Unbound: A Journal of Digital Scholarship
Digital Frontiers 2019: Tear Down the Walls
Vol. 2, No. 1, 2022

Transparency to Visibility (T2V)

Network Visualization in Humanities Research

S. Scott Graham, Zoltan Majdik, Dave Clark

Introduction

Humanities researchers have long studied how power 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 analytic artifacts 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 and toolkits 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 1) provide a brief summary of the current state of network visualization methods in the digital humanities; 2) describe the exigencies for the current project, and 3) detail our approach to network data extraction.

ISSN 2687-9018 (online)
https://journals.library.unt.edu/index.php/unbound/about

Background

Humanities Network Modeling: One of the central intellectual projects in the humanities over the past several decades has been developing robust theoretical accounts of power and influence within relational assemblages. 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 modeling can combine different graphical algorithms and other visual treatments to help make certain network features more visible. For example, a network map with a large central node indicates that the node in question exercises relatively more influence over the other nodes in the network than the peripheral nodes exercise over the central node.

Research using network modeling 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. While there are often significant and possibly irreconcilable differences among the various intellectual approaches available, Gilles Deleuze and Felix Guattari's rhizomatic theory,[1] Donna Haraway's technoscientific networks,[2] Bruno Latour's actor-networks,[3] and Karen Barad's theory of intra-action[4] (among many others) all highlight the importance of understanding the nature of relations and the types of circulation made possible within complex systems. These particular theoretical constructs are especially well-attuned to investigating network features like articulation density and complexity as primary sources of power and influence. Whether it is Latour's analysis of mundane objects,[5] Haraway's interrogation of transuranic elements,[6] Fox Keller's exploration of the material-semiotics of the gene,[7] or Barad's account of theoretical physics,[8] the importance of relationality among human, natural, technical, and economic systems is paramount. Multiple pathways of influence allow participants in complex networks to more effectively leverage multiple points of control and shift among them when a given program of action meets resistance. A multiplicity of social and/or economic connections allows for a broader range of more dynamic responses to changes in a given network.

Irrespective of the chosen theoretical construct or the ultimate aims of the inquiry, recent advances in visualization software provide researchers with new opportunities to better explore circulation within networks and cultural systems. Indeed, bibliometric, media studies and archival digital humanities scholarship have already made great strides in these areas. In recent years humanities journals have seen a veritable explosion in network mapping methodologies as applied to social media discourse, scholarly citation networks, and all manner of archival materials. However, those areas with the greatest attention no doubt owe that attention, in part,

---

[1] Gilles Deleuze and Félix Guattari, *A thousand plateaus: Capitalism and schizophrenia*. Bloomsbury Publishing, 1988.
[2] Donna Haraway, *Modest Witness@ Second_Millenium. FemaleMan© _Meets_On coMouse™: Feminism and Technoscience* (New York and London: Routledge, 1997.
[3] Bruno Latour, *Science in action: How to follow scientists and engineers through society*. Harvard university press, 1987.
[4] Karen Barad, *Meeting the universe halfway: Quantum physics and the entanglement of matter and meaning*. duke university Press, 2007.
[5] Bruno Latour, "Where are the Missing Masses? The Sociology of a Few Mundane Artifacts. Shaping technology/building society: studies in sociotechnical change. WE Bijker and J. Law." WE Bijker, J. Law, & American Council of Learned Societies (Eds.), *Shaping technology/building society: studies in sociotechnical change* (1992): 225-258.
[6] Haraway, ModestWitness, 53-55.
[7] Evelyn Fox Keller, *Refiguring life: Metaphors of twentieth-century biology*. Columbia University Press, 1995.
[8] Barad, Meeting.

to the ability to easily access data amenable to network visualization. Facebook friend networks, retweet networks, and citation networks, for example, are particularly easy to submit to network modeling 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*[9] project leverages Oxford's *Electronic Enlightenment Project* to visualize the geography of correspondence networks for key enlightenment thinkers. Much of this project revolves around digitizing the structured metadata from each letter (sender name, recipient name, mailing addresses, date, etc.).

A significant challenge for many humanities projects with respect to network modeling is that "data" is frequently neither retrievable nor structured. A scholar attempting to model the social networks in *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 modeling requires knowing that Alexei and Alyosha are the same person. Likewise, transforming the novel 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? Ultimately, establishing that Alexei and Alyosha are the same person and that he is Fyodor's son is easy if you are human, but challenging to implement computationally.

In sum, there are three key challenges that remain to be addressed before network modeling can be more widely and effectively adopted in the humanities: 1) Humanities researchers need methods and toolkits that support consistent and reliable identification of nodes in unstructured text. 2) Humanities researchers need approaches and techniques for determining when identified nodes are "in" a relationship. And, 3) Network modeling 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. As one would expect, some fairly advanced tools involving machine learning and/or NLP are required to meet these aims. The REDEN framework,[10] 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. REDEN thus 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.[11] 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 problems 1 and 2 above, the challenge of classifying relationships remains. The potential scale and scope of this challenge is exemplified in Pattuelli and Miller's "Semantic

[9] Mapping the Republic of Letters. http://republicofletters.stanford.edu/, 2013.
[10] Carmen Brando, Francesca Frontini, and Jean-Gabriel Ganascia, "REDEN: named entity linking in digital literary editions using linked data sets." (2016).
[11] Christopher N. Warren, Daniel Shore, Jessica Otis, Lawrence Wang, Mike Finegold, and Cosma Shalizi, "Six Degrees of Francis Bacon: A Statistical Method for Reconstructing Large Historical Social Networks." *DHQ: Digital Humanities Quarterly* 10, no. 3 (2016).

network edges: a human-machine approach to represent typed relations in social networks."[12] They too used an NER-based framework for node identification but ended up crowd-sourcing edge classification.

The T2V Project: Ultimately, the primary aim of the T2V project is to develop a method and toolkit for transforming unstructured text into relational network data. We opted to prototype our toolkit using conflict of interest 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.

In recent years, there has been increasing recognition that public disclosure of conflicts of interests is an essential part of efforts to safeguard against financial biases in health and medicine. Accordingly, disclosure laws like the Sunshine Act highlight the centrality of "transparency" in public accountability efforts. This focus on transparency is manifest in a wide variety of accountability efforts ranging from journal conflict of interest disclosure statements to databases like OpenSecrets.org, which tracks campaign finance data for American politicians. However, recent research in the humanities and social sciences suggests that transparency efforts, alone, are not enough. Indeed, a growing body of evidence indicates that conflict of interest disclosure statements may result in unintended and pernicious effects.[13] For example, disclosure statements have been shown to cause audiences to extend more trust to those holding conflicts of interest as disclosure provides an opportunity to display both honesty and expertise. Conflict disclosure can also lead to "moral licensing," a phenomenon whereby those who disclose conflicts become unduly confident in their objectivity because transparency obligations have been fulfilled.[14] In order to properly leverage disclosure statements in humanities research, scholars need not only access to financial relationship data, but also the means to analyze and present this data in ways that will be useful for both scholarly endeavors and to educate the broader public. Network visualization has great potential to be useful here, but since disclosure statements exist in a wide variety of unstructured prose formats, it is quite difficult to extract relationship data systematically.

A primary challenge to this work comes from the diversity of style guides for reporting conflicts of interest. Different journals might render the same conflict of interest quite differently. For example, various conflicts of interest 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

[12] Biswanath Dutta, Devika P. Madalli, M. Cristina Pattuelli, and Matthew Miller, "Semantic network edges: a human-machine approach to represent typed relations in social networks." *Journal of Knowledge Management* (2015).

[13] Andreas Lundh, Joel Lexchin, Barbara Mintzes, Jeppe B. Schroll, and Lisa Bero. "Industry sponsorship and research outcome." *Cochrane Database of Systematic Reviews* 2 (2017).

[14] Daylian M. Cain, George Loewenstein, and Don A. Moore, "The dirt on coming clean: Perverse effects of disclosing conflicts of interest." *The Journal of Legal Studies* 34, no. 1 (2005): 1-25.

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.

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 1). In order to effectively evaluate conflicts of interest, there must also be a way of normalizing differential 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.

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*

Table 1: Integrated Nodes and Edges Table derived from Conflict of Interest Statement.

Data Extraction and Parser Development

Our data comes from the MEDLINE database,[15] 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. PubMed, a service of the US National Institutes of Health, provides several protocols for accessing the MEDLINE database. The most well-known is the search engine at pubmed.com, but API and FTP interfaces are also available. To begin our study of conflict statements, we downloaded all MEDLINE XML files via the FTP locker. We then used a customized XML parser to load selected data on each of the 30 million indexed publication 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. (Articles are available at www.ncbi.nlm.nih.gov/pubmed/[insert PMID].) MEDLINE only began collecting conflicts of interest information in 2016, and not all journals participate in the program by reporting author conflicts of interest. Thus, of the 30 million collected articles, only 274,246 included conflicts of interest statements. Our analysis indicates that those 274,246 have a total of 159,878 individual conflicts of interest. 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 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. We refer to each version of the parser as the Pure Machine Learning (PML) Parser and the Hybrid-Metadata Assisted (HMA) Parser, respectively.

In short, the toolkit 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 conflicts of interest. For example, this sentence in the below COI:

"Simon Knight has received consultancy fees from OrganOx UK Ltd" is parsed into

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

Table 2: Simple COI Statement Parsed.

Those extracted conflicts are then passed to post-processing models that clean the data and render it in node and edge tables. Below, the individual components of the parser are described in more detail. Following the detailed explanation of parser components, we describe a more complicated parsing example.

Source Identification/Sponsor Tagger. A Natural Language Processing (NLP) method called Named Entity Recognition (NER) can reliably use grammatical and/or statistical techniques to extract and classify proper nouns, numbers, and dates from unstructured text. A sentence such as "Walter Sandulli and Jessica Goldenberg are employees of Akrimax," when parsed, would produce three "named entities":

[15] National Library of Medicine. PubMed Overview. https://www.nlm.nih.gov/bsd/pubmed.html. 2019.

Walter Sandulli, PERSON
Jessica Goldenberg, PERSON
Akrimax, ORG

NER approaches can work with significant accuracy on unknown texts and can achieve near-human levels of precision when trained using a machine learning approach. In the case of conflict of interest statements, the lack of consistent styling in the writing and editing of COI statements means that organization names are presented very differently, sometimes within the same COI statement (e.g., GlaxoSmithKline vs. Glaxo vs. GSK). COI statements are similarly inconsistent in presenting author names; often they use initials, but sometimes last names or other abbreviations will be present. Building a training corpus that is specific to the data set being studied can significantly improve the ability of the NER to correctly sort author names from organization names and present the organization names consistently.

Author Tagging/ Target Identification: In the metadata-assisted version of the parser, we used MEDLINE data on author names to increase recognition accuracy. To do so, this parser generates an author-name permutation table with 13 name permutations that correspond to author naming conventions from various journals. "Jane Alicia Doe," for example, would be rendered as "J.A.D.," "J. Doe," "J Doe," and ten other permutations of first, middle, and last name.

Relationship Types/ COI Classification Dictionary: The COI classification dictionary is based loosely on the International Committee of Medical Journal Editors (ICMJE) standardized conflicts of interest disclosure form. The ICMJE form is used by many major medical journals around the world and taxonomizes conflicts into five primary areas: 1) grant, 2) personal fees, 3) non-financial support, 4) other, and intellectual property. ICMJE guidance[16] 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.

[16] International Committee of Medical Journal Editors (ICJME). Conflicts of interest. http://www.icmje.org/conflicts-of-interest/. 2019.

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 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)'
```

Relationship Extraction: The parser assumes a standard syntax that almost all COI disclosure statements follow, where a name (or names) are followed by a COI disclosure type (like "is employed by"), which is followed by the COI source (the entity creating the conflict of interest). The parser extracts COI value(s) from each COI statement by stitching the three elements described above---NER, author permutations, COI classifications---together through Regex. For each PMID, (1) the parser first runs the COI disclosure through a spaCy NER function, which tags organizations through the updated language model, cleans results (e.g., removes words like "Inc."), and checks them against the complete author list. This last step helps avoid false positives in the NER tag list: because it can be difficult for an NLP/NER tagger to reliably identify a name like "Novartis" as ORG rather than PERSON, having a canonical author list against which to check ORG tags (and exclude them if they are matched against an author in the author list) provides cleaner data. (2) If ORG tags are present after these cleaning steps, a regular expression checks if any author name permutations associated with the PMID are followed by any COI term from the COI classification dictionary within 80 words, but not outside a sentence boundary. If so, (3) the regular expression checks if the author name permutation and COI word are followed, within the same sentence boundary, by the sponsor marked with the ORG tag.

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 3 shows the result of our parser's work on the example data from "Defining Priorities for Future Research: Results of the UK Kidney Transplant Priority Setting Partnership" (PMID: 27776143). The goal of the extraction is to parse the unstructured conflict of interest statements into a relatively standardized table of *sources* (e.g., names of pharmaceutical companies), *targets* (e.g., names of individual researchers), and *relationship types* (e.g., employment or grant funding). Each type of node requires a slightly different strategy to reduce ambiguity and inconsistency.

Target	**Relationship Type**	**Source**	**Conflict Weight**
Simon Ball	grant	"Oxfod"	2
Simon Knight	fees	"OrganOx UK"	1
Lorna Marson	fees	Novartis	1
Lorna Marson	fees	Astellas	1

Fiona Loud	fees	Merck	1
Graham Lipkin	fees	"Raptor Pharmaceuticals"	1
Graham Lipkin	fees	Alexion Pharma	1

Table 3: Complex COI Statement Parsed.

Parser Evaluation

There are many approaches to evaluating text analysis protocols. While precision and recall metrics are among the most popular, we opted for a machine-human interrater reliability approach, using an intraclass correlation coefficient (ICC). ICC metrics were originally developed to assess the extent to which human judgments were consistent and reliable across a pool of raters.[17] Since the ultimate goal of the T2V parser 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. In order to evaluate the effectiveness of the T2V parser, 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 conflicts of interest 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.

Recommendations for appropriate ICC thresholds vary somewhat across disciplines and contexts. The threshold of "low" agreement can be from below ICC = 0.04[18] to ICC = 0.05.[19] Fair to moderate agreement thresholds vary the most with recommend ranges from ICC= 0.40 to ICC = 0.75.[20] Most ICC schemata accept ICC > 0.6 as fair to good and ICC > 0.75 as good to excellent. Since identifying that no conflicts are present is an easier computational task than conflict 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. Nevertheless, parser reliability scores generally fell within ranges that would be classified as moderate to good.

HMA Parser: Using these ranges as a guide, the HMA parser was found to have a moderate to high degree of reliability between human and machine rating for each COI category. The average measure ICC for low-level conflicts was 0.722, with a 95% confidence interval from 0.69 to 0.751 ($F[998,903[= 6.27$, $p < .01$). The average ICC for medium weight conflicts was 0.773, with a 95% confidence level from 0.747 to 0.797 ($F[998,985] = 7.84$, $p < .01$). And, finally, the average ICC for high-level conflicts was 0.618, with a 95% confidence level from 0.578 to 0.656 ($F[998,923] = 4.28$, $p < .001$).

---

[17] John J. Bartko, "The intraclass correlation coefficient as a measure of reliability." *Psychological reports* 19, no. 1 (1966): 3-11.
[18] Terry K. Koo, and Mae Y. Li "A guideline of selecting and reporting intraclass correlation coefficients for reliability research." *Journal of chiropractic medicine* 15, no. 2 (2016): 155-163.
[19] Domenic V. Cicchetti, "Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology." Psychological assessment 6, no. 4 (1994): 284.
[20] Joseph Fleiss, The Design and Analysis of Clinical Experiments. Wiley, New York.1986.

PML Parser: In contrast to the ML+MD parser, the pure ML parser had a wider range of reliability scores. The average ICC for low-level conflicts was 0.772, with a 95% confidence interval from 0.745 to 0.797 ($F[998,916] = 7.86$, $p < 0.01$). The average ICC for medium weight conflicts was .834, with a 95% confidence interval ranging from 0.814 to 0.852 ($F[998,998] = 11$, $p < .01$). And, the average ICC for high-level conflicts was 0.506, with a 95% confidence interval ranging from 0.458 to 0.656 ($F[998,986] = 3.06$, $p < .01$).

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

Table 4: Number of Conflicts of Interest Identified by Human Rater or Parser

Table 4 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 conflicts of interest 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.

Applications and Future Directions

Ultimately, these data suggest that both the PML and HMA parsers have the potential to be extended productively both for additional research on conflicts of interest 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. The assemblages allow one to discern certain funding patterns that may be useful for further research into the influence of conflicts of interest on biomedical research. For example, the opioid network 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 conflicts of interest are generated by a single entity, in this case (Pfizer). In contrast, the more densely articulated HIV network shows that there are simply a greater variety of industry entities involved in supporting researchers. The nodes for Glided, ViiV, Merck, and AbbVie each demonstrate significant influence. Future scholarship in this area may be able to tie network features (such as network centrality or heterogeneity) to drug safety profiles. Results from this kind of inquiry might support more effective conflicts of interest management polices than current disclosure requirements.

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 *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 aesthetic forms 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, 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.