



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 21, 2017

Table of contents

1 General overview

2 Problem definition

3 Proposed approach

- Network generation
- Subset partition
- Learning

4 Use case

- Polymorphisms

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?

Network representation

Important network features can be captured 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.

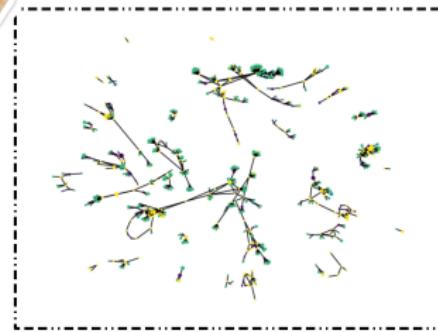
Problem definition (2)



Curated knowledge

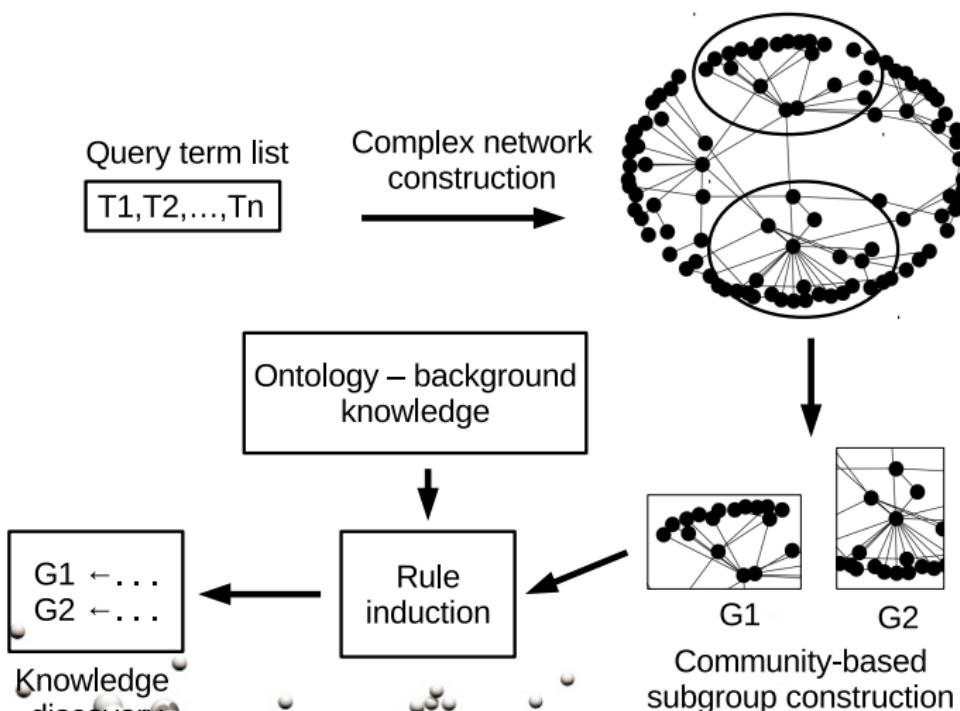


Complex networks



Fighting the combinatorial explosion

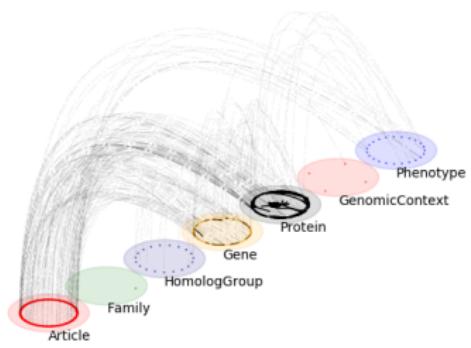
We argue there exists an efficient heuristic-based approach.



Fighting the combinatorial explosion - network construction

Basic procedure

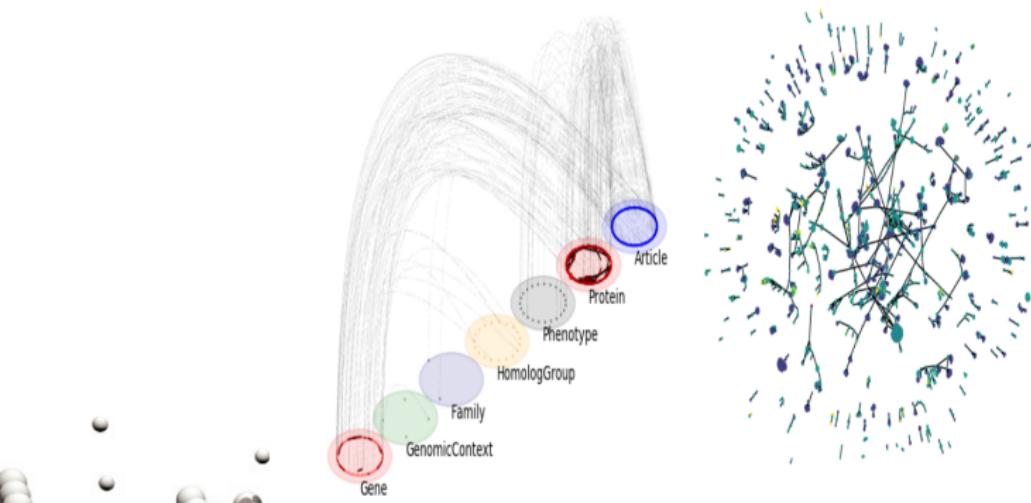
- collect network data on the studied phenomenon
- merge the data into a single heterogeneous network
- simplify the obtained network, or learn directly from it



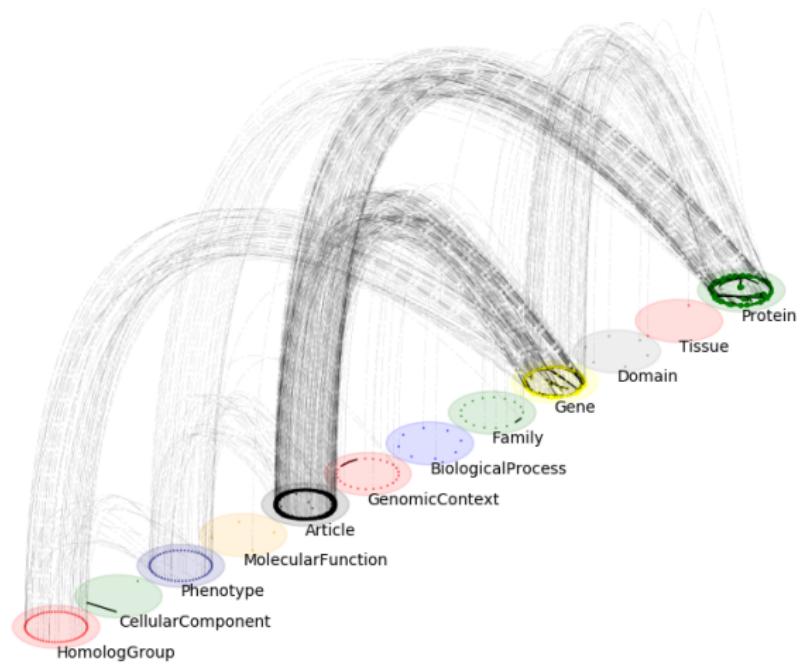
Example BioMine network

Algorithm

A *crawler unit* is spawned for individual terms in the input term list.
Result is a knowledge network associated with the input terms.



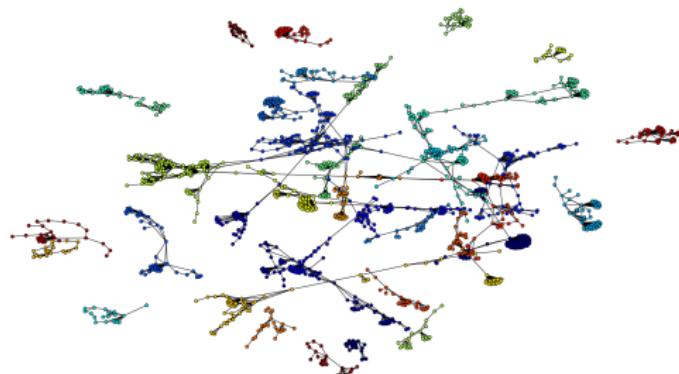
Up to 12 layers of interconnected information



Subset extraction

Extraction algorithm

A multiplex network is decomposed into individual communities via the use of *Louvain* or *multiplex InfoMap* algorithms. Initial term list $\Psi_{1,\dots,n}$ is then splitted according to community presence, such that each subset of terms $\Psi_{x \in \{1,\dots,n\}}$ is assigned to a target class ζ_x .



Intermediary result

Constructed $\zeta_{1,\dots,n}$ are interpreted as target classes for the semantic rule learning step. The objective function thus becomes:

Learning objective

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

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

Semantic rule learning

Background knowledge representation

- Individual term sets are first projected into $\Xi(BK)$ space
- Extensive knowledge from GO , $PFAM$, $KEGG$, .. used
- Rules can be generalized
- Supervised descriptive learning task

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

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

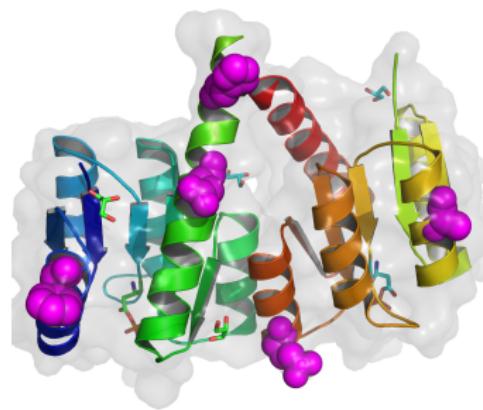
...

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

Practical example - SNPs

Description

SNPs within protein binding sites previously identified term set ψ indicates association with DNA-related processes, cancer development and membrane mechanisms. We were interested, whether there exist general explanations for latent, community-based patterns.



Practical example - SNPs (2)

General result

Largest communities corresponded to the most significant terms, identified in the previous study - proof of structure-based enrichment

Example

Many conjuncts emerged for more marginal communities - this is the new knowledge. For example: *CASR* gene was in previous study not directly associated with arginine-binding process.

Current performance

A system, capable of learning knowledge representations, **similar to the ones obtained directly by literature examination.**

Short-term development

Upgraded CD

The InfoMap equation is already used to incorporate the information on multiplex edges.

Ontology processing speedup

Recent work by *Kralj et al.* present a method capable of significant speedups (10-50x) at no cost with regard learned knowledge representations. This approach makes CBSSD scale better

Representation learning in streaming context

Can the CBSSD learn and update its knowledge on the fly?



Thank you for your attention!