Heterogeneous Subgraph Features for Information Networks

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

June 10, 2018 — GRADES-NDA, Houston, Texas, USA

Heidelberg University, Germany
Database Systems Research Group

Learning and Predicting in Heterogeneous Networks

Many information networks are heterogeneous

- Scientific publication networks
- Knowledge bases
- Metabolic networks
- ...

Learning and Predicting in Heterogeneous Networks

Many information networks are heterogeneous

- Scientific publication networks
- Knowledge bases
- Metabolic networks
- ...

How do you learn in heterogeneous networks?

Learning and Predicting in Heterogeneous Networks

Many information networks are heterogeneous

- Scientific publication networks
- Knowledge bases
- Metabolic networks
- ...

How do you learn in heterogeneous networks?

- With features, of course

Learning and Predicting in Heterogeneous Networks

Many information networks are heterogeneous

- Scientific publication networks
- Knowledge bases
- Metabolic networks
- ...

How do you learn in heterogeneous networks?

- With features, of course
- But how do you get the features?

Problems of Established Feature Extraction Approaches

Classic features:

- Require domain knowledge
- Are time-consuming to engineer
- Require metadata that may not be available

Problems of Established Feature Extraction Approaches

Classic features:

- Require domain knowledge
- Are time-consuming to engineer
- Require metadata that may not be available

Neural node embeddings:

- Sample neighbourhoods through random walks
- Require extensive parameter tuning

Problems of Established Feature Extraction Approaches

Classic features:

- Require domain knowledge
- Are time-consuming to engineer
- Require metadata that may not be available

Neural node embeddings:

- Sample neighbourhoods through random walks
- Require extensive parameter tuning

Alternative idea: use labeled subgraph counts as features

Heterogeneous Subgraph Features

Motivation: Heterogeneous Subgraph Features

Labeled subgraphs around a node:

- Encode neighbourhood information
- Are extremely diverse in heterogeneous networks

Motivation: Heterogeneous Subgraph Features

Labeled subgraphs around a node:

- Encode neighbourhood information
- Are extremely diverse in heterogeneous networks

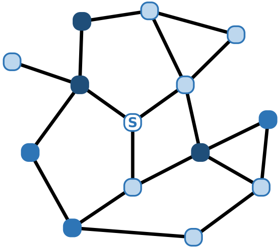
Conjecture:

The subgraph neighbourhood of a node is representative of its function and label.

Motivation: Heterogeneous Subgraph Features

Labeled subgraphs around a node:

- Encode neighbourhood information
- Are extremely diverse in heterogeneous networks



Conjecture:

The subgraph neighbourhood of a node is representative of its function and label.

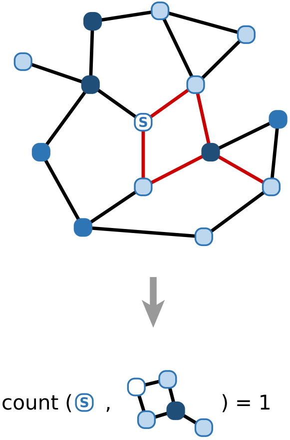
Motivation: Heterogeneous Subgraph Features

Labeled subgraphs around a node:

- Encode neighbourhood information
- Are extremely diverse in heterogeneous networks

Conjecture:

The subgraph neighbourhood of a node is representative of its function and label.



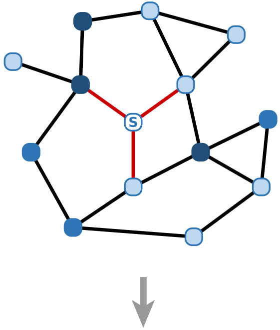
Motivation: Heterogeneous Subgraph Features

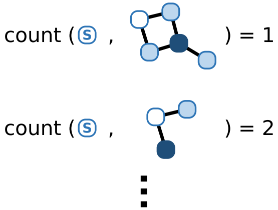
Labeled subgraphs around a node:

- Encode neighbourhood information
- Are extremely diverse in heterogeneous networks

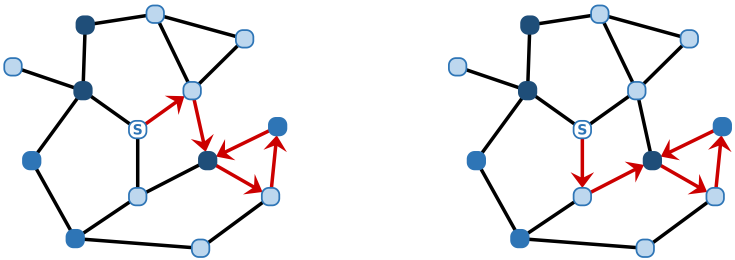
Conjecture:

The subgraph neighbourhood of a node is representative of its function and label.





Isomorphism of Subgraphs



Problem: depending on the iteration order, the nodes of structurally identical subgraphs may be visited in different order.

Heterogeneous Subgraph Encoding

Core approach:

- Explore the local neighbourhood around each node
- Represent subgraphs by their characteristic string
- Count subgraphs by hashing the characteristic string
- Use the counts of subgraphs as node features

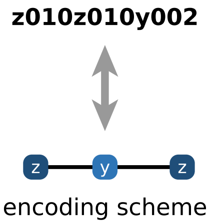
Heterogeneous Subgraph Encoding

Core approach:

- Explore the local neighbourhood around each node
- Represent subgraphs by their characteristic string
- Count subgraphs by hashing the characteristic string
- Use the counts of subgraphs as node features

Characteristic string construction:

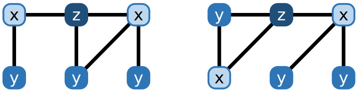
- Encode each node as a block
- Blocks start with the node label
- Subsequent entries denote neighbours of all given labels
- Blocks are sorted lexicographically



Encoding Collisions

Heterogeneous degree sequences:

- Are a **pseudo**-canonical encoding
- May result in colliding encodings

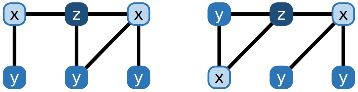


z210y101y100y100x021x011

Encoding Collisions

Heterogeneous degree sequences:

- Are a **pseudo**-canonical encoding
- May result in colliding encodings



z210y101y100y100x021x011

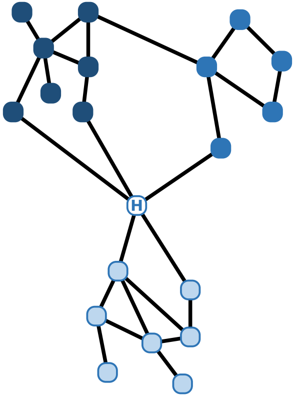
Encoding collisions:

- Can only be enumerated (no closed formula)
- Depend on the network structure and the labels
- Have negligible frequency in practice

Heuristic for Hub Mitigation

Real-world networks have:

- Skewed degree distributions
- Highly connected nodes (hubs)



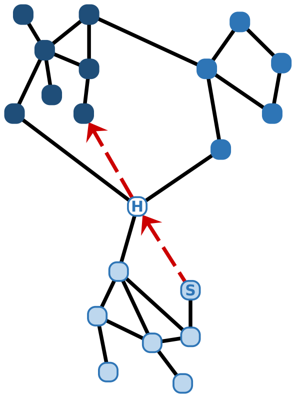
Heuristic for Hub Mitigation

Real-world networks have:

- Skewed degree distributions
- Highly connected nodes (hubs)

Due to hubs:

- Feature extraction time is strongly increased
- Random walks retrieve non-local information



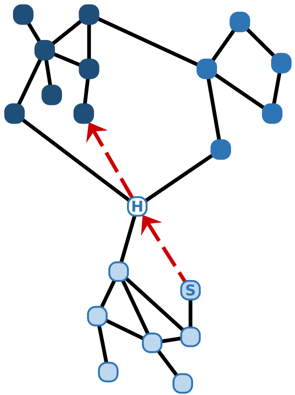
Heuristic for Hub Mitigation

Real-world networks have:

- Skewed degree distributions
- Highly connected nodes (hubs)

Due to hubs:

- Feature extraction time is strongly increased
- Random walks retrieve non-local information



Intuition: Do not explore beyond nodes with degree $> d_{max}$.

Evaluation: Label Prediction

Label Prediction: Task Definition

Given:

- Heterogeneous network
- Some nodes with missing labels

Predict:

- Missing node labels

Label Prediction: Task Definition

Given:

- Heterogeneous network
- Some nodes with missing labels

Predict:

- Missing node labels

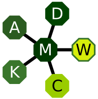
Formal approach:

- Model as a classification task using logistic regression
- Evaluate with F_1-score

Label Prediction: Data Sets

Movie network (IMDB):

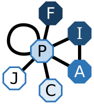
- Star-shaped structure around movies
- Low edge density



Movie Network
(IMDB)

Scientific publication network (MAG):

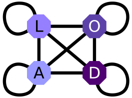
- Intermediate structure
- Papers form the core component



Microsoft Academic
Graph (MAG)

Entity cooccurrence network (LOAD):

- Cooccurrences of named entities in text
- Strongly connected structure
- High edge density



Entity Co-occurrence
Network (LOAD)

Feature Engineering and Extraction

Subgraph features:

- Maximum number of edges: 5
- No exploration beyond 10% of highest degree nodes
- Masked starting node label

Embedded features:

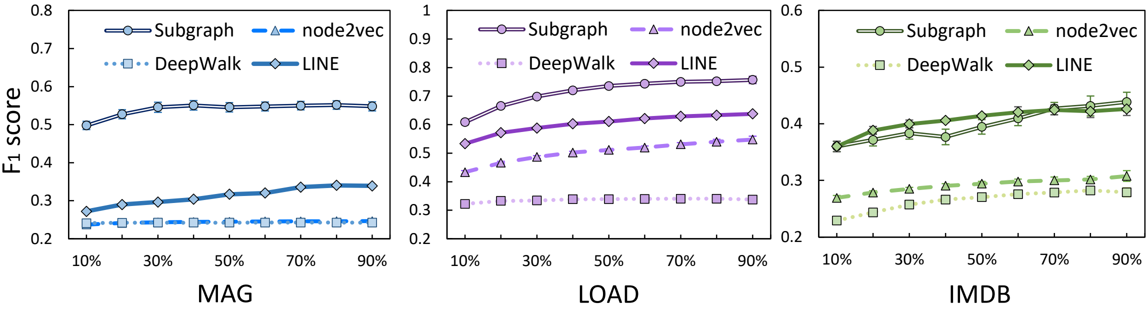
- DeepWalk
- LINE
- node2vec

Extraction Runtime Estimation (seconds per node)

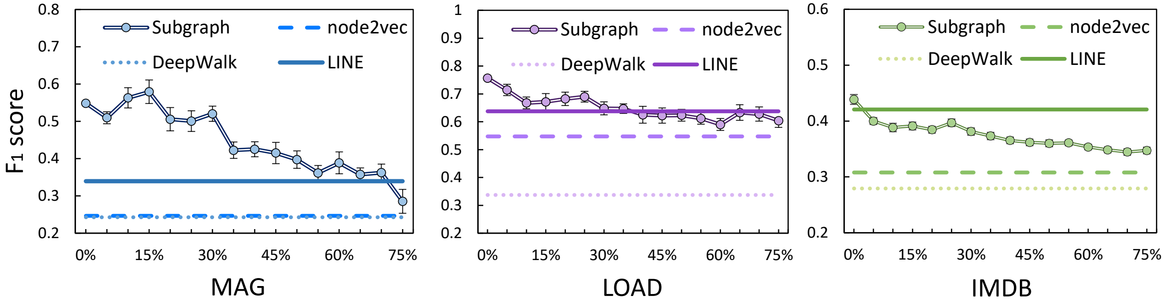
	subgraph features					node2vec	DeepWalk	LINE
	mean	75%	90%	95%	max		mean	
LOAD	32.1	19.6	29.7	53.0	1046	0.19	0.11	0.66
IMDB	2.6	1.7	3.0	6.7	47	0.01	0.01	0.64
MAG	25.2	10.4	11.0	19.5	2493	0.02	0.01	0.49

Percentages denote nodes for which the extraction finished in at most the shown time.

Evaluation Results (Training Size)



Evaluation Results (Missing Labels)



Evaluation: Institution Ranking

Institution Ranking: Task Definition

Given:

- Scientific publication network
- A range of years
- A set of conferences

Predict ranking of institutions:

- For upcoming conferences
- By accepted papers
- For the next conference

KDDCup 2016. `https://kddcup2016.azurewebsites.net`

Institution Ranking: Task Definition

Given:

- Scientific publication network
- A range of years
- A set of conferences

Predict ranking of institutions:

- For upcoming conferences
- By accepted papers
- For the next conference

Formal approach:

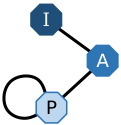
- Model as a regression task for the institution relevance score
- Evaluate with normalized discounted cumulative gain (NDCG20)

KDDCup 2016. `https://kddcup2016.azurewebsites.net`

Institution Ranking: Data Set

Subset of the Microsoft Academic Graph:

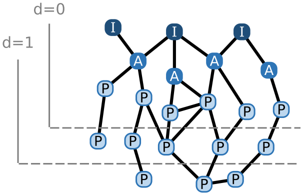
- Institutions I
- Authors A
- Papers P
- Publication data from 2011 - 2016



Microsoft Academic Graph
(MAG)

Data preparation:

- Focus on 5 conferences
  KDD, FSE, ICML, MM, MOBICOM
- Use citations to a depth of 3



Feature Types and Extraction

Classic features (manually engineered):

- Previous relevance scores, publication counts, etc. (8)
- Linguistic features (32)

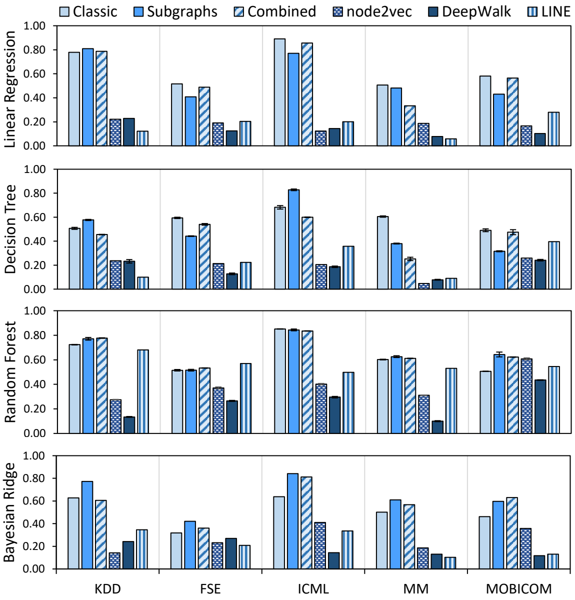
Subgraph features:

- Maximum number of edges: 5
- No maximum degree exploration limit

Embedded features:

- DeepWalk
- LINE
- node2vec

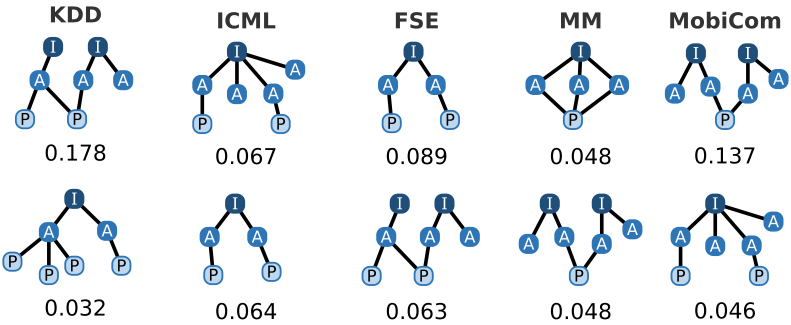
NDCG Scores for Institution Ranking



Average NDCG Scores for Institution Ranking

	LinRegr	DecTree	RanForest	BayRidge
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

Feature Importance Analysis (Random Forest)



Summary & Resources

Summary

Heterogeneous subgraph features:

- Extracted by local exploration and enumeration
- Avoid isomorphism test by encoding degree sequences

Summary

Heterogeneous subgraph features:

- Extracted by local exploration and enumeration
- Avoid isomorphism test by encoding degree sequences

In comparison to classic features:

- Similar performance
- Require no domain knowledge for extraction
- No engineering process necessary

Summary

Heterogeneous subgraph features:

- Extracted by local exploration and enumeration
- Avoid isomorphism test by encoding degree sequences

In comparison to classic features:

- Similar performance
- Require no domain knowledge for extraction
- No engineering process necessary

In comparison to embedded features:

- Better predictive performance
- Longer extraction time

Resources

The implementation is available online:

- C++ (core extraction routines)
- Python (wrapper)

`https://dbs.ifi.uni-heidelberg.de/resources/hsgf/`

Resources

The implementation is available online:

- C++ (core extraction routines)
- Python (wrapper)

`https://dbs.ifi.uni-heidelberg.de/resources/hsgf/`