FLORIDA TECH

FLORIDA'S STEM UNIVERSITY

Explorations of Connections in Biomedical Research through Knowledge Graphs

CSE 4510 Big Data: Group 4

Project Outline

- Introduction
- Literature Review
- Methods / Proposed Approach
- Data Description
- Data Cleaning and Management
- Implementation
- Results
- Conclusions and Future Work

Introduction

Objective: To explore ways in which we can use knowledge graphs via Neo4J to display the biomedical concepts found within the Semantic Medline Data base and their relationship

SemMedDB : Database of automatically gathered medical literature information

Neo4J: Graphical database software to visualize a database and relationships within

Research Questions

1. How can we employ the use of knowledge and relational graphs to explore connections among concepts, treatments, and diseases discussed in various published biomedical literature?

2. How can the creation of these knowledge graphs be used to optimize the retrieval of information in biomedical research and the communication of complex biomedical data to the general public?

Literature Review

Unified Medical Language System
A compilation of databases and files to enhance interpretability of biomedical concepts

- UMLS Metathesaurus
- Semantic Network
- SPECIALIST Lexicon

Biomedical Texts Abstracts sourced from Pub Med

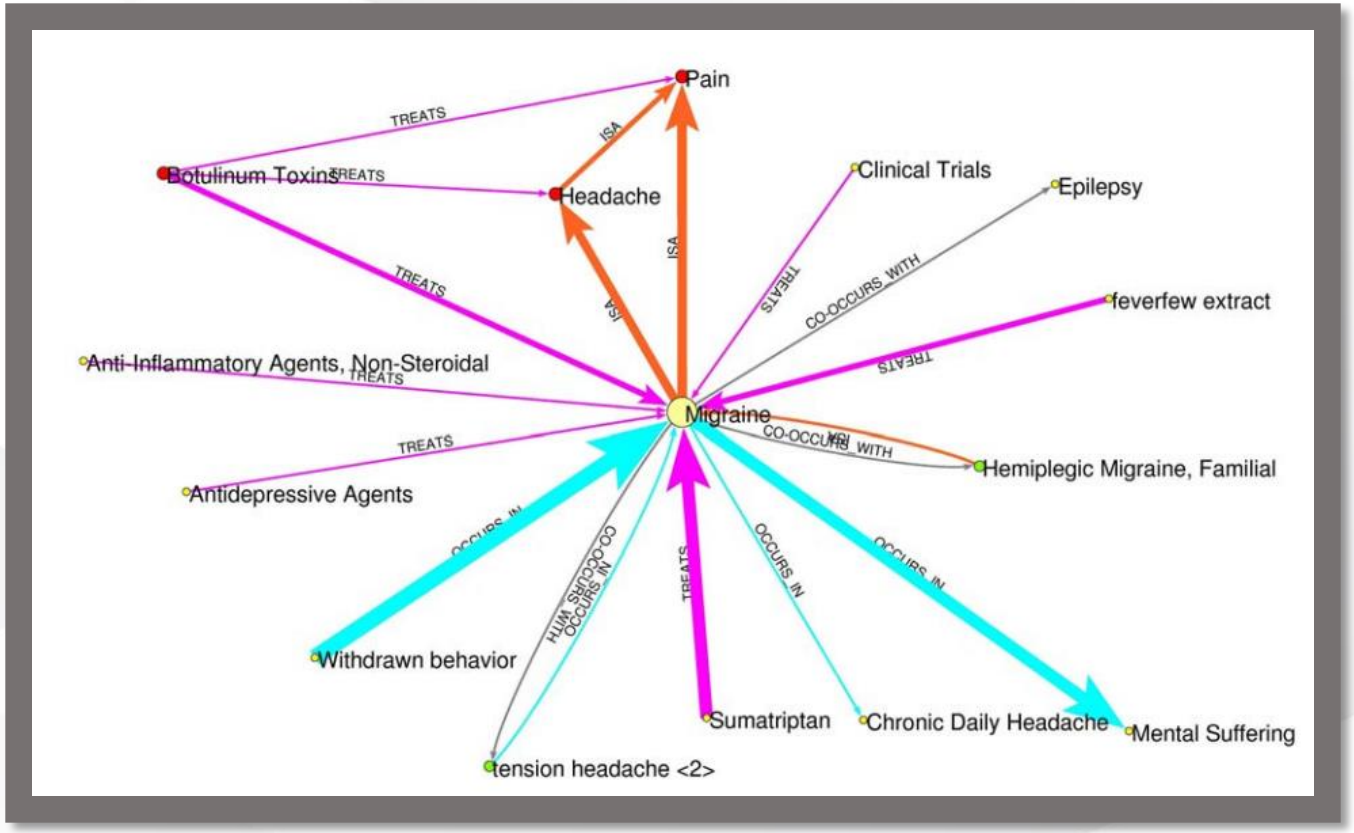
Sem Rep: NLP to mine semantic relationships in text

Sem Rep produces an SPO triple
{ Subject Predicate Object}

Literature Review

Previous Work : To transform the output produced from SemRep when reading in an abstract from a research article into a knowledge graph (M. Fiszman, T. Rindflesch, H. Kilicoglu 2004).

- Each node represents a biomedical entity, and the arrows represent the predication connecting the entities.

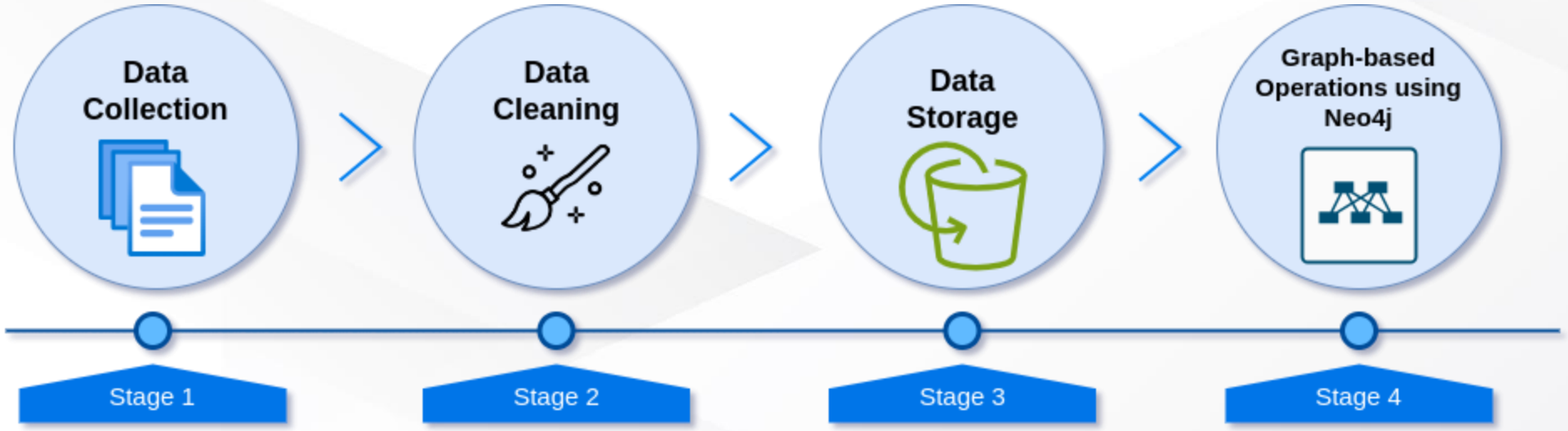


FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Methodology & Data

FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Procedural Workflow



FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Data Description

Sentence Table

- Contains information about sentences from literature.
- Size: 16 GB, 253,029,872 sentences

Entity Table

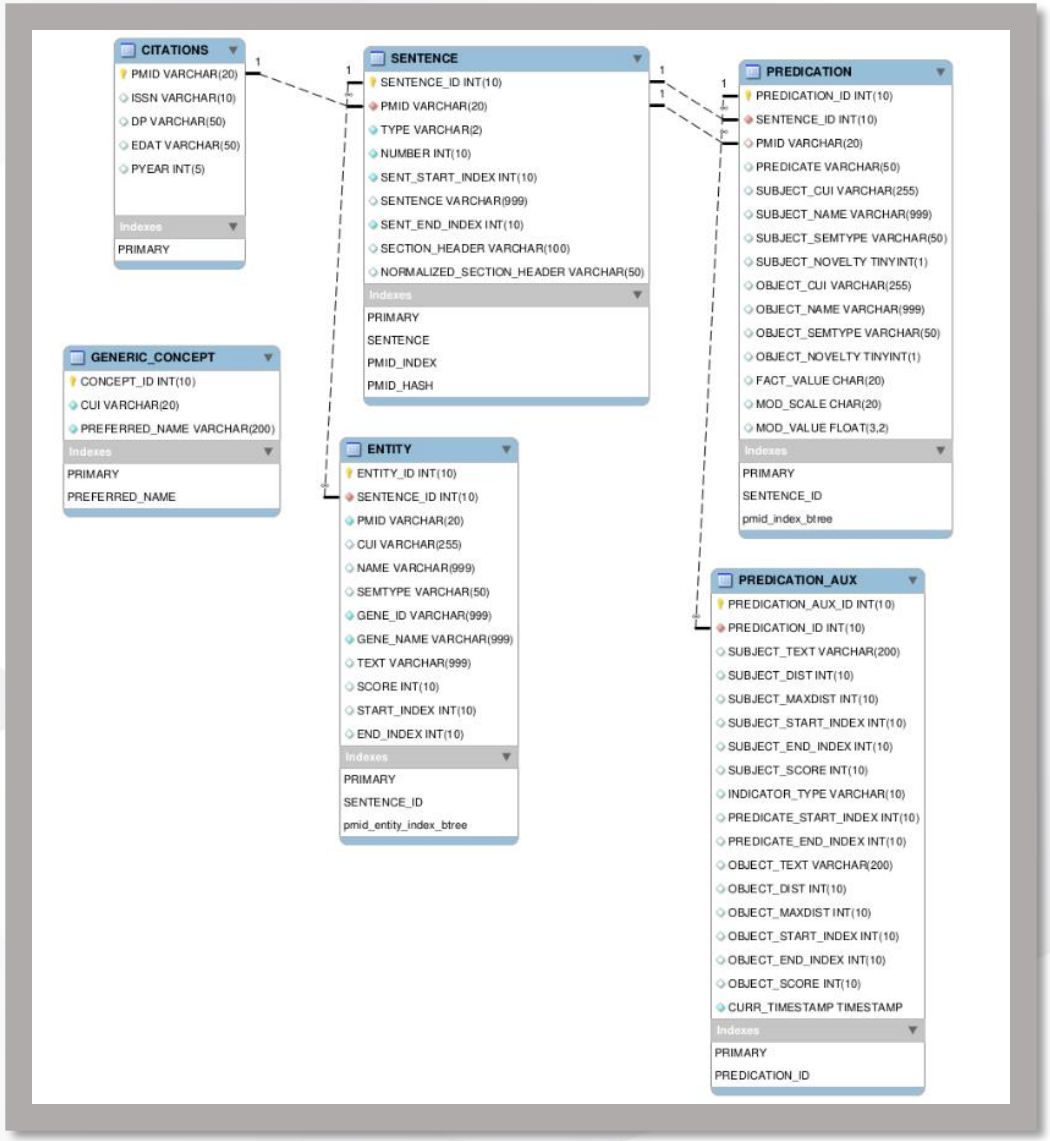
- Contains details about specific entities/nouns in sentences.
- Size:44 GB, 1,887,317,669 rows

Predication Table

- Represents verbs applied to entity nouns.
- Size:3 GB, 126,268,045 rows

Data Description

- The three data tables (Sentence, Entity and Predication) utilized in the projects are all inner joined together by their shared column “Sentence ID”.



FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Data Size and Reduction

Reduction methods were applied due to storage constraints.

Reasons for Reduction:

- Local project limitations without server storage.
- Ensuring data manageable for analysis.

Impact of Reduction:

- Maintain project feasibility.
- Focus on essential information.
- Efficient use of resources.

Data Cleaning

- Filled empty string and integer fields with "Not Available" and "-1", respectively.
- Corrections made for database version and schema mismatch.
- Identified and dropped columns with limited or unknown data (e.g., "SENT END INDEX," "NORMALIZED SECTION HEADER").
- Discovered sentence data in the "SENT END INDEX" column.
- Removed unnecessary columns to enhance data utility.

Implementation & Results

FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Relationship Building in Neo4J

Utilized the community version of Neo4J.

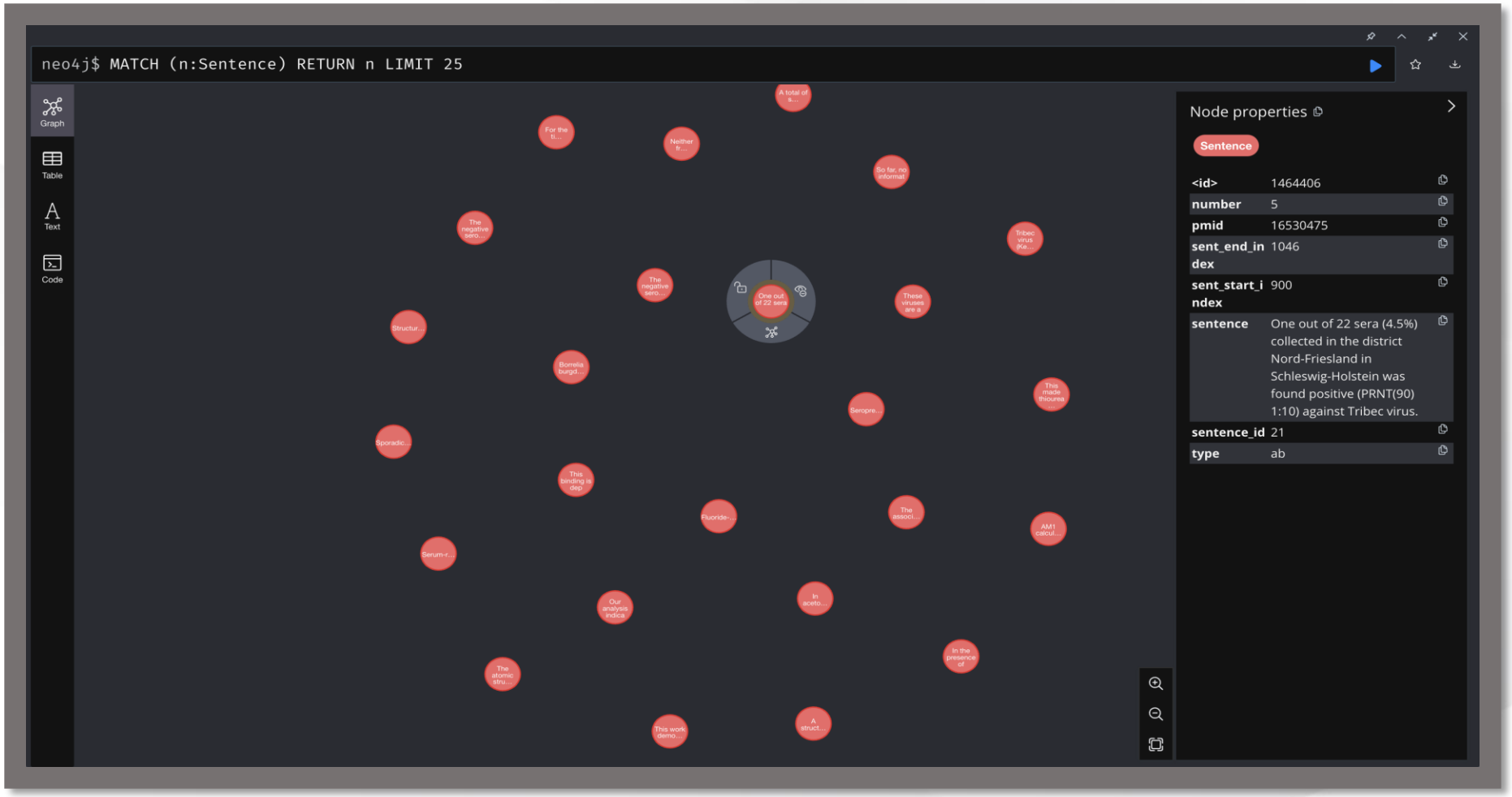
Main relations established via sentence foreign keys.

Predication and entity tables linked to the sentence table.

Neo4J relationships created: "SUBJECT OF" (entity to sentences) and "PREDICATES" (predication to entity).

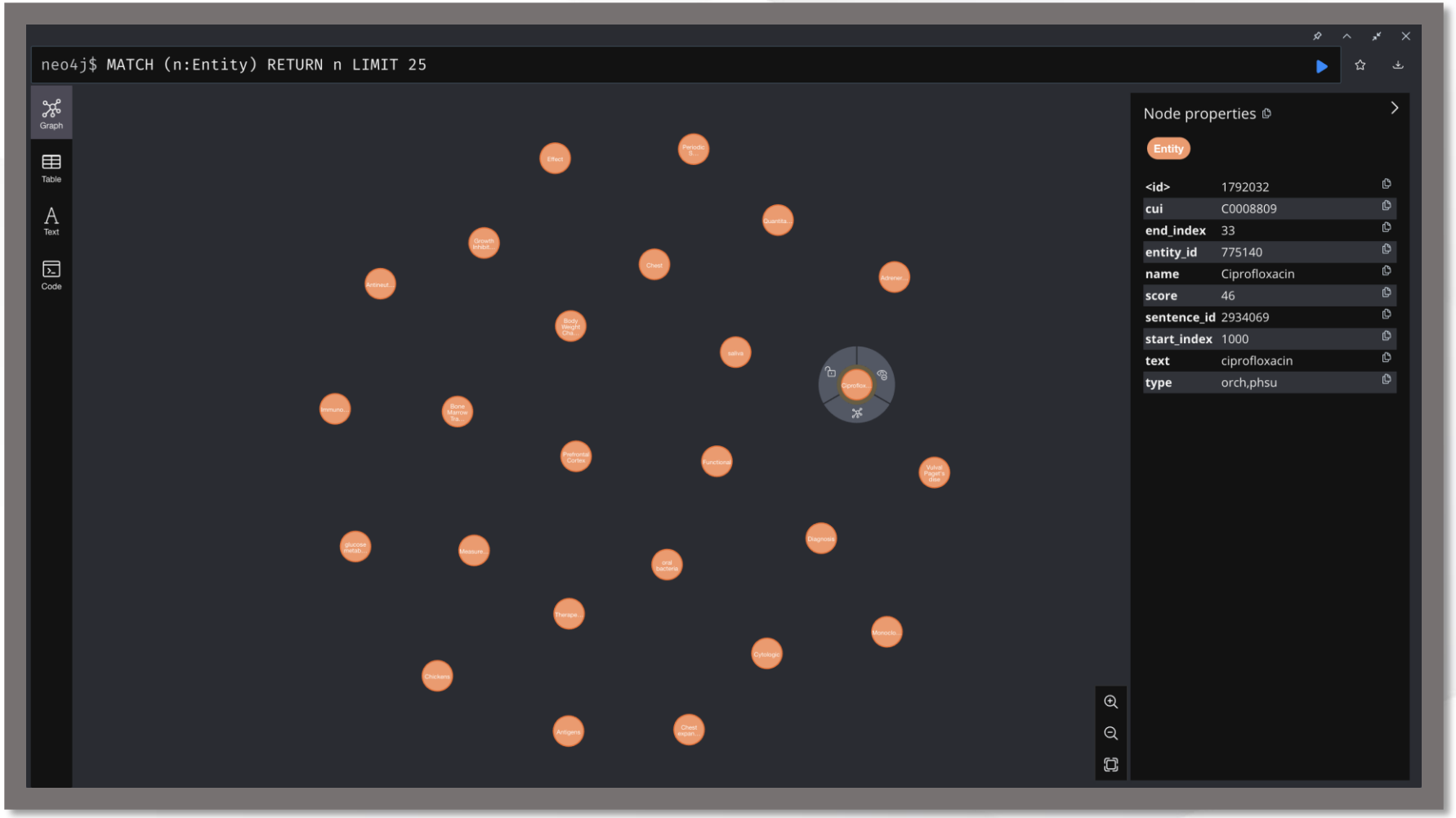
Three one-way types of relationships formed in the overall database.

Sentence Knowledge Graph



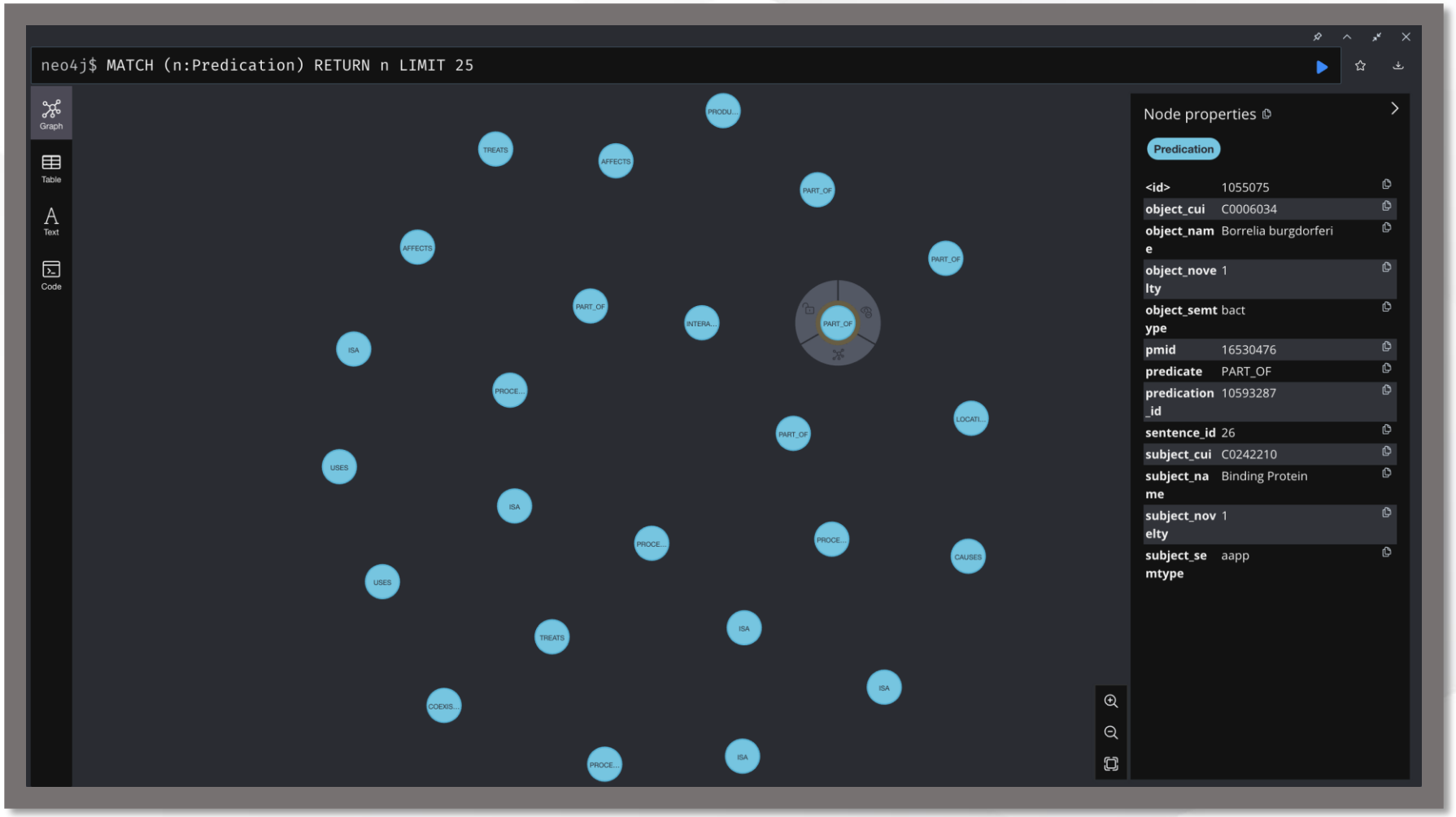
FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Entity Knowledge Graph



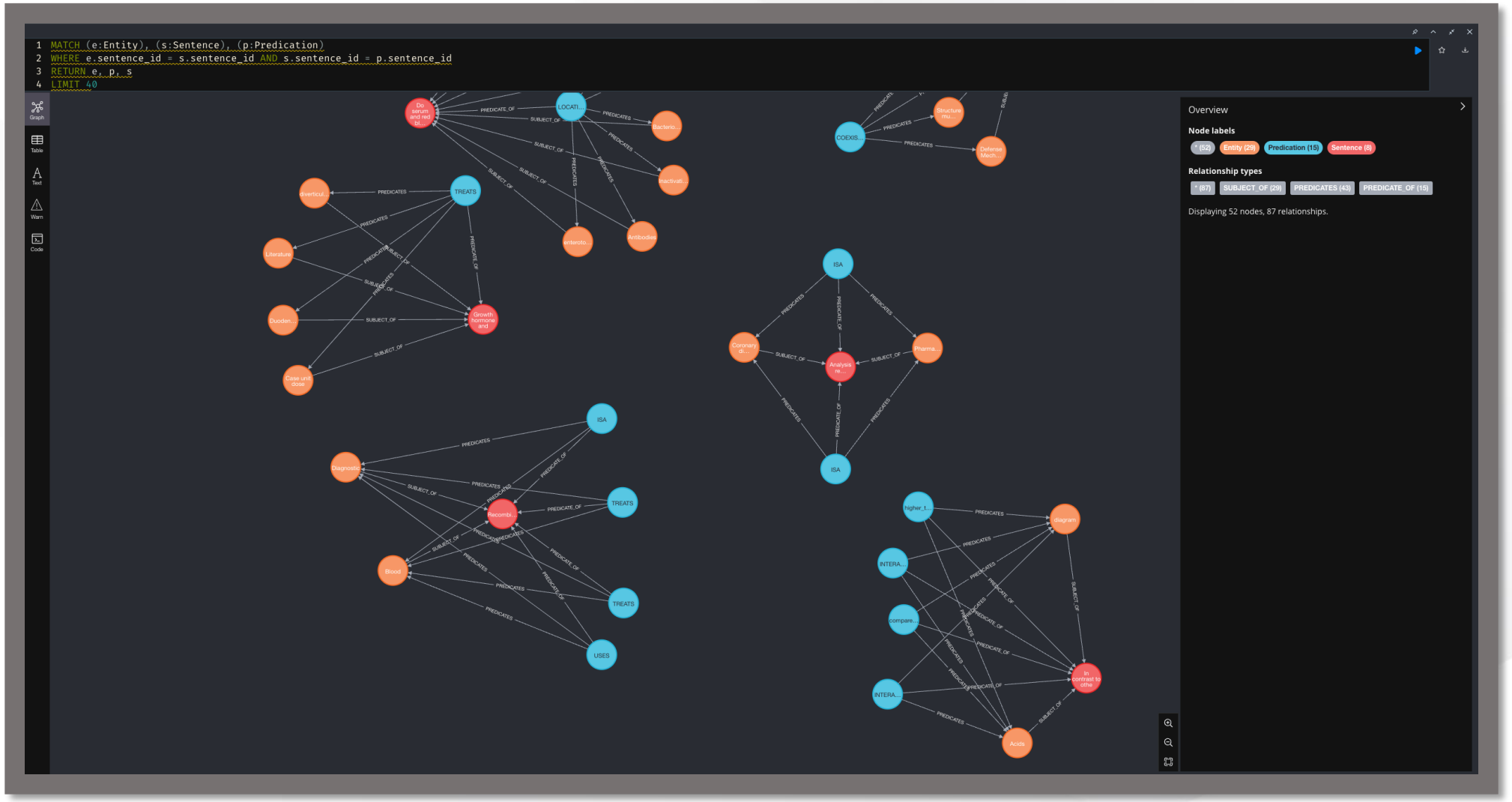
FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Predication Knowledge Graph



FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

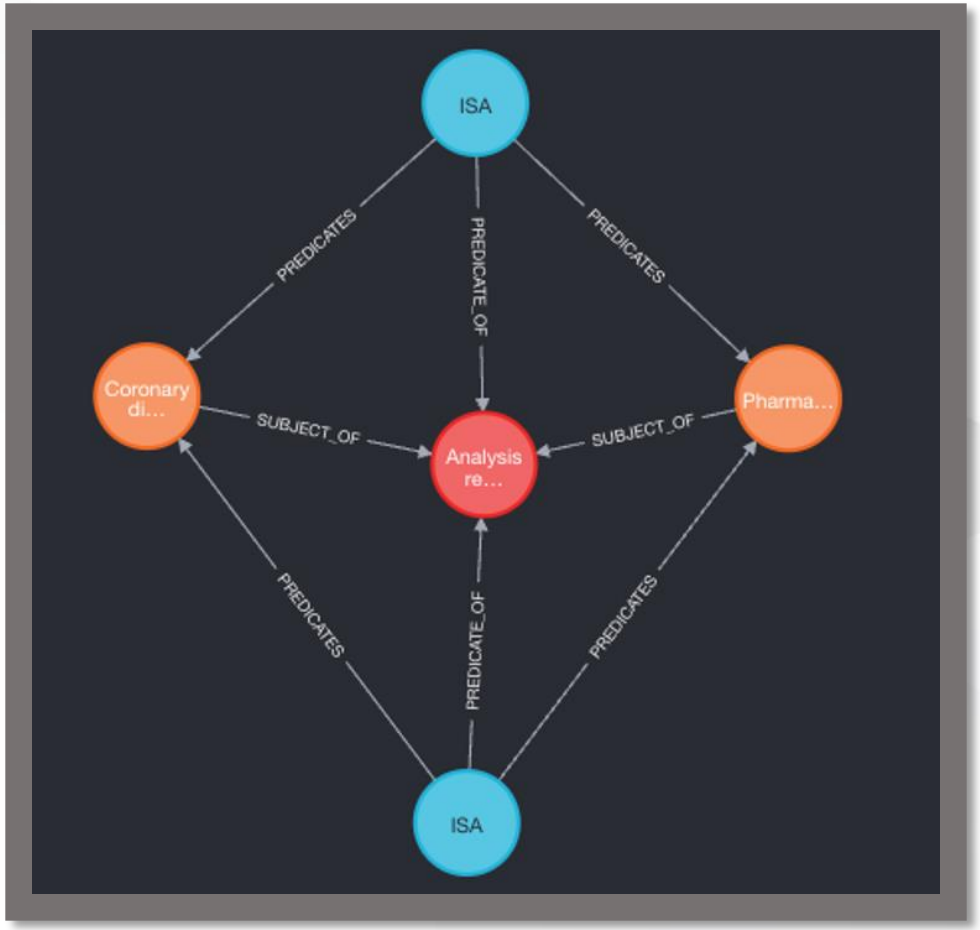
Fully Connected Knowledge Graph

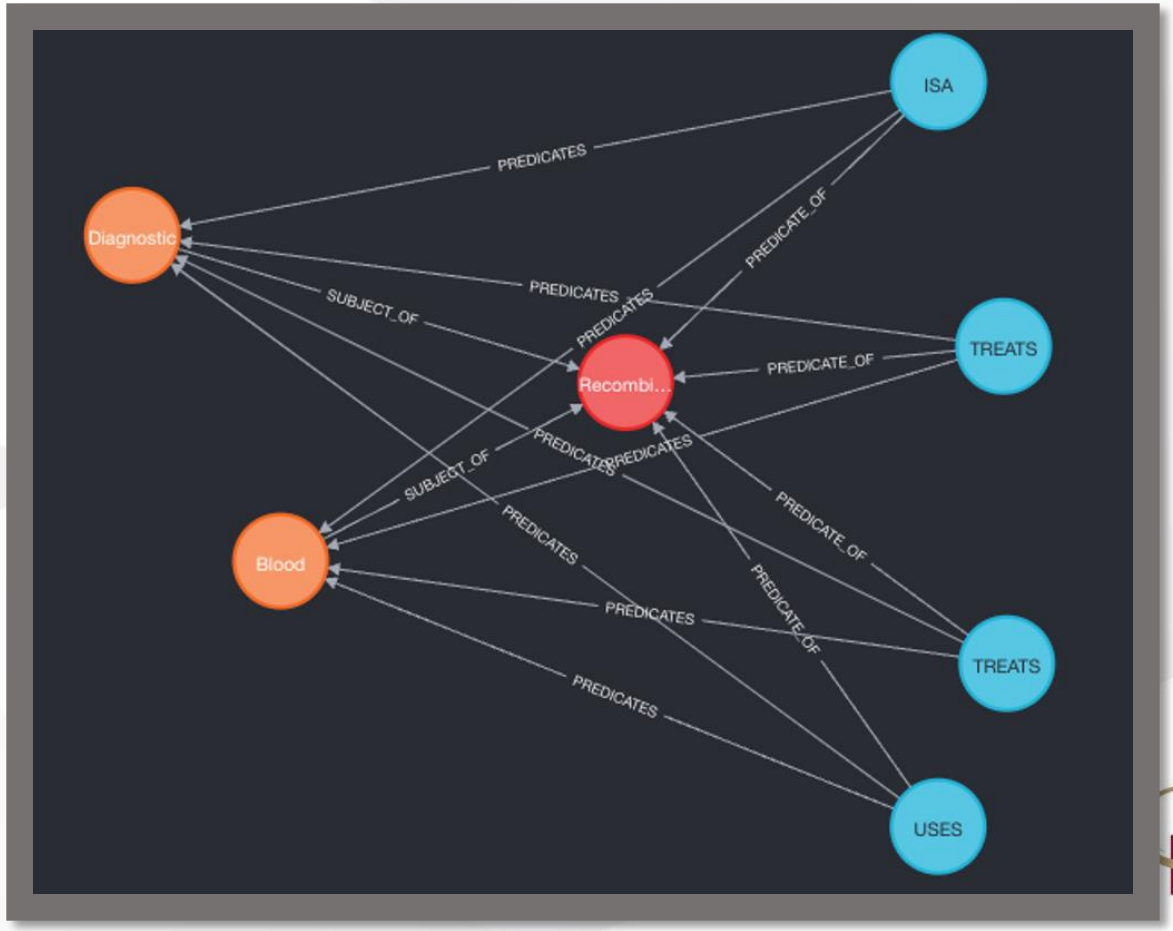


FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Fully Connected Knowledge Graph

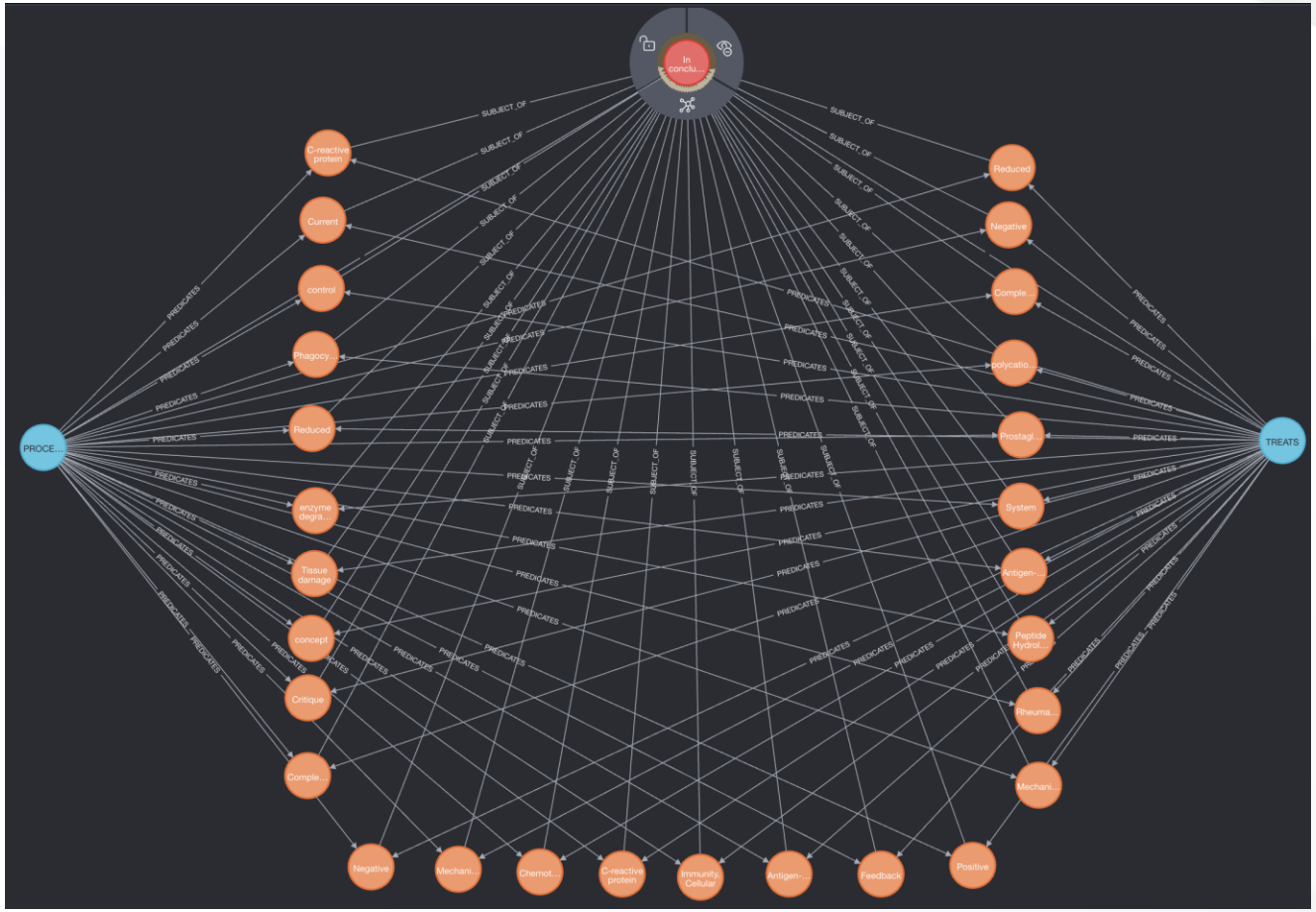
Below are the fully connected graphs with the predicates connecting various biomedical subjects and objects discussed in literature.





FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Example of a Query on the SemMed Graphical Database in Neo4j



Shows the shared connection between nodes with a subject semantic type of "hlca" or "menp".

FLORIDA TECH
FLORIDA'S STEM UNIVERSITY

Discussion

RQ1:

Utilizing SemRepDB, we build detailed knowledge graphs from structured biomedical literature. The user-friendly query system is vital, allowing swift identification of specific biomedical relationships or trends.

RQ2:

Hosting SemMed data on Neo4j a graph database-based platform enhances relationship navigation, enabling efficient querying of intricate patterns. This presents diverse industries with valuable insights, simplifying the analysis of connections across numerous articles.

Conclusion

- Had to upload smaller dataset
- Graph software provides powerful visualization tool for analysis and determining connections
- Future Work:
  - Upload full database to server
  - Investigate using indexes to speed up computation
  - Web applications, API, etc

References

- M Fiszman, T. Rindflesch, H Kilicoglu. "Abstraction Summarization for Managing the Biomedical Research Literature". Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics (2004).
- H Kilicoglu et al. "Constructing a semantic predication gold standard from the biomedical literature". BMC Bioinformatics (2011).
- A Burgun, O Bodenreider. "Comparing terms, concepts, and semantic classes in Word-Net and the Unified Medical Language System". Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations (2001), pp. 77–82.
- A. McCray. "Representing biomedical knowledge in the UMLS Semantic Network".High-Performance Medical Libraries: Advances in Information Management for the Virtual Era (1993), pp. 45–55.
- H. Kilicoglu, G. Rosemblat, M Fiszman. "Broad-coverage biomedical relation extraction with SemRep". BMC Bioinformatics (2020).
- Y Liu et al. "Using SemRep to label semantic relations extracted from clinical text".AMIA Annu Symp Proc (2012), pp. 587–595.
- R. Graciela et al. "A methodology for extending domain coverage in SemRep". Journal of Biomedical Informatics Vol. 46 (2013), pp. 1099–1107.
- Aronson AR. 2001 Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. AMIA Annu Symp Proc, pp 17-21.
- National Institutes of Health. (n.d.). Unified Medical Language System (UMLS). U.S. National Library of Medicine. https://www.nlm.nih.gov/research/umls/index.html

Questions?

FLORIDA TECH

FLORIDA'S STEM UNIVERSITY