



MEDNARODNA
PODIPLOMSKA ŠOLA
JOŽEFA STEFANA

JOŽEF STEFAN
INTERNATIONAL
POSTGRADUATE SCHOOL

Community-based semantic subgroup discovery (CBSSD)

Blaž Škrlj, Anže Vavpetič, Jan Kralj, Nada Lavrač

September 16, 2017

Table of contents

General overview

Problem definition

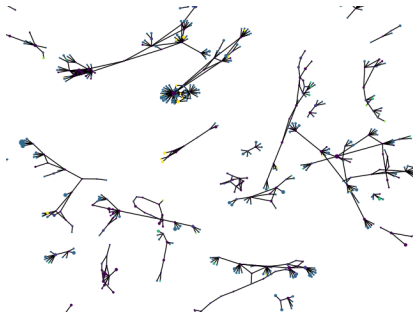
Proposed approach



Introduction

Properties of biological networks

- Multiple types of nodes and edges → heterogeneous networks
- Possible connections between distinct entities
- Large in some sub-domains
- Not trivial to interpret



How can an algorithm learn from a complex network?

Network representation

Important network features can be encapsured via community detection, graphlets, semantic clustering and other methods..

Example use

Such methodology is used to infer protein-RNA interactions, identify expression patterns, compare protein structures, fuse systems-level data etc.



Problem definition

Term-subset enrichment

Let t_1, t_2, \dots, t_n represent individual terms of interest from the whole term set ψ . Identify subsets $\Lambda_1, \Lambda_2, \dots, \Lambda_n \subseteq \psi$, which represent interpretable patterns, previously unknown to a human observer.

Example situation

Let G_1, G_2, \dots, G_n be n distinct genes we are interested in. Although individual genes, or the whole group of genes doesn't return any interesting results, we can further explore the subspace of n genes.

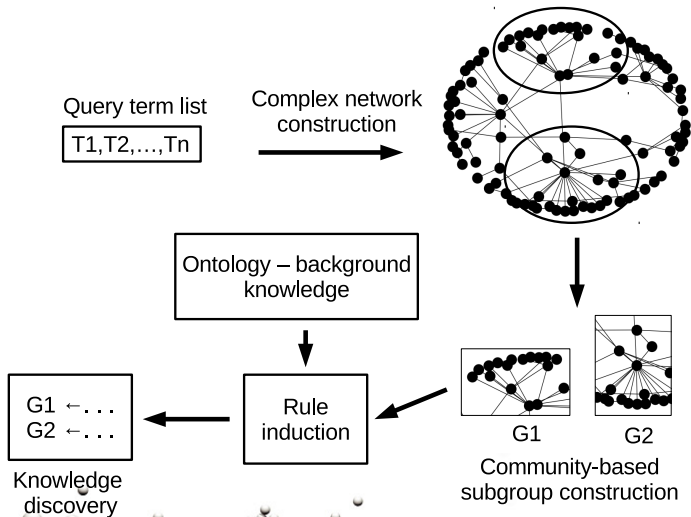
Problem

Exploring all possible combinations can be computationally expensive procedure, as there are $\sum_{i=2}^n \frac{n!}{(n-i)!i!}$ possible options.



Fighting the combinatorial explosion

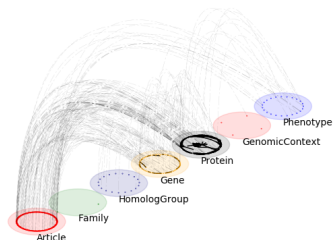
We argue there exists an efficient heuristic-based approach.



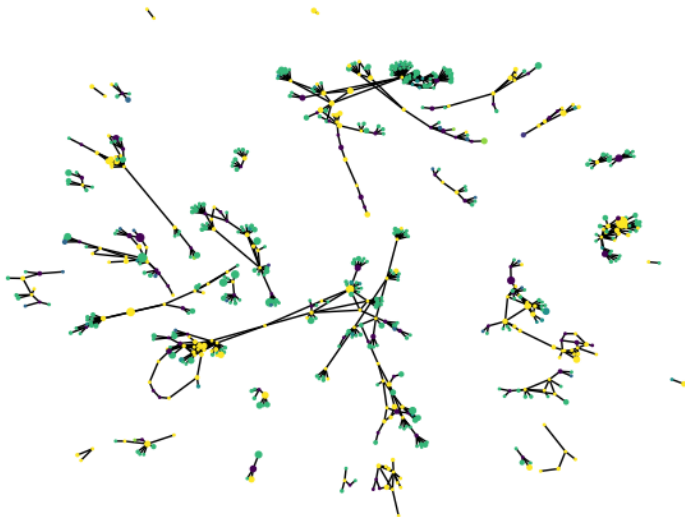
Fighting the combinatorial explosion - network construction

Basic procedure

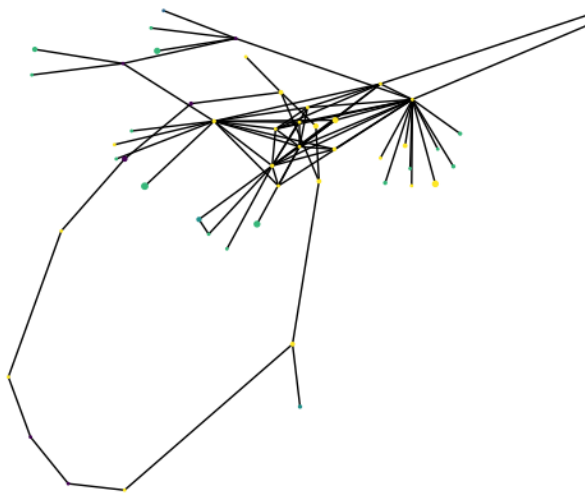
- collect network data on the studied phenomenon
- merge the data into a single heterogeneous network
- simplify the obtained network, or learn directly from it



Example BioMine network



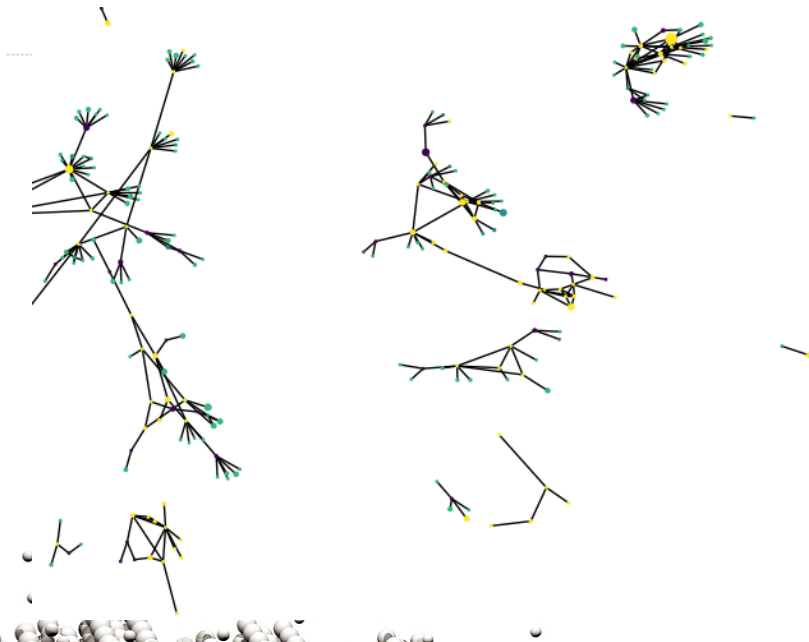
Example BioMine network



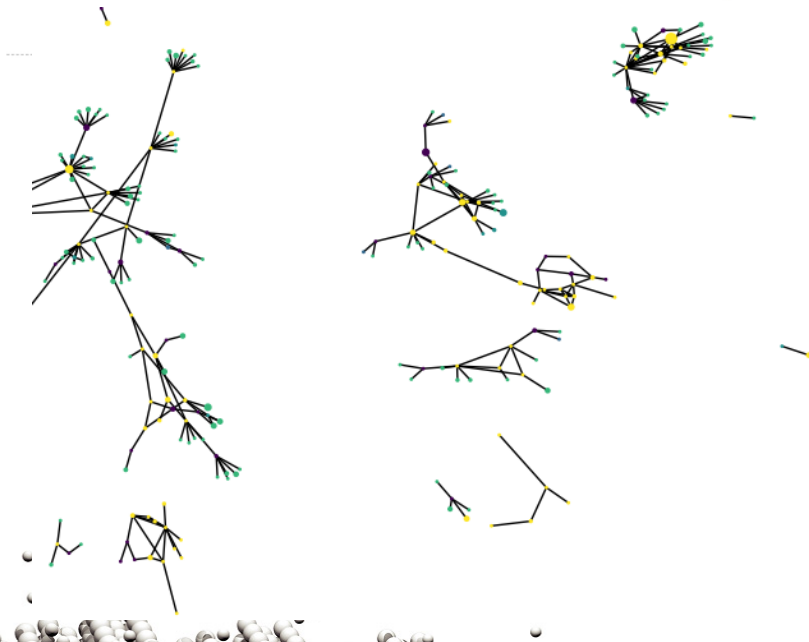
Example BioMine network



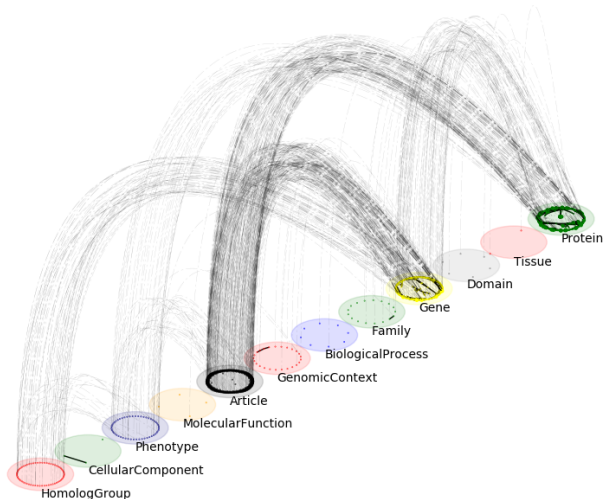
Example BioMine network



Example BioMine network



Up to 8 layers of interconnected information



Subset extraction

Extraction algorithm

A multiplex network Ψ is decomposed into individual communities via the use of *Louvain* algorithm. Initial term list ψ is then splitted according to community presence, such that initial terms t_1, \dots, t_n are assigned to a subset $\phi_{x,\dots,y} \in \psi$ iff $\exists t_{x,\dots,y} \in \Psi_{x,\dots,y}$.

