A Latent Model for Representation Learning on Networks

Abdulkadir Çelikkanat and Fragkiskos D. Malliaros

Center for Visual Computing
CentraleSupélec, University of Paris-Saclay and Inria Saclay

Introduction

- *Network representation learning* (NRL) aims to encode a given network structure into low-dimensional vectors
- Applications in network analysis: visualization, classification, community detection and link prediction
- In this work, we propose:
  - TNE – Topical Node Embeddings
  - Enriched feature vectors using node and cluster information

Problem Formulation

Objective

For a given graph $G = (\mathcal{V}, \mathcal{E})$, the goal is to find a mapping function

$$\Phi : \mathcal{V} \rightarrow \mathbb{R}^d,$$

where $\Phi(v)$ indicates the representation of the vertex v in $\mathbb{R}^d$

- The objective function of random walk-based methods is:

$$\max_{\Phi,\tilde{\Phi}} \sum_{v} \sum_{u \in N_\gamma(v)} \log \Pr(\Phi(u)|\tilde{\Phi}(v))$$

- Approximation of the function above:

$$\max_{\Phi,\tilde{\Phi}} \sum_{w \in \mathcal{W}} \sum_{v_i \in w} \sum_{-\gamma \leq j \leq \gamma} \log \Pr(\Phi(v_{i+j})|\tilde{\Phi}(v_i))$$

Random Walks and Communities

Many real-world networks can be expressed as a combination of nested or overlapping communities

- Each random walk can be represented as random mixtures over latent communities
- Each community can be characterized by a distribution over nodes

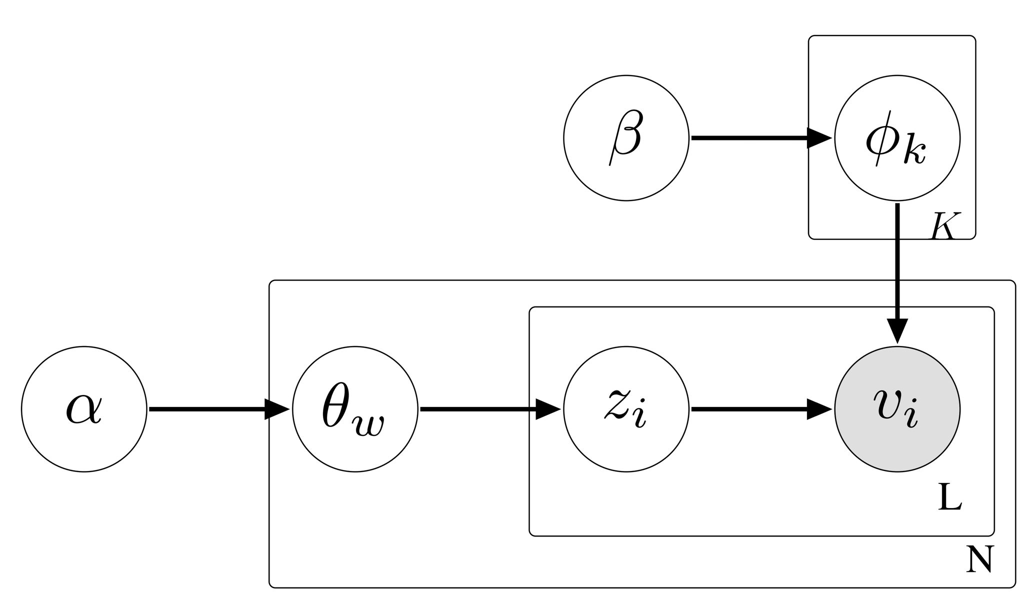

Figure: Graphical representation of the Latent Dirichlet Allocation (LDA) model

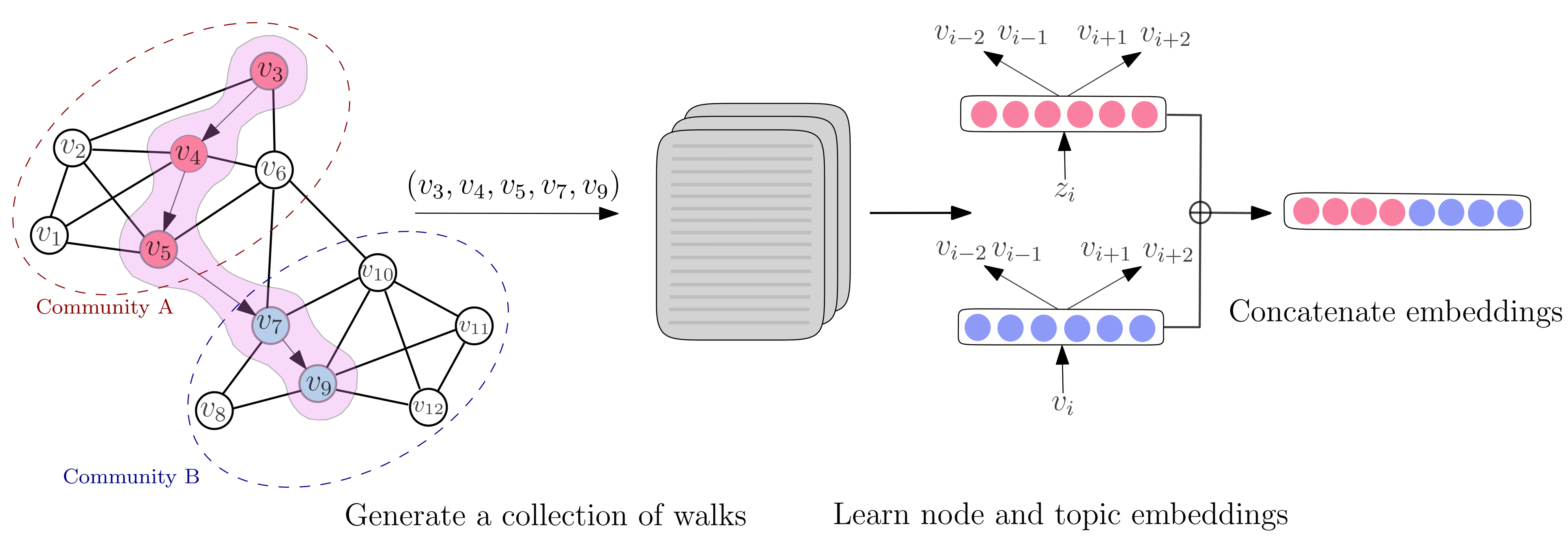

Figure: Schematic representation of the TNE model

Topical Node Embeddings (TNE)

- Let $t_w(v)$ be a community/topic assignment of a node v in the walk $w \in \mathcal{W}$
- The objective function to learn *topic embeddings* is:

$$\max_{\Psi,\tilde{\Psi}} \sum_{w \in \mathcal{W}} \sum_{v_i \in w} \sum_{-\gamma \leq j \leq \gamma} \log \Pr(\Psi(v_{i+j})|\tilde{\Psi}(t_w(v_i)))$$

- The final embedding vector is obtained by combining node and community embeddings

$$\Phi(v) \oplus \tilde{\Psi}(k^*),$$

where $k^* = \arg\max_k \Pr(\Phi(u)|\tilde{\Psi}(k))$

Experimental Results

- Downstream tasks: node classification and link prediction

	CiteSeer	DBLP	P2P	FB	arXiv
# Vertices	3,312	27,199	6,301	4,039	17,903
# Edges	4,598	66,832	2,077	88,234	197,031
# Clusters	6	4	-	-	-

Table: Networks used in the experiments

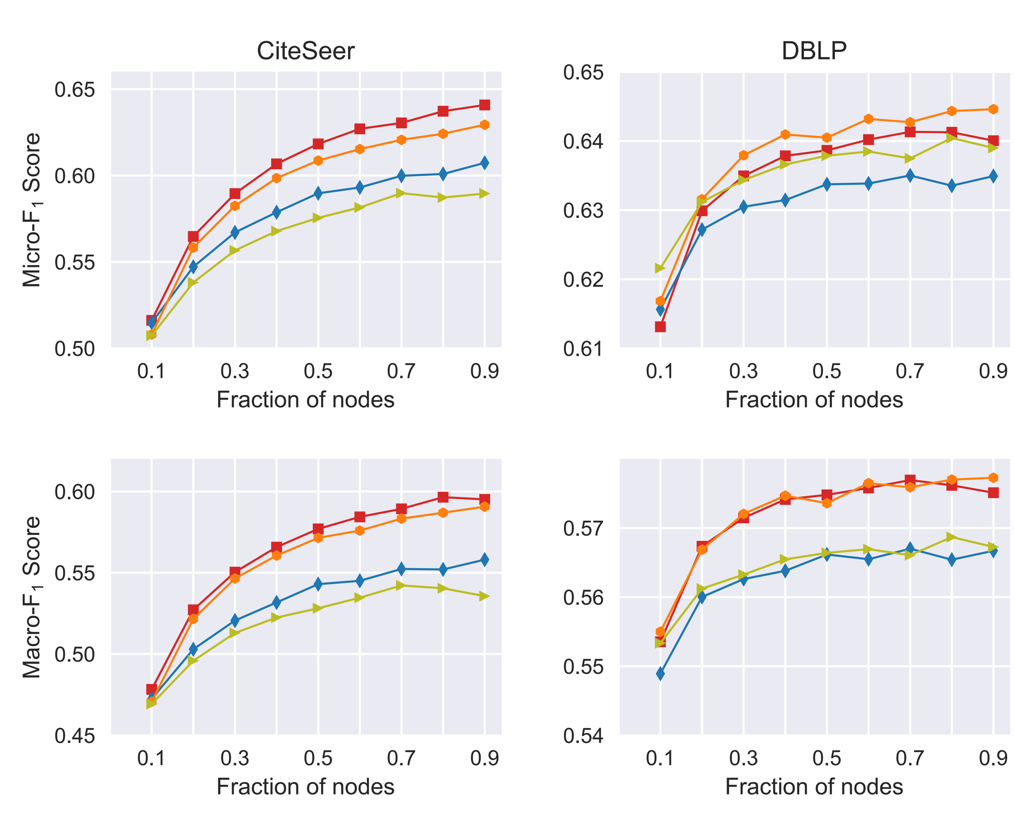

Figure: Micro-F_1 and Macro-F_1 scores for node classification over two different networks

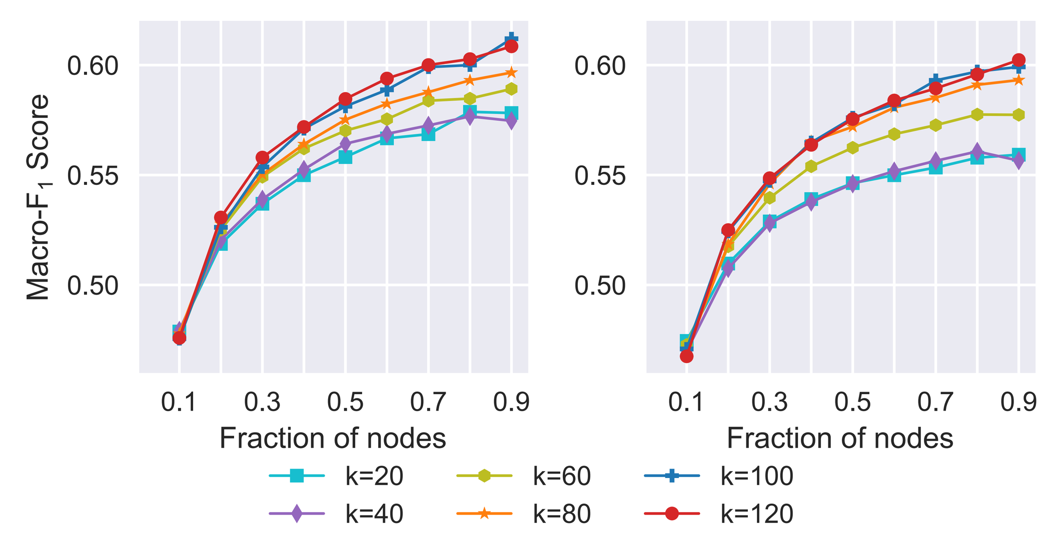

Figure: Varying number of topics over CiteSeer

	Algorithm	Dataset P2P	Facebook	arXiv
(a)	TNE(DW)	0.5684	**0.9834**	0.9270
	DeepWalk	0.5664	0.9812	0.9210
	TNE(N2V)	**0.5703**	0.9810	**0.9299**
	Node2Vec	0.5684	0.9809	0.9246
(b)	TNE(DW)	0.6928	**0.7686**	0.7851
	DeepWalk	**0.6943**	0.7626	**0.7854**
	TNE(N2V)	0.6914	0.7671	0.7847
	Node2Vec	0.6928	0.7634	0.7848
(c)	TNE(DW)	0.4931	**0.9319**	0.8403
	DeepWalk	0.4336	0.8863	0.7556
	TNE(N2V)	**0.5525**	0.9077	**0.8438**
	Node2Vec	0.4931	0.8988	0.7653

Table: **Link prediction:** AUC scores for (a) Hadamard, (b) Average and (c) Weighted-L1

Operator	Definition
Hadamard	$v \circ u$
Average	$0.5 \cdot (v + u)$
Weighted-L1	$\lvert v - u \rvert_1$

Table: Operators for learning edge features

References

[1] Yang Liu, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun.
Topical word embeddings.
In *AAAI*, 2015.

[2] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena.
Deepwalk: Online learning of social representations.
In *KDD*, 2014.

[3] Aditya Grover and Jure Leskovec.
Node2vec: Scalable feature learning for networks.
In *KDD*, 2016.

[4] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean.
Distributed representations of words and phrases and their compositionality.
In *NIPS*, 2013.

Project website: `github.com/abdcelikkanat/TNE`   `{firstname, lastname}@centralesupelec.fr`