

# Breaking Barriers in Research Projects: BeagleTM, a Powerful Python-based Text Mining Tool for Visual Discovery in Scientific Literature

Oliver Bonham-Carter, Ph.D.  
Dept of Computer and Information Science, Meadville, PA



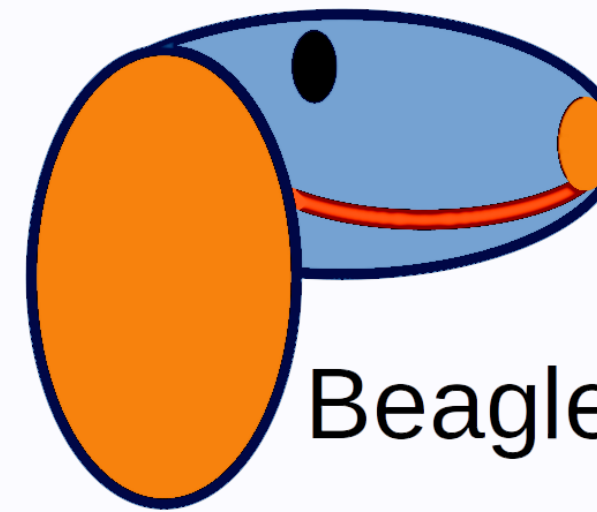
ALLEGHENY COLLEGE

<https://www.cis.allegheny.edu>  
obonhamcarter@allegheny.edu  
Presented at PyCon 2024

## PROJECT OBJECTIVES

### This project presents

- ▶ A study by text analysis of ethically inclined language in the literature of Bioinformatics research, and in related disciplines
- ▶ Results are amassed by use of the BeagleTM text analysis project
- ▶ We create networks to visualize relationships between articles of the literature that share common threads of language relaying ethical themes



BeagleTM

Figure: 1. BeagleTM is a robust, supervised, text analysis software written in Python that uses *bag of words approach*.

## METHODS

### Corpus

- ▶ Data: a corpus created from the non-commercial publication archives of National Center for Biotechnology Information (NCBI)  
<https://www.ncbi.nlm.nih.gov/>
- ▶ Over 19 GB of textual data: articles originating from over 2000 respected publishers (i.e., Science, Nature, Elsevier, IEEE, ACM and others) of diverse subjects in science

### Keywords

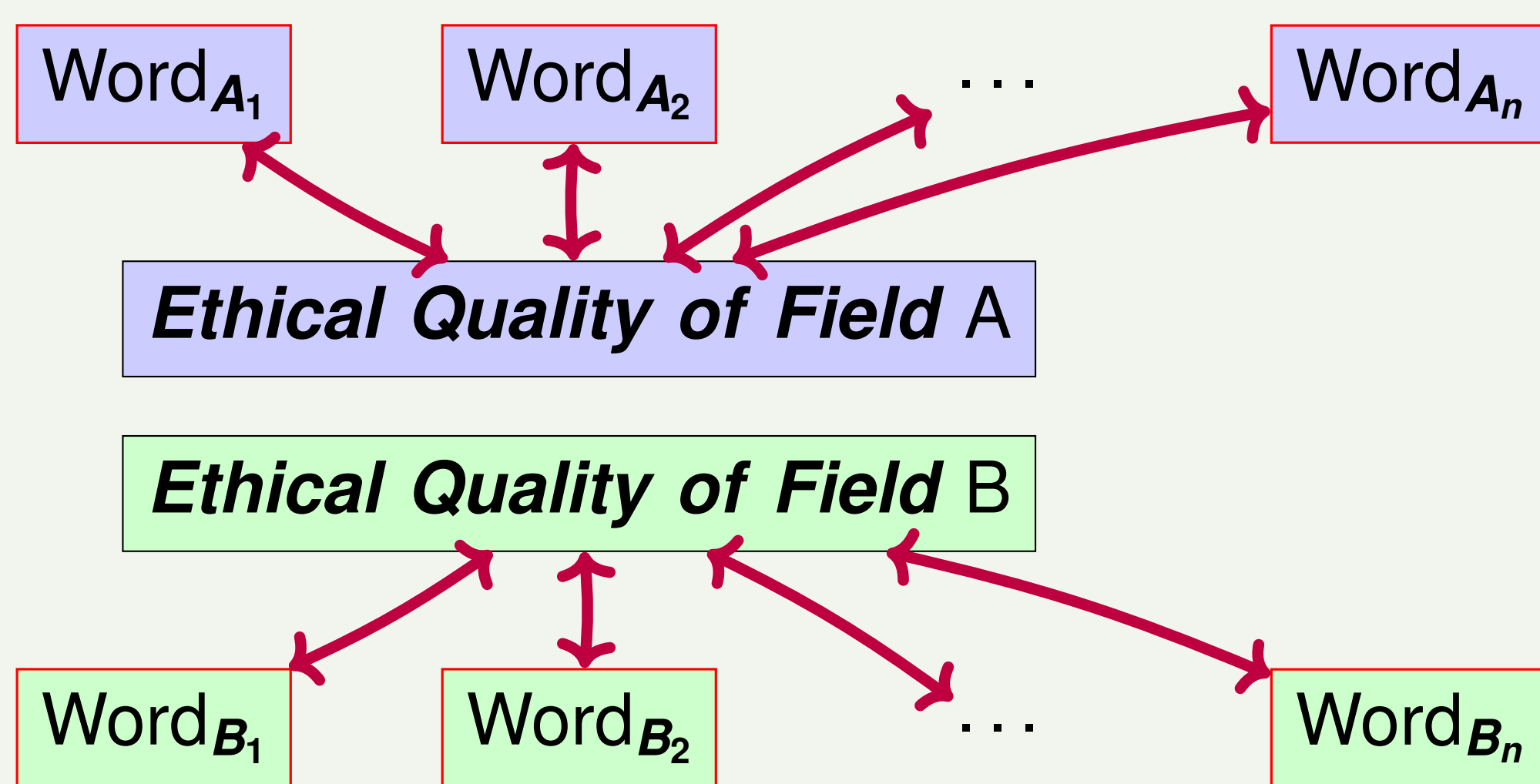


Figure: 2. **Determining relevant keywords to extract discussion of ethics in research.** Keywords are selected from two disciplines; A and B. We note that each discipline has own lexicon of words relating to ethical details. To capture the nuances of each area, our curated list of keywords concerns a host of simple definitions of ethical issues in articles.

- ▶ The keywords were chosen to be *generic* indicators of conversational ethics in scientific articles
- ▶ **Disciplines;** *Biology, Bioinformatics, Computer Science, and Informatics*, and **Terms** relating to ethical conduct in research (i.e., *ethics, liability, responsibility*, and similar)

## Parsing for Keywords in Literature to Create Networks

```
14965549, corpus/plosBio1/PMC348949.xml
+ Parsed: ['analysis', 'analytical', 'stem'] <- [2, 1, 6]
182 of 194
15824421, corpus/plosBio1/PMC368167.xml
+ Parsed: ['analysis', 'biology', 'computer science', 'informatics', 'stem'] <- [3, 1, 1, 2, 2]
184 of 194
14966539, corpus/plosBio1/PMC344483.xml
+ Parsed: ['analysis', 'analytical', 'liability', 'responsible', 'stem'] <- [7, 2, 2, 1, 16]
185 of 194
14737188, corpus/plosBio1/PMC348944.xml
+ Parsed: ['analysis', 'biology', 'stem'] <- [4, 3, 11]
186 of 194
15824426, corpus/plosBio1/PMC368171.xml
+ Parsed: ['analysis', 'stem'] <- [9, 6]
187 of 194
14737182, corpus/plosBio1/PMC368881.xml
+ Parsed: ['analysis', 'biology', 'responsible', 'stem'] <- [7, 1, 1, 3]
188 of 194
14551915, corpus/plosBio1/PMC212697.xml
+ Parsed: ['biology', 'computer science', 'informatics', 'responsible', 'stem', 'trust'] <- [9, 2, 1, 2, 16, 1]
189 of 194
14966535, corpus/plosBio1/PMC348944.xml
+ Parsed: ['analysis', 'biology', 'stem'] <- [4, 3, 11]
191 of 194
14965542, corpus/plosBio1/PMC348952.xml
+ Parsed: ['analysis', 'stem'] <- [1, 3]
193 of 194
12975657, corpus/plosBio1/PMC193685.xml
+ Parsed: ['analysis', 'stem'] <- [2, 1]
194 of 194

File saved to: 0_out/kw_genPurposeEthics_analysis_out_save-less.csv
-> Opening csv files: 0_out/kw_genPurposeEthics_analysis_out_save-less.sqlite3
-> Finished building sqlite3 database
DB File saved to: 0_out/kw_genPurposeEthics_analysis_out_save-less.sqlite3
-> Opening csv file: 0_out/kw_genPurposeEthics_analysis_out_save-less.sqlite3
-> Finished building sqlite3 database
DB File saved to: 0_out/kw_genPurposeEthics_analysis_out_save-less.sqlite3

Completed creation of SQL database from csv file
0_out/kw_genPurposeEthics_analysis_out_save-less.sqlite3
```



Figure: 3 Left; we used BeagleTM to parse all articles in the corpus created from PubMed archives for specific keywords. Articles in which keywords were found were used to build Relationship Networks to visualize connections of related ideas. Right; BeagleTM's network creator and browser.

- ▶ Python resources: Poetry (Parsing), Streamlit (Network Creation and Browsing), NetworkX, PyVis and Plotly (Network creation) and others

## METHODS CONTINUED

### Relationship Networks

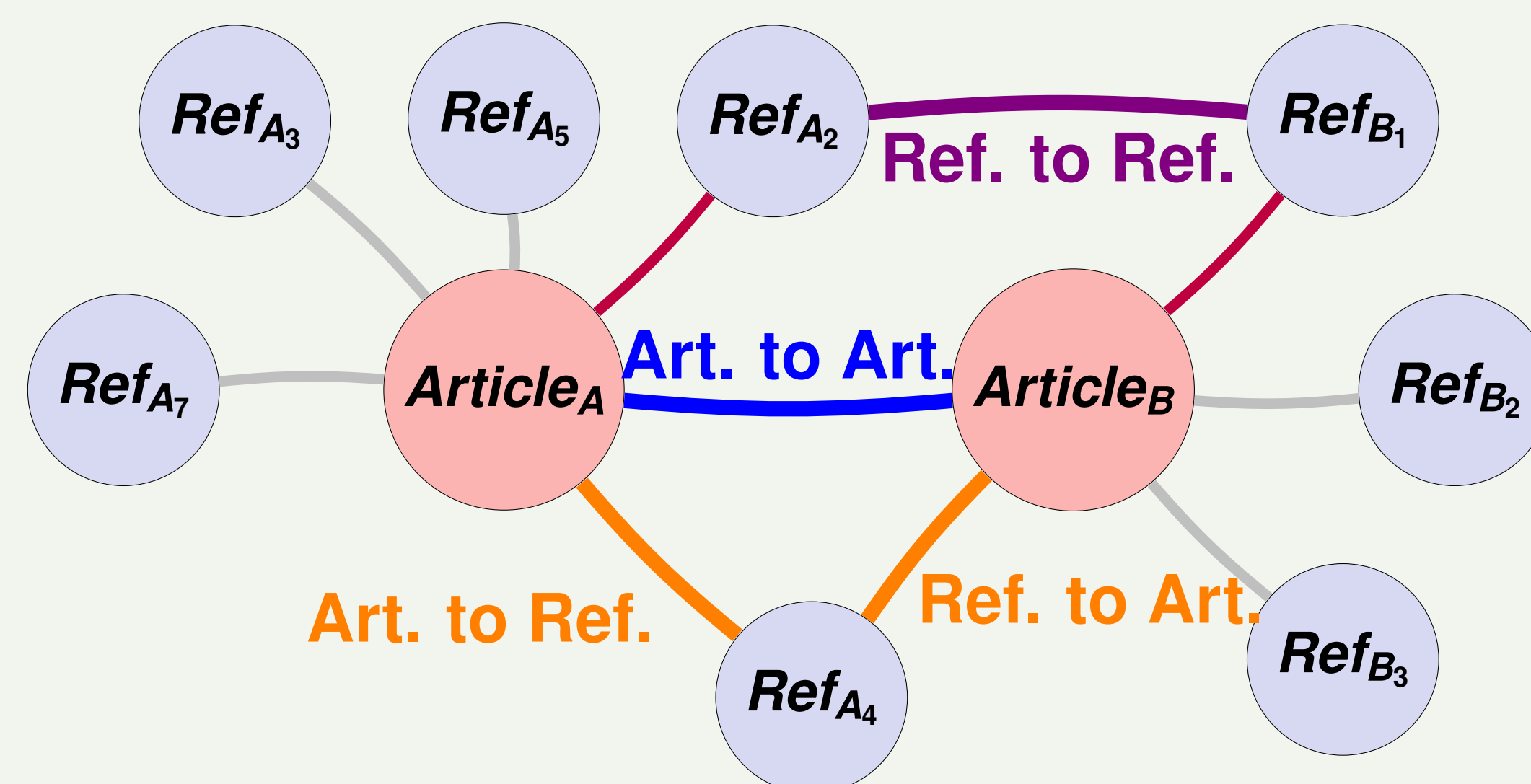


Figure: 4. Relationship networks. The larger (red) nodes represent the articles in which keyword occurrences were found. The smaller (blue) nodes represent the supporting documents comprising the article's bibliography. The edges connecting nodes imply that one publication is citing the other.

There are three types of connections to observe in these Relationship Networks

- ▶ **Type 1:** *Reference to Reference*
- ▶ **Type 3:** *Article to Article (Strongest)*
- ▶ **Type 2:** *Article to Reference to Article*

## RESULTS

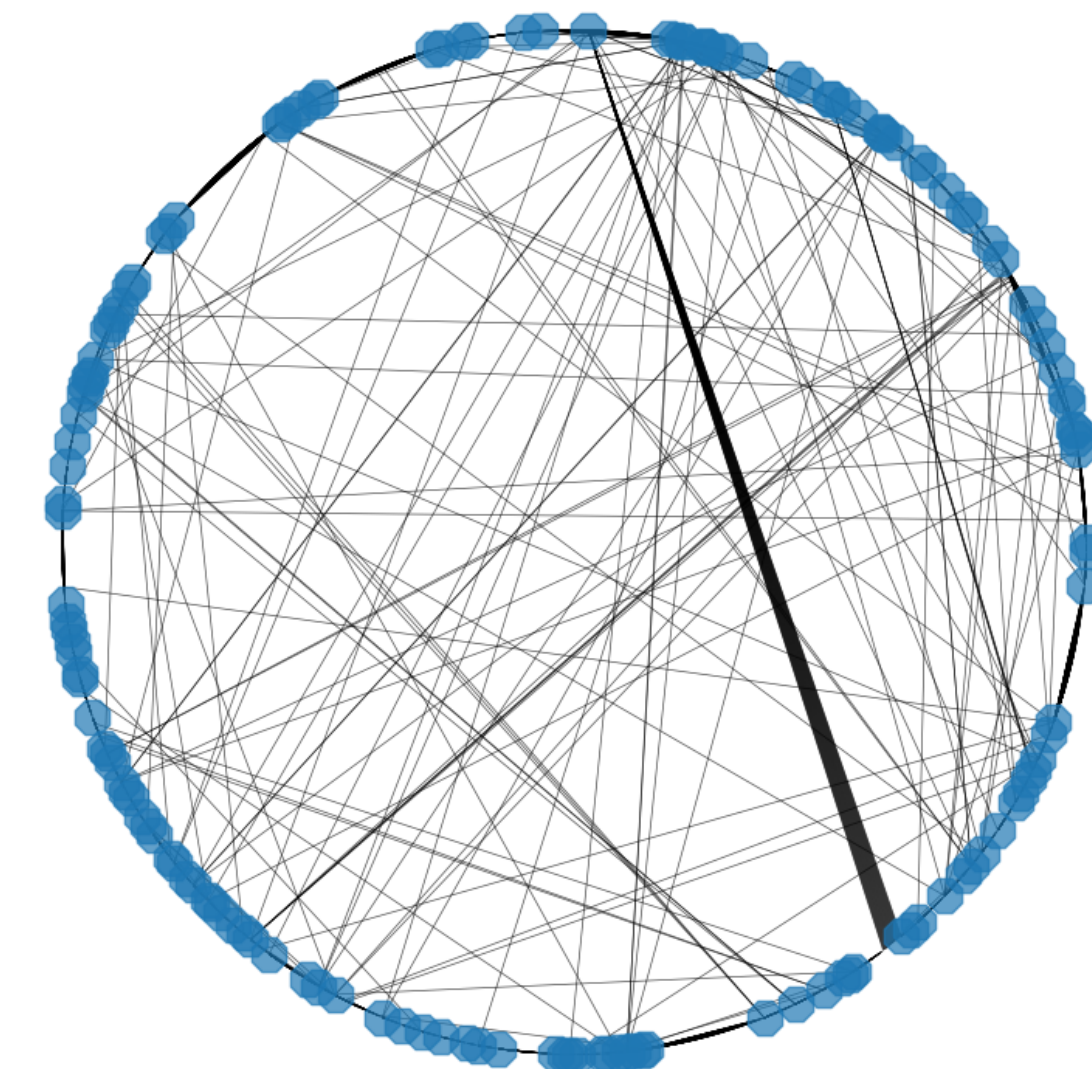
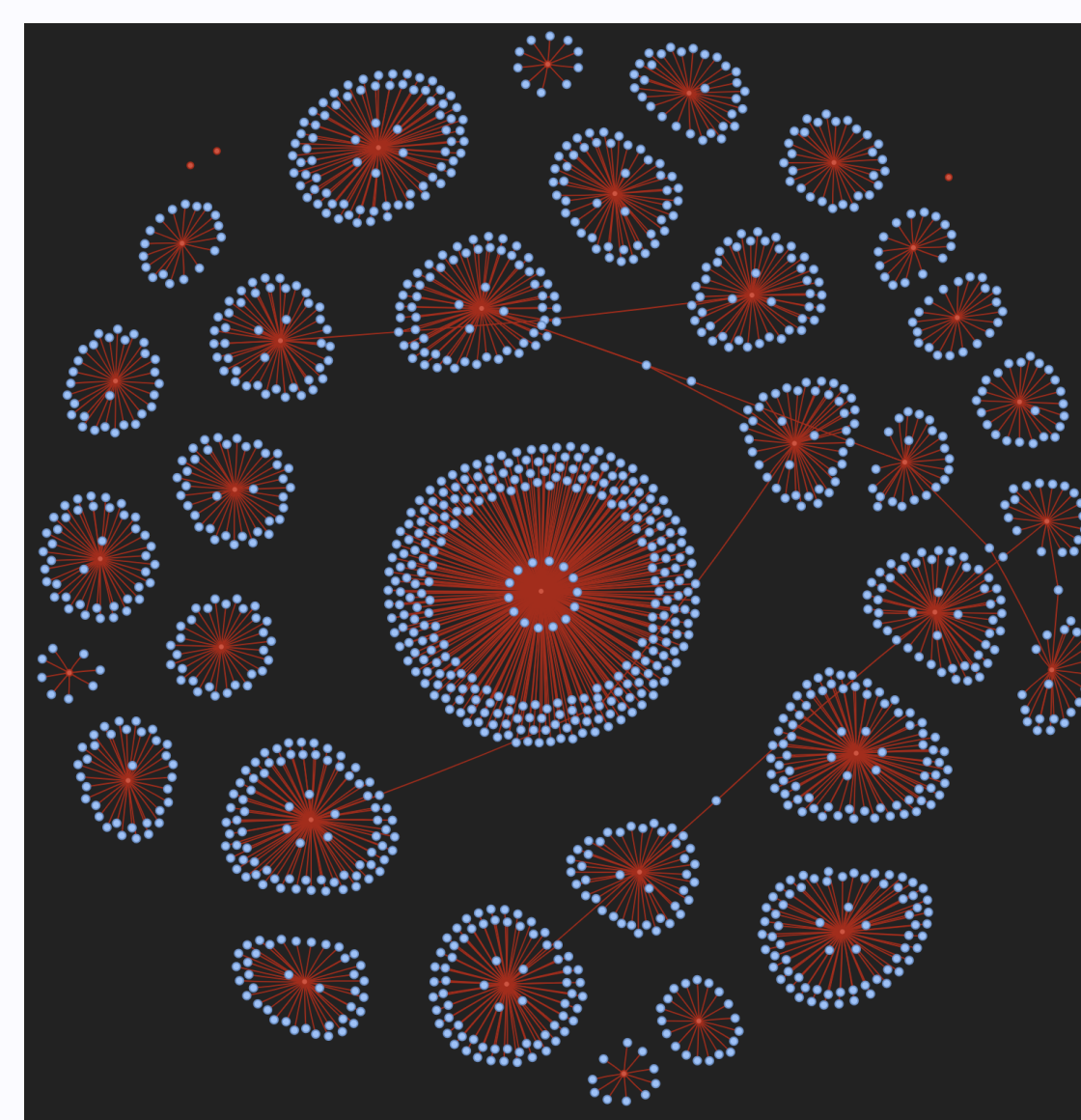


Figure: 5. **Left: A Relationship Network:** Groups represent articles (red nodes) sharing edges with bibliography references (blue nodes) according to common language stemming from contextual keywords. **Right: A general view of connectivity:** Nodes indicate articles, and the edges indicate common keywords. Here, all nodes must have the *same set of keywords* to be included the network, likely indicating common ideas.

## CONCLUSIONS

- ▶ By studying the keyword content in a discipline's articles, one may study the spread of word usage to infer a spread of ideas
- ▶ Articles (generally) that reference others having ethical inclinations, appear to also contain similar ethical language

## REFERENCES

- ▶ BeagleTM2: <https://github.com/developmentAC/beagleTM2>
- ▶ Bonham-Carter, Oliver. "Text Analysis of Ethical Influence in Bioinformatics and Its Related Disciplines." Future of Information and Communication Conference. Cham: Springer Nature Switzerland, 2024.