

Graph learning for prediction of drug-disease interactions

Preliminary results

Andrej Kastrin and Dimitar Hristovski

June 2, 2019

University of Ljubljana

Table of contents

1. Motivation
2. Methods
3. Results
4. Conclusions

Motivation

Classical tasks on complex networks

- Node classification
 - *Predict type of a given node*

Classical tasks on complex networks

- Node classification
 - *Predict type of a given node*
- Community detection
 - *Identify clusters of nodes*

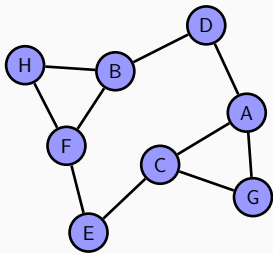
Classical tasks on complex networks

- Node classification
 - *Predict type of a given node*
- Community detection
 - *Identify clusters of nodes*
- Network similarity
 - *How similar are two (sub) networks*

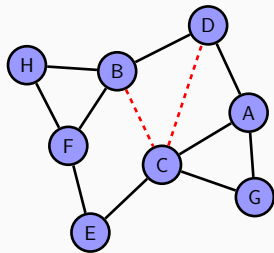
Classical tasks on complex networks

- Node classification
 - *Predict type of a given node*
- Community detection
 - *Identify clusters of nodes*
- Network similarity
 - *How similar are two (sub) networks*
- **Link prediction**
 - *Predict whether two nodes will be linked*
 - Knowledge graph completion

Link prediction task

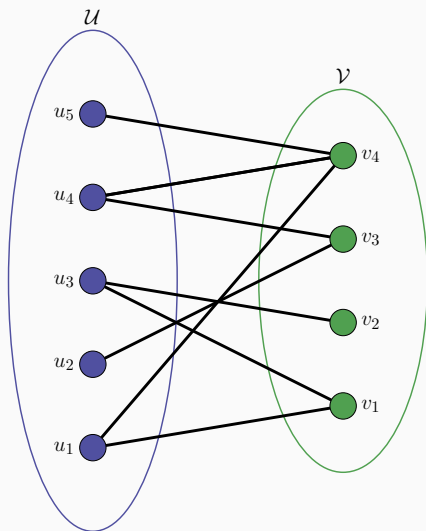


Training network $[t_1, t_2]$



Test network $[t_3, t_4]$

Bipartite network



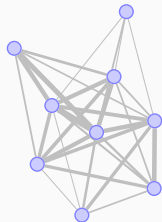
\mathcal{U} = diseases, \mathcal{V} = drugs

Methods

Methods overview



SemMedDB graph database



Network is constructed by selecting
TREATS relation from SemMedDB

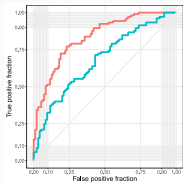


$$\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{pmatrix}$$

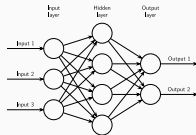
Graph representation as embedding



User evaluates the ranked list
of predicted relations



Prediction performance evaluation



Link prediction using neural link
predictor

Example

Second phase of a double-blind study clinical trial on Sibutramine for the treatment of patients suffering essential obesity

(PMID: 11360159)

SemRep

- Output:

(Sibutramine) – [:TREATS] – > (Obesity)

Example

Second phase of a double-blind study clinical trial on Sibutramine for the treatment of patients suffering essential obesity

(PMID: 11360159)

SemRep

- Output:

(Sibutramine) – [:TREATS] – > (Obesity)

- UMLS Concept

Example

Second phase of a double-blind study clinical trial on Sibutramine for the treatment of patients suffering essential obesity

(PMID: 11360159)

SemRep

- Output:

(Sibutramine) – [:TREATS] – > (Obesity)

- UMLS Concept



Example

Second phase of a double-blind study clinical trial on Sibutramine for the treatment of patients suffering essential obesity

(PMID: 11360159)

SemRep

- Output:

(Sibutramine) – [:TREATS] – > (Obesity)

- UMLS Concept
- UMLS Semantic relation



Example

Second phase of a double-blind study clinical trial on Sibutramine for the treatment of patients suffering essential obesity

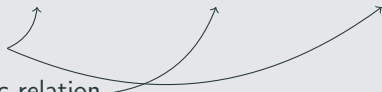
(PMID: 11360159)

SemRep

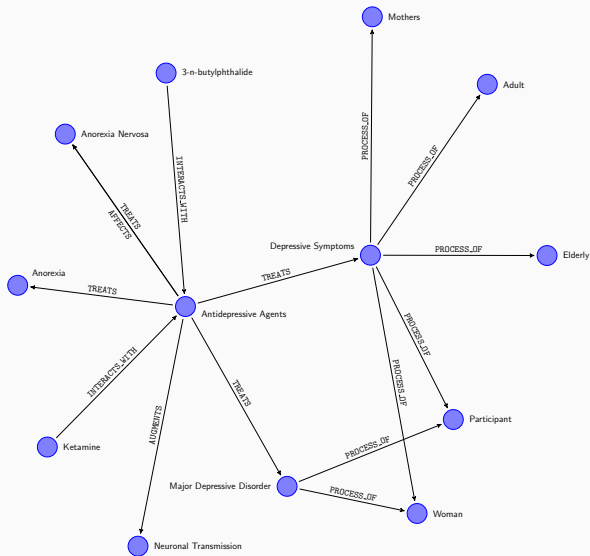
- Output:

(Sibutramine) – [:TREATS] – > (Obesity)

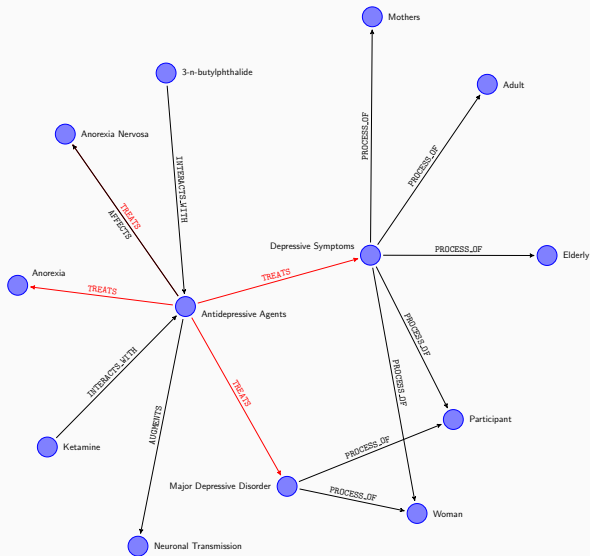
- UMLS Concept
- UMLS Semantic relation



Semantic MEDLINE



Semantic MEDLINE



Baseline predictors for bipartite networks

Definition

Define $\Gamma'(u) = \bigcup_{c \in \Gamma(u)} \Gamma(c)$ which is a set of neighbors of node u 's neighbors.

Baseline predictors for bipartite networks

Definition

Define $\Gamma'(u) = \bigcup_{c \in \Gamma(u)} \Gamma(c)$ which is a set of neighbors of node u 's neighbors.

Definition

1. Common neighbors:

$$s_{CN'}(u, v) = |\Gamma(u) \cap \Gamma'(v)|$$

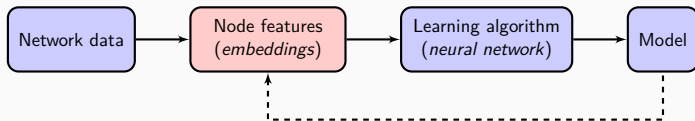
2. Jaccard coefficient:

$$s_{JC'}(u, v) = \frac{|\Gamma(u) \cap \Gamma'(v)|}{|\Gamma(u) \cup \Gamma'(v)|}$$

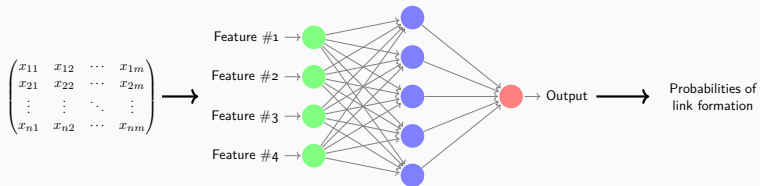
3. Adamic/Adar:

$$s_{AA'}(u, v) = \frac{1}{\log(|\Gamma(u) \cap \Gamma'(v)|)}$$

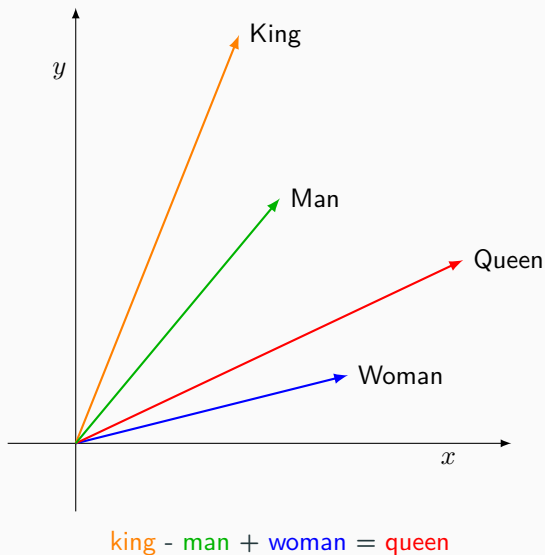
Machine learning workflow



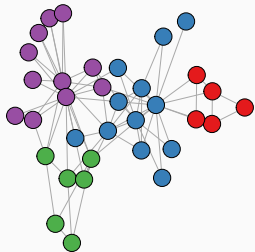
Machine learning workflow II



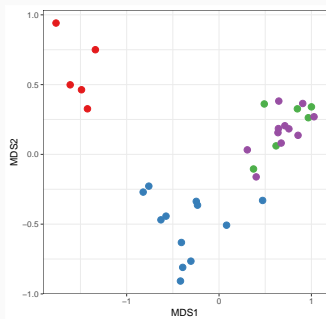
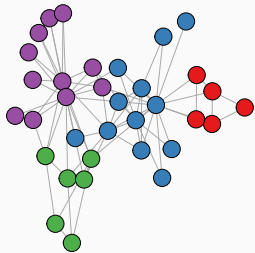
Word embeddings: word2vec



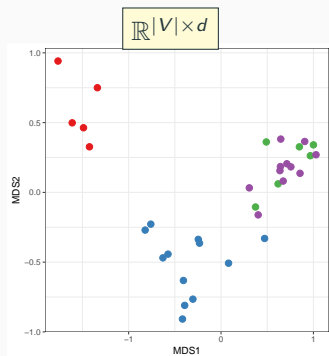
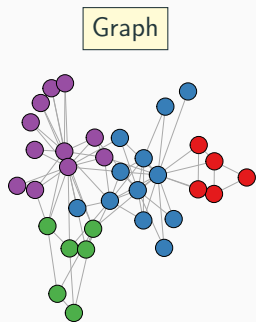
Network embedding idea



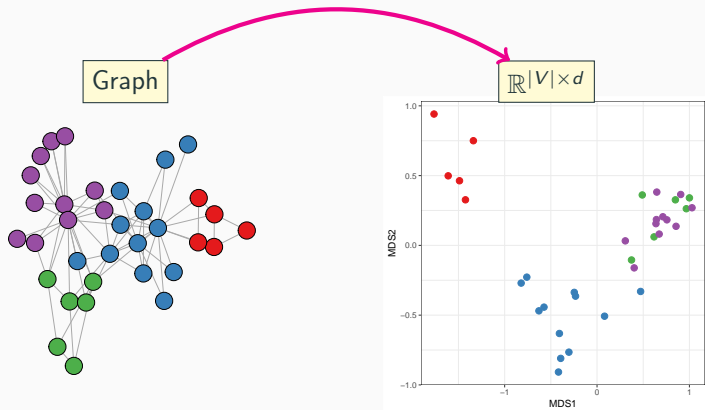
Network embedding idea



Network embedding idea



Network embedding idea

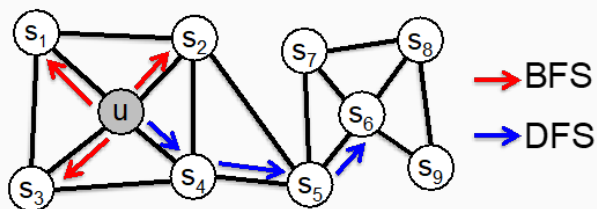


DeepWalk (Perozzi et al., 2013)

1. Simulate many short random walks starting from each node
2. For each node u , get nearby nodes as a sequence of nodes visited by random walks starting at u
3. For each node u , learn its embedding by predicting which nodes are in the neighborhood

Main idea

Use flexible, biased random walks that can trade off between **local** and **global** views of the network.



From node vectors (embeddings) to link vectors (embeddings)

Definition

Define binary operator over node vectors $f(u)$ and $f(v)$ to generate composite link representation $g(u, v)$. We consider:

1. Concatenation:

$$u_i + v_i$$

2. Average:

$$\frac{u_i \oplus v_i}{2}$$

3. Hadamard product:

$$u_i \odot v_i$$

From node vectors (embeddings) to link vectors (embeddings)

Definition

Define binary operator over node vectors $f(u)$ and $f(v)$ to generate composite link representation $g(u, v)$. We consider:

1. Concatenation:

$$u_i + v_i$$

2. Average:

$$\frac{u_i \oplus v_i}{2}$$

3. Hadamard product:

$$u_i \odot v_i$$

Example (Hadamard)

$$\begin{pmatrix} 1 \\ 2 \end{pmatrix} \odot \begin{pmatrix} 3 \\ 4 \end{pmatrix} = \begin{pmatrix} 1 \cdot 3 \\ 2 \cdot 4 \end{pmatrix} = \begin{pmatrix} 3 \\ 8 \end{pmatrix}$$

Results

Classification performance

Method	Binary operator	AUROC	AUPR	Prec k	mAP
CN	–	0.86	0.86	0.86	0.64
JC	–	0.85	0.84	0.86	0.62
AA	–	0.81	0.74	0.82	0.54
DeepWalk	Co	0.83	0.86	0.96	0.79
	Av	0.83	0.86	0.97	0.80
	Ha	0.72	0.72	0.82	0.65
node2vec	Co	0.83	0.86	0.96	0.80
	Av	0.83	0.86	0.97	0.81
	Ha	0.72	0.73	0.83	0.65

Note: CN = Common Neighbors, JC = Jaccard Coefficient, AA = Adamic/Adar; (Co)ncatenate, (Av)erage, and (Ha)damard binary operator; AUROC = area under ROC curve, AUPR = area under PR curve, Prec@ k = precision at k , mAP = mean average precision

Conclusions

Summary

- We investigate the representation learning in bipartite drug-disease network of semantic predications
- We design a deep neural network model that includes the graph structure into the embedding
- We found evidence that DeepWalk and node2vec outperformed baseline predictors in terms of $\text{Prec}@k$ and mAP
- Future work:
 - Extend the study to all predication types in SemMedDB
 - Domain expert for results interpretation

