Journal of Open Humanities Data

Graham SS, et al. 2020 Methods for Extracting Relational Data from Unstructured Texts Prior to Network Visualization in Humanities Research. *Journal of Open Humanities Data*, 6: 8. DOI: https://doi.org/10.5334/johd.21

RESEARCH PAPER

Methods for Extracting Relational Data from Unstructured Texts Prior to Network Visualization in Humanities Research

S. Scott Graham[1], Zoltan P. Majdik[2] and Dave Clark[3]

[1] Department of Rhetoric and Writing, University of Texas at Austin, Austin, US

[2] Department of Communication, North Dakota State University, Fargo, US

[3] Department of English, University of Wisconsin-Milwaukee, Milwaukee, US

Corresponding author: S. Scott Graham (ssg@utexas.edu)

Network modelling methodologies in the digital humanities have been be used to advance inquiry in a variety of areas—most commonly those having to do with correspondence, citation, and social media networks. While new technologies have made generating high-quality and even dynamic network visualizations relatively easy, key challenges remain for humanities researchers. Many common objects of humanistic inquiry, such as literary, historiographic, and biographical texts are often not easily transformed into the kinds of data structures necessary for network visualization. The Transparency to Visibility (T2V) Project was initiated to develop new methods and toolkits that can support humanistic researchers who need to extract relationship data from unstructured texts to support network visualization. The T2V team used bioethics accountability statements to pilot and evaluate different methods for extracting relationship data. The resulting machine-learning-enhanced natural language processing (NLP) and metadata-assisted approaches offer promising potential pathways for contemporary digital humanities and future toolkit development.

Keywords: bioethics; medical humanities; network modelling; natural language processing; machine learning
Funding statement: Funding was provided by a National Endowment for the Humanities Digital Humanities Advancement grant (HAA-261070).

1 Context and motivation

Humanities researchers have long studied how information and influence circulate through cultural systems. Advances in network visualization tools support this work, allowing scholars to create graphical representations of complex discursive and cultural systems. While both proprietary and open-source network mapping software have made generating high-quality and even dynamic network visualizations relatively easy, key challenges remain for humanities researchers. Primary among these challenges is the humanistic focus on unstructured textual data (novels, archives, poems, biographies, etc.). Creative, historiographic, biographical, and similar artifacts are usually not easily transformed into the kinds of data structures necessary for network visualization. Additionally, even when objects of study can be somewhat easily rendered into visualization-ready data formats, these transformations can be very time intensive and/or require advanced computational skills.

Thus, there is a significant need for the development of new methods that can support humanistic researchers who need to transform unstructured textual datasets into data structures that support useful and informative network visualization. The Transparency to Visibility (T2V) Project was initiated to pursue these goals. The T2V team used bioethics accountability statements to pilot and evaluate different methods for transforming and visualizing relational networks based on data in unstructured text. The resulting machine-learning-enhanced natural language processing (NLP) and metadata-assisted approaches offer promising potential pathways for contemporary digital humanities and future toolkit development. In what follows, we provide a brief summary of the current state of network visualization methods in the digital humanities (section 1); describe the exigencies for the current project (section 2); and detail our approach to data extraction for subsequent network visualization (section 3).

1.1 Humanities network modelling

In recent years humanities journals have seen an explosion in network mapping methodologies applied to social media discourse, scholarly citation networks, and all manners of archival and textual materials. The recent enthusiasm comes in part from the fact that network modelling

offers humanities researchers one powerful way of grappling with "the complexity of the objects of [our] disciplines" [14]. That is, network modelling provides methods for tracing and visualizing complex, multidimensional intra- and inter-textual relationships. In their simplest form, network models provide relationship data; they graphically represent the connections among nodes and edges (dots and lines in a network map). Scholars using network modelling can combine different graphical algorithms and other visual treatments to help make certain network features more visible. While appropriate use cases are myriad, Grandjean offers a rough taxonomy of scenarios where network modelling can be especially helpful (p. 2). **Table 1** summarizes this taxonomy.

Ultimately, research using network modelling has been instrumental in developing enhanced understandings of social media discourse, citation networks, socio-technical systems, historic social networks, and the circulation of textual forms within particular cultures.

Despite the uptake of network modelling methods in the digital humanities, there is a tendency to focus on a rather limited set of use cases, primarily those that would fall under the "circulation" type in Grandjean's taxonomy. Facebook friend networks, retweet networks, and citation networks, for example, are particularly easy to submit to network modelling because they are, by default, stored using data structures designed to highlight interrelationships among objects, e.g. relational databases. It is a relatively simple process to connect to the Twitter API or a public database and extract the kinds of data that can be readily transformed into nodes and edges tables. Even in cases where data is not conveniently stored in a relational database, there is a tendency to focus attention on the kinds of metadata that can be relatively easily extracted. For example, the *Mapping the Republic of Letters* (2013) project leverages Oxford's *Electronic Enlightenment Project* to visualize the geography of correspondence networks for key enlightenment thinkers [22]. Much of this project revolves around digitizing the structured metadata from each letter (sender name, recipient name, mailing addresses, date, etc.).

However, a significant challenge for many humanities projects with respect to network modelling is that "data" are frequently neither retrievable nor structured. A scholar attempting to model the character networks in [8] *The Brothers Karamazov*, for example, would not be able to easily download aggregate character interaction data. Additionally, individual characters, as presented in the novel, do not have preassigned unique identifiers that would make them easy to track. Preparing the data for network modelling requires knowing that Alexei and Alyosha are the same person, for example. Likewise, transforming the novel's text into a nodes and edges table requires establishing a framework for identifying relationships. Does something as simple as co-mentions per page constitute a "relationship"? Is it important to know the type of relationship for the analysis in question?

In sum, there are three key challenges that remain to be addressed before network modelling can be used more widely in humanistic research that cannot rely on pre-structured data. First, humanities researchers need methods that support consistent and reliable identification of nodes in unstructured text. Second, humanities researchers need approaches and techniques for determining when identified nodes are "in" a relationship. Finally, network modelling humanists need efficient and consistent ways of classifying relationship types within unstructured text.

A handful of digital humanities projects have made forays into addressing these areas, developing the kinds of advanced tools involving machine learning and/or NLP that are required to meet theses aims. The REDEN framework, developed by a group of linguists and literary historians, uses NLP named-entity recognition (NER) combined with structured and retrievable metadata to identify, distinguish, and connect different authors in French literary history [4]. REDEN makes important strides towards recognizing nodes of interest despite the challenges presented by multiple people having similar names (e.g., the multiple Baudelaires of French literary history). Another interesting example is the *Six Degrees of Francis Bacon* project [23]. This project combines NER to identify nodes (people) with an unsupervised machine-learning framework that estimates relationship strength based on document-level co-occurrence within a large corpus. While these projects offer promising approaches to addressing the first two problems above, the challenge of classifying relationships remains. The potential scale and scope of this challenge is exemplified in [9] "Semantic network edges: A human-machine approach to represent typed relations in social networks" They too used an NER-based framework for node identification but ended up crowdsourcing edge classification.

1.2 The T2V project

The primary aim of the T2V project is to develop a method for extracting affiliation network data from unstructured text. We opted to prototype our method using conflict of interest (COI) statements in medical publishing. These statements, which disclose financial relationships between medical researchers and biotech companies are only minimally structured, but contain relationships among writers and agencies that, while obvious to human readers, can be a challenge to capture in a database and visualize in a network. Thus, they represent an ideal test case for the T2V parser.

Table 1: Grandjean's (2020) taxonomy of network types.

Network Type	Description
Affiliation	Maps connections between companies, institutions, and groups of peoples.
Character	Maps relationships between fictional characters, character groups, and temporalities.
Circulation	Maps transportation means, places, goods, correspondence routes, etc.

Recent research in bioethics and related health policy suggests that network modelling approaches might be a promising avenue for future research in this area. Specifically, available data indicate that the biasing effects of COI may be magnified when authors and the study itself are funded by the same industry sources [1]. Similarly, an analysis using the approach described in this article has found that author COI rates are associated with certain industry payments to biomedical journals [13]. While a network modelling approach has not yet been used to further explore COI broadly, these findings suggest that such an approach might be quite promising. In turning toward this direction, we are inspired by approaches to network analysis from the humanities and humanistic science and technology studies. Although there are often significant and possibly irreconcilable differences among the various intellectual approaches available [7], rhizomatic theory [15], technoscientific networks [21], actor-networks, and [2] theory of intra-action (among many others) all highlight the importance of understanding the nature of relations and the types of circulation made possible within complex systems. These theoretical constructs are especially well-attended to investigating network features like articulation density and complexity as primary sources of power and influence.

While visualizing affiliation networks has the potential to be useful here, disclosure statements exist in a wide variety of unstructured prose formats, making it difficult to extract affiliation data systematically. For example, various COI style guides might represent a single disclosure as follows:

- Charles Winchester holds stock in GlaxoSmithKline.
- CE Winchester has equity interests in GSK.
- CEW holds equity shares in Glaxo.
- C.E.W. is a shareholder with GlaxoSmithKline Inc.
- Dr. Winchester has stock options with Glaxo Smith Kline.

The author holds equity interests with GSK India. In this case, the name of the researcher, the name of the company, and the type of relationship can each be represented in 3–5 different ways creating up to 100 possible textual permutations for the same three data points.

This issue is further complicated by the fact that many journal articles include numerous authors. It is not uncommon for large multicenter randomized controlled trials to include 50–100 named authors. Thus, individual sentences within conflicts of interest statements may group authors according to similar conflicts. For example, the following is an actual conflict-of-interest disclosure statement for an article with a relatively small number of authors:

> Frank Ernst, Peri Barr, and Riad Elmor are employees of Indegene, Inc., which received a fee for services related to the development and execution of this study, and for the tabulation, analysis, and reporting of its results. Walter Sandulli and Jessica Goldenberg are employees of Akrimax. Arnold Sterman has been a consultant for Akrimax, has contributed to research funded by Akrimax, and received an honorarium for his contributions to evaluating this study and to the development of this manuscript [10].

An effective relationship parser must be able to identify each individual relationship from this text:

- Frank Ernst are employees of Indegene, Inc.,
- Peri Barr are employees of Indegene, Inc.,
- Riad Elmor are employees of Indegene, Inc.,
- etc. ….

The identified relationships must then be parsed into source, target, and type categories (see **Table 2**). In order to effectively evaluate COI, there must also be a way of normalizing different representations of the same entity. That is, in the prior example, it would be important to know that GSK, GlaxoSmithKline, and GSK Inc are, in fact, the same entity. Otherwise, there will be at least three different GlaxoSmithKline nodes in any resulting network diagram. Given the unstructured nature of the current dataset, it is not possible to do this perfectly, but certain interventions will allow for increased reliability of results.

2 Method

Our data comes from the MEDLINE database, an online biomedical and life sciences bibliographic database. MEDLINE's database indexes more than 30 million journal articles, books, and scholarly reports, with selected records dating back to 1879. To begin our study of conflict statements, we downloaded all MEDLINE XML files.[1] We then used an XML parser to load selected data on each of the 30 million indexed publications into a local database that would support our project. In our custom database, each article is represented across four tables linked by a common PMID (or PubMed ID), which is also the index used by PubMed. For example, the PMID for Ernst et al. is 27798756 and appears at the end of its PubMed URL: https://pubmed.ncbi.nlm.nih.gov/27798756/. MEDLINE only began collecting COI information in 2016, and not all journals participate in the program of reporting author COI. Thus, of the 30 million collected articles, only 274,246 included COI statements. Our analysis indicates that those 274,246 have a total of 159,878 individual COI. Among those articles with conflicts, each article has an average of 10 reported conflicts.

Using this subset of the data and building on prior work in digital humanities and text analytics, we developed two variants of the T2V parser: the first uses a combination

Table 2: Integrated nodes and edges table derived from a COI statement.

Source	Target	Relationship Type
Indegene, Inc	Frank Ernst	Employment
Indegene, Inc	Peri Barr	Employment
Indegene, Inc	Riad Elmor	Employment
Akrimax	Indegene, Inc	Fee for Services
etc	*etc*	*etc*

of machine-learning enhanced named-entity recognition (NER) tagging and a conflict type dictionary to identify nodes (sponsors and authors) and edges (reported relationships). The second version uses PubMed/MEDLINE author metadata to improve overall parser performance. Two versions of the parser were developed in anticipation of future digital humanities projects that may not have pre-curated metadata available. We refer to each version of the parser as the Pure Machine Learning (PML) Parser and the Hybrid-Metadata Assisted (HMA) Parser, respectively. Each parser's logic model is below in **Figures 1** and **2**.

In short, the approach uses a trained language model to tag sponsors (e.g., pharmaceutical companies) in unstructured COI statements. When an organizational name is present in a COI statement, the parser then combines dictionaries of author name permutations (in the HMA model), or NER-tagged authors (in the PML model), and conflict types to extract individual COI. For example, this sentence in the below COI:

> "Simon Knight has received consultancy fees from OrganOx UK Ltd" is parsed into the following nodes and edges table in **Table 3**.

Those extracted conflicts are then passed to post-processing models that clean the data and render it in node and edge tables. Below we describe the PML parser in more detail. The PML parser will be most readily applicable for other projects in the humanities, and the functionality of the HMA parser is fully explained in [13]. Following the explanation of parser components, we offer a more complicated parsing example.

2.1 Author and sponsor identification

spaCy is a Python library that ueses pre-trained language models for natural language processing (NLP) [17]. spaCy supports multiple NLP applications including tokenization, tagging, and Named Entity Recognition (NER). The PMA parser leverages spaCy's NER capabilities to identify authors and sponsors in COI statements. A sentence such as "Walter Sandulli and Jessica Goldenberg are employees of Akrimax," when parsed through spaCy, would produce three "named entities":

Walter Sandulli, PERSON
Jessica Goldenberg, PERSON
Akrimax, ORG

Table 3: A nodes and edges table generated from the parsing of a simple COI statement.

Target	Relationship Type	Source	Conflict Weight
Simon Knight	fees	OrganOx UK	1

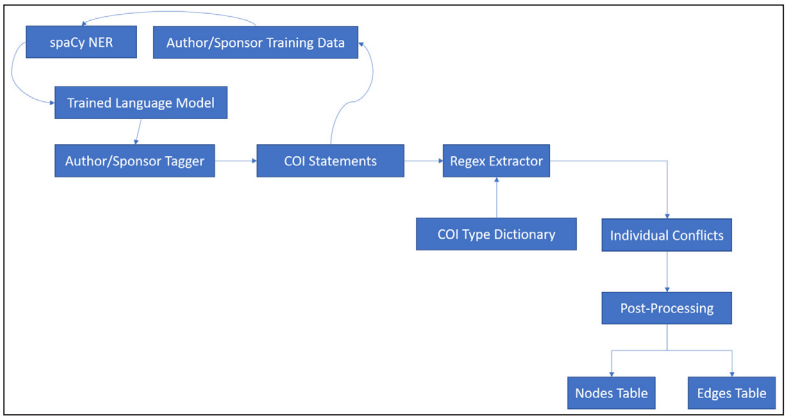

Figure 1: PML Logic Model.

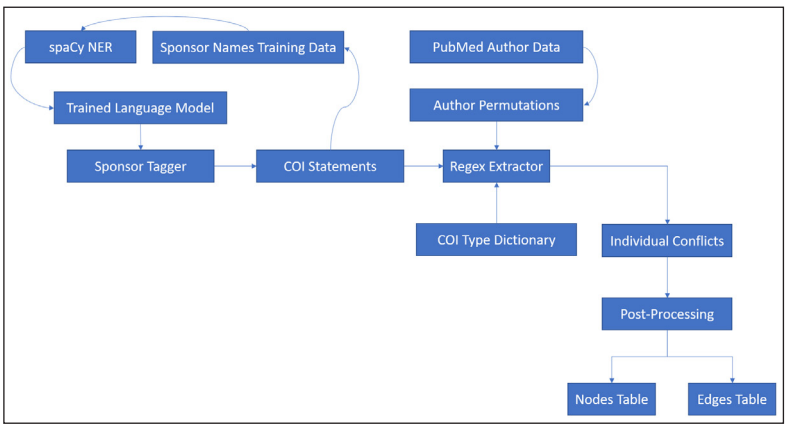

Figure 2: HMA Parser Logic Model.

As of 2020, spaCy v3, using the English Core Web Large (en_core_web_lg) model, achieved 85.4% accuracy on NER benchmarking assessments [18]. Because we are working with non-standardized and often idiosyncratic human language, we opted to improve the performance of the default spaCy NER-tagger with a training set specific to this project. spaCy's accuracy can be increased for individual projects by augmenting default language models with additional training data. In developing the PMA parser, we were able improve NER recognition by 25% using a small (n = 100) training set of human-tagged COI statements. In the HMA parser, author identification is accomplished by drawing author names and name abbreviations from the publication metadata in the MEDLINE database.

2.2 Relationship types and COI classification dictionary

Relationship classification for this project is based COI guidance provided by the International Committee of Medical Journal Editors (ICMJE). The ICMJE suggests that COI disclosures cluster around the following five broad areas: grant, personal fees, non-financial support, other, and intellectual property. [19] guidance for each category is listed below:

> **Grant:** A grant from an entity generally [but not always] paid to your organization.
> **Personal fees:** Monies paid to you for services rendered, generally honoraria, royalties, or fees for consulting, lectures, speakers bureaus, expert testimony, employment, or other affiliations.
> **Non-Financial Support:** Examples include drugs/equipment supplied by the entity, travel paid by the entity, writing assistance, administrative support.
> **Other:** Anything not covered under the previous three boxes.
> **Intellectual Property:** Patents and copyrights.

Our COI dictionary schema organizes these categories (plus "employment in industry") into a three-level schema based on potential benefit from a product's success. Specifically,

> **Low-Level COI** includes personal fees and non-financial support, as described by ICMJE.
> **Mid-Level COI** includes grants and research support.
> **High-Level COI** includes stock ownership and employment in industry.

The dictionary's implementation began with the terms provided by the ICMJE (e.g., for low-level COI, honoraria, consulting fees, speaking, fees) and expanded the dictionary based on the actual data available in the disclosure statements. The dictionary was implemented as part of the regular expression (Regex) parser described below.

```
LOW
r'(?:equity in|(?:owns?|owned|owned by)|patent|financial interest in|employ\w+\W|is (?:CEO|CFO)|is the (?:CEO|CFO)|inventor|found\w+|co-?found\w+)'
MID
r'(?:grant|fund\w+\W|support\w+\W|contract\w+\W|collaborat\w+\W|research)'
HIGH
r'(?:consul\w+\W|advi\w+\W|honorari\w+\W|fees?|edit\w+\W|travel\w*|member|panel)'
```

2.3 Relationship extraction

Both variants of the parser extract specific relationships through evaluating the proximity of identified sponsors, authors, and relationship types, as described above. Specifically, Regex expressions check to see if author names are followed (within 80 words, but not outside sentence boundaries) by COI relationships. If so, the parser checks to see if the author name and COI tag are followed by an ORG-tagged named entity. This process is repeated for each tagged sponsor in a COI statement. Outputs are assigned a numerical weight based on the COI classification dictionary. **Table 4** shows the result of our parser's work on the example data from [12] "PharmGKB summary: ivacaftor pathway, pharmacokinetics/pharmacodynamics." The original COI statement from which these results were derived is:

> RBA is a stockholder in Personalis Inc. and a paid advisor for Personalis Inc., Pfizer and Karius. TEK and MWC are paid scientific advisors to Rxight Pharmacogenetics. JPC has received research contract support to conduct clinical trials of ivacacftor at his institution.

In this example, the HMA parser was used so as to cross-reference author name abbreviations (e.g., RBA) with author names (Russ Altman).

3 Results and discussion

The PML and HMA parsers have several limitations that reduce overall accuracy. In the above example, a level two COI (research funding) is described in the last sentence. However, because the COI was described in terms of the drug studied (ivacacftor) and not the funder, no ORG tag

Table 4: Edges and nodes table generated from HMA-parsing of a complex COI statement.

Source	Target	COI Weight
Personalis Inc.	Russ Altman	1
Personalis Inc.	Russ Altman	3
Pfizer	Russ Altman	3
Karius	Russ Altman	3
Rxight Pharmacogenetics	Teri Klein	3
Rxight Pharmacogenetics	Michelle Whirl Carrillo	3

was assigned by spaCy, and therefore no relationship was identified by the parser. Additionally, as the vast majority of COI statements are written in plain English, the parse assumes that declared COIs will follow the expected Subject-Verb-Object sentence structure. As a result, the parsers are less reliable for COI statements that do not use complete sentences, e.g. "Conflict of Interest: C. Hohl: Consulting and honorarium: Cook Medical. A. Bro: Employee: Cook Medical" [16].

Despite these limitations, both parsers achieved moderate to high levels of reliability across COI categories. To assess overall reliability, a random sample of 1000 COI statements was submitted to human evaluation. Our sampling protocol excluded COI statements of fewer than 10 words. Our PubMed dataset includes 274,245 COI statements. However, the results of our analysis indicate that 258,871 of these are some version of "The authors report no conflicts of interest." Thus, a truly representative sample of 1000 COI statements would only provide 56 statements for the human or parser to evaluate. Since identifying that no conflicts are present is an easier computational task than COI classification, our approach here invariably resulted in lower ICC scores than would be expected in a truly representative sample. However, the benefit of this approach is that it ensured the parser evaluation would involve a much wider variety of conflict types.

Machine-human interrater reliability approach using an intraclass correlation coefficient (ICC) [3]. Since the ultimate goal of the project is to automate and extend the scale of human analyses, it is an appropriate metric for ensuring that the parser "codes like a human." Other digital humanities projects may be designed to perform tasks for the analysis itself that would be impossible for humans. However, in cases where the primary challenges are scale and scope, human-machine interrater reliability metrics as applied to appropriate samples offer the ideal evaluation framework. PML and HMA parser reliability results are provided in **Table 5**.

Recommendations for appropriate ICC thresholds vary somewhat across disciplines and contexts. The threshold of "low" agreement can be from below ICC = 0.04 [20] to ICC = 0.05 [6]. Fair to moderate agreement thresholds vary the most with recommend ranges from ICC = 0.40 to ICC = 0.75 [11]. Most ICC schemata accept ICC > 0.6 as fair to good and ICC > 0.75 as good to excellent.

Table 6 compares the number of high, medium, and low-level conflicts identified by the human rater and the HMA and PML parsers. In all categories, the human rater identifies significantly more COI than either of the automated parsers. However, our work to date strongly suggests that additional training of the PML model can bridge much of this gap for both parser types. Interestingly, while the HMA parser performed more reliably across categories, the pure ML parser outperformed the HMA parser for medium-level conflicts. This suggests that with sufficient training, our approach to node classification would be applicable in cases where there is no metadata available to assist the parser.

4 Implications/Applications

Ultimately, these data suggest that both the PML and HMA parsers have the potential to be extended productively both for additional research on COI and more broadly in the digital humanities. The data produced by the parsers can be readily converted into a nodes and edges table for subsequent visualization using one of many network visualization platforms. For example, **Figure 3** offers a network maps of COI in two biomedical research areas, opioids and HIV. The assemblages allow one to discern certain funding patterns that may be useful for further research into the influence of COI on biomedical research. For example, the opioid network

Table 5: PML and HMA parser reliability measured by average ICC, with lower and upper bounds for the 95% confidence interval.

COI Level	PML Parser			HMA Parser		
	ICC	Lower	Upper	ICC	Lower	Upper
Low	.772	.745	.797	.722	.69	.751
Medium	.834	.814	.852	.773	.747	.797
High	.506	.458	.656	.618	.578	.656

Table 6: Number of COI identified by human rater or parser.

	Hi	Medium	Low
Human	345	505	1046
HMD	192	351	552
PML	203	446	530

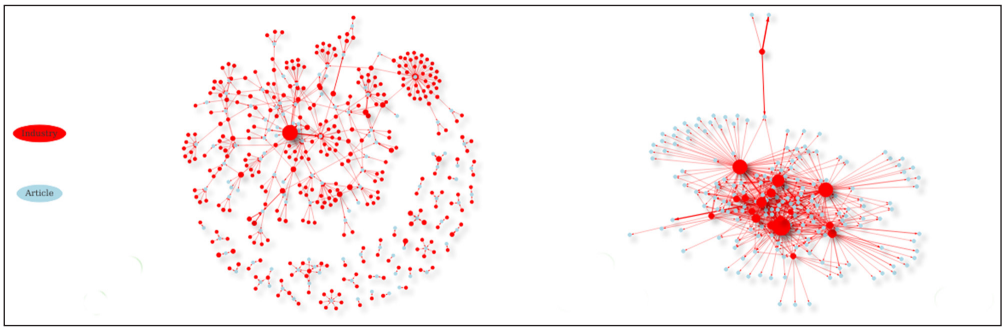

Figure 3: Network map of COI in opioid (left) and HIV (right) research.

map shows that the conflict network is relatively diffuse. However, a single large central node in the primary network neighborhood indicates that a significant proportional of COI are generated by a single entity, in this case (Pfizer). In contrast, the more densely articulated HIV network shows that there is simply a greater variety of industry entities involved in supporting researchers. The nodes for Gilead, ViiV, Merck, and AbbVie each demonstrate significant influence. Readers can create and explore dynamic COI network maps using the article network explorer at conflictmetrics.com. However, even dynamic network maps can be difficult to read and understand. While it is beyond the scope of this article, future digital humanities scholarship might explore supporting visualizations that can aid readers in processing complex network visualizations.

The data made available as a result of the T2V project provide multiple opportunities for humanistic inquiry in and around health and medicine. Future projects that map COI networks may provide insight into the particularities of funding circulation for different drugs, drug classes, or conditions. However, the greater potential for new insights will probably come from connecting these data with findings from other research projects. Future research might find that certain network profiles are associated with higher costs of care. It is also possible that we might discover that certain types of funding networks tend to dominate in biomedical areas marked by extreme socio-economic or racial disparities. Additionally, it is possible that similar network properties might be associated with gendered clinical practices. Fully evaluating the extent to which finding networks are related to other issues in biomedical research and clinical practice will require sustained interdisciplinary scrutiny. We hope that making these data widely available will help advance efforts in this area.

Beyond the particulars of industry funding and biomedical research, the results presented here suggest that this approach to extracting network data from unstructured text may be fruitful for other applicants in the humanities. Returning briefly to our example of relationship mining in Dostoyevsky's *The Brothers Karamazov*, the hybrid HMA approach could allow researchers to use a character permutation dictionary similar to our author permutations dictionary. Such a dictionary would allow the parser to know that Alexei Karamozov is the same entity as Alyosha, Alyoshka, Alyoshenka, Alyoshechka, Alexeichik, Lyosha, and Lyoshenka. Additionally, a customized Regex relationship dictionary could allow researchers to plot particular affiliations of interest for each of the characters. Of course, such work need not be limited to particular literary genres like the novel. New horizons of inquiry for this approach might include exploring intertextuality and/or citation-like attributions in texts that predate broadly accepted citation conventions, investigating Burkean ratios in dramatic texts [5], or locating and taxonomizing statements of moral obligation in ethical deliberation. Ultimately, the results presented here suggest there may be many promising future uses for the T2V approach.

Note

[1] Information on downloading MEDLINE XML files is available here: https://www.nlm.nih.gov/databases/download/pubmed_medline.html.

Additional File

The additional file for this article can be found as follows:

- **"Conflict of Interest: Article XML"**, https://doi.org/10.18738/T8/VSWAJY, Texas Data Repository Dataverse, V1. DOI: https://doi.org/10.5334/johd.21.s1

Competing Interests

The authors have no competing interests to declare.

Author roles

Graham: conceptualization, resources, data curation, software, formal analysis, supervision, funding acquisition, validation, investigation, visualization, methodology, project administration, writing—original draft, writing—review & editing.

Majdik: data curation, software, formal analysis, validation, investigation, methodology, writing—original draft, writing—review & editing.

Clark: data curation, software, formal analysis, validation, investigation, methodology, writing—original draft, writing—review & editing.

References

1. **Ahn R, Woodbridge A, Abraham A, Saba S, Korenstein D, Madden E, Boscardin WJ, Keyhani S.** Financial ties of principal investigators and randomized controlled trial outcomes: cross sectional study. *The BMJ*. 2017; 356: i6770. DOI: https://doi.org/10.1136/bmj.i6770
2. **Barad K.** *Meeting the universe halfway: Quantum physics and the entanglement of matter and meaning*; 2006. Durham, NC: Duke University Press. DOI: https://doi.org/10.1215/9780822388128
3. **Bartko JJ.** The intraclass correlation coefficient as a measure of reliability. *Psychological Reports*. 1966; 19(1): 3–11. DOI: https://doi.org/10.2466/pr0.1966.19.1.3
4. **Brando C, Frontini F, Ganascia J-G.** REDEN: named entity linking in digital literary editions using linked data sets. *Complex Systems Informatics and Modeling Quarterly*. 2016; 7: 60–80. DOI: https://doi.org/10.7250/csimq.2016-7.04
5. **Burke K.** *A grammar of motives*; 1945. Berkeley, CA: University of California Press.
6. **Cicchetti DV.** Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological assessment*. 1994; 6(4): 284–290. DOI: https://doi.org/10.1037/1040-3590.6.4.284
7. **Deleuze G, Guattari F.** *A thousand plateaus: Capitalism and schizophrenia*; 1988. London: Continuum.

8. **Dostoyevsky F.** *The brothers Karamazov*. Project Gutenberg Edition. Translated by Constance Garnett; 2009. New York, NY: The Lowell Press. URL: https://www.gutenberg.org/files/28054/28054-h/28054-h.html.
9. **Pattuelli MC, Miller M.** Semantic network edges: a human-machine approach to represent typed relations in social networks. *Journal of Knowledge Management*. 2015; 19(1): 71–81. DOI: https://doi.org/10.1108/JKM-11-2014-0453
10. **Ernst FR, Barr P, Elmor R, Sandulli W, Thevathasan L, Sterman AB, Goldenberg J, Vora K.** The economic impact of levothyroxine dose adjustments: the CONTROL HE study. *Clinical Drug Investigations*. 2017; 37(2): 71–83. DOI: https://doi.org/10.1007/s40261-016-0462-3
11. **Fleiss J.** *The design and analysis of clinical experiments*; 1986. New York, NY: Wiley. DOI: https://doi.org/10.1002/9781118032923
12. **Fohner AE, McDonagh EM, Clancy JP, Carrillo MW, Altman RB, Klein TE.** PharmGKB summary: ivacaftor pathway, pharmacokinetics/pharmacodynamics. *Pharmacogenetics and Genomics*. 2017; 27(1): 39–42. DOI: https://doi.org/10.1097/FPC.0000000000000246
13. **Graham SS, Majdik ZP, Clark D, Kessler MM, Hooker TB.** Relationships among commercial practices and author conflicts of interest in biomedical publishing. *PLoS ONE*. 2020; 15(7): e0236166. DOI: https://doi.org/10.1371/journal.pone.0236166
14. **Grandjean M.** A conceptual framework for the analysis of multilayer networks. *Paper presented at Digital Humanities 2020*, 22–24 July 2020; 2020. Ottawa, ON, Canada. URL: https://halshs.archives-ouvertes.fr/halshs-02650245.
15. **Haraway D.** *Modest_Witness@Second_Millenium.FemaleMan©_Meets_OncoMouse™: Feminism and Technoscience*; 1997. New York, NY: Routledge.
16. **Hohl C, Bro A, Friedberg RB.** Pain reduction in the recanalization of chronic iliofemoral venous occlusion with a new scoring balloon: a retrospective Series in 10 consecutive patients. *RöFo*. 2017; 189(11): 1086–1089. DOI: https://doi.org/10.1055/s-0043-111891
17. **Honnibal M.** *Introducing spaCy*; 2015. Retrieved from https://explosion.ai/blog/introducing-spacy.
18. **Honnibal M, Montani I, Van Landeghem S, Boyd A.** *Introducing spaCy v. 3.0 nightly*; 2020. Retrieved from https://explosion.ai/blog/spacy-v3-nightly.
19. **International Committee of Medical Journal Editors.** *Conflicts of interest*; 2019. Retrieved from http://www.icmje.org/conflicts-of-interest/.
20. **Koo TK, Li MY.** A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*. 2016; 15(2): 155–163. DOI: https://doi.org/10.1016/j.jcm.2016.02.012
21. **Latour B.** *Science in action: How to follow scientists and engineers through society*; 1987. Cambridge, MA: Harvard University Press.
22. **Findlen P, Edelstein D, Coleman N.** (eds.). *Mapping the Republic of Letters*; 2013. Retrieved from http://republicofletters.stanford.edu/.
23. **Warren CN, Shore D, Otis J, Wang L, Finegold M, Shalizi C.** Six degrees of Francis Bacon: A statistical method for reconstructing large historical social networks. *DHQ: Digital Humanities Quarterly*. 2016; 10(3). DOI: https://doi.org/10.17613/M6B020

How to cite this article: Graham SS, Majdik ZP, Clark D 2020 Methods for Extracting Relational Data from Unstructured Texts Prior to Network Visualization in Humanities Research. *Journal of Open Humanities Data* 6: 8. DOI: https://doi.org/10.5334/johd.21

Published: 19 November 2020

]u[*Journal of Open Humanities Data* is a peer-reviewed open access journal published by Ubiquity Press

OPEN ACCESS