EPFL

Blue Brain

BlueGraph

Co-occurence Analysis

Community Detection

Centralities

Representation Learning on Graphs

Shortest Paths Search

Unifying Python framework for graph analytics and co-occurrence analysis

Blue Graph is a Python framework for graph analytics and co-occurrence analysis consolidating capabilities from different graph processing backends under a unified API. Using Blue Graph, users can gain insights, uncover hidden patterns, infer implicit knowledge and build recommendations engines from graph-shaped data.

Graphs have become ubiquitous in many scientific domains, such as sociology or molecular biology. They provide a close-to-reality data representation that focuses on entities and their relationships. Graph analytics leverages this representation to provide insights on data, uncover hidden patterns and infer implicit knowledge. Consequently, graph analytics underlies many modern data science tasks, such as exploring and analysing Knowledge Graphs built from natural text enabling Knowledge Graph guided literature review.

There exists an ever-growing number of tools for graph analytics. While these provide a wide range of capabilities, they do not share common interfaces or data structures. Therefore, combining them into a single analytics pipeline is a lengthy task, which requires deep tool-specific knowledge.

To tackle this challenge, the Blue Brain Project built Blue Graph, a Python framework that unifies graph analytics capabilities from different graph processing backends (NetworkX, graph-tool, Neo4j, StellarGraph) and provides a unifying interface for common graph analytics tasks, such as community detection, path search and graph representation learning.

Key features of **Blue Graph**

Blue Graph enables users to perform various graph analytics tasks including:

- Co-occurrence analysis
- Computation of node centrality metrics (degree, PageRank, centrality)
- Community detection
- Path search
- Exploration and navigation of dense graphs through minimum spanning trees
- Node classification, node similarity search and link prediction through graph representation learning.

A set of unified interfaces that facilitate the process of building complex multi-backend graph analytics pipelines.

These interfaces include:

- Interfaces for computing graph metrics (such as degree, centrality, PageRank), performing path search and community detection through a graph analytics API
- Semantic property encoders and co-occurrence graph generators provided by pre-processing and co-occurrence analysis API
- Support for computing various graph embedding techniques through a graph representation learning API
- Perform graph node classification, similarity queries, link prediction through representation learning downstream tasks API.

Using Blue Graph for a Knowledge Graph guided review of the CORD-19 dataset

In response to the COVID-19 pandemic, the COVID-19 Open Research Dataset (CORD-19), the most extensive coronavirus literature collection was made open access and therefore, available for data mining.

Blue Graph was used as a part of Blue Brain's effort to perform a Knowledge Graph guided review of this vast COVID-19 related literature dataset to uncover the reported role of blood glucose levels in the severity of COVID-19. This project required BBP scientists to quickly build, curate, explore and analyze dense Knowledge Graphs obtained by applying Natural Language Processing techniques to the unstructured scientific texts of over 200'000 CORD-19 articles. The resulting graph was very dense and consisted of ~44 million edges representing entity co-occurrences.

Using the following **Blue Graph** capabilities users were able to reduce the knowledge graph complexity and make sense of the overwhelming amount of information the dataset contains.

1. **Co-occurrence analysis** allowed users to construct Knowledge Graphs whose nodes represented entities mentioned in the literature corpus and links represented entity co-occurrences quantified using mutual-information-based associations' scores

2. **Importance measures** guided the literature reviewer towards specific 'hub' entities

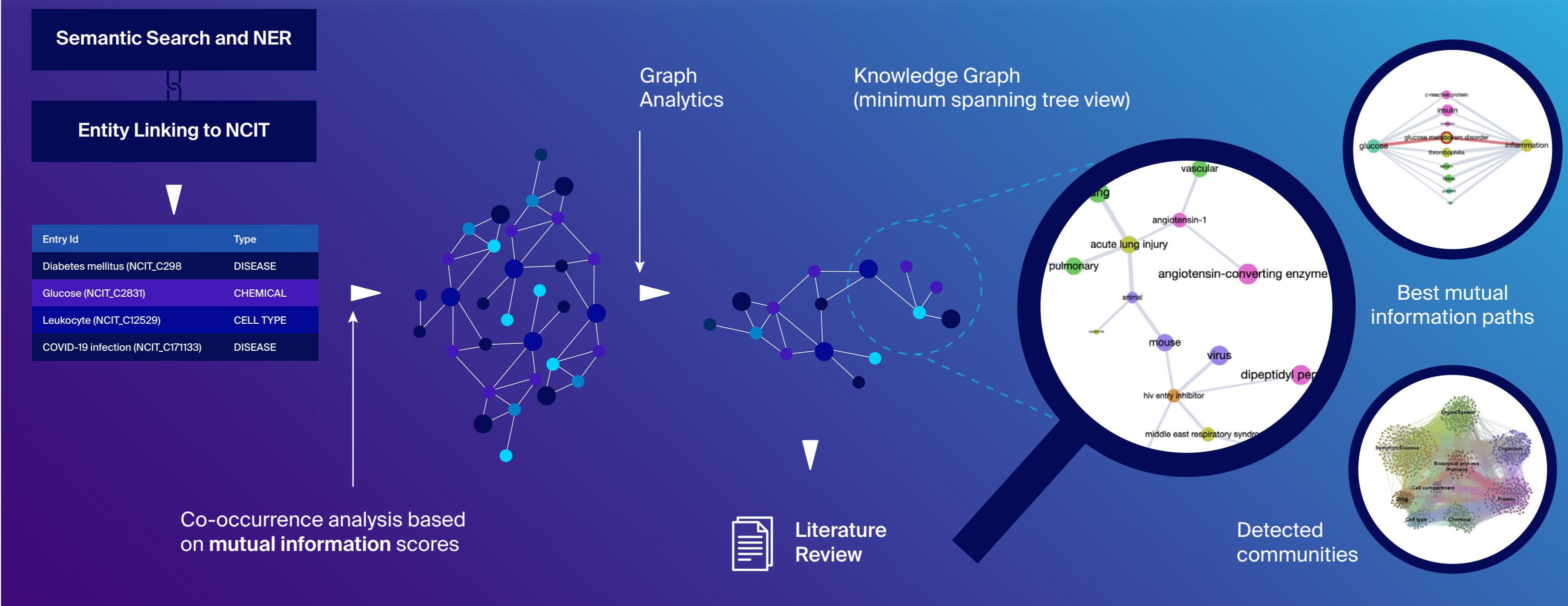
3. **Minimum spanning trees** allowed the user to find minimum informative substructures when dense Knowledge Graphs would become overwhelming

4. **A shortest path search coupled with mutual-information-based association measures** allowed users to explore the chains of entities forming association links between the nodes of interest

5. **Community detection methods** allowed users to discover groups of strongly associated entities

The Knowledge Graph guided review of COVID-19-related literature performed using Blue Graph enabled the BBP scientists to analyse the reported role of blood glucose levels in the severity of COVID-19. The resulting study was published in Frontiers in Public Health – **A machine-generated view of the role of blood glucose levels in the severity of COVID-19** https://doi.org/10.3389/fpubh.2021.695139

Blue Brain Search performs semantic literature search and entity extraction from text and **Blue Brain Nexus Forge** links extracted entities to the National Cancer Institute Thesaurus (NCIT) ontology to identify and disambiguate them.



Semantic Search and NER

Entity Linking to NCIT

Entry Id	Type
Diabetes mellitus (NCIT_C298	DISEASE
Glucose (NCIT_C2831)	CHEMICAL
Leukocyte (NCIT_C12529)	CELL TYPE
COVID-19 infection (NCIT_C171133)	DISEASE

Co-occurrence analysis based on **mutual information** scores

Graph Analytics

Knowledge Graph (minimum spanning tree view)

Literature Review

Best mutual information paths

Detected communities

About EPFL's **Blue Brain Project**

The aim of the EPFL Blue Brain Project, a Swiss brain research initiative founded and directed by Professor Henry Markram, is to establish simulation neuroscience as a complementary approach alongside experimental, theoretical and clinical neuroscience to understanding the brain, by building the world's first biologically detailed digital reconstructions and simulations of the mouse brain. As a research center at École polytechnique fédérale de Lausanne (EPFL), Blue Brain advocates EPFL's open science initiative, which seeks to maximize the reach and impact of research conducted at the school.

Blue Graph is open sourced in GitHub (*https://github.com/BlueBrain/BlueGraph*) under the Apache 2 license and released in PyPI.

Detailed Installation instructions are available in the Blue Graph documentation page (https://bluegraph.readthedocs.io/).

A set of hands-on tutorials are available:

- Covid-19 literature knowledge graph guided review use case:
  https://bluegraph.readthedocs.io/en/latest/tutorials/cord19kg_tutorials/getting_started_cord19kg.html
- Hands-on tutorials:
  https://bluegraph.readthedocs.io/en/latest/tutorials/bluegraph_tutorials/getting_started_bluegraph.html

Want to contribute ? An issue to report and a feedback to share ?
Reach out in Github and open a Github issue.

Software Authors

Eugenia Oshurko, Mohameth François Sy, Pierre-Alexandre Fonta, Anna-Kristin Kaufmann, and Jonathan Raël Lurie

Funding & Acknowledgements

The development of this project was supported by funding to the Blue Brain Project, a research center of the École polytechnique fédérale de Lausanne (EPFL), from the Swiss government's ETH Board of the Swiss Federal Institutes of Technology.

The authors thanks

Blue Brain's Molecular Biologist Dr. Emmanuelle Logette for her valuable feedback and review.

For more information, please contact:

Mohameth François Sy
Data and Knowledge Engineering Section Manager
mohameth.sy@epfl.ch

For press enquiries, please contact:

Kate Mullins
Communications Manager
Blue Brain Project
kate.mullins@epfl.ch