

textNet: Directed, Multiplex, Multimodal Event Network Extraction from Textual Data

Elise Zufall¹ and Tyler Scott^{1,2}

¹ Center for Environmental Policy & Behavior, UC Davis ² Department of Environmental Science & Policy, UC Davis

DOI: [10.xxxxxx/draft](https://doi.org/10.xxxxxx/draft)

Software

- [Review](#) ↗
- [Repository](#) ↗
- [Archive](#) ↗

Editor: [Open Journals](#) ↗

Reviewers:

- [@openjournals](#)

Submitted: 01 January 1970

Published: unpublished

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](#)).

Introduction

Network measurement in social science typically relies on data collected through surveys and interviews. Document-based measurement can be automated and scaled, providing opportunities for large scale or longitudinal research that are not possible through traditional methods. A number of tools exist to generate networks based on co-occurrence of words within documents (such as the [Nocodefunctions](#) app ([Levallois et al., 2012](#)), the “[textnets](#)” package ([Bail, 2024](#)), [InfraNodus](#) ([Paranyushkin, 2018](#)), and many more). But there is, to our knowledge, no open-source tool that generates network data based on the syntactic relationships between entities within a sentence. *textNet* allows a user to input one or more PDF documents and create arbitrarily complex directed, multiplex, and multimodal network graphs, enabling rich analysis of the relationships between verb attributes and tenses, entity types, structural motifs, and other network characteristics. For instance, [Zufall and Scott](#) demonstrate the use of *textNet* to identify which actors are involved in start-up versus ongoing management tasks, characterize patterns of information and funding flows, and compare the distribution of management tasks in networks from regions known to have contrasting characteristics ([2024](#)). *textNet* also works on arbitrarily long documents, making it well suited for research applications using long texts such as government planning documents, court proceedings, regulatory impact analyses, and environmental impact assessments.

textNet has applications in governance network scholarship, as demonstrated by [Zufall and Scott \(2024\)](#) and by ongoing work on water resources governance at the UC Davis Center for Environmental Policy and Behavior. Additional potential applications include legal scholarship, social-ecological network analysis, government planning documents, court proceedings, archival research, communication and media research, and other fields interested in exploring events and entity relationships in textual data.

Statement of Need

Network extraction from the kinds of documents listed above has typically required manual coding. Furthermore, existing network extraction methods that use co-occurrence leave a vast amount of data on the table, namely, the rich edge attribute data and directionality of each verb phrase defining the particular relationship between two entities, and the respective roles of the entity nodes involved in that verb phrase. We present an R package, *textNet*, designed to enable directed, multiplex, multimodal network extraction from text documents through syntactic dependency parsing, in a replicable, automated fashion for collections of arbitrarily long documents. The *textNet* package facilitates the automated analysis and comparison of many documents, based on their respective network characteristics. Its flexibility allows for any desired entity categories, such as organizations, geopolitical entities, dates, or custom-defined categories, to be preserved.

42 Directed Graph Production

43 As a syntax-based network extractor, *textNet* identifies source and target nodes. This produces
44 directed graphs that contain information about network flow. Methods based on identifying co-
45 occurring nodes in a document, by contrast, produce undirected graphs. Co-occurrence graphs
46 also tend to generate saturated subgraphs, since every co-occurring collection of entities has
47 every possible edge drawn amongst them. By contrast, *textNet* draws connections specifically
48 between pairs of entities that are mediated by an event relationship, rather than between every
49 entity in the document or even in the sentence.

50 Multiplex Graph Output

51 Syntax-based measurement encodes edges based on subject-verb-object relationships. *textNet*
52 stores verb information as edge attributes, which allows the user to preserve arbitrarily complex
53 topological layers (of different types of relationships) or customize groupings of edge types to
54 simplify representation.

55 Multimodal Graph Output

56 Multimodal networks, or networks where there are multiple categories of nodes, have common
57 use cases such as social-ecological network analysis of configurations of actors and environmental
58 features. Existing packages such as the *manynet* package (Hollway, 2024) provide analytical
59 functions for multimodal network statistics. *textNet* provides a structure for tagging and
60 organizing arbitrarily complex node labeling schemes that can then be fed into packages for
61 multi-node network statistical analysis. Node labels can be automated (e.g., the default entity
62 type tags for an NLP engine such as *spaCy* (Honnibal et al., 2021)), customized using a
63 dictionary, or based on a hybrid scheme of default and custom labels. Any node type is possible
64 (e.g., species, places, people, concepts, etc.) so this can be adapted to domain-specific research
65 applications by applying dictionaries or using a custom NER model.

66 Installation

67 The stable version of this package can be installed from Github, using the *pak* package (Csárdi,
68 Hester, et al., 2024):

```
69 pak::pak("ucd-cepb/textNet")
```

70 The *textNet* package suggests several convenience wrappers of packages such as *spacyr* (Benoit
71 et al., 2023), *pdftools* (Ooms, 2024), *igraph* (Csárdi, Nepusz, et al., 2024), and *network*
72 (Butts et al., 2023). To use the full functionality of *textNet*, such as pre-processing tools
73 and post-processing analysis tools, we recommend installing these packages, which for *spacyr*
74 requires integration with Python. However, the user may wish to preprocess and parse data
75 using their own NLP engine, and skip directly to the `textnet_extract()` function, which does
76 not depend on *spacyr* or Python integration.

77 Overview and Main Functions

78 The package architecture relies on four sets of functions around core tasks:

- 79 ■ [OPTIONAL] Pre-processing: `pdf_clean()`, a wrapper for the `pdftools::pdf_text()`
80 function which includes a custom header/footer text removal feature; and `parse_text()`,
81 which is a wrapper for the *spacyr* package and uses the *spaCy* natural language processing
82 engine (Honnibal et al., 2021) to parse text and perform part of speech tagging,
83 dependency parsing, and named entity recognition (NER). Alternatively, the user can
84 skip this step and load parsed text directly into the package. Externally produced data
85 must be converted to the format requirements outlined in the package manual.
- 86 ■ Network extraction: `textnet_extract()`, which generates a graph database from parsed text
87 based upon tags and dependency relations. The object returned from `textnet_extract()`

consists of a nodelist, an edgelist with a rich set of edge attributes, a verblist, and an appositivelist (containing potential coreferences such as acronyms and their full forms for disambiguation).

- Disambiguation: tools for cleaning, recoding, and aggregating node and edge attributes, such as the `find_acronyms()` function, which can be paired with the `disambiguation()` function to identify acronyms in the text and replace them with the full entity name.
- Exploration: the `export_to_network()` function for exporting the graph database to igraph and network objects, `top_features()` for viewing node and edge attributes, and `combine_networks()` for aggregating multiple document-based graphs based on common nodes.

The figure below summarizes the functionality of *textNet* and the flow of function outputs. Optional data cleaning features are shown with dotted arrows.

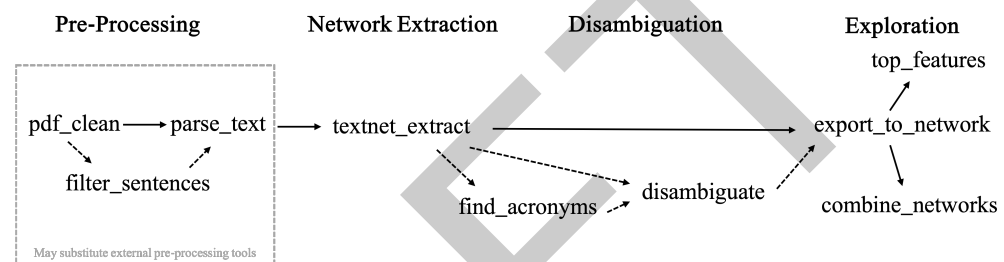


Figure 1: Workflow of *textNet* Functions

Downstream Analysis

textNet is compatible with standard network analysis tools in R. Functionality provided by *ggraph* (Pedersen & RStudio, 2024), *sna* (Butts, 2024), *igraph* (Csárdi, Nepusz, et al., 2024), *network* (Butts et al., 2023), and other network visualization and analysis packages can be used to further explore the extracted networks.

The *ggraph* package has been used to create the network visualization seen here, using a weighted version of an igraph constructed using the “old_new_parsed” sample data in *textNet*.

New Network



Figure 2: Representation of the Event Network of the New Plan

The network-level attributes output from `export_to_network` can also be analyzed against exogenous metadata that has been collected separately by the researcher regarding the different documents and their real-world context. The extracted networks can also be analyzed through a variety of network analysis tools, such as an Exponential Random Graph Model or a Temporal Exponential Random Graph Model.

Vignette

More information about the entity network extraction algorithm and an example start-to-finish data processing and analysis workflow can be found in the vignette for this package. The vignette uses sample data that travels with the *textNet* package.

Acknowledgements

The authors gratefully acknowledge the support of the Sustainable Agricultural Systems program, project award no. 2021-68012-35914, from the U.S. Department of Agriculture's National Institute of Food and Agriculture and the National Science Foundation's Dynamics of Integrated Socio-Environmental Systems program, grant no. 2205239.

References

- Bail, C. (2024). *Cbail/textnets* (Version 0.1.1). <https://github.com/cbail/textnets>
- Benoit, K., Matsuo, A., Gruber, J., & Council (ERC-2011-StG 283794-QUANTESS), E. R. (2023). *Spacyr: Wrapper to the 'spaCy' 'NLP' library* (Version 1.3.0). <https://cran.r-project.org/web/packages/spacyr/index.html>
- Butts, C. T. (2024). *Sna: Tools for social network analysis* (Version 2.8). <https://cran.r-project.org/web/packages/sna/index.html>

- 128 Butts, C. T., Hunter, D., Handcock, M., Bender-deMoll, S., Horner, J., Wang, L., Krivitsky, P.
129 N., Knapp, B., Bojanowski, M., & Klumb, C. (2023). *Network: Classes for relational data*
130 (Version 1.18.2). <https://cran.r-project.org/web/packages/network/index.html>
- 131 Csárdi, G., Hester, J., & Software, P. (2024). *Pak: Another approach to package installation*
132 (Version 0.8.0). <https://cran.r-project.org/web/packages/pak/index.html>
- 133 Csárdi, G., Nepusz, T., Traag, V., Horvát, S., Zanini, F., Noom, D., Müller, K., Salmon, M.,
134 Antonov, M., & details, C. Z. I. igraph author. (2024). *Igraph: Network analysis and*
135 *visualization* (Version 2.1.1). <https://cran.r-project.org/web/packages/igraph/index.html>
- 136 Hollway, J. (2024). *Manynet: Many ways to make, modify, map, mark, and measure myriad*
137 *networks* (Version 1.2.6). <https://CRAN.R-project.org/package=manynet>
- 138 Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2021). *spaCy: Industrial-strength*
139 *natural language processing in python* (Version 3.1.3). [https://github.com/explosion/](https://github.com/explosion/spaCy/tree/master)
140 [spaCy/tree/master](https://github.com/explosion/spaCy/tree/master)
- 141 Levallois, C., Clithero, J. A., Wouters, P., Smidts, A., & Huettel, S. A. (2012). Translating
142 upwards: Linking the neural and social sciences via neuroeconomics. *Nature Reviews*
143 *Neuroscience*, 13(11), 789–797. [https://nocodefunctions.com/cowo/semantic_networks_](https://nocodefunctions.com/cowo/semantic_networks_tool.html)
144 [tool.html](https://nocodefunctions.com/cowo/semantic_networks_tool.html)
- 145 Ooms, J. (2024). *Pdftools: Text extraction, rendering and converting of PDF documents*
146 (Version 3.4.1). <https://cran.r-project.org/web/packages/pdftools/index.html>
- 147 Paranyushkin, D. (2018). *InfraNodus*. Nodus Labs. <https://infranodus.com/>
- 148 Pedersen, T. L., & RStudio. (2024). *Ggraph: An implementation of grammar of graphics for*
149 *graphs and networks* (Version 2.2.1). [https://cran.r-project.org/web/packages/ggraph/](https://cran.r-project.org/web/packages/ggraph/index.html)
150 [index.html](https://cran.r-project.org/web/packages/ggraph/index.html)
- 151 Zufall, E., & Scott, T. A. (2024). Syntactic measurement of governance networks from
152 textual data, with application to water management plans. *Policy Studies Journal*, n/a.
153 <https://doi.org/10.1111/psj.12556>