Big Data Analysis

Deen Freelon
American University, Washington DC, USA

In International Encyclopedia of Communication, 2015.

The term "big data" lacks a single widely agreed-upon referent. Commercial interests often define such datasets according to "the three Vs"—velocity, variety, and volume (Laney, 2001)—but velocity is less relevant for communication researchers, who generally do not require real-time data processing capabilities. This invites the question of how much variety and volume are required for a dataset to qualify as "big," which has no context-independent answer. This in turn suggests that the key characteristic of big datasets is not size or variety per se, but rather tractability. In other words, from an analytical perspective, big data is data whose magnitude and complexity pose challenges for established data management practices. Because established practices differ widely between disciplines, more specific definitions (e.g., those that stipulate minimum dataset sizes) should be viewed with skepticism. The current definition implies that a particular discipline's threshold for "big" will increase over time as methodological best practices solidify and catch up to data availability and computing power (→ Research Methods). This has indeed proven to be the case (Jacobs 2009). This entry focuses on those aspects of big data as defined above that are most relevant to the analysis of communication data.

Methodological challenges of big data analysis

To fully understand the challenges big dataset present to communication research, it is essential to briefly review the field's methodological status quo. Quantitative communication scholars work mostly with datasets that number in the hundreds or low thousands, such as completed questionnaires, the results of in-person experiments, or texts for content analysis (→ Quantitative Methodology; Survey; Content Analysis, Quantitative). These can usually be entered into statistical or other analytical software without much difficulty, and where data formats are incompatible, the relatively small scale of the data renders the conversion labor tractable. Data are often analyzed using desktop programs such as Stata, SPSS, Microsoft Excel, Atlas.ti (for qualitative research), and others (→ Qualitative Methodology; Statistics, Descriptive). These programs all offer graphical, menu-based interfaces for ease of use, but most impose restrictions on the amount of data that can be analyzed in a reasonable amount of time using widely available hardware. For example, the latest version of Microsoft Excel as of this writing can accommodate a maximum of 1,048,576 rows of data, and slows down considerably when conducting certain operations on datasets smaller than this maximum. For this reason, some have proposed defining big data as datasets too large for Excel to handle (Ryan 2013), though this is problematic for reasons noted above. In any event, simply perusing datasets of over one million rows becomes difficult or impossible with traditional desktop-based applications, to say nothing of analysis or visualization.

Big datasets offer the communication researcher new analytical possibilities that justify the methodological reforms required to explore them. One of these is the ability to analyze complete datasets rather than samples (→ Sampling, Random; Sampling, Nonrandom). While not all big datasets are necessarily complete, many are (such as a collection of all tweets containing a given keyword that were posted between two dates), and these ensure that nothing of analytical value is omitted (→ Twitter). They also allow researchers to study specific subpopulations in greater detail, whereas a smaller sample might not contain subsamples large enough to be analyzed statistically. Those interested in rare but consequential events or individuals might need very large datasets to be able to isolate their targets. And many varieties of digital data are inherently longitudinal (such as → social media and email

data), facilitating analyses of how individuals, institutions, and social phenomena change over time (→ Facebook; Electronic Mail).

The digital revolution has enlarged the scale and scope of human communication to a much greater degree than any technological shift before it (→ Technology and Communication; Digital Media, History of). Many of the world's most popular digital communication platforms have made some or all of their data available for free in one or more machine-readable formats. Communication researchers interested in this massive trove of data face a number of *obstacles* of varying severity. First, in many cases merely collecting data from publicly accessible sources requires the ability to use and/or write scripts in programming languages such as Python or R. Because computer programming is not a part of communication research's traditional methodological toolkit, this requirement disqualifies many who might want to study certain types of digital communication data. Second, once the data has been collected, it often must be transformed in some way to conduct the desired analysis. For example, both Twitter and Facebook deliver data through their APIs (application programming interfaces) in a format called JSON (JavaScript Object Notation). Analyzing such data often requires converting the JSON into a different format, and this is very difficult to accomplish efficiently without a programming script. Third, many statistical procedures designed for use with small samples are not appropriate for datasets with Ns in the hundreds of thousands or millions. Perhaps the best-known example of this is that statistically significant results become increasingly likely as one's sample size grows. By the time the sample size reaches the millions, statistical significance is nearly omnipresent and therefore analytically useless (→ Statistics, Explanatory). Therefore, the analysis of datasets larger than the field has traditionally been used to handling should occasion a rethinking of statistical best practices.

Given such considerations, many researchers who wish to work with big datasets choose to learn one or more programming languages, with Python and R being two of the most popular. Another option is to subscribe to a big *data analytics platform* such as Sysomos or Crimson Hexagon that provides tools to analyze digital communication data without requiring programming knowledge. Such services typically host the data on their own servers, providing limited access, if any, to the raw data. The major advantage of analytics platforms is their ease of use: their interfaces are designed to allow users to produce detailed reports and visualizations quickly. Their major disadvantages are their high cost (which may run into the thousands of US dollars per month even with academic discounts) and their lack of flexibility. As with desktop software, users can only run those analytical routines that have been pre-programmed into the service. Working directly with code costs no money and allows maximum flexibility in research design, but its learning curve can be steep.

For researchers who choose to analyze their big datasets programmatically, *data cleaning* is a critical task, because big datasets are not always exactly what they purport to be. Whether purchased from a data vendor or collected on one's own, some expected data may be missing, other data that do not fit the sampling criteria may be present, and still other data may be valid but improperly formatted. Broadly speaking, data cleaning encompasses all tasks intended to standardize and repair defects in datasets so that they fit their advertised specifications as faithfully as is reasonable. When possible, it is usually preferable to clean data programmatically, although some techniques may need to be implemented manually. Fortunately, many of the most important programmatic cleaning techniques are fast, simple, accurate, and readily implementable. For example, when working with many social media posts sampled over an extended period of time, it is a good idea to check that no dates are completely absent of posts. On most of the larger social media platforms this will be highly unlikely with all but the rarest of search criteria. Days with no data may indicate problems with the data source or the data collection process. Also, whenever sampling text data based on a keyword(s), researchers should ensure that the keyword actually occurs in every unit of analysis in the dataset. This is especially important with secondhand data that may not have been properly cleaned the first time. Another issue that sometimes afflicts text-based datasets is what might be called "field bleed," where one or more data fields bleed

into one another because they have not been properly delimited. Field bleed can sometimes be fixed by programmatically reinserting the correct delimiters or re-exporting the data from its original source files with a more rigorous export function, but in some cases the offending rows may have to be manually repaired or even removed. This is far from an exhaustive list of data cleaning techniques: additional steps will be necessary in nearly all cases, but they will need to be tailored to the contours of the specific dataset.

Digital traces as big data: analytical considerations

Much, although not all, big data of interest to the communication researcher is digital trace data: records of online behavior recorded automatically and unobtrusively by web servers, cookies, mobile apps, and other monitorial devices. Digital traces are generated by nearly every online action imaginable: every share, every "like," every purchase, every star rating, and every click creates a trace somewhere. This kind of data breathes new life into old research questions in addition to raising entirely new questions, but they should be analyzed with caution. Most trace datasets were not generated with academic research in mind, and carelessness in handling them can lead well-intentioned researchers astray.

Researchers who wish to analyze trace data should keep several considerations in mind, some of the most important being the following. First, *interpreting the results* of trace-based analyses is not always as straightforward as it may seem. A Facebook "like," for example, may appear to indicate approval of whatever content to which it is appended. But this is not always the case: on some public pages, one must like a post to be able to leave a comment. Interpreting likes is also difficult when the liked object refers to an objectionable individual or situation such as a criminal or a war, as it would not generally be reasonable to assume that the likers actually endorse the object. Moreover, the fact that Facebook offers no affordance for disliking content may lead researchers to overemphasize positive reactions (in those contexts where such inferences can be properly substantiated) relative to negative reactions. In short, it is not always possible to programmatically calculate the distribution of likes (or any other trace) in a dataset and immediately draw inferences about mass public sentiment based on the results. Qualitative methods such as interviews and close textual readings can be effective in identifying some of the less intuitive messages digital traces can convey (Freelon 2014).

Second, researchers should bear in mind that findings based on datasets of digital traces *cannot always be generalized to broader populations.* The 90-9-1 (or the 1 percent) rule has long held that the proportion of online users that participates regularly is much smaller than that which merely lurks, or spectates without contributing (Nielsen 2006). But we should also remember that the set of all active users of a given platform is usually not representative of any given larger population. Of the most-used social media platforms, only Facebook can claim that a majority of Americans (to say nothing of other countries) are active users. This means that analyses of social media data should be generalized carefully if at all. Researchers should consider what users, perspectives, and behaviors are left out or underrepresented in their data and interpret their findings accordingly. When feasible, they should adapt their research designs to compensate for these limitations.

Third, *not all digital traces are created by humans*. Many are created by programs called "bots" that attempt to simulate human activity, often for deceptive purposes. For example, Twitter bots can do anything a human Twitter user can do, such as tweet, retweet, follow others, and post images. Unscrupulous individuals have been known to purchase social media followers at inexpensive bulk rates to fabricate the appearance of popularity. Bots can also attack human users (or one another) by flooding them with irrelevant or abusive content when particular keywords are posted. Bot-created traces pose a challenge to researchers interested in human communication activity because they are not always easy to distinguish from those created by humans. Depending on the specific context, researchers may need

to filter their data using → network analysis, machine learning, manual inspection, or some combination of the three.

The final consideration that will be discussed here refers to *trace data ownership and access*. Many social media services provide some level of free data access through their APIs, but none allow access to all their available data. IP addresses put user privacy at risk because they reveal a fair amount of personal information, so reputable communication platforms rarely if ever publish them (→ Research Ethics; Research Ethics: Internet Research). Social media companies also typically impose lengthy terms of service that govern how their data can be used by third parties. Twitter, for example, does not allow researchers to post most kinds of Twitter data publicly, although it does allow the posting of numerical IDs designating users and tweets. Twitter sells access to historical data that cannot be obtained through its public API, but (as of fall 2015) insists that this data either be re-licensed at cost after one year or deleted. Such restrictions are likely to seriously impair some kinds of Twitter research, but may be some researchers' sole option for accessing historical Twitter data.

Non-social media websites also generate a wealth of trace data about their visitors, but most of it is typically only available to site owners. Therefore, researchers interested in such a website's visitor access data would need to convince the site's owners to allow privileged access to their server logs. While this may be a feasible option for some small-scale studies, the logistics of securing such consent from scores or hundreds of website proprietors quickly become untenable.

SEE ALSO: Content Analysis, Quantitative; Digital Media, History of; Electronic Mail; Facebook; Network Analysis; Online Research; Qualitative Methodology; Quantitative Methodology; Research Ethics; Research Ethics: Internet Research; Research Methods; Sampling, Nonrandom; Sampling, Random; Social Media; Statistics, Descriptive; Statistics, Explanatory; Survey; Technology and Communication; Twitter

References and suggested readings

Freelon, D. (2014). On the interpretation of digital trace data in communication and social computing research. *Journal of Broadcasting and Electronic Media*, *58*(1), 59–75.

Jacobs, A. (2009). The pathologies of big data. *Communications of the ACM*, *52*(8), 36–44. At http://doi.org/10.1145/1536616.1536632, accessed September 19, 2015.

Laney, D. (2001, February 6). 3-D Data Management: Controlling Data Volume, Velocity and Variety. *META Group.* At http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf, accessed September 27, 2015.

Nielsen, J. (2006). The 90-9-1 rule for participation inequality in social media and online communities. At http://www.nngroup.com/articles/participation-inequality/, accessed September 19, 2015.

Ryan, K. (2013). Big data + big math = big mess or big money? At http://searchengineland.com/big-data-big-math-big-mess-or-big-money-2-163322, accessed September 19, 2015.