EAA e-Conference on
Data Science & Data Ethics

16 May 2023

Less is More: Including Network Features for Insurance Fraud Detection – A Case Study for a Belgian Insurance Company

Bruno Deprez

KU Leuven & UAntwerp

Agenda

1. INTRO TO NETWORK ANALYSIS
2. THE METHODS USED
3. DATA AND PERFORMANCE
4. THE RESULTS

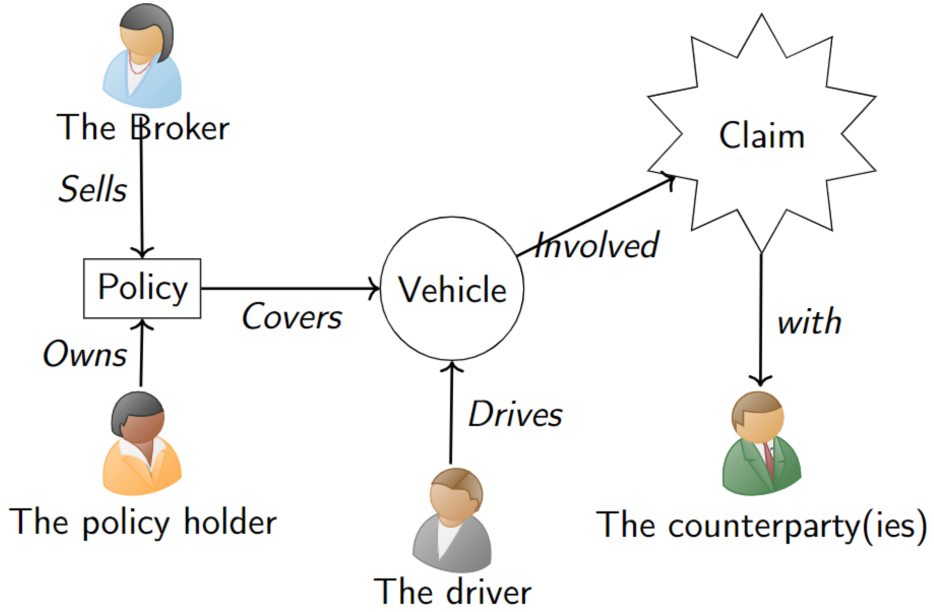
INTRO TO NETWORK ANALYSIS

01

THE NEED FOR NETWORK ANALYSIS

- Networks are a natural representation of many real-world phenomena

- Many actuarial problems can be formulated with networks, with applications in:
  - Pricing
  - Reserving
  - Fraud detection



THE NEED FOR NETWORK ANALYSIS IN FRAUD DETECTION

- Fraud detection done in two ways
  - Rule-based
  - Via machine learning

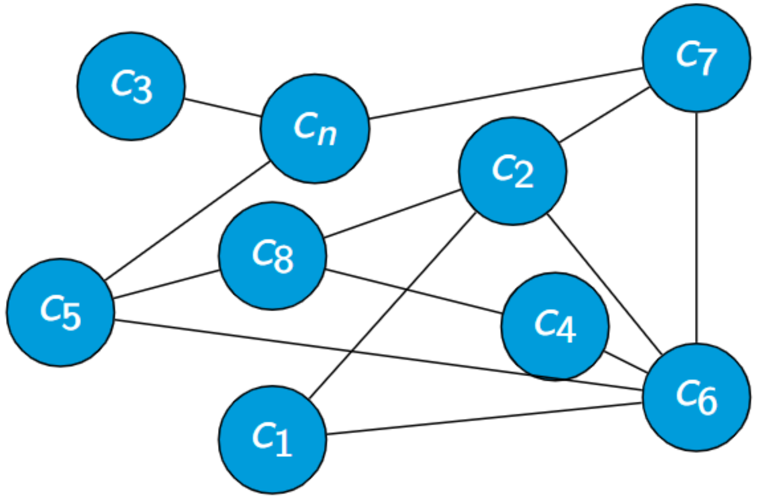
- Fraudsters adapt to this
  - Cover their tracks
  - New fraud methods
  - Avoid business rules, and ML models trained on historical data

- Use of network to uncover otherwise hidden relations
  - Much harder to blend in

THE DIFFICULTIES WITH NETWORKS

- Data used in actuarial problems often comes in tabular format
  - Machine learning methods are tailored to that

- Networks keep changing their structure
  - Not clear how to capture this as a table
  - Need to have as much network information as possible in tabular format



NETWORK EMBEDDING

- Definition:
  - Given network $G(V, E)$. Let $d \geq 1$ be the dimensionality of the node/network embedding. A node embedding function $f: V \rightarrow \mathbb{R}^d$ is a map that maps each node $v \in V$ to a real-valued feature vector in $\mathbb{R}^d$, where $d \ll |V|$.

- Intuitively:
  - Translate network into latent Euclidean space (table)
  - Capture as much network information in only a few features

GENERAL POINTS OF ATTACK

- Need for network embeddings is clear

- Main questions remain
  - Does network information improve the model?
  - Does network information lead to novel insights?

- Done in via two approaches
  - Model performance metrics
  - Complementarity of the results

THE METHODS USED

02

THE DIFFERENT MODELS

- Construct different models
  - Start with a base-line model with only intrinsic features
  - Based on gradient boosting classifier from sklearn (Python)

- Add different network features and embeddings to original features
  - Basic network features
  - BiRank (guilt-by-association)
  - Metapath2Vec (shallow representation learning)
  - GraphSAGE (deep representation learning: graph neural network)

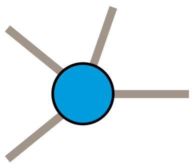
BASIC NETWORK FEATURES

- Can be directly extracted from the network
  - Based on neighbourhood and paths
- Highly interpretable
  - Capture the importance of the nodes
- We base us on the following
  - Degree
  - Betweenness centrality
  - Geodesic distance to itself
  - Number of cycles

OVERVIEW BASIC NETWORK FEATURES

Degree

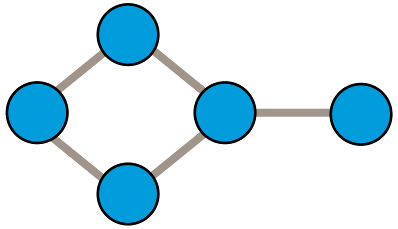
The number of connections



Betweenness Centrality

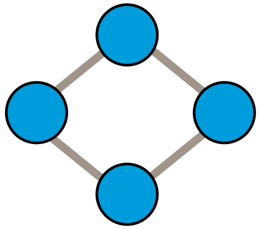
Percentage shortest paths going through it

Information flow



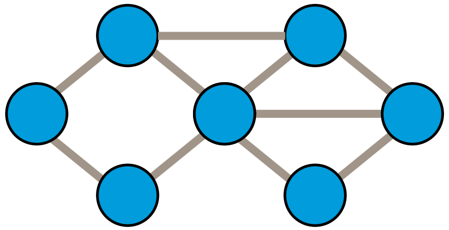
Geodesic to Itself

How closely/strongly related to itself



Number of Cycles

Can be used to uncover fraud rings



BIRANK

- Guilt-by-association algorithm for bipartite networks

- Incorporates fraud labels
  - Let information flow through network
  - Only set up for nodes of interest (claim/provider)

- Mathematically:

Aggregation of neighbouring scores

$$c_i = \alpha \boxed{\sum_{j=1}^{|P|} \frac{w_{ij}}{\sqrt{d_i d_j}} p_j} + (1-\alpha) \boxed{c_i^0}$$

Prior belief based on fraud labels

$$p_j = \boxed{\sum_{i=1}^{|C|} \frac{w_{ij}}{\sqrt{d_i d_j}} c_i}$$

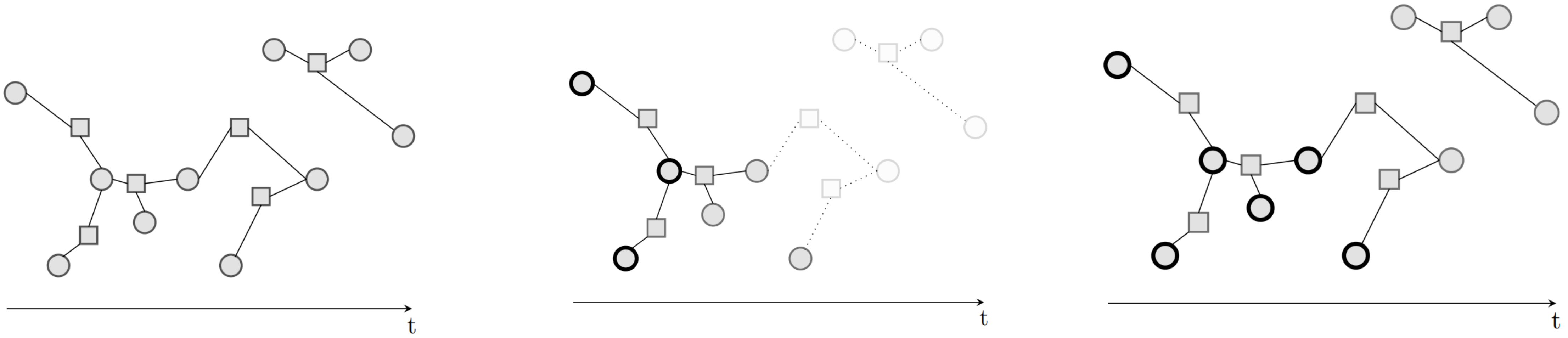
BIRANK

- Iteration using matrix-vector multiplication

$$\boldsymbol{p} = S^T \boldsymbol{c}$$
$$\boldsymbol{c} = \alpha S \boldsymbol{p} + (1+\alpha)\boldsymbol{c}^0$$

  - S is the rescaled adjacency matrix

- Important to avoid time-leakage
  - Only use labels up to a certain point



METAPATH2VEC

- Try to capture neighbourhood structure

- Does not use fraud data

- Takes meta-paths through network
  - We say what is allowed and what not

- Paths are seen as sentences, with nodes our vocabulary
  - Put into NLP to obtain the low-dimensional embedding

- Takes heterogeneity into account

GRAPHSAGE

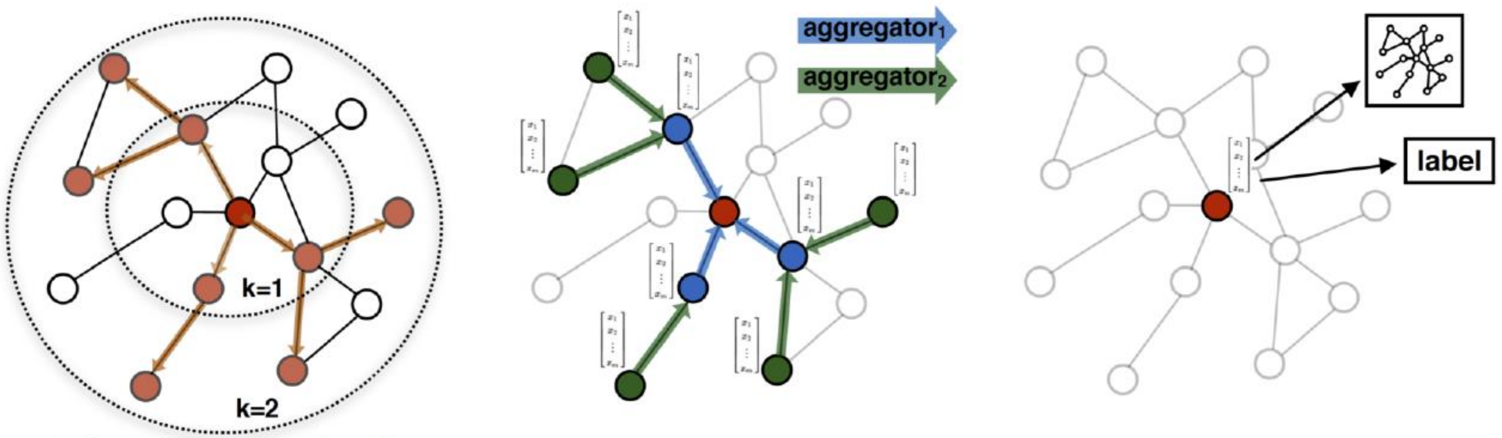
- Graph Neural Network
- Sample neighbours
- Aggregate from that sample
- Main Advantages
  - Inductive: generalisable to unseen nodes
  - Scalable: applicable on large graphs

GRAPHSAGE

- Graph Neural Network
  - $h_{N(v)}^{k} = \text{AGGREGATE}\left(\{h_u^{k-1}, \forall u \in N(v)\}\right)$
  - $h_v^{k} = \boxed{\sigma}\left(\boxed{\boldsymbol{W}^{k}} \cdot \text{CONCAT}\left(h_v^{k-1}, h_{N(v)}^{k}\right)\right)$

Non-linear activation funtion

Trainable weights



1. Sample neighborhood
2. Aggregate feature information from neighbors
3. Predict graph context and label using aggregated information

THE DATA AND PERFORMANCE METRICS

03

DATA SETS

- 2 data sets
  - Real-world motor insurance data set from an insurance company active on the Belgian market
  - Open-source data set from Kaggle on health care providers

- The first is to have a feeling of the performance at a company

- The second is to enhance reproducibility and knowledge sharing

DATA SETS

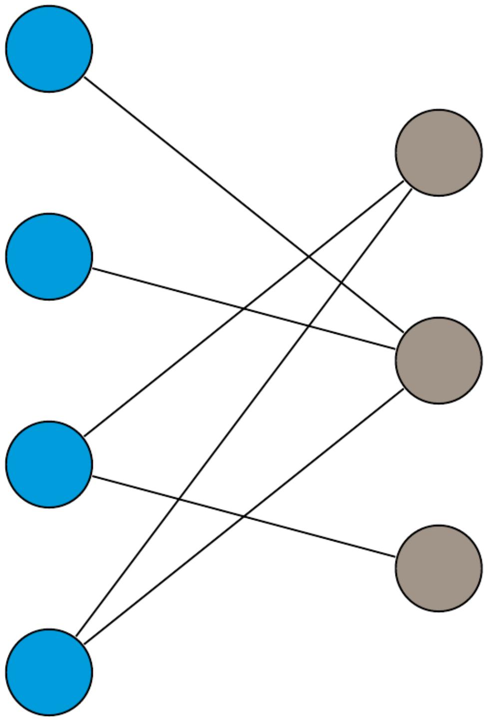
- Labels are highly imbalanced
  - Motor insurance: 3% investigated and 0.3% overall fraud
  - Health care providers: 9.4% labelled fraud

- Motor insurance has intrinsic features
  - Used for the base-line model

CONSIDERATIONS FOR BIRANK

- Need for bipartite graph

- Nodes of interest = group 1
- All other nodes = group 2



CONSIDERATIONS FOR METAPATH2VEC

- Not allowing random walks
  - Change to wander to much without going through node of interest

- Define meta-paths
  - Motor
    - $\mathcal{P}_1$: claim → contract → claim
    - $\mathcal{P}_2$: claim → counterparty → claim
    - $\mathcal{P}_3$: claim → broker → claim
  - Health
    - $\mathcal{P}_1$: provider → claim → provider
    - $\mathcal{P}_2$: provider → claim → physician → claim → provider
    - $\mathcal{P}_3$: provider → claim → beneficiary → claim → provider

CONSIDERATIONS FOR GRAPHSAGE

- No re-interpretation of the network

- Include feature data
  - Motor: claim-specific data available
  - Health: no data for providers, but data for patients and claims

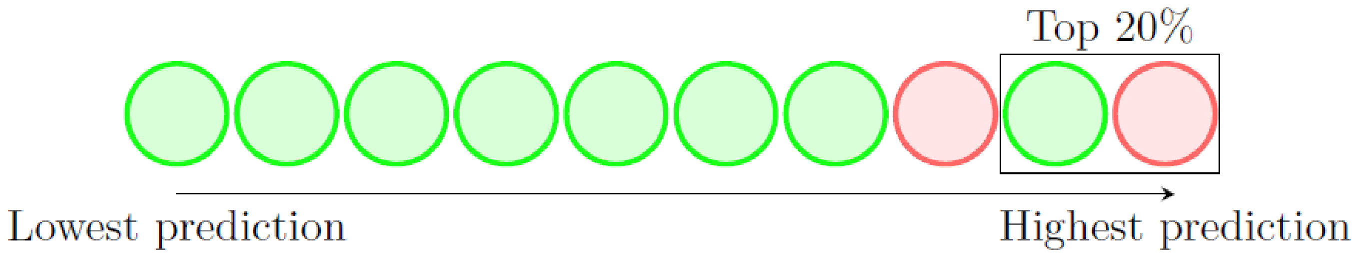
- When no feature data, set feature equal to 1

PERFORMANCE METRIC

- Three key metrics
  - Area under the ROC curve
  - Area under the precision-recall curve (average precision)
  - Lift curve
- AUC is more widely known
- Average precision is preferred when dealing with (highly) skewed data
- Lift curve looks locally
  - Important when putting model in production

LIFT CURVE

- Looks locally
  - At different levels, calculate the lift
- See how much more fraud is prevalent in top percentiles
- When put in production, only resources to investigate small part
- Example:
  - Red is real fraud
  - Lift is $2.5 = \frac{0.5}{0.2}$

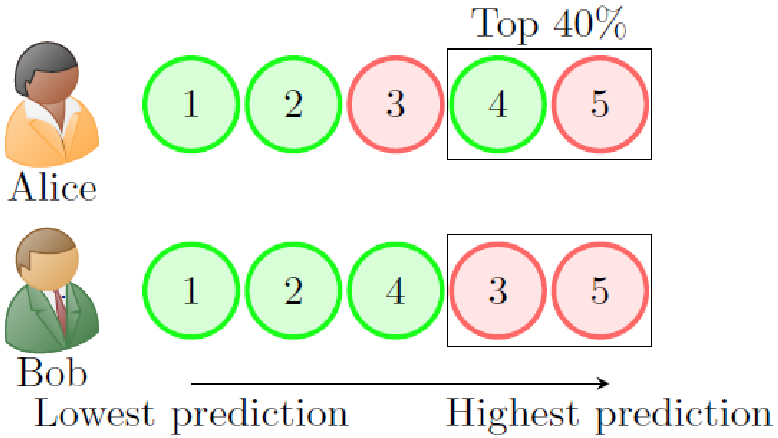


COMPLEMENATIRY

- Study added value of network features
  - Info not captured by intrinsic features

- Compare true positive between models
  - Do this at different levels

- Example:
  - Two models (red is real fraud)
  - Alice has 0% compl. to Bob
  - Bob has 50% compl. to Alice

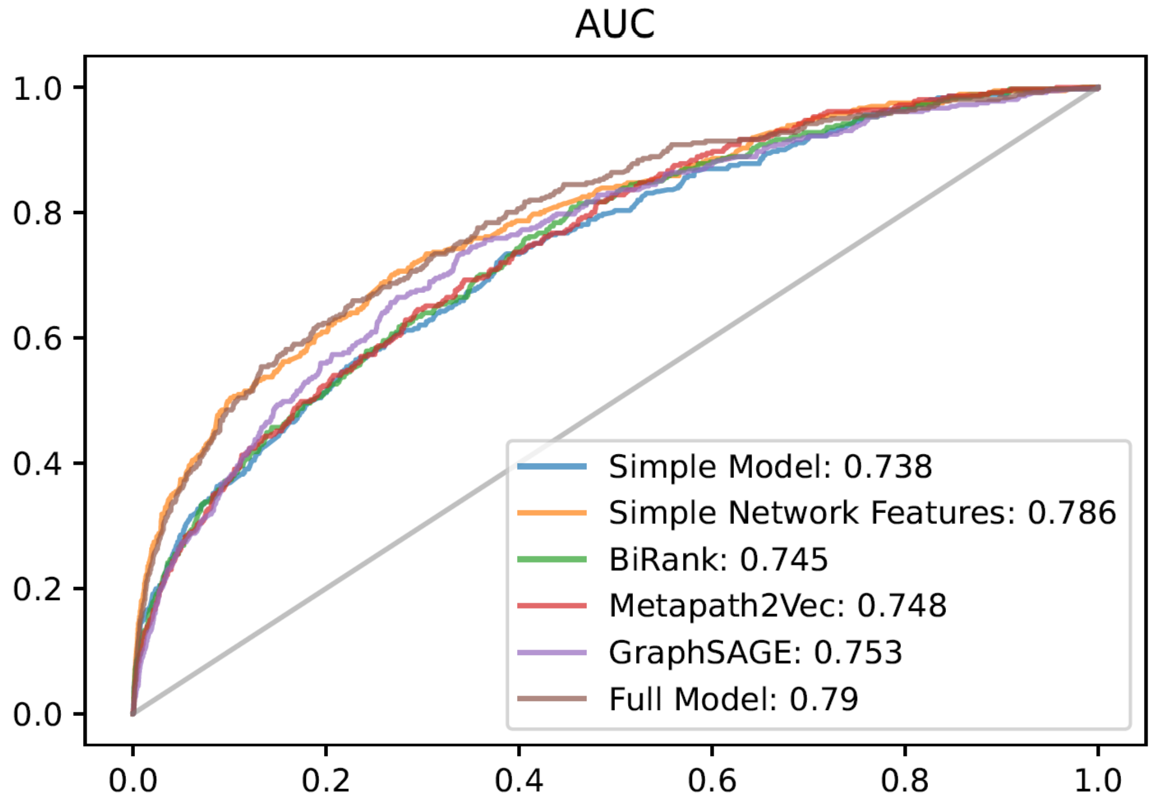


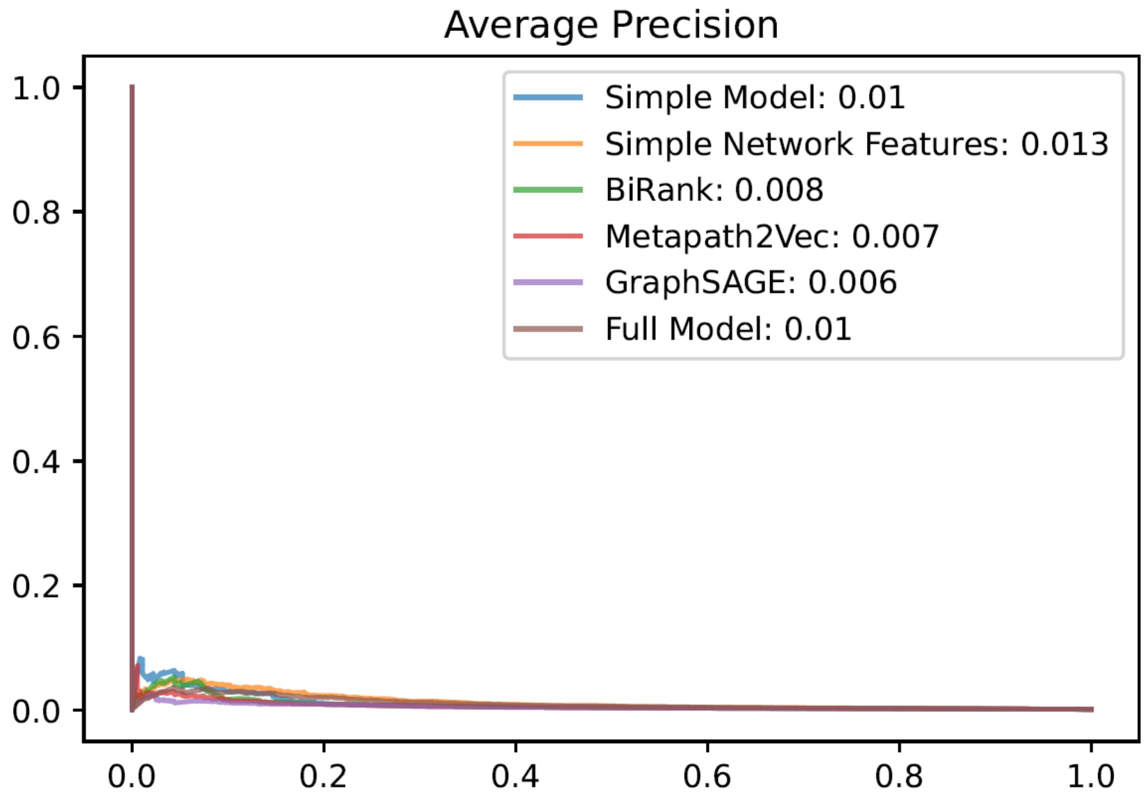
THE RESULTS

04

BELGIAN MOTOR INSURANCE DATA SET

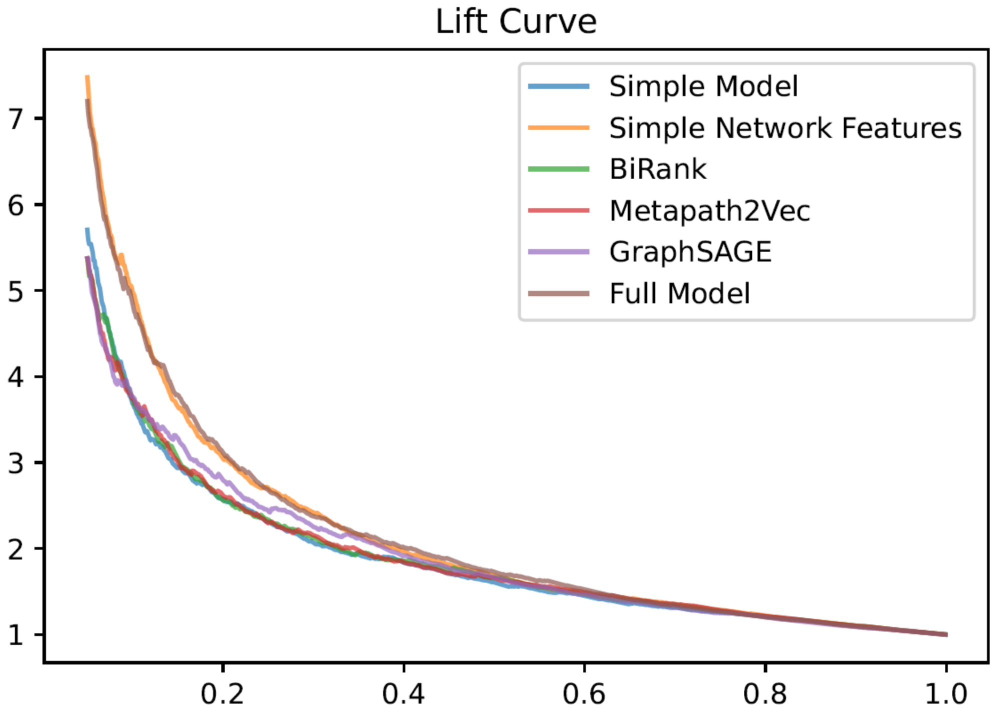
- Simple model: only intrinsic features
- All network features are added on top of those





BELGIAN MOTOR INSURANCE DATA SET

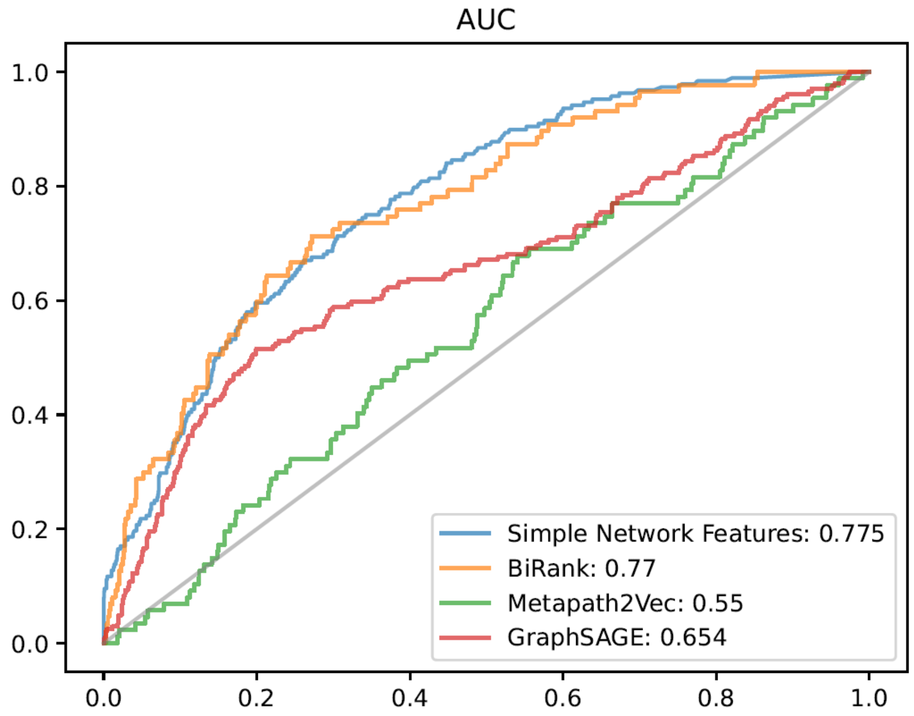
- Simple model: only intrinsic features
- All network features are added on top of those

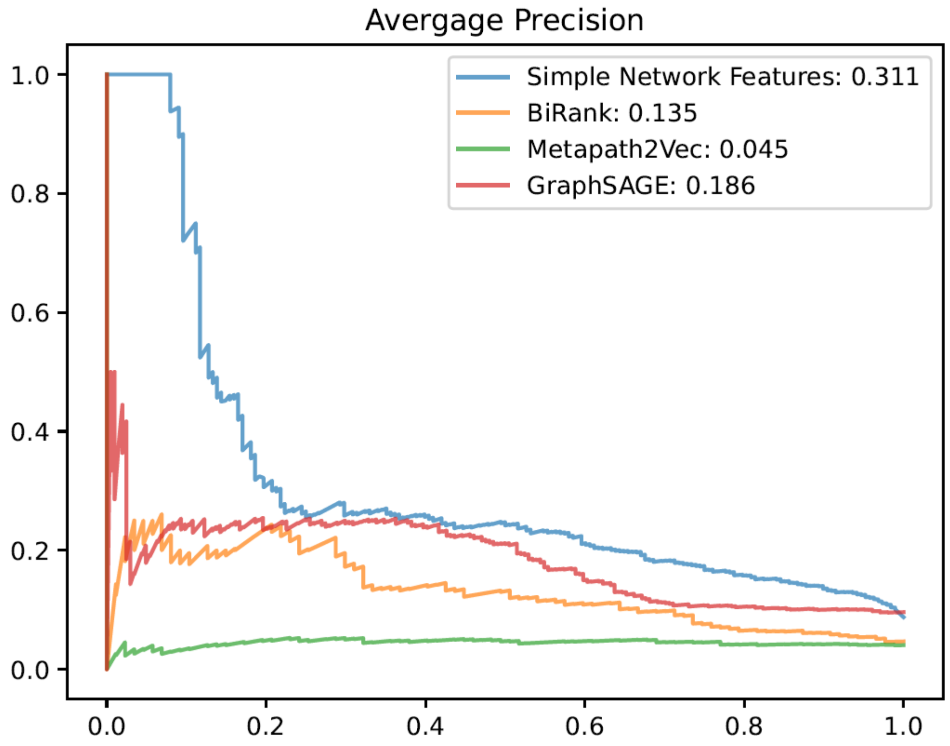


HEALTH CARE PROVIDER DATA SET

- No intrinsic features for health care providers

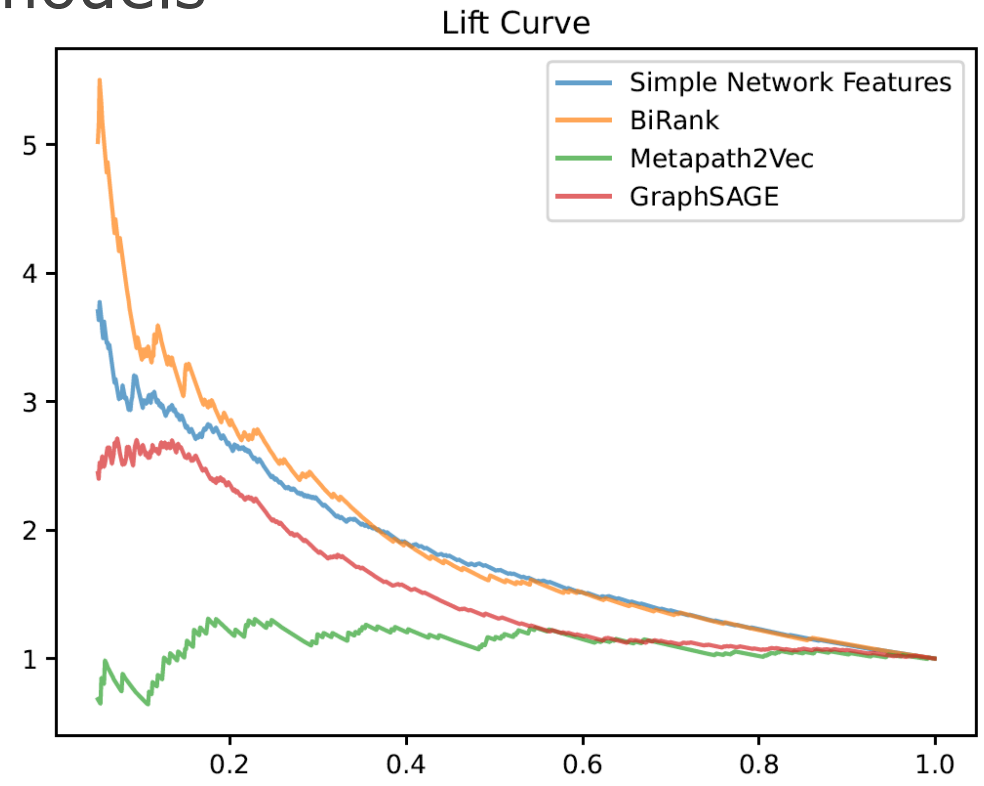
- Individual network models





HEALTH CARE PROVIDER DATA SET

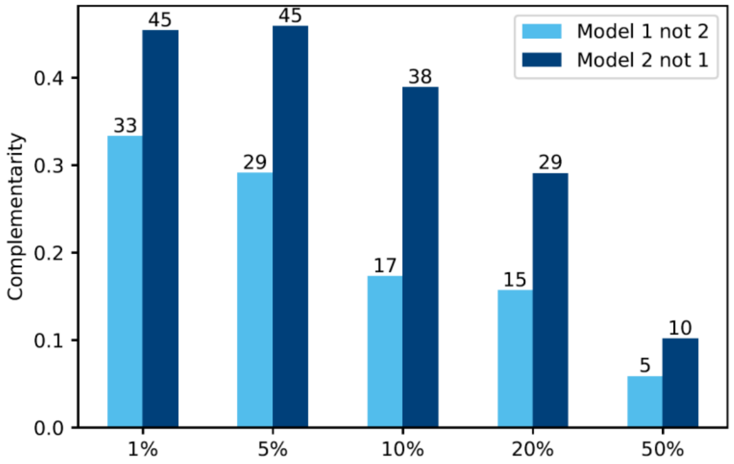
- No intrinsic features for health care providers
- Individual network models



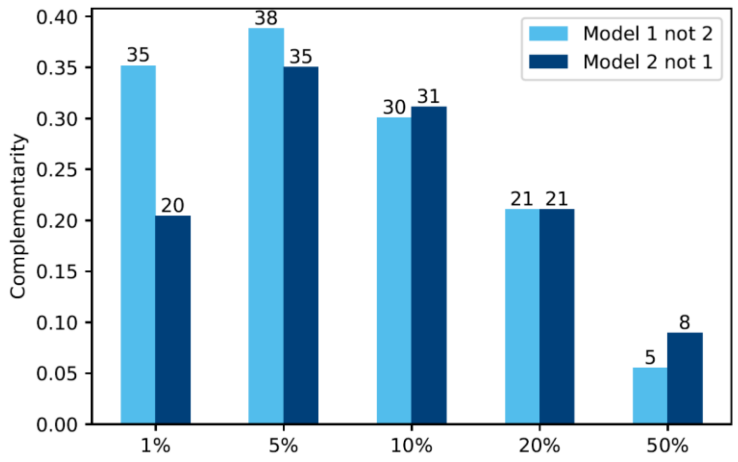
COMPLEMENTARITY

- Only relevant for the motor insurance dataset

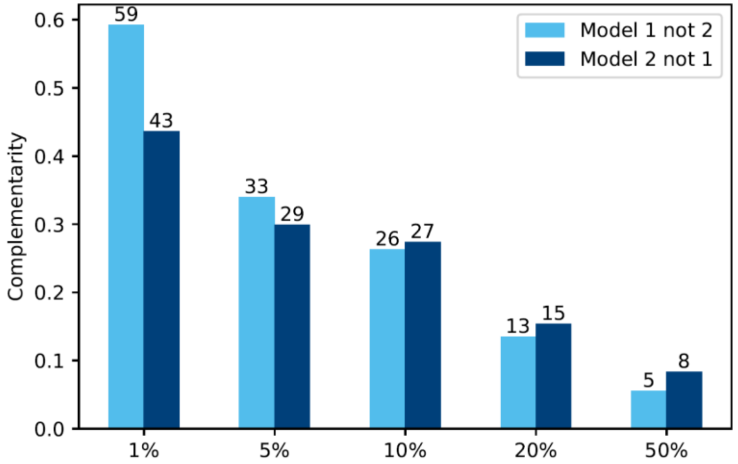
Simple

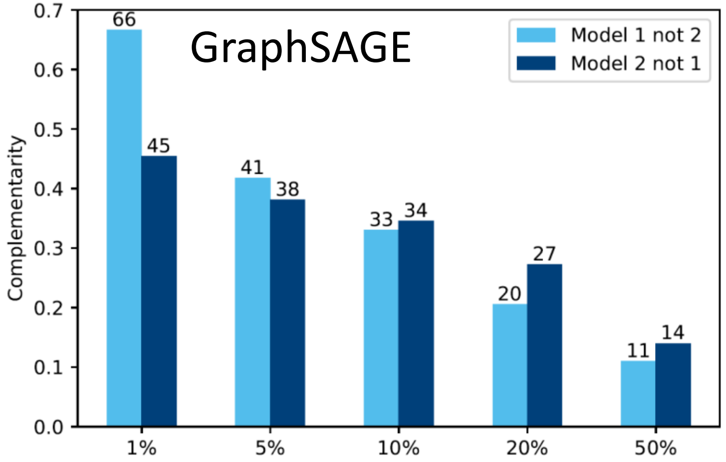


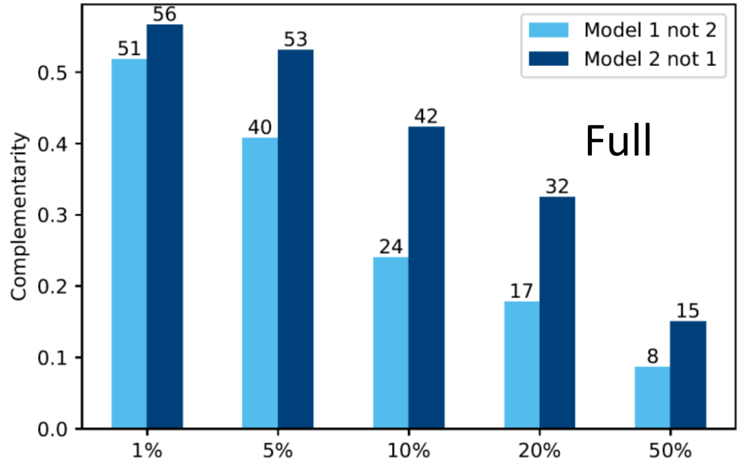
BiRank



Metapath2Vec







Conclusion

- Networks are a natural extension for fraud detection in insurance
- Vast variaty of methods available
  - Guilt-by-association
  - Shallow learners
  - Graph Neural Networks
- Complex methods are not necessarily better methods
- Network features uncover novel fraud patterns

REFERENCES

- Nino Arsov and Georgina Mirceva. Network embedding: An overview. arXiv preprint arXiv:1911.11726, 2019.
- Bart Baesens, Veronique Van Vlasselaer, and Wouter Verbeke. Fraud analytics using descriptive, predictive, and social network techniques: a guide to data science for fraud detection. John Wiley & Sons, 2015.
- Albert-Laszlo Barabase. Network Science. Cambridge University Press, 5 edition, 2020.
- Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In Proceedings of the 23rd international conference on Machine learning, pages 233–240, 2006.
- Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. metapath2vec: Scalable representation learning for heterogeneous networks. In Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, pages 135–144, 2017.
- Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. Advances in neural information processing systems, 30, 2017.
- Marıa Oskarsdottir, Waqas Ahmed, Katrien Antonio, Bart Baesens, Remi Dendievel, Tom Donas, and Tom Reynkens. Social network analytics for supervised fraud detection in insurance. Risk Analysis, 42(8):1872–1890, 2022.
- Rafael Van Belle, Charles Van Damme, Hendrik Tytgat, and Jochen De Weerdt. Inductive graph representation learning for fraud detection. Expert Systems with Applications, 193:116463, 2022.
- Wouter Verbeke, David Martens, and Bart Baesens. Social network analysis for customer churn prediction. Applied Soft Computing, 14:431–446, 2014.
- https://www.kaggle.com/datasets/rohitrox/healthcare-provider-fraud-detection-analysis
- https://github.com/B-Deprez/NetworkFraud_BiRank_M2V_SAGE

Bruno has a master's degree in Mathematics and in Actuarial and Financial Engineering, both at KU Leuven.

During his studies, he worked as an actuarial consultant at KPMG Belgium for 2.5 years, serving larger and smaller insurance companies on the Belgian market.

He is currently pursuing a joint PhD in fraud detection at the Research Centre for Information Systems Engineering (LIRIS) at KU Leuven and at the Department of Mathematics at the University of Antwerp, in Belgium.

ABOUT ME

Bruno Deprez

KU Leuven & UAntwerp

EAA e-Conference on
Data Science & Data Ethics

16 May 2023

Thank you very much for your attention

Contact

Bruno Deprez

KU Leuven & UAntwerp

+32 477/09 97 31

bruno.deprez@kuleuven.be