Link Prediction via Community Detection in Bipartite Multi-Layer Graphs

Maksim Koptelov, Albrecht Zimmermann and Bruno Crémilleux

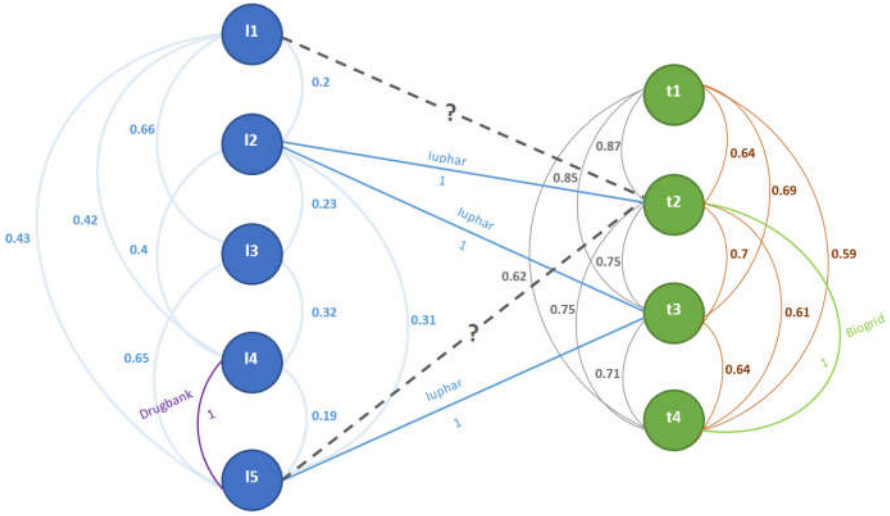
Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, France

Problem setting

Given: set of drugs, set of biological targets, drug-target interaction examples, drug-drug and target-target similarity information.

Problem: predict new drug-target interactions

Problem representation: link prediction in bipartite multi-layer graph



Challenges:

- Number of layers in a graph can be any
- Computation cost minimization
- Interpretability of the approach

State-of-the-art

- Vertices of graphs can be grouped into communities by a community detection approach (Spectral partitioning, The Louvain algorithm etc.)
- Connections within/between communities can be exploited to predict links between not-directly connected vertices (assuming that the two vertices are in the same community, or by exploiting that their neighbors are in the same community)
- Neighborhood measures described by Liben-Nowell and Kleinberg in 2007 can be used to measure the probability of the links.
- Existing community-based measures can be used as well:

  - CAR-based measures by Cannistraci et al. (2003)

  - Neighboring community-based measures by Xie et al. (2014)

  - Community relevance Jaccard coefficient by Ding et al. (2016).

Our approach

Community detection method:
- *Spectral partitioning* (Lescovec et al., 2014) or *Louvain algorithm* (Blondel et al., 2008)

Community matching:
- Some communities are *pure* (containing either type, drugs or targets only)
- *Mixed* communities (containing both types of vertices) are split into pure ones

Community to community (CC) or *Node to community (NC)*

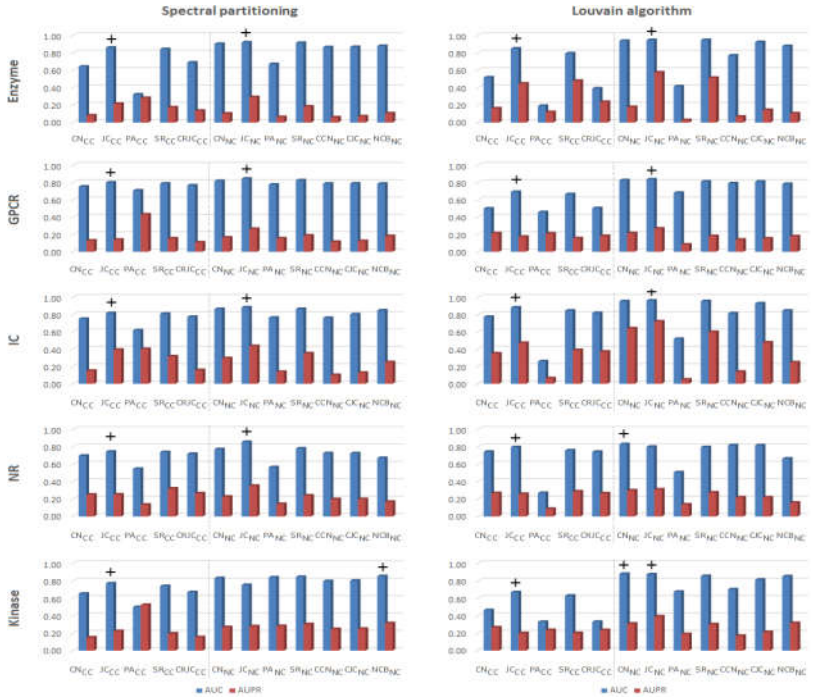
Link probability score:

1. *Common neighbors (CN)*: $CN(d_i, t_i) = |\{v \mid (d_i, v) \in E\} \cap \{u \mid (t_j, u) \in E\}|$
2. *The Jaccard coefficient (JC)*: $JC(d_i, t_j) = \frac{CN(d_i, t_j)}{|\{v \mid (d_i, v) \in E\} \cup \{u \mid (t_j, u) \in E\}|}$
3. *Preferential attachment (PA)*: $PA(d_i, t_i) = |\Gamma(d_i)| \cdot |\Gamma(t_j)|, \ |\Gamma(v)| = deg(v)$
4. *SimRank (SR)*: $SR(d_i, t_j) = \frac{CN(d_i, t_j)}{PA(d_i, t_j)}$
5. *CAR-based common neighbors (CCN)*: see Cannistraci et al., 2003
6. *CAR-based Jaccard coefficient (CJC)*: see Cannistraci et al., 2003
7. *Neighboring community-based (NCB)*: see Xie et al., 2014
8. *Community relevance Jaccard coefficient (CRJC)*: see Ding et al., 2016

Results

We test:

- **2 different community detection approaches:** *Spectral partitioning* and *Louvain algorithm*
- **8 link prediction measures:** *CN, JC, PA, SR, CCN, CJC, NCB, CRJC* (see poster for more details)
- **2 community-matching techniques:** *community-to community* and *node-to-community*
- **5 different data sets:** *Enzyme, GPCR, IC, NR* and *Kinase*



Main result:

- **Performance** of the approach with optimal parameters is **close to the state-of-the-art**