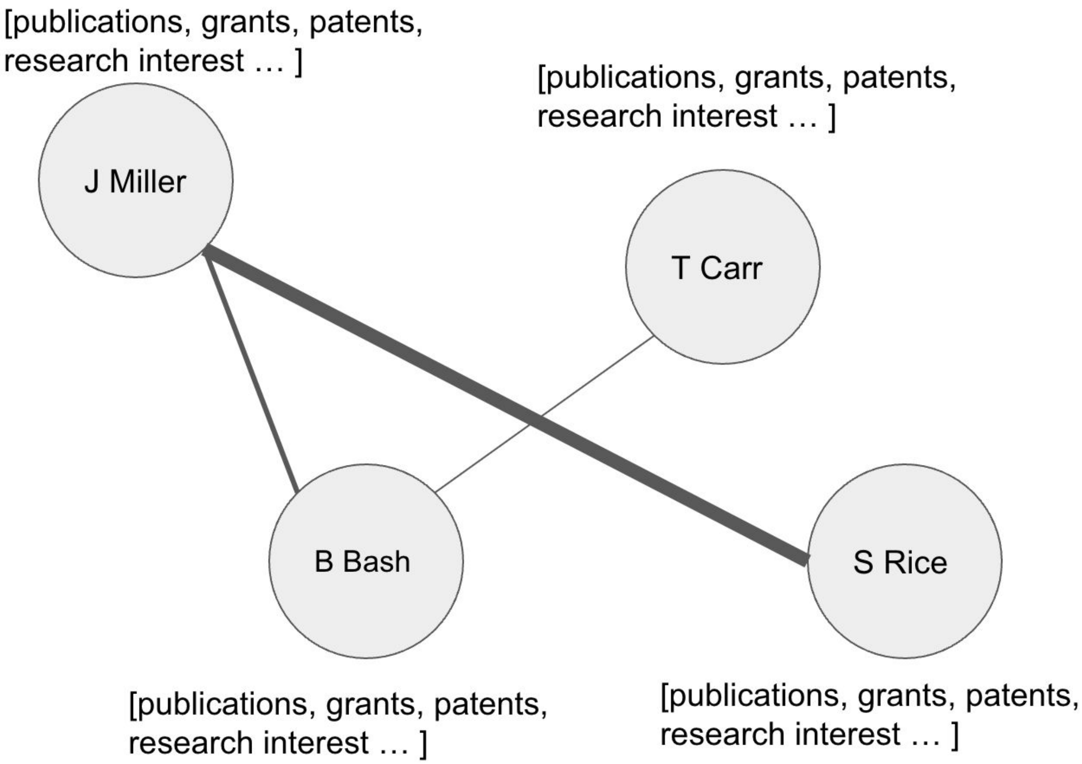
Project Status

Sammi Abida Salma

- Problem definition
- Previous work
- Proposed approach
- Evaluation Methodology

- Problem definition
- Previous work
- Proposed approach
- Evaluation Methodology



[publications, grants, patents, research interest …]

[publications, grants, patents, research interest …]

[publications, grants, patents, research interest …]

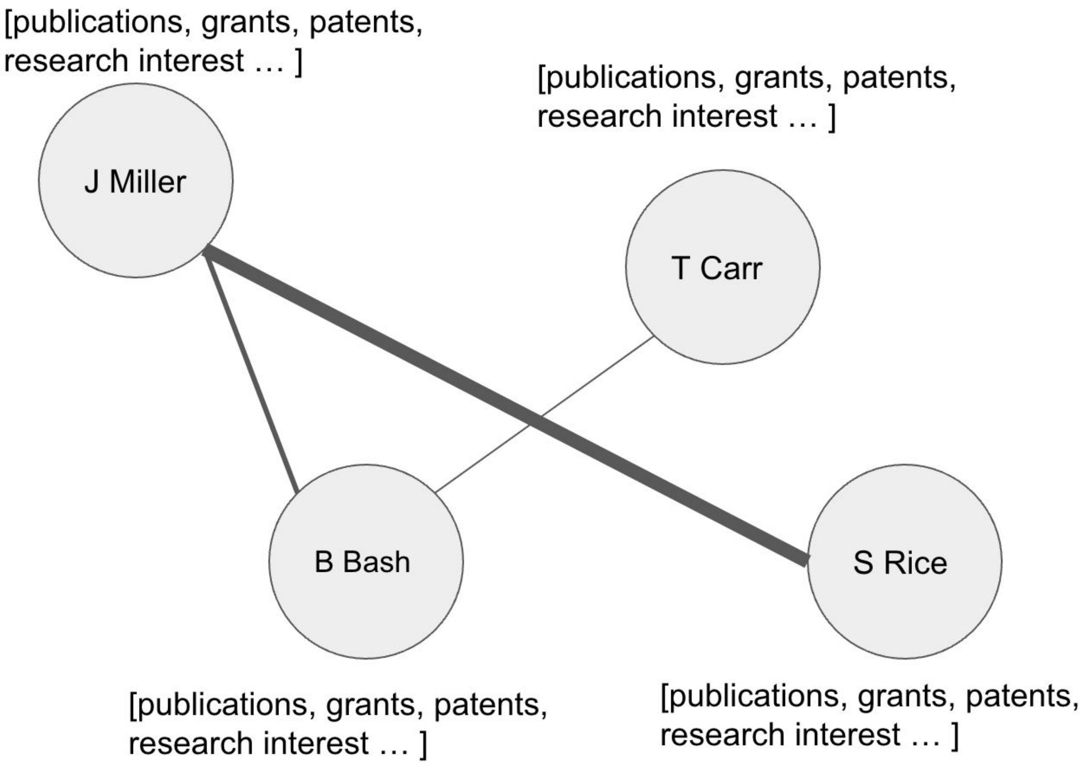
[publications, grants, patents, research interest …]

Node ->

Researchers

Link ->

collaborative research



[publications, grants, patents, research interest …]

J Miller

[publications, grants, patents, research interest …]

T Carr

B Bash

S Rice

[publications, grants, patents, research interest …]

[publications, grants, patents, research interest …]

Node ->

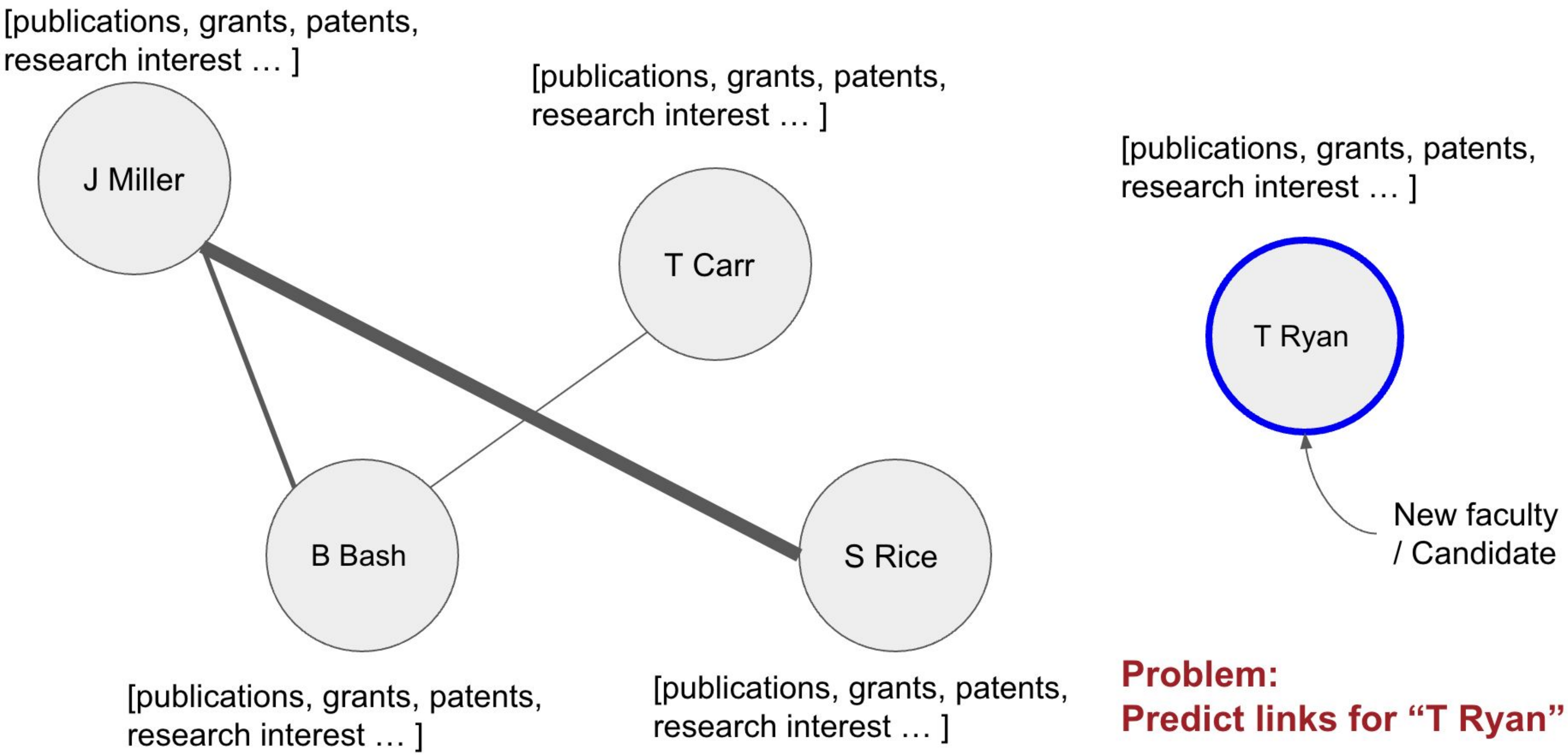
Researchers

Link ->

collaborative research

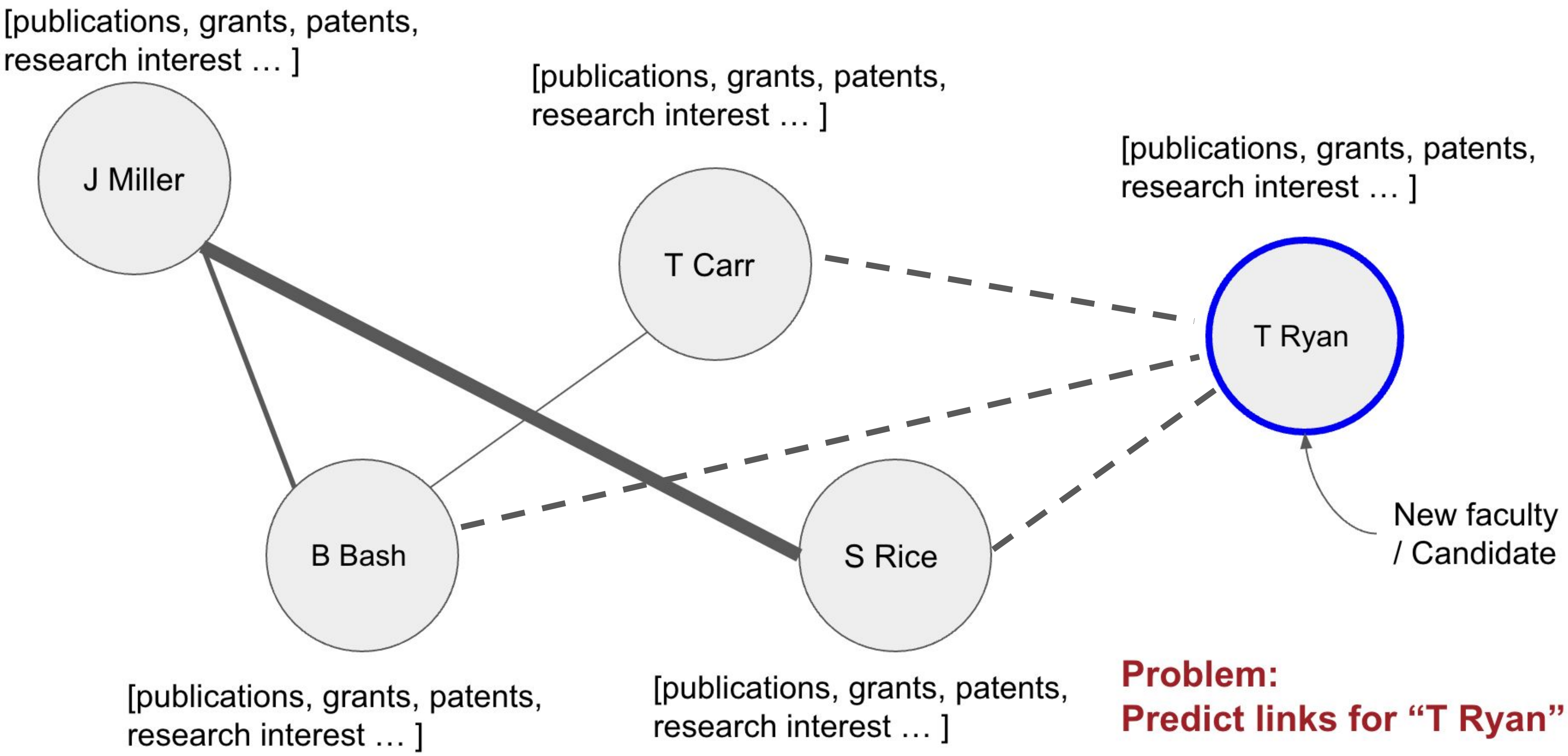
# nodes : **4,884**
# edges/links : **19,241**
Average degree : **7.8792**

Problem : *Infer Future Collaboration*


[publications, grants, patents, research interest …]
J Miller
[publications, grants, patents, research interest …]
T Carr
[publications, grants, patents, research interest …]
T Ryan
New faculty / Candidate
B Bash
S Rice
[publications, grants, patents, research interest …]
[publications, grants, patents, research interest …]

Problem:
Predict links for “T Ryan”

Link Prediction



[publications, grants, patents, research interest …]

[publications, grants, patents, research interest …]

[publications, grants, patents, research interest …]

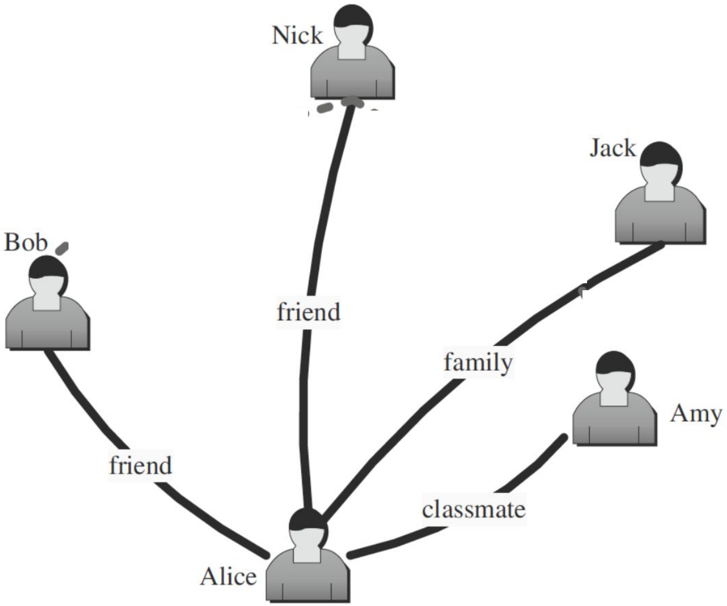
[publications, grants, patents, research interest …]

[publications, grants, patents, research interest …]

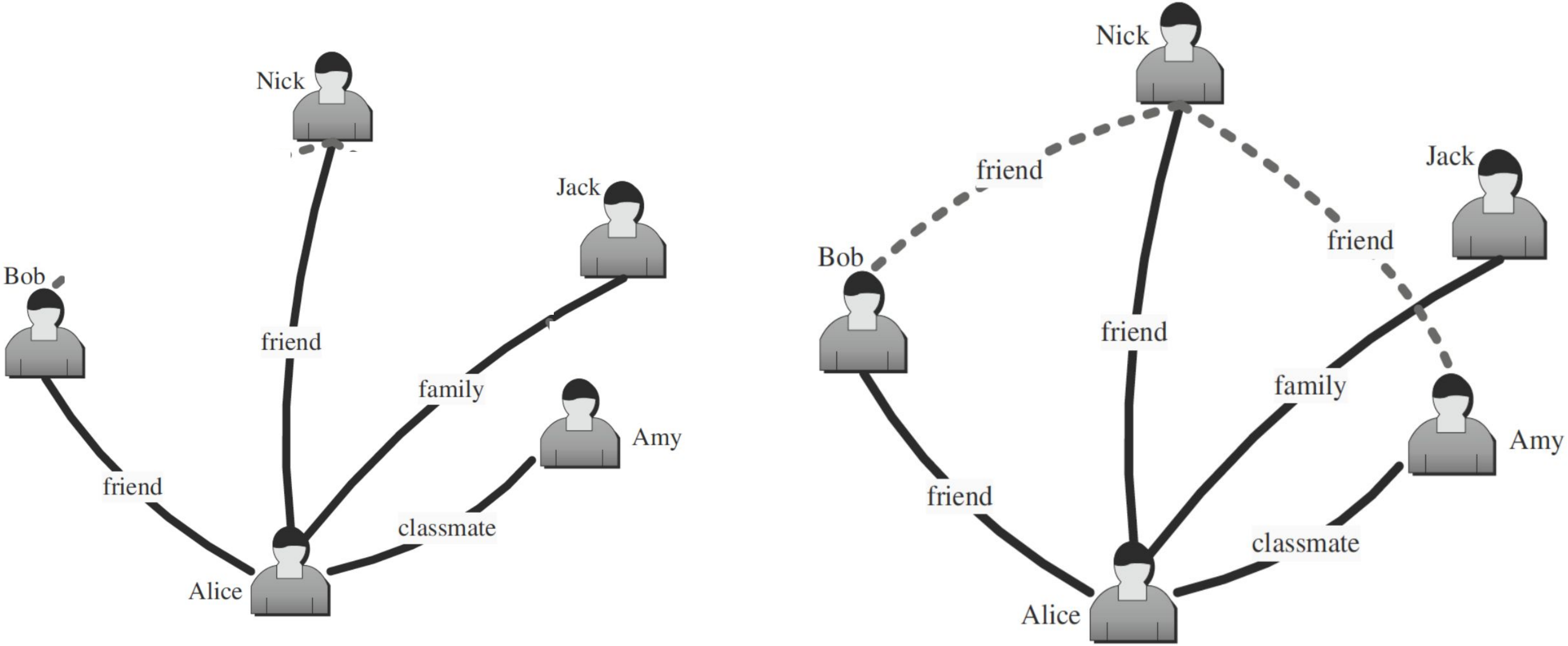
New faculty / Candidate

Problem:
Predict links for “T Ryan”

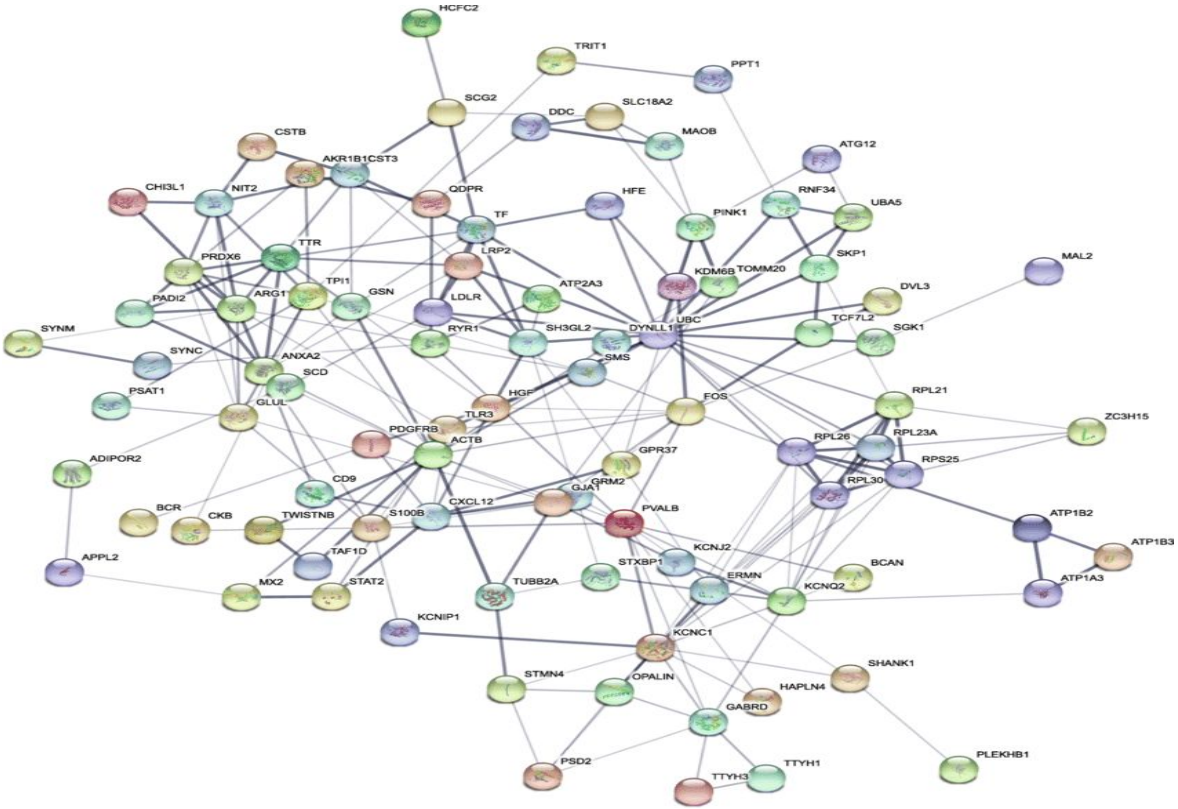
Link Prediction : *In Social Network*



Link Prediction : *Recommend Friend*



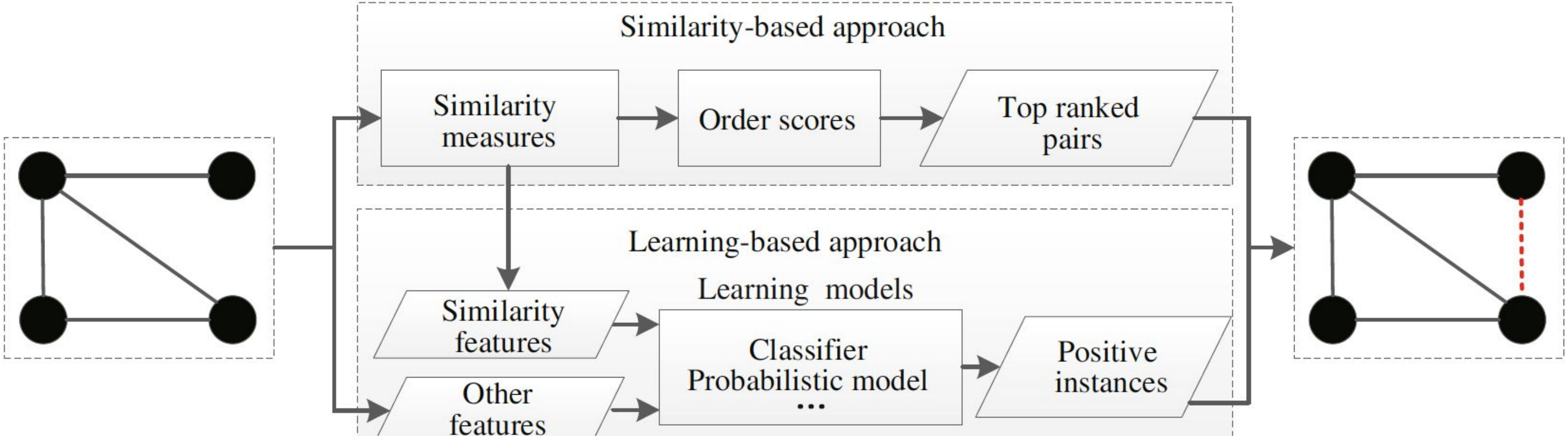
Link Prediction : *Infer Protein-protein interaction*



- ➢ Problem definition
- ➢ Previous work
- ➢ Proposed approach
- ➢ Evaluation Methodology

The generic link prediction framework

- Similarity-based approach (topological feature)
- Learning-based approach (topological feature + latent feature)



Similarity-based approach

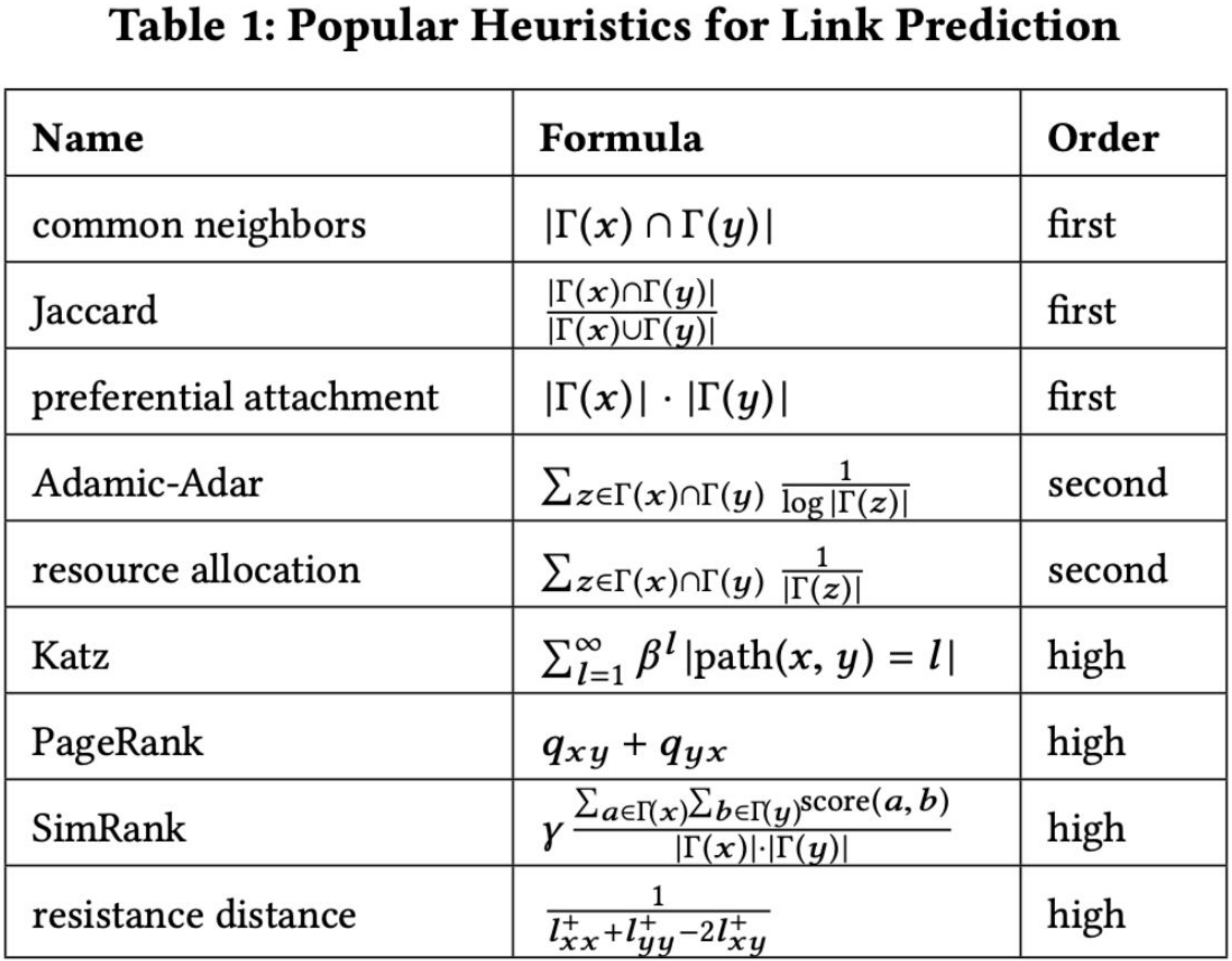


Table 1: Popular Heuristics for Link Prediction

Name	Formula	Order
common neighbors	$\|\Gamma(x) \cap \Gamma(y)\|$	first
Jaccard	$\frac{\|\Gamma(x) \cap \Gamma(y)\|}{\|\Gamma(x) \cup \Gamma(y)\|}$	first
preferential attachment	$\|\Gamma(x)\| \cdot \|\Gamma(y)\|$	first
Adamic-Adar	$\sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log \|\Gamma(z)\|}$	second
resource allocation	$\sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\|\Gamma(z)\|}$	second
Katz	$\sum_{l=1}^{\infty} \beta^{l} \|\text{path}(x, y) = l\|$	high
PageRank	$q_{xy} + q_{yx}$	high
SimRank	$\gamma \frac{\sum_{a \in \Gamma(x)} \sum_{b \in \Gamma(y)} \text{score}(a, b)}{\|\Gamma(x)\| \cdot \|\Gamma(y)\|}$	high
resistance distance	$\frac{1}{l^{+}_{xx} + l^{+}_{yy} - 2 l^{+}_{xy}}$	high

Similarity-based approach

- account (only) topological feature
  - degree of nodes
  - path information
- works well (*metric common neighbor*)
  - social network
- works poorly (*metric common neighbor*)
  - protein-protein interaction network
- Limitation
  - Non-universal
    - Different domains need different metrics
  - Fails to predict link where
    - **similarity scores do not capture the network’s latent formation mechanisms.**

Learning-based approach

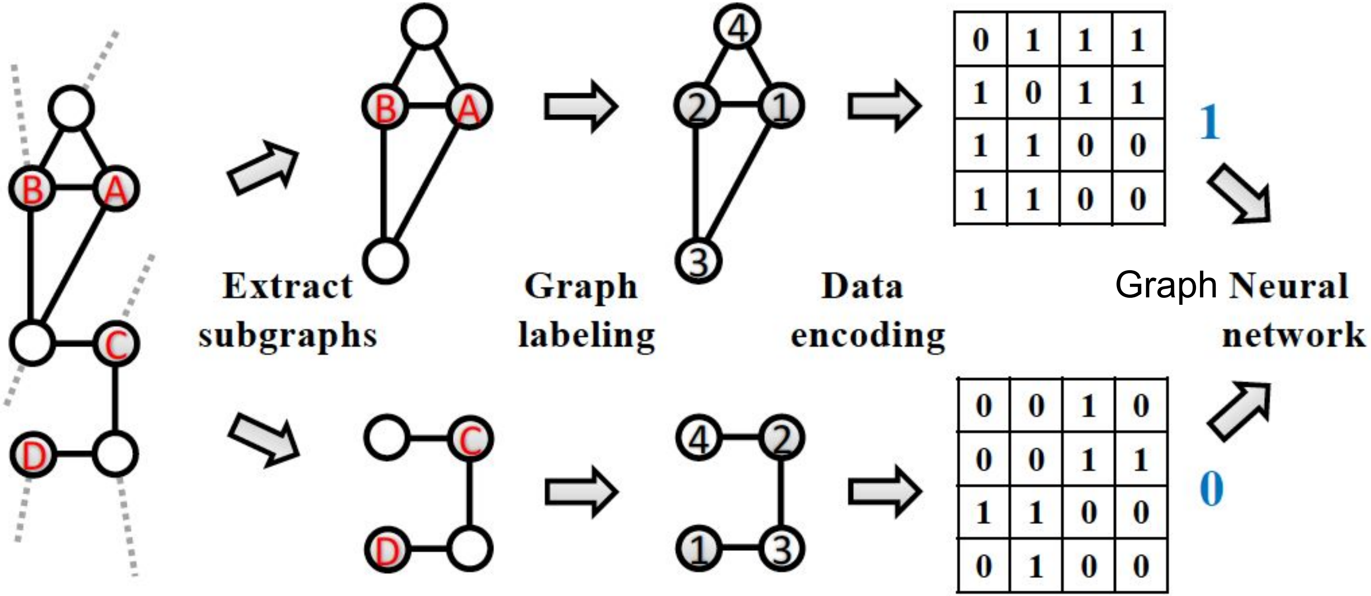
- ➔ Probabilistic Graphical Model
  - ◆ Ranking Method
  - ◆ Stochastic Block Model
- ➔ Matrix factorization
- ➔ Deep Learning
  - ◆ CNN (Convolution Neural Network)
  - ◆ GNN (Graph Neural Network)

- Problem definition
- Previous work
- Proposed approach
- Evaluation Methodology

Proposed approach

- Extract an enclosing subgraph of each target link
- Label the subgraphs using WL algorithm
- Encode the subgraph as an adjacency matrix
- Encode the features to vector
  - Publication title to a vector using word2vec transformation
- Split edges into train, validation, test set
- Feed the adjacency matrices along with feature vector to the graph neural network (GNN)
- Test and evaluate accuracy
- **Python -> pytorch-geometric**

Proposed approach



Extract subgraphs

Graph labeling

Data encoding

Graph Neural network

0	1	1	1
1	0	1	1
1	1	0	0
1	1	0	0

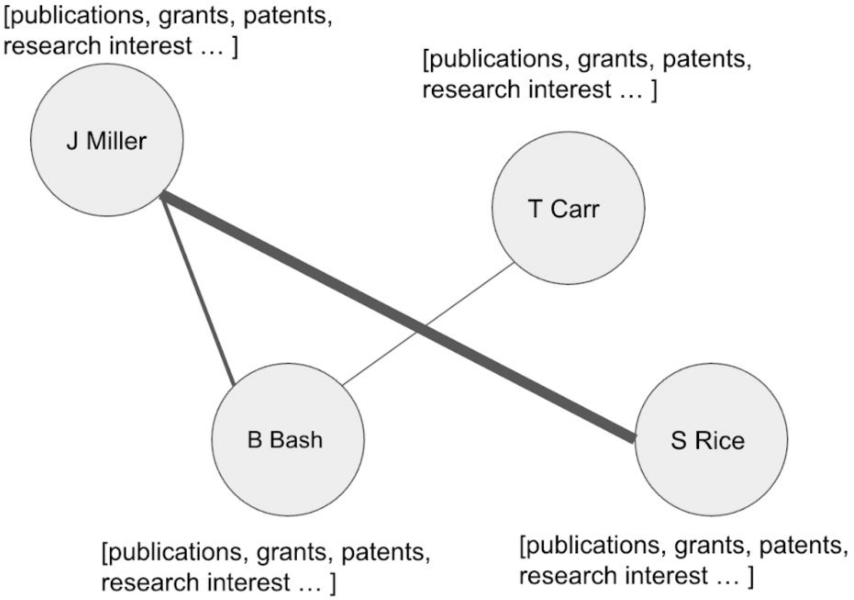
1

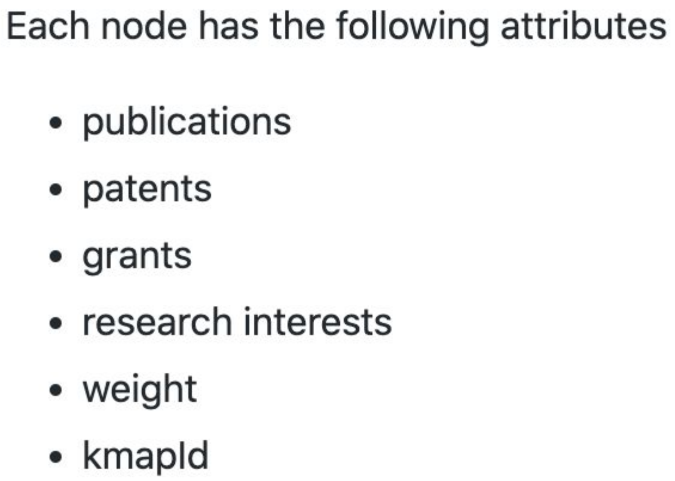
0	0	1	0
0	0	1	1
1	1	0	0
0	1	0	0

0

Data Encoding

- Encodes the subgraph as an adjacency matrix
- Encode **publication titles** (text data) to feature vector using word2vec

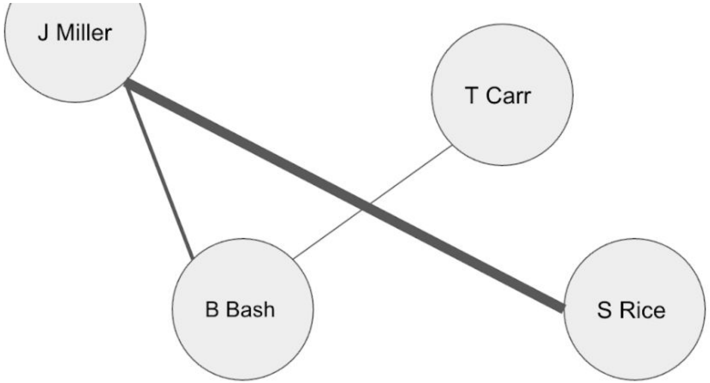

[publications, grants, patents, research interest …]
J Miller
[publications, grants, patents, research interest …]
T Carr
B Bash
S Rice
[publications, grants, patents, research interest …]
[publications, grants, patents, research interest …]

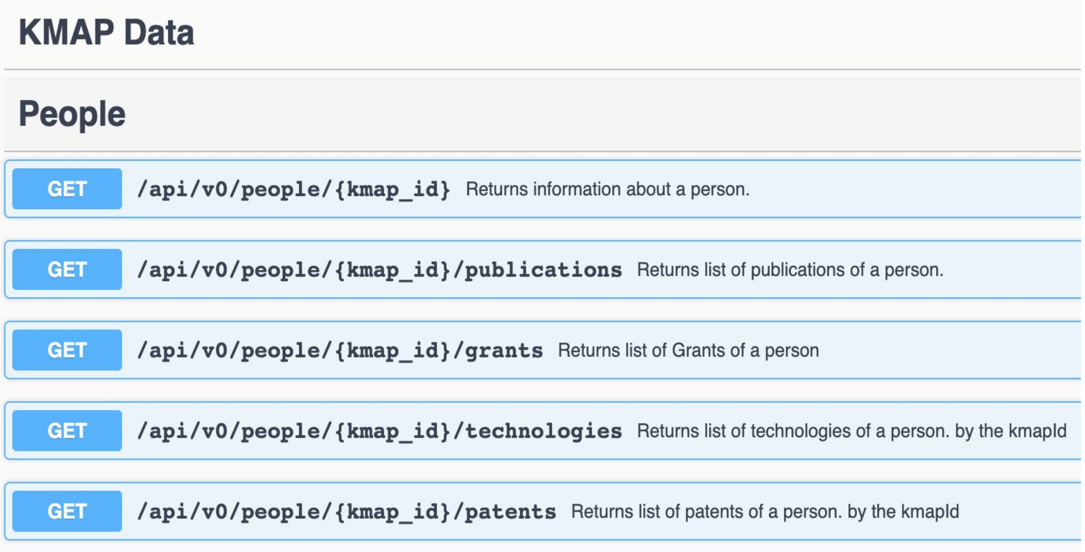


Each node has the following attributes

- publications
- patents
- grants
- research interests
- weight
- kmapId

Fetching Data





KMAP Data

People

GET `/api/v0/people/{kmap_id}` Returns information about a person.

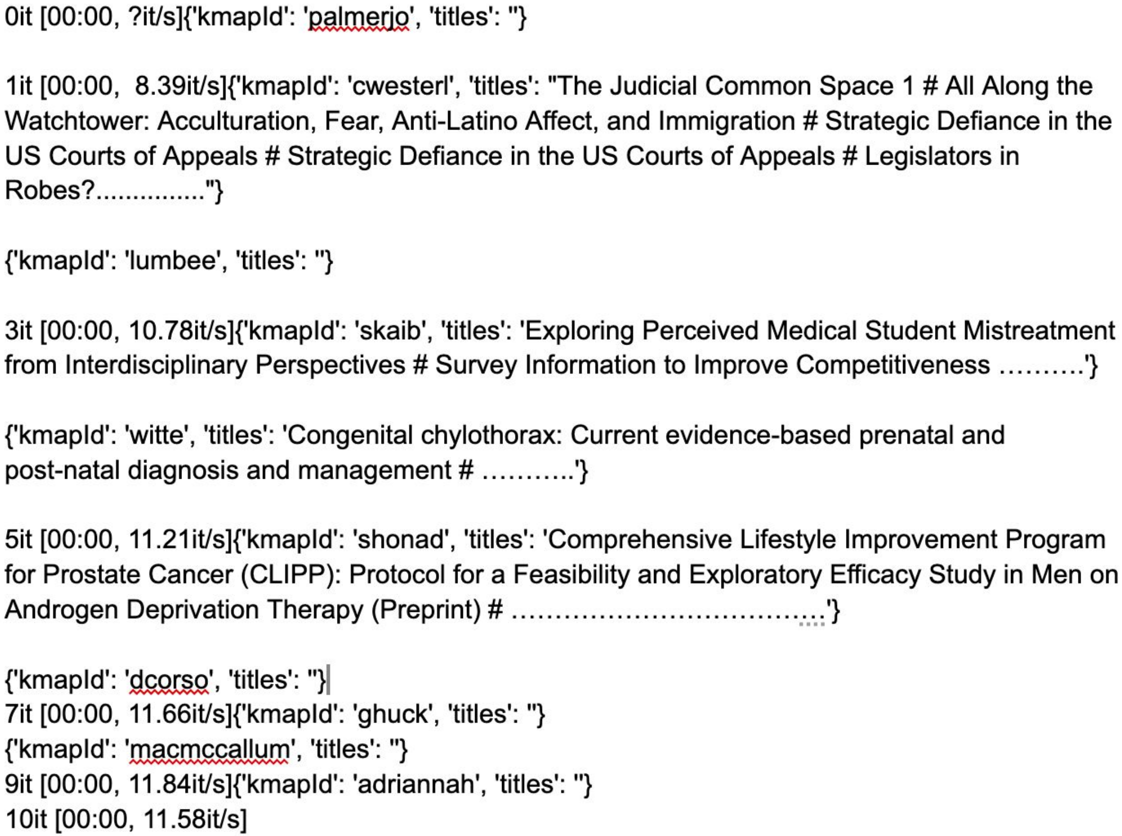
GET `/api/v0/people/{kmap_id}/publications` Returns list of publications of a person.

GET `/api/v0/people/{kmap_id}/grants` Returns list of Grants of a person

GET `/api/v0/people/{kmap_id}/technologies` Returns list of technologies of a person. by the kmapId

GET `/api/v0/people/{kmap_id}/patents` Returns list of patents of a person. by the kmapId

Data Cleaning



```
0it [00:00, ?it/s]{'kmapId': 'palmerjo', 'titles': ''}

1it [00:00,  8.39it/s]{'kmapId': 'cwesterl', 'titles': "The Judicial Common Space 1 # All Along the
Watchtower: Acculturation, Fear, Anti-Latino Affect, and Immigration # Strategic Defiance in the
US Courts of Appeals # Strategic Defiance in the US Courts of Appeals # Legislators in
Robes?..............."}

{'kmapId': 'lumbee', 'titles': ''}

3it [00:00, 10.78it/s]{'kmapId': 'skaib', 'titles': 'Exploring Perceived Medical Student Mistreatment
from Interdisciplinary Perspectives # Survey Information to Improve Competitiveness ……….'}

{'kmapId': 'witte', 'titles': 'Congenital chylothorax: Current evidence-based prenatal and
post-natal diagnosis and management # ………..'}

5it [00:00, 11.21it/s]{'kmapId': 'shonad', 'titles': 'Comprehensive Lifestyle Improvement Program
for Prostate Cancer (CLIPP): Protocol for a Feasibility and Exploratory Efficacy Study in Men on
Androgen Deprivation Therapy (Preprint) # ……………………………….'}

{'kmapId': 'dcorso', 'titles': ''}
7it [00:00, 11.66it/s]{'kmapId': 'ghuck', 'titles': ''}
{'kmapId': 'macmccallum', 'titles': ''}
9it [00:00, 11.84it/s]{'kmapId': 'adriannah', 'titles': ''}
10it [00:00, 11.58it/s]
```

Data Cleaning

# nodes : **4,884**
# edges/links : **19,241**
Average degree : **7.8792**

Remove nodes with empty publication data

- Problem definition
- Previous work
- Proposed approach
- Evaluation Methodology

Calculate correctness

given network G = (V, E)

- Take positive samples by selecting all edges (x,y) ∈ E.
- Take negative samples by randomly selecting α |E| pairs of x,y ∈ V such that (x,y) ∈/ E.
- Split both positive and negative samples to
  - 90% training set
  - 10% testing set
- Train GNN with training set
- Evaluate correctness using test set

Compare Performance

- Calculate accuracy using Stochastic Block Model
- Compare correctness

Thank You!