Hasso Plattner Institut
Digital Engineering · Universität Potsdam

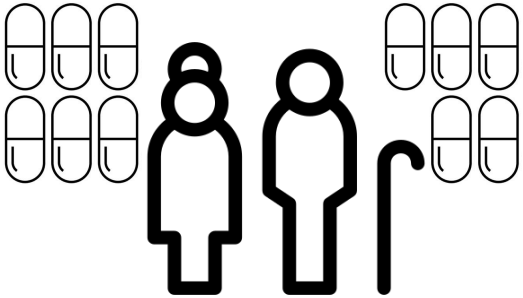
Exploring Graph Neural Networks for Drug Interaction Prediction

Henrik Wenck, Simon Witzke

Agenda

1. Context
2. Problem
3. Background - Link Prediction
4. Related work
5. Our data
6. Our approach
7. Results and evaluation
8. Discussion
9. Future work

Context



Simon's Grandparents

In 2011/12 **15%** of US population were affected by polypharmacy → increase in morbidity and mortality costing **>$177 billion** a year in treatment

Problem

Side-Effect prediction → Drug-Drug interaction prediction → **Multi-Relational Link Prediction** → possible to predict potential side-effects without medical tests

Approaches:

- Knowledge Graphs
- GNN Classifiers

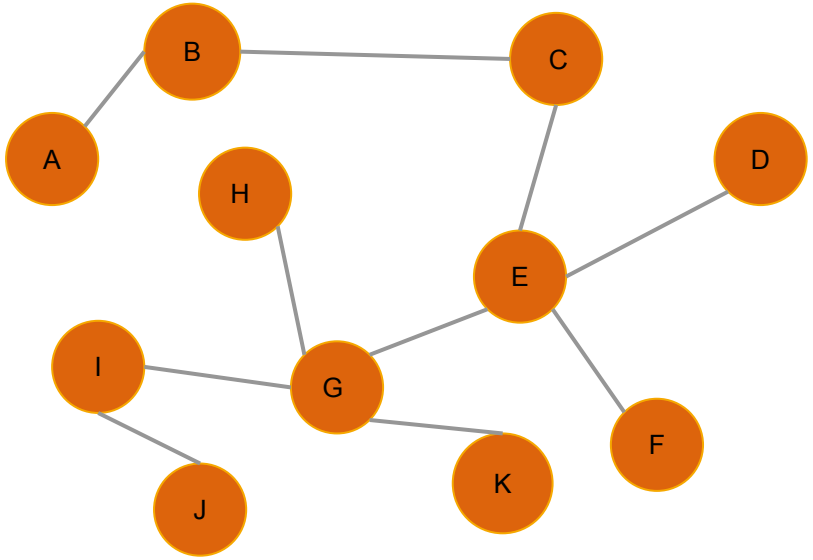
Challenges:

- Large graphs when taking for example protein reactions into account
- High quality outcomes necessary to make models useful

Background - the simple approach

A variety of problems can be structured as graph.

time t

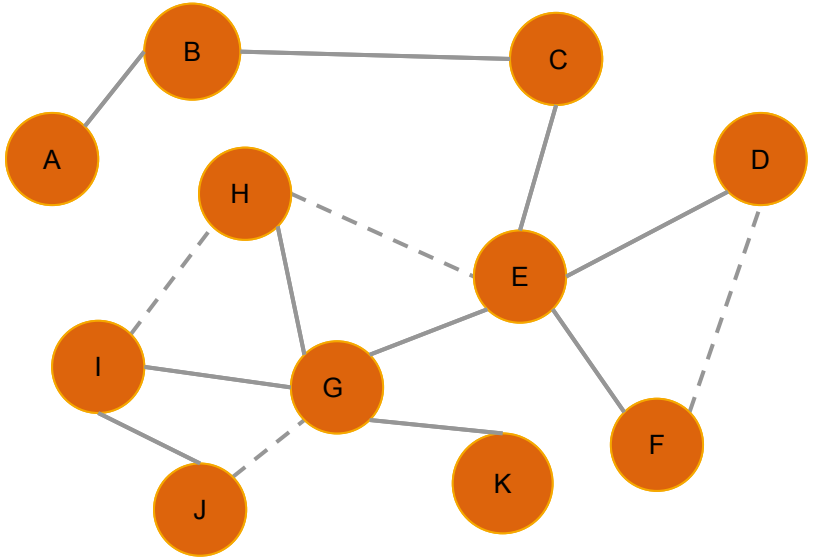


In our case we are interested in what connections might emerge between nodes in the future.

Background - the simple approach

A variety of problems can be structured as graph.

time t

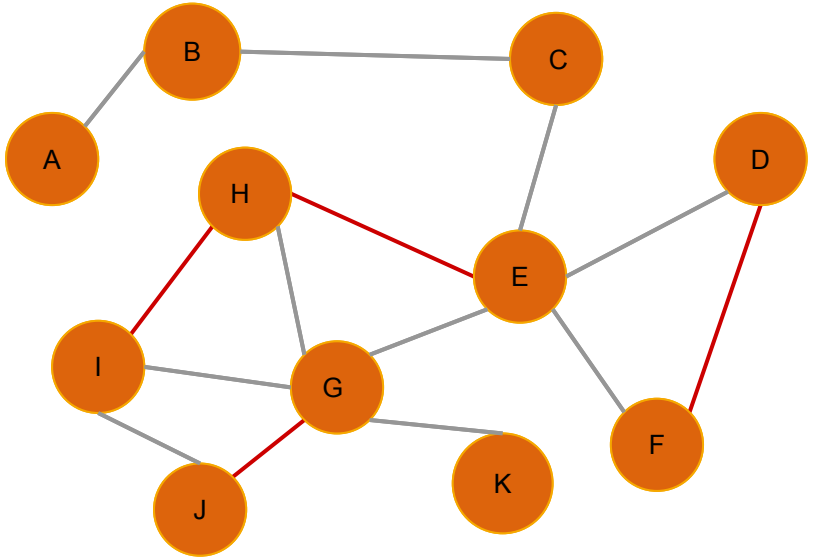


In our case we are interested in what connections might emerge between nodes in the future.

Background - the simple approach

A variety of problems can be structured as graph.

time t+1

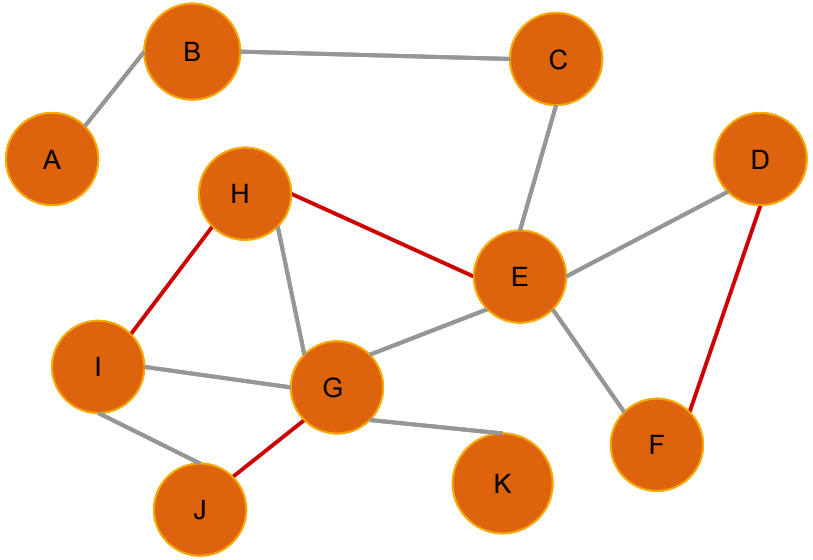


In our case we are interested in what connections might emerge between nodes in the future.

Background - the simple approach

A variety of problems can be structured as graph.

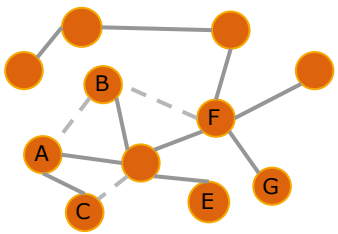
time t+1



How ?

In our case we are interested in what connections might emerge between nodes in the future.

Background - the simple approach



time t

Generate positive and negative samples

positive samples:
A - 1 - B
C - 1 - D
B - 1 - F

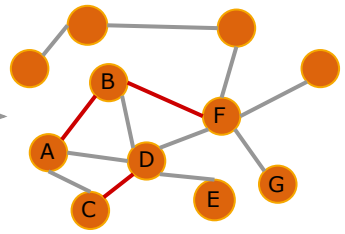
negative samples:
A - 0 - F
A - 0 - E
C - 0 - E
D - 0 - G
B - 0 - C
...

often highly imbalanced

Build Features

Prediction

Model



time t+1

Background - Latent Link Prediction Techniques

Tensor Decomposition	Geometric	Deep Learning
DistMult	TransE	ConvE
ComplEx	STransE	ConvKB
ANALOGY	TorusE	ConvR, CapsE, RSN
SimplE	CrossE, RotatE	
TriVec, TuckER		

t

$$\phi(\mathbf{h}, \mathbf{r}, \mathbf{t}) = \mathbf{h} \times \mathbf{r} \times \mathbf{t}$$

$$\phi(\mathbf{h}, \mathbf{r}, \mathbf{t}) = \delta(\tau(\mathbf{h}, \mathbf{r}), \mathbf{t})$$

Related work

- Approaches taking protein information into account:
  - Decagon
  - Knowledge-Graph Completion
- General Link Prediction approaches:
  - SEAL (Subgraphs, Embeddings & Attributes for LP)

We used general approaches to test their effectiveness on the polypharmacy problem.

Decagon

Two main components:

1. An encoder: a graph convolutional network (GCN) operating on the graph and producing embeddings for nodes,
2. A decoder: a tensor factorization model using these embeddings to model polypharmacy side effects.

Pros:

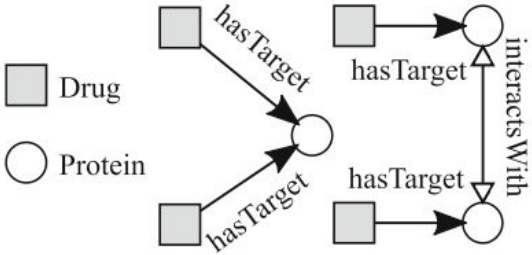
- Ability to predict multirelational links
- Sharing of information between edge type can improve performance

Cons:

- Performance varies with molecular basis → more data with higher complexity is required

Knowledge-Graph Completion

- Malone et al. use relational features to create an interpretable embedding by using KBLRN, a framework for end-to-end learning of knowledge base representations



Pros:

- Interpretability of features → suggest hypothesis for wet lab validation

Cons:

- Requires complex data including protein-protein interactions to properly make use of relational features

Source: Brandon Malone, Alberto García-Durán, and Mathias Niepert. 2018. Knowledge Graph Completion to Predict Polypharmacy Side Effects

DGCNN

1. **Graph convolutional layers**

$$\mathbf{Z}^{t+1} = f(\tilde{\mathbf{D}}^{-1}\tilde{\mathbf{A}}\mathbf{Z}^{t}\mathbf{W}^{t})$$

1.
2.
3.

$$\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$$

$$\tilde{\mathbf{D}}_{ii} = \sum_{j} \tilde{\mathbf{A}}_{ij}$$

$$\mathbf{Z}^{0} = \mathbf{X}$$

$$\mathbf{Z}^{t} \in \mathbb{R}^{n \times c_t}$$

$$\mathbf{W}^{t} \in \mathbb{R}^{c_t \times c_{t+1}}$$

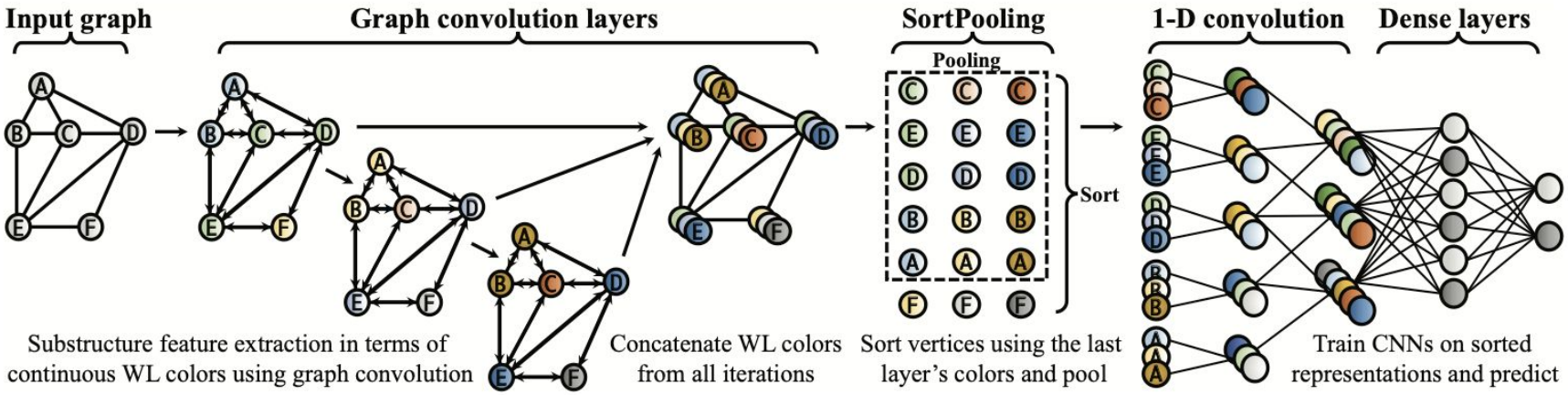
2. **SortPooling layer**

Input shape: $n \times \sum_{1}^{h} c_t$

Output shape: $k \times \sum_{1}^{h} c_t$

3. **Traditional convolutional and dense layers**
   - classic convolutional neural network for binary prediction (softmax)

DGCNN

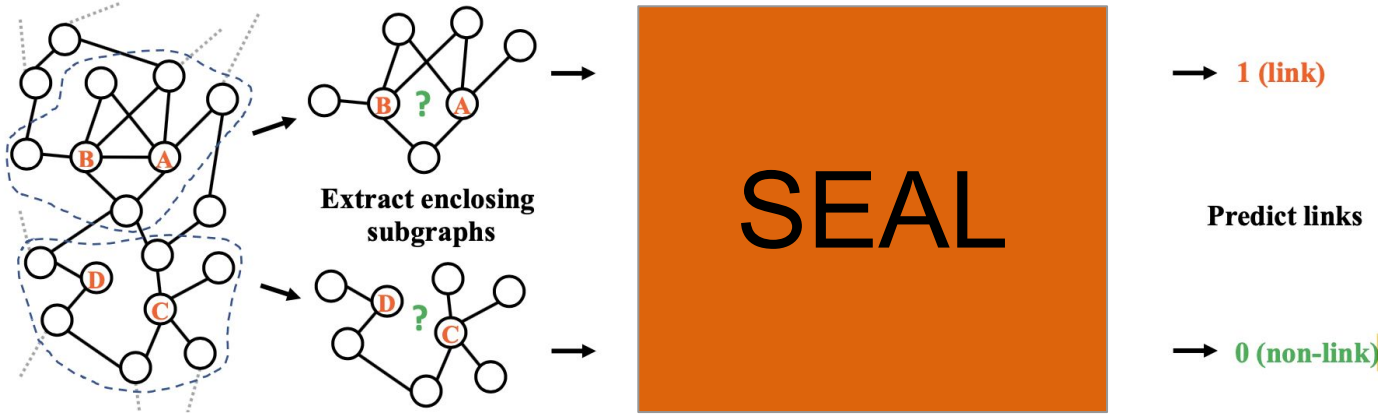


Source: Muhan Zhang and Yixin Chen. 2018. Link prediction based on graph neural networks.

SEAL

1. **Enclosing subgraph extraction**
2. **Node information matrix construction**
   - Double-Radius Node Labeling to assign integer label to every node in subgraph to represent structure
     - Advantage: perfect hashing function → allows fast closed-form computation
   - Incorporating latent and explicit features
     - Append data to corresponding row in X
     - Calculation of embeddings non-trivial → if we calculate embeddings on graph we encode link existing information of training links → easy to overfit on this → negative injection
3. **GNN Learning**
   - Use DGCNN to learn classifier on subgraphs

SEAL


B
A
D
C
B
?
A
D
?
C
Extract enclosing subgraphs
SEAL
1 (link)
Predict links
0 (non-link)

Our data*

TWOSIDES
63,473 drug combinations
1,318 side-effects
4,651,131 combination-side-effect associations

PPI-Network

SIDER

OFFSIDES

STITCH

Decagon Network

19,089 proteins

715,612 protein–protein edges

18690 drug-protein edges

645 drugs

10,184 single side effects (represented as node features)

4,651,131 drug-drug edges

1317 side-effects (drug-drug edge type)

split in 6 CSV files

Mumps-Network
79 drugs
600 interactions
0,95% occurrence

Emesis-Network
621 drugs
23043 interactions
36,30% occurrence

Carbuncle-Network
86 drugs
509 interactions
0,80% occurrence

Incr. body temp.-Network
625 drugs
21806 interactions
34,35% occurrence

Coccydynia-Network
95 drugs
508 interactions
0,80% occurrence

Bleeding-Network
608 drugs
14143 interactions
22,28% occurrence

*Sources of the original data sets can be found in references

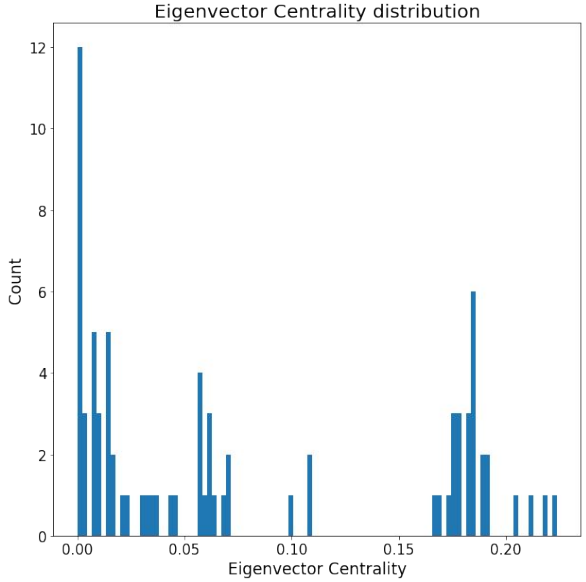
Mumps - Statistics

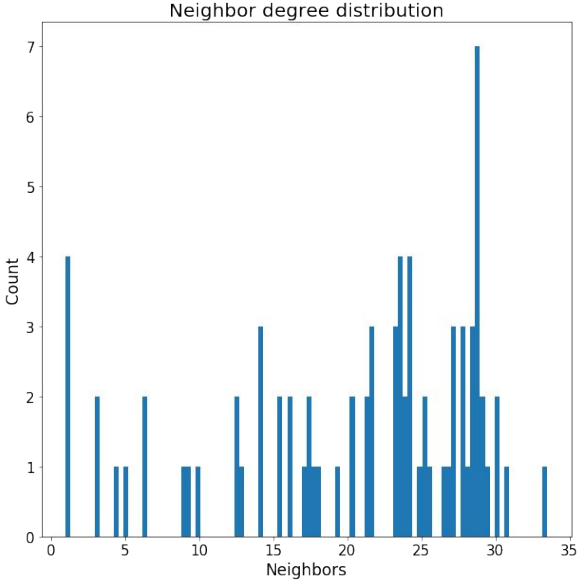
Average clustering = 0.64

Mean average neighbor degree = 20.41

Degree assortativity coefficient = 0.179

Diameter = ∞





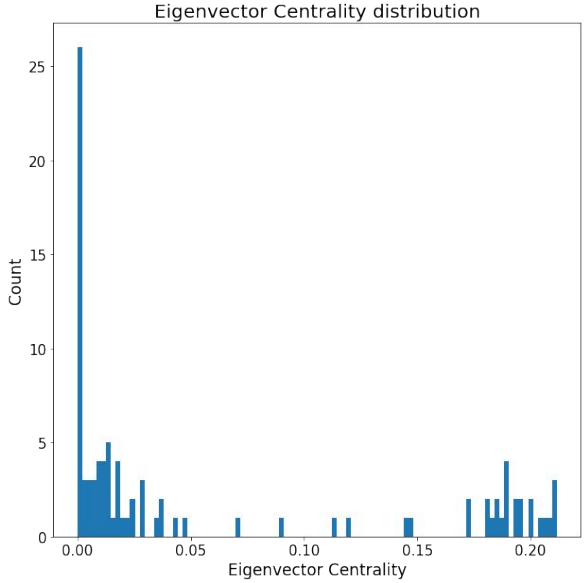
Coccydynia - Statistics

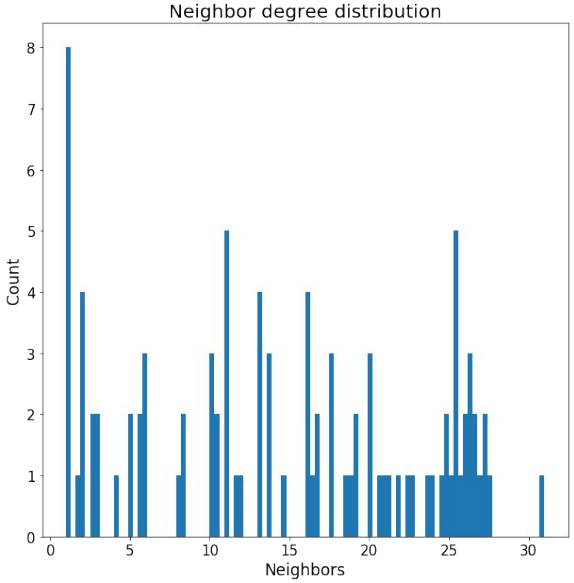
Average clustering = 0.63

Mean average neighbor degree = 14.64

Degree assortativity coefficient = -0.24

Diameter = ∞





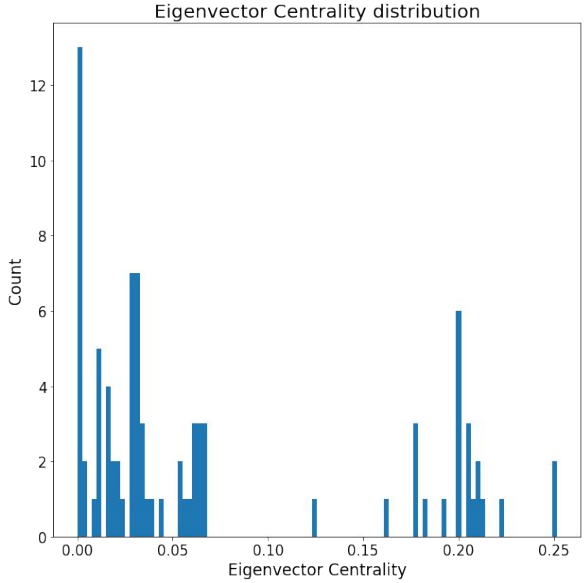
Carbuncle - Statistics

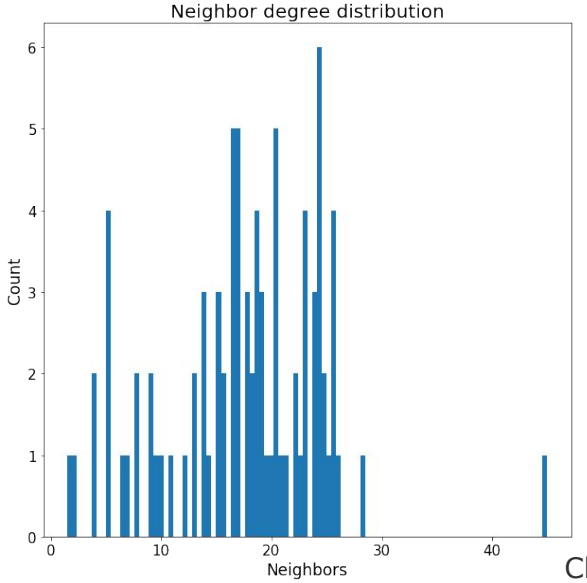
Average clustering = 0.74

Degree assortativity coefficient = 0.0676

Mean average neighbor degree = 17.53

Diameter= 7





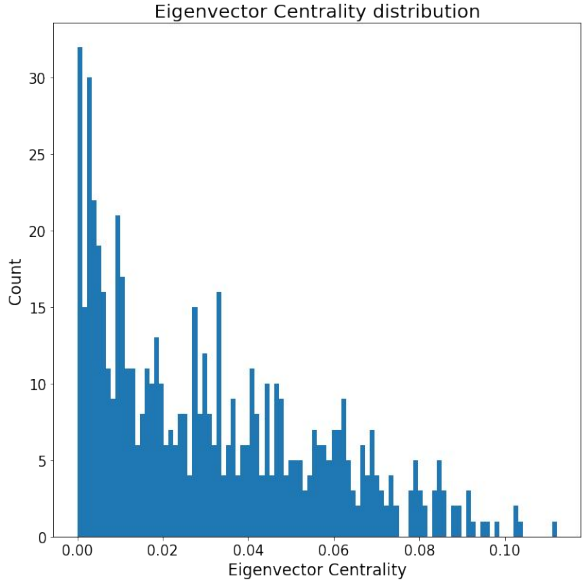
Emesis - Statistics

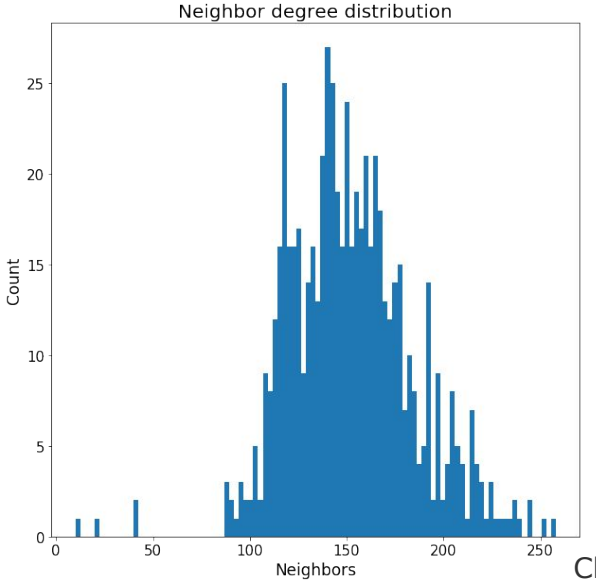
Average clustering = 0.43

Degree assortativity coefficient = -0.24

Mean average neighbor degree = 152.03

Diameter = 4





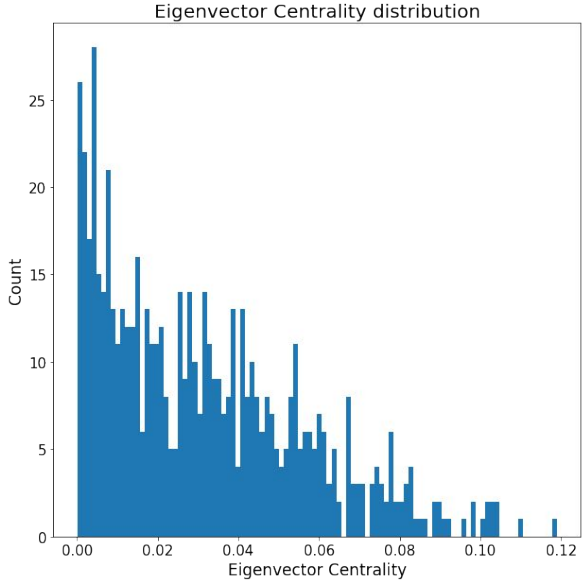
Increased body temperature - Statistics

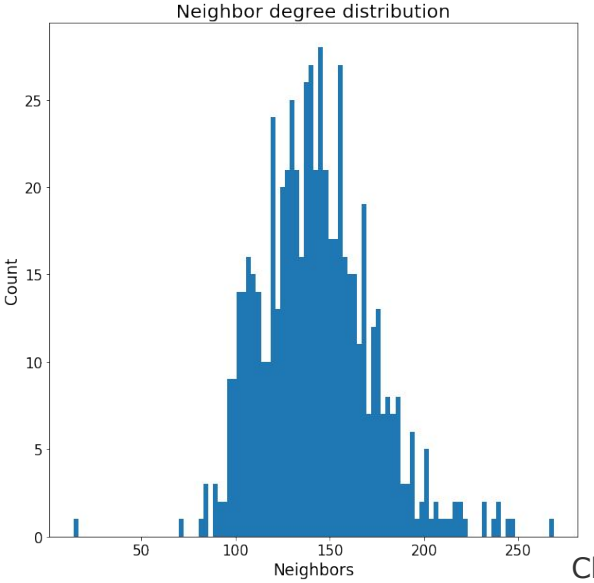
Average clustering = 0.40

Degree assortativity coefficient = -0.23

Mean average neighbor degree = 142.86

Diameter = 5





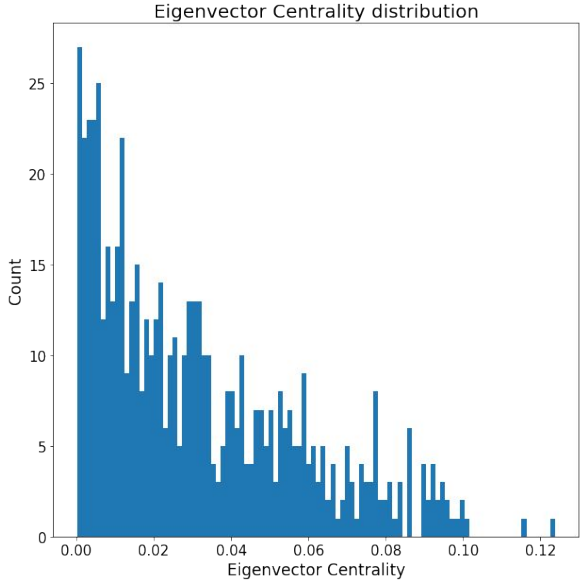
Bleeding - Statistics

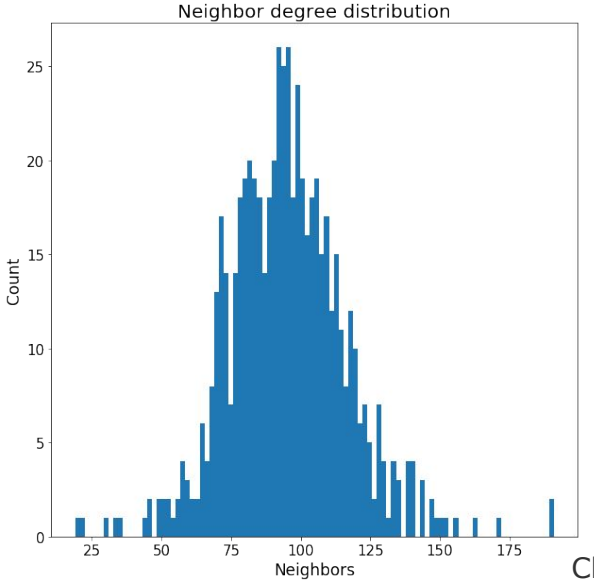
Average clustering = 0.29

Degree assortativity coefficient = -0.20

Mean average neighbor degree = 95.16

Diameter = 4





Our approach

Creation of drug-drug graph	Extraction of relevant side-effects	Creation of train/ test datasets	Application of algorithms
- Use preprocessed data from Decagon to build the drug-drug graph as list of sparse matrices	- Build networkx graph that only consists of edges of our 6 chosen side-effects - Convert graph to a dataframe and save as a csv	- Separately for each side-effect - Depending of algorithm to be tested apply different methods to create train/ test validation datasets - Main difference is format not content	- Train/ apply a total of 6 algorithms/ embeddings to our data - 10 experiment runs per algorithm and dataset - 360 runs with a runtime of about 106 hours

Metrics

For the evaluation of our models, we used 2 metrics: **AU-ROC** and **AUC-PR**.

AU-ROC:

- ROC curve shows true positive rate with respect to the false positive rate at all classification thresholds

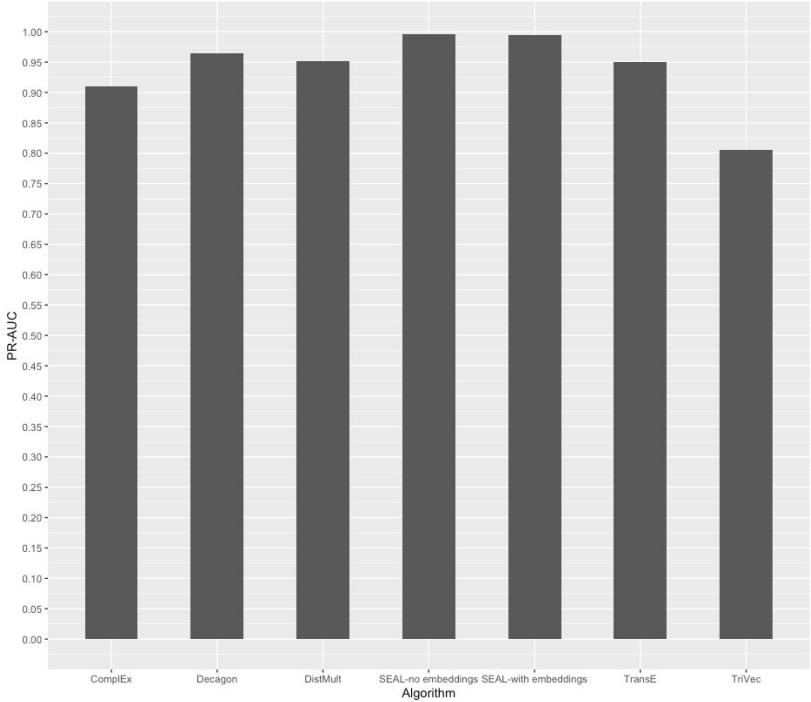
  → its area is equivalent to the probability of a randomly selected positive instance appearing above a randomly selected negative instance.

AUC-PR:

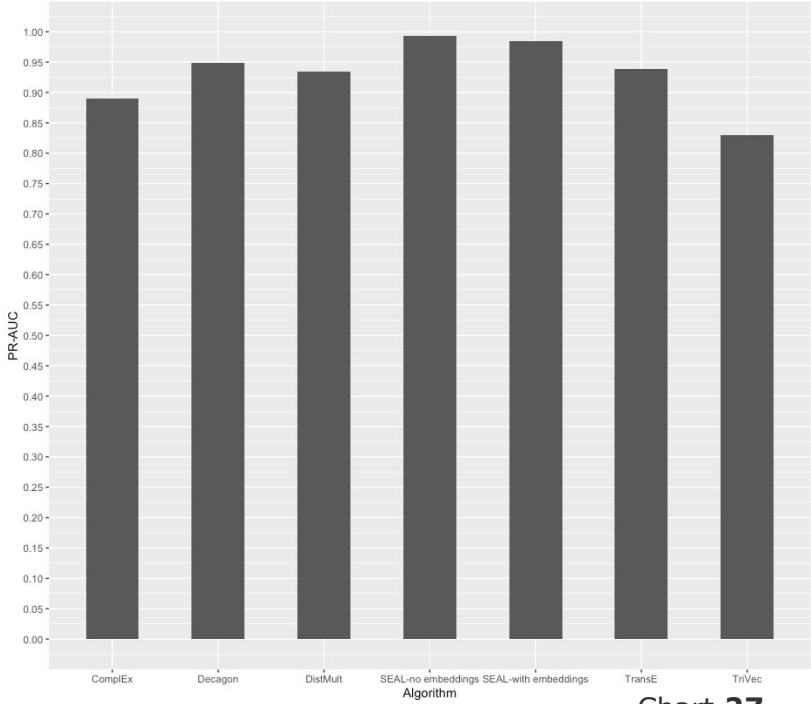
- The precision-recall curve shows precision with respect to recall at all classification thresholds.

Results: Mumps vs Carbuncle

Mumps - PR-AUC

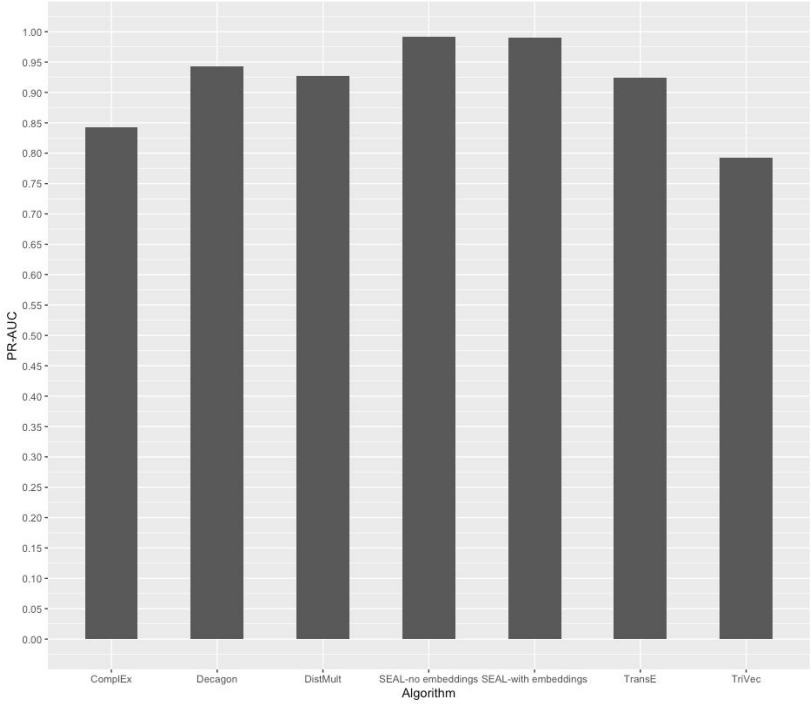


Carbuncle - PR-AUC

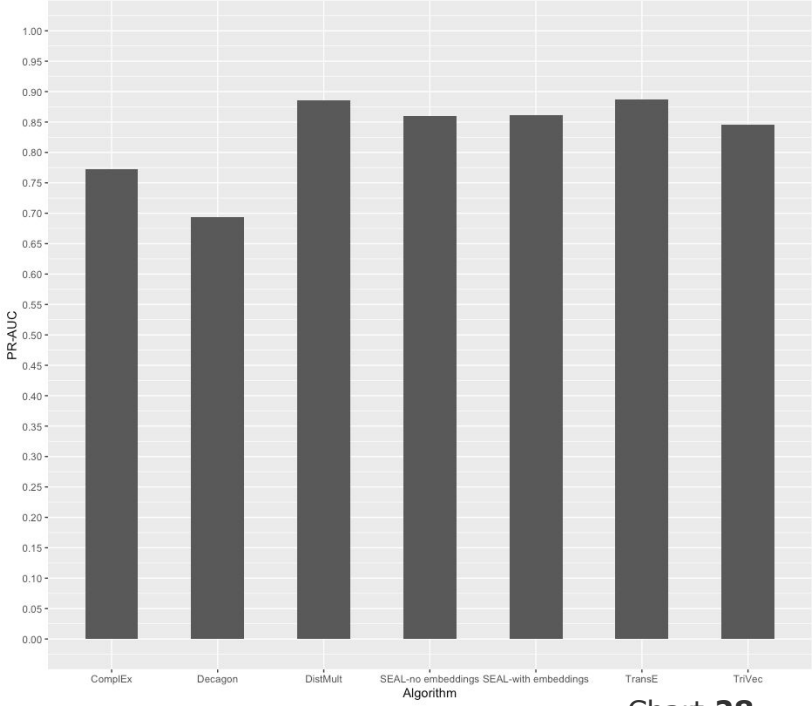


Results: Coccydynia vs Emesis

Coccydynia - PR-AUC

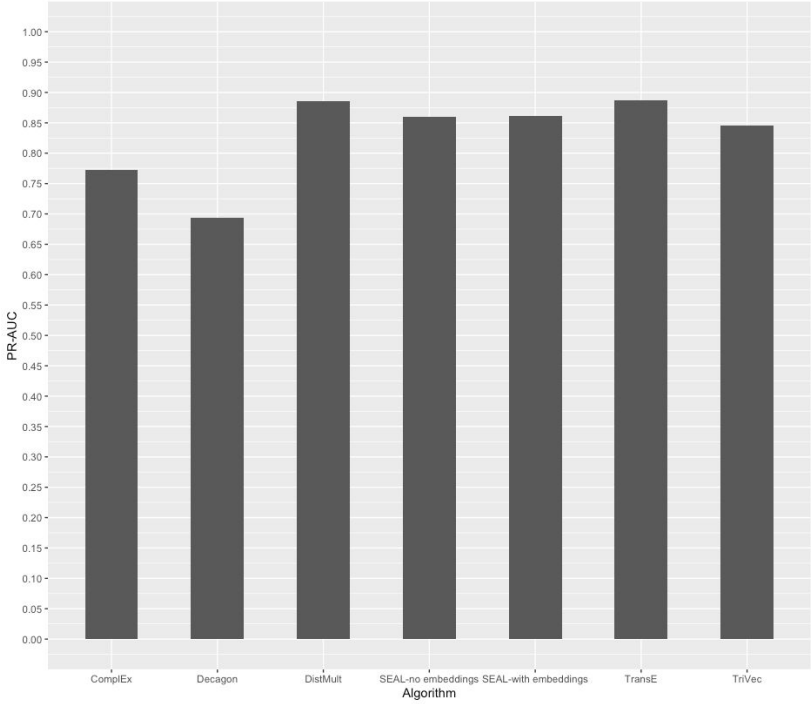


Emesis - PR-AUC

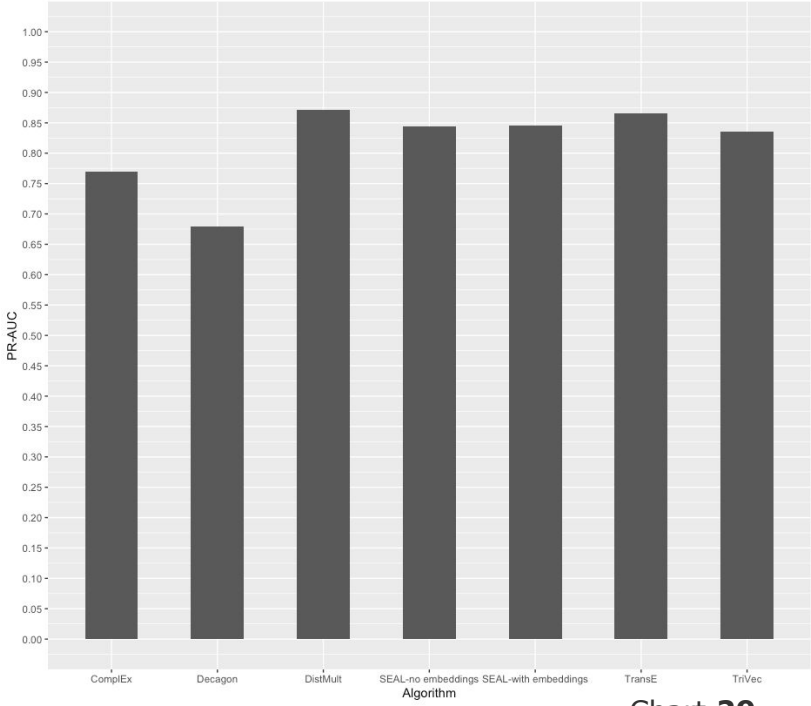


Results: Emesis vs Increased body temp.

Emesis - PR-AUC

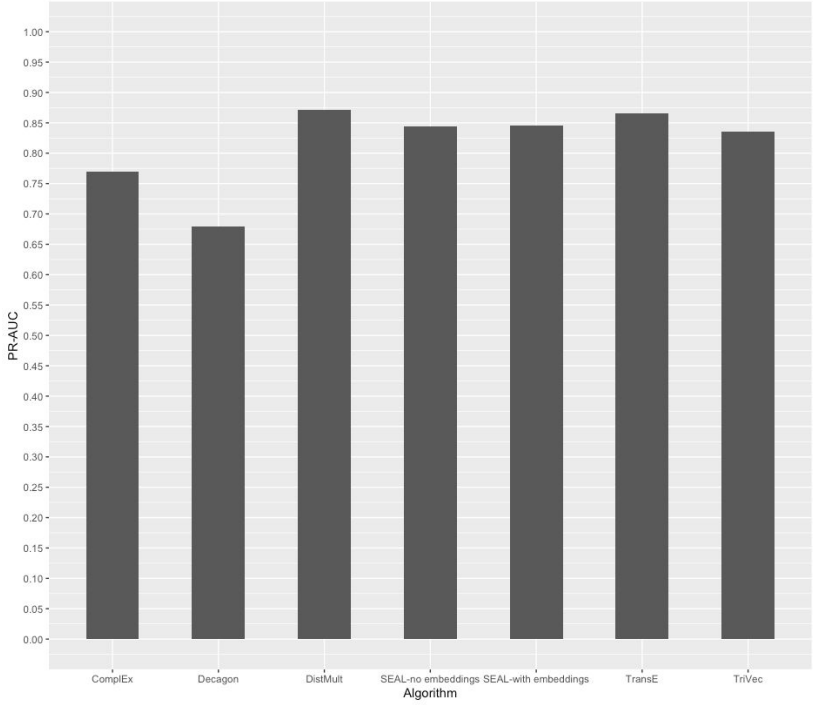


Increased body temp.- PR-AUC

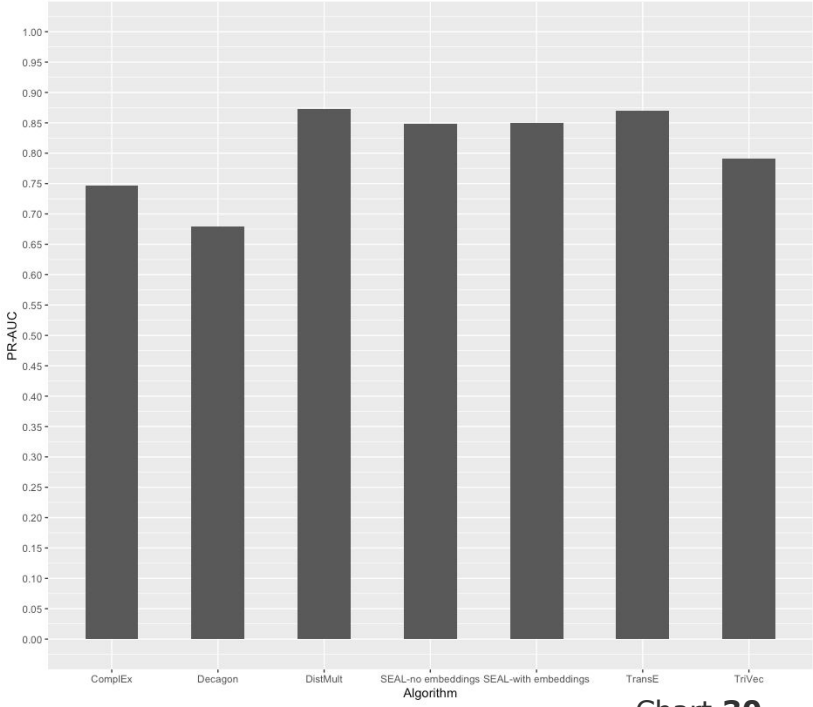


Results: Increased body temp. vs Bleeding

Increased body temp.- PR-AUC



Bleeding - PR-AUC



Discussion

Limitations

- Our approach is not suitable for rarer side effects or new drugs
- Potential overfitting for rarer side effects or very specific structures that are easy to learn
- Drug-Protein problem
- Unoptimized code

Drug-Protein and Protein-Protein interactions contain extremely important information and existing approaches that make use of them could be improved by using “better” parts

Future work

- Further DGCNN (and SEAL) ablation studies by removing GCLs/ 1-D convolutional or MaxPooling layers to evaluate their impact

- Evaluation if DGCNN is applicable to the multi-relational problem space
- Build an architecture based on SEAL/ DGCNN that incorporates single side-effects, drug-protein and protein-protein interactions
  → ablation studies to verify relevance of different parts

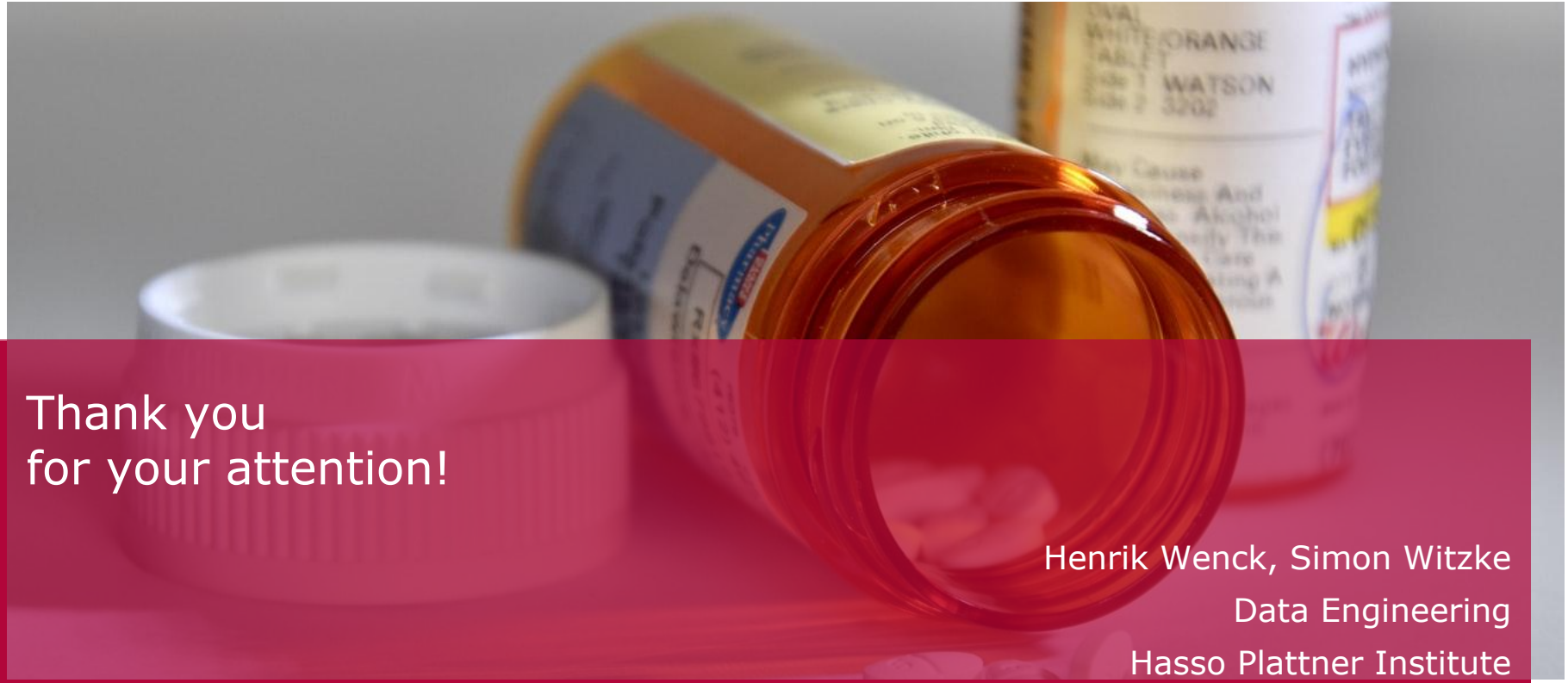
- Maybe include relational features in SEAL

- Modelling of side effects between more than two nodes (medical question)

References

1. Andreea Deac, Yu-Hsiang Huang, Petar Veličković, Pietro Liò, and JianTang. 2019. Drug-Drug Adverse Effect Prediction with Graph Co-Attention.arXiv:1905.00534 [stat.ML]
2. Giovanna Maria Dimitri and Pietro Lió. 2017. DrugClust: A machine learning approach for drugs side effects prediction.Computational Biology and Chemistry 68 (June 2017), 204–210. https://doi.org/10.1016/j.compbiolchem.2017.03.008
3. Elizabeth D. Kantor, Colin D. Rehm, Jennifer S. Haas, Andrew T. Chan, and Edward L. Giovannucci. 2015. Trends in Prescription Drug Use Among Adults in the United States From 1999-2012. JAMA 314, 17 (Nov. 2015), 1818. https://doi.org/10.1001/jama.2015.13766
4. Simon Kocbek, Primoz Kocbek, Andraz Stozer, Tina Zupanic, Tudor Groza, and Gregor Stiglic. 2018. Building interpretable models for polypharmacy prediction in older chronic patients based on drug prescription records.PeerJ6 (Oct. 2018),e5765. https://doi.org/10.7717/peerj.5765
5. Brandon Malone, Alberto García-Durán, and Mathias Niepert. 2018. Knowledge Graph Completion to Predict Polypharmacy Side Effects. In Lecture Notes in Computer Science. Springer International Publishing, 144–149. https://doi.org/10.1007/978-3-030-06016-9_14
6. S. Mizutani, E. Pauwels, V. Stoven, S. Goto, and Y. Yamanishi. 2012. Relating drug-protein interaction network with drug side effects. Bioinformatics 28, 18(Sept. 2012), i522–i528. https://doi.org/10.1093/bioinformatics/bts383
7. Vít Nováček and Sameh K. Mohamed. 2020. Predicting Polypharmacy Side-effects Using Knowledge Graph Embeddings. AMIA Jt Summits Transl Sci Proc 2020(2020), 449–458.
8. B. Reason, M. Terner, A. Moses McKeag, B. Tipper, and G. Webster. 2012. The impact of polypharmacy on the health of Canadian seniors.Family Practice 29, 4(Jan. 2012), 427–432. https://doi.org/10.1093/fampra/cmr124
9. Itay Shaked, Matthew A. Oberhardt, Nir Atias, Roded Sharan, and Eytan Ruppin.2016. Metabolic Network Prediction of Drug Side Effects.Cell Systems 2, 3 (March 2016), 209–213. https://doi.org/10.1016/j.cels.2016.03.001
10. Ruiyi Wang, Tong Li, Zhen Yang, and Haiyang Yu. 2020. Predicting Polypharmacy Side Effects Based on an Enhanced Domain Knowledge Graph. In Applied Informatics, Hector Florez and Sanjay Misra (Eds.). Springer International Publishing,Cham, 89–103.
11. Hao Xu, Shengqi Sang, and Haiping Lu. 2020. Tri-graph Information Propagation for Polypharmacy Side Effect Prediction. arXiv:2001.10516 [cs.LG]
12. Yoshihiro Yamanishi, Edouard Pauwels, and Masaaki Kotera. 2012. Drug Side-Effect Prediction Based on the Integration of Chemical and Biological Spaces.Journal of Chemical Information and Modeling 52, 12 (Dec. 2012), 3284–3292.https://doi.org/10.1021/ci2005548
13. Wen Zhang, Yanlin Chen, Shikui Tu, Feng Liu, and Qianlong Qu. 2016. Drug side effect prediction through linear neighborhoods and multiple data source integration. In 2016 IEEE International Conference on Bioinformatics and Biomedicine(BIBM). IEEE. https://doi.org/10.1109/bibm.2016.7822555
14. Xian Zhao, Lei Chen, and Jing Lu. 2018. A similarity-based method for prediction of drug side effects with heterogeneous information.Mathematical Biosciences 306 (2018), 136 – 144. https://doi.org/10.1016/j.mbs.2018.09.010
15. Marinka Zitnik, Monica Agrawal, and Jure Leskovec. 2018. Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics 34, 13 (June 2018), i457–i466. https://doi.org/10.1093/bioinformatics/bty2943
16. Muhan Zhang and Yixin Chen. 2018. Link prediction based on graph neural networks. In Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18). Curran Associates Inc., Red Hook, NY, USA, 5171–5181.
17. Zhang, M., Cui, Z., Neumann, M., & Chen, Y. (2018). An End-to-End Deep Learning Architecture for Graph Classification. AAAI.
18. Menche,J. et al. (2015) Uncovering disease-disease relationships through the incomplete interactome. Science, 347, 1257601.
19. Chatr-Aryamontri,A. et al. (2015) The BioGRID interaction database: 2015 update. Nucleic Acids Res., 43, D470–D478.
20. Szklarczyk,D. et al. (2016) STITCH 5: augmenting protein–chemical interaction networks with tissue and affinity data. Nucleic Acids Res., 44, D380–D384.
21. Szklarczyk,D. et al. (2017) The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic Acids Res., 45, D362–D368.
22. Kuhn,M. et al. (2016) The SIDER database of drugs and side effects. Nucleic Acids Res., 44, D1075–D1079.
23. Tatonetti,N.P. et al. (2012) Data-driven prediction of drug effects and interactions. Sci. Transl. Med., 4, 125ra31.
24. Rossi, A., Firmani, D., Matinata, A., Merialdo, P., & Barbosa, D. (2020). Knowledge Graph Embedding for Link Prediction: A Comparative Analysis. ArXiv, abs/2002.00819.

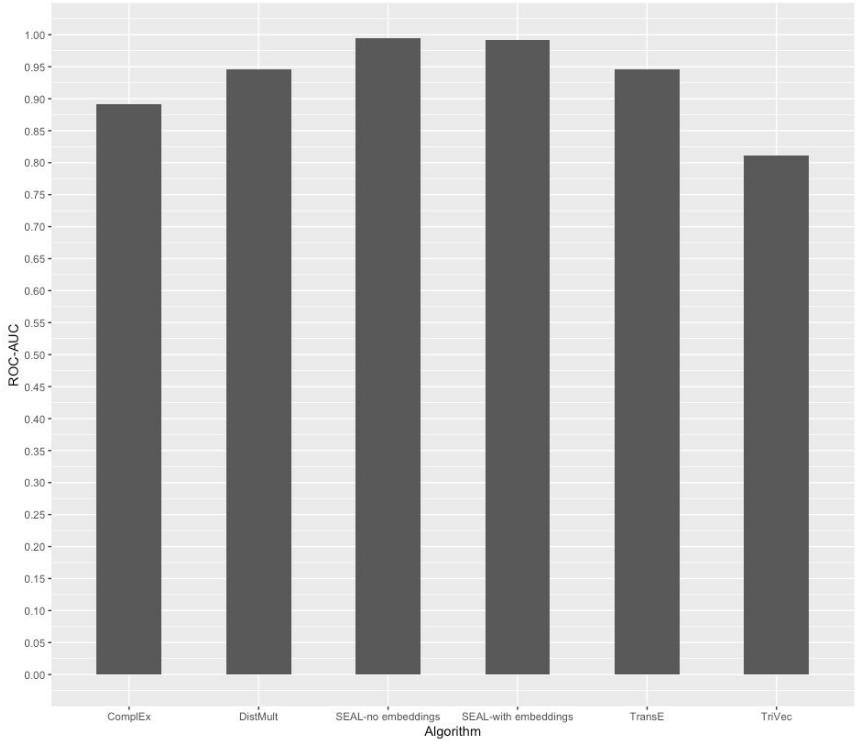
Hasso Plattner Institut
Digital Engineering · Universität Potsdam



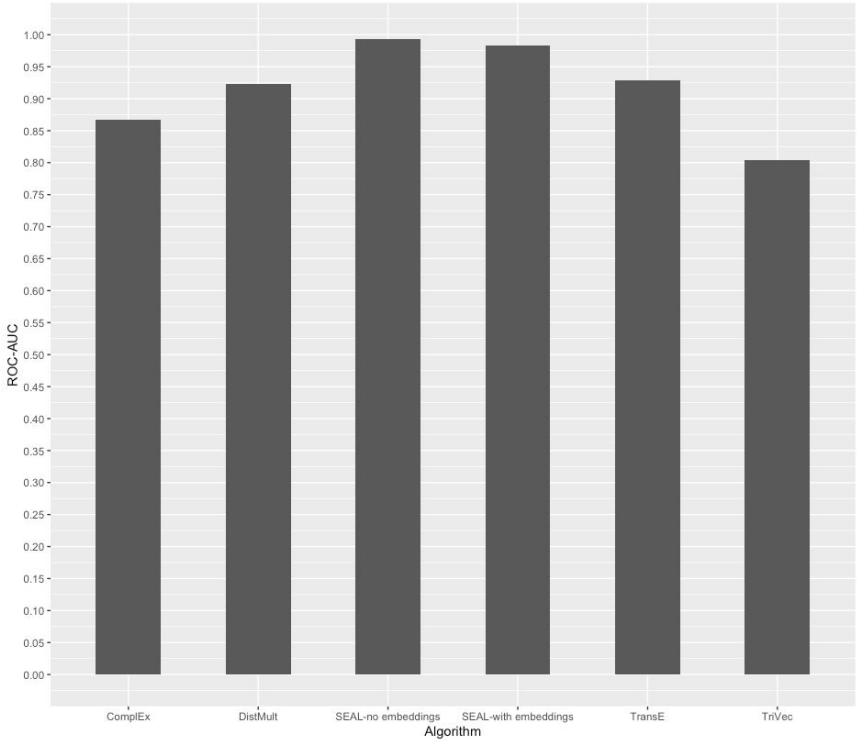
Thank you
for your attention!

Henrik Wenck, Simon Witzke
Data Engineering
Hasso Plattner Institute

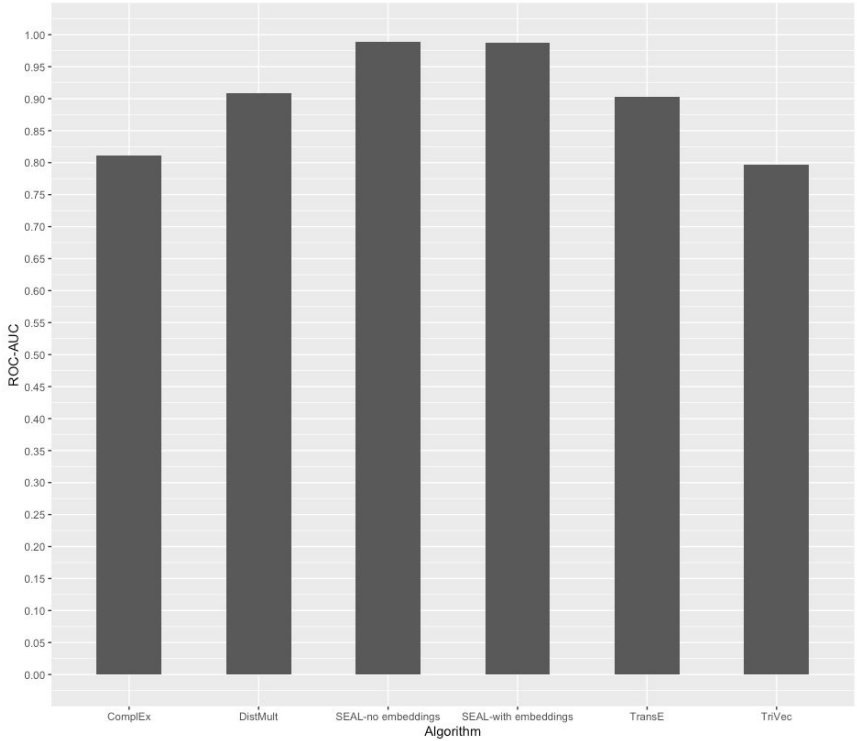
Mumps


ROC-AUC
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
ComplEx
DistMult
SEAL-no embeddings
SEAL-with embeddings
TransE
TriVec
Algorithm

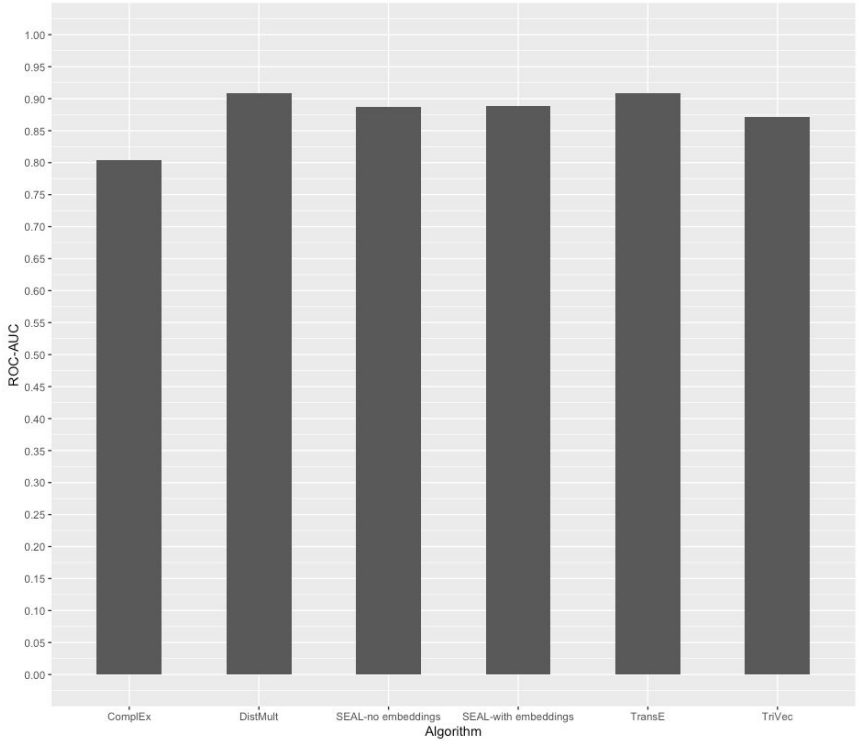
Carbuncle


ROC-AUC
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
ComplEx
DistMult
SEAL-no embeddings
SEAL-with embeddings
TransE
TriVec
Algorithm

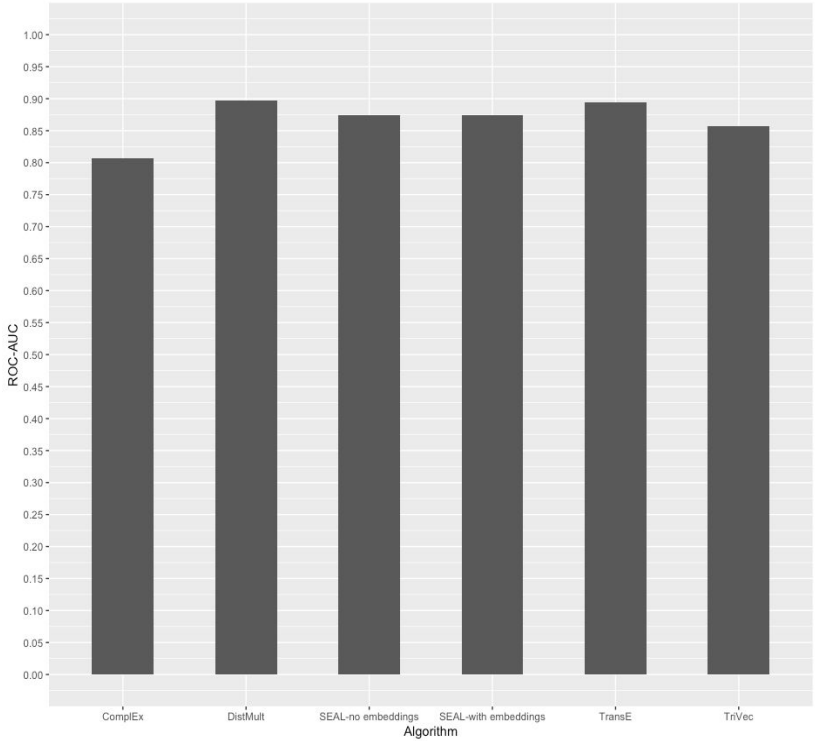
Coccydynia


ROC-AUC
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
ComplEx
DistMult
SEAL-no embeddings
SEAL-with embeddings
TransE
TriVec
Algorithm

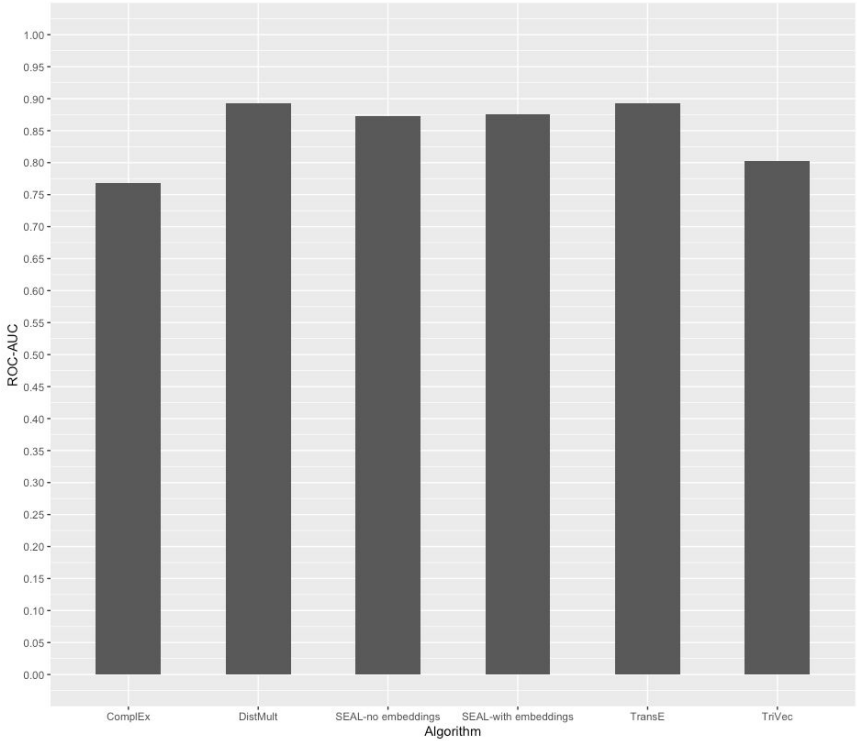
Emesis


ROC-AUC
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
ComplEx
DistMult
SEAL-no embeddings
SEAL-with embeddings
TransE
TriVec
Algorithm

Increased body temp.



Bleeding


ROC-AUC
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
ComplEx
DistMult
SEAL-no embeddings
SEAL-with embeddings
TransE
TriVec
Algorithm

Background

TriVec (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7233093/)

TransE (https://proceedings.neurips.cc/paper/2013/file/1cecc7a77928ca8133fa24680a88d2f9-Paper.pdf)

DistMult (https://arxiv.org/pdf/1412.6575.pdf)

ComplEx (https://arxiv.org/pdf/1606.06357.pdf - glaube ich)

how are they calculated

Weisfeiler-Lehmann Neural Machine ?

WLNM has several drawbacks. Firstly, WLNM trains a fully-connected neural network on the subgraphs' adjacency matrices.

-drawbacks am Ende ? (siehe 2 am Ende)

Our data*

PPI-Network
19,055 Proteins
719,402 interactions

STITCH
8,934 proteins
519,022 chemicals (drugs)
8,083,600 drug-protein interactions

SIDER
1556 drugs
5,868 side-effects
286,399 drug-side-effect associations

OFFSIDES
1332 drugs
10,097 side-effects
487,530 drug-side-effect associations

TWOSIDES
63,473 drug combinations
1,318 side-effects
4,651,131 combination-side-effect associations

Decagon preprocessing

Decagon Network
19,089 proteins
715,612 protein–protein edges
18690 drug-protein edges
645 drugs
10,184 single side effects (represented as node features)
4,651,131 drug-drug edges
1317 side-effects (drug-drug edge type)

split in 6 CSV files

Mumps-Network
79 drugs
600 interactions
0,94528% occurrence

Carbuncle-Network
86 drugs
509 interactions
0,80192% occurrence

Coccydynia-Network
95 drugs
508 interactions
0,80034% occurrence

Emesis-Network
621 drugs
23043 interactions
36,30363% occurrence

Incrsd. body temp.-Network
625 drugs
21806 interactions
34,35477% occurrence

Bleeding-Network
608 drugs
14143 interactions
22,28192% occurrence

Chart **43**

*Sources of the original data sets can be found in references

Our data

@Chris would it be helpful to show graph and node level metrics and distributions again, such as:

- Average node degree
- average clustering
- node connectivity
- Assortativity Coefficient
- Eigenvector centrality
- Page Rank

for each of our six side-effect graphs (or for a selected 2)

Results and Evaluation

Rest Tonspur
Vielleicht macht es Sinn für einen Algorithmus die Zeiten auch unter den Verschiedenen side effects zu vergleichen

@Chris We haven't conducted a concrete ablation study (we ran SEAL with and without embeddings which is probably what comes closest)

Results aus decagon und malone paper im Vergleich diskutieren

Summary statistics for metrics
Wie bauen wir das konkret auf. Also mit der sensitivity und ablation sache