# Big Data, Collection of (Social Media, Harvesting)

HAI LIANG
*The Chinese University of Hong Kong, Hong Kong*
JONATHAN J. H. ZHU
*City University of Hong Kong, Hong Kong*

Social media (such as blogs, online social networks, microblogs) has become one of the major data sources for quantitative communication research over the past decade. In addition to survey, experiment, and content analysis, social media harvesting is the fourth primary method of data collection in social sciences. The increasing social media use has generated rich information online that is archived instantly on web servers. Although much of this information is often very useful, it is rarely in a form researchers can use directly. Social media harvesting is a way to collect data from social media platforms unobtrusively and automatically. By employing simple programing tools, researchers can extract relevant messages from social media platforms for various research purposes.

Harvesting social media data differs from traditional communication methods (e.g., survey, experiment, and content analysis) in several ways. First, it is unobtrusive in nature. That means researchers are not actively intervening in the data collection process. It is a method of collecting and studying social behavior without affecting it. Therefore, harvested data is more objective and reliable than that obtained from surveys and experiments. Yet, the disadvantage is that it is hard to know the demographic and psychological variables of social media users.

Second, harvesting social media is an automatic process. Although some traditional methods, such as content analysis and historical research, are also unobtrusive, they usually require human manual coding. Differently, a harvesting approach employs computer programs to collect and code data automatically. Therefore, using a harvesting approach will save effort, time, and money. Moreover, it can handle large-scale data sources more easily than do traditional methods. However, the disadvantage is that computer programs are not as smart as human beings, which can lead to systematic errors.

Third, social media data is usually fine grained, real time, and on a global scale. Collecting data unobtrusively and automatically is not new for social scientists. For example, firms track purchases of customers and banks collect massive data from credit card transactions. Social media harvesting stands out for the macroscopic global scale and microscopic behavioral extensiveness of the data (Golder & Macy, 2014). On social media, all behaviors (e.g., every click and key press) on social media

*The International Encyclopedia of Communication Research Methods*. Jörg Matthes (General Editor), Christine S. Davis and Robert F. Potter (Associate Editors).
© 2017 John Wiley & Sons, Inc. Published 2017 by John Wiley & Sons, Inc.
DOI: 10.1002/9781118901731.iecrm0015

platforms are recorded, never forgotten, and available for collecting and analyzing in real time.

## Types of data

In general, three kinds of data are available for harvest on social media platforms. The first type is content data. It could be user comments on web forums, user profiles on social networking sites (SNS), news articles on news-sharing websites, photos on photo-sharing websites, videos on video-sharing websites, and so forth. The format of content data could be structured and unstructured. The structured data, like data tables organized in rows and columns, can be analyzed directly. Unfortunately, most data on social media is unstructured in the original form. Researchers need to use text mining techniques to preprocess raw texts and use image and video processing techniques to preprocess photos and videos for further analyses. Current communication studies are more focused on texts than videos and images.

The second type is behavior data. Users read online news; comment on posts, blogs, and videos; write reviews for products; listen to music; and watch videos, among many other daily behaviors that are recorded on social media with precise timestamps. Online behaviors can be classified in three categories: (i) User–entity behavior refers to the interaction between a user and an entity in social media, such as sharing a news article on Facebook. This data is very useful for communication scholars because it is directly related to media use and audience analysis in traditional communication research. (ii) User–community behavior refers to the interaction between a user and an online community, such as participating in a political discussion forum. (iii) User–user behavior refers to the interaction between two individual users, such as sending private messages. The last category is more well-known as structural data.

The third type is network structure data, that is, visual or hidden hyperlinks among users and/or content. An advantage of social media harvesting is that it is plausible to obtain a whole social graph (i.e., network) about the underlying user/content population from social media. It makes sophisticated social network analysis possible and popular in recent communication studies. Before the emergence of social media, social network analysis has been limited to very small groups by requirements of direct observation of interpersonal interactions. Now, it is much easier to obtain detailed measures of network structure and network processes at the population level. The simplest structural data comes from the hyperlinks among the web pages. On SNS, following relationships among users naturally construct a social graph. Besides, communication scholars have analyzed the network structures of individual discussions on politics, health, among other topics.

## Sampling strategies: selecting seed(s)

Sampling seeds refer to the initial cases of a sample under study. Sampling for traditional communication research (e.g., survey or content analysis) usually does not require seeds

because the samples are small enough to be collected in one batch. In the era of Big Data, some people have argued that there is no need for sampling anymore given the availability of population data, while other scholars believe that sampling is still not only necessary but also desirable for most researchers of social media who usually do not have access to the population data and/or do not have the capability to handle the massive population data. Therefore, it is important to plan ahead which seeds to target and how to obtain them.

There are in general two strategies for selecting sampling seeds for harvesting social media data: *convenience-based* seeds versus *randomly selected* seeds. A convenience-based approach selects a few users or content pages that are easy to locate, for example, from the homepage of the targeted website, or from search results based on targeted keywords. Once the seeds are located, additional users or webpages will then be followed through successively based on the existing hyperlinks among them, which is traditonally known as "snowballing" in social network analysis or "breadth-first search" in information retrieval. This approach is obviously easy to implement and fast to finish. As such, it has been widely used in harvesting social media data. However, convenience samples are known to involve selection biases. In particular, when sampling social media data, convenience-based seeds are likely to lead to nodes (i.e., users or webpages) that are more active, visible, and connected. Therefore, communication scholars are encouraged to use the random-selection approach that is more difficult to carry out but less subject to selection biases.

Of the three types of social media data defined earlier, sampling of content data and user behavior data follows the same principles and techniques whereas sampling of networked data involves completely different issues.

### *Harvesting content and behavior data*

Random selection of seeds requires a *sampling frame*, which is an operational version of the population (i.e., given social media) under study. In traditional survey or content analysis, sampling frames are often readily available from physically existing records, for example, roster of registered voters, directory of residential telephones, and date of publications. Even if such records do not exist, it is quite easy to construct one based on relevant information, for example, devising a list of housing units based on a map or a list of random telephone numbers based on knowledge of the telephone service. These methods remain largely applicable to constructing sampling frames for the online populations (i.e., users or webpages) under study. However, various technical features of social media platforms (e.g., massive size, hierarchical structure) do present new challenges. It is often necessary and beneficial to combine manual inspections and computerized detection to work around these issues.

One of the methods using this combined strategy is random digit search (RDS), for identifying sampling frames and selecting sampling seeds for social media platforms (Zhu, Mo, Wang, & Lu, 2011). RDS is in fact an extension of the traditional random digit dialing (RDD) method that has been widely used in social science telephone surveys. RDD uses the known range of available telephone numbers in an area as the sampling frame (e.g., from area code-prefix-0000 to area code-prefix-9999), from which a set

of random numbers are generated for dialing. Of course, some (or many) of the random numbers may not exist at all. However, the method works well in the absence of telephone directories and, more importantly, avoids many self-selection biases (e.g., unlisted numbers) in telephone directions. The RDD method has been widely applied to construct sampling frames for a variety of social media platforms ranging from blogsites (e.g., blog.sina.com.cn, see Zhu et al., 2011), social networks (e.g., twitter.com, see Liang & Fu, 2015), which use sequential numbers (plus characters sometimes) as user IDs. Different from RDD in which the range of potential telephone numbers is easy to detect, RDS usually involves multiple rounds of manual examination of the targeted website and computerized search (hence the name "random digit search") to determine the boundaries of a sampling frame for the website and to optimize the initial sample size (which is usually a dozen or even 100,000 times the final sample size), the weights assigned to different bins (i.e., subsets) of the sampling frame (which is often necessary to account for potential selection biases across the frame), and other key decision matters for sampling.

Experimental results clearly demonstrate drastic discrepancies in key social media metrics (e.g., number of posts, comments, views) between convenience-based samples and random samples based on RDS, with users in convenience samples posting 5+ times more, receiving 400+ times more comments, and 10,000+ times more viewing (Zhu et al., 2011). Nevertheless, RDS is limited to social media websites where sequential user/page IDs are used (which we believe to be true for most, if not all, cases) and can be detected (which is sometimes difficult or even impossible). When sequential IDs are not available, one may consider using search keywords, posting timelines, IP addresses, and other information of systematic patterns to form sampling frames. However, little empirical experiment has been reported to show the feasibility and quality of these approaches.

### *Harvesting network structure data*

As discussed earlier, network structure data refer to explicit links (e.g., hyperlinks) or implicit links (e.g., cooccurrences) among nodes (i.e., users or pages). A variety of sampling methods have been employed to study the structure of social media. Despite different names used, the methods come essentially from three branches of the sampling family (see Figure 1): probability (or uniform) sampling, breadth-first search (BFS) sampling, and random walk (RW) sampling. However, the methods differ on how the initial seeds are identified and how the subsequent nodes are followed through.

*Uniform* sampling is the same as the random sampling of nodes in the previous section. As such, it requires a sampling frame in advance, which can often be created using RDS or other methods (e.g., timeline-based, search query-based). With a carefully constructed sampling frame, even the term "uniform" (i.e., equal weight to each node) can be relaxed with a weighting scheme based on information of distribution density of node IDs in the sampling frame. As mentioned earlier, construction of a representative sampling frame is time consuming. However, once the sampling frame is in place, uniform sampling of the remaining nodes is fast because multiple spiders (i.e., search program) can be deployed simultaneously from different machines.

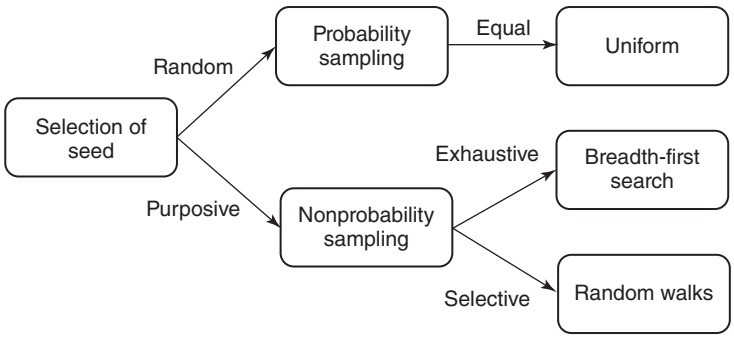


**Figure 1** A family tree of network sampling methods.

BFS sampling is evidently a convenience-based method as it starts with seeds that are purposively or conveniently selected (e.g., from the homepage of a targeted website, the list of top-*n* users, top-*n* posts, top-*n* keywords), rather than randomly selected from a sampling frame representing the entire nodes. Once started, BFS exhausts all explict links from the seeds until a predetermined sample size is reached, which makes the approach extremely efficient, for example, reaching 90%+ of Facebook, Twitter, and other global giant social networks in about a dozen steps. If the resulting sample size approximates the online population under study, BFS is of course the best option to adopt because it delivers both efficiency and quality. Unfortunately, it is not a practical option for most researchers in communication.

RW shares the same strategy as BFS in selecting seeds. As such, RW is a member of the nonprobability sampling family, despite the word "random" in its name. However, there is indeed a random element in the subseqent steps of RW. Rather than exhausting all nodes linked directly or indirectly by the seeds as in BFS, RW usually follows only one neighbor from each of the seeds, successively doing so until reaching the target sample size. More importantly, the subsequent nodes are picked based on a simple random method (hence the name "random walk"). A variety of weighting schemes, such as metropolis-hasting random walk (MHRW), reweighted random walk (RWRW), forest fire (RF), biased random walk (BRW), and self-adjustable random walk (SARW), have been developed to improve the simple random walk method (see Xu & Zhu, 2016 for a review). All variants of the RW method result in samples with significantly better quality than BFS, but are markedly slower than BFS and uniform sampling.

Table 1 summarizes key features across the three branches of the network sampling family. The most important message appears in the last panel (estimated properties of the sample) which was based on a series of experiments we carried out on a real large social network with 10+ million users and 200+ links (Xu & Zhu, 2016). Node properties refer to the commonly used metrics of user behavior and user-generated content (e.g., number of articles posted, number of comments received, number of clicks/views received). As such, these are metrics of user behavior data or content data rather than network structure data. As discussed earlier, uniform sampling produces unbiased results whereas convenience sampling (both BFS and RW) do not. However, all three approaches perform unsatisfactorily on network properties as measured by mean degree (i.e., number of neighbors that measure the density of a network),

**Table 1** Comparison across prevailing sampling methods.

| | *Uniform sampling* | *Breadth-first search* | *Random walk* |
|---|---|---|---|
| Prior knowledge required | UIDs of all nodes | UIDs of all nodes | UIDs of all nodes |
| Probability to sample each node | Equal for all nodes of the network | Equal for all nodes of the seed's ego-network | Varying for all nodes of the seed's ego-network |
| No. of steps to arrive at a desired sample[a] | | | |
| • $10K$ nodes | 1 | 3 | 7 (if $n = 5$)[b] |
| • $1M$ nodes | 1 | 4 | 9 (if $n = 5$)[b] |
| Estimated sample characteristics | | | |
| • Node properties | Unbiased | Biased | Biased |
| • Network properties | Biased | Biased | Partially biased |

[a] Assuming that one seed is used for both BFS and RW sampling and the mean degree is 100.

[b] $n$ is the number of nodes sampled in each step, calculated by $N = \frac{\log(n^q - 1)}{\log(n-1)}$, where $N$ is the desired sample size and $q$ the number of steps sampled.

clustering coefficient (i.e., fraction of triad closure of a network), and assortativity (i.e., correlation between pairs of connected nodes to measure mixing patterns of a network). Why does uniform sampling fail the test? Simply because it (and all other classic probability methods) assumes independence among sampling units (i.e., nodes), which is no longer true for networked data.

In short, sampling of networked data has remained an open challenge with no perfect solution in sight. The best advice we could offer at this stage is to be cautious about estimated network properties based on samples, no matter what the sampling method or sample size.

## Web harvesting

There are three ways to obtain large-scale data on social media platforms. First, the most direct way is to download the data from databases on the web servers. This kind of data is known as "log-file" data. Log-file data is unique for getting behavioral data, such as log-in information, browsing history, which are not visible on web pages. However, it is nearly impossible to get this kind of data unless researchers have close collaborations with the social media companies.

The second method is collecting data through application programming interfaces (APIs). An API is basically an interface of a computer program that allows the software to interact with other software. It is a small script file written by users, following the rules specified by the web owner, to download data from its database other than web pages. To put it simply and informally, APIs are special URLs that web owners intentionally provided for developers to download data from their databases.

As part of their business model, social media companies often make their APIs available to third parties. The primary purpose of providing APIs is to enable the development and enhancement of social media services. However, the API is also an interface for researchers to collect data from social media platforms. Through small software scripts, researchers can access the API to retrieve, store, and manipulate digital traces left by the users of a service for further empirical analysis.

The third method is web scraping. This method is particularly useful for those websites that do not provide APIs. Where data available on the website are not available through the API, an alternative method is to crawl the social media website with an automated script that explores the website and collects data using HTTP requests and responses. Web scraping is the process of taking unstructured information from web pages and turning it into structured information that can be used in a subsequent stage of analysis.

Web scraping involves writing computer programs to do automatically what people do manually when they select, copy, and paste. The process is similar to the "save as" function when people are browsing web pages. They then parse the web pages, extract the useful information, and organize the data in csv or txt format, which can be imported to other statistics software.

Theoretically, web scraping does not require any help from the social media companies. Researchers can scrape everything that is visible on the web pages. In other words,

scraping interacts with web pages to obtain data. Yet, both log-file and API approaches interact directly with the database on web servers. In this sense, web scraping might be less robust, less efficient, and less informative than the API method.

### *Via APIs*

The most popular way to collect social media data today is to use the API. Relevant queries are sent to the social media with the API to collect large-scale data. Three technical issues are involved in the process of data collection via APIs: authorization, generating API links, and parsing JSON (JavaScript Object Notation) data. In plain language, log in, submit data requests, and save the responses in structured data tables.

Before making any API requests, many social media platforms (including Facebook and Twitter) require a formal procedure of authorization. Researchers need to create an application on their websites and authorize the application to access relevant data using a standardized protocol called Open Authorization (OAuth). It is similar to the log-in function for general browsing purposes. The protocol is a social web standard at this point. Both Facebook and Twitter use this protocol.

Put less technically, OAuth is a means of allowing users to authorize third-party applications to access their account data without needing to share sensitive information like a password. Russell (2013) provides a slightly broader overview of how OAuth works, and Twitter's OAuth documentation offers specific details about its particular implementation. For simplicity of development, the key pieces of information that researchers need to take away from the application's settings are its consumer key, consumer secret, access token, and access token secret. These four credentials provide everything that an application would ultimately be getting to authorize itself through a series of redirects involving the user granting authorization, so they should be treated with the same sensitivity that you would give to a password.

How to interact with APIs to access data? The most popular social media providing API services also have an API document online. The API document lists all technical details about how researchers can download different types of data. With small variations, most platforms provide their APIs in similar ways. For example, clicking this URL (http://www.reddit.com/user/screenname/comments/) will direct you to all comments posted by the author named "screenname" on www.reddit.com. The corresponding API link is very intuitive: http://www.reddit.com/user/screenname/comments/.json. The only difference between the original URL and the API link is the postfix ".json", which means that this link provides all data in the original page in the JSON format.

Similarly, the normal URL for a typical Facebook page looks like https://www.facebook.com/username, while the corresponding API link is https://graph.facebook.com/username?fields=id,name&access_token=ACCESS_TOKEN. The ACCESS_TOKEN was generated via OAuth as a courtesy for the logged-in user. The normal URL for Twitter search is https://twitter.com/search?q=%40twitterapi (the search keyword is "twitterapi"), while the equivalent API link is https://api.twitter.com/1.1/search/tweets.json?q=%40twitterapi. The Twitter API link is not clickable like the Facebook and Reddit ones. It requires more complicated URL requesting methods for authentication.

Fortunately, many existing tools can complete this task easily as listed at the end of this entry.

Instead of visiting www.twitter.com (or www.facebook.com) like normal users, data harvesters visit api.twitter.com (graph.facebook.com). So, the first task of data collection using APIs is to find the rules for generating this kind of API link. Then, by requesting API links in web browsers or via programing scripts, it will return the data usually in the JSON format. Once the JSON data has been obtained, the second issue is how to parse it automatically. JSON is a compact, text-based format for computers to exchange data. It can be very easily processed by popular programming and statistical languages, such as R and Python. It is loaded once into Python just like a dictionary (or a list in R). Simply save the JSON data in txt files, and then use software to convert them into data tables.

Using public APIs can obtain rich information for communication research. A research strategy that is currently in vogue is the use of APIs to extract large amounts of behavioral data. Several kinds of behavioral data could be accessed via APIs. First, the distribution and frequency of content generation on social media platforms are provided. In marketing research, it has been argued that 20% users have produced 80% messages on the Internet. The distribution of message posting is highly skewed. It is straightforward to examine this claim since most social media APIs provide some summary statistics about how many messages a user has posted. Research on Twitter has found that 2% users produced more than 80% messages and more than half of users did not post any. In political communication research, the unequal distribution has been interpreted as unequal political participation—few users are more active in voicing online.

Another important behavioral data in communication research is information-sharing behavior, such as the retweeting behavior on Twitter. Retweet is an effective means to relay diffusion information beyond adjacent neighbors. Twitter APIs explicitly provide retweeting information by indicating the retweeted (original) tweet ID. Researchers treated retweet trees as communication channels of information diffusion and observed that retweets reach a large audience and spread fast on Twitter (Kwak, Lee, Park, & Moon, 2010). The number of retweeting/sharing times has also been used as an indicator of popularity of online messages, especially for measuring marketing effects.

It is also very easy to get the accurate time stamp for each post via API, whereas it would be very unstable via web scraping. Time stamps can be used to investigate the circadian rhythms and the sustainability of using social media platforms. APIs always return time stamps in standard forms (e.g., Greenwich Mean Time/GMT on Twitter) as well as time differences for local time adjustment. Using this information, researchers have found that posting activity rises in the morning and increases throughout the day until the evening on Twitter and Facebook. Moreover, weekend use shows a much lower activity and less distinct time of day patterns (Golder, Wilkinson, & Huberman, 2007).

In addition to behavioral data, large-scale following–follower networks are available via APIs. For example, Twitter API provides full access to follower and following lists of public user accounts, whereas Facebook and LinkedIn are reluctant to provide full

friendship lists due to privacy concerns. Social media following networks are used as important data to analyze the formation of social networks, which would be hard to observe in offline situations. It has been found that popularity, reciprocity, transitivity, and homophily are the key mechanisms in forming online social networks (Golder & Yardi, 2010).

Further, researchers have constructed discussion networks according to the "reply-to" and "mention" relationship and constructed information-sharing networks according to the "retweet" relationship on Twitter. This information was successfully incorporated into political communication to measure political fragmentation and polarization. Researchers found that social media users are more inclined to discuss politics with politically similar others. However, the degree of fragmentation may vary across the nature of network ties and time (Conover et al., 2011). Moreover, network analysis based on API data has been used to visualize thematic threads in the network and to analyze the social connections and the diffusion of tweets. It is particularly valuable as a tool to mine and visualize patterns from large sets of relational data.

The API method is also popular for collecting data for content analysis to examine semantic patterns in social media communication. Using the Twitter Streaming API, a random sample of public tweets that contain rich information for text mining can be extracted. Based on content-based categorization of Twitter messages, researchers have distinguished between different communicative purposes. Based on semantic network analysis, research has investigated the consumption and production patterns of hard versus soft news information (Horan, 2013). The hashtags and URLs embedded in tweets could help researchers to track the paths of news, health, and crisis information diffusion. Other studies have examined the function of specific features of Twitter, for example, by using the @reply sign as a marker of addressivity and interactional coherence between tweets (see a review by Lomborg & Bechmann, 2014).

### *Via web scraping*

In addition to API, web scraping or screen scraping is another common approach to collect web data. The procedure is straightforward. It actually mirrors the process whereby a user views a web page, finds the relevant information, and copy-and-pastes the elements into data tables. However, letting the computer do the tasks automatically is more complicated than copying and pasting, because it requires some basic understanding of the language the web is written in, and a programming language.

*HTML*. The first thing researchers need to know is the language used to write web pages. Web pages are written in Hyper Text Mark-up Language (HTML). Although a researcher will not be producing any HTML web pages, an understanding of the structure of HTML is needed, because HTML elements will be used as signposts along the way to extracting the data on the web.

The basics of HTML are simple. The HTML code is viewable within most popular web browsers, for example, "view source" in Internet Explorer and "inspect elements" in Google Chrome. An HTML element is an individual component of an HTML document or web page, once this has been parsed into the Document Object Model (DOM). HTML is composed of a tree of HTML elements and other nodes, such as text nodes.

Elements can also have content, including other elements and text. Many HTML elements represent semantics or meanings. For example, the <title> … </title> element represents the title of the document.

HTML is composed of elements called tags. Tags are enclosed in left and right-pointing angle brackets, for example <html> … </html>. Some tags are paired, and have opening and closing tags. For example, <html> is an opening tag which appears at the beginning of each HTML document, and </html> is the closing tag which appears at the end of each HTML document. Other tags are unpaired. For example, the <img> tag, used to insert an image, has no corresponding closing tag.

Two tags in particular are worth noting: *span* and *div*. In HTML, *span* and *div* elements are used to define parts of a document so that they are identifiable when no other HTML element is suitable. The tag *span* represents an inline portion of a document, for example, words within a sentence. The tag *div* represents a block-level portion of a document such as a few paragraphs. A full list of HTML tags are available on the w3schools.com website (http://www.w3schools.com/tags/). This reference will help researchers have a better understanding of how a web page is organized.

In addition to the tag, HTML elements can have attributes. Attributes provide additional information about an element. Attributes come in name/value pairs like: name="value." For example, a hyperlink is defined with <a> tag and the link address is specified in the *href* attribute. It normally looks like this: <a href="http://www.example.com">Click here please</a>. The <a> tag encloses the text element—Click here please—and has an attribute with the name of *href*. The value of that *href* attribute is "http://www.example.com," which is the destination URL. Most HTML elements can take any of several common attributes.

For scraping purpose, two attributes are worth pointing out. First, the *id* attribute provides a document-wide unique identifier for an HTML element. It can also be used by scripting languages to reference the element. That means links can be made directly to an element with a specific *id*. Second, the *class* attribute provides a way of classifying similar elements. Any number of elements can have the same value (unlike the *id* attribute). This can be used for semantic purposes, or for presentation purposes. A complete list of the attributes for each HTML element is also available at http://www.w3schools.com/tags/.

In practice, researchers are not required to understand the specific meaning of each tag or attribute. Instead, researchers need to know how to use the combinations of tags and attributes to determine the positions of any relevant items in a web page. If only a single item on a web page is of research interest, using the corresponding *id* attribute will be sufficient to determine the position. If a set of similar items is of interest, using the *class* attribute might be workable in most situations.

*XPath*. The second thing researchers need to know is how to use XPath to extract relevant information from HTML pages by combinations of tags and attributes (see Table 2). XPath can be used to navigate through elements and attributes in Extensible Markup Language (XML) documents. These path expressions look very much like the expressions when working with a traditional computer file system. XML was designed to describe data, whereas HTML was designed to display data on a web page. However, similar to HTML, XML documents are treated as trees of nodes. Therefore,

XPath is applicable for extracting content in HTML documents when they are parsed appropriately.
In Table 2, **class="title may-blank"** is an attribute node; the body element is the parent of the two hyperlink elements; the two element nodes are called siblings. The most useful XPath expressions to select nodes are listed in Table 3.
For example, **//a[@href]** selects all the **a** elements that have an attribute named **href** (i.e., hyperlinks) in Table 1 example; **//a[@href="http://www.cnn.com"]** selects the second hyperlink element (with the attribute value ="http://www.cnn.com"); and **//a[0]/@href** returns the URL (attribute value) of the first hyperlink element.

After introducing HTML and XPath, scraping web pages is straightforward. First, download the HTML pages and save them onto your hard disk. Second, many tools (including common web browsers) are available to parse the HTML documents and create a corresponding structure. Usually, this displays as a tree-like structure, which is very similar to when you click the "view source" icon in web browsers. Third, based on this structure, researchers can use XPath expressions to retrieve relevant messages. For some web browsers, such as Google Chrome, they provide a "copy XPath" function. Select an element on a web page, right click inspect elements, and then click copy XPath. The XPath expression of the selected element is copied. Finally, save the messages in data tables on your disk. Again, all the steps could be done automatically using most computer languages.

In addition to XPath, there are alternative methods to extract HTML elements, such as using CSS (Cascading Style Sheets) selectors, regular expressions and search

**Table 2** Example of an HTML document.

```
<html> (root element node)
<body>
<p>
   <a class="title may-blank" href="http://www.nyu.com"> Li Ka-shing, Asia's richest man
     </a> (element node)
   <a class="title may-blank" href="http://www.cnn.com"> URA Director quits </a>
     (element node)
</p>
</body>
</html>
```

**Table 3** Useful XPath expressions.

| *Expression* | *Description* |
|---|---|
| *nodename* | Selects all nodes with the name "nodename" |
| / | Selects from the root node |
| // | Selects nodes in the document from the current node that match the selection no matter where they are |
| @ | Select attributes |
| * | Matches any element node |

functions by treating HTML documents as raw texts. No matter what, the basic logic is the same—selecting content according to its associated HTML tags and attributes.

Web scraping is the most popular approach to collect hyperlink networks. In a hyperlink network, the nodes are web pages (or a set of web pages, i.e., websites), while ties are the hyperlinks between any two websites. For some time now researchers have been using web crawlers, which are software tools that automatically traverse a website by first retrieving a single web page (e.g., entry or the home page on a site) and then recursively retrieving all web pages that are referenced (e.g., following hyperlinks throughout the site). A web crawler can save a copy of each web page that it encounters, but it can also parse the HTML content, extracting the hyperlinks and text content. In order to automatically extract data from web pages, web crawlers rely on the web pages being written to comply with the HTML specification. Hyperlinks could be easily extracted from HTML codes by XPath. Let *AllLinks* = **//a[@herf]**, where *AllLinks* contains all hyperlink elements. And then, for each *Link* in *AllLinks*, **Link/@herf** will return a URL address.

Web scraping is a good method for collecting web content when APIs are unavailable. This is particularly useful for collecting posts in blogs, web discussion forums, and news web portals. The text content is usually embedded between the paragraph elements (i.e.,  … ). Normally, using XPath **//p** will get all main text paragraphs in a HTML web page. However, scraping online content is less convenient and accurate than collecting via APIs. Even though, the text is certainly in the page, so is a lot of other irrelevant content such as navigation bars, headers, footers, advertisements, and other sources of noise. In addition, HTML pages might be written in quite different structures. Sometimes, the developers may not follow the HTML standards very well. All the problems can increase the harvesting time and cause unexpected errors.

Web scraping is very challenging for collecting behavioral data. The time stamps displayed on web pages are usually less precise and reliable than those obtained from APIs. Sometimes, it only displays a rough time, for example "posted two days ago." Moreover, the screen names could be changed frequently on most social media platforms. Because of this, it is hard to merge multiple screen names used by a single user at different time periods. Even worse, it is nearly impossible to get any information about user browsing, clicking, and log-in behaviors without access to the log files via APIs. This kind of information is totally not observable from web pages. The best practice, therefore, is to use the scraping method only when official APIs are unavailable.

## Ethical issues

A major legal and ethical aspect of web harvesting is the use of content that may be considered private by human subjects. On services such as Facebook, content is technically semiprivate—more than half users set their profiles to friends-only status. On Twitter, content is by default public, though nearly 10% of users have protected their tweets. Legally, even technically public social media data can be personal and sensitive, and must be handled according to privacy laws (Lomborg & Bechmann, 2014).

Harvesting social media data may lead to an unwanted and unforeseen exposure of private data for participants. Large quantities of personal information on social media

are collected and can easily be manipulated and taken out of context (boyd, 2006). This raises a set of key ethics questions. For example, how can we ensure that participants are adequately informed and protected when collecting data? Twitter profiles are publicly available online materials, so does that mean we can conduct any kind of research using this data without permission? Although there are no definitive answers to all ethical questions, two major concerns for human subject research in social science are applicable for social media harvesting: informed consent and anonymization.

### *Informed consent*

The issue of informed consent in social media harvesting has two aspects: one concerns the informed consent that users give to the service providers, and the other concerns the specific use of personal data for academic research. When users sign up for a social media platform, they have to accept the terms of use, which often implies letting the service providers use their information to improve and monetize their product.

Researchers make use of this informed consent when accessing the data through APIs. On Facebook the researcher needs to ask for additional permissions to collect data. Twitter is more generous in this aspect. Even though, Twitter requires that researchers do not share any identifiable information as well as the raw content of public tweets publicly. However, this does not mean that an informed consent agreement with individual users is no longer required. Whether to ask for participants' permission before data harvesting depends on what is being extracted and how the data is going to be used.

First, it depends on environments that are primarily designed for information sharing (e.g., news portals, blog sites, newsgroups, Twitter) and those that are primarily designed for social networking (e.g., social network sites). With information-sharing websites, it is hard to justify that informed consent is always required (Ackland, 2013). For example, news articles published by media organizations and other public institutions do not involve human subjects. Therefore, harvesting this type of messages does not require informed consent for academic purposes.

However, other publicly available archived materials created by individual users, for example, forum posts, tweets, should be used with caution. "Fair use" needs to be determined because of copyright issues involved. One perspective states that if content is being used for academic purposes, then it is unlikely that this violates fair use. Another perspective is that online postings, while public, are written with expectation of privacy and thus informed consent is required.

Most social media platforms have privacy settings to let users regulate their disclosure strategies. To harvest the messages that were explicitly protected, informed consent is certainly required. In this situation, obtaining informed consent agreement has additional merit for data collection. Usually, public APIs do not allow researchers to collect information from protected users. If those users are the research subjects, obtaining permissions directly from the users is the only way.

Second, whether informed consent is required depends on whether the subsequent analyses involve identifiable information or merely aggregated information.

In quantitative studies using large-scale datasets, there is no direct contact between researcher and research participant. Most of these studies are interested in structural analysis, pattern recognition, and prediction. In this situation, it may be appropriate not to seek informed consent, simply because there is a greater distance between the analysis and the actual users involved. In addition, it is infeasible for large-scale studies to gain informed consent from each person. Social media users may have left the platform or have ceased posting but the data are still available online. In this scenario, how data are anonymized both to the researcher and when presenting results becomes critical.

### *Participant anonymity*

In quantitative studies using large-scale data, researchers anonymize their data by default and test units are not publicly searchable. Unlike face-to-face surveys, web harvesting is conducted by computer programs automatically. Therefore, researchers usually do not know who the subjects are. In this sense, "respondents" in web harvesting are truly anonymous, even though identification is possible by additional analysis.

Preserving anonymity can be very difficult when quoting texts in social media research. What is worse, it may be possible to deanonymize individuals, not only through text strings, but also through social graphs (see Zimmer, 2010). Researchers need to take this problem seriously when sharing data publicly. Both Twitter and Facebook have established strict policies on data sharing. For example, according to Twitter, sharing tweets is not allowed, although sharing anonymized graphs is permissible.

The challenges associated with anonymizing data collected via harvesting have implications not only for data sharing among researchers, but also for the kinds of analysis. For example, in network analyses of structural patterns of communications among users, the information and analytic depth that may be contained in disclosing the users behind individual nodes must be weighed against the risk of violating their privacy expectations.

In addition, there are particular ethical issues associated with the use of web scraping (Thelwall & Stuart, 2006). First, scraping a website intensively can potentially use a lot of the resources (e.g., bandwidth) of the website owners, which might lead to loss of service quality. Some universities also have specific rules to limit the frequency of web requests. For this reason it is important that web crawlers should be used responsibly, and also by limiting the crawler so there are delays between each page request. Second, it is important that web crawlers obey the robots.txt protocol, which is used by web masters to inform crawlers which parts of the website can be crawled (Ackland, 2013).

## Tools

Tools for social media harvesting could be summarized according to whether programming knowledge is required and whether it is commercial. In terms of open source programming tools, most social scientists use Python or R languages for web harvesting and analysis. To collect web data via APIs, Python modules, such as "urllib2"

and "requests," could be used for HTTP requests; then, JSON data will be returned. Researchers could use the "json" module to convert JSON to other data formats. In R, researchers can use the "RCurl" package to request web pages, while use the "rjson" package for parsing the returned JSON data.

To collect web data via scraping, Python modules, such as "mechanize" and "scrapy," are used to read HTML pages. The advantages of these packages to the "requests" module are that they are even more user-friendly and have more convenient functions. Researchers then usually use the "BeautifulSoup" module to parse HTML pages and extract relevant content via search functions. In R, the "XML" package is an alternative to the Python "BeautifulSoup" module. Although these tools are sufficient in most situations, they cannot deal with the content written in JavaScript, for example, executing the mouse over and scroll down functions. A powerful tool in Python and R for executing JavaScript is the "Selenium" module. It can be used to write an automated web browser, which exactly imitates the browsing behaviors of human beings.

This is the generic solution for harvesting social media content. Instead of writing codes from the very beginning, some handy modules have been developed for harvesting specific social media platforms. For example, researchers can use the Python "twitter" module and the R "twitteR" library to interact with Twitter APIs. Using the "Facebook Python SDK" and "RFacebook" can extract Facebook data. These packages are user-friendly and can significantly reduce the cost of programming. An even easier way to harvest social media is to use the "pull-down menu" applications. NodeXL is a free Excel plug-in to collect, analyze, and visualize social media data. NodeXL can harvest Flickr, YouTube, Facebook, and Twitter data via APIs.

Besides, there are many commercial tools for social media data harvesting. First, social data vendors are available, such as Gnip and Datasift, which claim that they can provide complete access to Twitter, Tumblr, Foursquare, among other social media platforms. Sometimes, this is the only way to obtain the historical archives. Second, some commercial applications for web scraping are widely used in the marketing and public relations industries, such as Visual Web Ripper and Crimson Hexagon Forsight.

## Acknowledgment

The preparation of the chapter was supported in part by General Research Fund (CityU154412) from Hong Kong Research Grants Council, and Grant Writing Fund (9618003) from City University College of Liberal Arts and Social Sciences.

SEE ALSO: Big Data, Analysis of; Content Analysis, Automatic; Data, Types of; NodeXL; R (Software); Social Network Analysis (Social Media)

## References

Ackland, R. (2013). *Web social science: Concepts, data and tools for social scientists in the digital age*. London: SAGE.

boyd, d. (2006). Friends, friendsters, and myspace top 8: Writing community into being on social network sites. *First Monday, 11*(12). doi:10.5210/fm.v11i12.1418

Conover, M., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., & Flammini, A. (2011). *Political polarization on Twitter*. Paper presented at the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM), Barcelona.

Golder, S. A., & Macy, M. W. (2014). Digital footprints: Opportunities and challenges for online social research. *Annual Review of Sociology, 40*(1), 129–152. doi:10.1146/annurev-soc-071913-043145

Golder, S. A., Wilkinson, D. M., & Huberman, B. A. (2007). Rhythms of social interaction: Messaging within a massive online network. In *Communities and technologies 2007* (pp. 41–66). London: Springer.

Golder, S. A., & Yardi, S. (2010). *Structural predictors of tie formation in Twitter: Transitivity and mutuality*. Paper presented at the 2010 IEEE Second International Conference on Social Computing (SocialCom), Minneapolis, MN.

Horan, T. J. (2013). "Soft" versus "hard" news on microblogging networks: Semantic analysis of Twitter produsage. *Information, Communication & Society, 16*(1), 43–60. doi:10.1080/1369118X.2011.649774

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). *What is Twitter, a social network or a news media?* Paper presented at the Proceedings of the 19th International Conference on World Wide Web, Raleigh, NC, USA.

Liang, H., & Fu, K. W. (2015). Testing propositions derived from Twitter studies: Generalization and replication in computational social science. *PloS ONE, 10*(8), e0134270. doi:10.1371/e0134270

Lomborg, S., & Bechmann, A. (2014). Using APIs for data collection on social media. *The Information Society, 30*(4), 256–265. doi:10.1080/01972243.2014.915276

Russell, M. A. (2013). *Mining the social web: Data mining Facebook, Twitter, LinkedIn, Google+, GitHub, and more*. Sebastopol, CA: O'Reilly.

Thelwall, M., & Stuart, D. (2006). Web crawling ethics revisited: Cost, privacy, and denial of service. *Journal of the American Society for Information Science and Technology, 57*(13), 1771–1779. doi:10.1002/asi.20388

Xu, X. K. & Zhu, J. J. H. (2016). Flexible sampling large-scale social networks by self-adjustable random walk. *Physica A, 463*, 356–365. doi:10.1016/j.physa.2016.07.055

Zhu, J. J. H., Mo, Q., Wang, F., & Lu, H. (2011). A random digit search (RDS) method for sampling of blogs and other web content. *Social Science Computer Review, 29*(3), 327–339. doi:10.1177/0894439310382512

Zimmer, M. (2010). "But the data is already public": On the ethics of research in Facebook. *Ethics and Information Technology, 12*(4), 313–325. doi:10.1007/s10676-010-9227-5

## Further reading

Hanretty, C. (2013). *Scraping the web for arts and humanities*. Retrieved from https://www.essex.ac.uk/ldev/documents/going_digital/scraping_book.pdf (accessed October 16, 2016).

Munzert, S., Rubba, C., Meißner, P., & Nyhuis, D. (2015). *Automated data collection with R: A practical guide to web scraping and text mining*. Chichester, UK: John Wiley & Sons.

Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science, 346*(6213), 1063. doi:10.1126/science.346.6213.1063

**Hai Liang** is an assistant professor in the School of Journalism and Communication, the Chinese University of Hong Kong. His research interests and background cut across science and social science, ranging from computational media studies to political communication and public health. He has published numerous articles in indexed journals such as *Communication Research, Human Communication Research, Journal of Computer-Mediated Communication, New Media & Society, International Journal of Public Opinion Research, Social Science Computer Review*, and *PLoS ONE*.

**Jonathan J. H. Zhu** is a professor of media and communication at City University of Hong Kong. He has conducted research on diffusion, use and impact of the Internet and other social media using traditional and computational social science methods with results published in social science journals such as *New Media & Society*, *Journal of Computer-Mediated Communication*, *Cyberpsychology, Behavior and Social Networking*, *Social Science Computer Review*, as well as science and engineering journals such as *Communications of the AMC*, *IEEE Transactions on Visualization and Computer Graphics*, *Journal of the American Association for Information Science and Technology*, *Nonlinear Dynamics*, *Mathematical Problems in Engineering*, *Physica A*, and *International Journal of Medical Informatics*.