Online Appendix for "Political Science and Big Data: Structured Data, Unstructured Data, and How to Use Them"

1. Content and Citation Analysis of Political Science Articles about Big Data

To analyze the state of the discipline, we examined bibliometric data from Clarivate Analytics' *Web of Science* (WoS) academic index. Despite the known shortcomings of WoS, which is a proprietary database whose inclusion criteria are not entirely transparent,[1] we consider it among the best available sources for citation data, especially when exploring the mainstream bibliography of a discipline.[2]

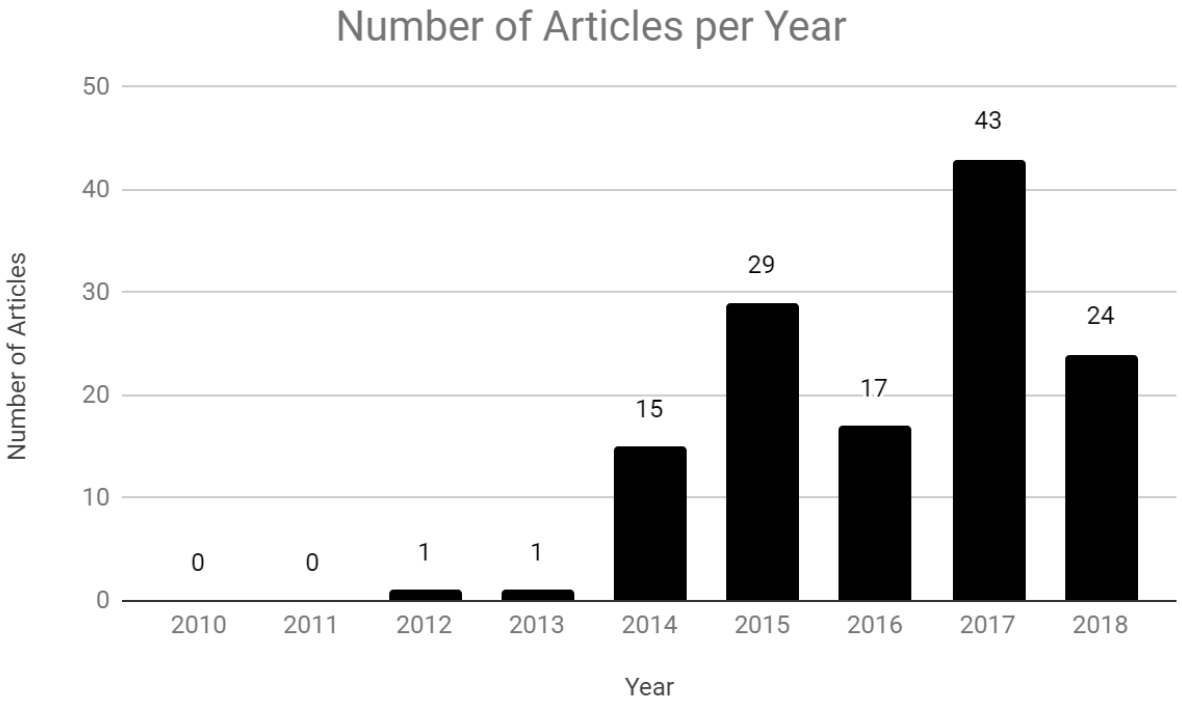
We searched the *Web of Science* on 1 January 2019 for papers whose topics included the expression "big data." We downloaded all *research articles* (that is, not conference proceedings, reviews, editorials, etc.) that were published in journals whose topic was categorized by WoS as political science (n=87), international relations (n=35), and public administration (n=36).[3] All in all, there are 133 such articles, 132 of which are in English and one is in Spanish, with the

[1] Diana Hicks et al., "Bibliometrics: The Leiden Manifesto for Research Metrics," *Nature* 520, no. 7548 (April 2015): 429–431; Philippe Mongeon and Adèle Paul-Hus, "The Journal Coverage of Web of Science and Scopus: A Comparative Analysis," *Scientometrics* 106, no. 1 (January 2016): 213–228, at 218–19.
[2] Peter Marcus Kristensen, "International Relations at the End: A Sociological Autopsy," *International Studies Quarterly* 62, no. 2 (June 2018): 245–259, at 246–47.
[3] These numbers are not exclusive. The *Web of Science* may assign up to three categories to the same journal. Further, WoS periodically updates its indexing, constantly adding and removing articles and categorizations. Therefore, the numbers presented here are likely to change retroactively every once in a while.

earliest article being published in 2012.[4] As Figure 1 shows, since 2014 there has been a marked increase, although not a dramatic or a steady one, in the number of political science articles about big data. Some of this growth was instigated by special issues: 30 articles were published in the same issue with at least two other articles from the database; 41 articles share the same volume with at least two other articles.

[4] The full list of articles and all the variables in this analysis can be found at https://github.com/for-anonymous-review/Big-Data-in-Political-Science. The actual number of articles may be somewhat higher: a few political science and policy articles about big data were not included in the *Web of Science* search results for unknown reasons. The two examples that we were able to find using other search methods are Sandra González-Bailón, "Social Science in the Era of Big Data," *Policy & Internet* 5, no. 2 (June 2013): 147–160; Maureen A. Pirog, "Data Will Drive Innovation in Public Policy and Management Research in the Next Decade," *Journal of Policy Analysis and Management* 33, no. 2 (Spring 2014): 537–543.

FIGURE 1: POLITICAL SCIENCE ARTICLES ABOUT BIG DATA, BY YEAR



Source: chart created with data from a Web of Science search conducted on 1 January 2019 with the following specifications: DOCUMENT TYPE = "Articles"; WEB OF SCIENCE CATEGORIES = "Political Science," "International Relations," and "Public Administration."

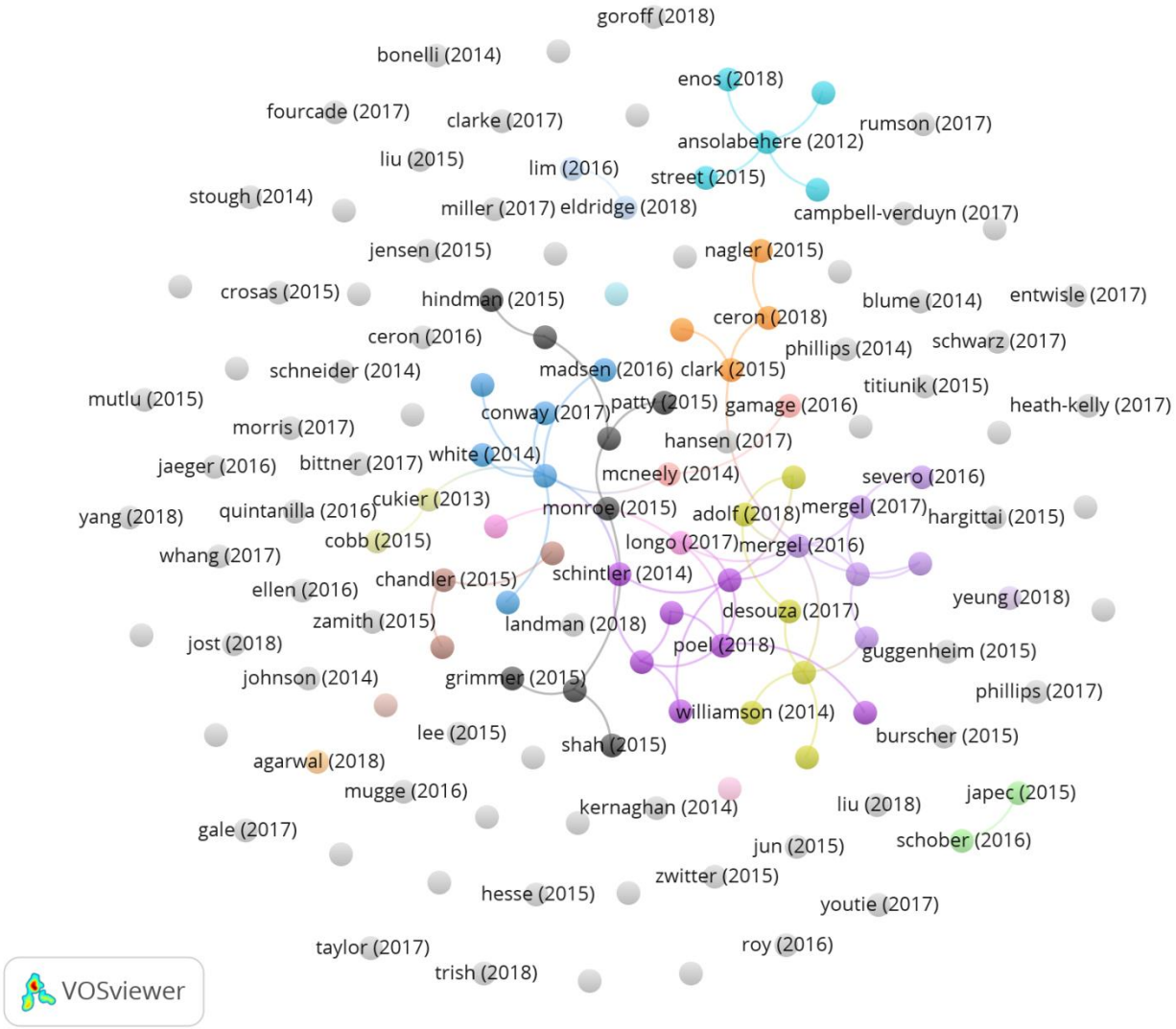
Next, we reviewed the content of each one of the 133 articles, both manually and with computer-assisted qualitative data analysis software (*NVivo*), and checked how many times and in what parts of the article the term "big data" occurred. The analysis revealed that, in some cases, the term served as a buzzword (or, perhaps, as clickbait) to catch the attention of readers or journal editors without much actual discussion of big data. In sixteen cases, the expression "big data" only occurred in the article title, abstract, list of references, or list of keywords

(including keywords designated by the *Web of Science* when a journal failed to provide such a list). Fifteen articles mentioned big data only once; in 35 articles, the term occurred three times or less. Only 52 articles – less than 40 percent of the 133 articles – offered a definition of big data that was workable to some extent or that cited existing definitions.

A citation network analysis of the 133 articles (using VOSviewer software[5]) suggests that the discussion of big data in political science is highly fragmented. Less than half of the articles in the database are linked to each other through citation (Figure 2). This includes even articles that gathered a relatively high number of citations by papers not in the database. The few articles that cite one another are often those that were published in the same special issue or by the same journal. As a co-citation analysis (Figure 3) demonstrates, many of the 133 articles draw on a small pool of seminal works from other disciplines about big data, while seldom referring to the rather limited literature on big data in political science.

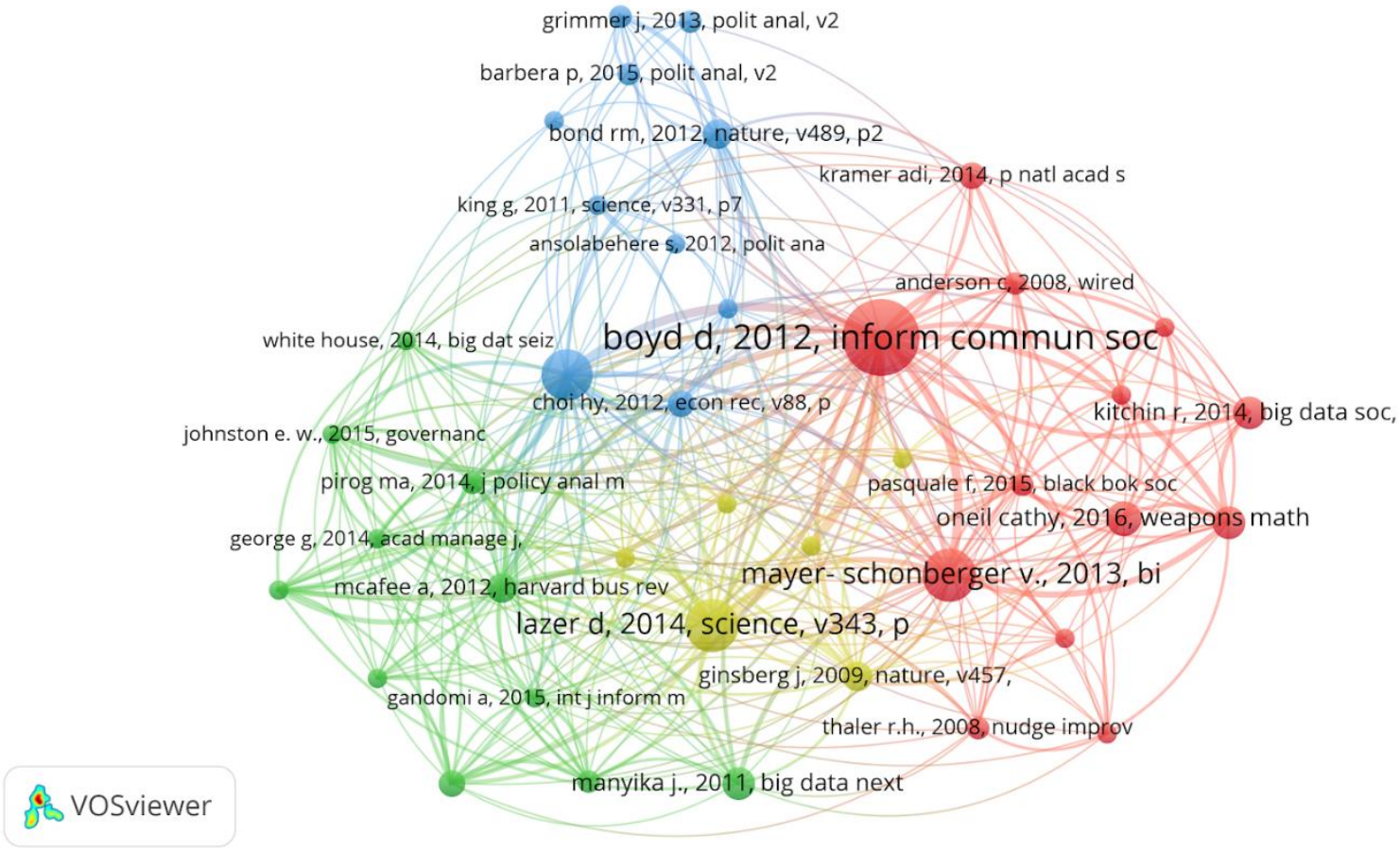
[5] http://www.vosviewer.com/. See Nees Jan van Eck and Ludo Waltman, "Software Survey: VOSviewer, a Computer Program for Bibliometric Mapping," *Scientometrics* 84, no. 2 (August 2010): 523–538.

FIGURE 2: A CITATION NETWORK OF POLITICAL SCIENCE ARTICLES ABOUT BIG DATA



Source: figure created using VOSviewer software and Web of Science bibliometric data (1 January 2019). A link between two nodes means that the more recent paper cites the earlier one. The size of each node represents the total number of citations that the paper has (according to WoS). The total number of citation links in the database is 59.

FIGURE 3: A CO-CITATION NETWORK OF POLITICAL SCIENCE ARTICLES ABOUT BIG DATA



Source: figure created using VOSviewer software and Web of Science bibliometric data (1 January 2019). Only papers that are cited by at least four articles in the database are shown (n=40). The cited works are represented by nodes. A link between two nodes means that at least one article in the database includes both articles in its list of references.

2. Finding Dates, Events, Actors, and Locations in Unstructured Big Data Sources

When large collections of unstructured data are available, it would be impractical to carefully read each and every document, article, or book and to manually identify and record all the temporal units, actors, locations, and events that occur in the text. Unless one has a huge, devoted team of highly trained research assistants at one's command, such a task will take far too much time.

To address this issue, we propose an additional approach: *regular expressions*. This concept refers to common patterns in a text as well as the ability to recognize such patterns.[6] Using regular expressions requires some digital literacy skills but not necessarily any programming knowledge; to find patterns in a body of text, it is possible to use one of the many online resources that allow uploading or pasting text into a designated box and specifying the terms of the search.[7] However, scholars with a basic background in programming languages such as Python can automate the process of text recognition without much effort.[8] A more sophisticated option would be to use a learning algorithm that, through a process of trial and error, would eventually detect patterns of interest.

How can we use regular expressions to identify temporal units, events, actors, and locations in big unstructured data? Dates and times would be the easiest to find, as they have a unique textual pattern that is distinguishable from other parts of the text. The format of presentation may vary – for example, the same date can appear as "26 September 1959," Sep. 26,

[6] Jeffrey E. F. Friedl, *Mastering Regular Expressions*, 3rd ed. (Sebastapol, CA: O'Reilly, 2006), 4.
[7] For example, https://regexr.com/.
[8] Al Sweigart, *Automate the Boring Stuff with Python: Practical Programming for Total Beginners* (San Francisco, CA: No Starch Press, 2015), chap. 7.

1959," "09/26/59," "26.9.1959," "1959-09-26," and so forth. In fact, the year may not appear at all. However, the number of possible formatting options is finite. By searching for all these patterns in an unstructured source, researchers can pinpoint the day or moment in which something that was worth reporting transpired. Based on their expertise and contextual knowledge, they can then decide which of those events are relevant to their questions and hence worthy of inclusion in the timeline. Once again, it is of the utmost importance that researchers document the exact search terms that they had entered as well as their inclusion criteria in order to be able to fix errors and mistakes or replicate the process at a later stage.

The names of places and people constitute a somewhat different category. While they appear to be an integral part of the text, they do have some characteristics (such as capitalization patterns and common suffixes) that can distinguish them from other parts of the text. Political scientists with a background in programming and knowledge in the field of natural language processing may pursue additional automated approaches. For example, information extraction systems can mine text on the basis of preloaded dictionaries of common names or locations.[9] Further, they can carry out *named entity recognition* tasks, whereby the system automatically identifies entities such as date and time, people, locations, and organizations.[10] By writing such an algorithm or installing an available package, researchers can extract from the data a list of actors, locations, and activities that are candidates for inclusion in the codebook and hence in the analysis. While the named entity recognition approach involves a high degree of automation, it is still an auxiliary method. Researchers must decide which entities to include and how to carry out the analysis.

[9] Ronen Feldman and James Sanger, *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data* (Cambridge; New York: Cambridge University Press, 2007), 106–107.
[10] Ibid., 96–97; Gregor Wiedemann, *Text Mining for Qualitative Data Analysis in the Social Sciences* (New York: Springer Berlin Heidelberg, 2016), 33–34.

Alternatively, and more deductively, researchers can run a text search for specific words of interest. Many software packages, including free document readers, can perform advanced searches in multiple documents. Researchers who know what they are looking for can create a list of relevant search terms in all possible variations – for example, "Dwight Eisenhower," "Dwight D. Eisenhower," "Dwight David Eisenhower," "'Ike' Eisenhower," etc. Searching for these words, the researchers can isolate the chunks of text in which they occur to read them more closely. Doing the same with verbs of interest (for example, "visit," "declare," "fight"), they may find information about events. Some software packages offer more sophisticated search capabilities, such as stemming and automatic detection of synonyms. However, this method is only valid for unearthing additional information on known entities rather than discovering new ones. This limitation underlines, once again, the importance of background knowledge, context, regional expertise, theory, and clearly defined questions and hypotheses for causal inference in political science.