podemos identificar correlaciones en todas las direcciones y encontrar que una relación o muchas son significativas sin que esto implique que realmente estemos explicando algo. La correlación puede mostrar un patrón, pero no significa causalidad.

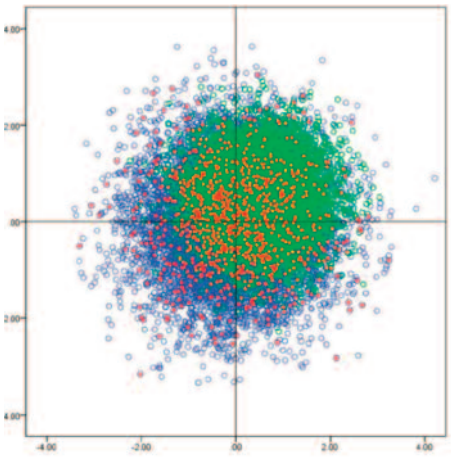
De igual forma, algunas de las técnicas tradicionalmente aplicadas para identificar niveles de correlación entre variables, no son pertinentes de acuerdo al tipo de datos que estemos analizando. Medidas de correlación comúnmente usadas en Ciencias Sociales como la Pearson, que parten de una supuesta normalidad y que privilegian relaciones lineales que en muchos casos la distribución de nuestros datos no cumplen, no sería técnicamente preciso aplicarlas. Esto ha hecho que medidas no paramétricas vuelvan a adquirir relevancia para identificar niveles de correlación, por ejemplo en distribuciones de potencia (caso Spearman) e incluso en el caso de masivas bases de datos se están desarrollando medidas no paramétricas para identificar posibles relaciones entre variables que presenten diferentes tipos de distribución: es el caso del coeficiente de máxima información (Reshef *et al.,* 2011).

Pero lo más importante es no perder de vista las preguntas que nos hacemos, los presupuestos teóricos, así como los metodológicos, de los que partimos. No es sólo tener un gran volumen de datos, éstos no se explican por sí solos. La interpretación del personal científico sigue siendo central (Boyd y Crawford, 2012).

Representatividad y características sociodemográficas de la población bajo estudio

¿Cuán representativos son los datos con los que contamos de la población bajo estudio? ¿Qué tipo de muestra es la que tenemos? Podemos tener millones de datos, pero no necesariamente son representativos. Incluso más que una gran cantidad o volumen de datos nos dice más una muestra aleatoriamente distribuida, que un alto porcentaje de datos de la población que está sesgada y en este sentido no sea representativa. Tal como podemos observar en el siguiente gráfico, es mucho más representativa una muestra aleatoria del 5% de la población (puntos rojos), que una sub-muestra poblacional del 80% sesgada (puntos verdes) que no incluye una parte de la realidad poblacional, por ejemplo, que se representa en el cuadrante inferior izquierdo del gráfico (Zhu, 2013).

GRÁFICO 5
TIPO DE MUESTRA CON RESPECTO A LA POBLACIÓN



Fuente: Conferencia Universidad de Hong Kong. 15 de marzo del 2013, Jonathan Zhu.

En este sentido es fundamental cómo se obtiene nuestra muestra de datos; si no es posible acceder al *firehose* (por ejemplo para el caso de Twitter) podemos intentar medidas reiterativas en diferentes momentos del evento o cuestión que estemos siguiendo. La triangulación de métodos también nos puede dar más robustez en los resultados que finalmente obtengamos. Los APIs, por lo menos para el caso de Twitter, permiten obtener muestras representativas (González-Bailón *et al.,* 2012; De Choudhury *et al.,* 2010) si la recolección de datos es sistemática, aplicando múltiples términos de búsqueda (Morstatter *et al.,* 2013) y continuada desde el comienzo de los eventos; el problema es lograr anticipar el inicio de un evento.

También es clave aclarar representatividad frente a qué población. Es importante precisar a nivel poblacional cuántas personas utilizan Twitter; la representatividad puede ser ante los usuarios de Twitter, pero difícilmente la podremos extrapolar al conjunto de la población. Ésta es la conclusión de un estudio del Centro de Investigación Pew (2013) que durante el año electoral en EEUU compararon los resultados de las encuestas nacionales frente al tono de los *tweets* en respuesta a ocho eventos noticiosos importantes, incluyendo el resultado de las elecciones presidenciales, el primer debate presidencial y los principales discursos de Barack Obama durante el 2012. A veces, la conversación en Twitter es más a la izquierda que las respuestas de la encuesta, mientras que otras veces es más conservadora. A menudo es la negatividad general lo que destaca. Gran parte de la diferencia puede tener que ver tanto con la estrecha franja de la población representada en Twitter, como con qué parte de los usuarios de Twitter deciden finalmente participar en determinada conversación. Los resultados muestran un claro sesgo ideológico más demócrata y de un alto nivel de formación (licenciados, master, doctorado) de los usuarios de servicios como Twitter (Pew Research Center, 2013). En España son las personas con mayor nivel de formación entre 25 y 35 años los más populares en este servicio (IAB Spain y Elogia, 2013).

En este sentido, la pregunta sobre a qué población representa la muestra que tenemos es más que pertinente. No es lo mismo estudiar países con un alto nivel de penetración de Internet en todas las franjas de edad de la sociedad como puede ser Noruega o Suecia en Europa (Eurostat, 2011) que el caso español donde las personas de mayor edad y con menor nivel de cualificación tienen un alto riesgo de exclusión digital (CIS, 2012). También de acuerdo al canal del que se estén obteniendo los datos, su nivel de uso y perfil de usuarios no es el mismo, lo que tendrá implicaciones directas en las extrapolaciones que se deseen realizar. Por ejemplo los jóvenes son los principales usuarios de Tuenti en España, mientras que los usuarios más populares de Twitter se ubican en la franja de los 25-35 años (The Cocktail Analysis, 2012).

Aspectos éticos

Es central considerar que si por el hecho de estar disponible en la Red es ético disponer y analizar estos datos en cuanto a la vulneración de derechos que puede presentarse. Quizás el individuo de quien tratan los datos no desea que sean analizados y que se publiciten, tal vez en el momento que él o ella decidió volcar estos datos en la Red, su circunstancia vital era otra, o incluso la persona no fue quien decidió conscientemente poner los datos en la Red, sino que un tercero fue el que puso la información sobre el individuo. Aproximaciones como el Big Data, tal como se mencionó, también se caracterizan por su carácter relacional, lo que en términos de privacidad y de-

recho al buen nombre puede tener repercusiones. Para el caso de los datos individuales que dispongamos es posible que decidamos no incluir su nombre en la investigación —para respetar su privacidad—. No obstante, si contamos con los datos sociodemográficos del individuo, sabemos sus preferencias, su domicilio, lo que estudia o estudió, en qué trabaja y lo que opina, hay una alta probabilidad de identificar qué persona es o qué personas cercanas a él puedan saber de quién se está hablando. Según el tipo de investigación, hay información que las personas no desean que otros sepan, por más que la podamos encontrar en la Red.

En este sentido el carácter anónimo requiere pensar nuevas medidas para este tipo de enfoque, que el simple hecho de no incluir el nombre en la investigación. Otros autores plantean la necesidad de destruir después de cierto tiempo los registros utilizados para evitar que puedan caer en manos equivocadas que den un uso no adecuado a cierta información personal (Davis, 2012). Sin embargo, el riesgo es la imposibilidad de poder replicar las investigaciones, elemento necesario para evaluar la calidad de cualquier práctica científica; además la contrastación es esencial para el avance mismo de la ciencia.

Ante esta situación es necesario un ejercicio de responsabilidad y transparencia por parte de la comunidad científica y un debate que permita garantizar la confidencialidad de los datos con las garantías a todos los derechos individuales, así como que el avance de la ciencia se vea lo menos afectado. Se puede sugerir la necesidad de un consentimiento informado utilizado en enfoques investigativos anteriores, pero el carácter relacional de los datos implicaría que el individuo esté dispuesto a que su vida o una buena parte de ella se vuelva totalmente pública.

Surge también la cuestión de si existe un comportamiento individual ajeno a la Red. ¿Es diferente el perfil del individuo que es posible recrear con todas las huellas que él o ella va dejando en la Red? A partir de los rastros que él o ella deja en redes sociales, en las búsquedas que realiza, en las páginas que consulta, en los bienes que compra, en la información que consume, en los comentarios que publica o difunde se puede crear un perfil individual. ¿Es acaso diferente este perfil, del perfil del individuo fuera de la Red? Castells (2009) nos habla de que en sociedades altamente desarrolladas, la Red ya es parte de la vida de las personas: se vive también a través de la Red, por cuestiones del trabajo, para comunicarse o quedar con amigos, para consumir entretenimiento o para realizar compras, entre otros muchos comportamientos. En esa medida, cada vez es más difusa la diferencia entre un comportamiento en la Red o fuera de ella; a través del móvil inteligente las personas están conectadas de forma constante. También es muy posible que no se desee que todas las facetas del individuo sean transparentes y conocidas, anque técnicamente sea factible conocerlas. Si no se expresa el carácter público de la información, se solicita el consentimiento de los afectados explicándoles posibles consecuencias, no sería ético acceder a ésta. De igual forma debería respetarse el derecho a desear ser anónimo, aunque parezca paradójico en un mundo cada vez más interconectado y donde millones de personas cada vez que se conectan publican sobre su vida en la Red, tanto consiente como inconscientemente (Mayer-Schönberger, 2011).

También existe la cuestión de quién será el dueño de los datos. Son pocas las personas que leen los términos o condiciones de las redes sociales cuando deciden volcar sus datos personales (Mayer-Schönberger, 2011). ¿Quiénes son los dueños de estos datos, la red social , el titular de

la cuenta o el científico que los obtiene y procesa para poder hacer una lectura de los mismos? Legalmente, en la mayoría de los casos, la red social puede disponer de los mismos. Y de hecho es lo que se observa en la mayoría de servicios corporativos, que utilizan esta información con propósitos publicitarios al ofrecer estudios de mercado y publicidad cada vez más personalizados o en ciertos regímenes políticos ofrecer información considerada cómo sensible a las autoridades del régimen, por ejemplo opositores, personas subversivas o potenciales amenazas al régimen (Morozov, 2012).

7. ÚLTIMAS CONSIDERACIONES

Después de presentar este panorama de las alternativas de investigación que nos ofrecen las TIC, y en particular Internet, destacamos cómo Internet ha dejado de ser un objeto de investigación en sí mismo para convertirse en un espacio en el cual es posible estudiar diversos fenómenos sociales. En el texto realizamos un recorrido desde los métodos virtuales, a los digitales y al denominado Big Data en estudios transmedia. Un primer punto que resaltamos es que por más datos que tengamos no se debe olvidar el importante y subjetivo ejercicio de interpretación que exigen los datos. Tal como lo exponen Boyd y Crawford (2012), existe la falsa creencia de que por el hecho de manejar grandes volúmenes de datos, los científicos sociales nos acercamos más al anhelo del cuantitativismo de las mal llamadas ciencias exactas. Más cuando uno de los desafíos centrales está en la difusión de prácticas, contenidos y formas de aprendizaje que difícilmente pueden ser ajenos de la lectura del científico que interpreta las observaciones.

Evidentemente para cada caso particular los algoritmos se podrán ajustar y determinar qué palabras clave presentadas de cierta manera pueden tener un significado diferente de su definición literal, lo que significa que el criterio del investigador y su experiencia tienen un rol crítico.

En esta línea, los investigadores debemos ser capaces de dar cuenta de los sesgos en la interpretación de los datos Todos los investigadores somos intérpretes de datos. Los procedimientos para obtener y procesar los datos pueden partir de modelos matemáticos de un alto nivel de sofisticación y aparente precisión, pero tan pronto como el investigador trata de comprender lo que significa, el proceso de interpretación ha comenzado. También las decisiones del diseño que determinan lo que se medirá, se derivan de la interpretación (y a veces esto se nos olvida).

En este sentido es necesario por lo menos conocer las técnicas de obtención y análisis de datos mediante métodos digitales y Big Data por parte de los científicos sociales, para poder contribuir en el desarrollo y crítica a este enfoque. Es importante no dejarse deslumbrar por la cantidad de datos que se pueden obtener, también son necesarias voces críticas que mejoren las técnicas, que garanticen la representatividad de los datos, que formulen las preguntas pertinentes y que desarrollen mecanismos para hacer cumplir los desafíos éticos de este enfoque investigativo. Además, es esencial la aproximación multi-método dejando a un lado divisiones artificiales entre enfoques cuantitativos y cualitativos, en la medida en que los métodos digitales son integradores y exigen aproximaciones desde diversos enfoques. En este sentido, este texto más que pretender ser un manual que va al detalle, plantea todo un panorama del amplio mundo por explorar y construir.

8. REFERENCIAS

Andrews, K.T. y Biggs, M. (2006). "The dynamics of protest diffusion: Movement organizations, social networks, and news media in the 1960 sit-ins". *American Sociological Review* 71 (5): 752.

Bakshy, E.; Rosenn, I.; Marlow, C. y Adamic, L. (2012). "The role of social networks in information diffusion". En *Proceedings of the 21st international conference on World Wide Web:* 519–528. http://dl.acm.org/citation.cfm?id=2187907.

Baños, R.; Borge-Holthoefer, J.; Wang, N.; Moreno, Y. y González-Bailón, S. (2013). "Diffusion Dynamics with Changing Network Composition". *ArXiv* e-print 1308.1257. http://arxiv.org/abs/1308.1257.

Barabási, A. L. (2003). *Linked: How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life.* New York, NY: A plume book.

Bauman, Z. (2011). *Culture in a Liquid Modern World.* Polity Press.

Benkler, Y. (2006). *The wealth of networks: how social production transforms markets and freedom.* New Haven [Conn.]: Yale University Press.

Bennett, W. L. y Segerberg, A. (2011). "Digital Media and the Personalization of collective Action". *Information, Communication & Society* 14 (6) (septiembre): 770-799. doi:10.1080/1369118X.2011.579141

Berg-Schlosser, D. (2004). "The Quality of Democracies in Europe as measured by current indicators of democratization and good governance". *Journal of Communist Studies and Transition Politics,* 20(1): 28-55.

Bimber, Brian, Stohl, y Flanagin (2008). "Technological Change and Political Organization". En: Andrew Chadwick y Philip N Howard (eds). *Routledge Handbook of Internet Politics.* London: Routledge: 72-85.

Bollen, K. A. (1990). "Political Democracy: Conceptual and Measurement Traps". *Studies in Comparative International Development,* 25(1): 7-24.

Bollen, K. A. y Paxton, P. (2000). "Subjective Measures of Liberal Democracy". *Comparative Political Studies,* 33(1): 58-86.

Boyd, D. y Crawford, K. (2012). "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon". *Information, Communication & Society* 15 (5): 662-679.

Bruns, A. (2012). "How Long Is a Tweet? Mapping Dynamic Conversation Networks on Twitter Using Gawk and Gephi". *Information, Communication & Society* 15 (9): 1323-1351. doi:10.1080/1369118X.2011.635214.

Calderaro, A. (2010). Empirical analysis of political spaces on the Internet: The role of e-mailing lists in the organization of alter-globalization movements. International *Journal of E-Politics (IJEP),* 1, 73-87.

Calenda, D. y Lyon, D. (2006). "Culture e tecnologie del controllo: riflessioni sul potere nella società della rete". [Culture and technology of control: Reflexions about the power of network society] *Rassegna Italiana di Sociologia,* 4: 583-612.

Castells, M. (2009). *Comunicación y poder.* Madrid: Alianza.

Chadwick, A. (2008). "Web 2.0: New Challenges for the Study of E-Democracy in Era of Informational Exuberance". *I/S: A Journal of Law and Policy for the Information Society* 5: 9-42.

Chadwick, A. y Howard, P. N. (ed) (2008). *Routledge Handbook of Internet Politics.* London: Routledge.

CIS (2012). Barómetro junio 2012. estudio 2948.0.0 según edad. 01. http://www.cis.es/cis/export/sites/default/Archivos/Marginales/2940_2959/2948/Cru294800EDAD.html

Clinton, K.; Purushotma, R.; Robison, A. y Weigel, M. (2006). "Confronting the challenges of participatory culture: Media education for the 21 st century". *MacArthur Foundation Publication* 1 (1): 1–59.

Costanza-Chock, S. (2011). "Digital popular communication: Lessons on information and communication technologies for social change from the immigrant rights movement". *National Civic Review* 100 (3): 29-35.

Costanza-Chock, S. (2012). "Mic check! Media cultures and the Occupy Movement". *Social Movement Studies* 11 (3-4): 375-385.

Cristancho, C. y Salcedo (2013). "El estudio de la movilización social en la era del Big Data". En *IX Congreso Internacional Internet, Derecho y Política (IDP 2013): Big Data: Retos y Oportunidades.* Barcelona España: UOC-Huygens Editorial: 387-404. http://www.tandfonline.com/doi/abs/10.1080/1369118X.2013.808360.

Davis, R. (1999). *The web of politics: The internet's impact on the American political system.* Oxford, UK: Oxford University Press.

Davis, K. (2012). *Ethics of Big Data.* O'Reilly Media, Inc.

De Choudhury, M.; Lin, Y. R.; Sundaram, H.; Candan, K. S.; Xie, L. y Kelliher, A. (2010). "How does the data sampling strategy impact the discovery of information diffusion in social media". En *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media:* 34–41. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1521/1832.

De Landtsheer, C., Krasnoboka, N. y Neuner, C. (2001). "La facilidad de utilización de los websites de partidos políticos. Estudio de algunos países de Europa del Este y Occidental" [The facility of use of websites of political parties. Study of some countries of Western and East Europe]. *Cuadernos de Información y Comunicación* (CIC), 6: 107-140.

Dean, J. y Sanjay Ghemawat (2008). "MapReduce". *Communications of the ACM 51* (enero 1): 107. doi:10.1145/1327452.1327492.

Della Porta, D. y Mosca, L. (2006). *Report on WP2 – Searching the net. Project Democracy In Europe and the mobilization of society.* Retrieved from http://demos.eui.eu

Della Porta, D. y Mosca, L. (2009). "Searching the net. Web sites' qualities in the Global Justice Movement". Information, *Communication & Society,* 12, (6): 771 - 792.

Diamond, L.. y Morlino, L. (2004). "The Quality of Democracy. An Overview". *Journal of Democracy,* 15(4): 20-31.

Diani, M. (2002). *Network Analysis. In B. Klandermans and S. Staggenborg. Methods of Social Movement Research.* Minneapolis: The University of Minnesota Press.

Diani, M. (2004). "Cities in the World: Local Civil Society and Global Issues in Britain". En D. Della Porta and S. Tarrow (Eds.). (2004). *Transnational protest and global activism.* Lanham (MD): Rowman & Littlefield.

Donk, W. (2004). *Cyberprotest: New Media, Citizens, and Social Movements.* London: Routledge.

Earl, J. (2010). "The Dynamics of Protest-Related Diffusion on the Web". *Information, Communication & Society* 13 (2): 209-225. doi:10.1080/13691180902934170.

Eurostat (2011). "Eurostat-Information Society Statistics". http://epp.eurostat.ec.europa.eu/portal/page/portal/information_society/data/main_tables.

Fundación Telefónica (2013). *La Sociedad de la Información en España 2012.* Madrid: Fundación Telefónica. http://e-libros.fundacion.telefonica.com/sie12/.

Fuster Morell, M. (2005). *El activismo asociativo pro-wifi en el Estado Español* [Pro-wiffi associacionism activism in the Spanish State]. Archivo Observatorio para la CiberSociedad. Retrieved from http://www.cibersociedad.net/archivo/articulo.php?art=210

Fuster Morell, M. (2007). *Strumenti tecno-politici. In Transform! Italia, Parole di una nuova politica.* Roma: Edizioni XL: 113 - 121.

Fuster Morell, M. (2010). *Governance of online creation communities: Provision of infrastructure for the building of digital commons.* (Unpublished dissertation). Florence: European University Institute.

Fuster Morell, M. (2011). "Advantages, Challenges and New Frontiers in Using Information Communication Technologies in Societal and Social Movement Research". *tripleC: Communication, Capitalism & Critique. Open Access Journal for a Global Sustainable Information Socie*ty 9 (2): 632–643.

Gibson, R.; Nixon, P. y Ward, S. (Eds.) (2003). *Political Parties and the Internet. Net gain?* New-York and Londres: Routledge.

Gibson, R. K. (2009). "New Media and the Revitalization of Politics". *Representation* 45 (3): 289-299. doi:10.1080/00344890903129566.

Ginsberg, J.; Mohebbi, M.; Patel, R.; Brammer, L.; Smolinski, M. y Brilliant, L. (2009). "Detecting Influenza Epidemics Using Search Engine Query Data". *Nature* 457 (7232) (febrero 19): 1012-1014. doi:10.1038/nature07634.

Gonzalez-Bailon, S. (2013). "Online Social Networks and Bottom-Up Politics". SSRN *Scholarly Paper* ID 2246663. Rochester, NY: Social Science Research Network. http://papers.ssrn.com/abstract=2246663.

Gonzalez-Bailon, S.; Wang, N.; Rivero, A.; Borge-Holthoefer, J. y Moreno, Y. (2012). "Assessing the Bias in Communication Networks Sampled from Twitter". SSRN *eLibrary* (diciembre 4). http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2185134.

Granovetter, M. (1983). "The strength of weak ties: A network theory revisited". *Sociological theory* 1: 201–233.

Hanneman, R. A. y M. Riddle (2005). *Introduction to social network methods*. Riverside, CA. USA: University of California Press. http://faculty.ucr.edu/~hanneman/.

Hansen, D.; Shneiderman, B. y Smith, M. A. (2010). *Analyzing Social Media Networks with NodeXL: Insights from a Connected World*. Morgan Kaufmann.

Hine, C. (2000). *Virtual ethnography*. Sage. http://books.google.es/books?hl=es&lr=&id=X5w1P2_jMNYC&oi=fnd&pg=PP9&dq=Hine+2000&ots=IjVvGz-QIt&sig=eZRbnrGGy-EcPvhn6rKSKA28VYk, ed. 2005. Virtual Methods: Issues in Social Research on the Internet. Oxford, UK: Berg.

IAB Spain y Elogia (2013). "IV estudio anual Redes Sociales | IAB Spain". *IV. Madrid: IAB*. http://www.iabspain.net/redes-sociales/.

Jenkins, H. (2004). "The cultural logic of media convergence". *International journal of cultural studies* 7 (1): 33–43.

Johns, M.; Chen, S. y Hall, G. (2004). *Online Social Research: Methods, Issues & Ethics*. New York: P. Lang.

Kavada, A. (2006). "The 'alter-globalization movement' and the Internet: A case study of communication networks and collective action". Paper presented at the *Cortona Colloquium 2006–Cultural Conflicts, Social Movements and New Rights: A European Challenge*. Cortona, Italy.

Kavada, A. (2007). "Email lists as multiple sites of identity construction: The case of the London 2004 European Social Forum". Paper prepared for the *Symposium Changing politics through digital networks: The role of ICTs in the formation of new social and political actors and actions*. Florence, Italy.

Kivits, J. (2005). "Online interviewing and the research relationship". En C. Hine (ed). *Virtual Methods: Issues in Social Research on the Internet*. Oxford, UK: Berg.

Kleinman, S. (2004). "Researching OURNET: A case study of a multiple methods approach". En M. D. Johns, S. S. Chen, and G. J. Hall (Eds.). *Online Social Research: Methods, Issues & Ethics*. New York: Peter Lang Publishing Inc.

Koopmans, R., y Zimmermann, A. (2007). "Visibility and communication networks on the Internet: The role of search engines and hyperlinks". En C. De Vrees y H. Schmidt (Eds.). *A European public sphere: How much of it do we have and how much do we need*. Mannheim, Germany: Connex: 213-264.

Luis Fer Mtz. (2010). "Twitter API *versus* Twitter Firehose cual es mejor?" http://www.dosensocial.com/2010/12/13/twitter-firehose-vs-twitter-api-las-diferencias-que-debes-conocer/

Manovich, L. (2011). "Trending: the promises and the challenges of big social data". *Debates in the digital humanities:* 460-75.

Marres, N. y Rogers, R. (2008). "Subsuming the ground: how local realities of the Fergana Valley, the Narmada Dams and the BTC pipeline are put to use on the Web". *Economy and Society* 37 (2): 251–281.

Mayer-Schönberger, V. (2011). *Delete: The Virtue of Forgetting in the Digital Age*. Princeton University Press.

McAdam, D. (1983). "Tactical Innovation and the Pace of Insurgency". *American Sociological Review* 48 (6) (diciembre 1): 735-754. doi:10.2307/2095322.

McLuhan, M. (1996). *Comprender los medios de comunicación: las extensiones del ser humano*. Editorial Paidós.

Morlino, L. (2004). "What is a 'good' democracy?" *Democratization,* 11(5): 10-32.

Morozov, E. (2012). *The net delusion: The dark side of internet freedom*. PublicAffairs.

Morstatter, F.; Liu, H.; Kathleen, M. y Pfeffer, J. (2013). "Is the Sample Good Enough? Comparing Twitter's Streaming API with Twitter's Firehose | Follow the Crowd". En Massachusetts, USA. http://crowdresearch.org/blog/?p=6596&utm_source=feedburner&utm_medium=email&utm_campaign=Feed%3A+FollowTheCrowd+%28Follow+the+Crowd%29.

Munck, G. L. y Verkuilen, J. (2002). "Conceptualizing and Measuring Democracy. Evaluating Alternative Indices". *Comparative Political Studies* 35(1): 5-34.

Navarria, G. (2007). "Reflections on beppegrillo.it: A successful attempt of innovation and active promotion of political participation through the web?" Paper prepared for the *4th ECPR General Conference,* Pisa (Italy), 6-8 September, 2007. Section: "Emerging Patterns of Collective Action" Panel: "The use of ICTs for innovative forms of participation".

Norris, P. (2003). "Preaching to the converted? Pluralism, Participation and Party Websites". *Party Politics,* 9 (1), 21-45.

Norris, P. (2001). *Digital Divide: Civic Engagement, Information Poverty, and the Internet Worldwide*. Cambridge: Cambridge University Press.

O'Reilly, T. (2005, September 20). *What is Web 2.0? Design patters and business models for the next generation of software*. Retrieved from http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html

Ocampo, J. A. y Stiglitz, J. E. (2008). *Capital market liberalization and development*. OUP Oxford.

Oliver, P. E., y Myers, D. J. (1998). "Diffusion Models of Cycles of Protest as a Theory of Social Movements". *National Defense University.* http://www.ssc.wisc.edu/~oliver/PROTESTS/ArticleCopies/isaf.pdf.

ONTSI (2013). *Perfil sociodemográfico de los internautas (datos INE 2012) | ONTSI.* Madrid: ONTSI INE. http://www.ontsi.red.es/ontsi/es/estudios-informes/perfil-sociodemogr%C3%A1fico-de-los-internautas-datos-ine-2012.

Pang, B.; Lee, L. y Vaithyanathan, S. (2002). "Thumbs up?: sentiment classification using machine learning techniques". En *Proceedings of the ACL-02 conference on Empirical methods in natural language processing*-Volume 10: 79–86. http://dl.acm.org/citation.cfm?id=1118704

Pardo, P. (2010). "Las máquinas que controlan la economía | Mundo | elmundo.es", *El Mundo* edición. http://www.elmundo.es/elmundo/2010/12/29/internacional/1293605644.html.

Passy, F. (2003). "Social networks matter. But how?" *Social movements and networks: Relational approaches to collective action:* 21-48.

Pew Research Center (2013). "Twitter Reaction to Events Often at Odds with Overall Public Opinion". Pew Research Center. 04. http://www.pewresearch.org/2013/03/04/twitter-reaction-to-events-often-at-odds-with-overall-public-opinion/.

Reagle, J., Jr. (2005). *Do as I do: Leadership in the Wikipedia.* Retrieved from: http://reagle.org/joseph/2005/ethno/leadership.html

Reshef, David N., Yakir A. Reshef, Hilary K. Finucane, Sharon R. Grossman, Gilean McVean, Peter J. Turnbaugh, Eric S. Lander, Michael Mitzenmacher, y Pardis C. Sabeti. (2011). "Detecting Novel Associations in Large Datasets". *Science* (New York, N.y.) 334 (6062) (diciembre 16): 1518-1524. doi:10.1126/science.1205438.

Rogers, E. M. (2010). *Diffusion of Innovations,* 4th Edition. Simon and Schuster.

Rogers, R. (2008). "The politics of web space". http://www.google.es/search?q=the+politics+of+web+space&ie=utf-8&oe=utf-8&aq=t&rls=org.mozilla:es-ES:official&client=firefox-a.. 2013. Digital methods. MIT press. USA.

Rogers, R. (2009). *The End of the Virtual Digital Methods.* University of Amsterdam. http://www.govcom.org/publications/full_list/oratie_Rogers_2009_preprint.pdf.

Rogers, R. (2013). *Digital methods.* USA, Cambridge: MIT press.

Romero, D.; W. Galuba, S.; Asur, S. y Huberman, B. (2011). "Influence and passivity in social media". *Machine Learning and Knowledge Discovery in Databases:* 18-33.

Römmele, A. (2003). "Political parties, party communication and new information and communication technologies". *Party Politics,* 9: 7-20.

Rutter y Smith (2005). "Ethnographic Presence in a Nebulous Setting". En Christine Hine (ed). *Virtual Methods: Issues in Social Research on the Internet.* Oxford, UK: Berg: 81-92.

Sánchez, C.M. (2013). "Los pistoleros de Wall Street». mayo 26.
http://www.finanzas.com/xl-semanal/magazine/20130526/pistoleros-wall-street-5471.html.

Skocpol, T. (2004). *Diminished democracy: from membership to management in American civic life*. University of Oklahoma Press.

Sudulich, M. L. (2006). "ICT and SMO: something new?" Paper presented at the *Cortona Colloquium 2006 – Cultural Conflicts, Social Movements and New Rights: A European Challenge"*, 20-22 October 2006, Cortona, Italy.

Tarrow, S., y Della Porta, D. (2005). "Transnational Protest and Social Activism: An Introducction". En Donatella della Porta y Sidney G Tarrow (ed). *Transnational Protest and Global Activism*. Lanham [etc.]: Rowman & Littlefield Publishers: 1-20.

The Cocktail Analysis (2012). "4º Oleada Observatorio de Redes Sociales" abril 9.
http://www.slideshare.net/TCAnalysis/4-oleada-observatorio-de-redes-sociales.

Thelwall (2009). *Introduction to webometrics quantitative Web research for the social sciences*. [San Rafael, Calif.]: Morgan & Claypool Publishers,.

Trechsel, A.; Kies, R.; Mendez, F. y Schmitter, P. (2003). *Evaluation of the use of new technologies in order to facilitate democracy in Europe: E-democratizing the parliaments and parties of Europe*. Retrieved from:
http://www.erepresentative.org/docs/6_Main_Report_eDemocracy-inEurope-2004.pdf.

Van Aelst, P. y Walgrave, S. (2004). "New Media, new movements? The role of the Internet in shaping the anti globalization movement". En Van De Donk, W., Loader, B., Rucht, D., y Nixon, P. *Cyberprotest: New Media, Citizens and Social Movements*. London: Routledge.

Vedres, B.; Bruszt, L. y Stark, D. (2005a). "Organizing technologies: Genre forms of online civic association in Eastern Europe". *The Annals of the American Academy of Political and Social Science* 597: 171-188. doi: 10.1177/0002716204270504

Vedres, B.; Bruszt, L. y Stark, D. (2005b). "Shaping the web of civic participation: Civil society websites in Eastern Europe". *The Journal of Public Policy,* 25: 149-163.

Wasserman, S. y Faust, K. (1994). *Social network analysis: methods and applications*. Cambridge University Press.

Watts, D. J. (2003). *Six Degrees: The Science of a Connected Age*. New York [etc.]: W.W. Norton.

Weng, L.; Flammini, A.; Vespignani, A. y Menczer, F. (2012). "Competition among Memes in a World with Limited Attention". *Scientific Reports* 2 (marzo 29).
doi:10.1038/srep00335. http://www.nature.com/srep/2012/120329/srep00335/full/srep00335.html.

Wimmer y Dominick (2010). *Mass Media Research, International Edition*. 9th Revised edition. Wadworth.

Witten, Ian H. y Eibe Frank (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.
http://books.google.es/books?hl=es&lr=&id=QTnOcZJzlUoC&oi=fnd&pg=PR17&dq=types+of+machine+learning&ots=3goCbqVhQa&sig=wN9IfiNL7a1gMjTEjbiEGjZn6as.

Xie, J., S. Sreenivasan, G. Korniss, W. Zhang, C. Lim, y B. K. Szymanski (2011). "Social consensus through the influence of committed minorities". *Physical Review* E 84 (1) (julio 22): 011130. doi:10.1103/PhysRevE.84.011130.

Zhu, J. (2013). "Big data for social science research". *Research Methodolgy* University of Honk Kong College conference Room.

Zikopoulos, P. y Eaton, C. (2011). *Understanding big data: Analytics for enterprise class hadoop and streaming data*. McGraw-Hill Osborne Media.