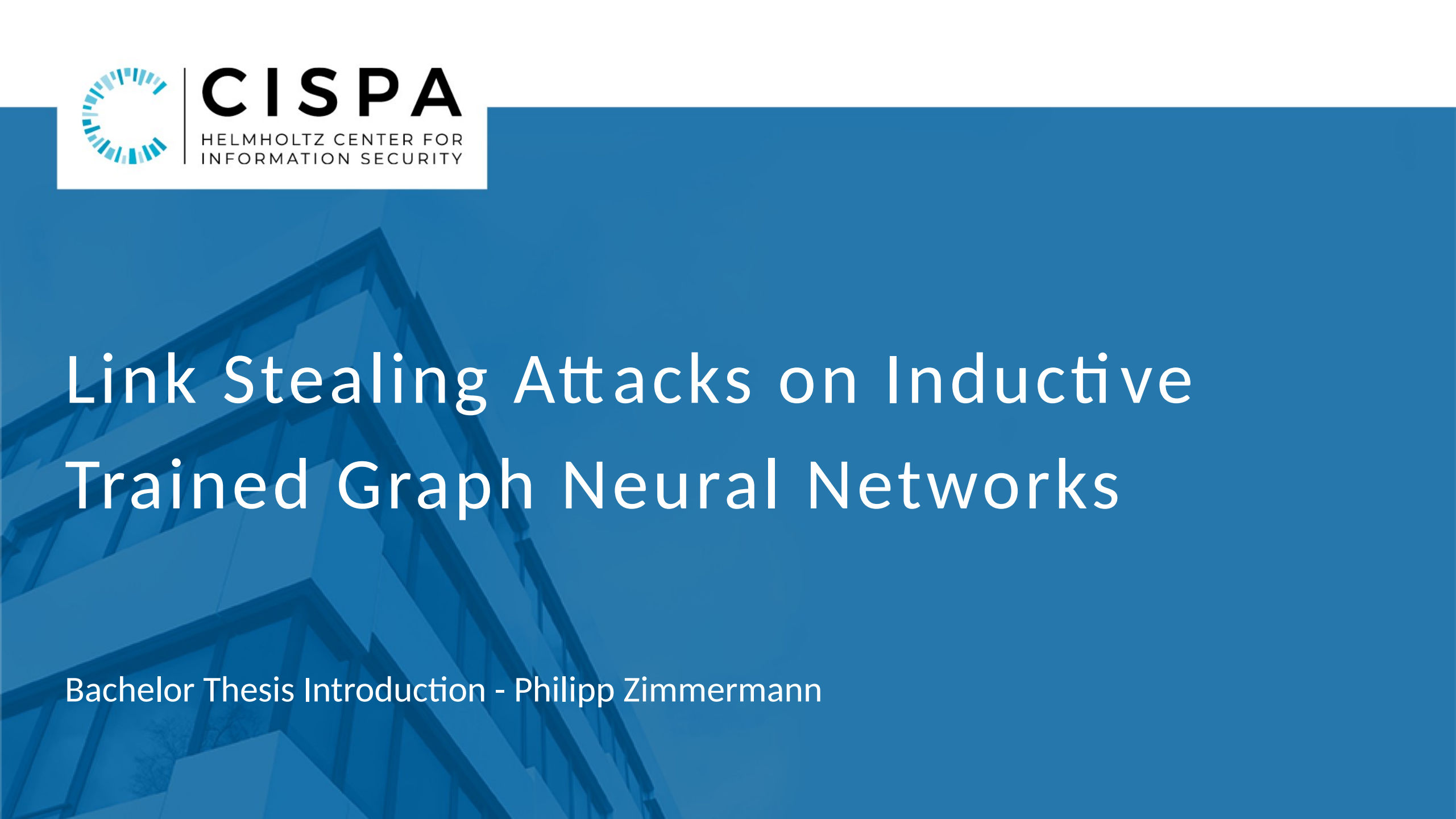


Link Stealing Attacks on Inductive Trained Graph Neural Networks

Bachelor Thesis Introduction - Philipp Zimmermann

Outline

- Graphs
- Graph Neural Networks
- Our Approach: Link Stealing Attacks on Inductive Trained Graph Neural Networks
- Experimental Setup
- Goal

Graphs

Graphs

- Data Structure
  - Model large data and relationships between entities
  - Nodes with features
  - Edges

Graphs

- Data Structure
  - Model large data and relationships between entities
  - Nodes with features
  - Edges

- Chemical Networks
  - Protein-protein interactions

Graphs

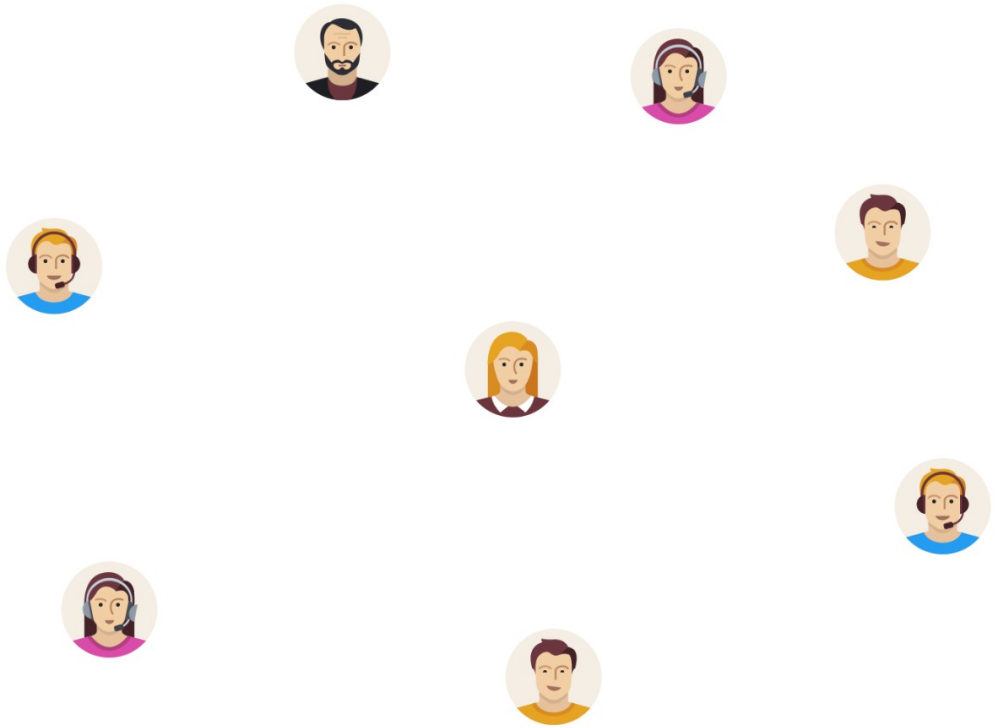
- Data Structure
  - Model large data and relationships between entities
  - Nodes with features
  - Edges

- Chemical Networks
  - Protein-protein interactions

- Social Networks
  - Instagram
  - Facebook
  - Twitter

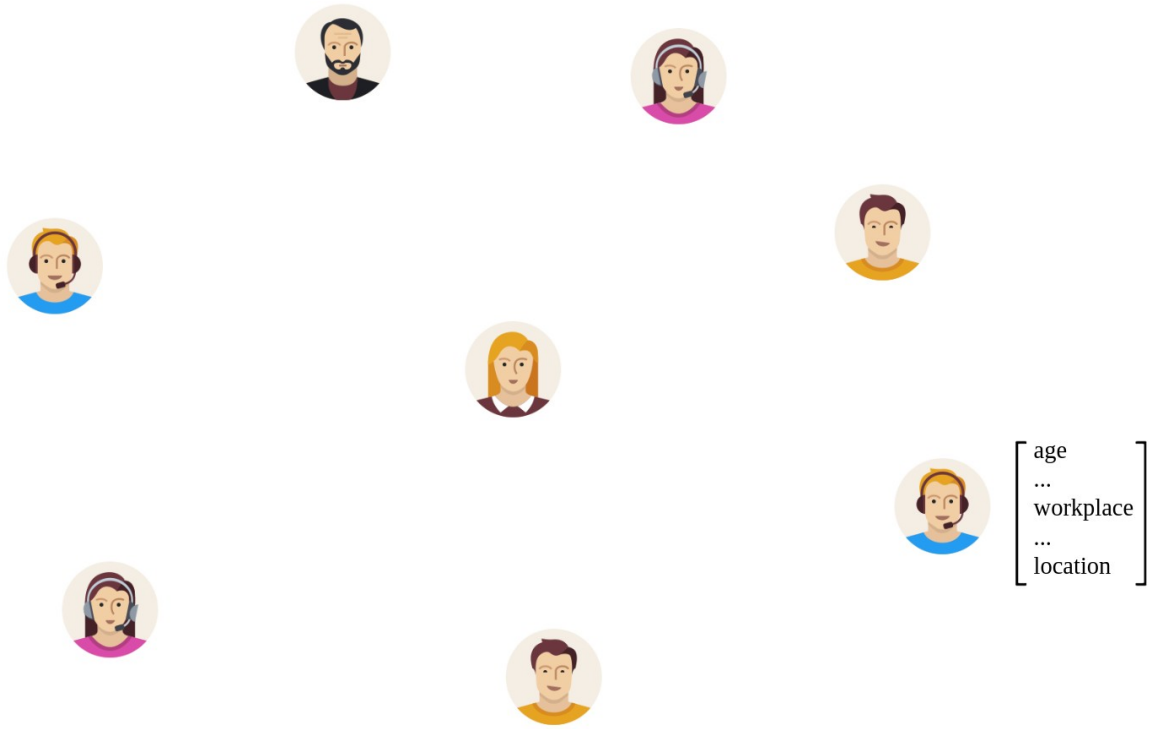
Graphs

- Data Structure
  - Model large data and relationships between entities
  - Nodes with features
  - Edges
- Chemical Networks
  - Protein-protein interactions
- Social Networks
  - Instagram
  - Facebook
  - Twitter



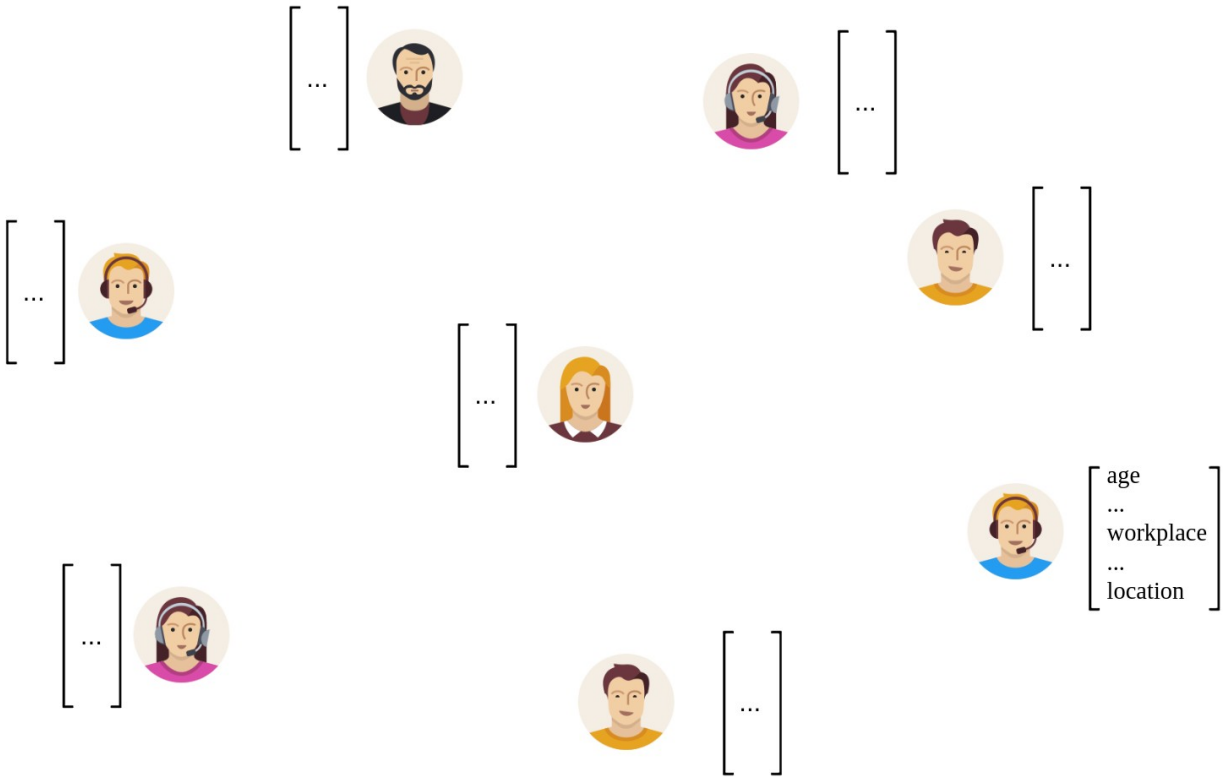
Graphs

- Data Structure
  - Model large data and relationships between entities
  - Nodes with features
  - Edges
- Chemical Networks
  - Protein-protein interactions
- Social Networks
  - Instagram
  - Facebook
  - Twitter



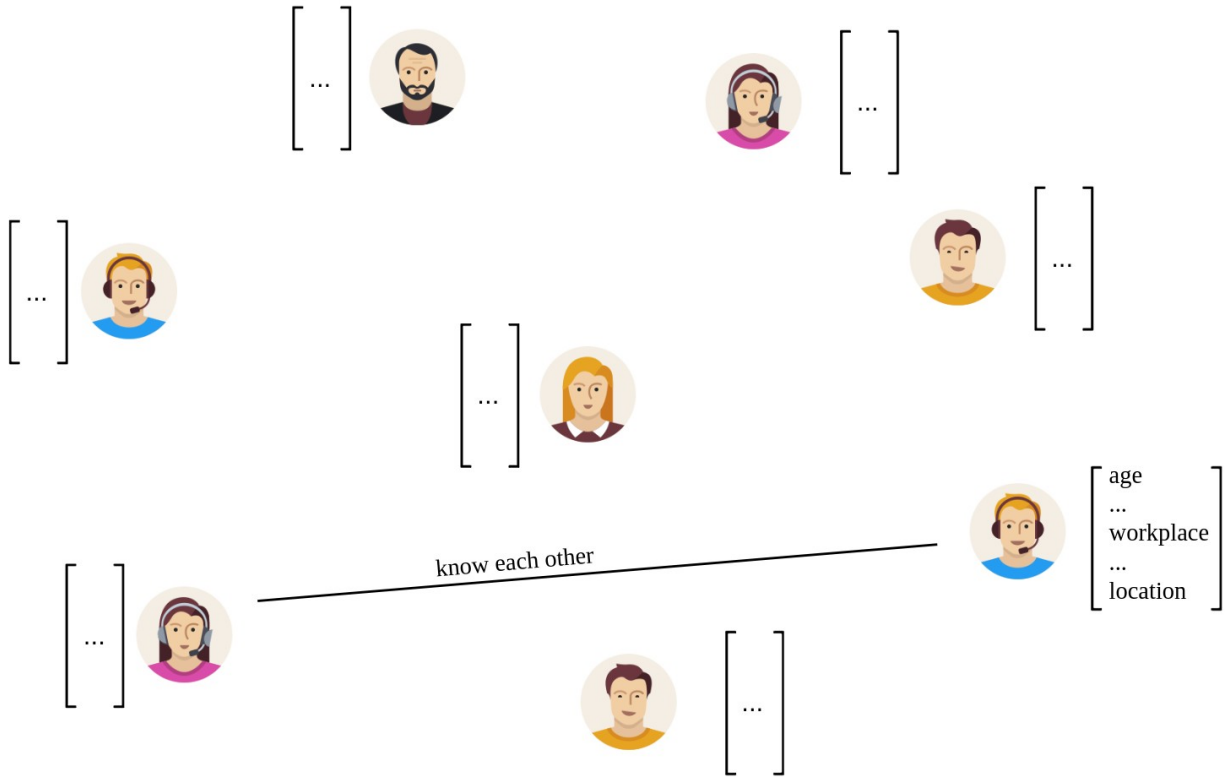
Graphs

- Data Structure
  - Model large data and relationships between entities
  - Nodes with features
  - Edges
- Chemical Networks
  - Protein-protein interactions
- Social Networks
  - Instagram
  - Facebook
  - Twitter



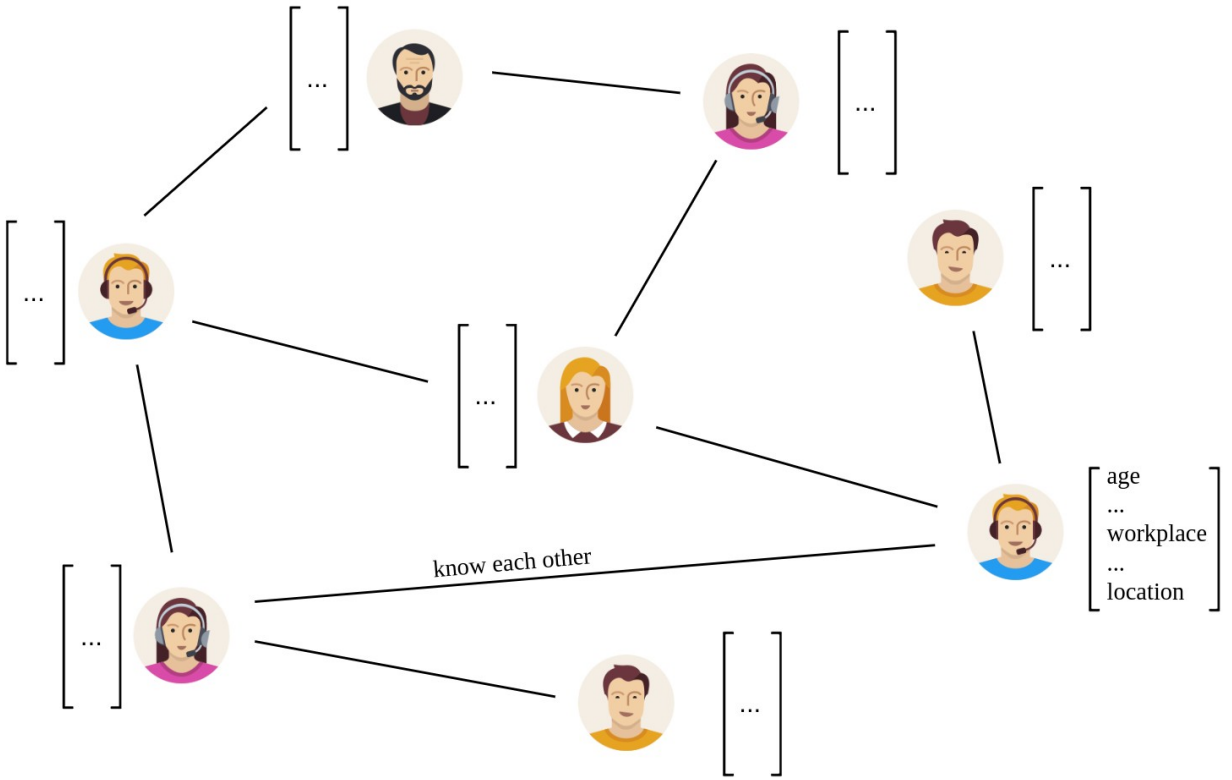
Graphs

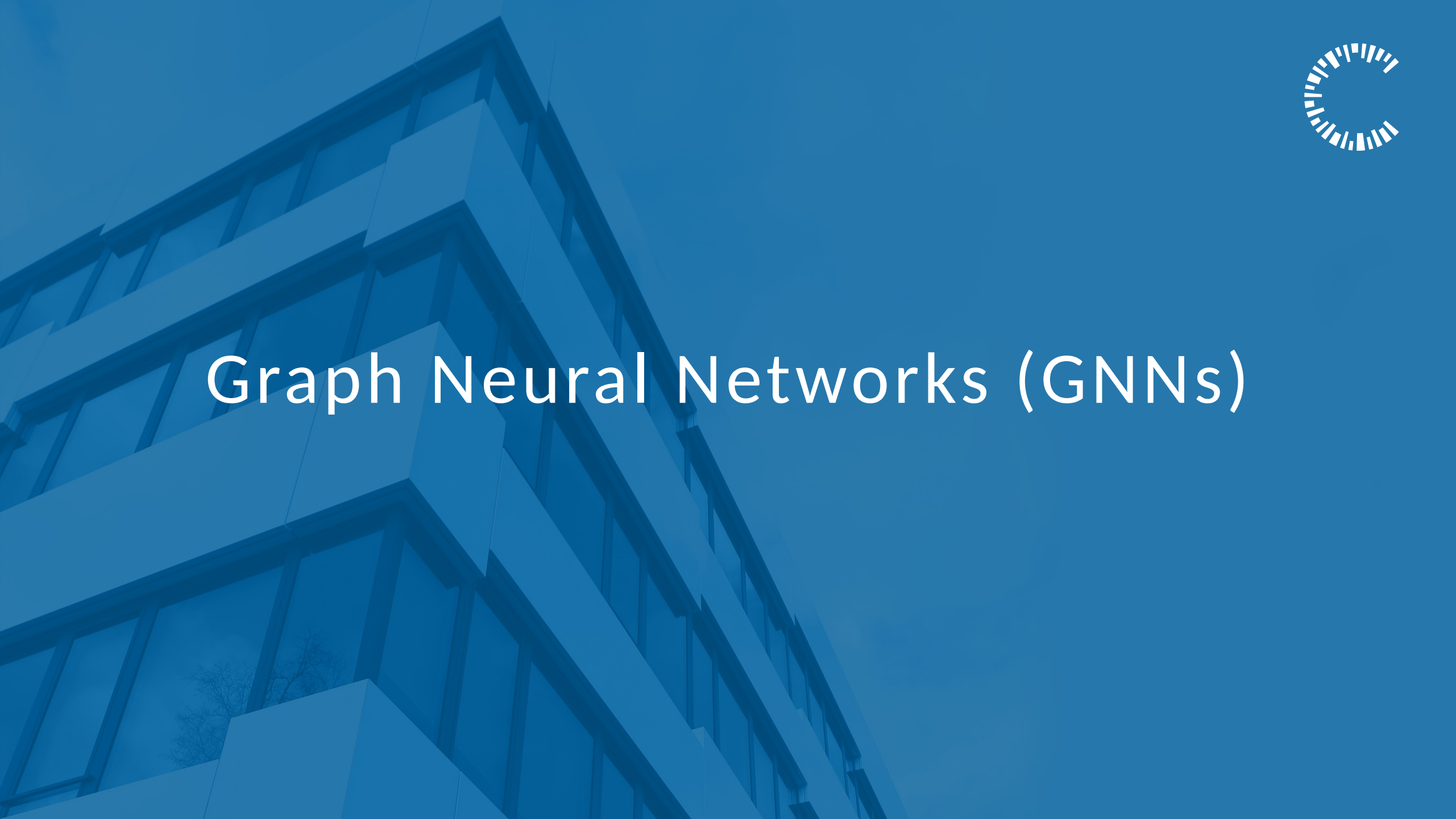
- Data Structure
  - Model large data and relationships between entities
  - Nodes with features
  - Edges
- Chemical Networks
  - Protein-protein interactions
- Social Networks
  - Instagram
  - Facebook
  - Twitter



Graphs

- Data Structure
  - Model large data and relationships between entities
  - Nodes with features
  - Edges
- Chemical Networks
  - Protein-protein interactions
- Social Networks
  - Instagram
  - Facebook
  - Twitter





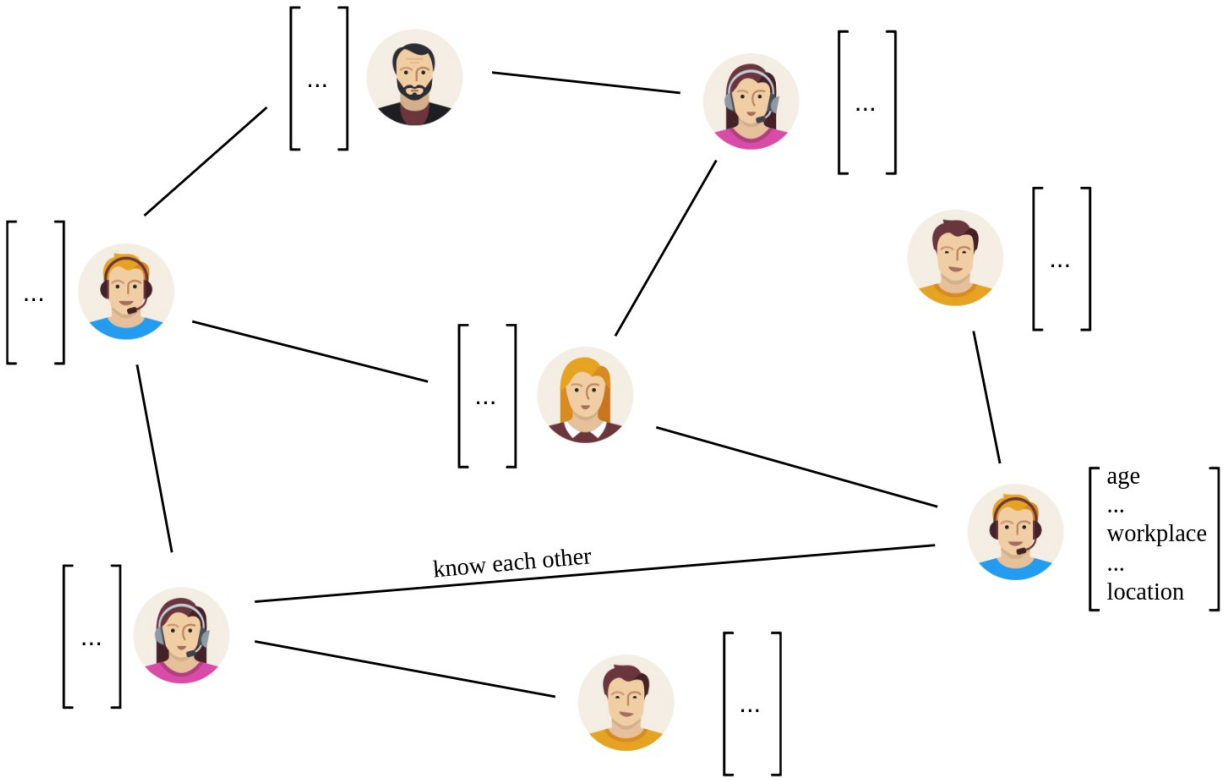
Graph Neural Networks (GNNs)

Graph Neural Networks

- Machine Learning Model over Graphs

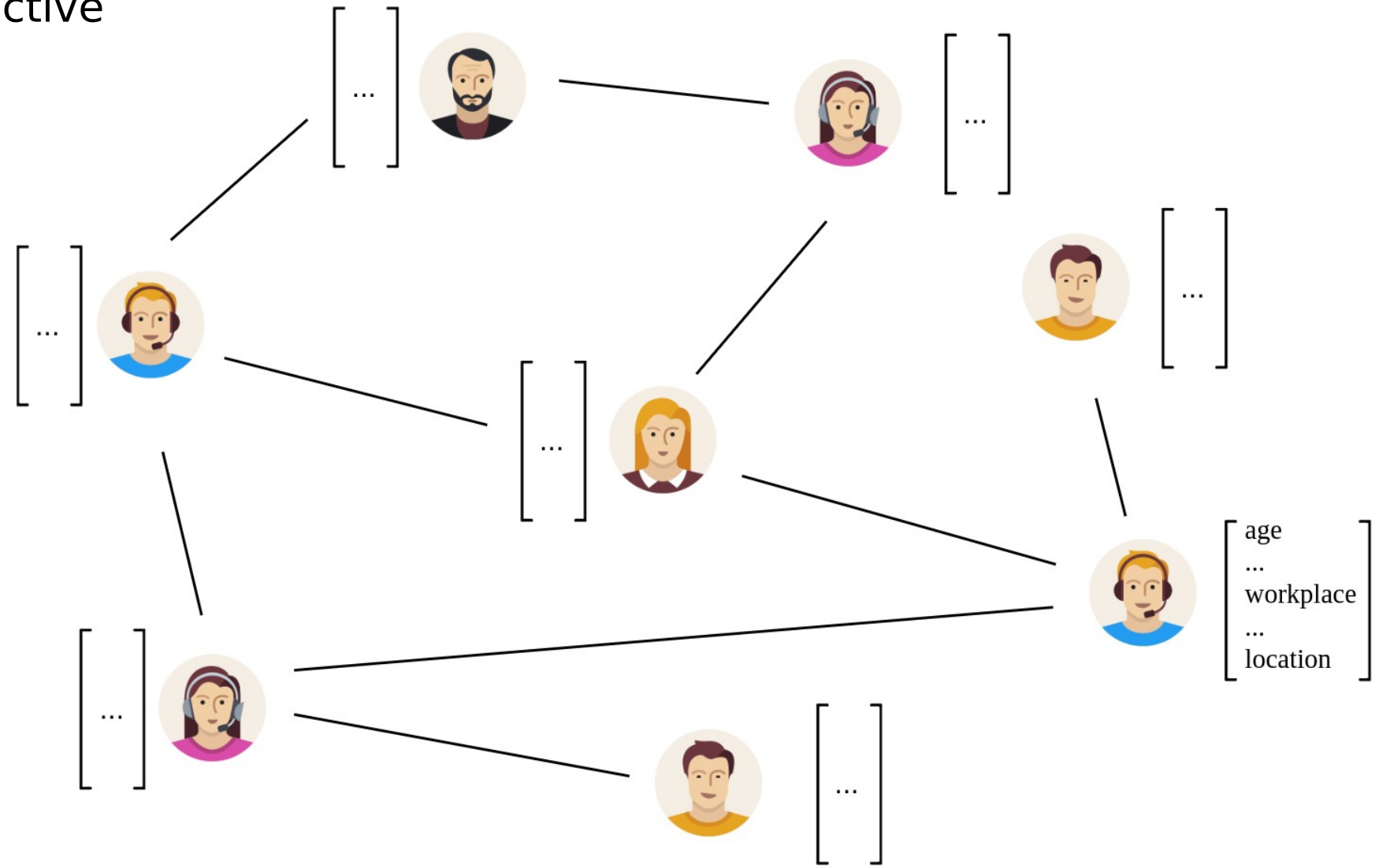
- Different Tasks
  - Node classification
  - Graph classification
  - Link prediction

- Different Learning Methods
  - Transductive
  - Inductive



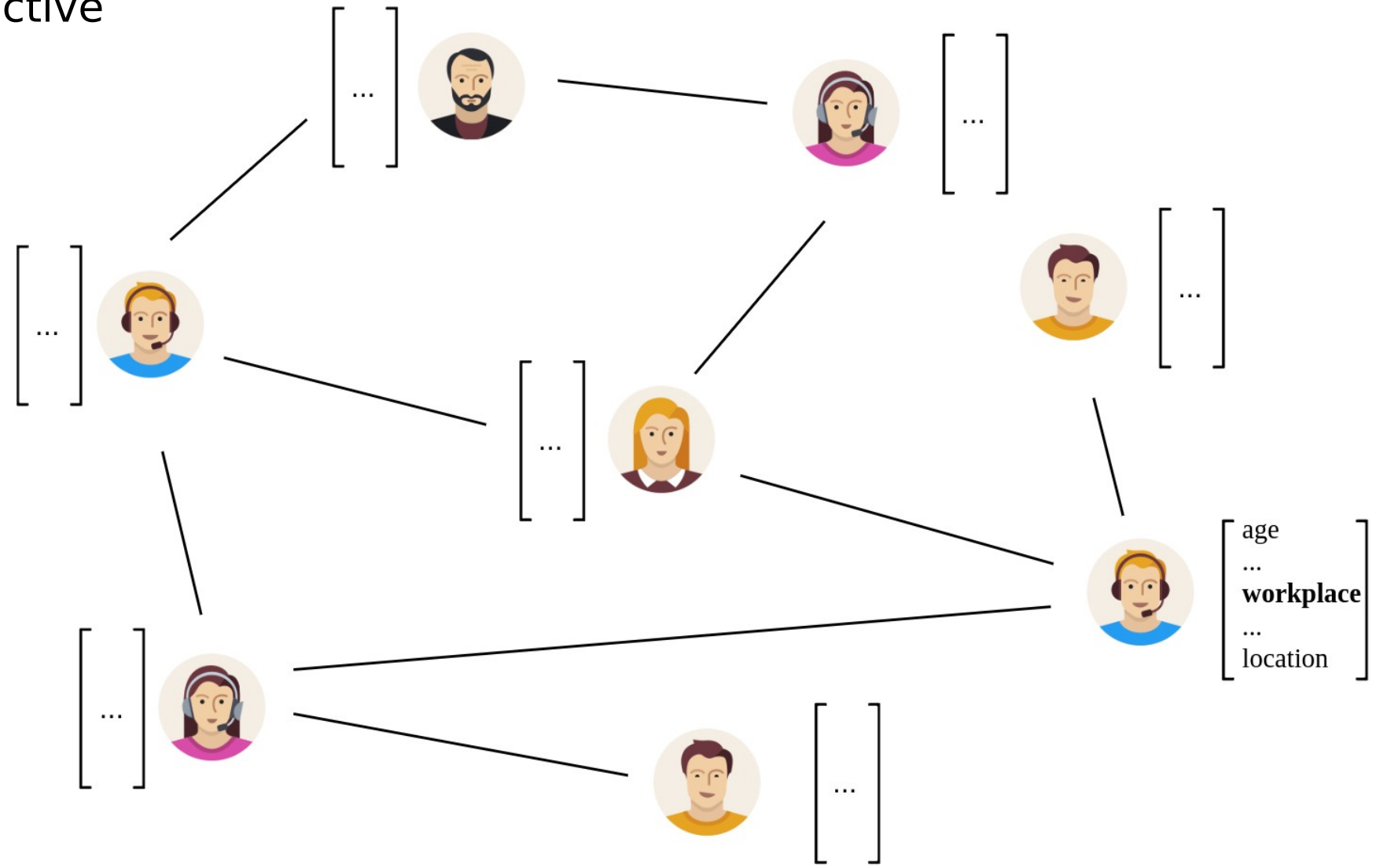
Graph Neural Networks

- Transductive



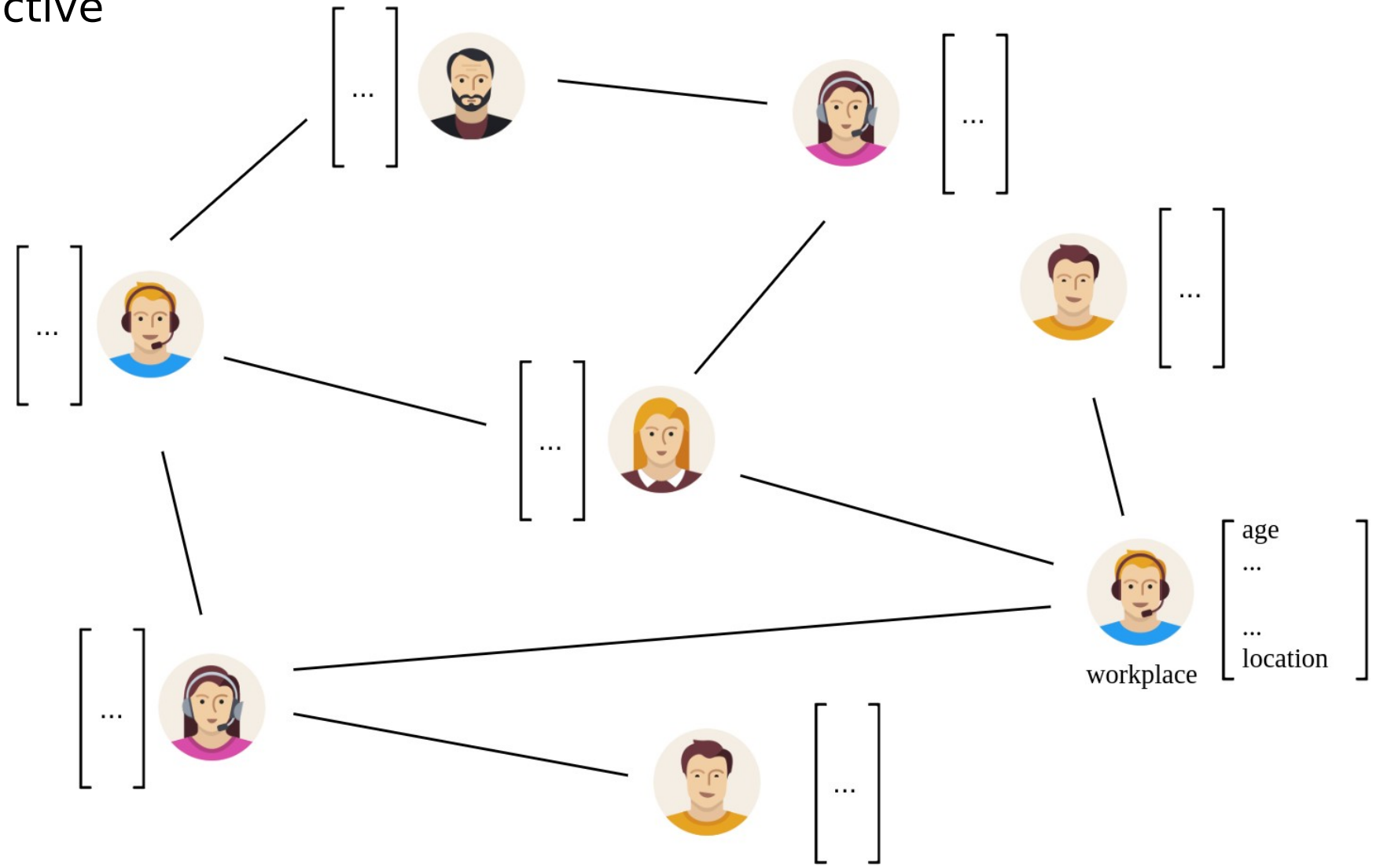
Graph Neural Networks

- Transductive



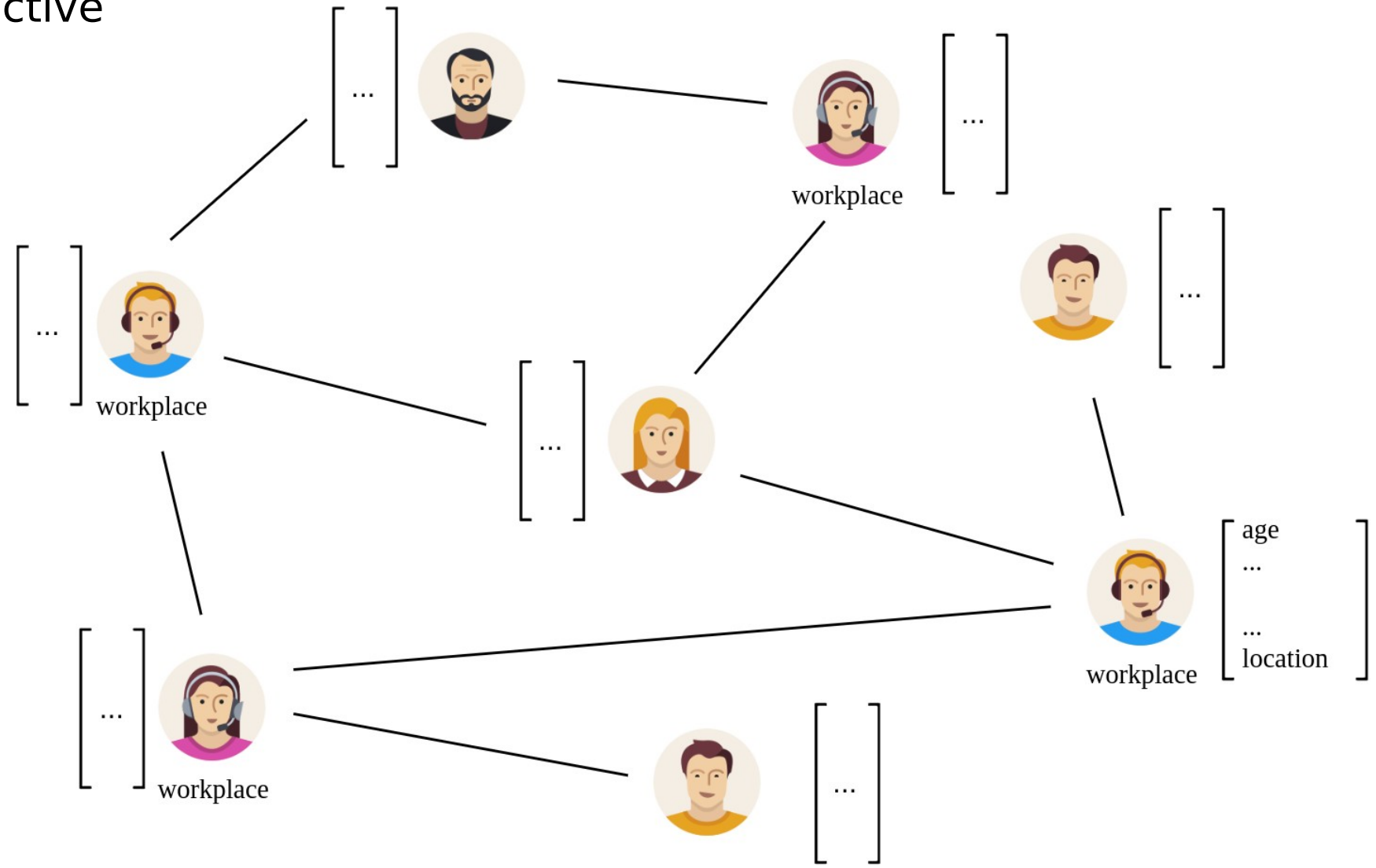
Graph Neural Networks

- Transductive



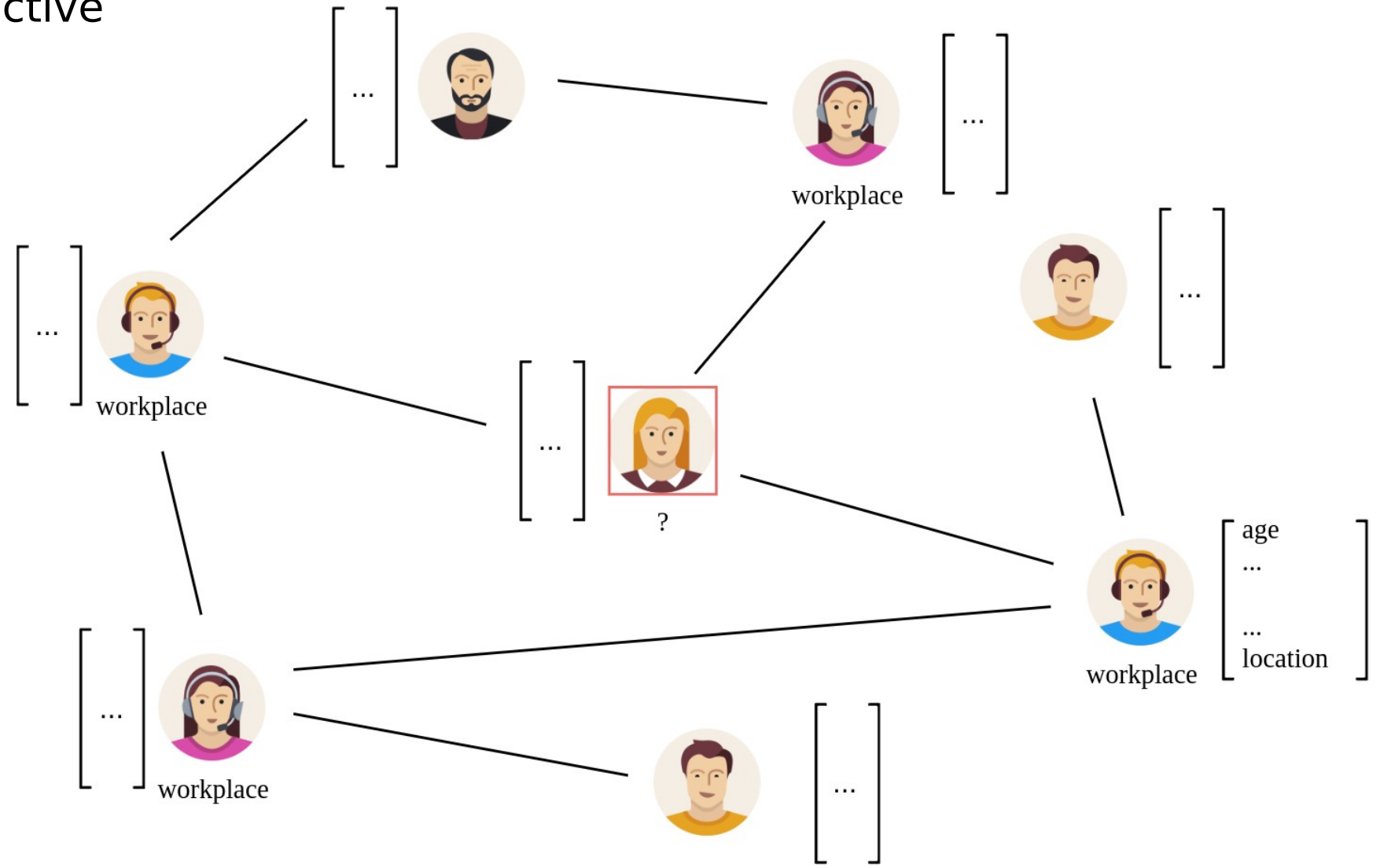
Graph Neural Networks

- Transductive



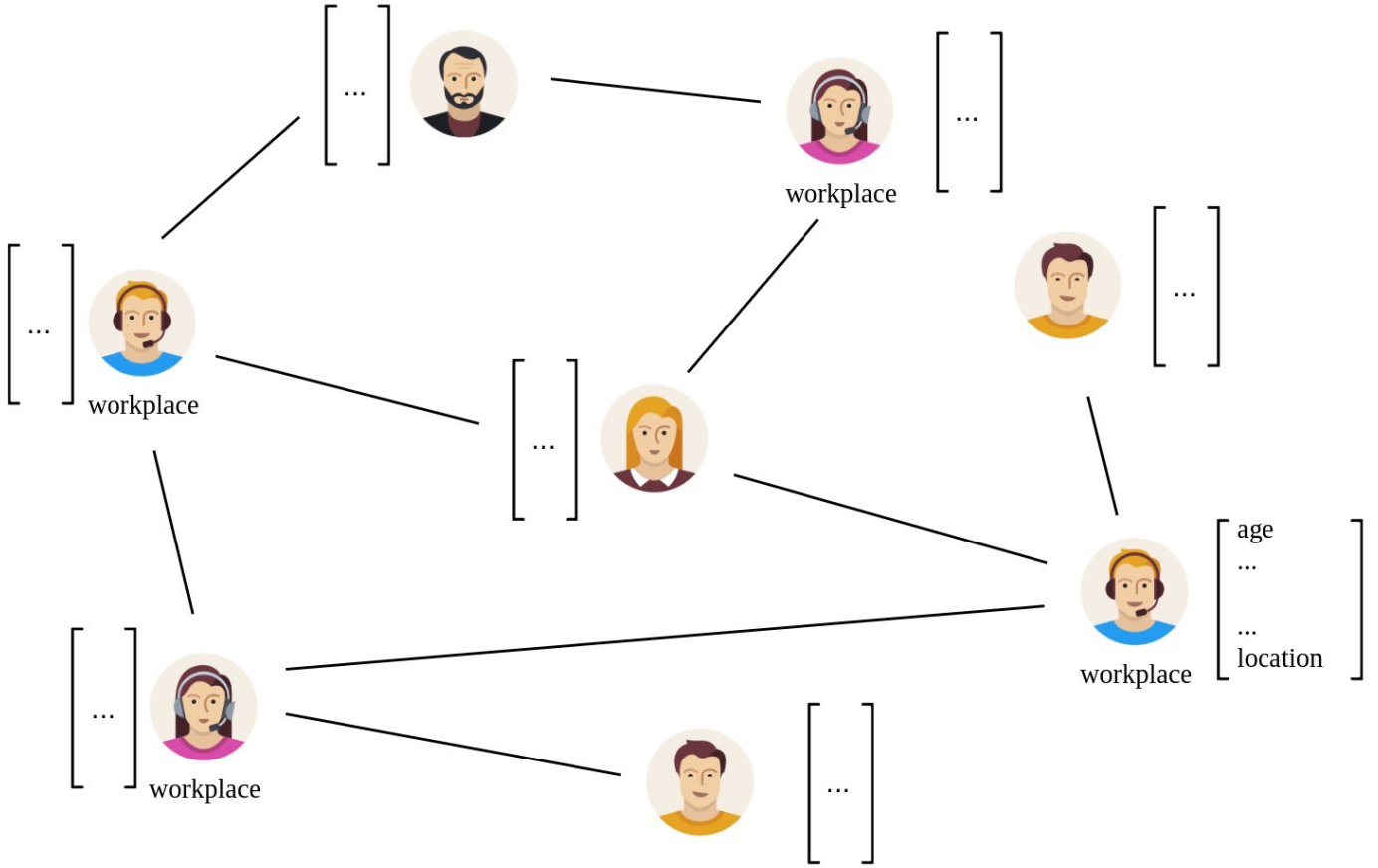
Graph Neural Networks

- Transductive



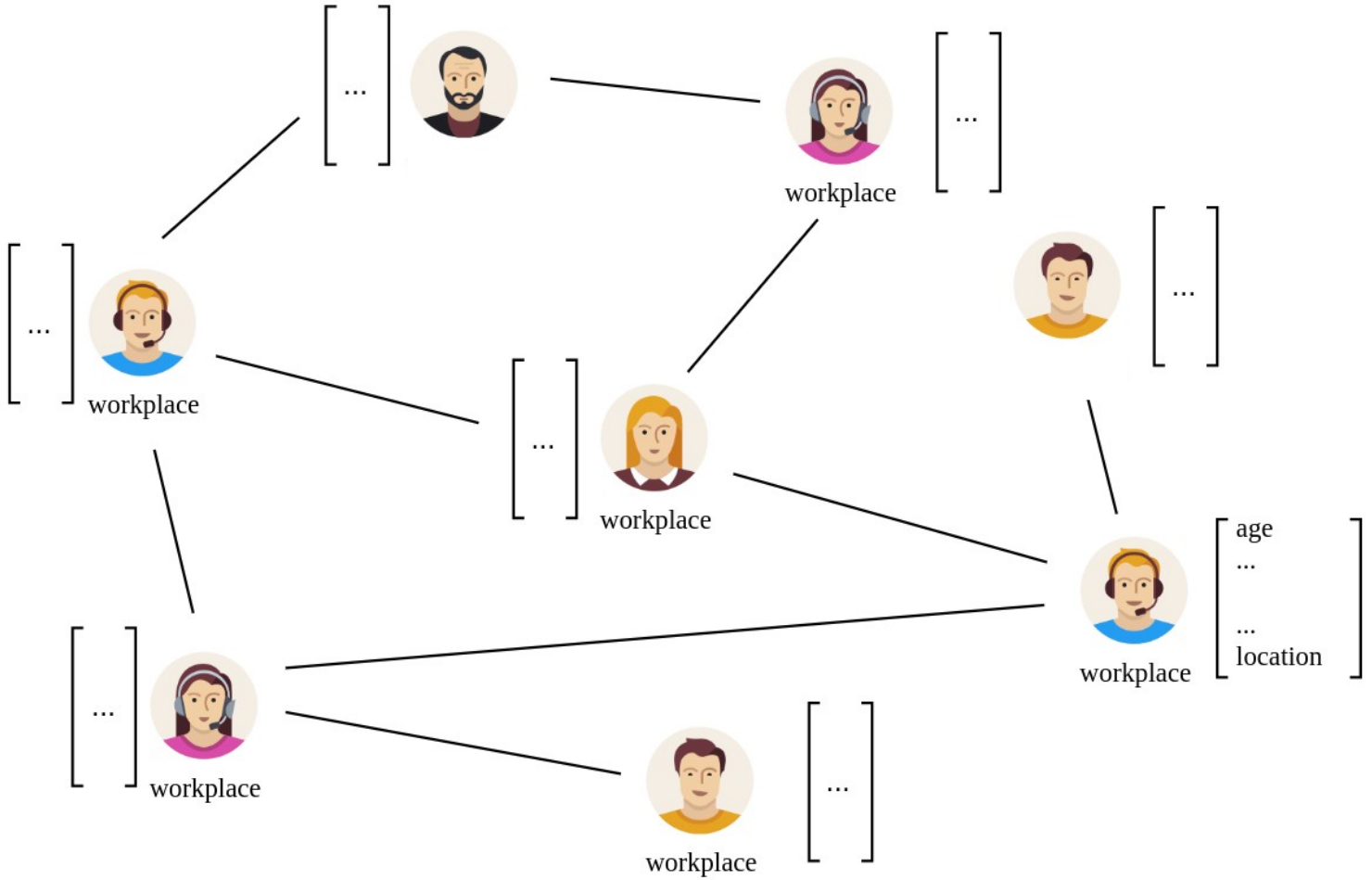
Graph Neural Networks

- Transductive
  - Fix graph (features and link)
  - Some nodes' labels are missing
  - Limited scenario



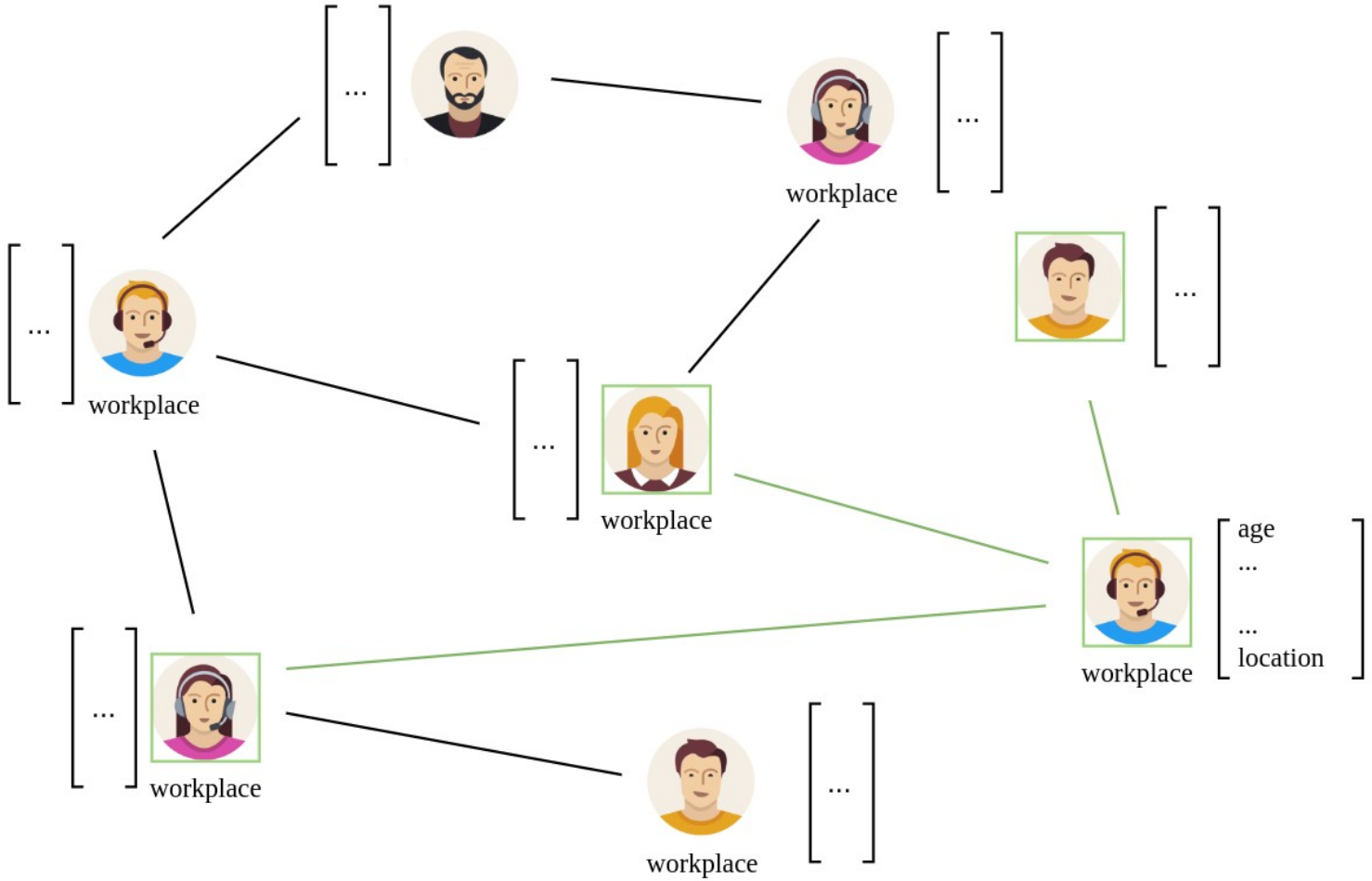
Graph Neural Networks

- Inductive



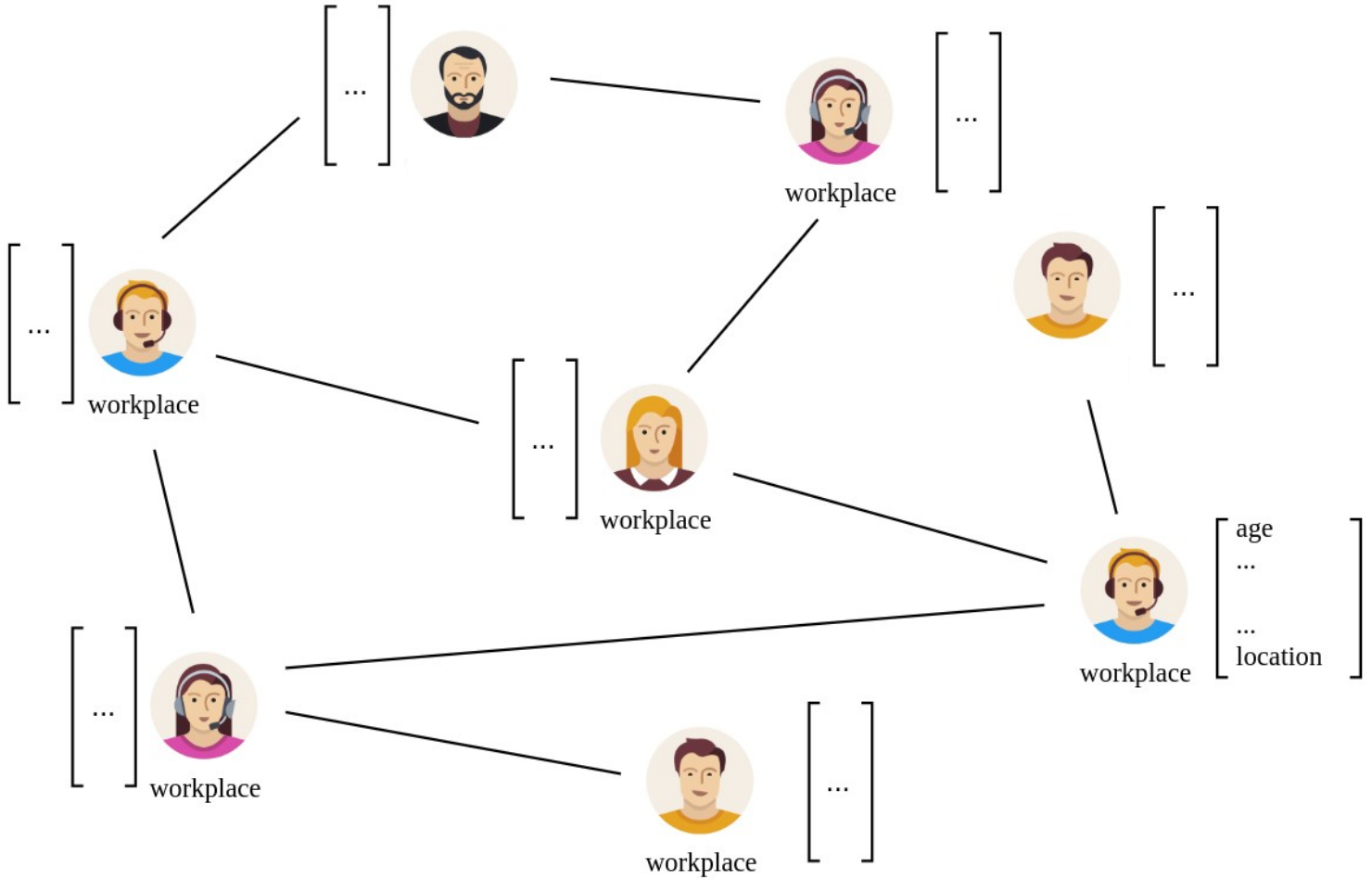
Graph Neural Networks

- Inductive



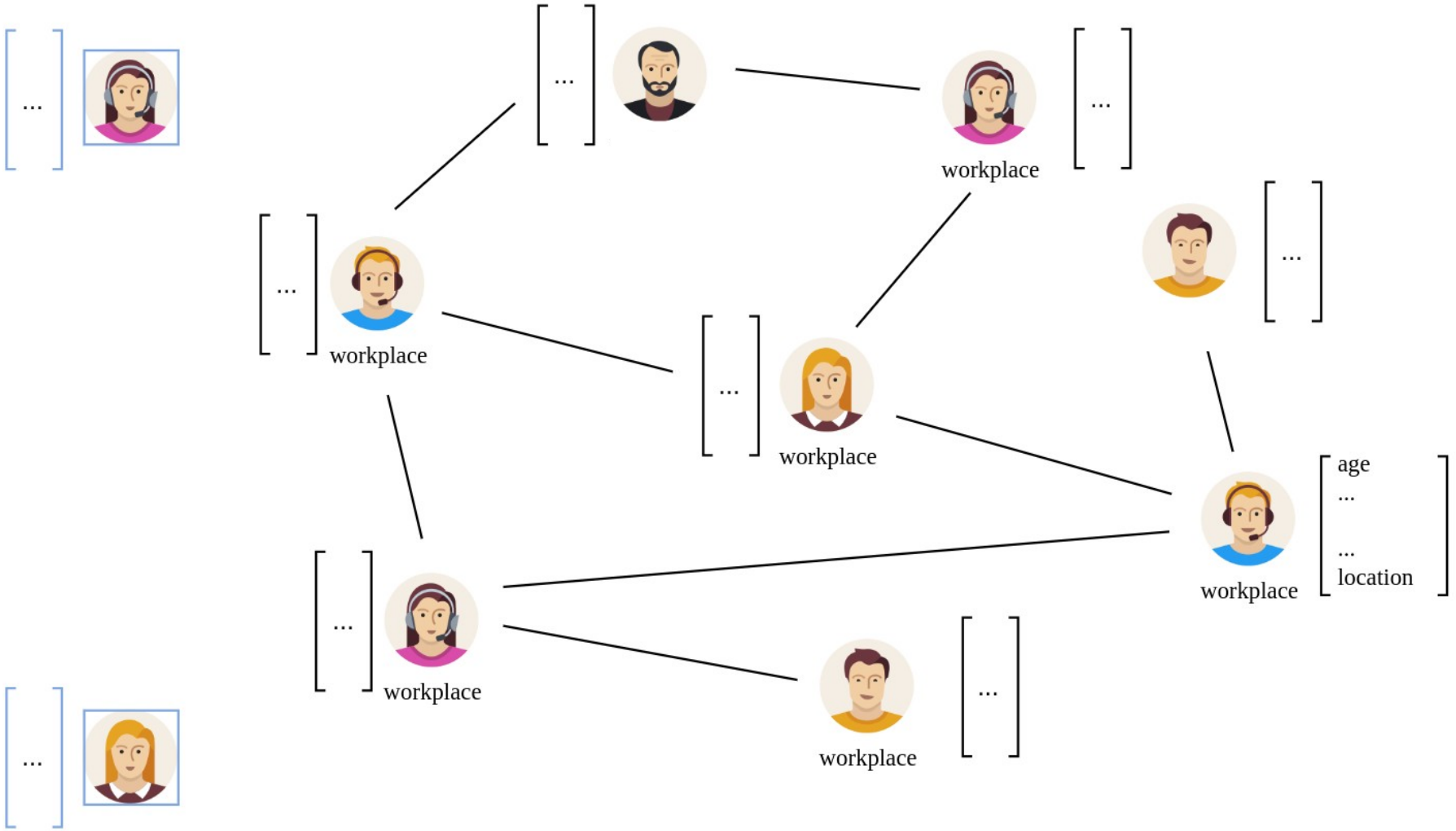
Graph Neural Networks

- Inductive



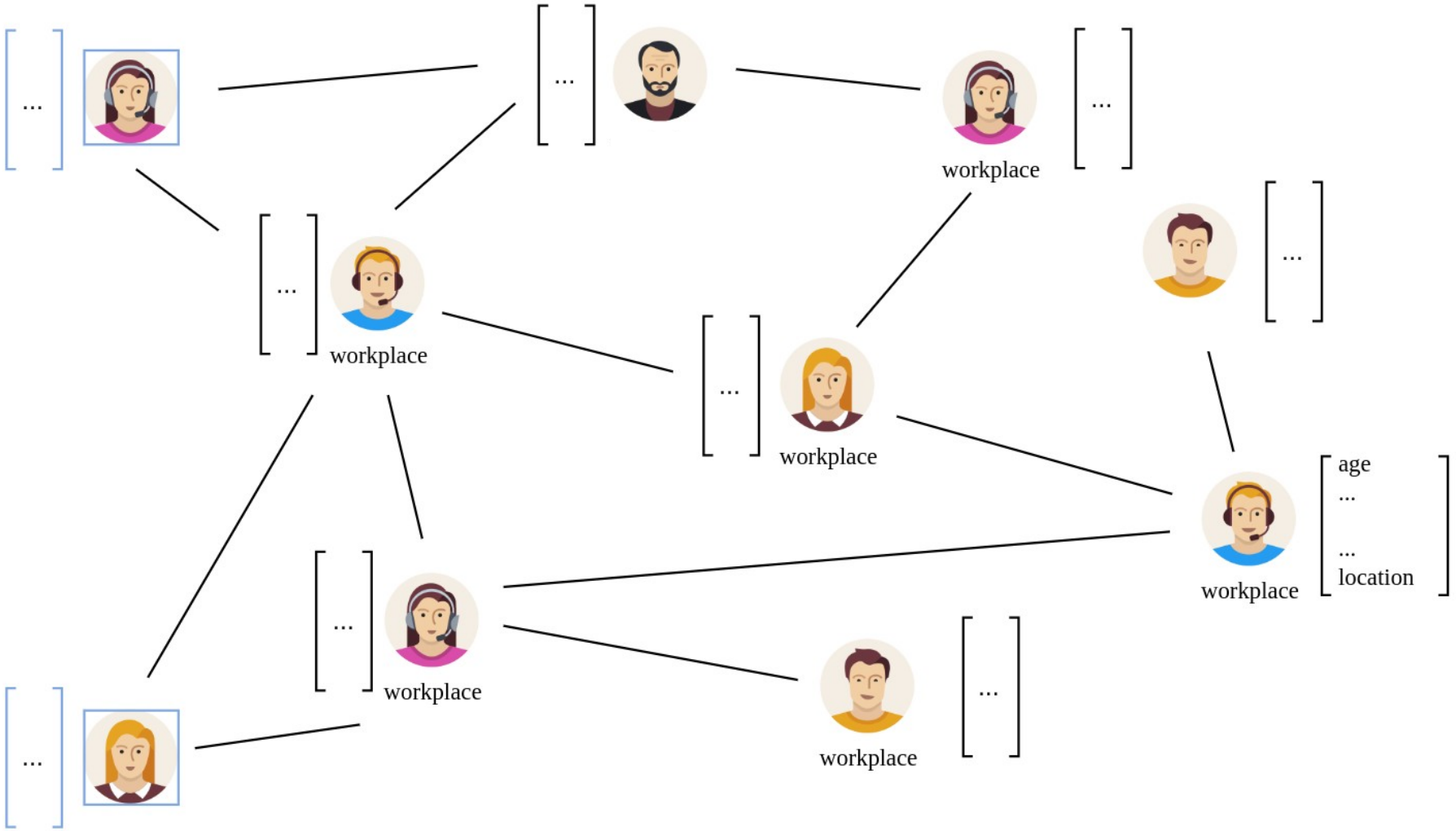
Graph Neural Networks

- Inductive



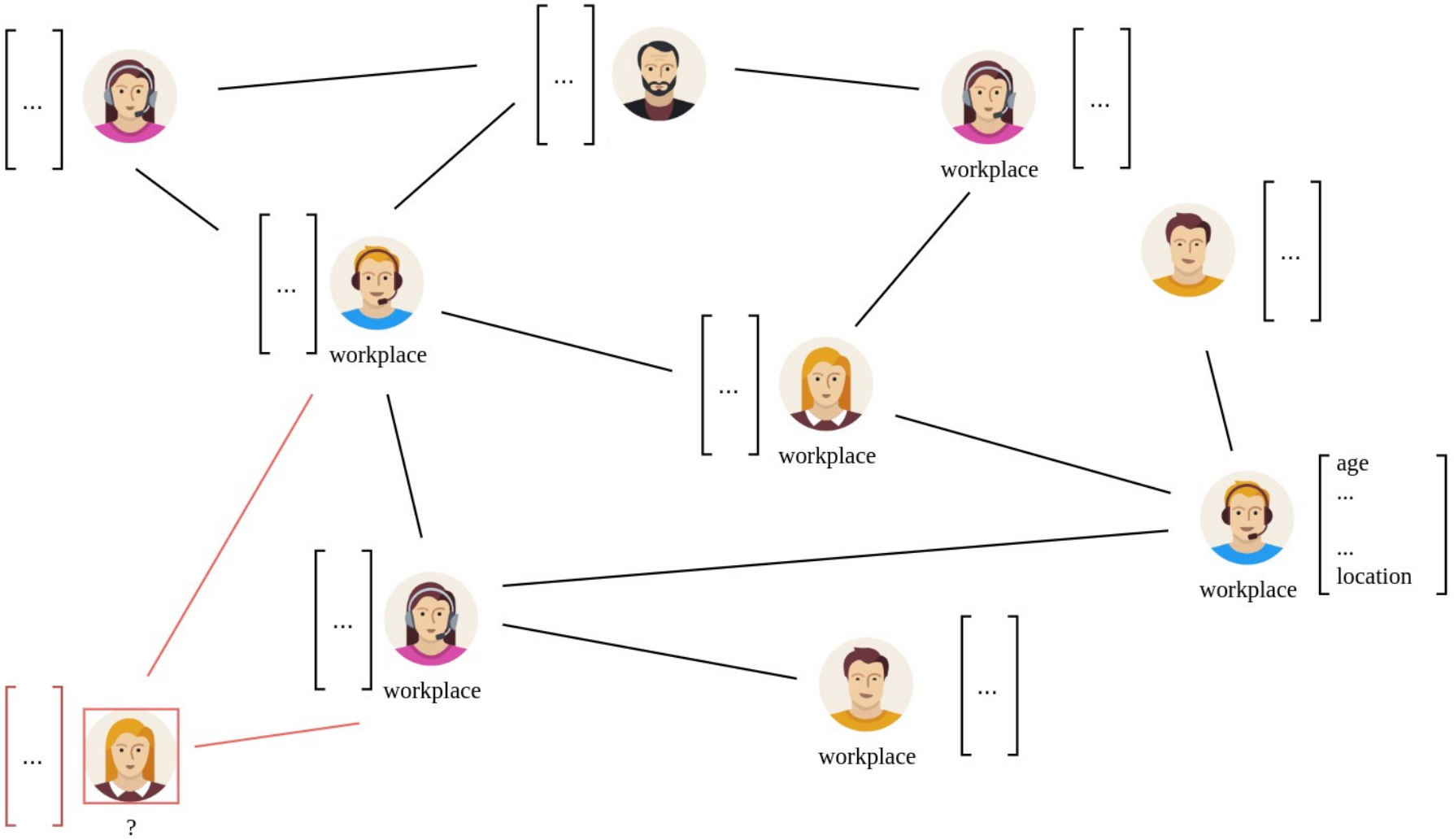
Graph Neural Networks

- Inductive



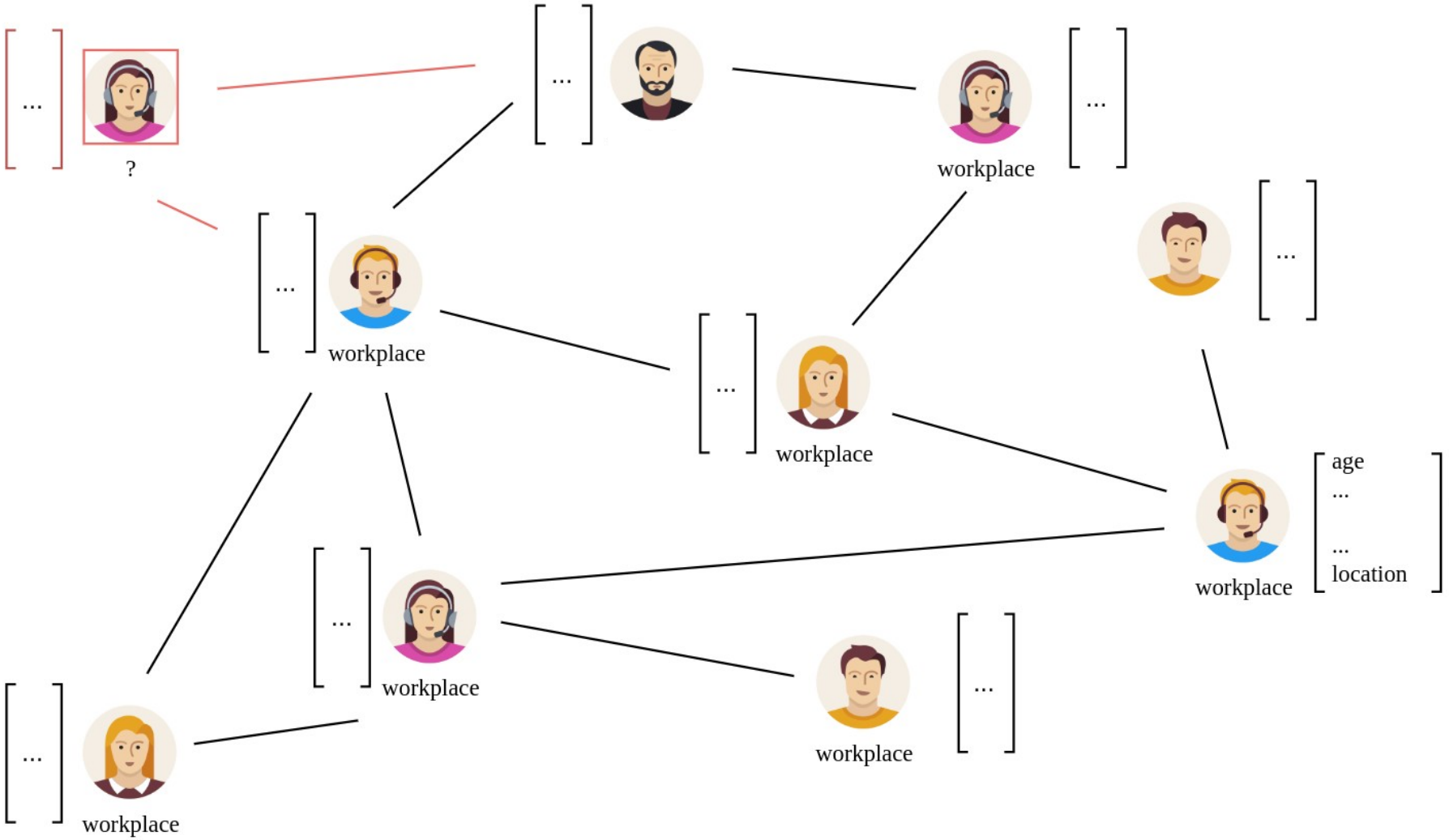
Graph Neural Networks

- Inductive



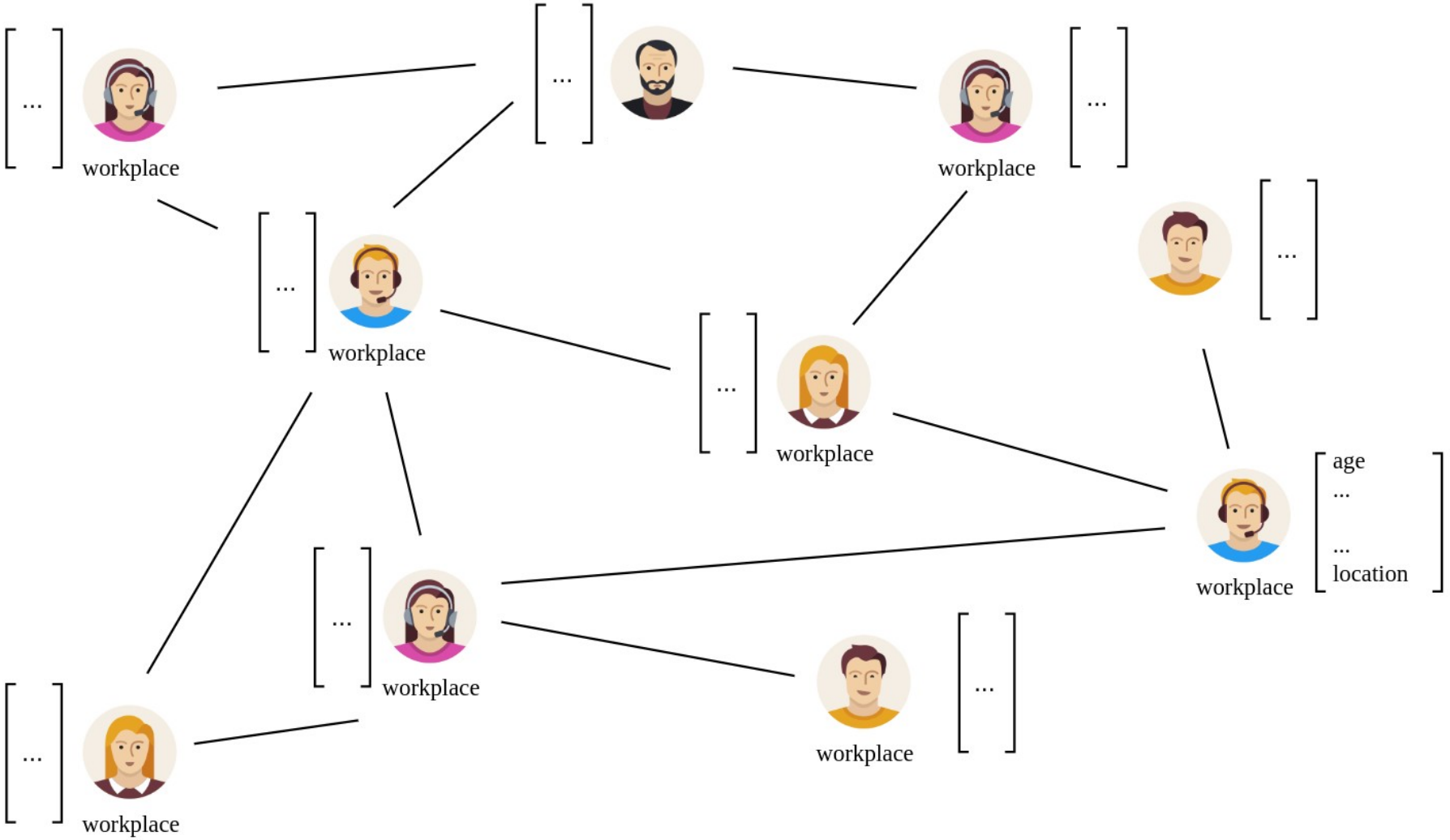
Graph Neural Networks

- Inductive



Graph Neural Networks

- Inductive



Graph Neural Networks

- Inductive
  - Extends transductive setting
  - Able to generalize to unseen nodes
  - Unnecessary to retrain the model
  - Broader scenario

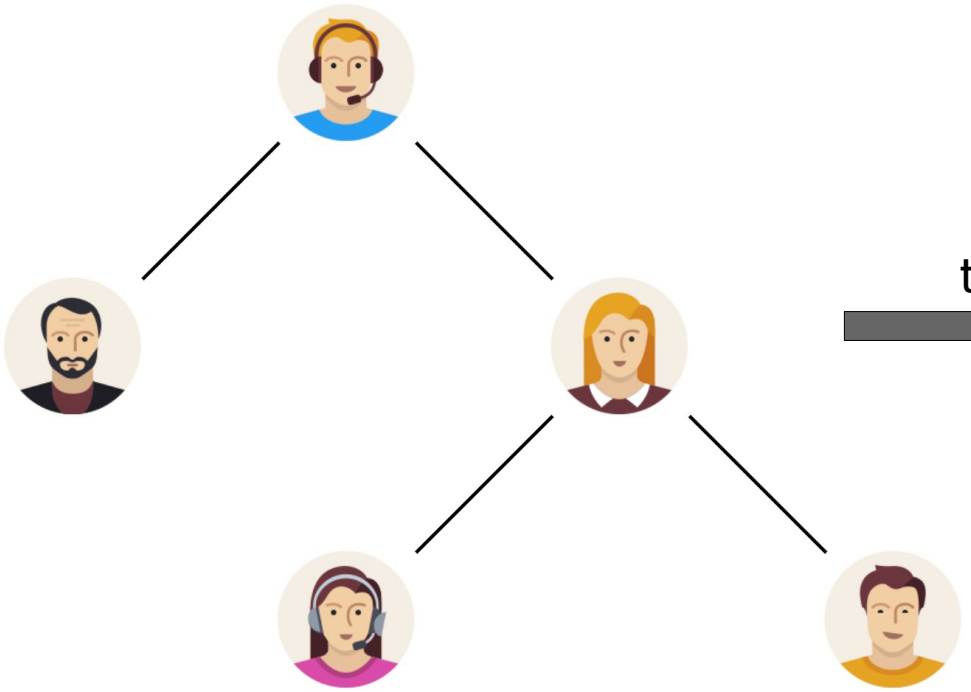
Our Approach:
Link Stealing Attacks

Link Stealing Attacks

- Scenario:
  - GNN trained on graph G to perform downstream task
  - Attacker
    - Black box access to target model
    - Partial graph with incomplete set of edges

- Goal:
  - Recover missing links from partial graph

Link Stealing Attacks

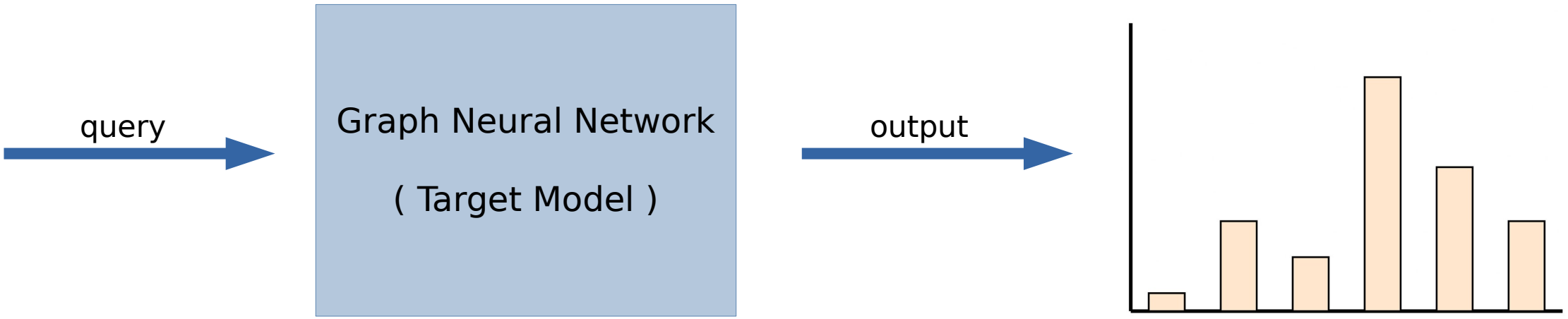


train

Graph Neural Network

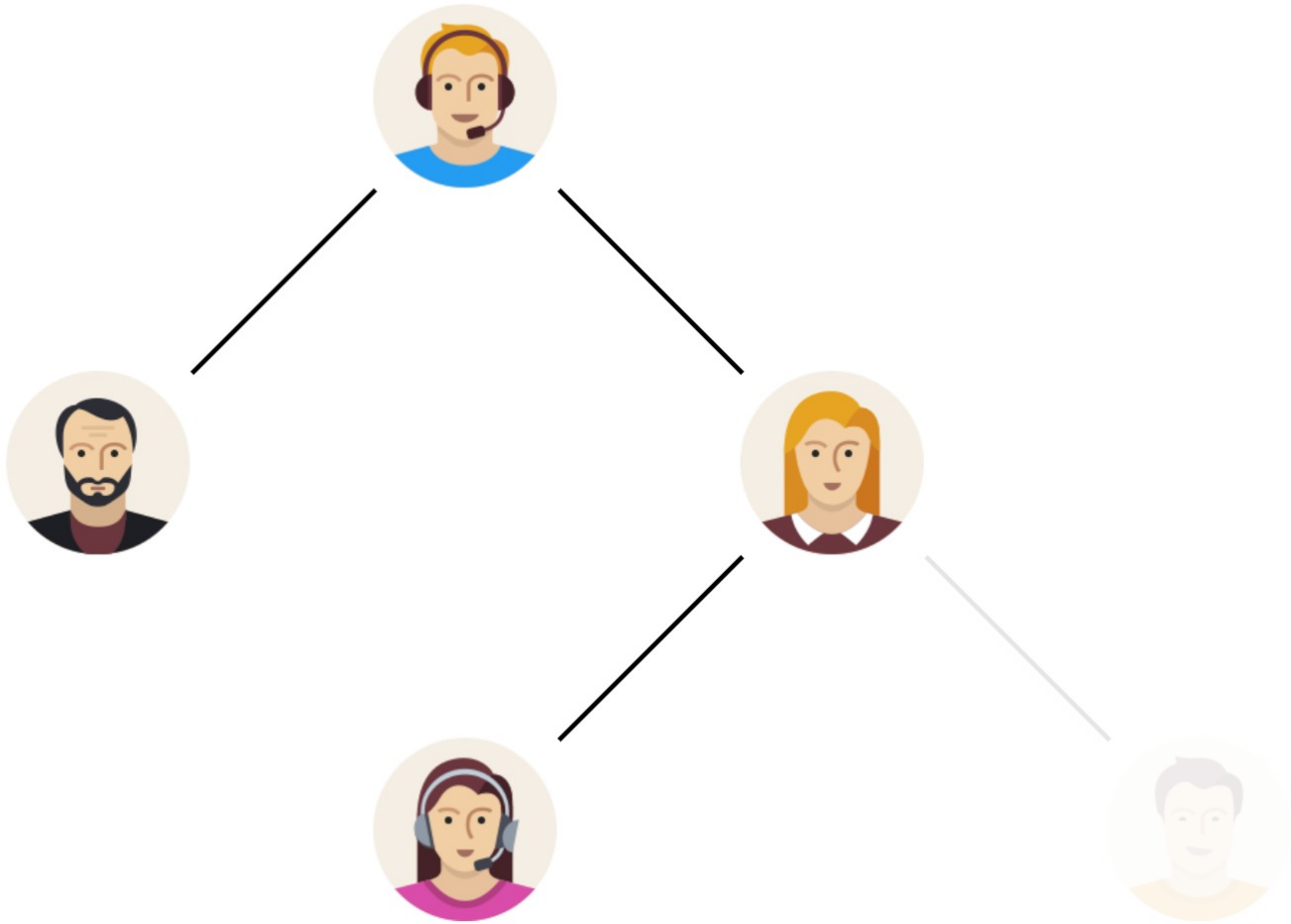
(Target Model)

Link Stealing Attacks



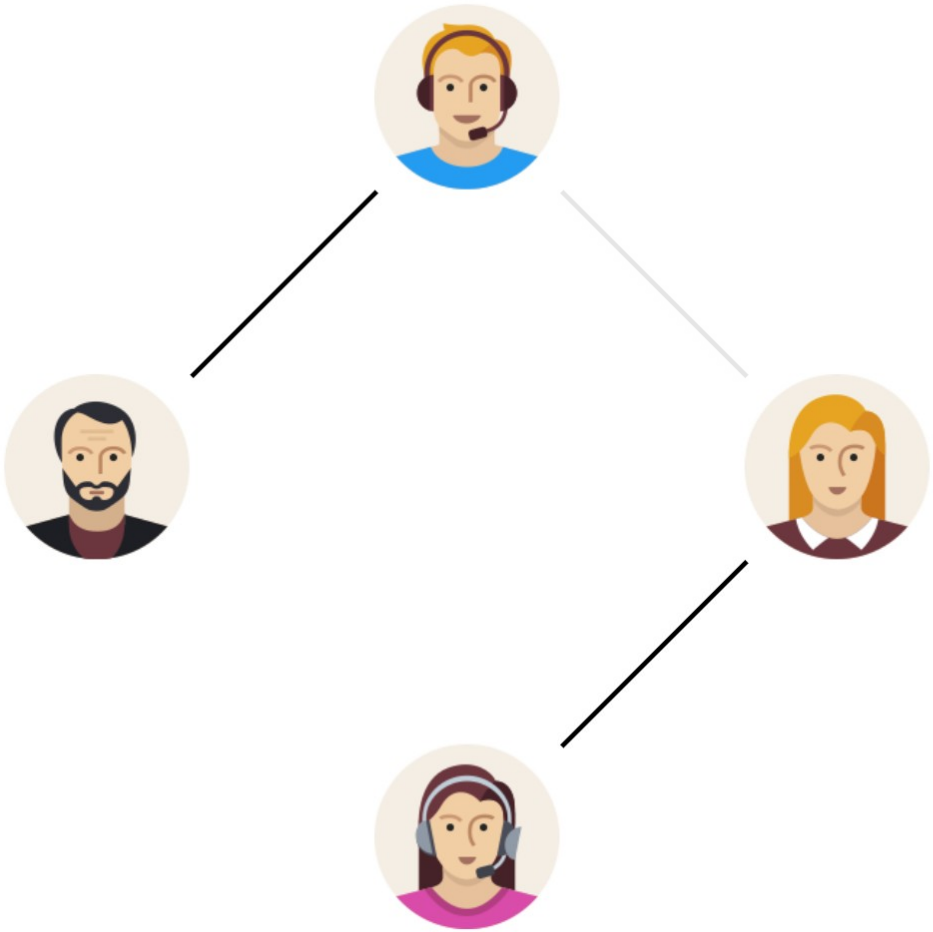
Link Stealing Attacks

- Attacker Graph



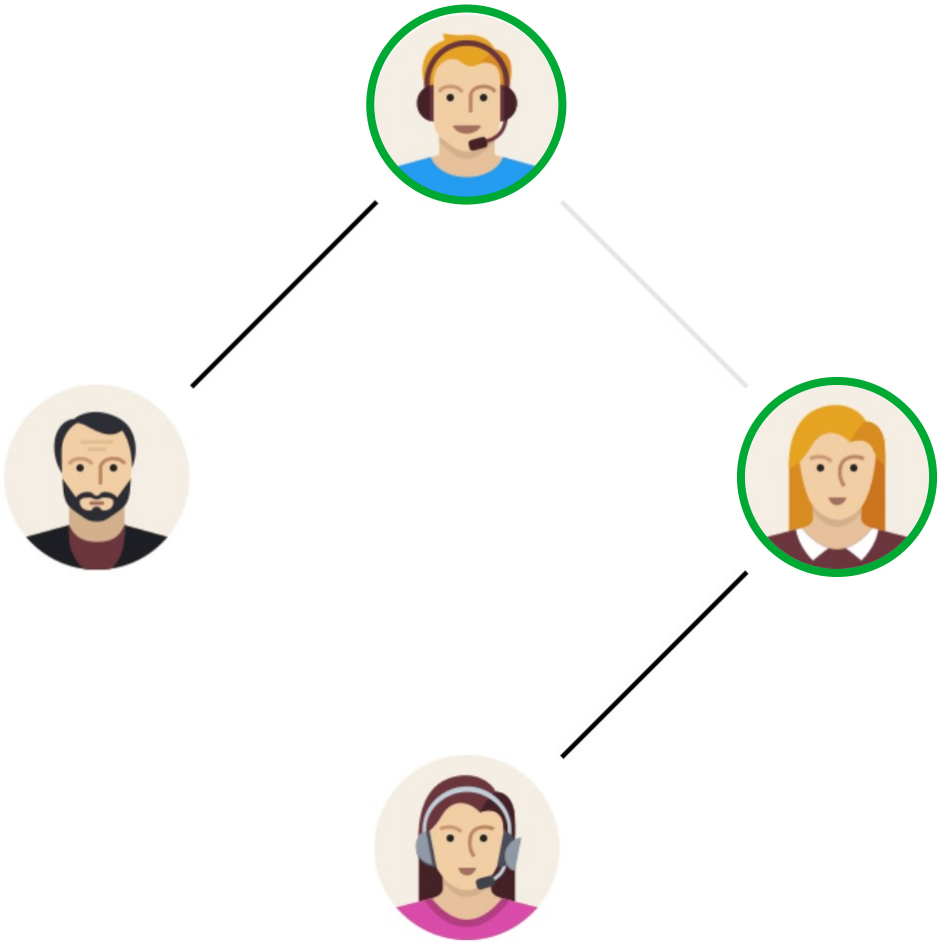
Link Stealing Attacks

- Attacker Graph with One Missing Link

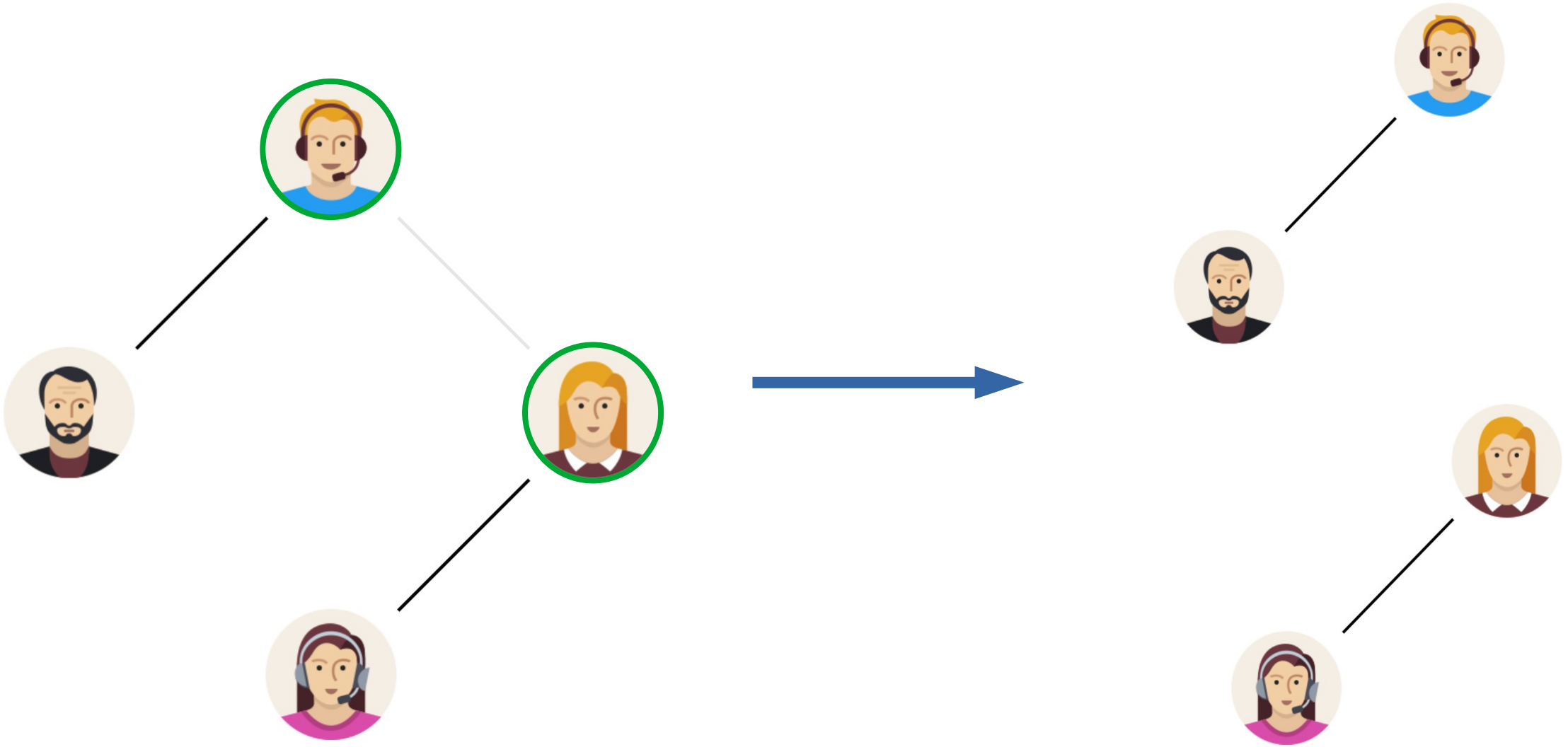


Link Stealing Attacks

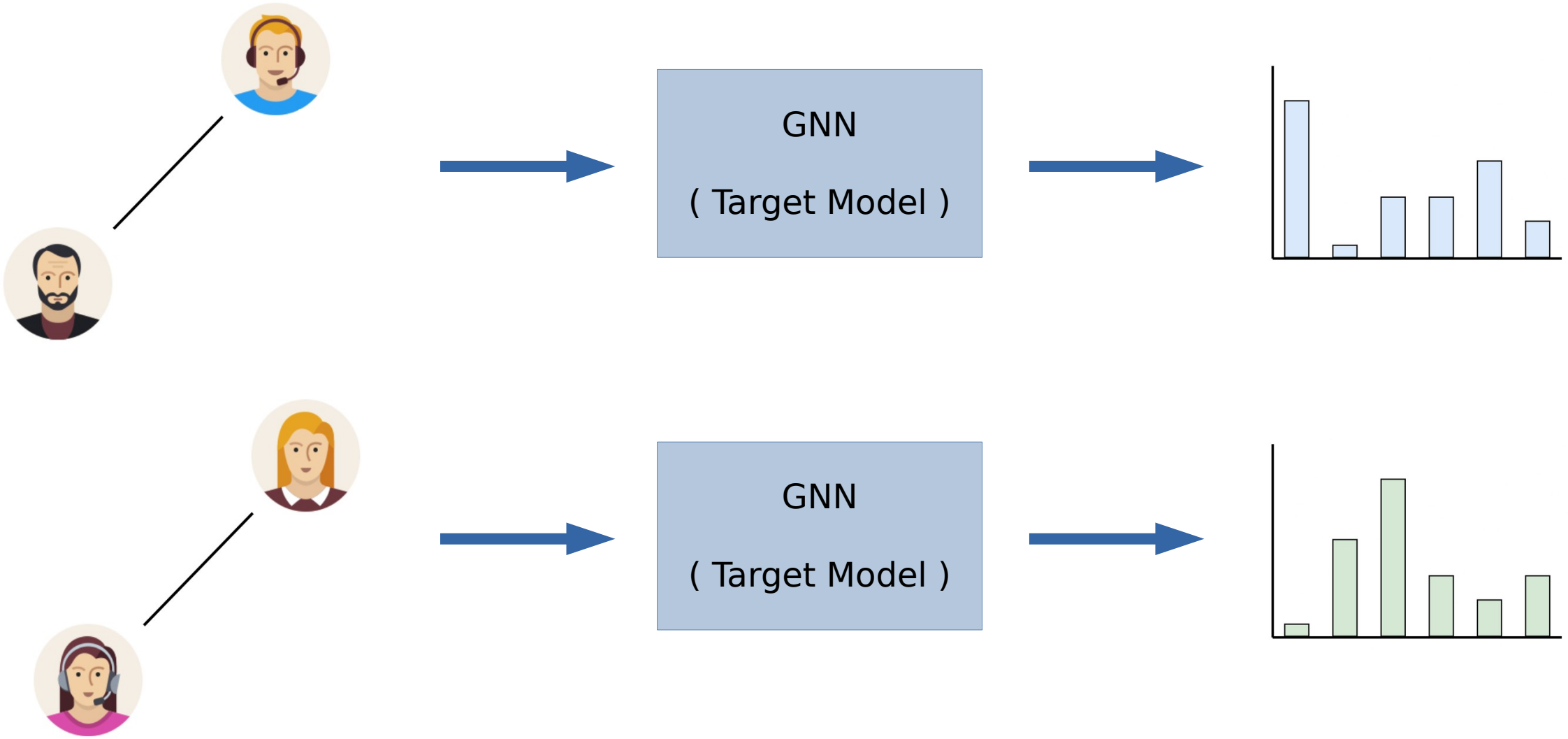
- Attacker Graph



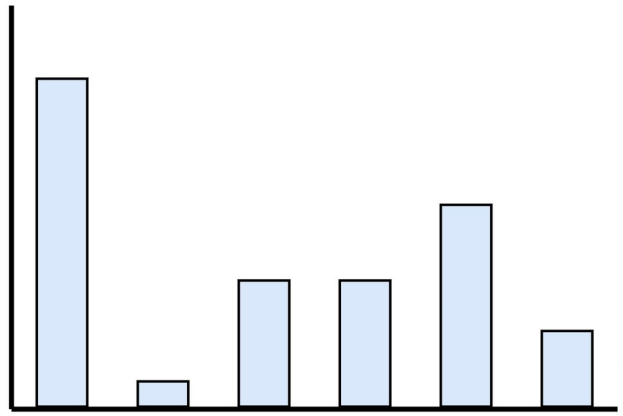
Link Stealing Attacks

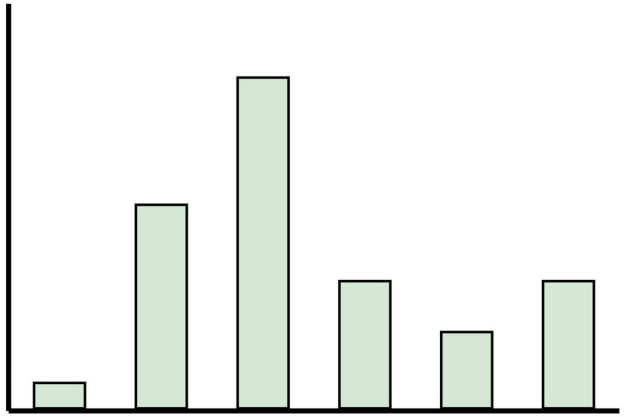


Link Stealing Attacks

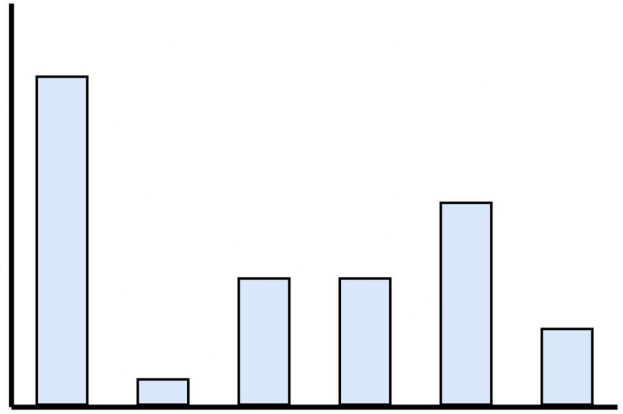


Link Stealing Attacks

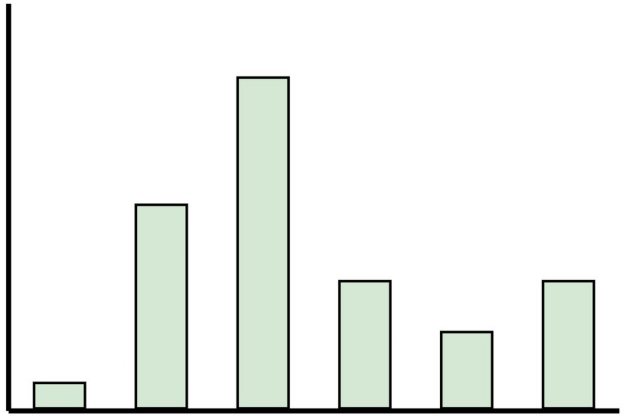




Link Stealing Attacks

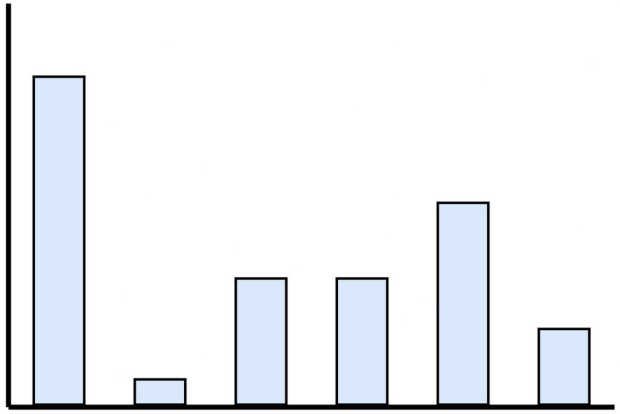


[0.3, 0.2, 0.3, …, 0.1]

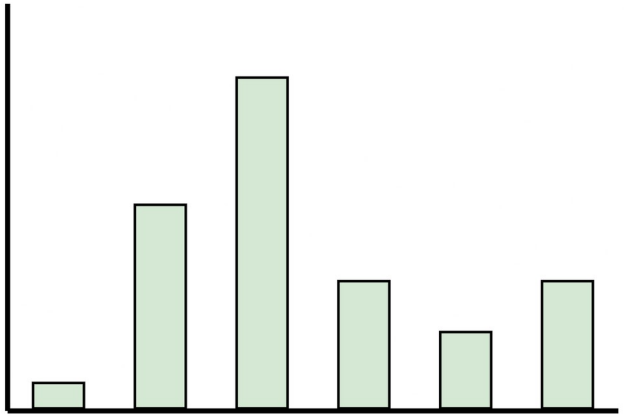


[0.2, 0.1, 0.4, …, 0.1]

Link Stealing Attacks

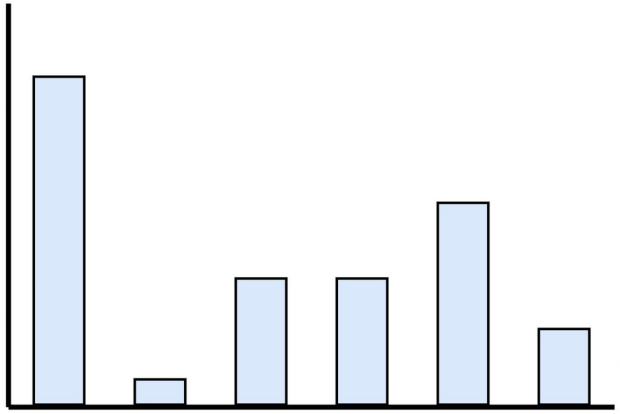


[0.3, 0.2, 0.3, …, 0.1]

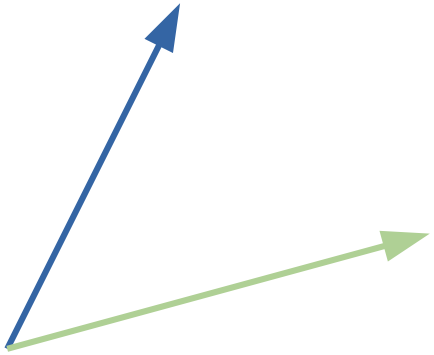


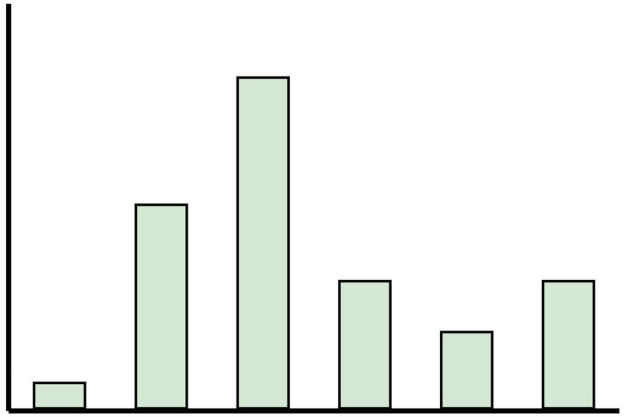
[0.2, 0.1, 0.4, …, 0.1]

Link Stealing Attacks



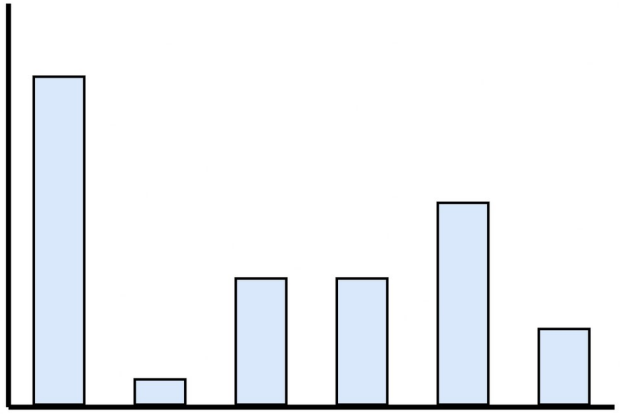
[0.3, 0.2, 0.3, …, 0.1]



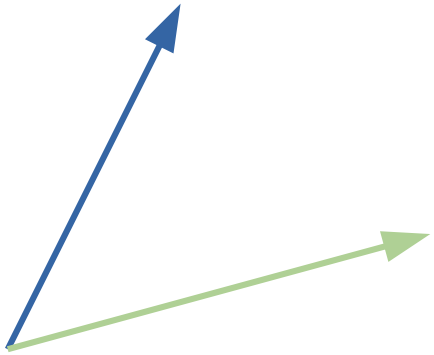


[0.2, 0.1, 0.4, …, 0.1]

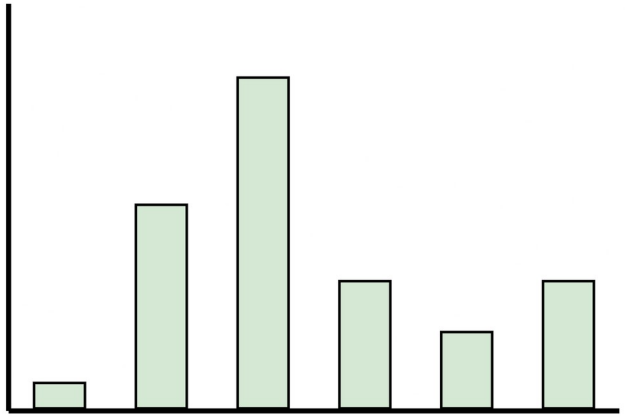
Link Stealing Attacks



[0.3, 0.2, 0.3, …, 0.1]



[Cosine, Manhattan, …, Euclidean]



[0.2, 0.1, 0.4, …, 0.1]

Link Stealing Attacks

[Cosine, Manhattan, …, Euclidean]

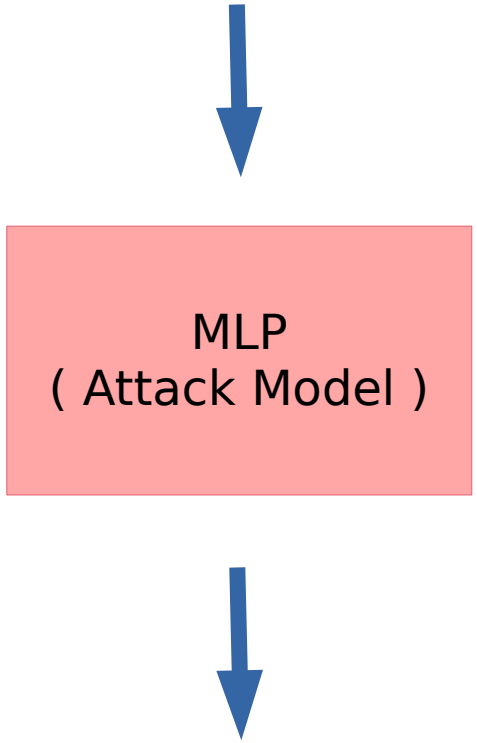
Link Stealing Attacks

[Cosine, Manhattan, …, Euclidean]

MLP
(Attack Model)

Link Stealing Attacks

[Cosine, Manhattan, …, Euclidean]



Prediction whether two nodes are connected or not

Experimental Setup

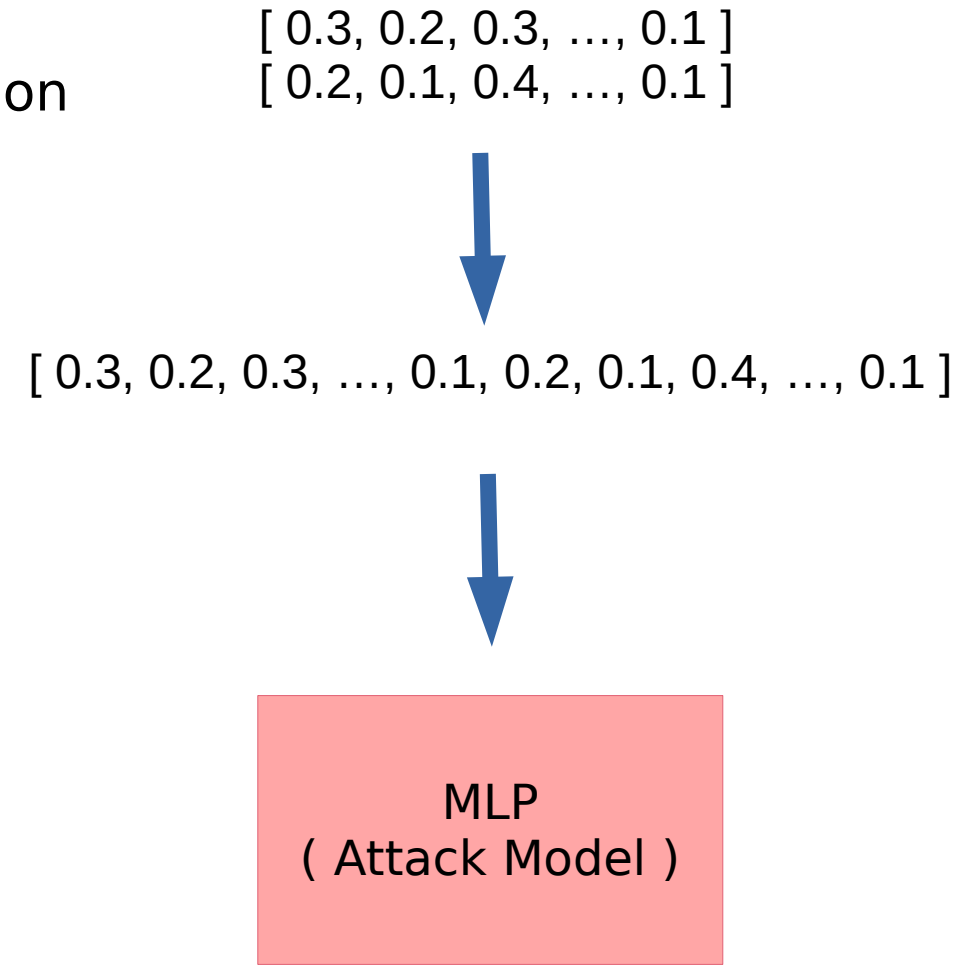
Experimental Setup

- Three Datasets
  - Cora
  - CiteSeer
  - Pubmed

- Three Graph Neural Network Types
  - GraphSAGE
  - GAT
  - GCN (inductive)

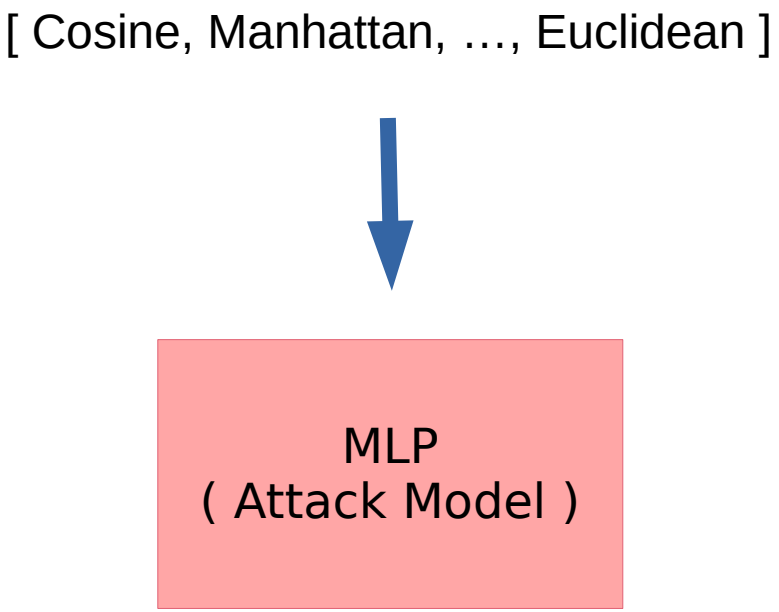
Experimental Setup

- Attack 1
  - Same distribution



Experimental Setup

- Attack 2
  - Same Distribution



Experimental Setup

- Attack 3
  - Different Distribution

[Cosine, Manhattan, …, Euclidean]

MLP
(Attack Model)

Goal

Goal

- Observation
  - Inductive trained GNNs are likely to reveal sensitive information about their training graph

- Serious Concerns
  - Intellectual property
  - Confidentiality
  - Privacy


Questions?