



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

---

0 1 2 3 4

General overview

Problem definition

Proposed approach

Use case



# Introduction

## Properties of biological networks

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



# How can an algorithm learn from a complex network?

Navigation icons: back, forward, search, etc.

## Network representation

Important network features can be encaptured 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  $\psi$ . 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.

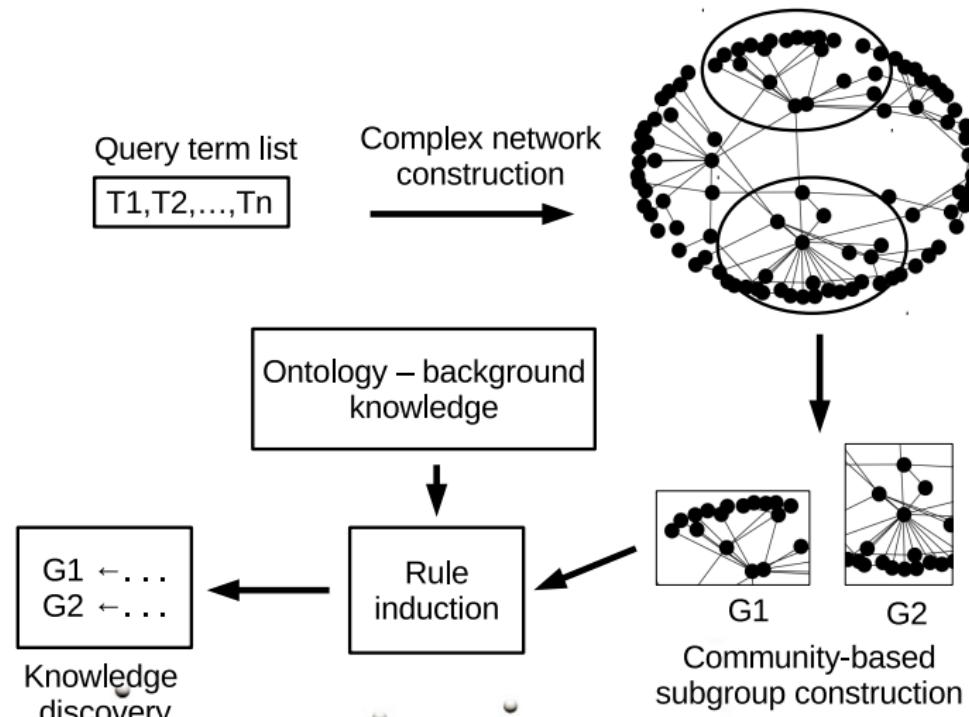
## Problem

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

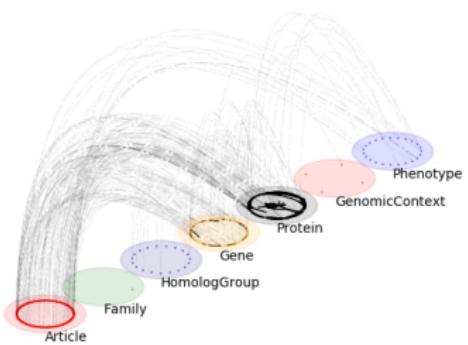
We argue there exists an efficient heuristic-based approach.



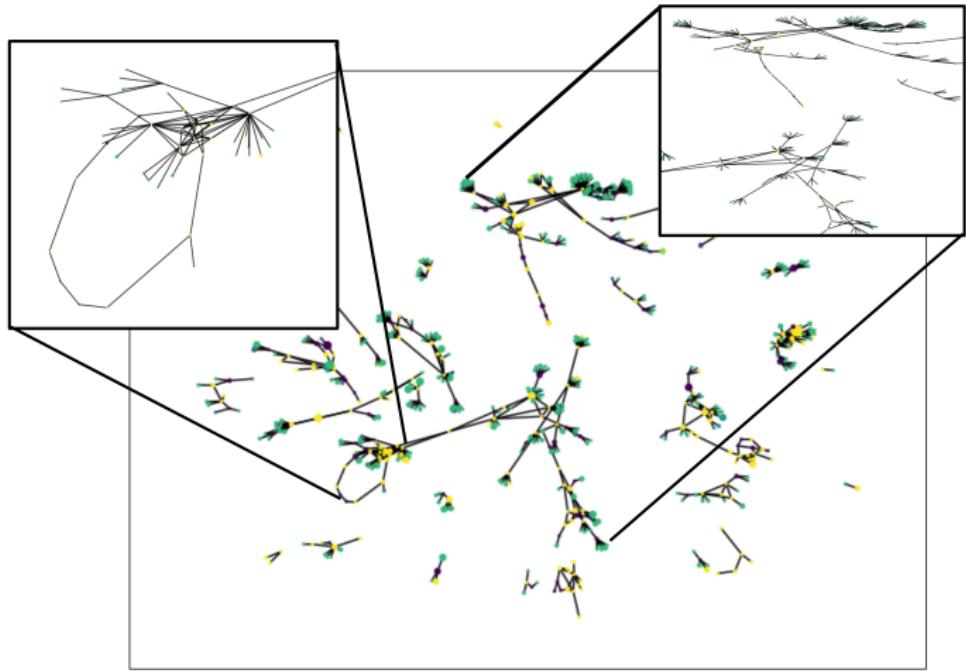
# Fighting the combinatorial explosion - network construction

## Basic procedure

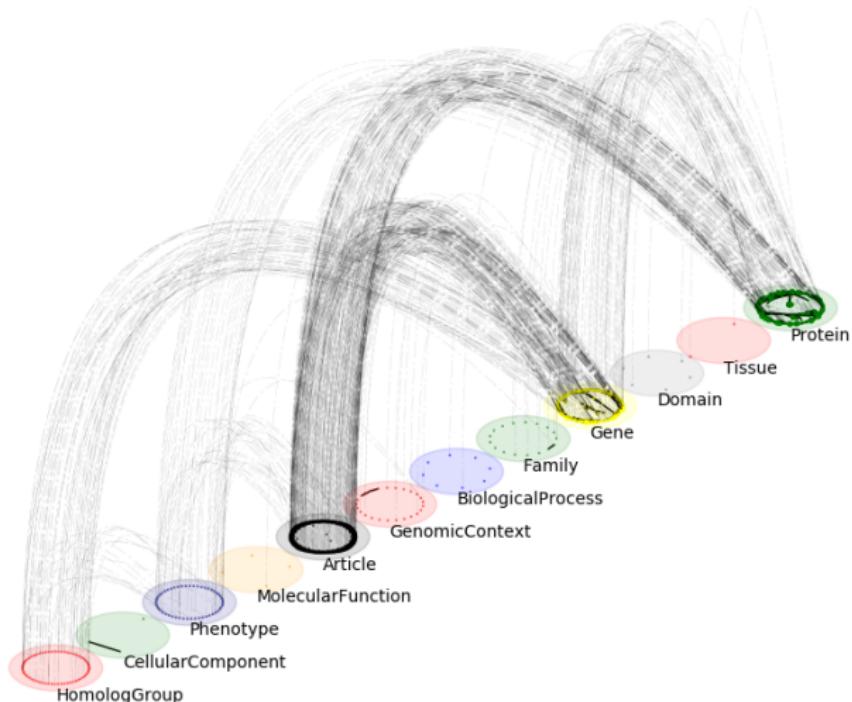
- collect network data on the studied phenomenon
- merge the data into a single heterogeneous network
- simplify the obtained network, or learn directly from it



# Example BioMine network



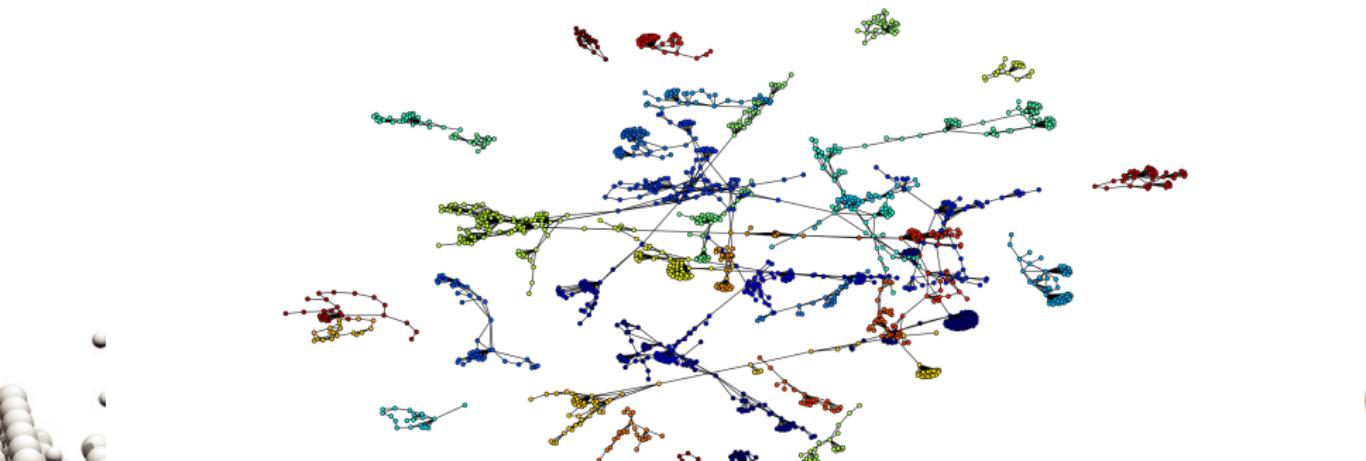
# Up to 12 layers of interconnected information



# Subset extraction

## Extraction algorithm

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



# Intermediary result

After community detection, the initial term list  $\Psi$  is splitted into n smaller subsets  $\zeta_{1,\dots,n}$ , which are interpreted as target classes for the semantic rule learning step. The objective function thus becomes:

## Learning objective

Learn a rule set  $\Delta_{1,\dots,n}$  for individual classes  $\zeta_{1,\dots,n}$  using background knowledge ( $\Xi$ ) in form of ontologies, such that the likelihood of individual class  $\zeta_x; x \in \{1, \dots, n\}$  is maximized.

$$\Delta_{\zeta_1, \dots, \zeta_n} = \underset{i \in R_1, \dots, R_n}{\operatorname{argmax}} \left[ P(R_i | \Xi) \right] \quad (1)$$

# Semantic rule learning

## Background knowledge representation

- Individual term sets projected into  $\Delta(BK)$  space
- Extensive knowledge from  $GO$ ,  $PFAM$ ,  $KEGG$ , .. used
- Rules can be generalized according to  $BK$
- Semi-supervised, multi-class learning problem

$$R_1 \rightarrow \bigwedge_i GO_i; i \in \{GO_{R_1}\}$$

$$R_2 \rightarrow \bigwedge_i GO_i; i \in \{GO_{R_2}\}$$

...

$$R_n \rightarrow \bigwedge_i GO_i; i \in \{GO_{R_n}\}$$



# Practical example - SNPs

## Description

SNPs within protein binding sites previously identified term set  $\psi$  indicates association with DNA-related processes, cancer development and membrane mechanisms. We were interested, whether there exist more detailed, yet generalized properties of the studied phenomenon.

# Practical example - SNPs (2)

