Node Clustering with Traditional and Neural Network Approaches

Ruizhe XIA, Fa ZHANG

HKUST

May 18, 2023

Introduction

- Graph is a widely existing type of data that can capture the dependency between samples.

Introduction

- Graph is a widely existing type of data that can capture the dependency between samples.
- Let $G = (V, E)$ be an undirected graph with vertex set $V = \{v_1, \cdots, v_n\}$, we use $i \sim j$ to denote that $v_i \in V$ is a neighbor of $v_j \in V$, i.e. $(i, j) \in E$.

Introduction

- Graph is a widely existing type of data that can capture the dependency between samples.
- Let $G = (V, E)$ be an undirected graph with vertex set $V = \{v_1, \cdots, v_n\}$, we use $i \sim j$ to denote that $v_i \in V$ is a neighbor of $v_j \in V$, i.e. $(i, j) \in E$.
- The adjacency matrix A of the graph G is the matrix

$$A_{ij} = \begin{cases} 1 & i \sim j \\ 0 & \text{otherwise} \end{cases}. \tag{1}$$

Introduction

- Graph is a widely existing type of data that can capture the dependency between samples.
- Let $G = (V, E)$ be an undirected graph with vertex set $V = \{v_1, \cdots, v_n\}$, we use $i \sim j$ to denote that $v_i \in V$ is a neighbor of $v_j \in V$, i.e. $(i, j) \in E$.
- The adjacency matrix A of the graph G is the matrix

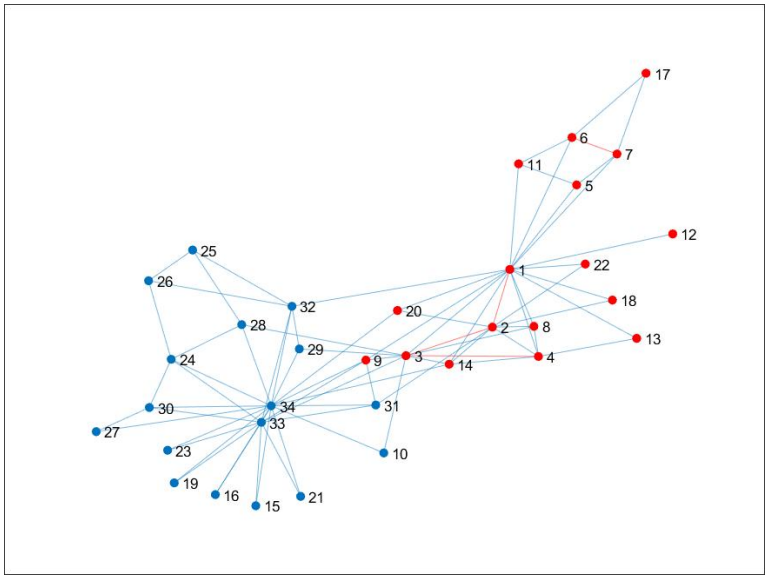
$$A_{ij} = \begin{cases} 1 & i \sim j \\ 0 & \text{otherwise} \end{cases}. \tag{1}$$

- The degree of a vertex $v_i \in V$ is defined as

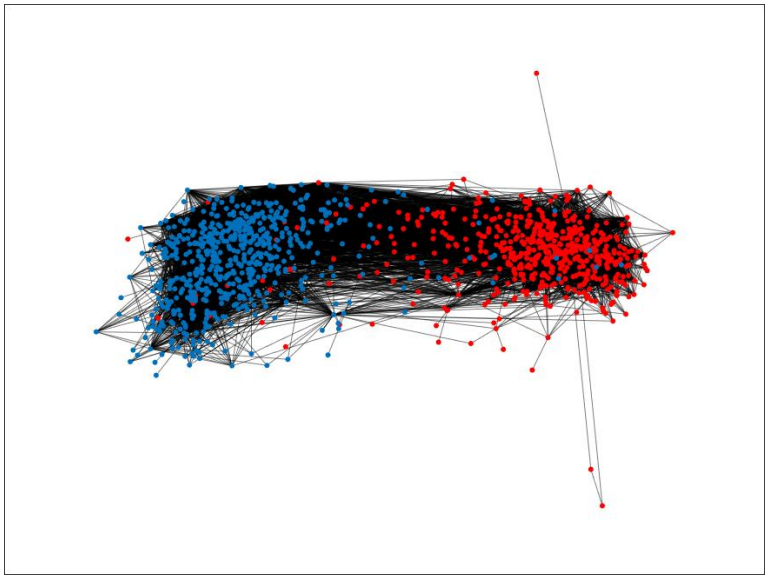
$$d_i = \sum_{j=1}^{n} A_{ij}, \tag{2}$$

and then we define a diagonal matrix $\mathrm{D} = \mathrm{diag}(d_i)$.

Datasets



(a) Zachary's Karate Club Dataset
Coach(red) & Owner(blue)



(b) Political Blogs Dataset
Liberal(red) & Conservative(blue)

Figure: Visualization

Problem

- Given an undirected and connected graph $G = (V, E)$, partition into two disjoint sets A, B, where $A \cup B = V$ and $A \cap B = \emptyset$.
- Metric

$$\text{Accuarcy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}. \tag{3}$$

Traditional Methods

- Spectral Clustering
  The Normalized Graph Laplacian

$$L_{sym} = D^{-1/2} L D^{-1/2}.$$

Traditional Methods

- Spectral Clustering
  The Normalized Graph Laplacian

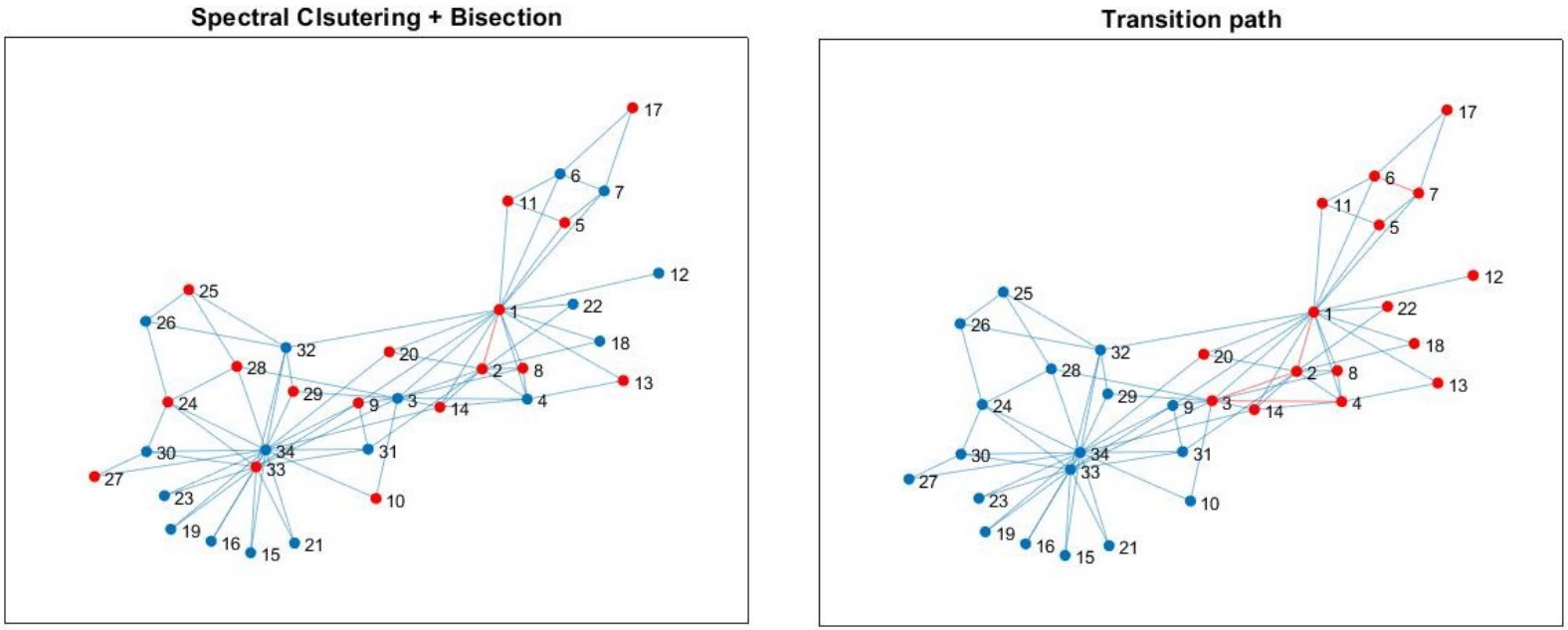
$$\mathrm{L}_{sym} = \mathrm{D}^{-1/2}\mathrm{L}\mathrm{D}^{-1/2}.$$

- Transition Path Theory
  The transition probability matrix

$$\mathrm{P} = \mathrm{D}^{-1}\mathrm{A}.$$

  Let $V = V_0 \cup V_1 \cup V_u$ be a partition of V. The committor function $q(i)$ gives the probability of first hitting V_1 before V_0.

Node Clustering

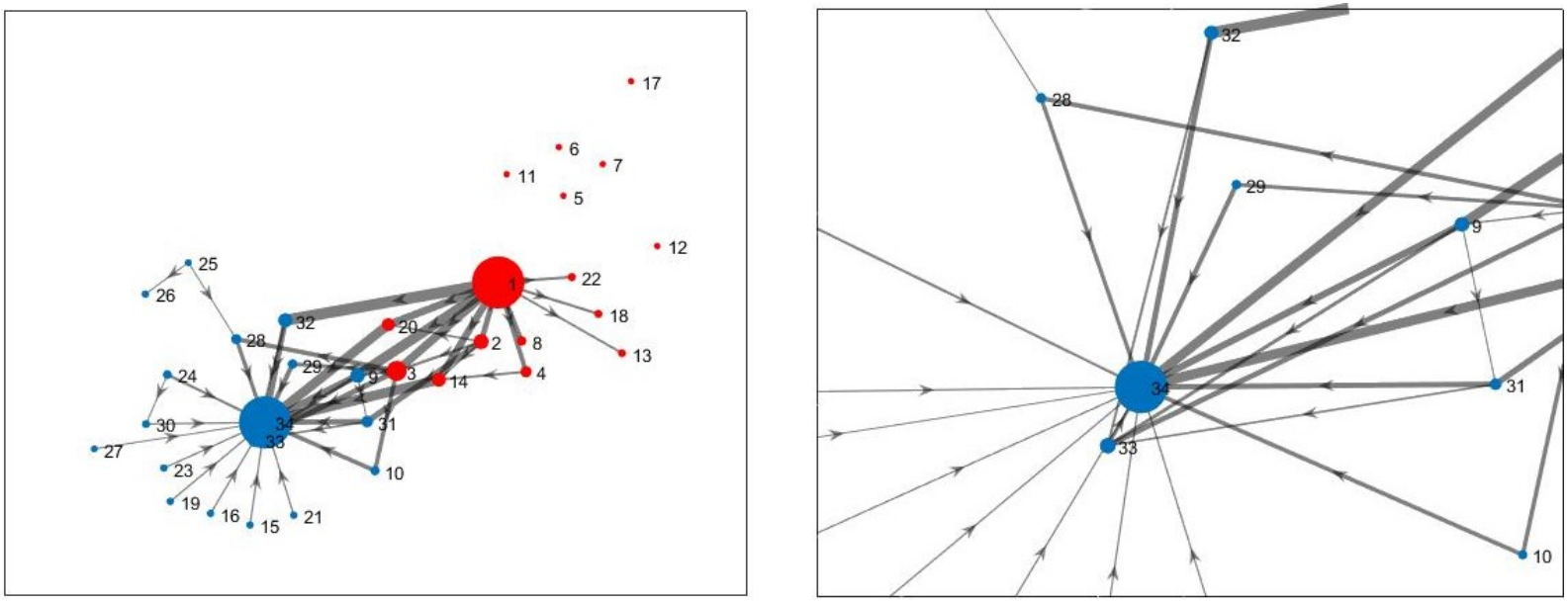


(a) Bisection(Acc = 58.82%) (b) Transition Path(Acc = 97.06%)

Figure: Traditional Methods

Transition Path (Zachary's Karate Club Dataset)

Important nodes: 2,3,9,14,20,32,33
Isolated nodes: 5,6,7,11,12,17



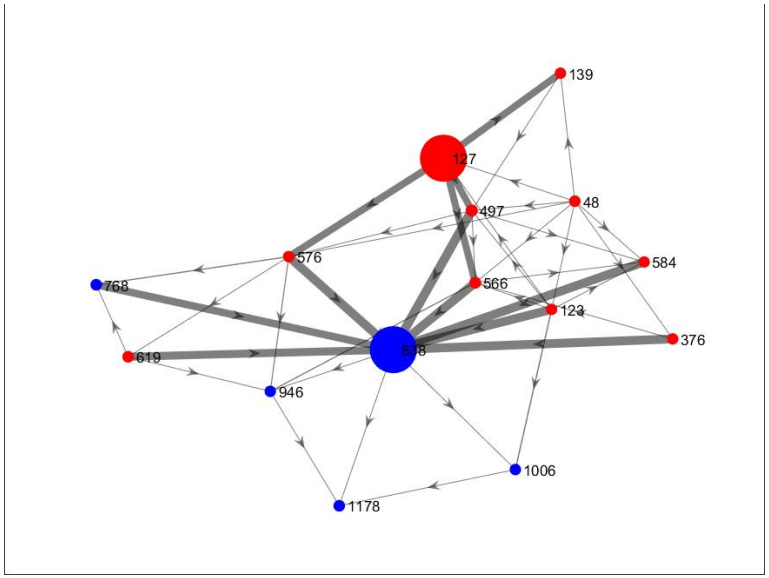
(a) Transition Path

(b) Zoom in Node 34

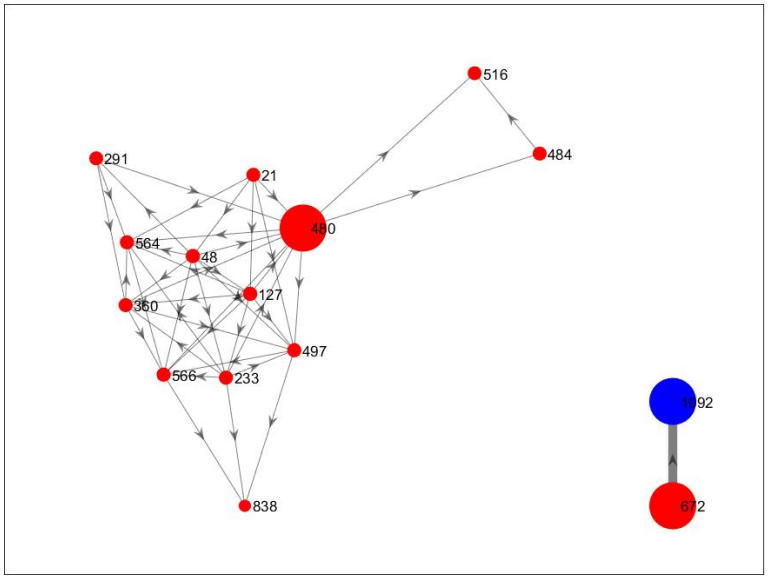
Figure: Effective/Transition Flux of Zachary's Karate Club Dataset

Transition Path(Political Blogs Dataset)

The labels matter.



(a) Good source/target state (Acc = 94.68%)



(b) Random source/target state (Acc = 48.04%)

Figure: Subgraph of Top 15 Nodes in Political Blogs Network

Neural Network Methods

- DeepWalk
  A method for learning embeddings of nodes. DeepWalk first uses random walks on a graph to generate sequences of nodes. And then uses the skip-gram model on these sequences to learn embeddings for each node.

Neural Network Methods

- Node2Vec
  Follow a similar process to DeepWalk, but Node2Vec uses a biased random walk strategy that captures both the local and global structure of the graph.

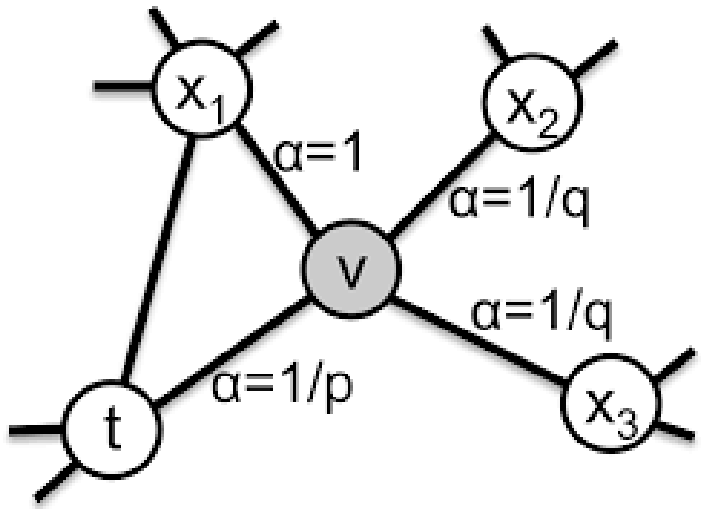


Figure: The walk transitioned from t to v and is evaluating its next step

Accuracy

Accuracy % Methods / Dataset	SC	TPT	DeepWalk	Node2Vec
Karate Club	58.82	97.06	97.06	97.06
Political Blogs	51.80	94.68*	95.51	96.16

Table: Accuracy

Perturbation Analysis

	1	2	3	4	5	6	7
SC	50.92+10.60	47.86+10.68	46.90+10.26	46.20+10.14	44.40+9.75	44.61+9.98	44.62+9.99
TPT	96.39+1.54	96.13+1.68	95.75+1.92	95.60+2.15	95.56+2.22	95.33+2.35	95.14+2.47
DeepWalk	96.76+0.01	96.53+0.01	96.24+0.01	96.20+0.02	96.12+0.01	96.12+0.01	96.02+0.02
Node2Vec	96.88+0.01	96.59+0.01	96.82+0.01	96.41+0.01	96.41+0.02	96.82+0.01	96.12+0.02

Table: Missing Edges

Summary

- The advantage of traditional methods is their interpretability. However, it loses the flexibility to capture complex nonlinear relationships, and it's also computationally expensive.

- The data-driven approach makes neural network methods very flexible. Besides, they are scalable and efficient.