Heterogeneous Subgraph Features for Information Networks

Andreas Spitz, Diego Costa, Kai Chen, Jan Greulich, Johanna Geiß, Stefan Wiesberg, and Michael Gertz

Heidelberg University, Im Neuenheimer Feld 205, 69120 Heidelberg, Germany

Motivation

For predictive analyses of heterogeneous networks with diversely **labelled nodes**, the extraction of node features is of central importance. **Feature engineering** can be difficult or impossible since external metadata is often missing. As an alternative, node embeddings incorporate the **topological structure** of a network into node features, but require extensive parameter tuning, encode only limited neighbourhood information, and do not include node labels.

Idea: Given a heterogeneous network, use the census of labelled subgraphs around a node as its features.

Drawbacks of Previous Features

Classic features

- Require domain knowledge
- Time consuming to engineer
- Metadata may not be available

Node embeddings [1], [2], [3]

- Compact vector representations
- Sample through random walks
- Numerous parameters require tuning

Network motifs [4]

- Use global subgraph counts
- Do not represent node neighbourhoods
- Importance calculated from null model

Evaluation Data

Movie network (IMDB)

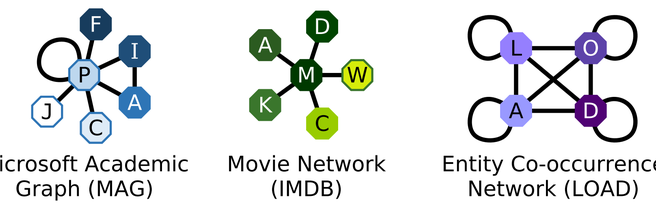
- Star-shaped structure around movies
- Low edge density

Entity cooccurrence network (LOAD)

- Strongly connected structure
- High edge density

Scientific publication network (MAG)

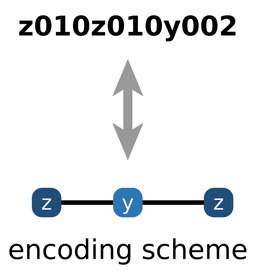
- Intermediate structure
- Papers form the core component



Heterogeneous Subgraph Features

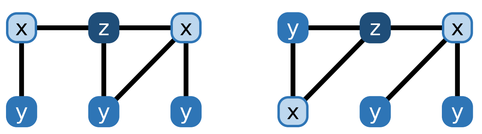
Extraction

- Use DFS exploration
- Represent subgraphs by characteristic string
- Replace isomorphism tests by hashing



Limitations

- Use only small subgraphs: number of nodes or edges ≈ 5
- Avoid exploration beyond hubs
- Colliding encodings: possible but rare

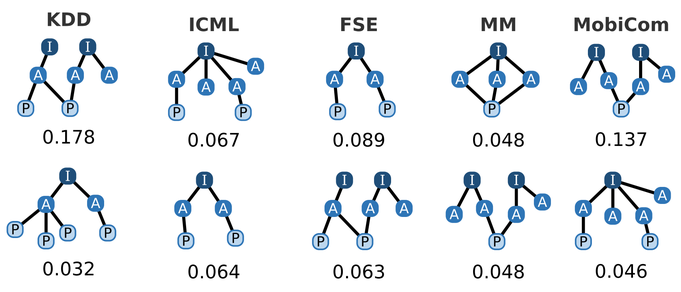


z210y101y100y100x021x011

Feature Importance

Example of interpretability

The most important subgraphs for predicting institution ranks in the MAG data:



Subgraph features

- Can be decoded (unlike embeddings)
- Are interpretable structures
- Highlight relations between labels
- Provide insights into the relevant substructures of the network

Evaluation: Rank Prediction

Task definition

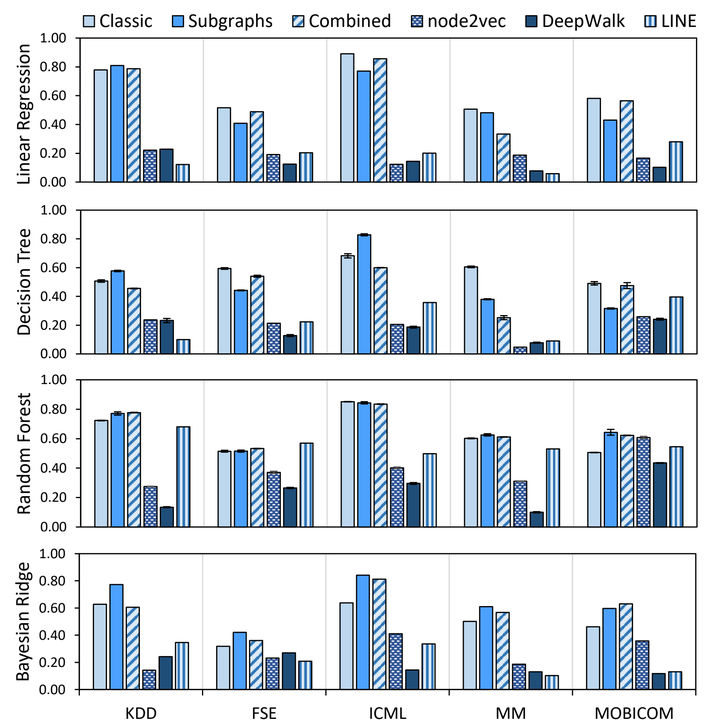
Given the publications of scientific institutions in previous years, predict a ranking of institutions by their future publications.

Data

Subset of the MAG data for 5 major conferences. Classic features are derived from metadata, the ACM digital library, and linguistic features from the abstracts.

Evaluation metric

Normalized discounted cumulative gain (NDCG) for the 20 top institutions.



	LinReg	DecTr	RanFo	BayReg
classic	0.65	0.58	0.64	0.51
subgraph	0.58	0.51	**0.68**	0.65
combined	0.62	0.46	**0.68**	0.60
node2vec	0.18	0.19	0.39	0.27
DeepWalk	0.14	0.17	0.25	0.18
LINE	0.17	0.23	0.56	0.23

Evaluation: Label Prediction

Task definition

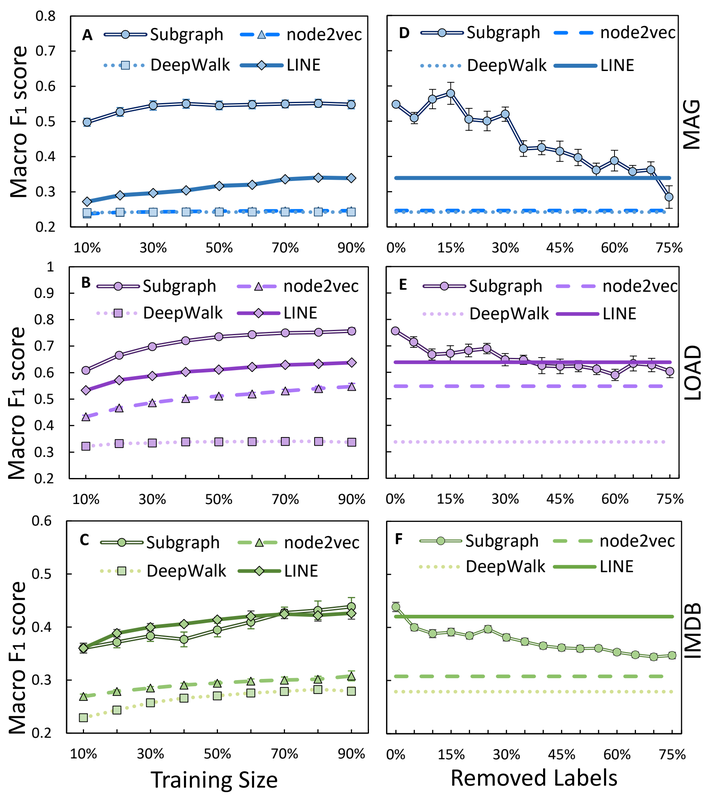
Given a heterogeneous network in which some nodes are missing labels, predict the missing label information.

Data

All three evaluation data sets. Only structural features are used since classic features are largely unavailable.

Evaluation metric

Macro F_1 score of predicted labels.



Outlook

Open research questions

- Modelling directed edges
- Modelling heterogeneous edges
- Sampling strategies

References

[1] B. Perozzi, R. Al-Rfou, and S. Skiena. **DeepWalk: Online Learning of Social Representations**. *KDD'14*.

[2] A. Grover and J. Leskovec. **node2vec: Scalable Feature Learning for Networks**. *KDD'16*.

[3] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei. **LINE: Large-scale Information Network Embedding**. *WWW'15*.

[4] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, U. Alon. **Network Motifs: Simple Building Blocks of Complex Networks**. *Science, 298(5594), 2002*.

Contact Information:

Andreas Spitz

spitz@informatik.uni-heidelberg.de

http://dbs.ifi.uni-heidelberg.de/

This work was presented at the 1st GRADES-NDA Workshop at SIGMOD. June 10 – 15, 2018, Houston, USA.