DR. B. KHALDI
Assoc. Prof.
email: b.khaldi@esi-sba.dz

Part III

Advanced GNN Techniques

Lecture 08

Predicting Links with Graph Neural Networks

- Understanding the different methods used for Link Prediction.
- Activity: Implement Basic Predicting Links approaches on a common Graph datasets.

DR. B. KHALDI
Assoc. Prof.
email: b.khaldi@esi-sba.dz

Introduction to Link Prediction

Link Prediction Defined:

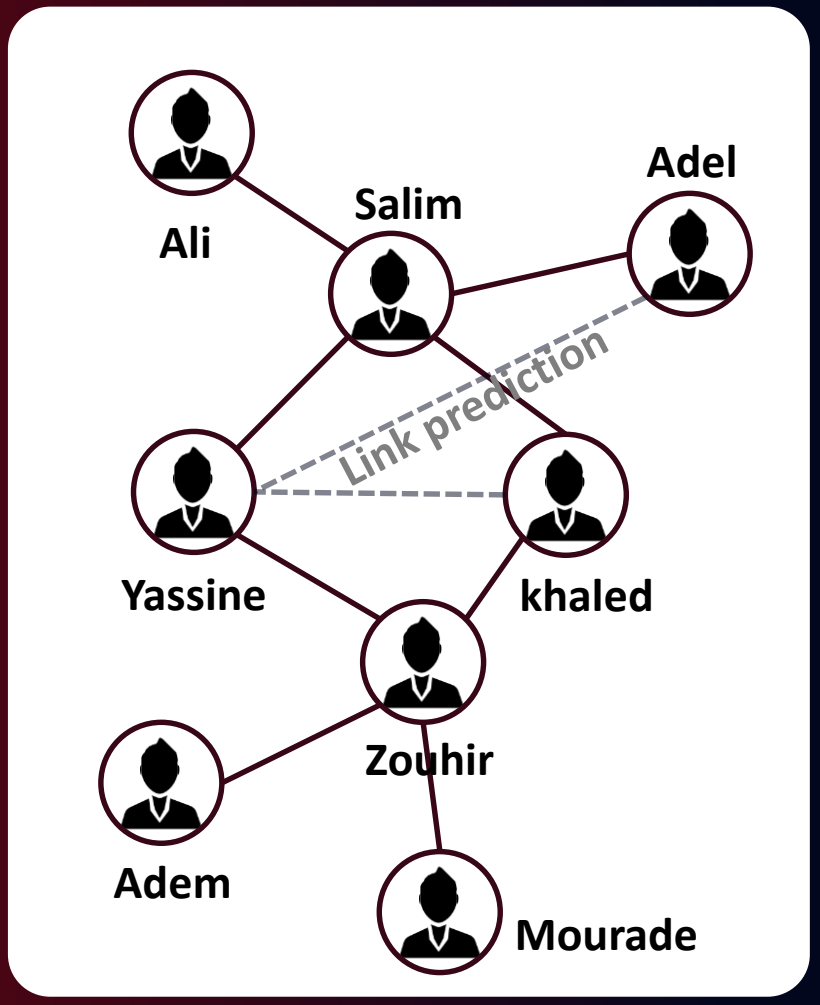
- Popular task in graph analysis.
- Predicting existence of a link between two nodes.

Core Applications:

- Integral for social networks and recommender systems.
- Example: Social media displaying mutual friends/followers.

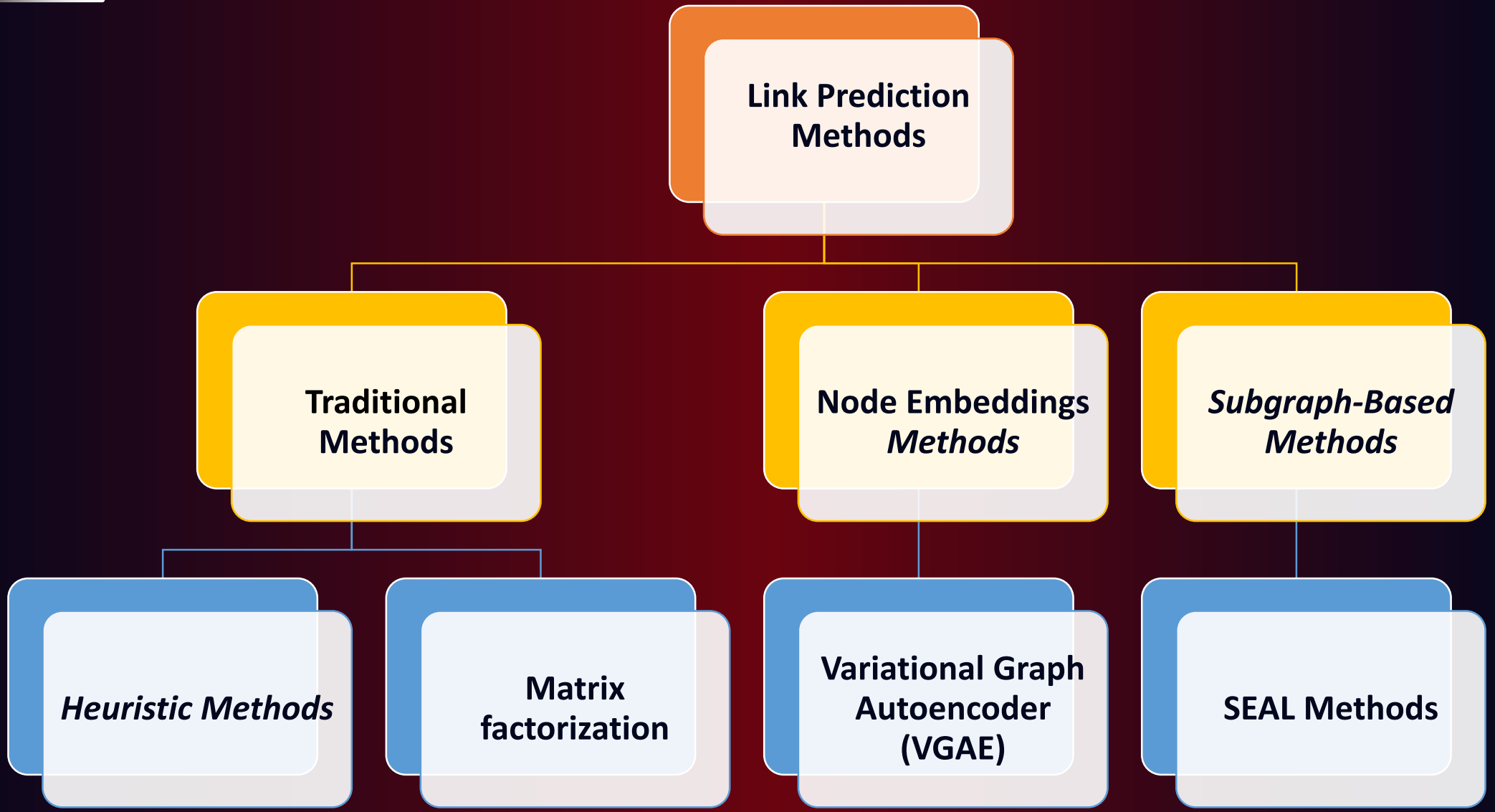
Objective of Link Prediction:

- Estimate likelihood of connection between nodes.
- Crucial for understanding relationships in networks.



DR. B. KHALDI
Assoc. Prof.
email: b.khaldi@esi-sba.dz

Overview of Link Prediction Methods



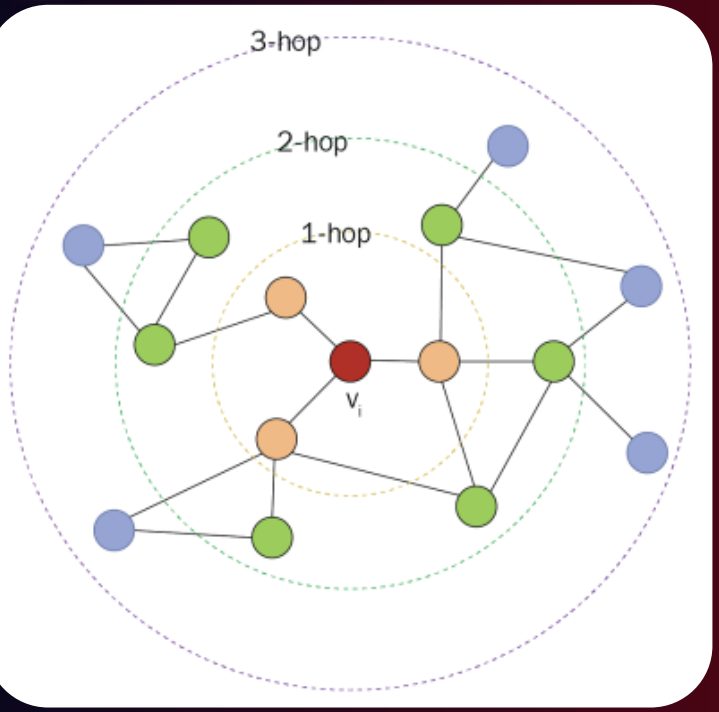
DR. B. KHALDI
Assoc. Prof.
email: b.khaldi@esi-sba.dz

Overview of Link Prediction Methods

Traditional Methods

01 Heuristic Methods

02 Matrix factorization



Introduction:

Heuristic techniques are simple yet effective for link prediction.

Use **node similarity scores** as the likelihood of links.

Classifiable based on the number of **hops** they consider.

Classification of Heuristics:

Divided into **local** (1-hop and 2-hop) and **global** heuristics.

Local heuristics focus on **node neighborhoods**.

Global heuristics consider the **entire network**.

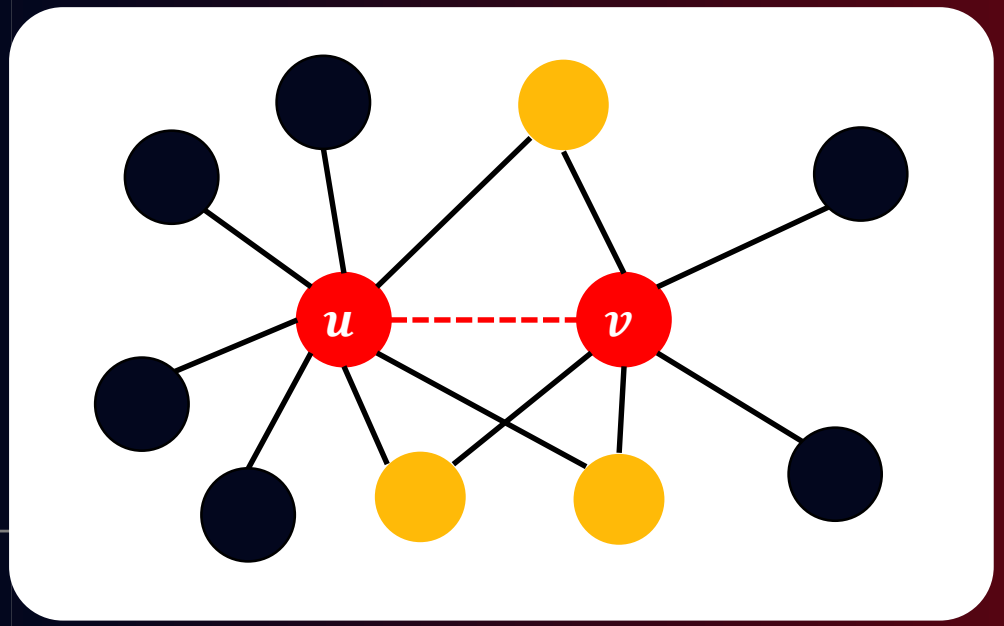
DR. B. KHALDI
Assoc. Prof.
email: b.khaldi@esi-sba.dz

Overview of Link Prediction Methods

Traditional Methods

01 Heuristic Methods

02 Matrix factorization



- **Local Heuristics:**
  - Measure the **similarity** between two nodes by considering their **local neighborhoods**.
  - **Popular Local Heuristics:**

Common Neighbors: Counts the number of 1-hop neighbors two nodes share.

$$f_{CN}(u, v) = |\mathcal{N}(u) \cap \mathcal{N}(v)|$$

Reflects the idea that more common neighbors imply a higher likelihood of connection.

CN is widely used in social network friend recommendation. It assumes that the more common friends two people have, the more likely they themselves are also friends.

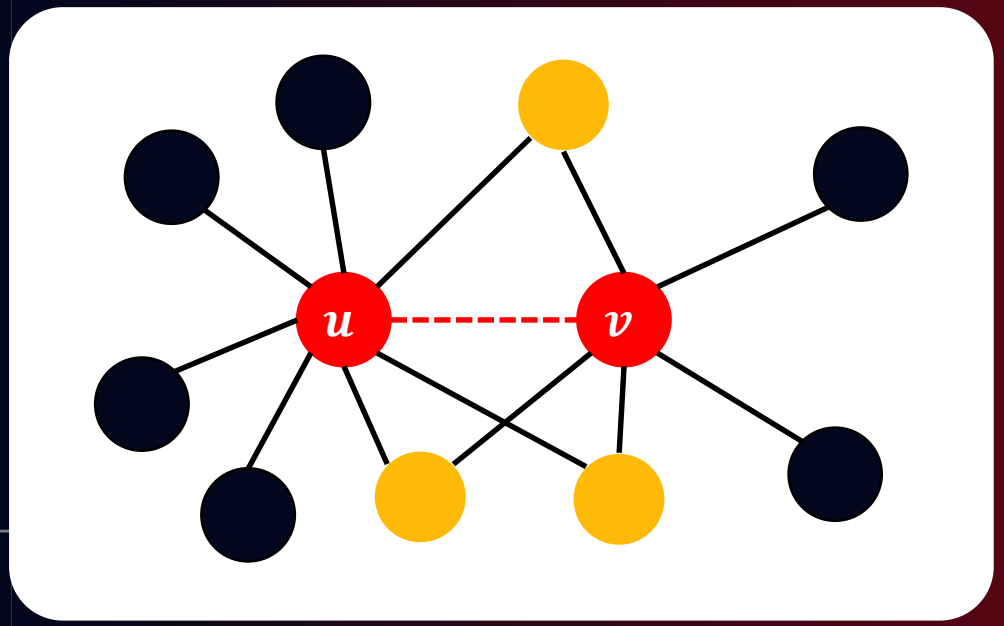
DR. B. KHALDI
Assoc. Prof.
email: b.khaldi@esi-sba.dz

Overview of Link Prediction Methods

Traditional Methods

01 Heuristic Methods

02 Matrix factorization



- **Local Heuristics:**
  - Measure the **similarity** between two nodes by considering their **local neighborhoods**.
  - **Popular Local Heuristics:**

Jaccard's Coefficient: Measures the proportion of shared 1-hop neighbors.

$$f_{Jaccard}(u,v) = \frac{|\mathcal{N}(u) \cap \mathcal{N}(v)|}{|\mathcal{N}(u) \cup \mathcal{N}(v)|}$$

Normalizes results by the total number of neighbors, rewarding nodes with few interconnected neighbors.

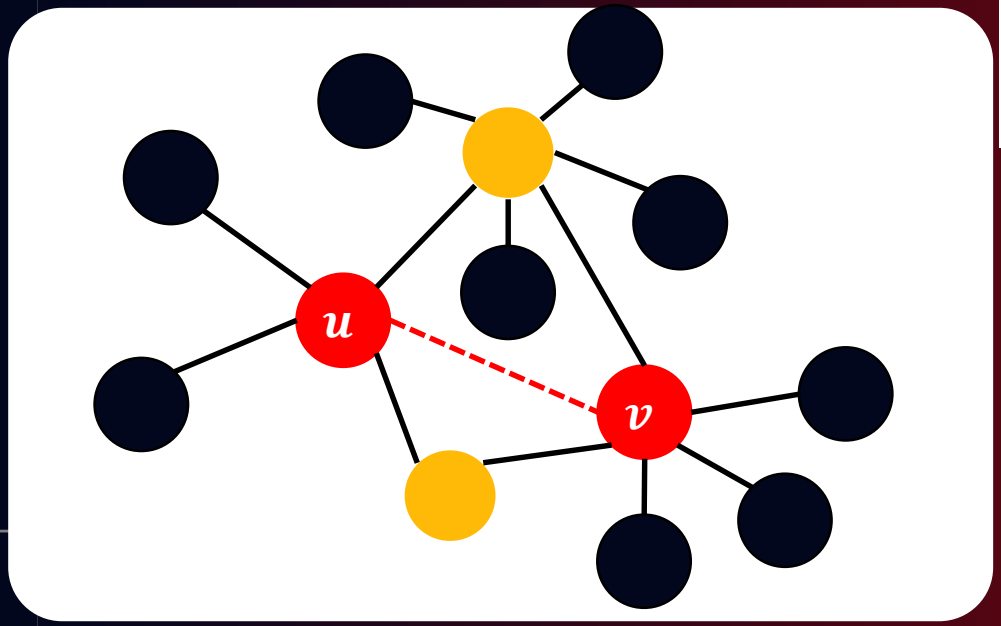
DR. B. KHALDI
Assoc. Prof.
email: b.khaldi@esi-sba.dz

Overview of Link Prediction Methods

Traditional Methods

01 Heuristic Methods

02 Matrix factorization



- **Local Heuristics:**
  - Measure the **similarity** between two nodes by considering their **local neighborhoods**.
  - **Popular Local Heuristics:**

Adamic–Adar Index: Considers inverse logarithmic degree of shared 2-hop neighbors.

$$f_{AA}(u,v) = \sum_{z \in (\mathcal{N}(u) \cap \mathcal{N}(v))} \frac{1}{log|\mathcal{N}(z)|}$$

Rewards nodes with small neighborhood sizes, reducing importance of common neighbors with large neighborhoods.

DR. B. KHALDI
Assoc. Prof.
email: b.khaldi@esi-sba.dz

Overview of Link Prediction Methods

Traditional Methods

01 Heuristic Methods

02 Matrix factorization

- **Global Heuristics:**
  - **Limitations of Local Heuristics:**
    - **Dependency on Node Degrees:**
      - Local heuristics rely on direct or indirect node degrees.
      - Speed and explainability benefits, but limits complexity of relationships.
  - **Global Heuristics** ➔ Offer a solution to the limitations of local approaches.

Katz Index: Computes weighted sum of all possible paths between two nodes.

$$f_{Katz}(u,v) = \sum_{l=1}^{n} \beta^l |paths^{\langle l \rangle}(u,v)|$$

Weights based on a discount factor, $\boldsymbol{\beta} \in [\mathbf{0}, \mathbf{1}]$.

Penalizes longer paths, emphasizing short paths for higher connection likelihood.

DR. B. KHALDI
Assoc. Prof.
email: b.khaldi@esi-sba.dz

Overview of Link Prediction Methods

Traditional Methods

01 Heuristic Methods

02 Matrix factorization

Introduction: Matrix factorization for link prediction is inspired by recommender systems.

Indirectly predicts links by predicting the entire adjacency matrix, $\hat{A}$ using node embeddings.

Prediction Logic: Similar nodes, u and v, should have similar embeddings, z_u and z_v respectively.

Dot product logic $z_v^T z_u$: Maximal if nodes are similar,

Minimal if different.

DR. B. KHALDI
Assoc. Prof.
email: b.khaldi@esi-sba.dz

Overview of Link Prediction Methods

Traditional Methods

01 Heuristic Methods

02 Matrix factorization

Approximating Adjacency Matrix:

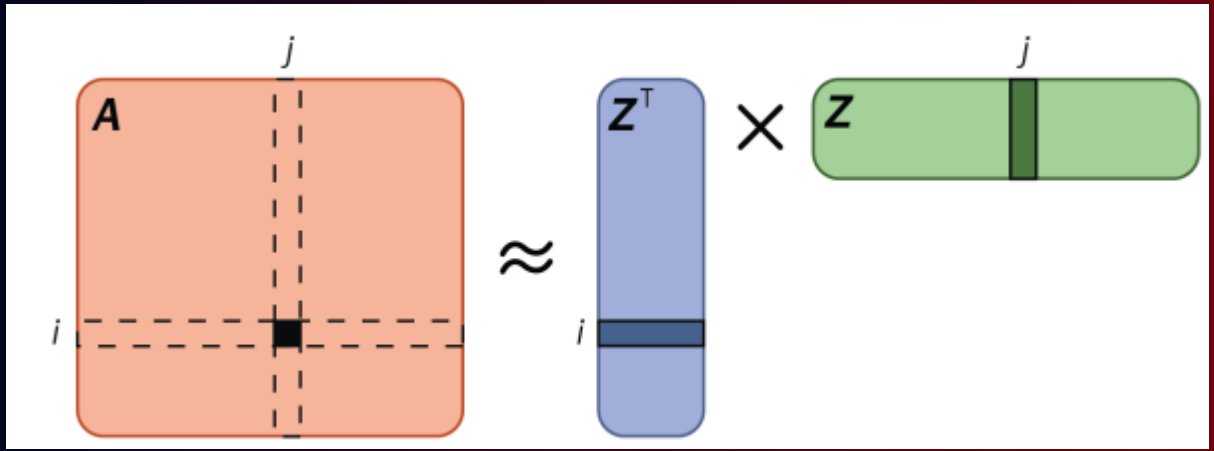
Dot product used to approximate each element (link) of the adjacency matrix, $\hat{A}$.

$$A_{uv} \approx z_v^T z_u$$

Matrix multiplication: Z is the node embedding matrix.

$$A \approx Z^T Z$$

Goal: Learn node embeddings minimizing $L2$ norm between true and predicted elements for the graph.



$$L_2 = \underset{Z}{minimize} \sum_{i \in V, j \in V} (A_{uv} - z_v^T z_u)^2$$

DR. B. KHALDI
Assoc. Prof.
email: b.khaldi@esi-sba.dz


Overview of Link Prediction Methods

Traditional Methods

01 Heuristic Methods

02 Matrix factorization

❑ Advanced Matrix Factorization Variants

Incorporating Laplacian Matrix: Advanced variants include **Laplacian matrix** and powers of A.

DeepWalk and **Node2Vec** are alternative solutions producing node embeddings for link representations.

Implicit Approximation: Algorithms like DeepWalk implicitly approximate and factorize complex matrices.

Example: Matrix computed by DeepWalk.

$$log\left(\sum_{i\in V}\sum_{j\in V} A_{ij}\left(\frac{1}{T}\sum_{r=1}^{T}\left(D^{-1}A\right)^{r}\right)D^{-1}\right) - log\ b$$

DR. B. KHALDI
Assoc. Prof.
email: b.khaldi@esi-sba.dz


Overview of Link Prediction Methods

Traditional Methods

01 Heuristic Methods

02 Matrix factorization

❑ Limitations of Matrix Factorization

Limitations of DeepWalk and Node2Vec:	Cannot use node features, relying solely on topological information.
	Lack inductive capabilities; cannot generalize to nodes not in the training set.
	Struggle to capture structural similarity, leading to vastly different embeddings for structurally similar nodes.
Motivation for GNN-based Techniques:	GNNs address limitations of matrix factorization.
	Next sections will explore GNN-based techniques for link prediction.

DR. B. KHALDI
Assoc. Prof.
email: b.khaldi@esi-sba.dz

Overview of Link Prediction Methods

Node Embeddings Methods

01 Graph Autoencoder (GAE)

02 Variational Graph Autoencoder (VGAE))

Architecture Overview:

Introduced by Kipf and Welling in 2016.

GAE is the GNN counterpart of the autoencoder.

Composed of two modules: **Encoder** and **Decoder**.

Encoder:

Two-layer Graph Convolutional Network (**GCN**) computes node embeddings, Given the adjacency matrix $\boldsymbol{A}$ and node feature matrix $\boldsymbol{X}$ of a graph:

$$\boldsymbol{Z = GCN(X, A)}$$

Decoder:

Approximates $\widehat{\boldsymbol{A}}$ using matrix factorization and a sigmoid function $\boldsymbol{\psi}$.

$$\widehat{\boldsymbol{A}} = \boldsymbol{\psi(Z^T Z)}$$

Training objective: Binary cross-entropy loss between elements of both adjacency matrices:

$$\boldsymbol{L_{BCE} = \sum_{i \in V, j \in V} \left(-A_{ij} log \widehat{A}_{ij} - \left(1 - A_{ij}\right) log\left(1 - \widehat{A}_{ij}\right) \right)}$$

DR. B. KHALDI
Assoc. Prof.
email: b.khaldi@esi-sba.dz

Overview of Link Prediction Methods

Node Embeddings Methods

01 Graph Autoencoder (GAE)

02 Variational Graph Autoencoder (VGAE))

Difference from GAE:	Similar to the difference between autoencoders and variational autoencoders.
	VGAE learns normal distributions and samples to produce embeddings.
Encoder:	Two GCNs sharing their first layer.
	Learn parameters of latent normal distribution: mean ($\boldsymbol{\mu}$) and variance ($\boldsymbol{\sigma}^2$).
Decoder:	Samples embeddings from $\mathbb{N}(\boldsymbol{\mu}, \boldsymbol{\sigma 2})$
	Inner product between latent variables approximates the $\widehat{\boldsymbol{A}}$. $\widehat{\boldsymbol{A}} = \boldsymbol{\psi}(\boldsymbol{Z}^{\boldsymbol{T}}\boldsymbol{Z})$
Loss Function:	Includes **Kullback-Leibler (KL) divergence** to ensure the encoder's output follows a normal distribution.
	Loss function: Evidence Lower Bound (ELBO). $\boldsymbol{L_{ELBO} = L_{LBC} - KL[q(Z\|X,A)\|\|p(Z)\,]}$

Here, $\boldsymbol{q(Z|X,A)}$ represents the encoder and $\boldsymbol{p(Z)}$ is the prior distribution of .

DR. B. KHALDI
Assoc. Prof.
email: b.khaldi@esi-sba.dz


Overview of Link Prediction Methods

Subgraph-Based Methods

01 SEAL Method

Introduction to SEAL Framework:

Introduced in 2018 by Zhang and Chen.

SEAL (**S**tructural **E**dge **A**nomaly **L**earning) focuses on learning graph structure features for link prediction.

The framework involves three steps:

Enclosing subgraph extraction, which consists of taking a set of real links and a set of fake links (negative sampling) to form the training data.

Node information matrix construction, which involves three components – node labels, node embeddings, and node features.

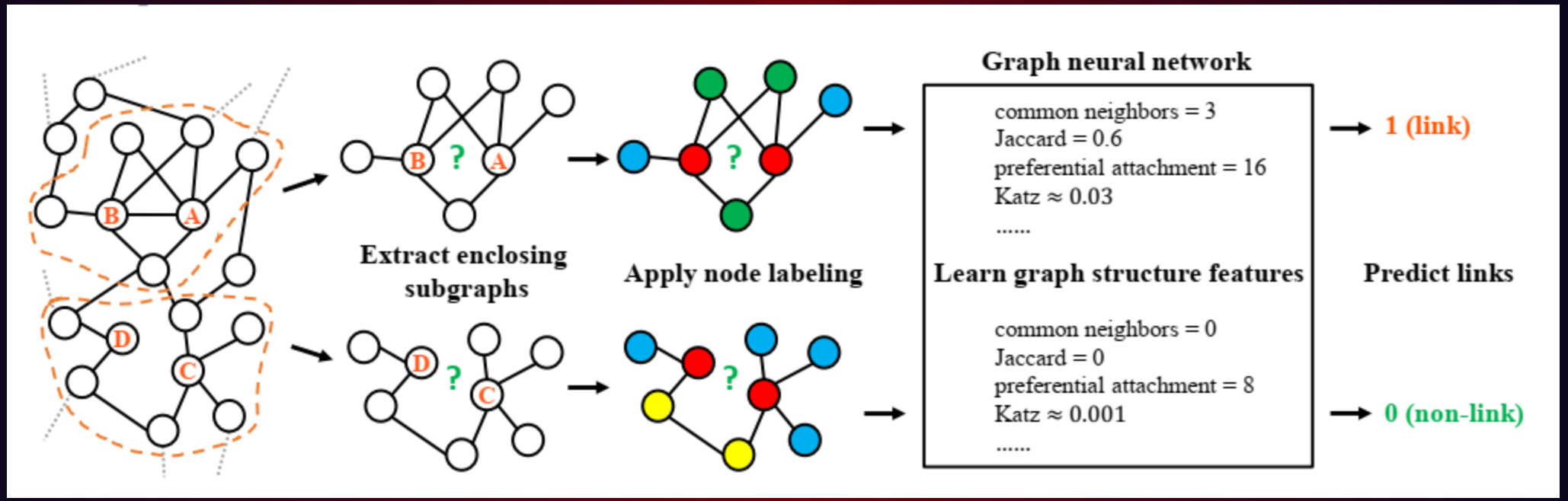
GNN training, which takes the node information matrices as input and outputs link likelihoods.

DR. B. KHALDI
Assoc. Prof.
email: b.khaldi@esi-sba.dz

Overview of Link Prediction Methods

Subgraph-Based Methods

01 SEAL Method



DR. B. KHALDI
Assoc. Prof.
email: b.khaldi@esi-sba.dz


THANK YOU