Extracting Relational Network Data from Unstructured Texts

Lessons from the Transparency-to-Visibility (T2V) Project

S. Scott Graham, University of Texas at Austin
Zoltan P. Majdik, North Dakota State University
Dave Clark, University of Wisconsin-Milwaukee
Tristin Brynn Hooker, University of Texas at Austin

NATIONAL ENDOWMENT FOR THE Humanities

XSEDE
Extreme Science and Engineering Discovery Environment

SAGA Simulation and Game Applications Lab
THE UNIVERSITY OF TEXAS AT AUSTIN

Table of Contents

Executive Summary

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 (nodes and edges tables) 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 the kinds of data structures that support useful and informative network visualization. The Transparency to Visibility (T2V) Project was initiated to pursue these goals. Accordingly, the T2V team used bioethics accountability statements to pilot and evaluate different methods for transforming and visualizing relational networks in unstructured text. The resulting machine-learning-enhanced natural language processing and metadata-assisted approaches offer promising potential pathways for contemporary digital humanities and future toolkit development.

Introduction

In September 2018, the T2V project was awarded a National Endowment for the Humanities (NEH) Level II Digital Humanities Advancement Grant to fund the continued development and enhancement of a prototype relational network data extraction and visualization toolkit. The project aim was to integrate two pre-existing toolkit prototypes developed by the research team: 1) A Medical Conflicts of Interest Analyzer that extracts financial relationship information from the text of medical journals, and 2) A Clinical Trials Sponsors Network Visualizer that creates relational network graphs from financial relationships data. The T2V team of scholars in digital humanities, medical humanities, rhetoric, and library studies used a dataset of approximately 275,000 conflicts-of-interest disclosure statements indexed by PubMed to develop a prototype of the new technology and methodology. The primary aims of developing the integrated T2V toolkit were three-fold:

- to develop and promulgate an intellectual framework, grounded in humanistic scholarship, for visualizing complex systems (including financial networks) related to matters of humanistic and public concern.
- to allow medical humanities scholars to access and visualize large datasets on funding and conflicts of interest for a variety of biomedical domains.
- to use a flexible, open-source infrastructure to support easy redevelopment for a wide variety of humanities research projects, both within and beyond biomedicine.

The T2V team pursued these three aims in such a way as to support the ongoing development of robust humanistic inquiry into relational networks and complex systems. Humanists have long explored how power and influence circulate through discursive and cultural systems, and recent digital humanities achievements and related advances in network visualization tools have the potential to substantively support this work. However, a key challenge for humanities scholars is that the novels, epistolary archives, biographies and other texts that are common humanities artifacts are difficult to map as networks. Popular network visualization toolkits such as Gephi, or certain R and Python packages rely on coordinated data tables that provide a comprehensive list of network entities (nodes) with unique identifiers and map the relationships (edges) among those entities. Yet, relationships of interest to humanities researchers seldom come structured into succinct nodes and edges tables. The unstructured prose of novels, biographies, letters, and journals must be transformed, first, into appropriate data tables before network visualization is possible. The T2V project was designed to develop a framework supporting just this kind of work. In what follows, this whitepaper:

- Describes the intellectual and technological exigencies for the T2V project;
- Outlines our approach to extracting relational network data from unstructured text;
- Summarizes T2V data visualization initiatives; and
- Makes recommendations for future research.

Background

Humanities Network Modeling: One of the central intellectual projects in the humanities over the past several decades has been to develop robust theoretical accounts of power and influence within relational assemblages. This kind of research 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 (1988) rhizomatic theory, Donna Haraway's (1997) technoscientific networks, Bruno Latour's (1987) actor-networks, and Karen Barad's (2007) 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 particular theoretical constructs are especially well-attended 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, Haraway's interrogation of transuranic elements, Fox Keller's (1995) exploration of the material-semiotics of the gene, or Barad's account of theoretical physics, 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 much of that attention, in part, 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 to 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. Much of this project revolves around digitizing the structured metadata from each letter (sender name, recipient name, mailing addresses, date, etc.).

Conflicts of Interest Statements: The T2V project focused on developing a toolkit that can aid humanities researchers by providing an easy-to-customize automated framework for converting unstructured text into nodes and edges, capturing the relationship among nodes using a combination of NER, machine learning, and regular expressions. As a test data set, T2V uses conflict of interest statements in medical publishing; these statements 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.

In recent years, there has been increasing recognition that public disclosure of potential conflicts of interest is an essential part of efforts to safeguard against financial biases in health and medicine (Lundh, et al., 2017). 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 (Cain, Loewenstein, & Moore, 2005). 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. In order to properly leverage disclosure statements in humanities research, scholars need access not only 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 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 relationships from this text:

> Frank Ernst are employees of Indegene, Inc.,
> Peri Barr are employees of Indegene, Inc.,
> Riad Elmor are employees of Indegene, Inc.,
> 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 are employees of Akrimax.
> Jessica Goldenberg are employees of Akrimax.
> Arnold Sterman has been a consultant for Akrimax,
> Arnold Sterman has contributed to research funded by Akrimax
> Arnold Sterman received an honorarium for his contributions to evaluating this study and to the development of this manuscript.

The identified relationships must then be parsed into source, target, and relationship 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
Akrimax	Walter Sandulli	Employment
Akrimax	Jessica Goldenberg	Employment
Akrimax	Arnold Sterman	Consulting
Akrimax	Arnold Sterman	Grant Funding

Table 1: Fully-parsed relationships from sample COI statement (above)

T2V Data Extraction

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. 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 https://www.ncbi.nlm.nih.gov/pubmed/[insert PMID].)

Table	Columns
Article Metadata	PMID Title Publication Date
Author Data	PMID Last Name First Name Middle Initial
Conflict Statements	PMID Statement
Journal	PMID Journal Name

Table 2: T2V Database schema

For example, "Defining Priorities for Future Research: Results of the UK Kidney Transplant Priority Setting Partnership," which is available at https://www.ncbi.nlm.nih.gov/pubmed/27776143, looks like this in our database:

Table	Column Data
Article Metadata	27776143 Defining Priorities for Future Research: Results of the UK Kidney Transplant Priority Setting Partnership. October 2016
Author Data	27776143 Knight Simon R Metcalfe Leanne O'Donoghue Katriona Ball Simon T Beale Angela Beale William Hilton Rachel Hodkinson Keith Lipkin Graham W Loud Fiona Marson Lorna P Morris Peter J
Conflict Statements	27776143 Angela Beale, William Beale, Leanne Metcalf, Keith Hodkinson, Peter Morris and Katriona O'Donoghue have no conflicts of interest to declare. Simon Knight has received consultancy fees from OrganOx UK Ltd. Lorna Marson has received lecture fees from Astellas and Novartis. Fiona Loud has received consultancy fees from Merck and Galderma. Graham Lipkin has received lecture fees from Raptor Pharmaceuticals and consulting fees from Alexion Pharma. Rachel Hilton has received lecture fees from Roche Pharmaceuticals and consultancy fees from Novartis. Simon Ball has received research grants from Oxford Immunotec Ltd. This does not alter our adherence to PLOS ONE policies on sharing data and materials.
Journal	27776143 PloS one

Table 3: Sample data for a single entry (PMID: 27776143) in the T2V database

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 conflicts of interest. Among those articles with conflicts, each article has an average of 10 reported conflicts.

Parser Development

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. Each parser's logic model is below in Fig. 1 and Fig. 2.

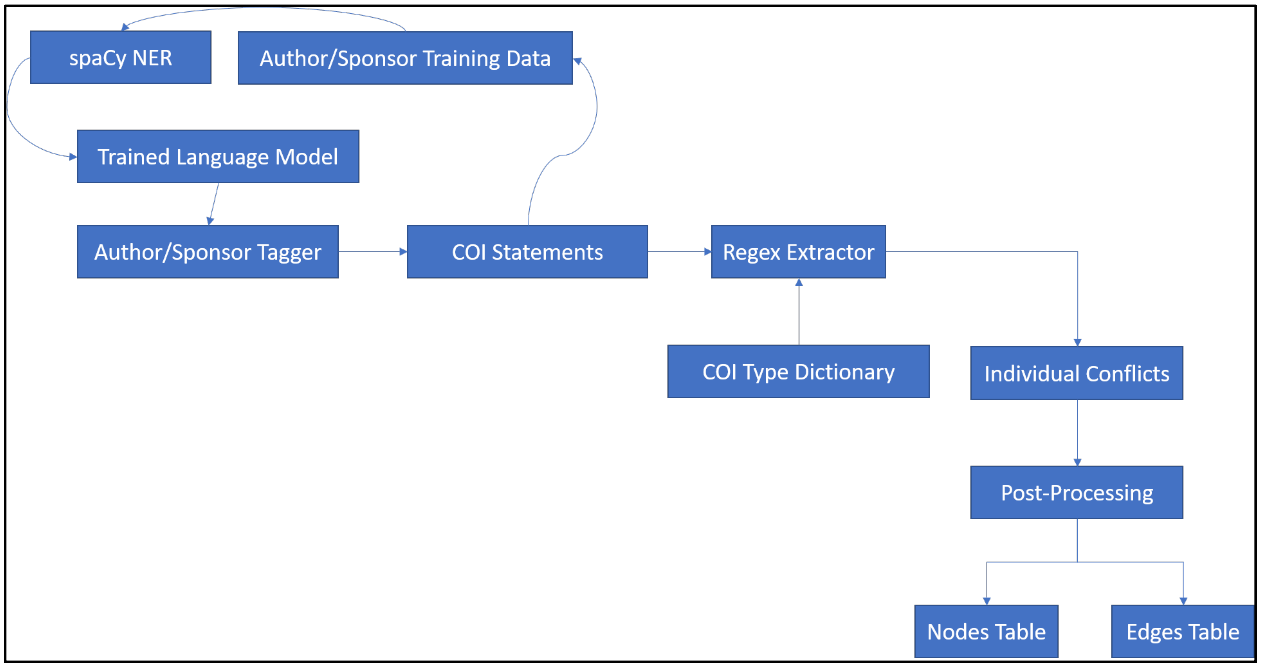


Fig. 1. PML Parser Logic Model

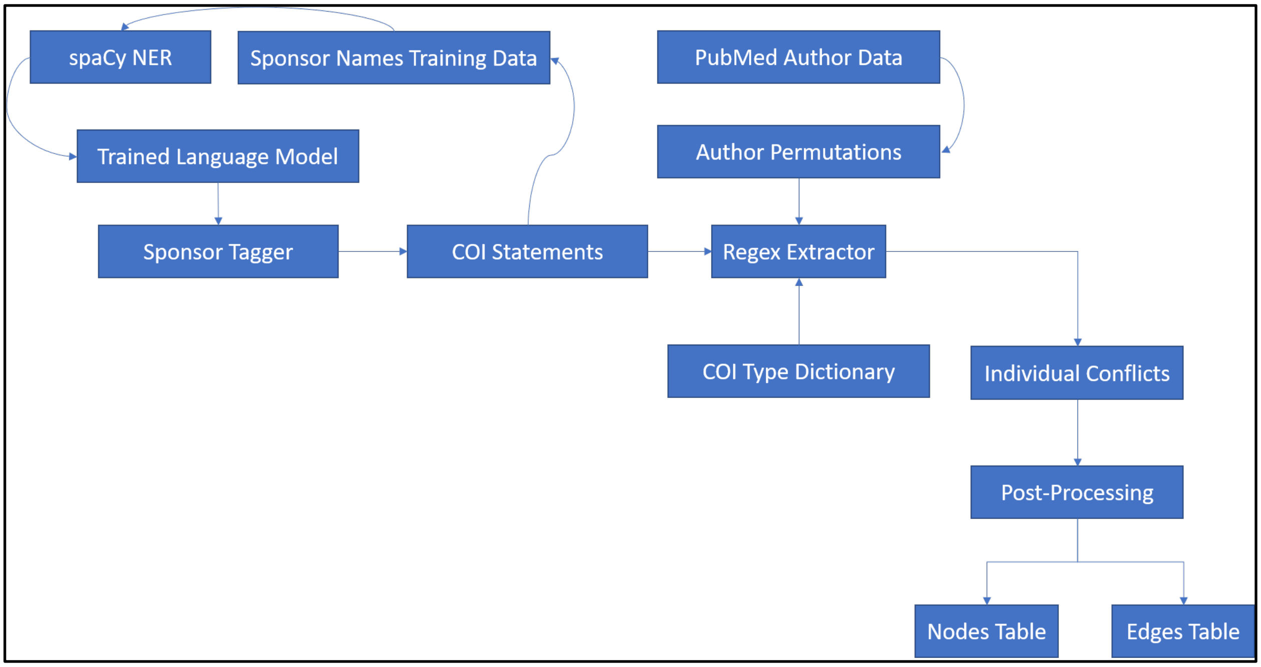


Fig. 2 HML Parser Logic Model

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 combines dictionaries of author name permutations and conflict types to extract individual conflicts of interest; we then used a dictionary of "conflict weights" to assign scores to particular kinds of conflict (e.g., an ownership interest is weighted more heavily than an honorarium). For example, this sentence in the above 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 5: Fully-parsed data for sample COI statement (above)

Those extracted conflicts are then passed to post-processing scripts 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":

> 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.

T2V relied on the spaCy library for NLP, and spaCy's NER feature can be trained by providing it with specially formatted, hand-selected COI statements that correctly identify and categorize named entities:

> ("IB and AJS were employees of Incyte at the time the research took place.", {"entities": [(0, 2, "PERSON"), (7, 10, "PERSON"), (29, 35, "ORG")]})

We were able improve sponsor recognition 25% using a small (n=100) training set. When the parser identifies an organization in a conflict statement, it then looks for author information that can suggest a target for a potential conflict of interest.

Author Tagging/ Target Identification: Our approach used MEDLINE data on author names to further 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 journal style guides for disclosure statements. “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 and/or initials. If the parser has identified a source organization in a conflict of interest, it then looks for sets of initials or partial names that might indicate an author name is present (e.g., the presence of “SK” in the text). It then compares that entity with the associated permutation table to pull the author’s full name (e.g., “SK” is swapped out, and “Simon Knight” is stored in the database.

Relationship Types/ COI Classification Dictionary: The COI classification dictionary is based loosely on the International Committee of Medical Journal Editors’ (ICMJE, 2017) 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 5) intellectual property. ICMJE 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, etc.
>
> **Other:** Anything not covered under the previous three boxes.
>
> **Intellectual Property:** Patents and copyrights.

Our COI dictionary schema organizes these categories (as well as 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 Regex parser described below.

```
coi_type_1 = 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+)'

coi_type_2 =
r'(?:grant|fund\w+\W|support\w+\W|contract\w+\W|collaborat\w+\W|research)'

coi_type_3 =
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, and COI classifications—together through a regular expression. For each PMID, (1) the parser first runs the COI disclosure through a custom 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 collated as rows in a new dataframe, and assigned a numerical weight based on the COI classification dictionary.

Target	Relationship Type	Source	Conflict Weight
Simon Ball	grant	Oxford	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 6: Fully-parsed disclosure statement for "Defining Priorities for Future Research: Results of the UK Kidney Transplant Priority Setting Partnership" (PMID: 27776143).

Table 6 above 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.

Parser Evaluation

There are many approaches to evaluate 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 (Bartko, 1966). 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 "parses like a human reader." Other digital humanities projects may be designed to perform tasks which would be impossible for human readers. 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 assess the reliability 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 50 words. Our curated 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 (Koo & Li, 2016) to ICC = 0.05 (Cicchetti, 1994). Fair to moderate agreement thresholds vary the most with recommend ranges from ICC= 0.40 to ICC = 0.75 (Fleiss, 1986). 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.

HMD Parser: Using these ranges as a guide, the hybrid ML+MD 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).

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 ranting from 0.458 to 0.656 (F[998,986] = 3.06 , p < .01).

	Hi	Med	Low
Human	345	505	1046
Hybrid ML-MD	192	351	552
Pure ML	203	446	530

Table 7: Number of Conflicts of Interest Identified by Human Rater or Parser

Table 7 (above) 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.

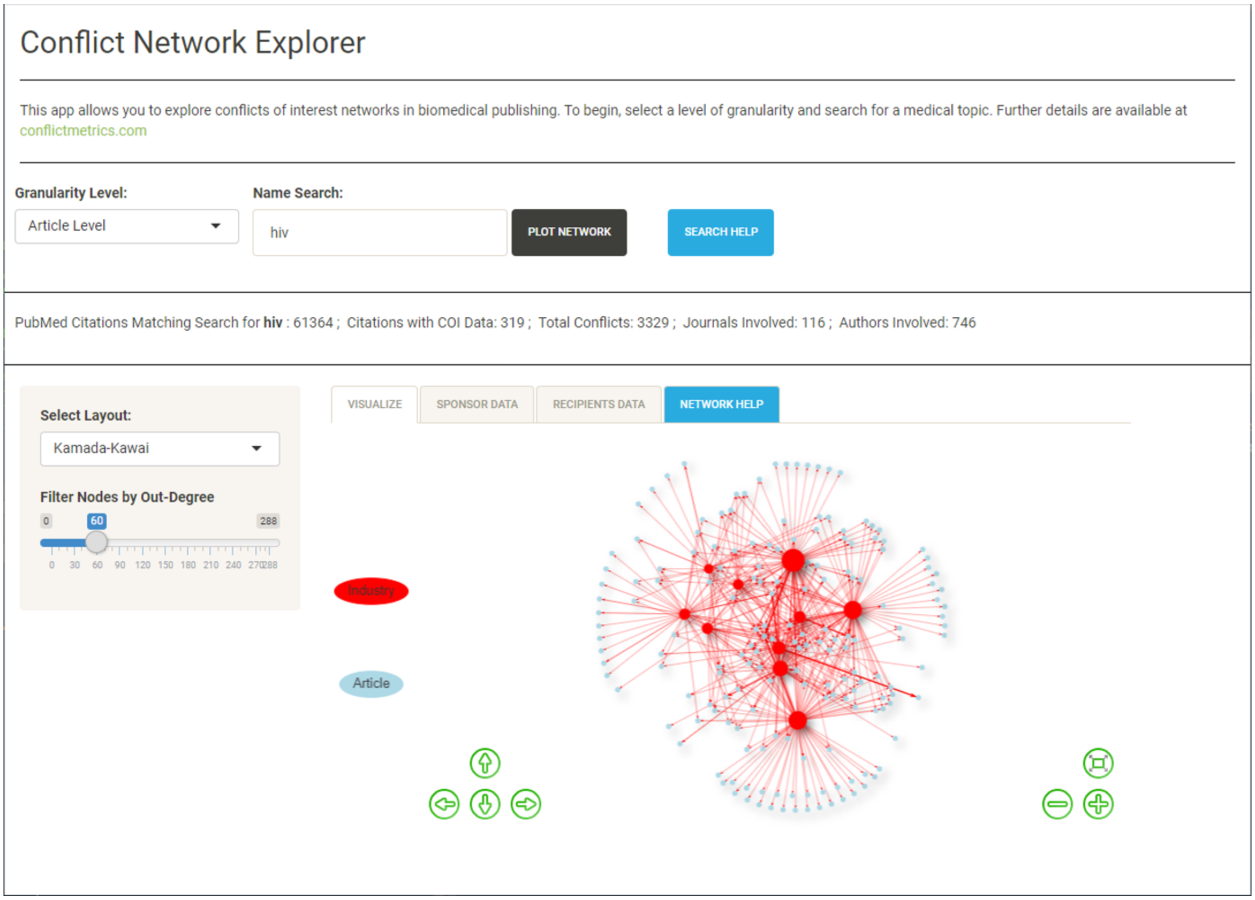
T2V Network Visualization

There is an increasingly wide variety of network modeling applications and toolkits available. Beyond proprietary options, Gephi, and certain third-party packages for R or Python are among the most popular. Given the overall aims of the T2V project, we required a framework that would allow for reliable, replicable, and repeated visualizations of a wide variety of biomedical COI networks. That is, for the T2V project to effectively communicate our ideas about how financial relationships circulate within biomedical subareas, we needed to adopt a framework that would allow users to specify subareas of interest and dynamically map the corresponding COI networks. We also wished to develop a framework that would allow users (regardless of technical proficiency) to create and manipulate network visualizations. Even though Gephi, among the most popular open-source options, allows for the creation of the most visually appealing network models, it is not designed to allow the easy creation of multiple networks based on freeform user queries. Ideal options for visualization creation and dissemination would require a server-side implementation. Two primary options included R Shiny and Python Flask, each of which allow developers to create R or Python environments that users can access through an HTML interface. Since R currently has a more robust suite of third-party network visualization packages, we opted to pursue that development pathway.

Sponsor Network Explorer App

With the assistance of the University of Texas at Austin Simulation and Gaming Applications (SAGA) Lab and additional support from XSEDE's jetstream infrastructure, we developed a Sponsor Network Explorer app using the R Shiny framework. The app framework allows users to enter a freeform query (relating to biomedical topics) in the app search box. The user query is routed through the National Center for Bioinformatics API which queries PubMed for post-2016 publications related to the topic and returns the 100,000 most relevant hits. The data returned are used to populate the SQL queries to the T2V COI database. R then renders the results returned as nodes and edges tables and subsequently creates the network model using the VisNetwork package.

VisNework is a third-party R package that creates dynamic, manipulable network visualizations. This package uses the vis.js javascript library and allows for user interactivity controls for display. Users can zoom in and out on the network, drag-and-drop nodes, and highlight network neighborhoods of interest. Additionally, we leveraged reactive functionality in the R Shiny framework to allow for increased user control of network visualizations. Specifically, we added an out-degree filter and the ability to change the network visualization algorithm.



Conflict Network Explorer

This app allows you to explore conflicts of interest networks in biomedical publishing. To begin, select a level of granularity and search for a medical topic. Further details are available at conflictmetrics.com

Granularity Level: Article Level

Name Search: hiv

PubMed Citations Matching Search for **hiv** : 61364 ; Citations with COI Data: 319 ; Total Conflicts: 3329 ; Journals Involved: 116 ; Authors Involved: 746

Select Layout: Kamada-Kawai

Filter Nodes by Out-Degree: 60

Fig. 3: Screen Shot of Conflict Network Explorer App

Visualization App Evaluation

We conducted three iterative rounds of user testing with both in-person and online protocols (see the appendix), with volunteers from a range of backgrounds similar to our likely user base. We identified our users as people who are likely to have basic library and database search skills, and who would be familiar with web-based search tools, but who come from a variable level of familiarity or expertise with biomedical research. As such, user testing volunteers include those who would have spent some time with related tasks, but who also had varying degrees of familiarity with the subject matter: the user-testers included two humanities PhD students, two yoga/fitness instructors, two corporate professionals (a web marketing director and an office manager who has done extensive PubMed research), and a licensed nurse practitioner who works in physician education in the pharmaceutical industry.

Developing the user-testing protocols: To develop the user-testing protocol, we followed best practices in user testing development, and asked users to first identify and describe their understanding of the purpose, ownership, and capabilities of the tool, as well as their perception of the level of expertise they would need to use it. We then asked users to perform a specific search for a specific kind of term (a medical condition), at a given level, and then asked users to identify and describe the results. We asked users both to identify the information that the search yielded, and to describe their understanding of the relationships between the information in the search (for example, the relationship between the number of available citations and the number of conflicts of interest and journals or authors involved). Finally, we asked users to filter the results by out-degree and to experiment with changing the layout of the graph. We asked users to interpret or describe their perceptions of how performing these functions altered their understanding of their search results.

Results: In the initial round of testing, our users had fairly consistent results. Most users could identify the general purpose of the site—that it was a tool for searching—but they had varying degrees of understanding of what they would be searching, for what purpose, or who the site belonged to/what it was part of. Most users were able to determine from the example search term given (leukemia), that the search terms should be related to "disease" or "medicine," but only the user who had previously conducted PubMed research identified that they were searching for conflicts of interest in research sponsorship and potential bias. Most users concluded that they would need to have some idea of what they were looking for, in both search terms and results, in order to use the tool effectively.

Additionally, most users did not spend much time working through the help navigations on the screen that linked them to the help/documentation sites on conflictmetrics.com, and most had questions about the terminology both for the search results, and for the filters and layout options. Some of the initial users did not realize there was a table beneath the graphic representation of search results, and, once they did, did not understand the meaning of the labels in the table or how they related to the graphic representation. Most users were able to identify the relevant data points that searches revealed, but struggled to articulate the relationship between citations, conflicts, journals, and authors in the graphic representation, particularly users that were not familiar with biomedicine.

From this testing, we concluded that the tool needed more clear help documentation on the site, more clear links back to the documentation on conflictmetrics.com, and a more noticeable/accessible explanation of the search terms. We also concluded that users might need to be directed to the table that displays beneath the graph, and that the table might need to be organized to show more consistent terminology with the graphic search results and to help provide more explanations. The tool was consequently revised with more explanation of its purpose at the top, a clearer link to conflictmetrics.com for further information, links to definitions of terminology and documentation of the graphs, and two "Help" tools: one for help with initial searching and one with help for interpreting the network diagram. The Sponsor Network Explorer app data table was also re-formatted with more consistent terminology and results that displayed sponsors and recipients of funding in separate tabs.

In the second round of in-person user testing, users were able to identify the purpose of the tool easily, but we found that users unfamiliar with biomedical publishing still needed help with understanding what to search for, how to read the diagram, and with definitions of terms. This was largely the result of users failing to look at the links to documentation regarding help with reading network diagrams. Instead, users stayed on the main screen of the tool and simply performed searches and attempted to interpret the results. Users did find it easier to understand the graphic representation and the table, in terms of representing research sponsorship. One user noted that, even after she scrolled up again and noted the "Help" tabs, that it would not have occurred to her to look at them during her initial use. In the process, we also identified some difficulties with the display of hovering help-text on small screens. From this round, we concluded that, while many of the users' questions were answered in the help documentation, we needed to make the documentation more prominent, visible, and clickable.

During the final round of user-testing, we tested the revised version of the tool with previous user testers and with new testers who had not been exposed to the tool before. Among previous users, the universal consensus was that, while some of the terminology—particularly for interpreting the network diagrams and their interfaces—was still unfamiliar and did not carry much meaning, their understanding of what the tool was for and how to use it was markedly improved. Most returning users found the help sections to be genuinely helpful, and to improve their experience.

Among new users, the experience was much more varied, and seemed to depend in large part on users' previous familiarity with biomedical publishing and the amount of time the user spent looking at the site. The users who had the most success with the tool were the two users who consulted the "Help" sections directly, evident both from their description of their interaction and their mirroring of the language in those sections. These users were able to comprehend the purpose and most of the results yielded through searching, though both commented that they did not fully understand what some of the layout options for the network visualizations represented, and one noted that the term "funding" should appear on the main screen of the app, since it appears in the "Search Help" button. The other user with the most success with this tool, who does not have an extensive background in biomedical publishing or research, suggested that, while the "Help" sections made it possible to understand the purpose and use of the tool, the process of working through all of them to understand the tool was a little time-consuming and that it would have been more helpful to have a link to a brief video walk-through of how to use the site, especially for users who were not familiar with the project.

Conclusions and Recommendations: From this final round of testing, we concluded that, while the help documentation is present and useful for those who consult it, users still seem to need some background and orientation to put this tool into context. The returning users found the helps to be helpful, but this is perhaps because they had already spent time with the tool and identified what confused them. New users tended to move quickly through the test using only the main interface, not clicking through to help, and relying on their own expertise and familiarity to guide them. Moving forward with the project, this is useful information in terms of disseminating both the app and this research. It is unlikely that most users will be consulting this tool with no context for what they hope to do: it will primarily be used by researchers interested in this topic. However, when presenting the tool and the research that develops from it, contextualization will still be necessary. Moving forward, it might be worth considering developing the kind of visual walk-through that one of our new users suggested.

Recommendations for Future Development

Ultimately, the successes of the T2V project suggest that it has established a useful foundation for future research in humanities network modeling. Specifically, the inter-rater reliability data indicate 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 in the nodes and edges table for subsequent visualization using one of many network visualization platforms. Despite the overall success of the T2V and related projects, there remain some significant needs in terms of continued development of new methods and toolkits that can support humanistic researchers who need to transform unstructured textual datasets into the kinds of data structures that support a broad range of network visualization projects. 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.

The recent proliferation of distant reading techniques in DH (see, for example, Underwood, 2019; Moretti, 2013; and Majdik 2019.) notwithstanding, few are optimized to produce the kinds of data necessary to create useful and effective network graphs. 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 (Brando, Frontini, Ganascia, 2016), 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 (Warren, et al., 2016). 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 (2015) "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 crowd-sourcing edge classification.

Future development building on the T2V framework has the potential to overcome many of these issues. In so doing, researchers should investigate the applicability of T2V protocols for a wider variety of humanities data forms. Conflicts of interest statements are unique to bioethics. Furthermore, they are essentially entirely relationship data. Humanities corpora for literature, history, philosophy, and other areas do not always describe relationships. Thus, an improved T2V framework will have to be successful in an environment where not all data is relational. It will also have to be adaptable to a wider variety of node and edge types. Future work in this area should evaluate the approaches herein described on other humanities data. A few possible new horizons of inquiry for this approach might include: 1) exploring intertextuality and/or citation-like attributions in texts that predate broadly accepted citation conventions, 2) identifying and classifying character relationships in written narratives, 3) investigating Burkean ratios in dramatic texts, or 4) locating and taxonomizing statements of moral obligation in ethical deliberation.

References

Barad, K. (2007). *Meeting the universe halfway: Quantum physics and the entanglement of matter and meaning*. Durham, NC: Duke University Press.

Bartko, J.J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports*, 19(1), 3-11.

Cain, D.M., Loewenstein, G., & Moore, D.A. (2005). The dirt on coming clean: possible effects of disclosing conflicts of interest. *Journal of Legal Studies*, 34, pp. 1-24; Loewenstein, G., Sah, S., & Cain, D. M. (2012). The unintended consequences of conflict of interest disclosure. *JAMA*, 307(7), 669-670.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological assessment*, *6*(4), 284.

Deleuze, G., & Guattari, F. (1988). *A thousand plateaus: Capitalism and schizophrenia*. Bloomsbury Publishing.

Fleiss, J. (1986). *The Design and Analysis of Clinical Experiments*. New York: Wiley.

Haraway, D. (1997). Modest. *Witness@ Second_Millenium. FemaleMan (cLMeets_Onco Mouse: Feminism and Technoscience*.

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

Keller, E. F. (1995). *Refiguring life: Metaphors of twentieth-century biology*. Columbia University Press.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, *15*(2), 155-163.

Latour, B. (1987). *Science in action: How to follow scientists and engineers through society*. Cambridge, MA: Harvard University Press.

Lundh, A., Lexchin, J., Mintzes, B., Schroll, J. B., & Bero, L. (2017). Industry sponsorship and research outcome. *Cochrane Database of Systematic Reviews*, (2).

Majdik, Z. P. (2019). A Computational Approach to Assessing Rhetorical Effectiveness: Agentic Framing of Climate Change in the Congressional Record, 1994–2016. *Technical Communication Quarterly*, 1-16.

Mapping the Republic of Letters. (2013). http://republicofletters.stanford.edu/.

Moretti, F. (2013). *Distant reading*. Verso Books.

Pattuelli, M. C. & Miller, M. (2015). Semantic network edges: A human-machine approach to represent typed relations in social networks. *Journal of Knowledge Management*, 19(1), 71-81.

Underwood, T. (2019). *Distant Horizons: Digital Evidence and Literary Change*. Chicago, IL: University of Chicago Press.

Warren, C. N., Shore, D., Otis, J., Wang, L., Finegold, M., & Shalizi, C. (2016). Six Degrees of Francis Bacon: A Statistical Method for Reconstructing Large Historical Social Networks. *DHQ: Digital Humanities Quarterly*, *10*(3).

Appendix: Usability Testing Protocol

Introduction / Instructions

Thank you for participating in our test, today. The test itself should take approximately ten minutes to complete. We are working with a prototype of a network mapping tool related to biomedical research and sponsorship/funding. We are asking you to use the prototype so we can learn how it works for future users. That said, we are testing how the *tool* functions, not *you*. Don't worry about making mistakes, and don't worry about hurting anyone's feelings. We want to improve the tool, so your feedback and your experiences are helpful to us.

Demo Questions

Thank you. Let's start with a few questions to give us some information about you as a user:

1. First, what is your occupation?
2. Roughly how much time would you say you spend interacting with websites, web forms, or searching online in a week, either for work or personal use? An estimate is fine.
3. What kind of experience do you have (if any) with doing research in libraries, databases, or any other kind of network?

First Responses to the Prototype

1. Thank you. Now, let's look at the tool, itself. Follow this link to the tool prototype: http://129.114.17.166/network_explore/
2. Take a look at the page. You can scroll or click on drop-down menus, but don't type anything or make any selections, yet. As you look at this page, what do you make of it? Who do you think it belongs to? What purpose does it serve? What can you do here?
3. What information do you think you might you need in order to use this tool?

Key Tasks

Thank you. Now, we're going to ask you to perform a few different tasks with the tool. To perform these tasks you can scroll, select, type, and search on the site, or click anything that is clickable. We're not going to give you much information beyond the task description, because we will learn more about how the tool works that way. After you've performed each task, please answer the questions that follow, giving us as much feedback about your experience and thought process as you can.

Task One: Plot the network for a medical condition at the ARTICLE level. Use the default layout and filter settings.

> What do you make of the results? Can you identify the total number of available citations, conflicts of interest, journals involved, and authors involved in this network? How would you describe the relationship between those numbers, based on looking at the search results?

Task Two: Plot the network for a medical condition at the AUTHOR level. Once you have your results, search within the results for either an industry name or an author name.

> Can you determine the frequency of this name in the network? How would you describe what this frequency number means?

Task Three: Plot the network for any medical condition at any granularity level. Once you have the results, change the "Filter Nodes by Out-Degree" setting.

How does this change your results? What does this change tell you?

Task Four: Plot the network for any medical condition at any granularity level. Once you have the results, change the layout.

How does this change your results? What does this change tell you?

Final Thoughts and Questions

Do you have any final feedback or questions about the tool, now that you are finished with the test?