

 / Research

Graph-Based Anomaly Detection: Methods and Hands-On Implementation



Organization

- Internet connection
- Google account
- Questions
- Presentation - <https://github.com/seznam/MLPrague-2026/tree/main/presentation>

Agenda

- About Seznam.cz (5 min)
- Introduction to Graph-based systems (5 min)
- Graph-based systems at Seznam.cz (5 min)
- Introduction to Graph-based anomaly detection + hands-on (35 min)
- Supervised methods + hands-on (40 min)
- *Break* (30 min)
- Supervised methods hands-on (25 min)
- Unsupervised methods + hands-on (45 min)

About Seznam.cz

Technological company and media house

About **7.6 million unique users** visit Seznam.cz services every month

The product portfolio consists of highly popular sites such as:



88%

**Monthly reach of the
Czech online population**



Jakub Chynoradský

Identity solutions



Adam Jurčík

Identity solutions



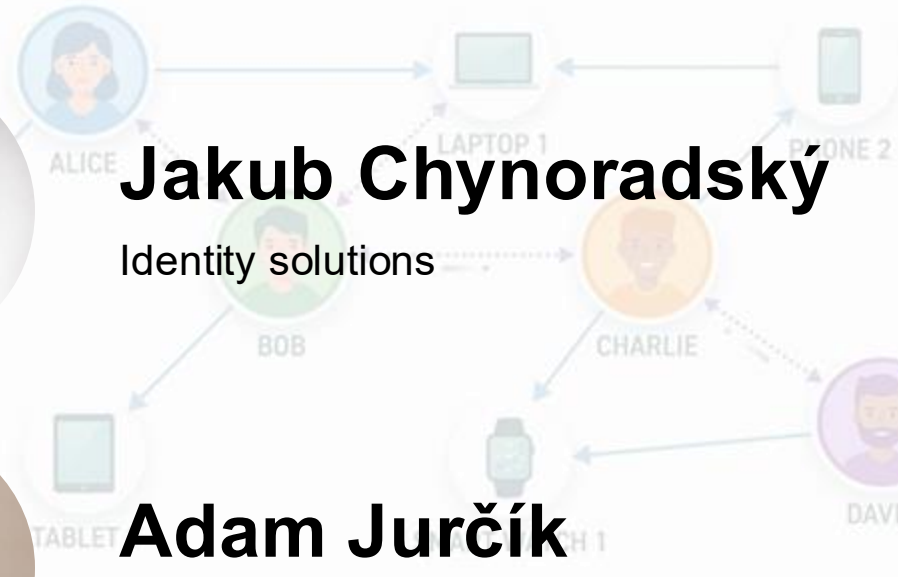
Jaroslav Kuchař

Identity & Geolocation supervision



Marek Šrank

Geolocation solutions

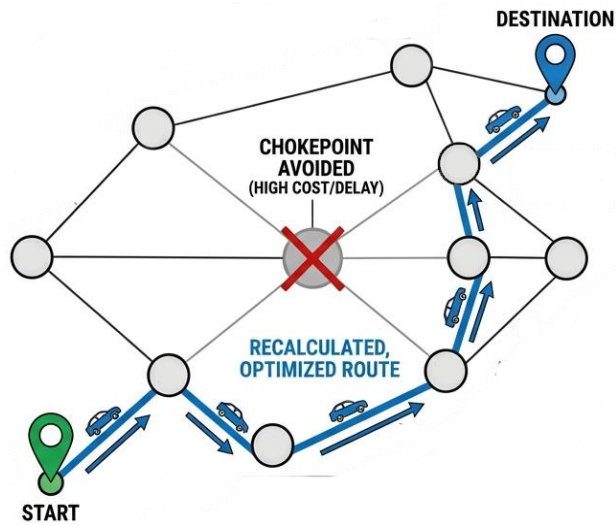


Agenda

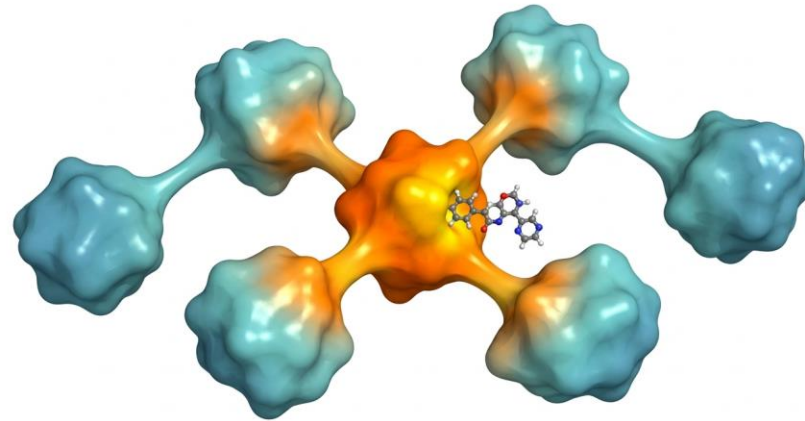
- About Seznam.cz
- **Introduction to Graph-based systems**
- Graph-based systems at Seznam.cz
- Introduction to Graph-based anomaly detection + hands-on
- Supervised methods + hands-on
- *Break*
- Supervised methods hands-on
- Unsupervised methods + hands-on

Real-world Graphs

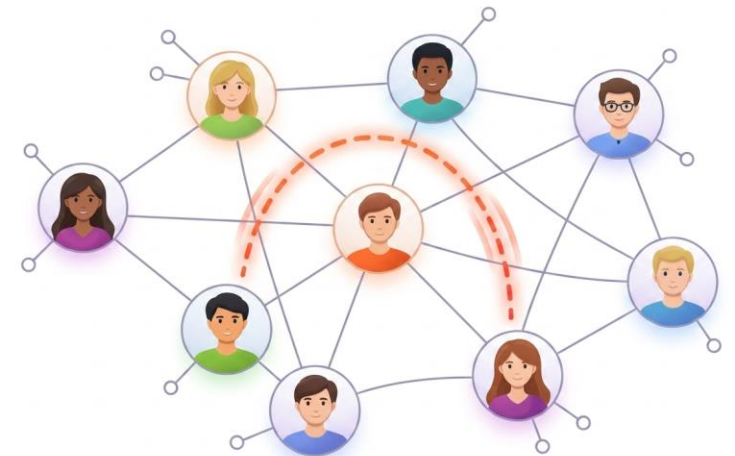
Navigation



Protein-protein interactions



Social network



Online Graphs



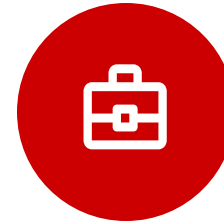
>3B

active users
Facebook
(social graph)



~400B

documents
Google index
(web graph)



~100B

nodes
LinkedIn graph
(professional network)

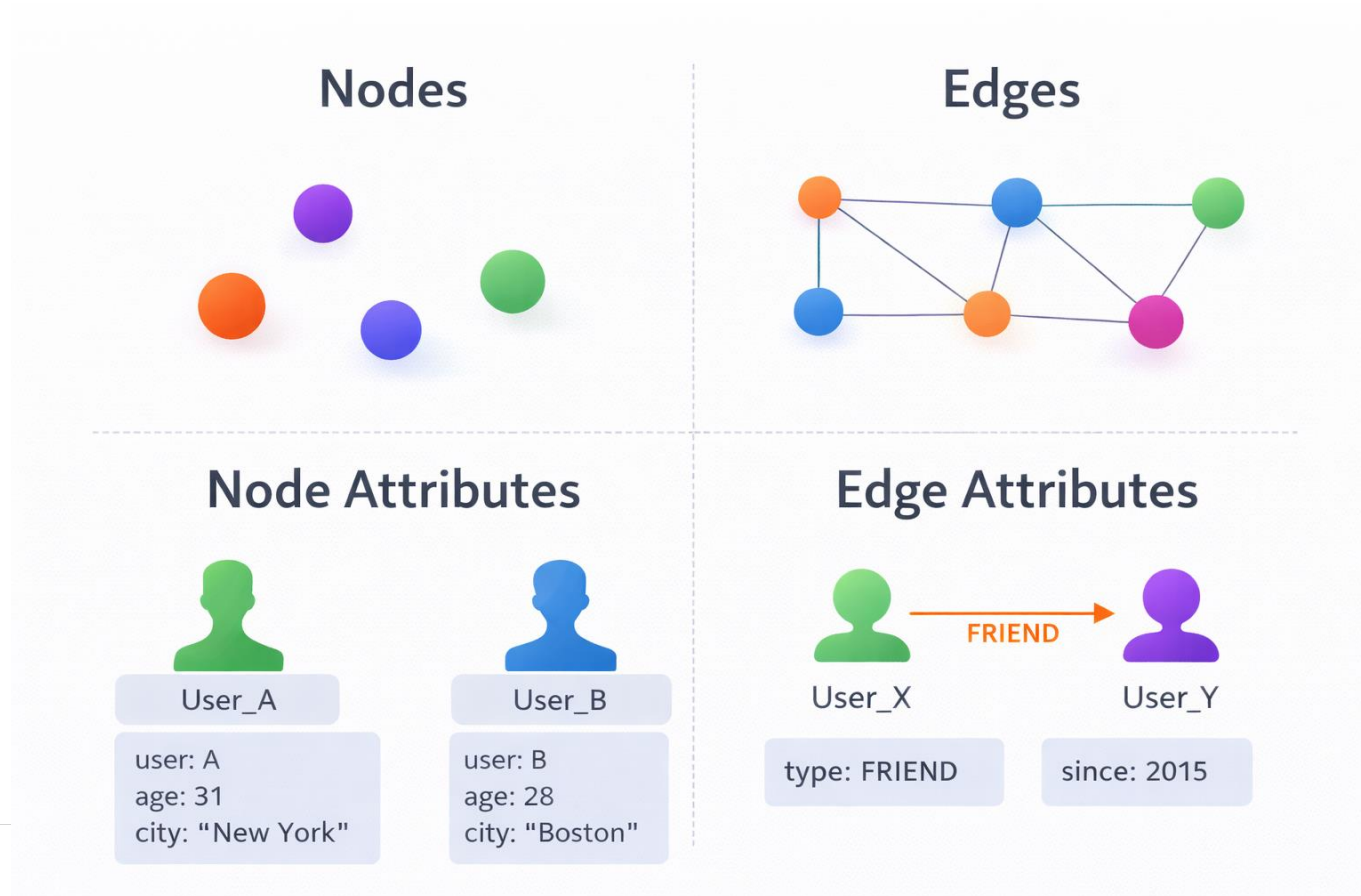


~600M

products
Amazon
(co-purchase network)

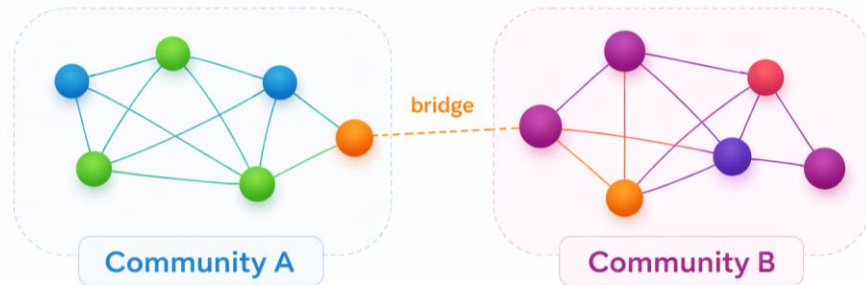
The value comes from relationships between entities
Graphs are natural representation

Graph Fundamentals

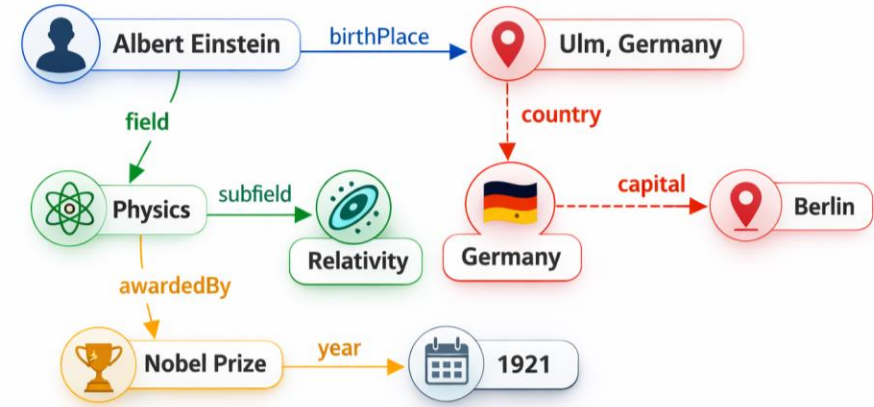


Graph Applications

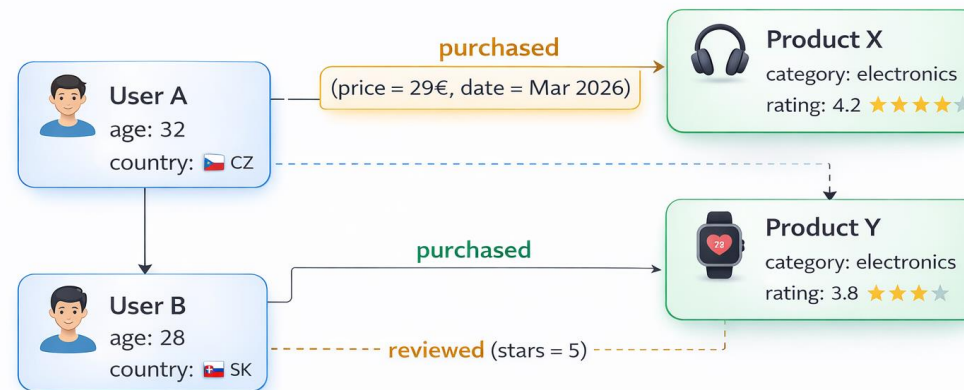
Social networks: community detection, node importance, link prediction...



Knowledge graph: Reasoning, entity & relation prediction, QA...



E-commerce: recommendations, rankings, fraud detection...

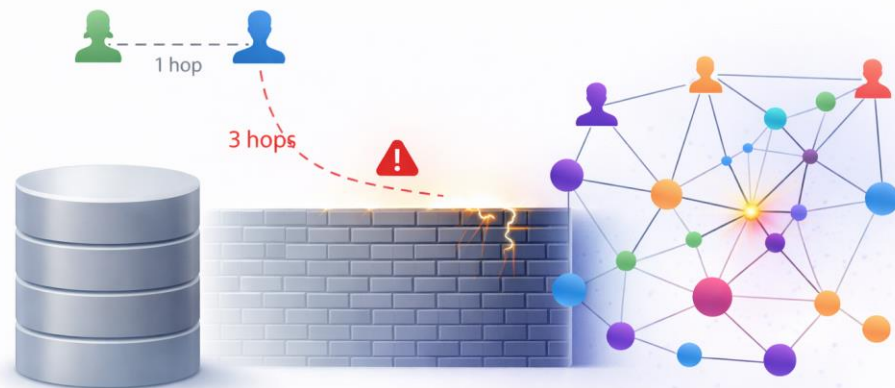
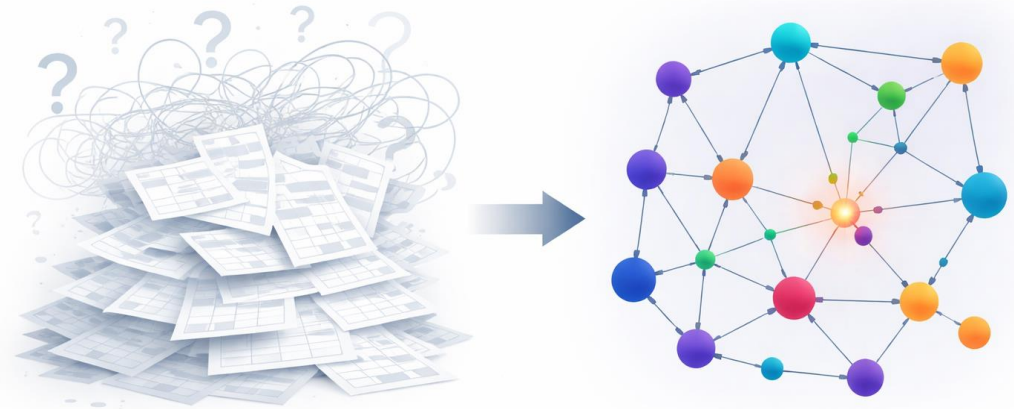


Modelling Data as Graphs

Views

- Tabular – each sample is independent
- Graph – samples are connected

Efficient "multi-hop" operations



Agenda

- About Seznam.cz
- Introduction to Graph-based systems
- **Graph-based systems at Seznam.cz**
- Introduction to Graph-based anomaly detection + hands-on
- Supervised methods + hands-on
- *Break*
- Supervised methods hands-on
- Unsupervised methods + hands-on

Discussions

medium

PROŽENY

GARÁŽ.CZ

News, lifestyle, topics, blogs

SUPER.CZ

SPORT.CZ

Seznam Zprávy

Novinky.cz

Graph of authors – normal and quality

Quality propagation for comments



~700K

Reactions
per day



~70K

Comments
per day



~600K

Users
per day



~1K

Harmful content
per day

Diskuze k článku na Seznam
Zprávy: Trik se složenkou na
131 korun oživil starou exekuci.
Soud to zarazil

AJ

Napište komentář

Dluhy by měly být placeny, ale
systém je nespravedlivý

35%

Exekutoři by měli být trestáni za
neetické jednání

28%

Exekutoři jsou vnímáni
jako zneužívající systém

25%

Oddlužení je vnímáno jako...

6%

Lichva a vysoké úroky jsou...

6%

Vše 358

Moje komentáře 0

Od nejlepších

sort by relevance



Sledovat

Před 18 hodinami

Dluhy se mají platit. Jsem rád, že z článku vyplývá, že musí uhradit alespoň původní částku 40 000 plus úroky. Nemravná exekuce je věc druhá, ale základní částku, kterou prokazatelně dostal zaplatit musí.

Proto nemám rád institut oddlužení, kde stačí zaplatit jen zlomek dluhu.

6 161

Odpovědět



Sledovat

Před 17 hodinami

Přesně, jak k tomu projdou věřitelé...nechápu systém oddlužení...každý mameluk si nepůjčuje a kašle na to...pak dostane dar od státu na úkor věřitelů. Jiná je pokud zkrachuje nějaká firma, podnikatel...ale ti salamisti, kteří si půjčují už s tím, že prostě platit nebudou a že jich je...prd nic bych jim nepovolila

11 51

Odpovědět

Zobrazit další reakce

Assistant (BETA)



Powerful search and problem solving

Goods knowledge graph 

Item relevance for shopping



>500K

Users
per month



Millions

Conversations
per month



5-10%

Goods interaction

Rád bych si koupil horské kolo.

I would like to buy a mountain bike

Vygenerováno na základě 2 kroků >

Zde je přehled tří modelů horských kol, které jste zmínil, včetně jejich základních parametrů a vlastností pro lepší porovnání:

 **Author Rival 27,5" — Spolehlivé a univerzální horské kolo vhodné pro začátečníky a juniory**

- Rám: duralový 6061, lehký a odolný
- Kola: 27,5" s pláště 2.10" pro dobrý grip a stabilitu
- Odpružení: zamykatelná vidlice RST Gila 100 mm zdvihu
- Brzdy: mechanické kotoučové Tektro
- Převedy: 24 rychlostí, Shimano TX800
- Hmotnost: 15 kg
- Cena od: 7 998 Kč



Author Rival 27,5" černé/
červené 2023/24 19"

unisex • dural • počet rychlostí: 24
• odpružená vidlice • typ brzd:...

7 998 – 10 990 Kč



Leader Fox Esent 29" bílé/
černé 2023 18"

unisex • dural • počet rychlostí: 8 •
odpružená vidlice • typ brzd:...

7 354 – 11 990 Kč



Trek Procaliber 9.5 Gen 3 29"
Matte Keswick Green/Mercu...

unisex • karbon • počet rychlostí: 12
• odpružená vidlice • typ brzd:...

41 990 – 46 990 Kč

   Zdroje

How to choose the right bike wheel size?

Jak vybrat velikost kola?



Identity



User experience personalization



Interactaction with multiple Seznam accounts:

- Personal, Organization, Occasional

Activity linking and isolation



~7.6M

Users per month



~7.4M

Accounts



~3.6M

Active accounts per day



Agenda

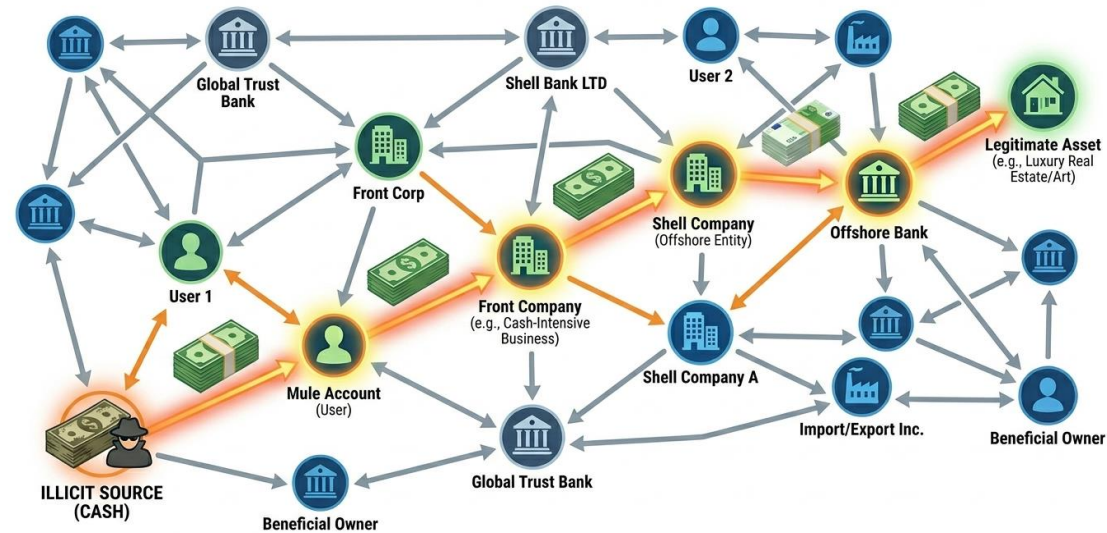
- About Seznam.cz
- Introduction to Graph-based systems
- Graph-based systems at Seznam.cz
- **Introduction to Graph-based anomaly detection + hands-on**
- Supervised methods + hands-on
- *Break*
- Supervised methods hands-on
- Unsupervised methods + hands-on

Anomaly detection

Anomalies – data that deviate significantly from the expected behavior of the majority

Graph anomalies may arise from:

- Rare behavior
- Coordinated malicious actions
- Structural irregularities



Graph anomaly detection leverages both connectivity patterns and attribute information

Where graph anomalies hide

Social platforms

- Spam accounts
- Fake reviews

Financial systems

- Credit card fraud

Cybersecurity

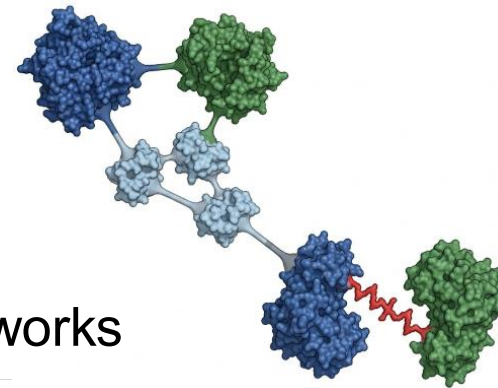
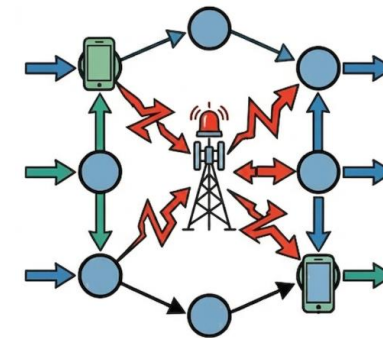
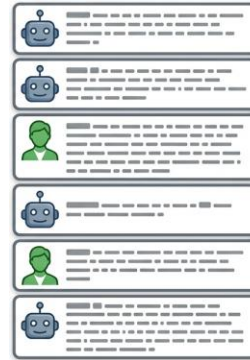
- Abnormal access patterns

Biological

- Anomalous interaction patterns in protein networks

Infrastructure

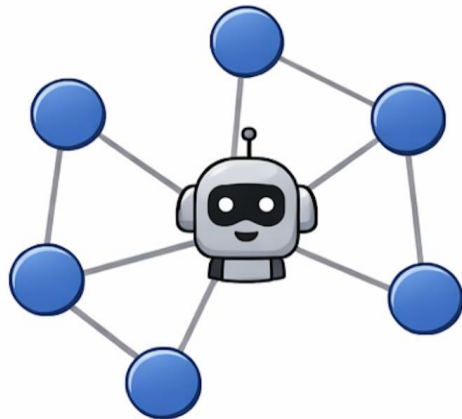
- Unusual traffic flows in telecom or power grids



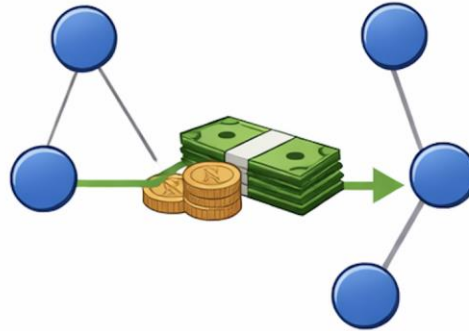
Different forms of anomalies

Anomalies can take various forms:

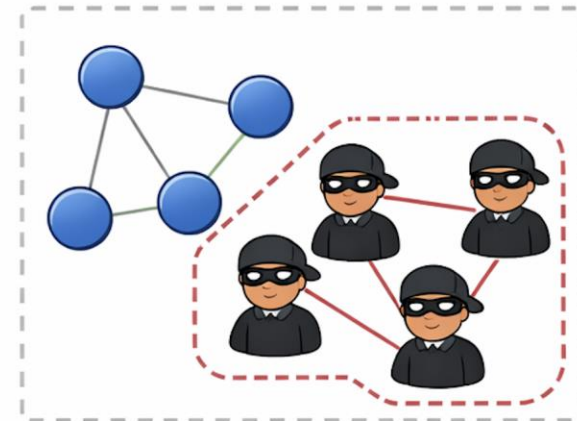
Node Anomaly



Edge Anomaly



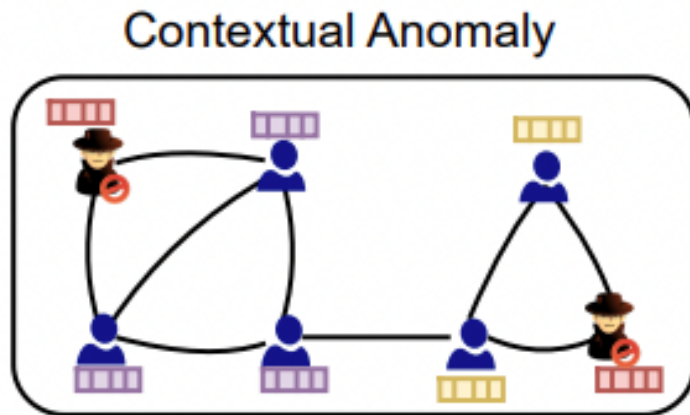
Subgraph Anomaly



Different types of anomalies

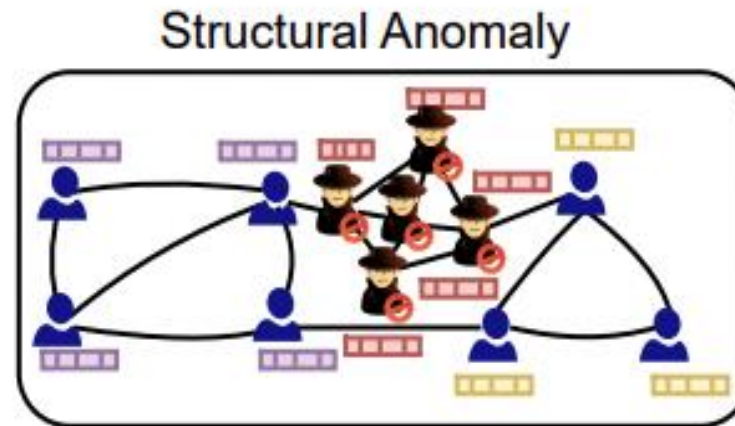
Contextual

- have natural neighboring structures but their attributes are corrupted

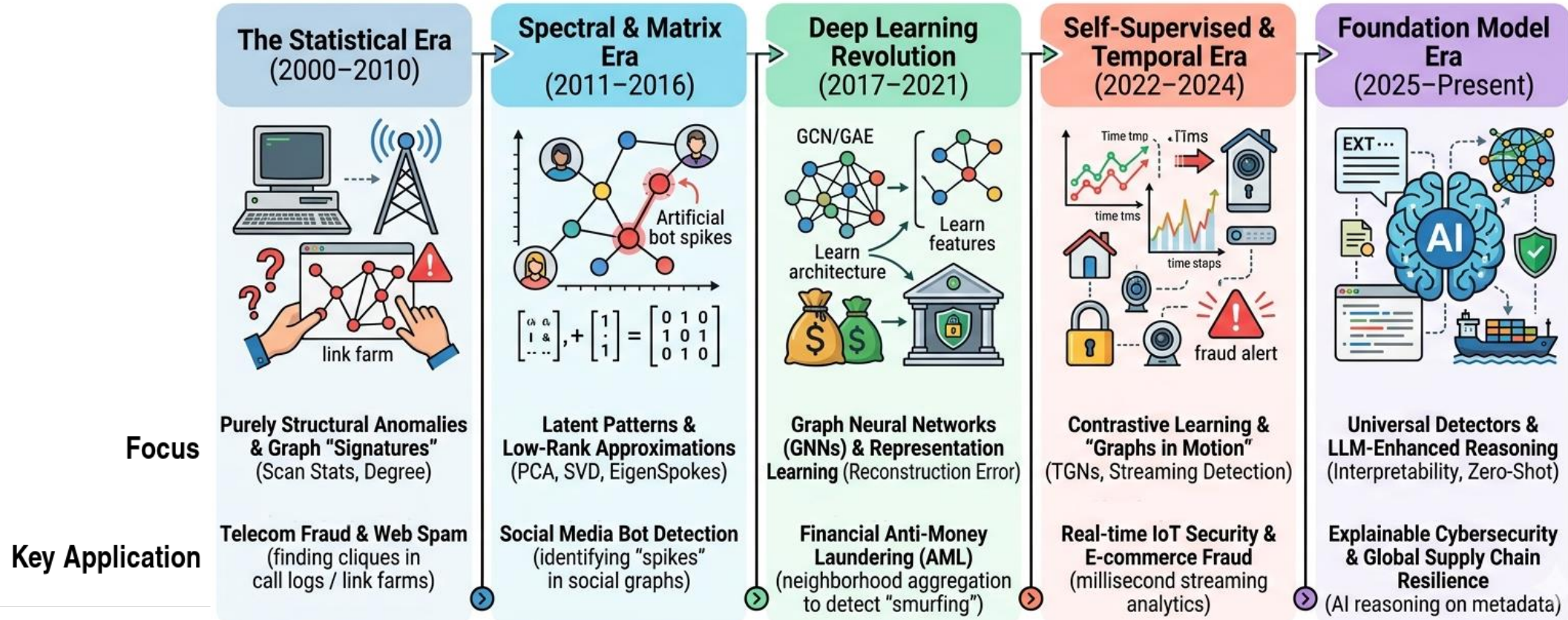


Structural

- attribute information of the structural anomalies is often normal, while they have several abnormal links to other nodes



Graph Anomaly Detection Evolution



Summary: Graph Anomalies

Relations among data points are part of many real-world problems

Graphs enable considering multi-hop relations

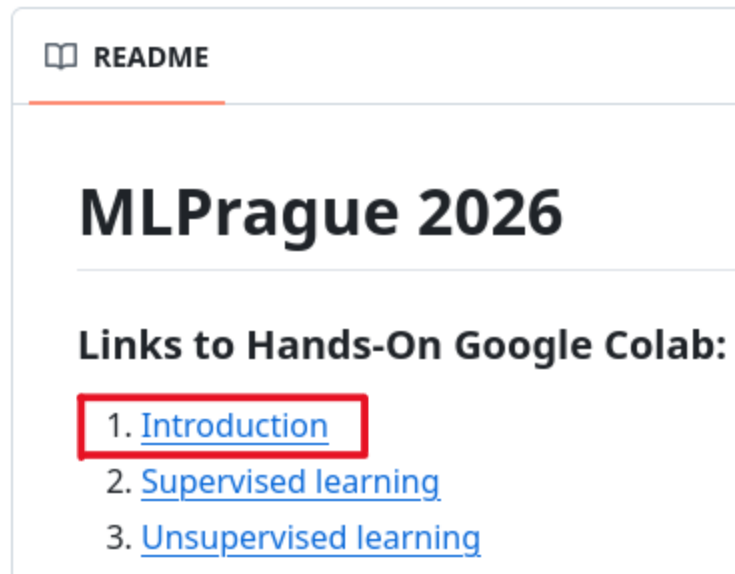
Graph anomaly detection enables mining of relations for outlier detection



How are relations modelled in a graph?

Hands-on: Introduction

- Open <https://github.com/seznam/MLPrague-2026>
- Ask us if you have any issue



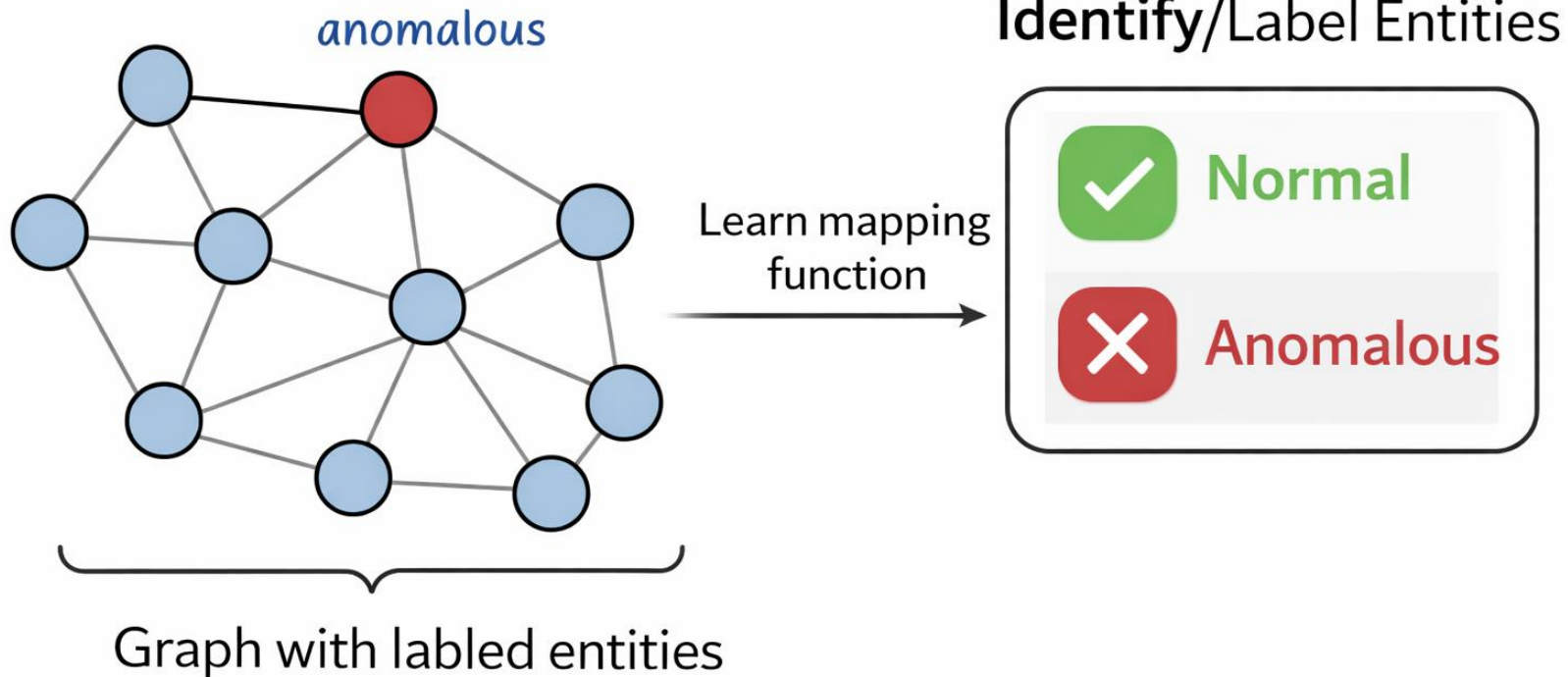
A screenshot of a GitHub README page for the repository 'MLPrague 2026'. The page title is 'MLPrague 2026' and it includes a section titled 'Links to Hands-On Google Colab:'. Under this section, there is a list of three links: '1. Introduction', '2. Supervised learning', and '3. Unsupervised learning'. The '1. Introduction' link is highlighted with a red rectangular box.

Agenda

- About Seznam.cz
- Introduction to Graph-based systems
- Graph-based systems at Seznam.cz
- Introduction to Graph-based anomaly detection + hands-on
- **Supervised methods + hands-on**
- *Break*
- Supervised methods hands-on
- Unsupervised methods + hands-on

Supervised methods – task definition

Supervised = we have labels.



Supervised methods - overview

Handcrafted features

+ ML

Embeddings

+ ML

End-to-end learning:

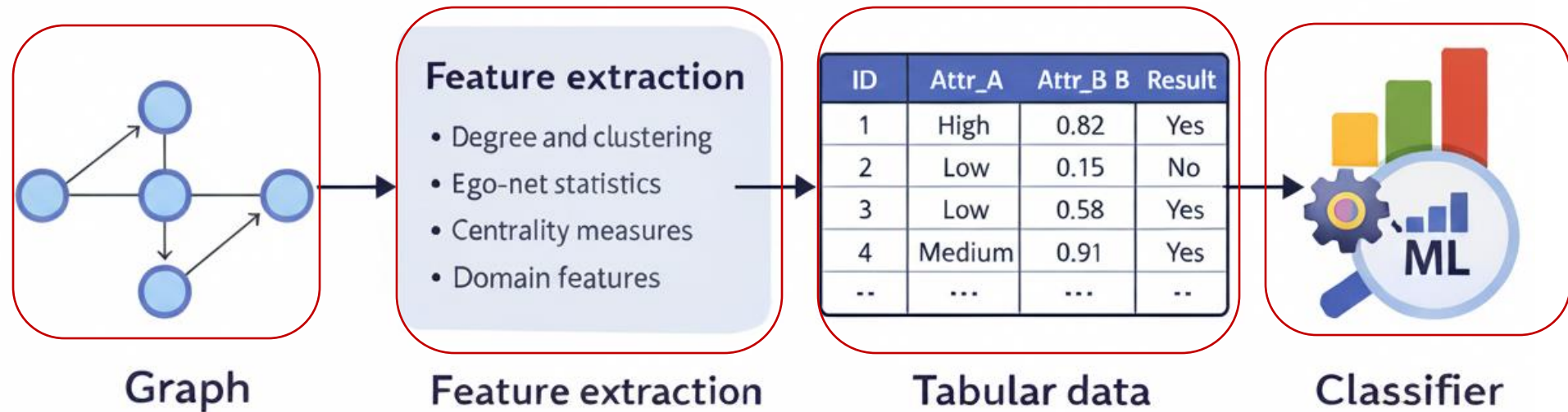
GNN



Increasing integration of graph structure into learning

Handcrafted features + ML

Treat it like tabular data.



Handcrafted features + ML

Pros

- Simple
- Fast
- Interpretable
- Well-understood
- Works with any classifier

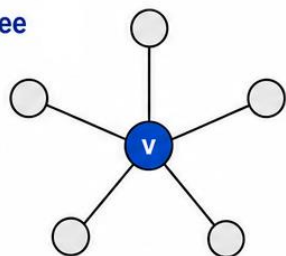
Cons

- Manual feature engineering
- Need to decide what matters (bias)
- Can lose structural information
- Won't capture complex patterns

Common Graph Features

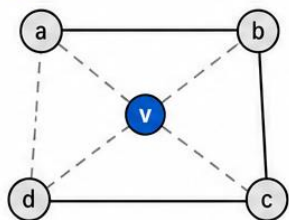
Local

Node degree



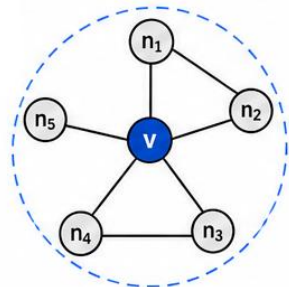
$$\text{deg}(v) = 5$$

Clustering coefficient



$$\text{CC}(v) = \frac{3}{6} = 0.5$$

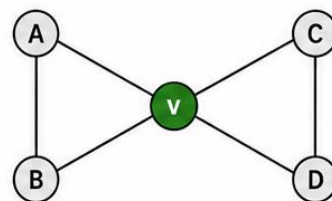
Ego net features



$$\begin{aligned} |V_{\text{ego}}| &= 6 \\ |E_{\text{ego}}| &= 7 \\ \text{density} &= \frac{2|E|}{|V|(|V|-1)} \\ &= \frac{2 \cdot 7}{6 \cdot 5} = 0.47 \end{aligned}$$

Global

Betweenness centrality

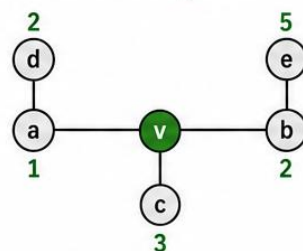


shortest paths

- A-B (not through v)
- A-C (through v)
- A-D (through v)
- B-C (through v)
- B-D (through v)
- C-D (not through v)

$$\text{BC}(v) = \frac{4}{6} \approx 0.67$$

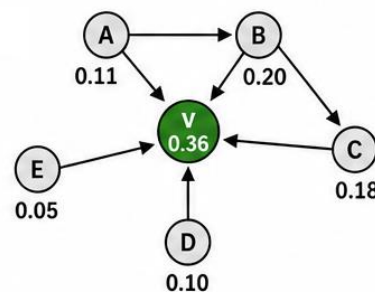
Closeness centrality



$$\text{avg dist} = \frac{1+2+2+3+5}{5} = 2.6$$

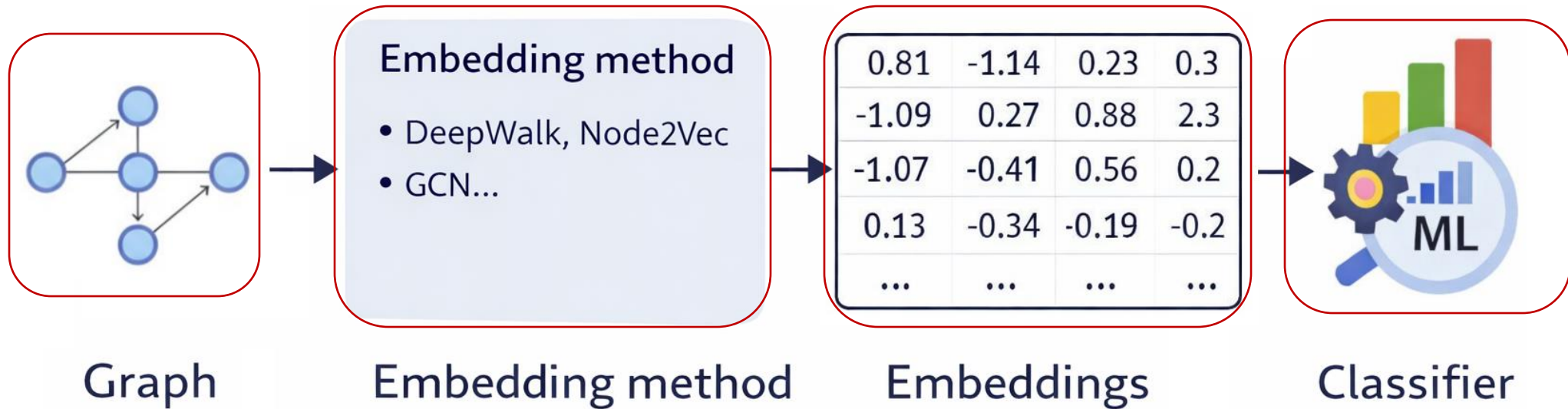
$$\text{closeness} = \frac{1}{2.6} \approx 0.38$$

PageRank



Embeddings + ML

Let embeddings do the feature engineering.



Embeddings + ML

Pros

- Learned features
- Can capture more subtle structural patterns
- Flexible
- Easier to debug than GNN

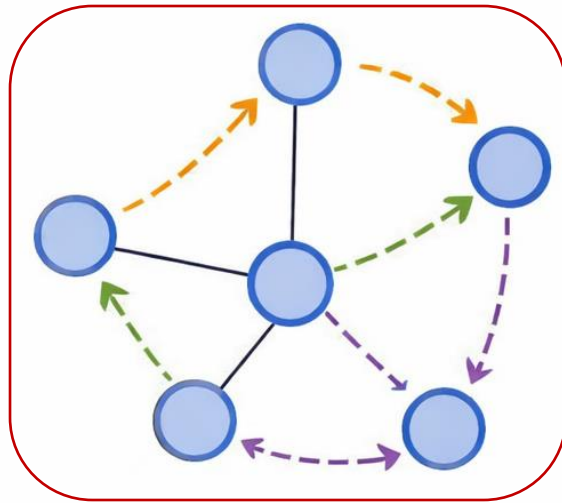
Cons

- Embeddings not optimized for task
- Information loss in the decoupling

Node2Vec: Embeddings via Biased Random Walks

Nodes from similar contexts get similar embeddings.

① Generate random walks



② Treat the walks as sentences of words and feed them to Word2Vec

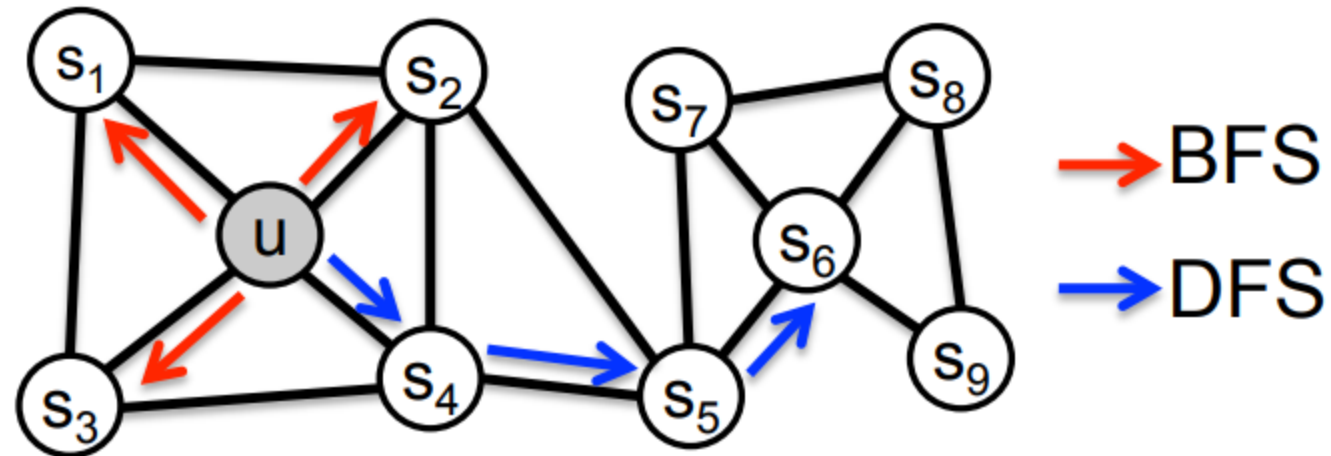
A B E
C D E
C B E



Node2Vec: Embeddings via Biased Random Walks

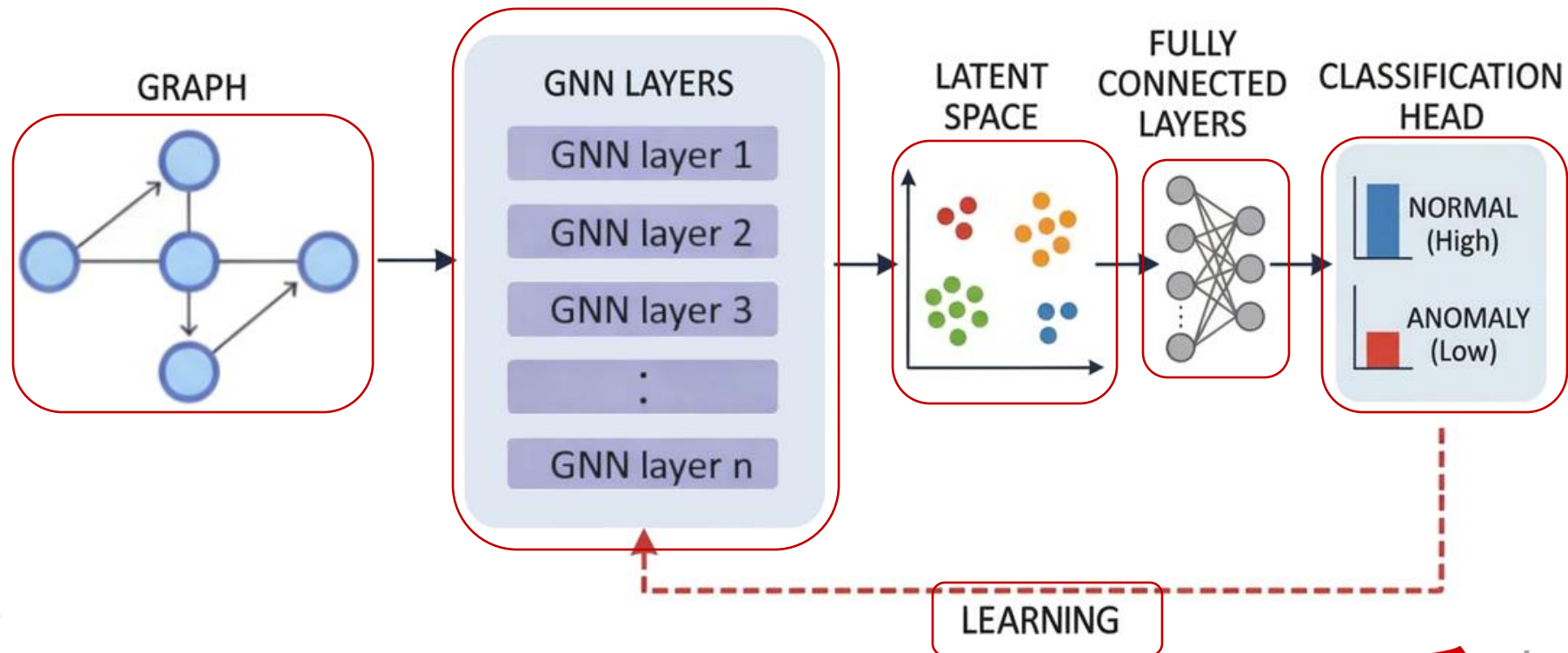
Walks are biased:

- **Parameter p** (return): controls backtracking – low p keeps the walk local
- **Parameter q** (in-out): controls exploration depth – BFS-like vs DFS-like exploration



Graph Neural Networks

End-to-end learning.



Graph Neural Networks

Pros

- SOTA performance
- Task-optimized representation
- Can capture complex patterns

Cons

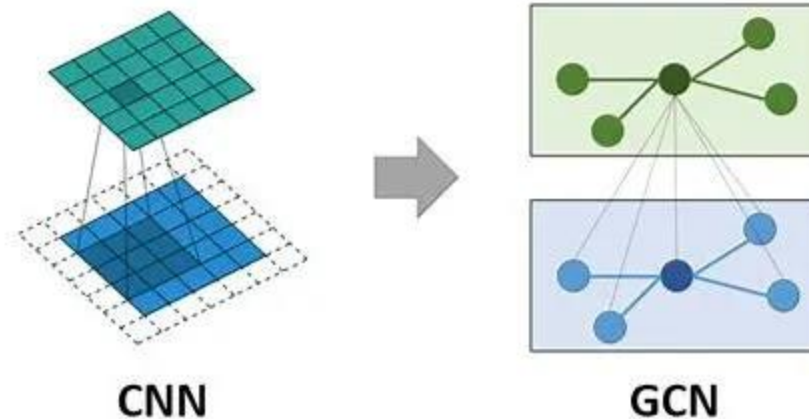
- Computationally heavier
- May require more data
- Harder to interpret

Graph Convolutional Neural Network (GCN)

Transforms aggregated features from graph neighbors using shared weights.

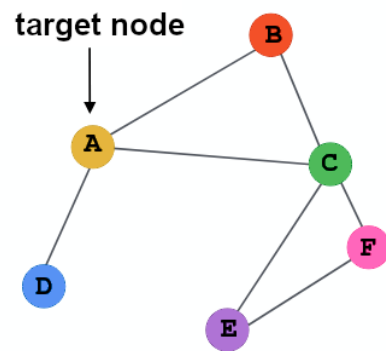
Similar to CNNs for images, but instead of using fixed position-based weights for pixels:

- 1) Aggregates features from graph neighbours (mean/sum)
- 2) Transforms with shared weights – same for all nodes

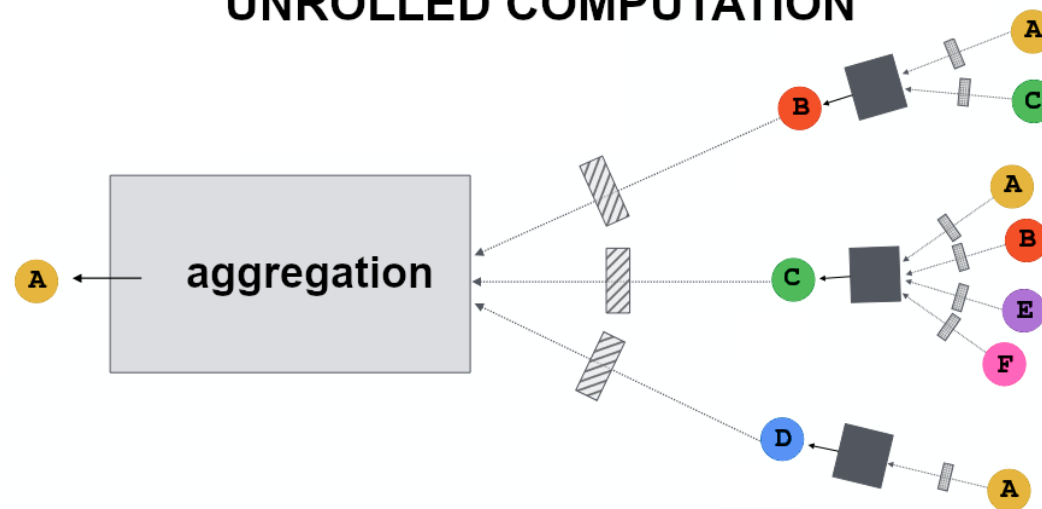


Graph Convolution Example

INPUT GRAPH



UNROLLED COMPUTATION



Summary: Supervised learning

Needs training labels

Graph features: handcrafted → embeddings → neural nets

Graph neural networks might offer SOTA performance



What approach would you try first?

Agenda

- About Seznam.cz
- Introduction to Graph-based systems
- Graph-based systems at Seznam.cz
- Introduction to Graph-based anomaly detection + hands-on
- Supervised methods + hands-on
- *Break*
- Supervised methods hands-on
- **Unsupervised methods + hands-on**

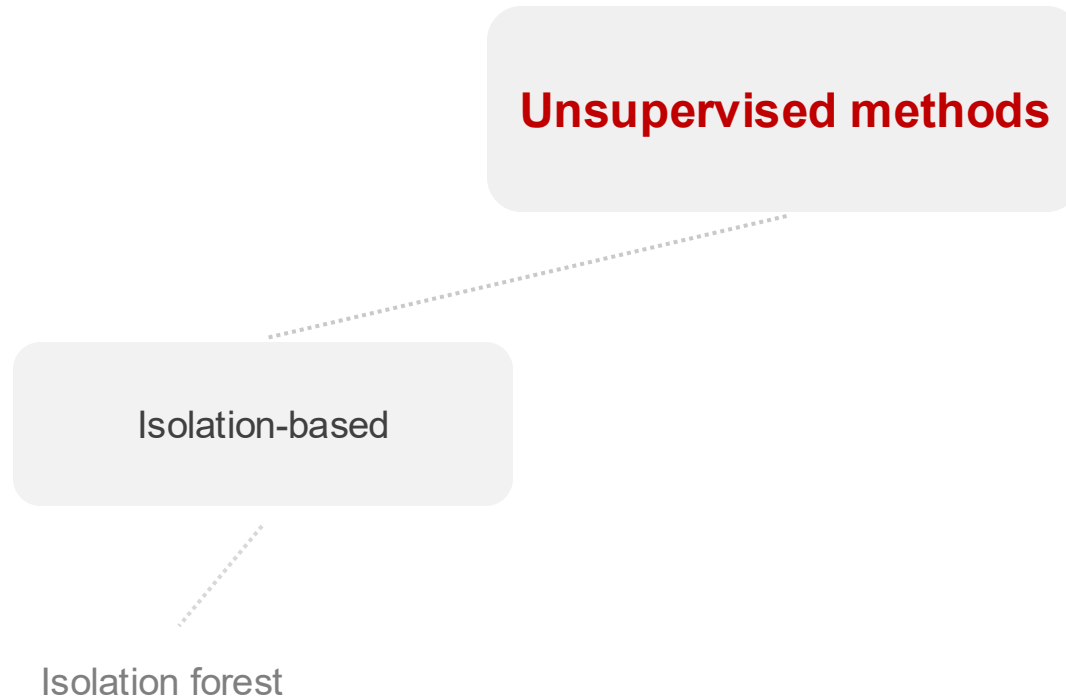
Why unsupervised learning?

Anomalies are:

- rare and expensive to obtain
- diverse and evolving

Unsupervised methods don't need training labels...

Method classification



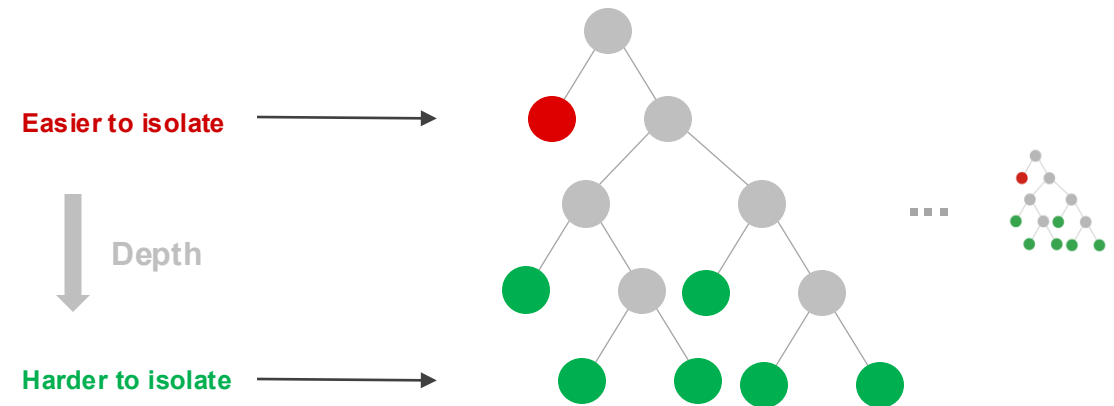
Isolation forest

Tree-based anomaly detector

Idea - anomalies are easier to isolate

How it works:

- **Random partitioning**
- **Path length measurement**
- **Anomaly scoring**



Isolation forest – in graph context

Ignores graph structure

Possible solution:

- Use structural attributes of a nodes as tabular input

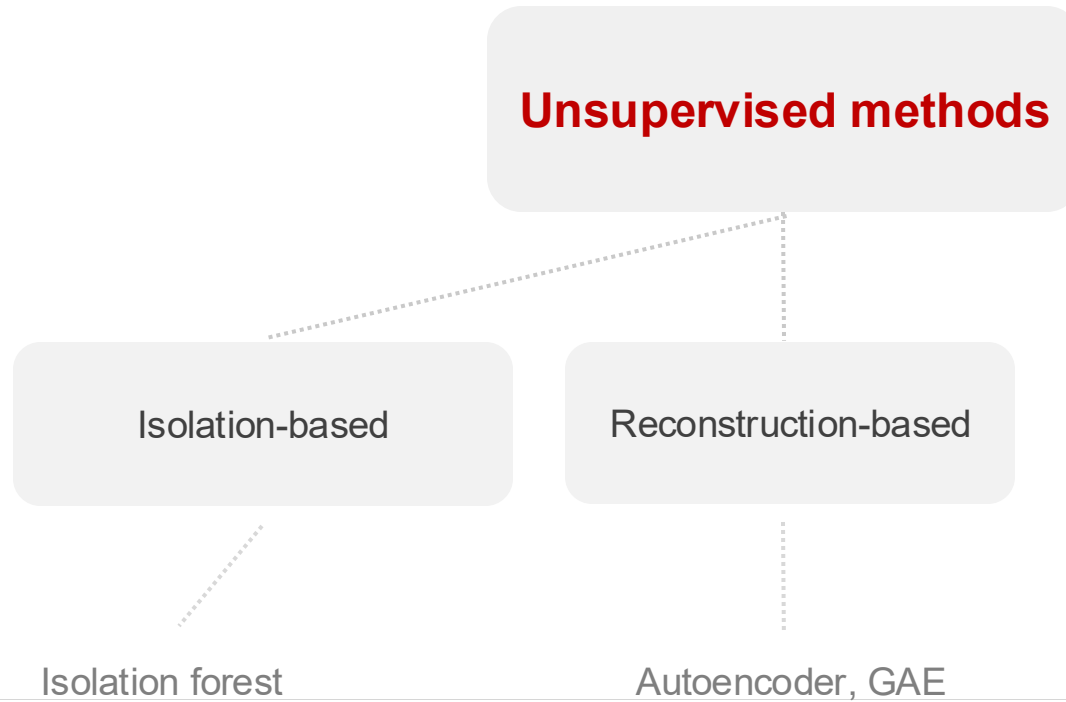
Pros

- Fast
- Scalable

Cons

- Blind to structural anomalies

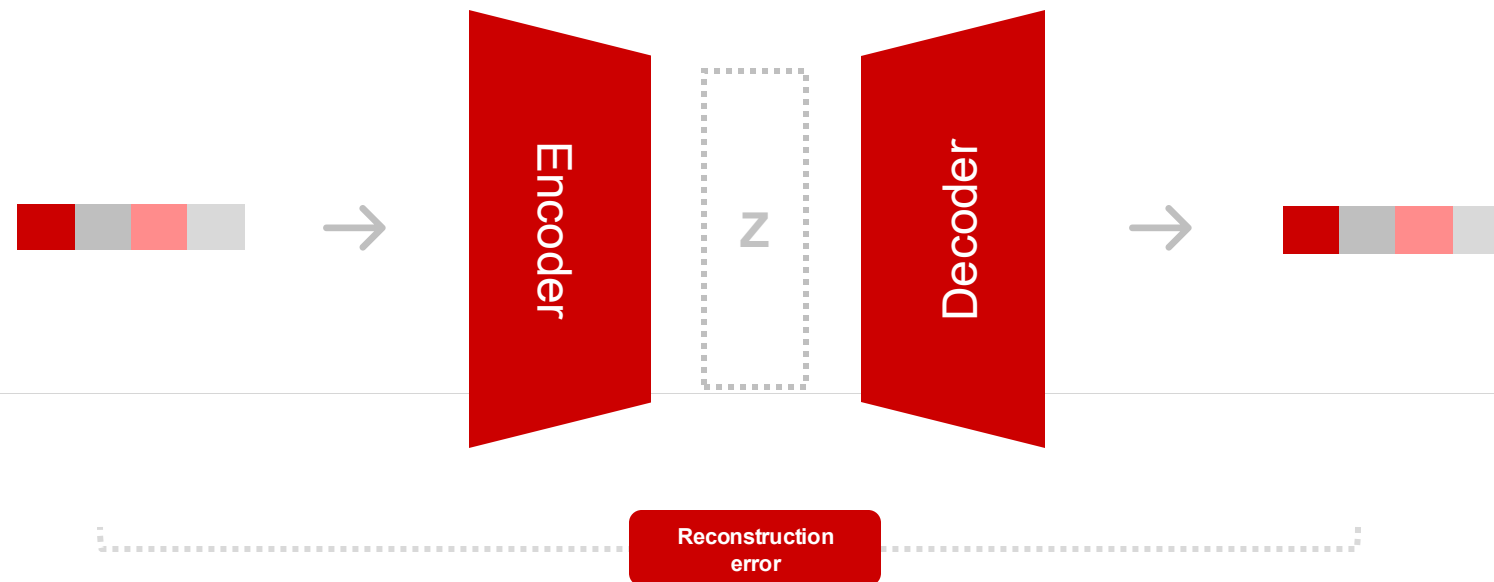
Method classification



Reconstruction based approach

Leveraging autoencoder architecture for anomaly detection

Idea – model learns to compress and reconstruct **normal patterns** well and anomalies will produce higher reconstruction error.

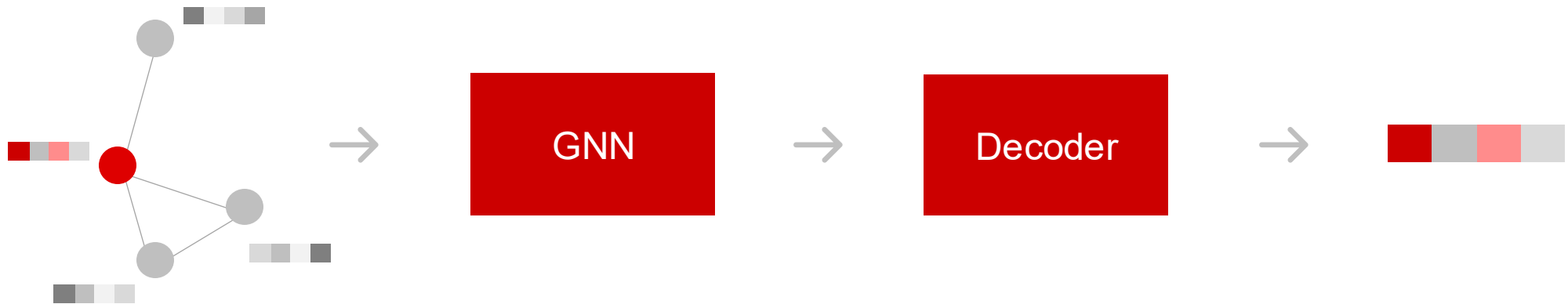


Graph Auto-Encoder

Replaces encoder with GNN

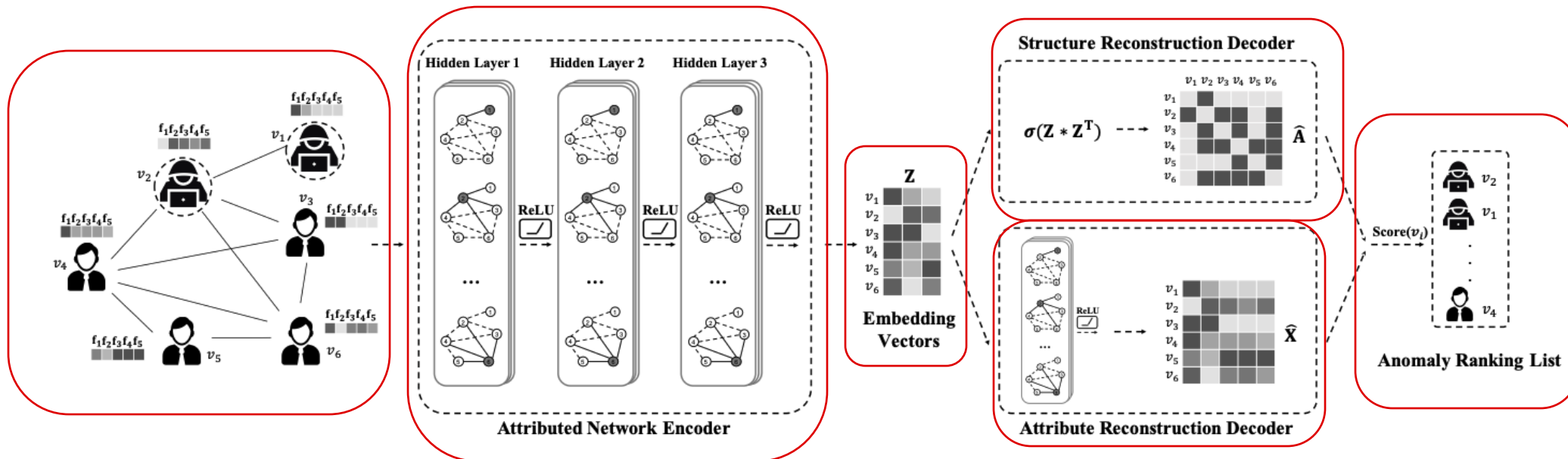
Objective: Reconstructs node features

Alternative objective: reconstruction of adjacency matrix



DOMINANT

Deep Anomaly Detection on Attributed Networks



Reconstruction based approach

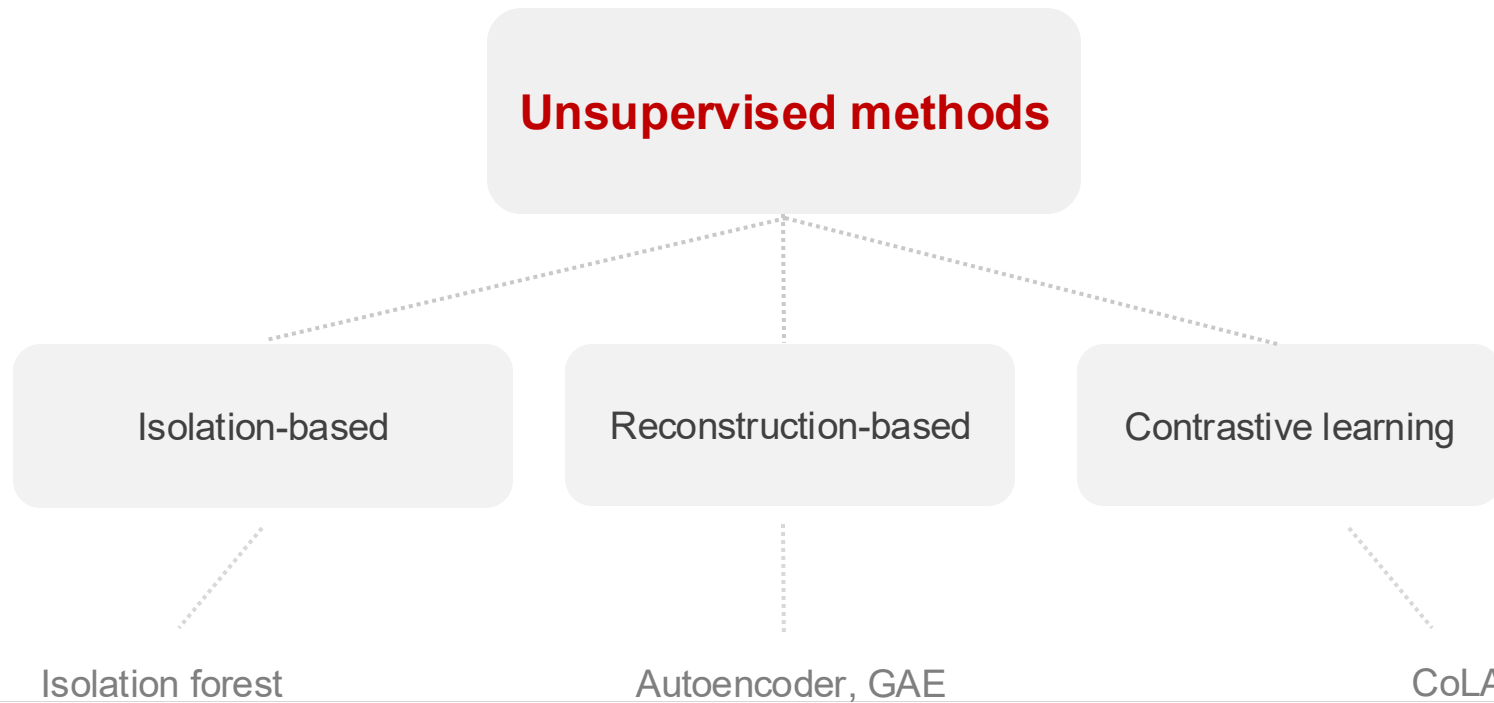
Pros

- Intuitive anomaly score (reconstruction error)
- Captures both feature and structural patterns

Cons

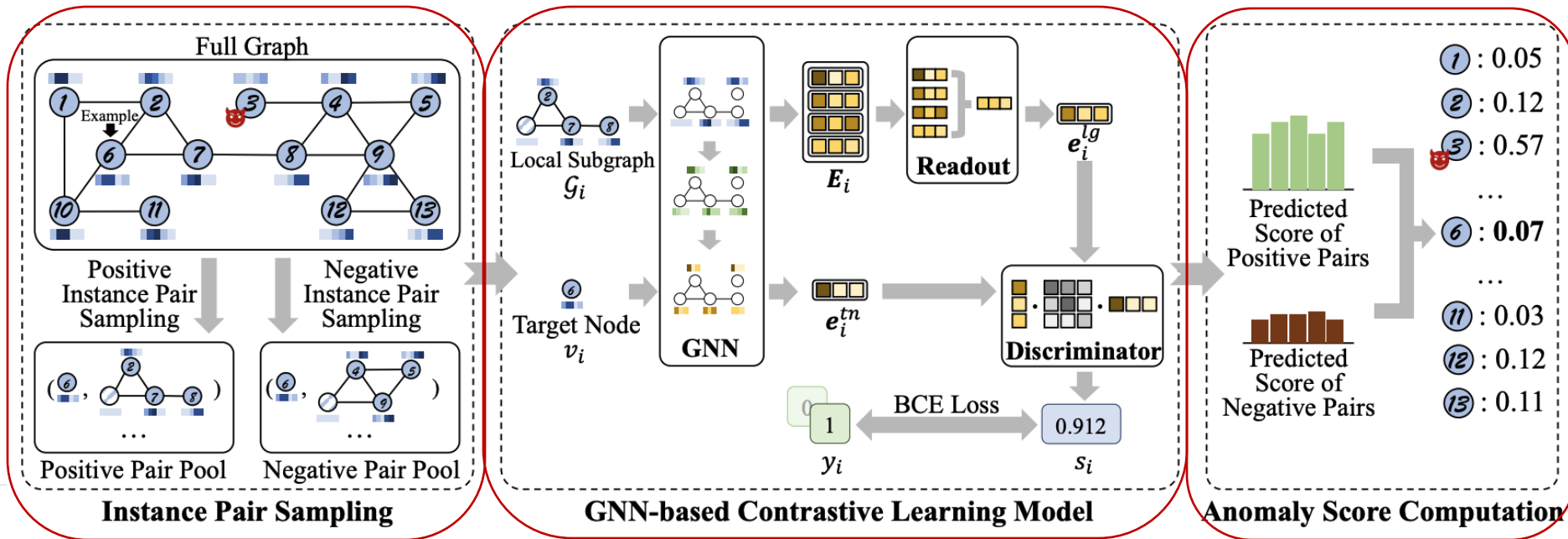
- Can be memory-hungry at scale
- Sensitive to hyperparameters

Method classification



CoLA

Anomaly Detection on Attributed Networks via Contrastive Self-Supervised Learning



Contrastive learning

Pros

- Lightweight

Cons

- Noisy contrasts can hurt performance
- Harder to interpret

Summary: Unsupervised learning

With unsupervised learning, we don't need training labels

Isolation forest is a strong baseline, but ignores graph structure

We can combine autoencoders and GNN to detect anomalies



What are the objectives that GAE can be trained on?

Summary

When possible treat graph data as graphs

Train a supervised anomaly detector if you have labels

Unsupervised detectors can be trained even if labels are unavailable or scarce

You learned how to leverage graphs for anomaly detection

Thank you!

Questions & inquiries
adam.jurcik@firma.seznam.cz

We are hiring!
o-seznam.cz/kariera/vyzkum



Copyright © 1996–2026 Seznam.cz, a. s.



References

Motivation:

<https://www.icij.org/investigations/fincen-files/mining-sars-data/>

Graph machine learning:

<https://snap.stanford.edu/node2vec/>

<https://snap.stanford.edu/graphsage/>

Graph anomaly detection:

<https://github.com/squareRoot3/GADBench>

<https://docs.pygod.org/en/latest/>

Workshop GitHub repository:

<https://github.com/seznam/MLPrague-2026>

Seznam IT positions

Machine learning:

[In-house LLM](#)

[NLP application](#)

[Fulltext search](#)

Engineering:

[In-house LLM](#)

[In-house social network](#)

[AI for Mapy.com](#)

Product management:

[In-house LLM product manager](#)

[In-house LLM junior product manager](#)

Feedback

Please, leave us your feedback: <https://forms.gle/8CPVcrC6vJNkzgMBA>

