Local Intrinsic Dimensionality and Graphs:
Towards LID-aware Graph Embedding Algorithms

Miloš Savić, Vladimir Kurbalija, Miloš Radovanović

University Novi Sad, Faculty of Sciences
Department of Mathematics and Informatics
Novi Sad, Serbia
`{svc, kurba, radacha}@dmi.uns.ac.rs`

https://graphsinspace.net/

Introduction

- LID: characterize complexity of data space around a data point
  - Theoretical LID framework by Houle [2]
  - LID ≡ indiscriminability of the distance function
- Various LID applications: clustering, outlier detection, deep learning (robustness against adversarial attacks), etc.
- Graph embeddings

Our motivation & contributions

1. Discussion of potential LID applications to graphs
2. NC-LID: LID-related measure for nodes in a graph that is based on their natural (local) communities
3. Two LID-elastic extensions of Node2Vec [1] based on NC-LID

LID and Graphs

- Existing LID models and estimators: tabular dataset (data points in Euclidean space), smooth distance functions
- LID estimators based on distances from a reference data point x to its k closest neighbors
  - MLE-based LID estimator
  - Estimating LID within tight localities
- **Two ways for applying LID estimators to graphs**
  - By applying LID estimators directly on graph-based distances
  - By estimating LID of nodes on graph embeddings → LID-based evaluation of graph embedding algorithms

NC-LID: LID-related Measure for Graph Nodes based on Natural Communities

- Ball around a data point → subgraph S around a node n
- GB-LID: local intrinsic discriminability of a graph-based distance function *dist* considering S as the observed locality of n

$$\text{GB-LID}(n) = -\ln\left(\frac{|S|}{T(n, S)}\right),$$

- $|S|$ – the number of nodes in S
- r – the maximal distance between n and any node from S
- $T(n, S)$ – the number of nodes whose distance from n is smaller than or equal to r

- NC-LID is an instance of GB-LID
  - S – natural (local) communities [3]
  - *dist* – shortest-path distance
- NC-LID(n) = 0 → n has a "convex" natural community
- Higher NC-LID implies more "concave" natural communities

LID-elastic Node2Vec Variants

- **Main idea**: Hyper-parameters of graph embedding algorithms personalized for nodes / pairs of nodes and adjusted according to NC-LID
- **Main premise**: high NC-LID nodes will have higher link reconstruction errors in embeddings due to more complex natural communities

- **Node2Vec hyperparameters**
  - NRW: the number of random walks starting from each node
  - LRW: the length of each random walk
  - p and q: parameters controling random walk biases
- **Our LID-elastic Node2Vec extensions**
  - `lid-n2v-rw`: personalizes NRW and LRW per node according to NC-LID
  - `lid-n2v-rwpq`: extends `lid-n2v-rw` by personalizing p and q for each pair of connected nodes according to NC-LID values

Experiments and Results

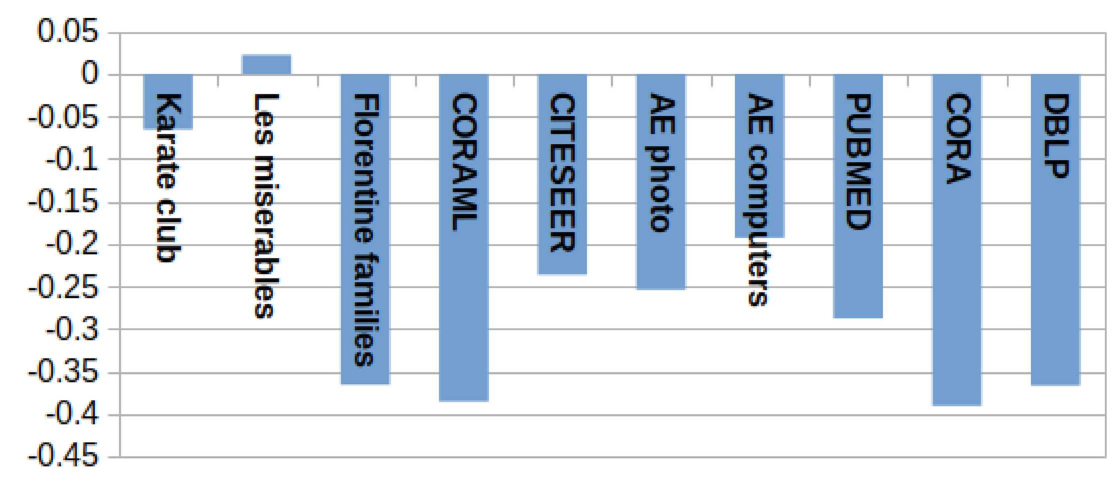

Figure 1: The Spearman correlation between NC-LID of nodes and their F_1 scores.

	`n2v`		`lid-n2v-rw`		`lid-n2v-rwpq`			
Graph	F_1	Dim.	F_1	Dim.	F_1	Dim.	Best	I[%]
Karate club	0.78	100	0.83	50	0.85	100	`lid-n2v-rwpq`	9.4
Les miserables	0.81	100	0.80	100	0.83	200	`lid-n2v-rwpq`	2.7
Florentine families	0.96	100	0.96	100	0.96	100	all	0.0
CORAML	0.65	25	0.66	50	0.63	25	`lid-n2v-rw`	1.3
CITESEER	0.24	10	0.25	10	0.28	10	`lid-n2v-rwpq`	18.7
AE photo	0.50	50	0.52	50	0.49	50	`lid-n2v-rw`	4.9
AE computers	0.45	50	0.47	100	0.42	50	`lid-n2v-rw`	4.7
PUBMED	0.39	50	0.43	50	0.42	50	`lid-n2v-rw`	9.4
CORA	0.57	25	0.60	50	0.59	50	`lid-n2v-rw`	3.9
DBLP	0.40	25	0.44	25	0.53	50	`lid-n2v-rwpq`	31.7

Table 1: Comparison of Node2Vec and LID-elastic Node2Vec embeddings.

Conclusions and Future Work

- NC-LID can point to weak parts of Node2Vec embeddings
- Node2Vec embeddings can be improved by LID-elastic extensions based on NC-LID (lower link reconstruction errors)
- LID-related metrics based on expanding subgraph localities
- Correlations between LID-related scores and centrality metrics
- Biased random walk strategies based on natural communities

References

[1] Aditya Grover and Jure Leskovec. Node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD'16, page 855–864, 2016.

[2] Michael E. Houle. Dimensionality, discriminability, density and distance distributions. In *2013 IEEE 13th International Conference on Data Mining Workshops*, pages 468–473, 2013.

[3] Andrea Lancichinetti, Santo Fortunato, and János Kertész. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11(3):033015, 2009.

Acknowledgements

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

Science Fund
of the Republic of Serbia