Clustering of Graph Embeddings

Dušica Knežević, Jela Babić

Department of Mathematics and Informatics
Faculty of Sciences, University of Novi Sad, Serbia

Overview

- Introduction
- Graph Embedding Methods
- Evaluation Methods
- Results
  - Intrinsic Evaluation
  - External Evaluation
- Conclusion and Future Work

Overview

- **Introduction**
- Graph Embedding Methods
- Evaluation Methods
- Results
  - Intrinsic Evaluation
  - External Evaluation
- Conclusion and Future Work

Introduction

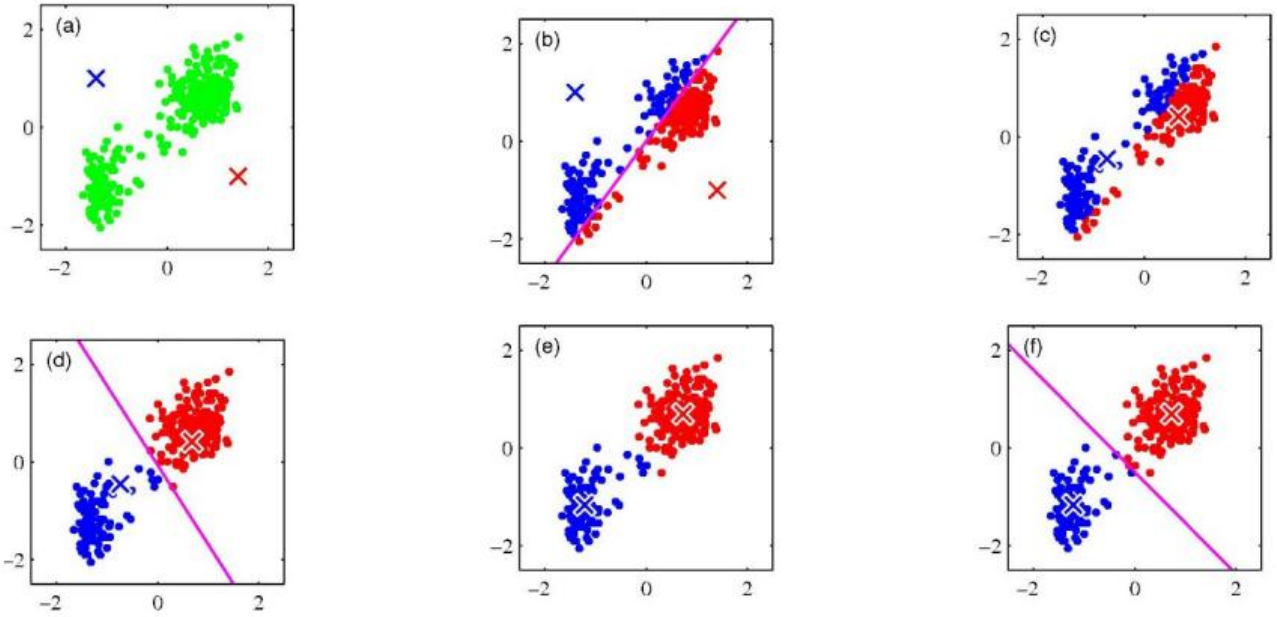
- **Graph embeddings**
  - Learning vector-based representations of graphs
  - Applications of tabular-based ML algorithms to graphs

- **Local Intrinsic Dimensionality (LID)**
  - Complexity of data space around data point
  - Theoretical LID framework by Houle (2013), based on distance distributions
  - NC-LID measure for graph nodes by Savić et al. (2021)
  - LID-elastic node2vec extensions based on NC-LID (Savić et al., 2021)

- **Motivation & contributions**
  - Savić et al. (2021): primary, general-purpose evaluation of LID-elastic node2vec extensions (graph reconstruction errors)
  - Evaluation of LID-elastic node2vec for a concrete application (node clustering with KMeans)

Introduction

- **Clustering**
  - Detecting groups of similar data points (nodes)
- **KMeans**
  - initial centroids: select K points randomly
  - assign each data point to the closest centroid
  - recompute centroids and repeat the previous step



Overview

- Introduction
- **Graph Embedding Methods**
- Evaluation Methods
- Results
  - Intrinsic Evaluation
  - External Evaluation
- Conclusion and Future Work

Graph Embedding Methods

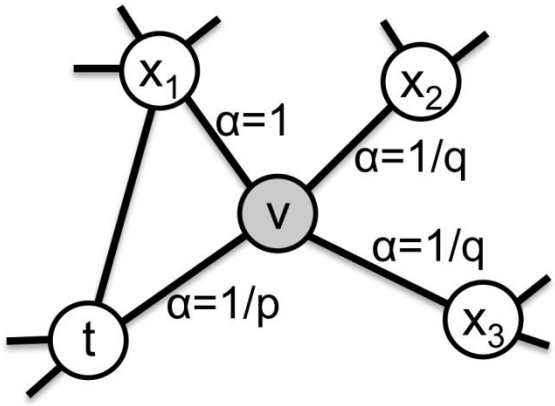
- **Taxonomy of graph embedding methods**
  - Methods based on matrix factorization
  - **Methods based on random walks**
  - Methods based on deep learning techniques (autoencoders, graph convolutional networks, etc.)

- **Random walk methods**
  - A certain number of random walks sampled from each node
  - Random walks treated as ordinary text
    → graph embedding reduced to text embedding
  - Representative methods: DeepWalk (Perozzi et al., 2014) and **node2vec** (Grover and Leskovec, 2016)
  - DeepWalk: unbiased random walk sampling
  - **Node2vec**: biased random walk sampling interpolating between BFS and DFS traversals

LID-elastic Node2vec Variants

- Node2vec hyper-parameters
  - NRW – the number of random walks starting from each node
  - LRW – the length of each random walk
  - *p* – parameter controlling the probability of immediately going back to the previous node in the random walk
  - *q* – parameter controlling the probability of going more into depth

- **lid-n2v-rw**
  - Personalizes NRW and LRW per node based on NC-LID so that high-LID nodes have a larger number of shorter walks



- **lid-n2v-rwpq**
  - Extension of lid-n2v-rw
  - Personalizes *p* and *q* for each pair of connected nodes according to their NC-LID values so that both coming back and leaving the natural communities of high-LID nodes is discouraged

Overview

- Introduction
- Graph Embedding Methods
- **Evaluation Methods**
- Results
  - Intrinsic Evaluation
  - External Evaluation
- Conclusion and Future Work

Clustering Evaluation Methods

- **Intrinsic evaluation**
  - No labels indicating cluster assignments
  - Silhouette score S
    - $-1 \leq S \leq 1$
    - Larger $S \rightarrow$ better clustering
    - Negative $S \rightarrow$ no clusters
- **External evaluation**
  - Labels indicating cluster assignments are explicitely given
  - Normalized mutual information (NMI)
    - compares two partitions into groups
    - $0 \leq \text{NMI} \leq 1$
    - Larger NMI $\rightarrow$ more similar partitions
  - Compared partitions
    - partition induced by explicit labels present in graph data
    - partitions obtained by community detection (greedy modularity optimization)
    - partitions identified by KMeans for different K values

Overview

- Introduction
- Graph Embedding Methods
- Evaluation Methods
- **Results**
  - Intrinsic Evaluation
  - External Evaluation
- Conclusion and Future Work

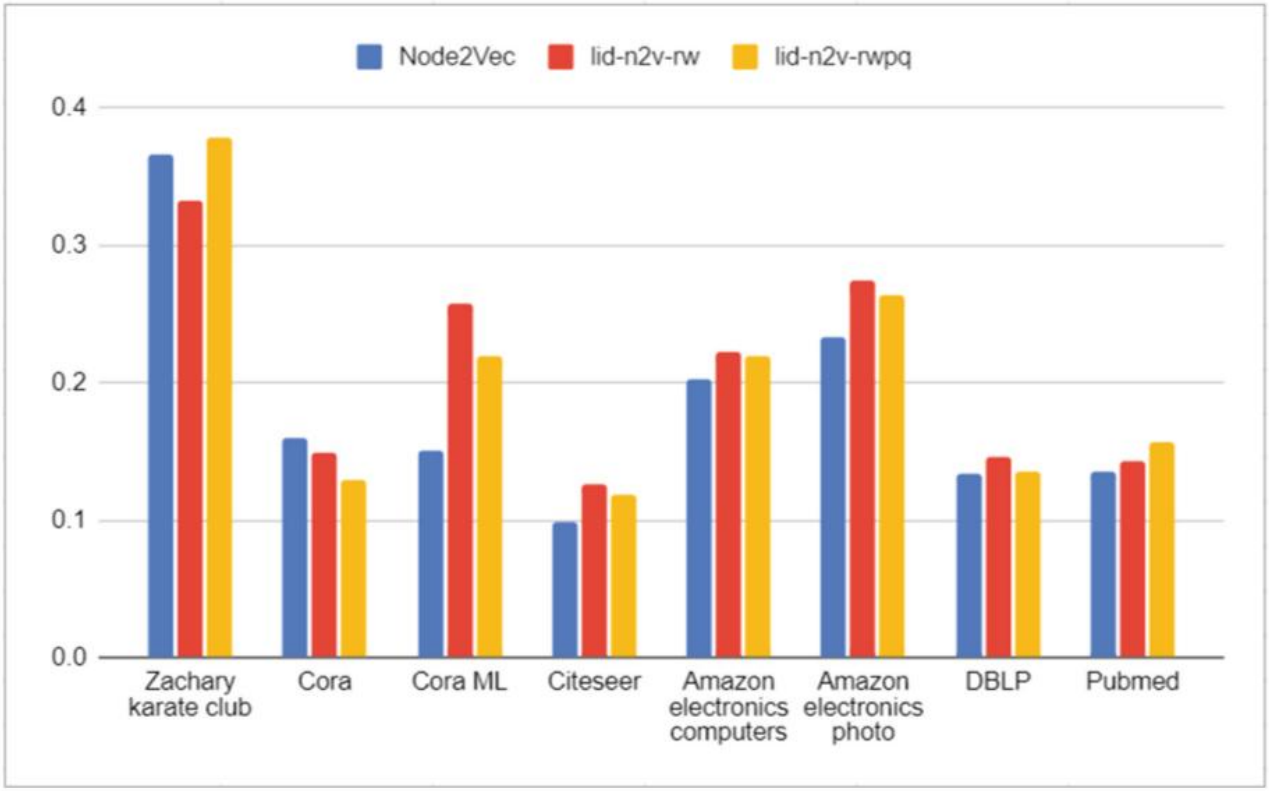
Results

- Experimental datasets: 8 real-world networks used in Savić et al. (2021)
- Hyperparameter tuning of graph embedding algorithms according to graph reconstruction errors as in Savić et al. (2021)
- Graph embeddings generated in five dimensions (10, 25, 50, 100, 200), the best NMI and Silhouette scores obtained in dimension 10
- Best values of hyperparameters p and q in dimension 10:

Dataset	node2vec		lid-n2v-rw		lid-n2v-rwpq	
	p	q	p	q	p	q
Zachary karate club	0.25	2	0.25	2	0.25	2
Cora	4	0.25	4	0.25	4	0.25
Cora ML	4	0.25	4	0.25	4	0.25
Citeseer	0.5	0.25	0.5	0.25	0.5	0.25
Amazon electronics computers	4	0.5	4	0.5	4	0.5
Amazon electronics photo	4	0.5	4	0.5	4	0.5
DBLP	4	1	4	1	4	1
Pubmed	2	0.5	2	0.5	2	0.5

Intrinsic evaluation

- The best Silhouette scores for KMeans clustering when K ≤ 10



- Lid-n2v-rw tends to have the highest Silhouette scores (5 out of 8 wins)
- Original node2vec wins only on one dataset (Cora)

External evaluation

- NMI scores for explicit labels and KMeans when K ≤ 10

Dataset	node2vec	lid-n2v-rw	lid-n2v-rwpq
Zachary karate club	0.693	0.826	0.727
Cora	0.545	0.523	0.418
Cora ML	0.548	0.583	0.597
Citeseer	0.489	0.577	0.572
Amazon electronics computers	0.554	0.554	0.569
Amazon electronics photo	0.649	0.675	0.657
DBLP	0.557	0.478	0.471
Pubmed	0.368	0.479	0.483

- Node2vec wins on 2 datasets, LID-elastic extensions win on 6 datasets
- Considerable improvements: Zachary, Citeseer and Pubmed

External evaluation

- NMI scores for explicit labels and KMeans when K is equal to the number of detected communities

Dataset	node2vec	lid-n2v-rw	lid-n2v-rwpq
Zachary karate club	0.727	0.826	0.861
Cora	0.546	0.548	0.545
Cora ML	0.640	0.639	0.651
Citeseer	0.857	0.855	0.858
Amazon electronics computers	0.403	0.404	0.401
Amazon electronics photo	0.489	0.489	0.486
DBLP	0.574	0.557	0.557
Pubmed	0.574	0.529	0.523

- Node2vec wins on 2 datasets, LID-elastic extensions win on 5 datasets
- Equal NMI scores for node2vec and lid-n2v-rw on AE photo

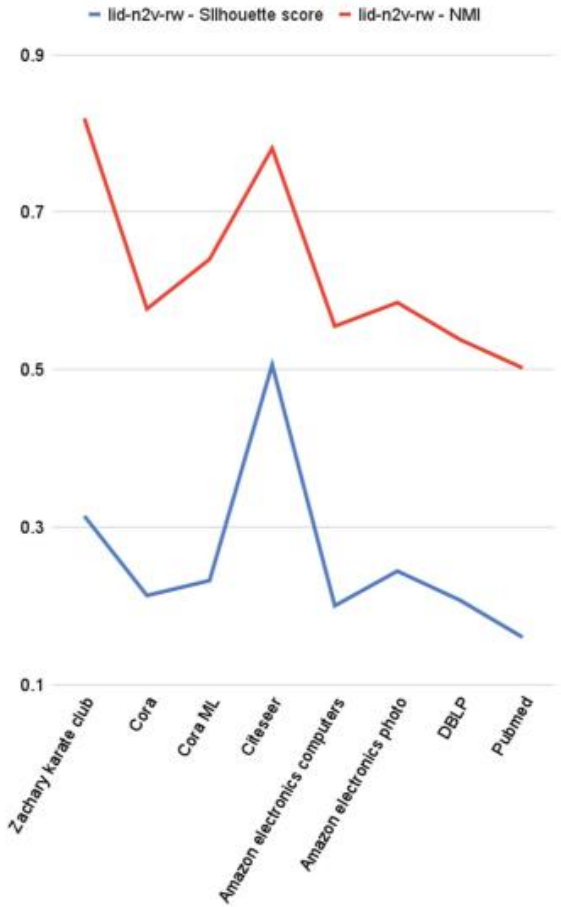
External evaluation

- The best NMI scores when K ≤ 10 and K is equal to the number of detected communities

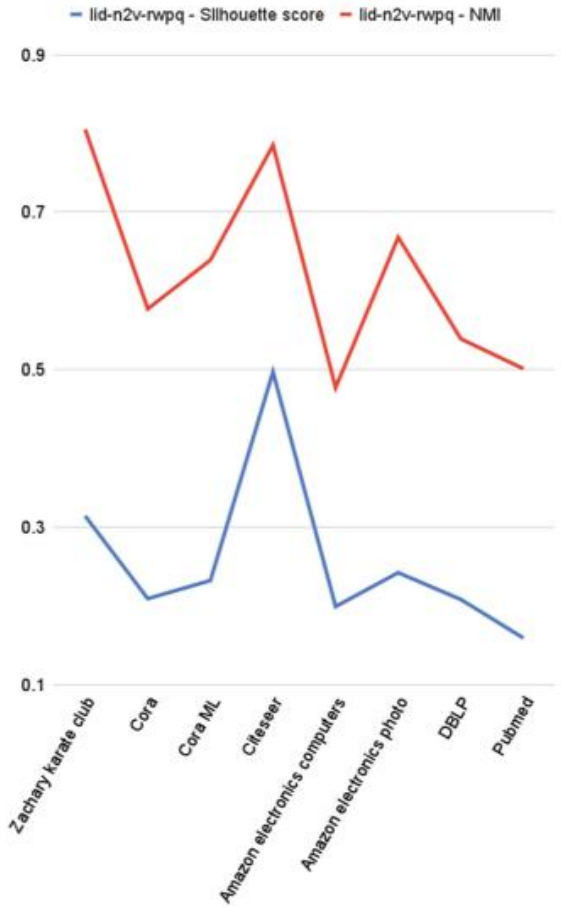
Dataset	Best NMI score	Method
Zachary karate club	0.861	lid-n2v-rwpq
Cora	0.548	lid-n2v-rw
Cora ML	0.651	lid-n2v-rwpq
Citeseer	0.858	lid-n2v-rwpq
Amazon electronics computers	0.569	lid-n2v-rwpq
Amazon electronics photo	0.675	lid-n2v-rw
DBLP	0.574	node2vec
Pubmed	0.574	node2vec

Evaluation consistency

- Silhouette and NMI scores are correlated



(a) lid-n2v-rw



(b) lid-n2v-rwpq

Overview

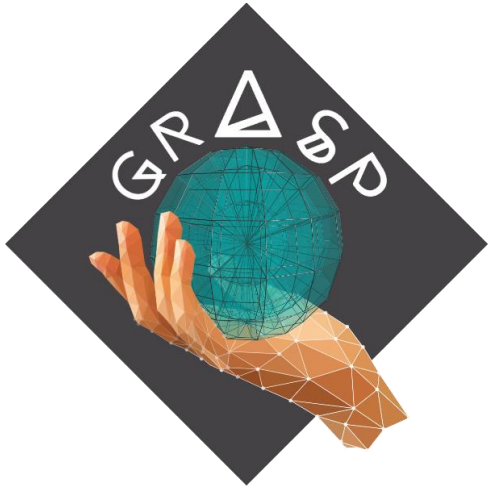
- Introduction
- Graph Embedding Methods
- Evaluation Methods
- Results
  - Intrinsic Evaluation
  - External Evaluation
- **Conclusion and Future Work**

Conclusion and Future Work

- LID-elastic extensions improve node2vec in node clustering
  ... especially when detecting a small number of clusters
- Intrinsic evaluation should be considered more reliable
  (explicitly given labels do not necesserily identify real clusters)
- In our case results of internal evaluation are consistent with results of external evaluation
- Future work:
  - Evaluation extended by including additional datasets, additional community detection algorithms and additional clustering algorithms for tabular data
  - New graph embedding methods based on biased random walks considering global communities
  - Random walk sampling strategies focused on nodes belonging to overlapping communities

Thank You!

QUESTIONS?



This research is supported by
the Science Fund of the Republic of Serbia
#6518241, AI – GRASP

Science Fund
of the Republic of Serbia