Network Representation Learning

First Year Review

Benedek András Rózemberczki

The University of Edinburgh

October 30, 2018

Overview

Motivation – Challenges

1. Large network sizes.
2. Expressiveness – signed edges, features, temporal patterns.
3. Noisy nature of data.
4. Transferring knowledge is hard.

Application – Drug Discovery

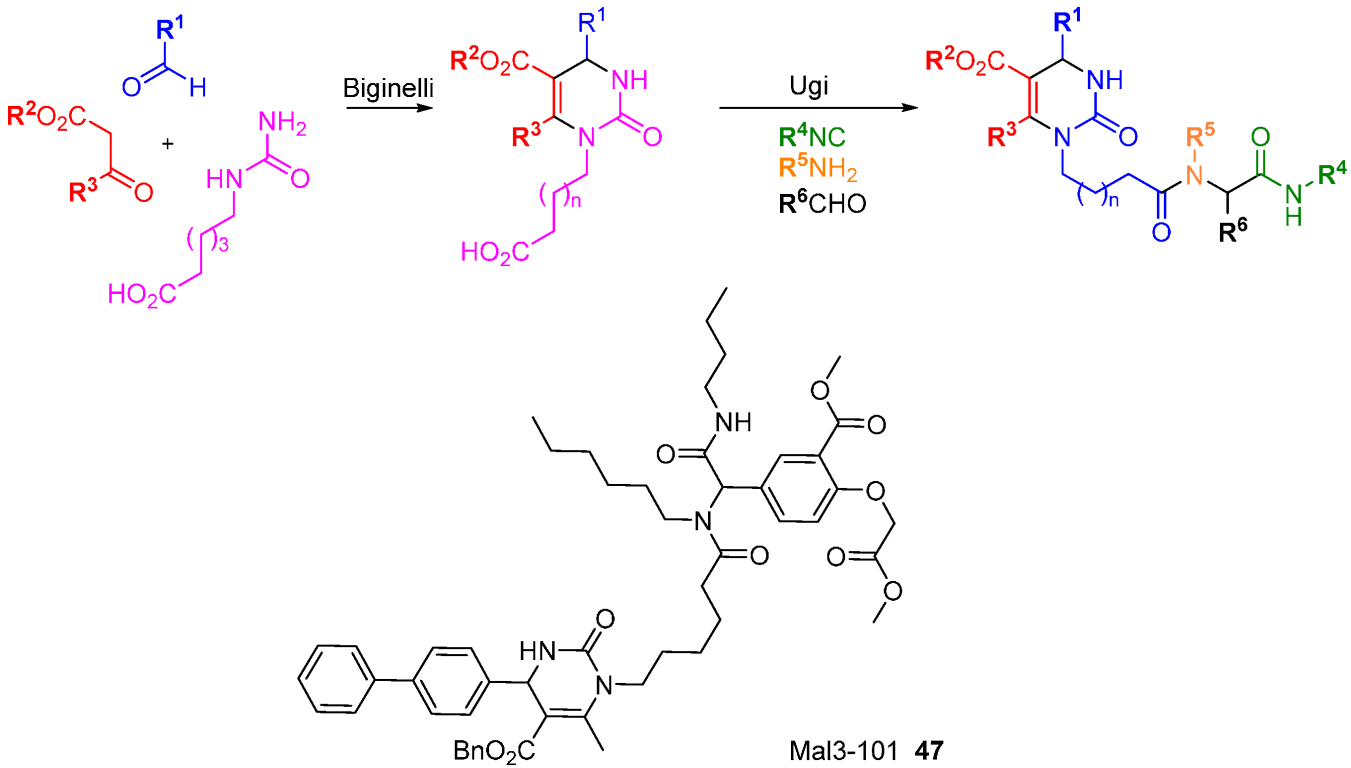

Figure 1: We could find drugs or just make inference about properties of molecules (Gärtner et al., 2003).

Application – Data Imputation

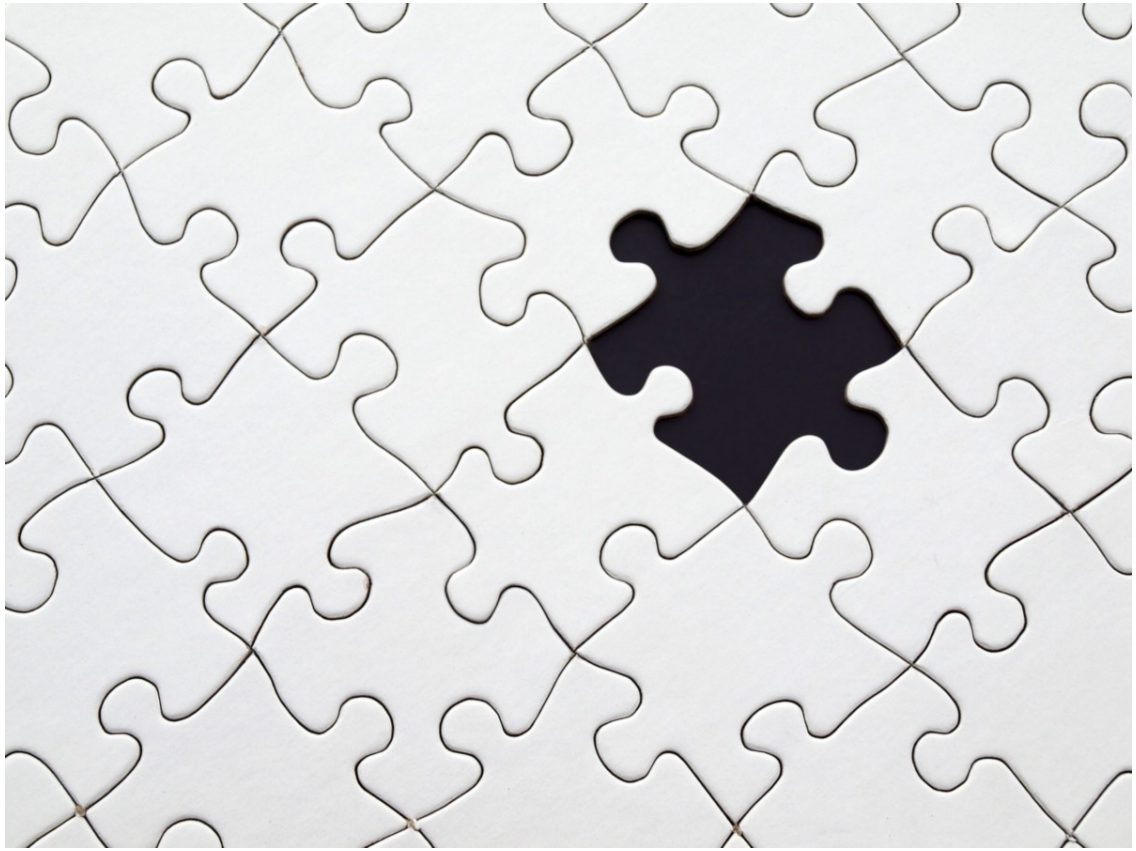

Figure 2: We could impute missing feature values (Perozzi, Skiena, 2015).

Application – Recommendations

Figure 3: We could recommend items to people (Chen et al., 2016).

Application – Fraud Detection

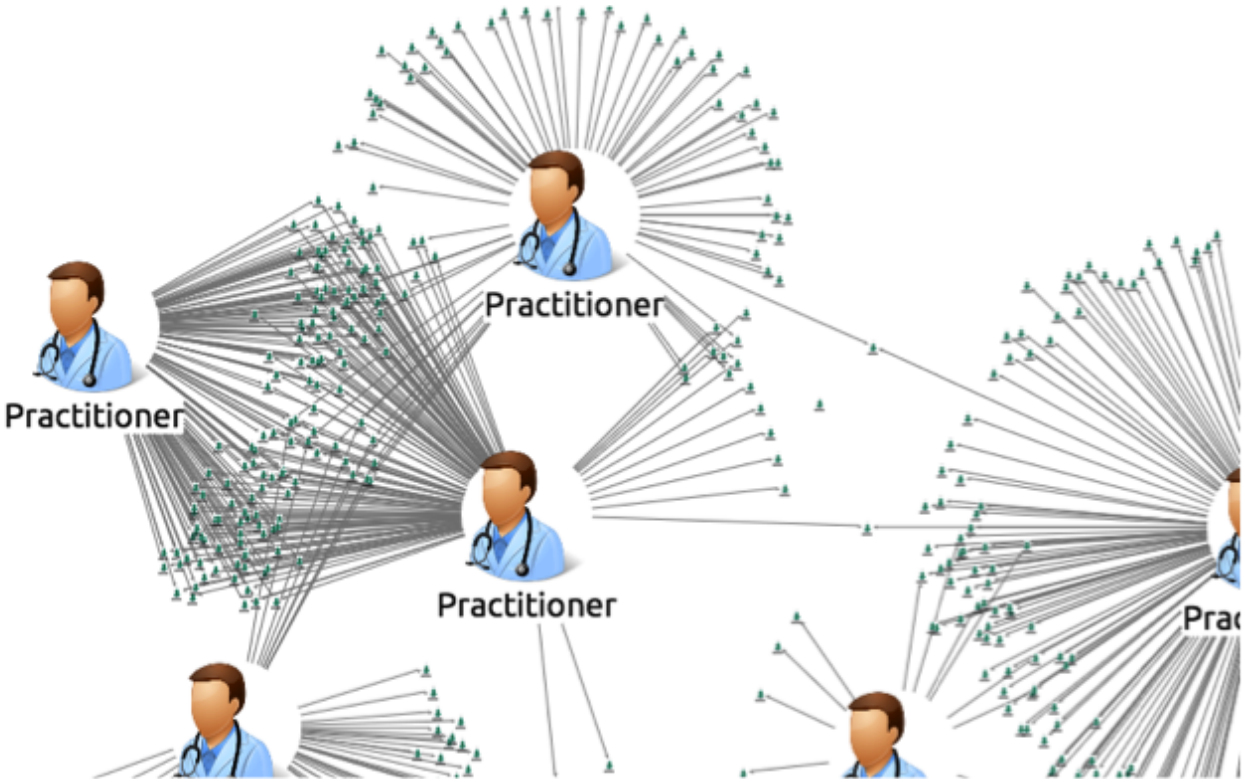

Figure 4: We could classify fraudulent transaction series (Kriege et al., 2016).

Future research projects

1. Graph Wavelet Sketching.
2. Multi-scale Attributed Node Embedding.
3. Graph Embedding with Structural Regularization.

Research Project I.: Graph Wavelet Sketches

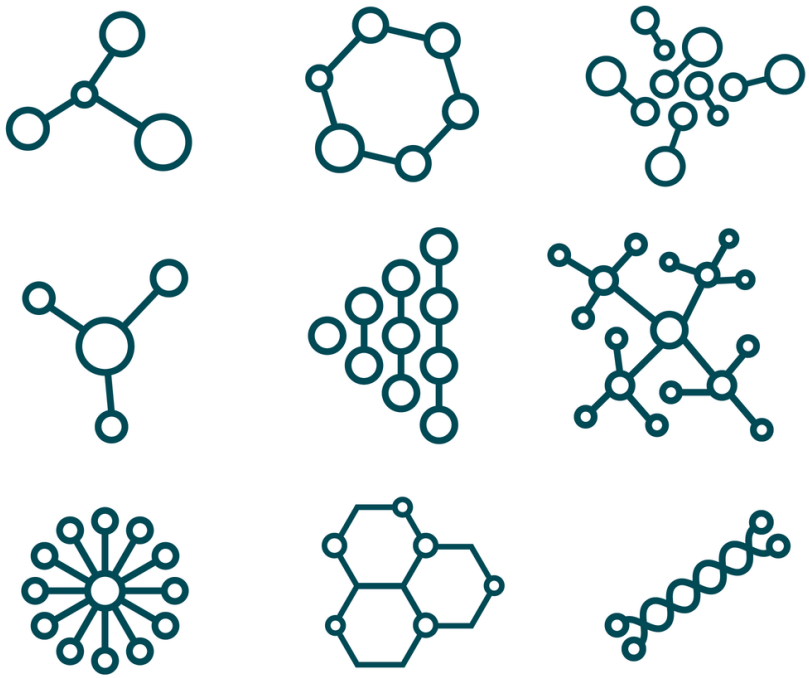

Figure 5: How can we measure the similarity of graphs? How can we capture structural similarity? Molecules might have different sizes, node features might be missing.

The motivation

1. Graph kernels are slow and features are non expressive.
2. Current methods (e.g. factorization) are transductive.
3. In addition these methods do not work in a decentralized setting.
4. How can we conserve node level information?

What is a Graph Wavelet? (Donnat et al., 2018)

A wavelet of a node is a distribution. It characterizes the ability of the node to transfer information to other nodes in the graph. Now L is the Laplacian of $G(V, E)$:

$$L = D - A = U\Lambda U^T$$

Where U is the set of eigenvectors and $\Lambda = Diag(\lambda_1, \ldots, \lambda_N)$. Let us define the heat kernel with scale s as:

$$g_s(\lambda) = e^{-\lambda s}$$

The wavelet specific to node v is:

$$\Psi(v) = U \cdot \text{Diag}(g_s(\lambda_1), \ldots, g_s(\lambda_N)) \cdot U^T \cdot \delta_v$$

The characteristic function transform at $t \in T$ of the wavelet is:

$$\psi(v)_t = \frac{1}{|V|} \sum_{u \in V} e^{i \cdot t \cdot \Psi_u(v)}$$

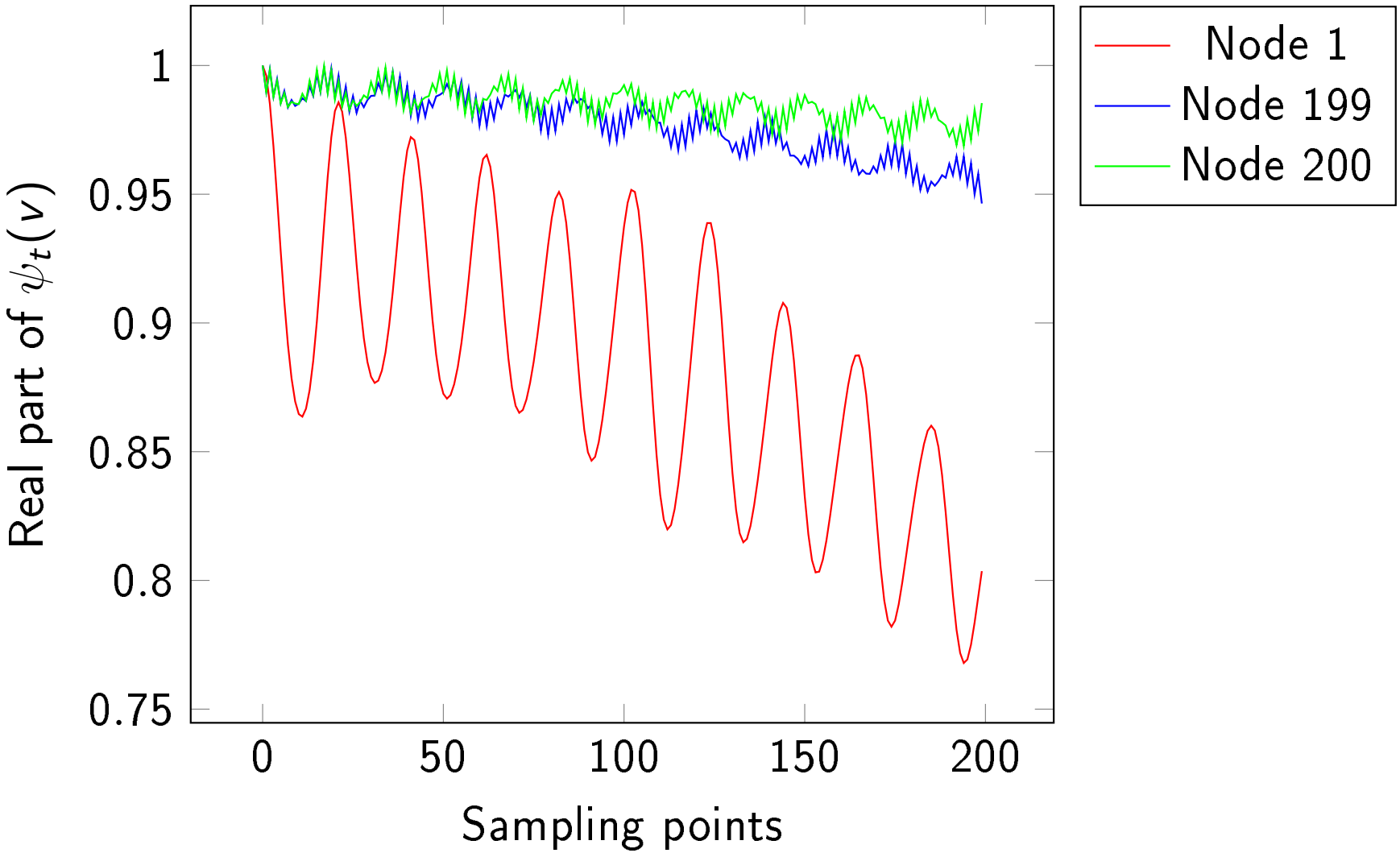

Figure 6: Characteristic function transform of three wavelets for a Barabasi-Albert tree with 200 nodes.

The Wavelet Sketch Algorithm

How can we represent graphs of a graph database so they are comparable?

- For each $G \in \mathcal{G}$:
  - Create a sketch matrix/tensor.
  - For each $v \in V_G$:
    - Calculate or approximate $\Psi(v)$.
    - Sketch $\psi(v)_t$ at each $t \in T$ and increment the corresponding matrix/tensor values of the sketch.
  - Normalize the sketch.
  - Save the sketch.

Application – Molecular Graph Classification

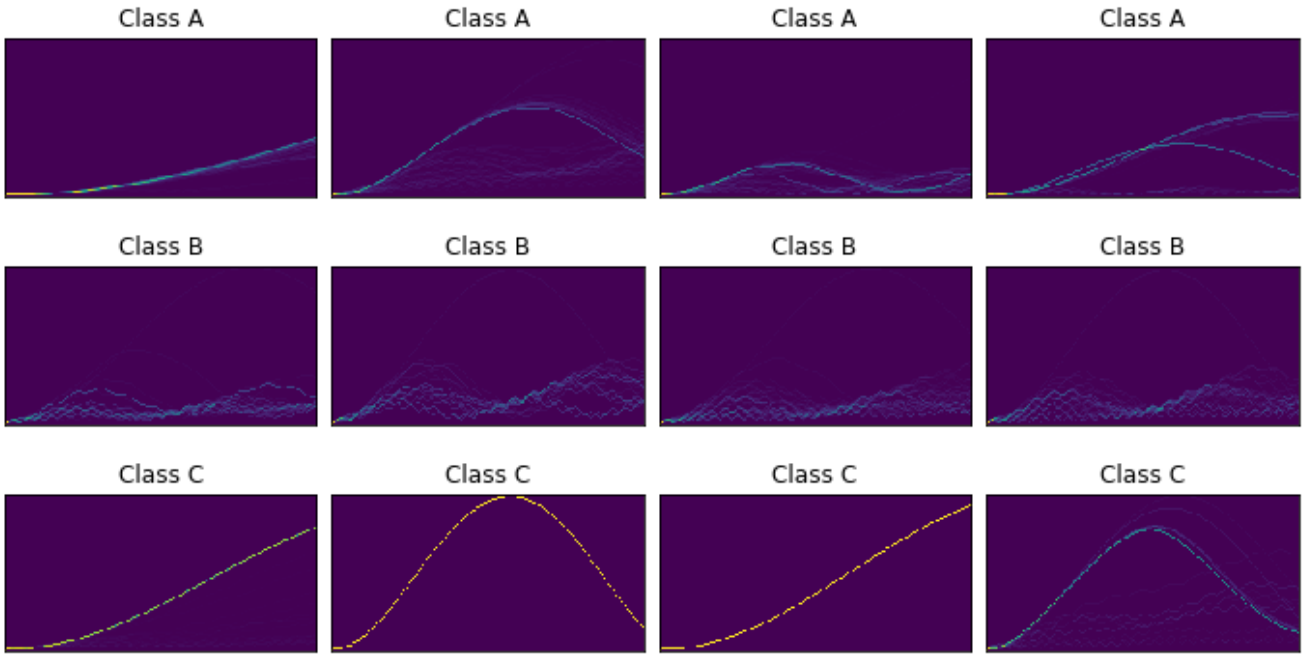

Figure 7: Graph Wavelet Sketches of molecular graphs in MUTAG () characterize graphs in an interpretable way.

Research Project II.: Multi-scale Attributed Embedding

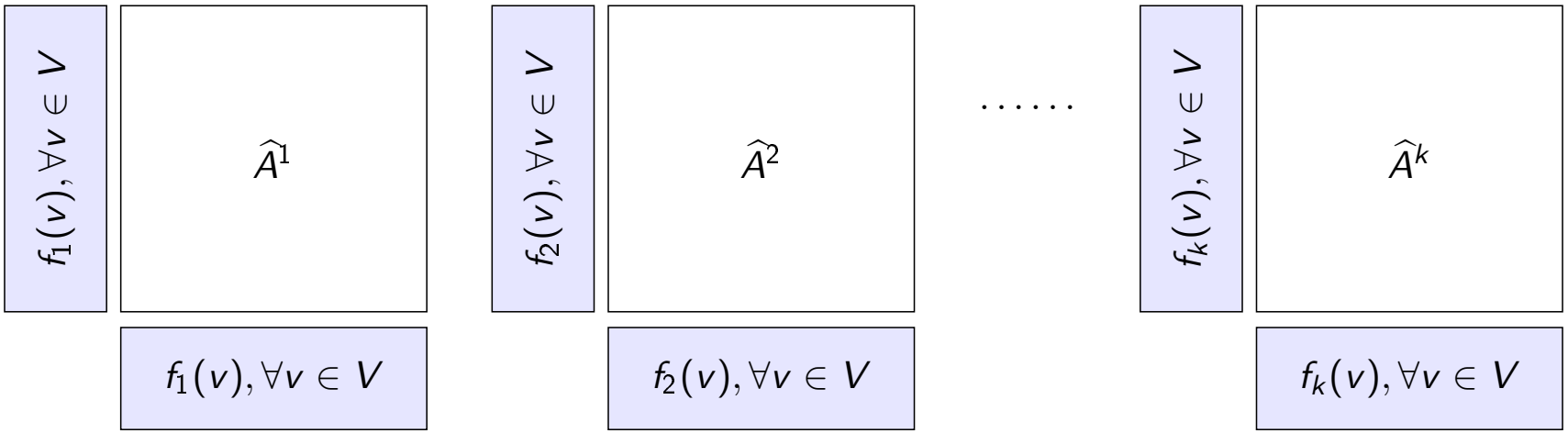

Figure 8: A multi-scale neighborhood preserving node embedding factorizes each power of the adjacency matrix up to order k independently with d dimensional factors. See for example Perozzi et al. (2017).

Attributed Node Embedding

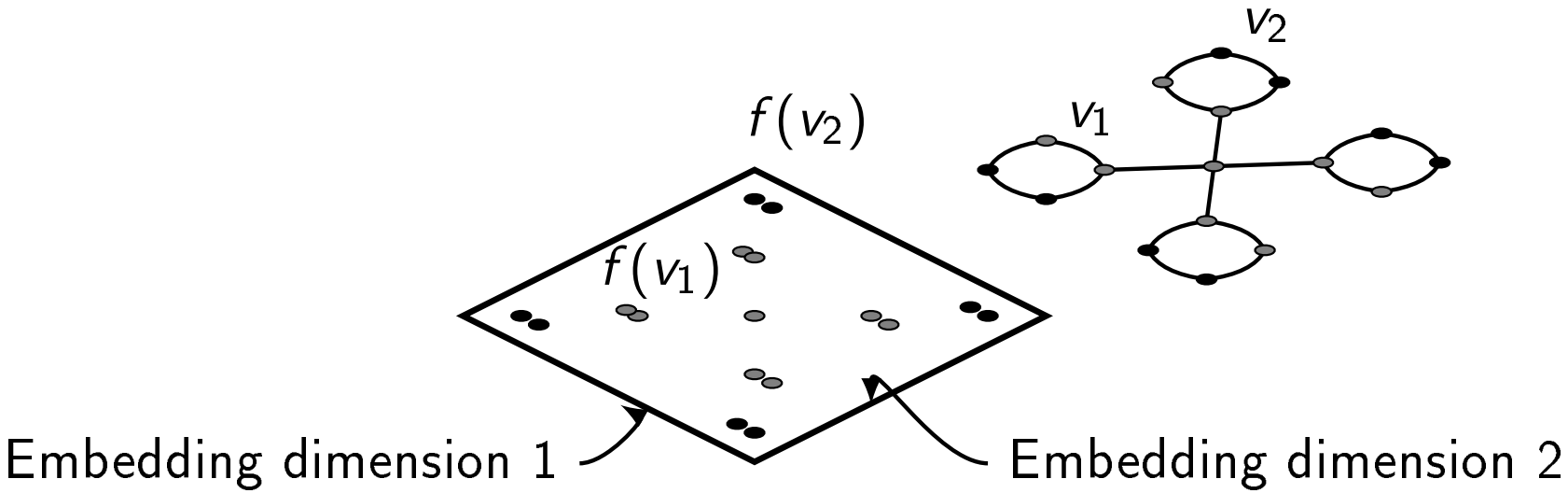

Figure 9: Attributed neighborhood preserving node embedding procedures preserve proximity on the graph between pairs of nodes in the embedding space and take the spatial autocorrelation of node features into account.

Multi-Scale Attributed Embedding – The Intuition

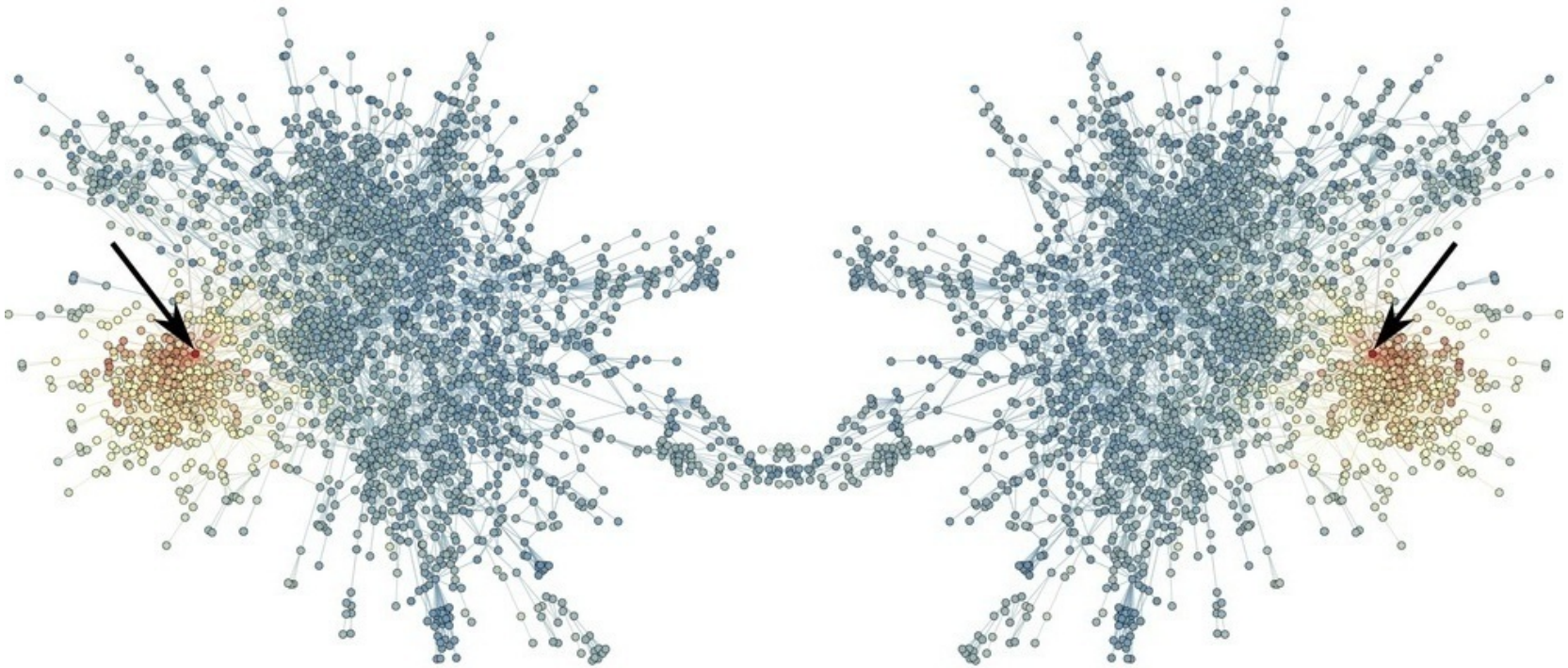

Figure 10: We want to characterize the features that neighbors at different proximities have. Two nodes from very different communities in fact might have similar feature distributions in their proximity.

MUSAE Model

The loss function of the model:

$$\|(\mathbf{D}^{-1}\mathbf{A})^r\mathbf{X} - \mathbf{W_r}\mathbf{H_r}\|_F^2 \qquad \forall r = 1, \ldots, k$$

The notation:

- $\mathbf{D}$ – Diagonal degree matrix.
- $\mathbf{A}$ - Adjacency matrix.
- $\mathbf{X}$ - Feature matrix.
- r – Proximity index.
- $\mathbf{W}_r$, $\mathbf{H}_r$ Node and feature embedding at proximity r.

Predictive Performance – Lifetime Prediction on Twitch

Let us have a case study on using MUSAE.

We compare the MUSAE predictive performance to neighbourhood based and attributed node embeddings:

1. Attributed: TADW, AANE, ASNE, SINE
2. Neighbourhood based: LINE, DeepWalk, Walklets
3. Feature matrix factorization.

Research project III.: Structural Regularization of Graph Embeddings

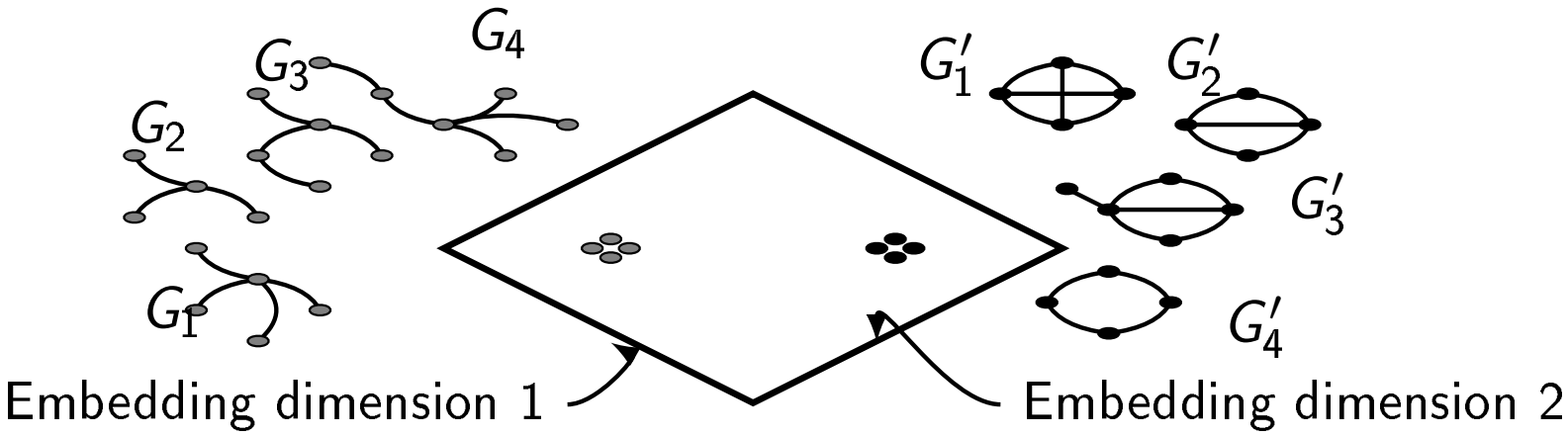

Figure 11: Graph embedding algorithms create low dimensional representations of whole graphs in an embedding space. Those graphs that have similar structural properties (labeled random walks, shortest paths, subtree patterns, graphlets) are located close in the embedding space.

Model Description – Based on Narayanan et al. (2016)

let us define a probabilistic model:

$$\min_{g} \quad \sum_{G \in \mathcal{G}} -\log P(\mathcal{F}_G | g_G).$$

Assuming inner product parametrization and conditional independence:

$$\min_{g,h} \sum_{G \in \mathcal{G}} \left[\ln \left(\sum_{f \in \mathcal{F}} \exp(h_f \cdot g_G) \right) - \sum_{f_i \in \mathcal{F}_G} h_{f_i} \cdot g_G \right]$$

This can be regularized such that:

$$\min_{g,h} \sum_{G \in \mathcal{G}} \left[\ln \left(\sum_{f \in \mathcal{F}} \exp(h_f \cdot g_G) \right) - \sum_{f_i \in \mathcal{F}_G} h_{f_i} \cdot g_G \right] + \underbrace{\sum_{(u,v) \in G_F} \lambda \cdot w_{u,v} \cdot \| h_{f_u} - h_{f_v} \|_2}_{\text{Regularization term}}$$

How Can I Get the Weights?

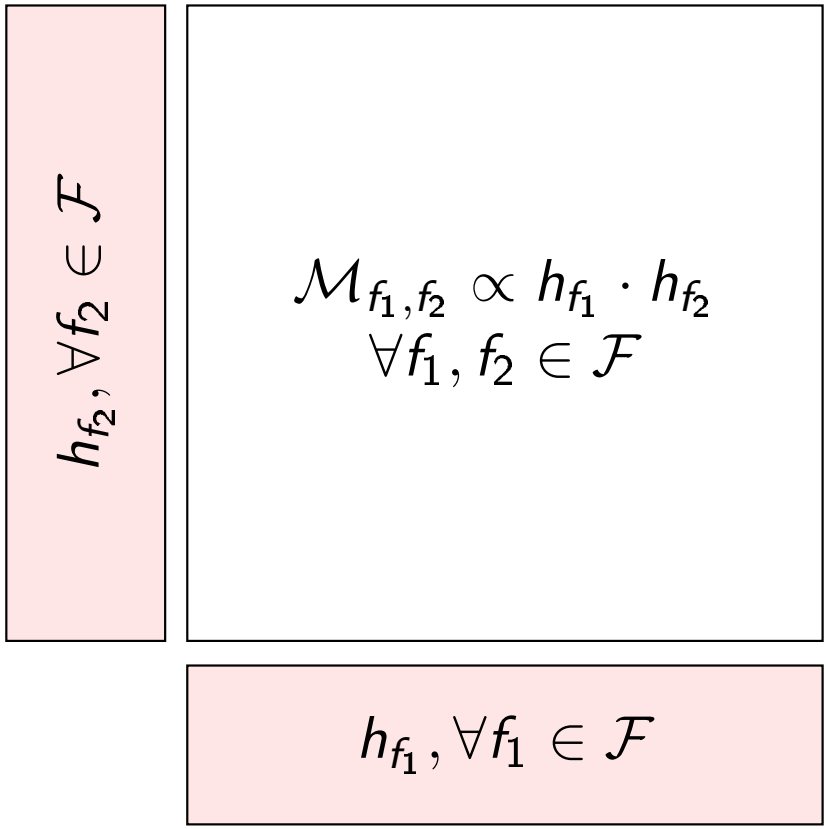

Figure 12: We can create an empirically sampled feature co-occurrence matrix. Finally, we can decompose this matrix and create a weighting scheme for regularization.

Graph Embedding with Self Clustering

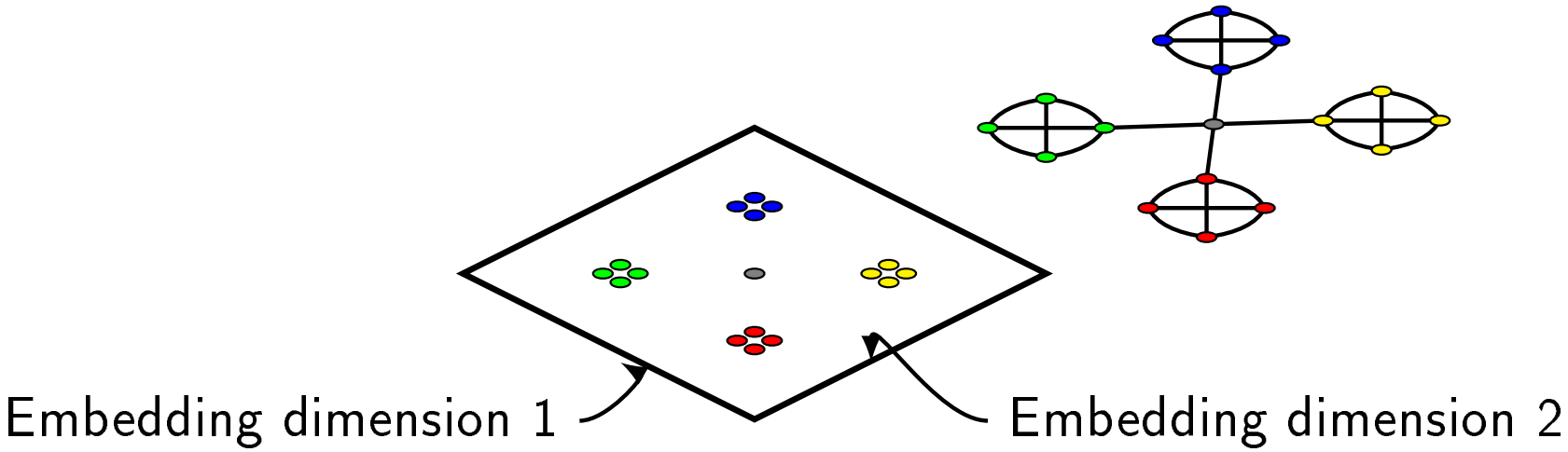

Figure 13: Community aware neighborhood preserving node embedding procedures maintain proximity on the graph between pairs of nodes in the embedding space and force dense components to be embedded in low volume spaces.

The basic node embedding model:

$$\min_{f} \quad \sum_{v \in V} -\log P(N_S(v)|f(v))$$

Adding a clustering term:

$$\min_{f,\mu} \quad \sum_{v \in V} -\log P(N_S(v)|f(v)) + \underbrace{\gamma \cdot \sum_{v \in V} \min_{c \in C} \|f(v) - \mu_c\|_2}_{\text{Clustering cost}}$$

Using smoothness regularization:

$$\Lambda = \lambda \cdot \sum_{(v,u) \in E_S} w_{(v,u)} \cdot \|f(v) - f(u)\|_2$$

How Does it Work in Practice?

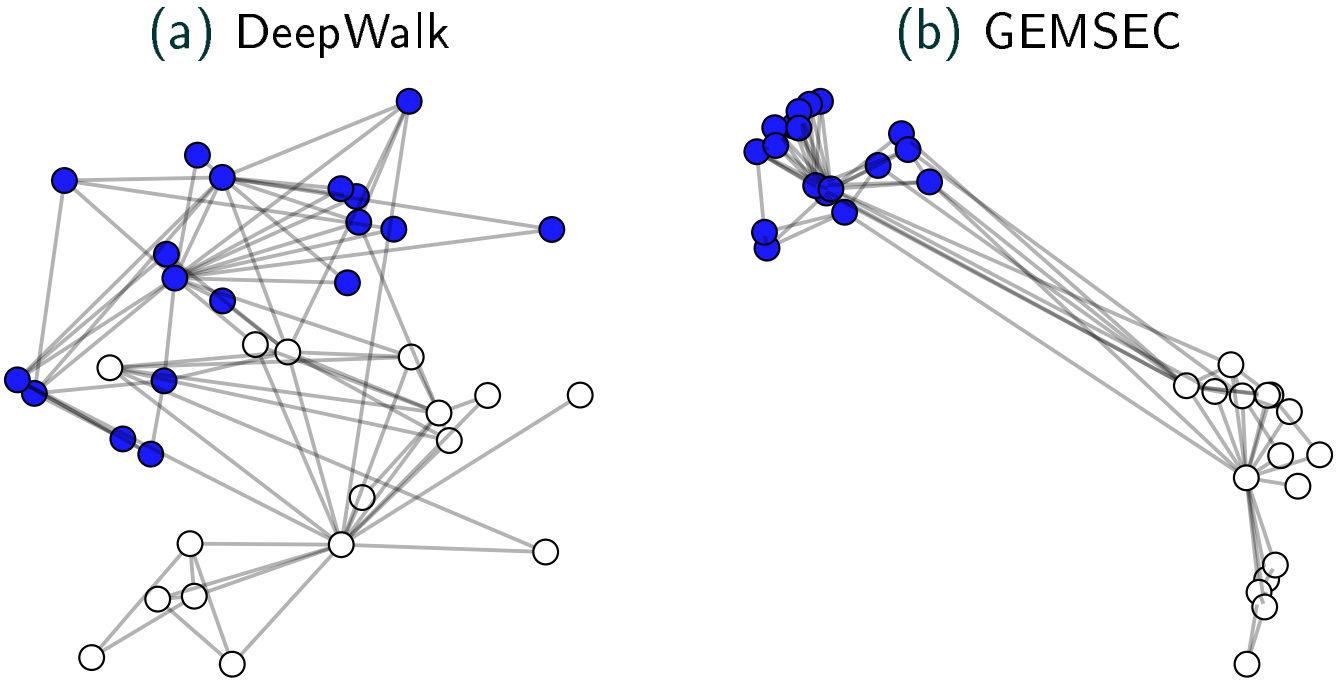

Figure 14: Zachary's Karate club visualization. White nodes: instructor's group; blue nodes: president's group. GEMSEC gives tighter embedding of the intrinsic communities.

	Croatia	Hungary	Romania
Overlap factorization	0.319 (±0.017)	0.361 (±0.007)	0.275 (±0.025)
DeepWalk	0.321 (±0.006)	0.361 (±0.004)	0.307 (±0.008)
LINE	0.331 (±0.013)	0.374 (±0.007)	0.332 (±0.007)
Node2Vec	0.348 (±0.012)	0.393 (±0.008)	0.346 (±0.008)
Walklets	0.363 (±0.013)	0.397 (±0.007)	0.361 (±0.011)
ComE	0.326 (±0.012)	0.363 (±0.010)	0.323 (±0.008)
M-NMF	0.336 (±0.005)	0.369 (±0.015)	0.330 (±0.016)
Smooth DeepWalk	0.329 (±0.006)	0.375 (±0.006)	0.321 (±0.008)
GEMSEC	0.328 (±0.006)	0.377 (±0.004)	0.332 (±0.008)
Smooth GEMSEC	0.333 (±0.006)	0.379 (±0.006)	0.334 (±0.008)
GEMSEC$_2$	**0.381** (±0.007)	0.407 (±0.005)	**0.378** (±0.009)
Smooth GEMSEC$_2$	0.373 (±0.005)	**0.409** (±0.004)	0.376 (±0.008)

Table 1: Multi-label node classification performance of the embedding extracted features on the Deezer genre likes datasets.

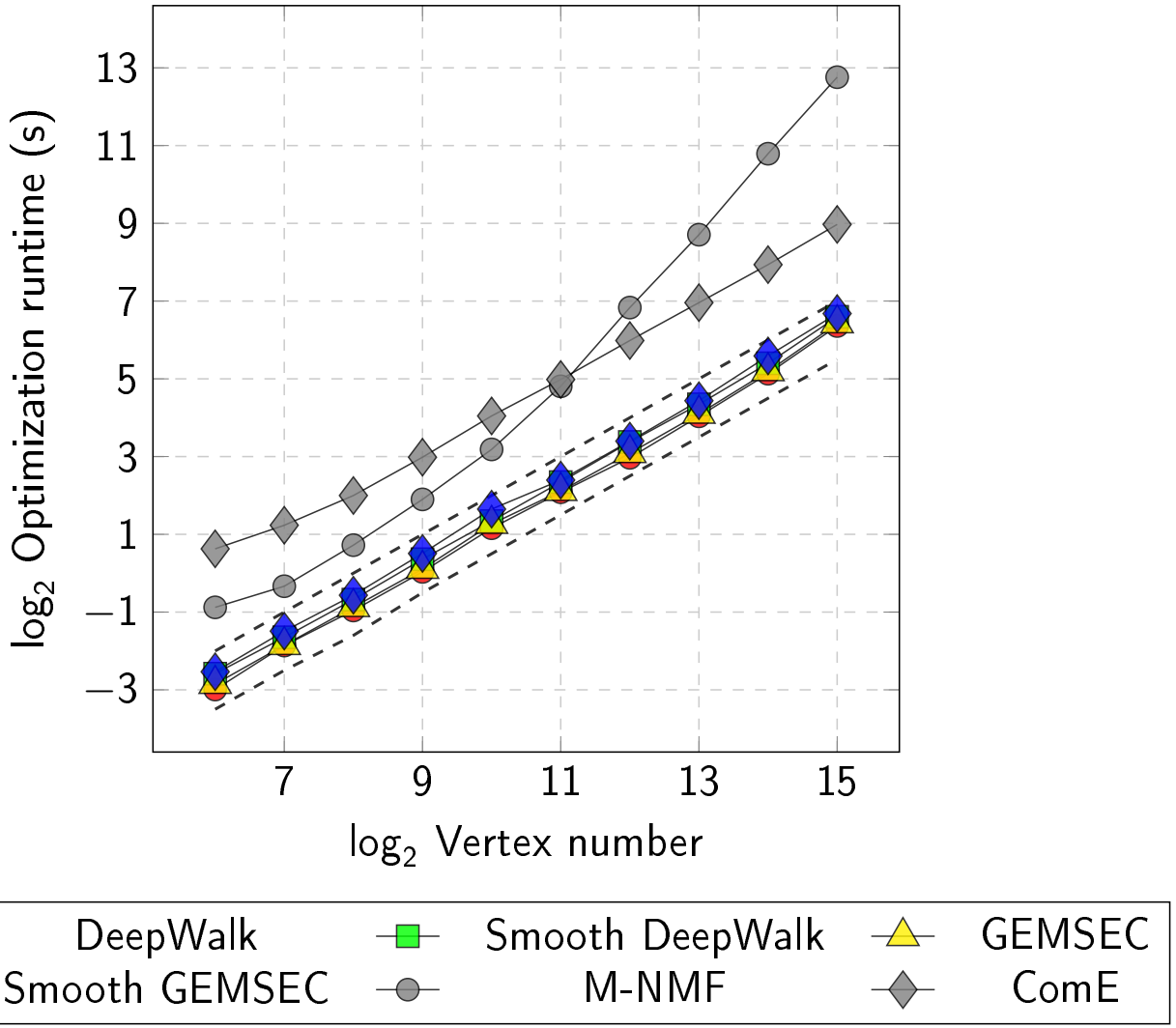

Figure 15: Sensitivity of optimization runtime to graph size measured by seconds. The dashed lines are linear references.

Thank You for the kind attention!

Bibliography I

Chih-Ming Chen, Chun-Yao Yang, Chih-Chun Hsia, Yian Chen, Ming-Feng Tsai. Music playlist recommendation via preference embedding. In *RecSys Posters*, 2016.

Claire Donnat, Marinka Zitnik, David Hallac, Jure Leskovec. Learning structural node embeddings via diffusion wavelets. 2018.

Thomas Gärtner, Peter Flach, Stefan Wrobel. On graph kernels: Hardness results and efficient alternatives. In *Learning Theory and Kernel Machines*, pages 129–143. Springer, 2003.

Nils M Kriege, Pierre-Louis Giscard, Richard Wilson. On valid optimal assignment kernels and applications to graph classification. In *Advances in Neural Information Processing Systems*, pages 1623–1631, 2016.

Annamalai Narayanan, Mahinthan Chandramohan, Lihui Chen, Yang Liu, Santhoshkumar Saminathan. subgraph2vec: Learning distributed representations of rooted sub-graphs from large graphs. *arXiv preprint arXiv:1606.08928*, 2016.

Bibliography II

Bryan Perozzi, Steven Skiena. Exact age prediction in social networks. In *Proceedings of the 24th International Conference on World Wide Web*, pages 91–92. ACM, 2015.

Bryan Perozzi, Vivek Kulkarni, Haochen Chen, Steven Skiena. Don't walk, skip!: Online learning of multi-scale network embeddings. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 258–265. ACM, 2017.