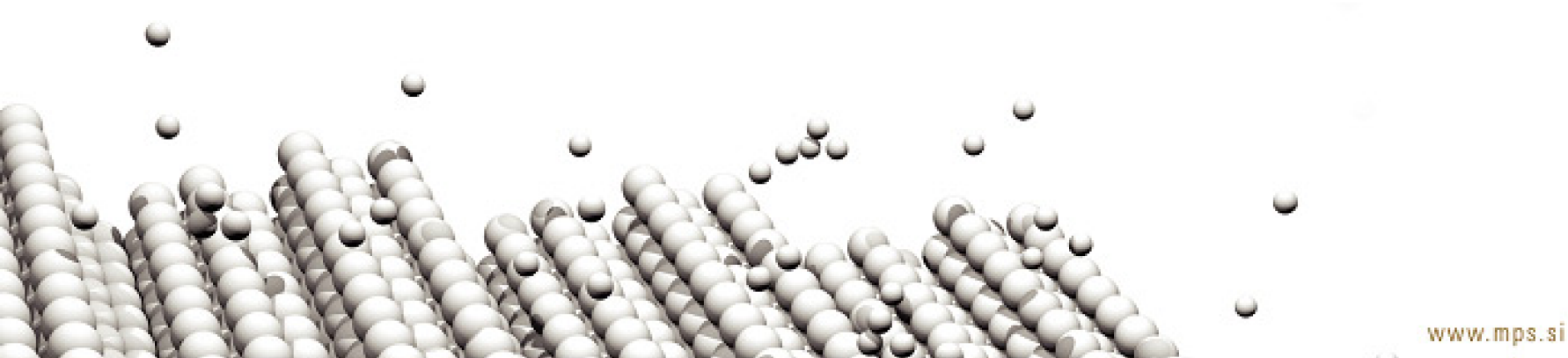

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA

JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL

# Community-based semantic subgroup discovery (CBSSD)

Blaž Škrlj, Anže Vavpetič, Jan Kralj, Nada Lavrač

September 15, 2017



# Table of contents



# Introduction

## Properties of biological networks

- Multiple types of nodes and edges $\rightarrow$ 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?

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

## Example situation

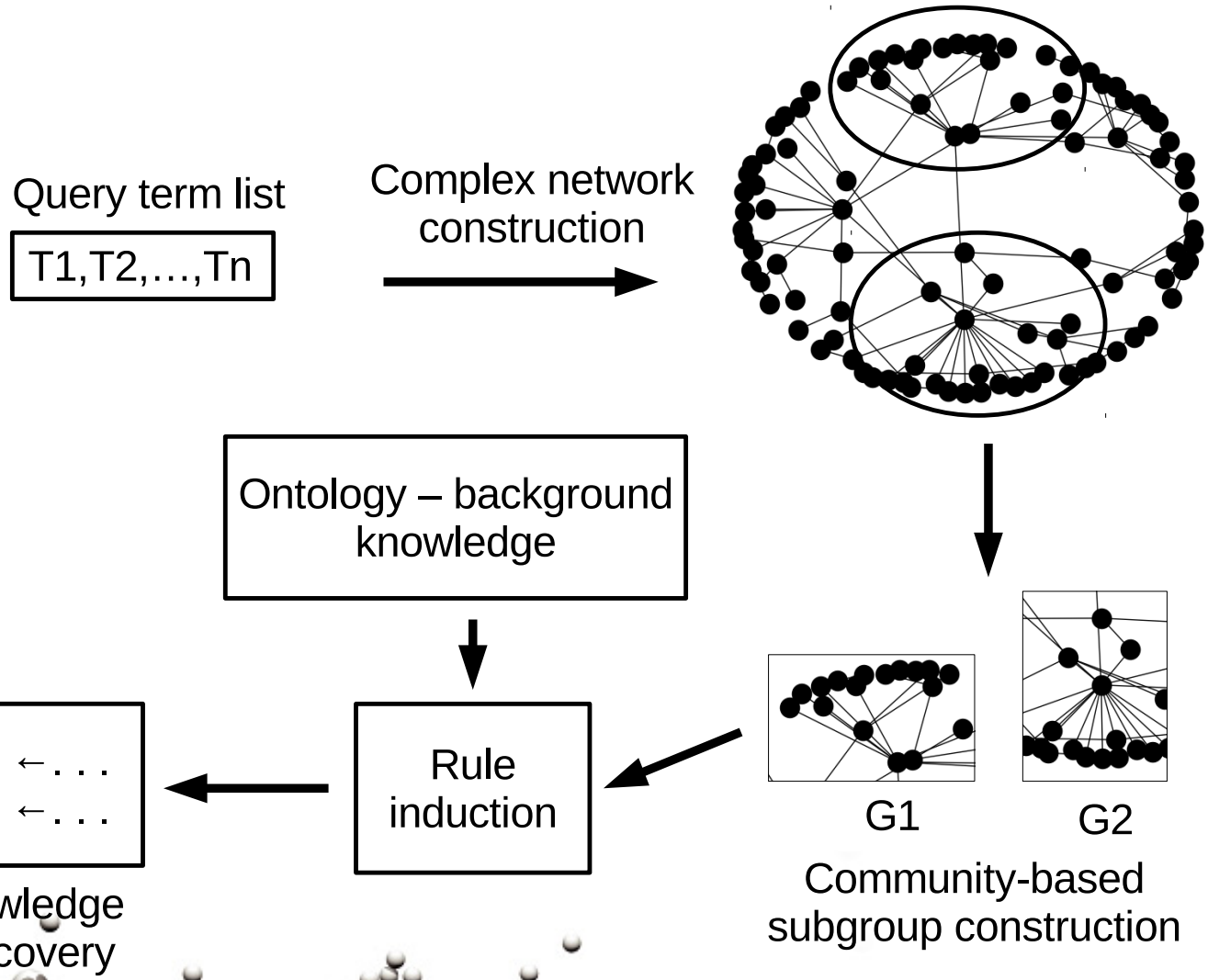
Let $G_1, G_2, .., G_n$ be $n$ distinct genes we are interested in. Although individual genes, or the whole group of genes doesn't return any interesting results, we can further explore the subspace of n genes.

## 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 heterognous network
- simplify the obtained network