Intrinsic Dimensionality and Graph Representations

Jaishree Janu
Supervisors: Dr. Maximilian Stubbemann,
Prof. Dr. Dr. Lars Schmidt-Thieme

Table of contents

Introduction

Intrinsic Dimensionality (ID)	Graph representations
Innate or natural behavior Helps understand true complexity, visualizations, computationally less expensive	Low-dimensional vector representations of graphs

- We want to study the association of graph IDs with embedding IDs
- Study the impact of IDs on downstream tasks
- Motivated towards contribution in graphs analysis and ML

Research questions

How does a graph's ID correlate to its embeddings' ID?

How do graph embedding techniques affect the resulting embeddings’ IDs?

How do graph IDs impact the performance metrics of downstream tasks?

Is there any linear correlation between the graph metrics and a node's embeddings' ID?

How can we create appropriate synthetic datasets to study the above relationships?

Related work

ID across layers in DNNs

- ID << no. of units $\forall$ layer
- ID first increases and then decreases in final layers
- TwoNN: ID depends on nearest-neighbour statistics

ID of objective landscape

- Train networks on random linear subspace, *d-dim* subspace of $\mathbb{R}^D$
- ID: dimension at which solution appears

LID aware embedding algorithm

- LID aware node2vec algorithm outperforms vanilla node2vec
- NC-LID correlated to link reconstruction errors than other node centrality measures

Unbiased graph embedding

- Node sensitive attributes passed down to node embeddings, affect downstream tasks
- Structural properties unbiased
- Reconstruction loss defined on bias-free graph

ID estimators for graphs

NCLID

- Accounts for : no. of nodes in a neighbourhood (*S*) and no. of nodes located from *n* in relevant radius
- Maintains two sets B (border nodes) and C (community nodes)
- Quantifies discriminability of shortest-path distance considering NCs

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

$$f_C = \frac{k_{in}(C)}{(k_{in}(C) + k_{out}(C))^{\alpha}}$$

GEOL

- Incorporates geometric properties of graphs
- Concentration of measure phenomenon: most important features concentrate near their median/mean, non-discriminating
- Defines partial diameter for each feature to find the extent of discriminating subsets
- Approximations for computational feasibility

Graph embedding techniques

Node2vec

- Flexible biased random walk
- Incorporates both BFS and DFS
- *return p* and *in-out q* parameters

GraphSAGE

- Sample neighbourhood
- Aggregate feature information from neighbours
- Predict label
- Leverages nodes features

GCN

- Layer wise propagation rule for NN on graphs
- first-order approximation of localized spectral filters on graphs

GAT

- Mask self-attention layers
- hidden representations of each node, by attending over its neighbors

ID estimators in Euclidean Spaces

PCA

- Identifies *basis* (meaningful frame of reference)
- Identifies principal components: the most variance directions
- Orthogonal projection of data

DanCo

- Normalized nearest neighbour distances and angles computed on neighbouring points
- Compares the joint *PDF* related to angles and norms estimated on the dataset, with those estimated on synthetic datasets of known id

Mind_ML

- Correction method based on the comparison between pdfs.
- PDF related to the normalized nearest neighbour distances

Graph measures

Degree Centrality

DC(u): fraction of nodes it is connected to,

- normalized by the maximum possible degree in graph G

Closeness Centrality

CC(u): reciprocal of the avg. shortest path distance to u over all n-1 reachable nodes

$$C(u) = \frac{n-1}{\sum_{v=1}^{n-1} d(v,u)}$$

Betweenness Centrality

BC(u): sum of the fraction of all-pairs shortest paths that pass through u

$$c_B(v) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)}$$

Node Homophily Ratio

Measures the extent to which similar nodes are connected to each other in a graph

Downstream ML tasks and evaluation metrics

Node Classification

- Multi-class classification
- State-of-the-art models: GraphSAGE, GCN, GAT
- Evaluation metric: Accuracy

Link Prediction

- Encoder: Graph Convolutional Network
- Decoder: *inner product* of node embeddings
- Randomly add negative links
- Binary classification problem
- Evaluation metric: ROC-AUC score

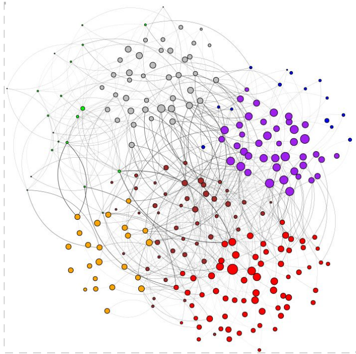
Anomaly Detection

- Unsupervised Model: DOMINANT
- Encoder: three convolutional layers
- Structure reconstruction and attribute reconstruction decoder
- Evaluation metrics: ROC-AUC score and avg. precision score

Data foundations: types of graphs

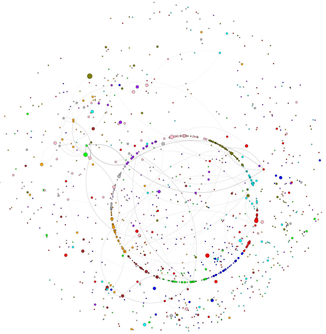
SBM

- Stochastic Block Model
- Nodes from same class belong to the same cluster/block
- Edge probabilities



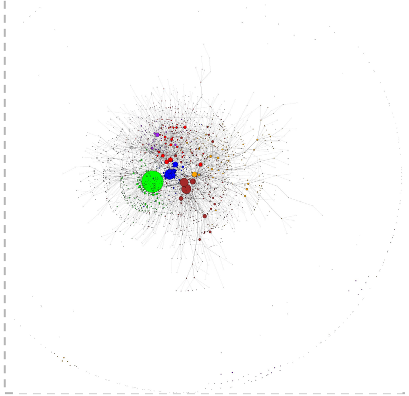
RPSBM

- Random Partition SBM
- Nodes from same class may not belong to the same cluster
- Node homophily, avg degree



Real graphs

- Citation graphs: CORA, CORAML, CiteSeer, PubMed, DBLP

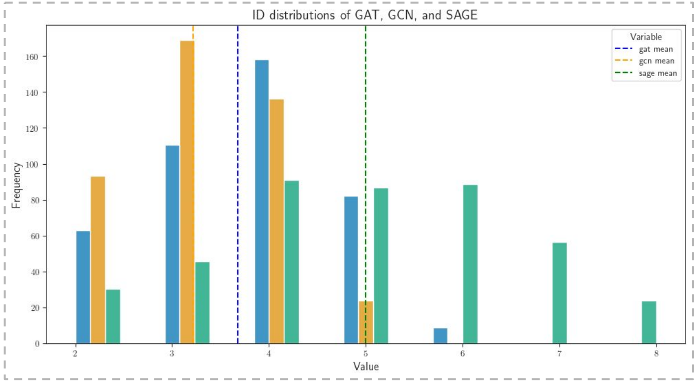


SBM graphs

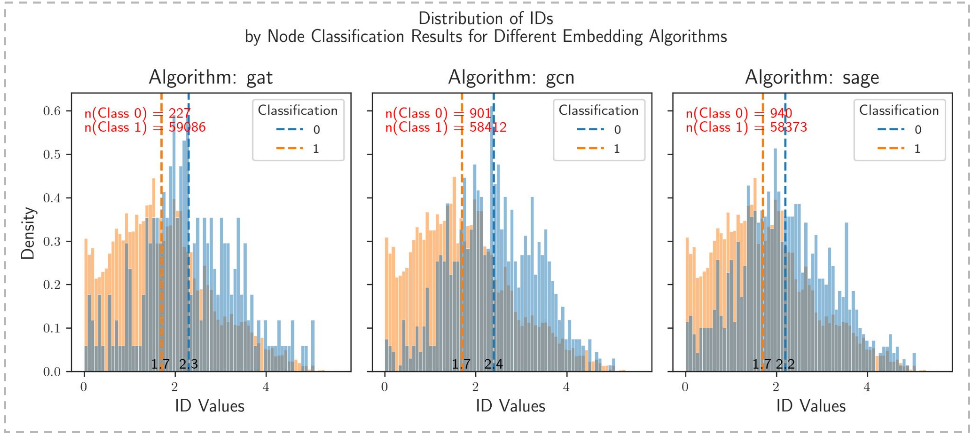
Pearson correlation

Analysis	Variable 1	Variable 2	Correlation	p-value
Graph IDs & Node Class.	avg_nclid_graph	sage_test_acc	-0.33	0.00
Graph IDs & Anomaly Det.	dim_graph_geol	avg_precision_score	-0.31	0.00
Graph IDs & Node Class.	avg_nclen	sage_test_acc	0.31	0.00
Graph IDs & Anomaly Det.	avg_nclid_graph	avg_precision_score	-0.28	0.00
Graph IDs & Node Class.	avg_nclid_graph	gcn_test_acc	-0.25	0.00
Graph IDs & Node Class.	avg_nclid_graph	gat_test_acc	-0.22	0.00
Graph IDs & Node Class.	avg_nclen	gat_test_acc	0.21	0.00
Graph IDs & Node Class.	dim_graph_geol	gcn_test_acc	-0.21	0.00
Graph IDs & Anomaly Det.	dim_graph_geol	roc_auc	-0.2	0.01
Graph IDs & Anomaly Det.	avg_nclid_graph	roc_auc	-0.19	0.01
Graph IDs & Node Class.	dim_graph_geol	sage_test_acc	-0.18	0.01
Graph IDs & Node Class.	avg_nclen	gcn_test_acc	0.17	0.02

Distribution of embedding IDs



Distribution of LID by whether a node is accurately classified



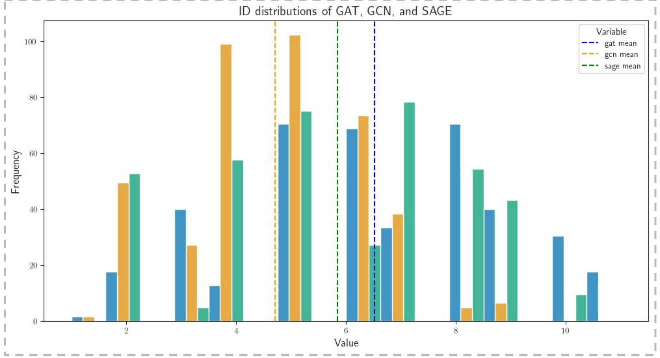
Both Welch's t-test and Mann-Whitney U test have p-values < 0.05. Cohen's D: -0.54, -0.63, -0.43

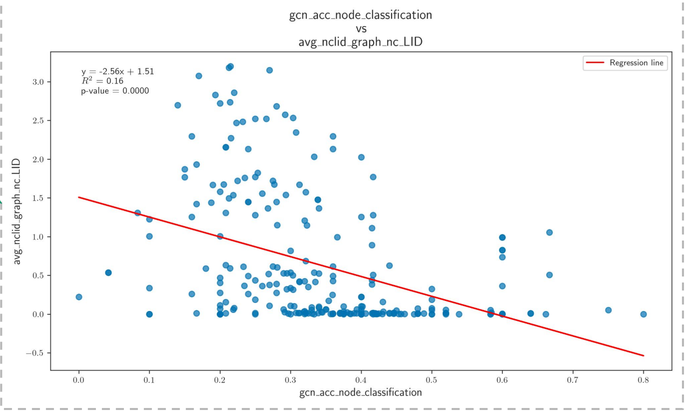
RPSBM graphs

Pearson correlation

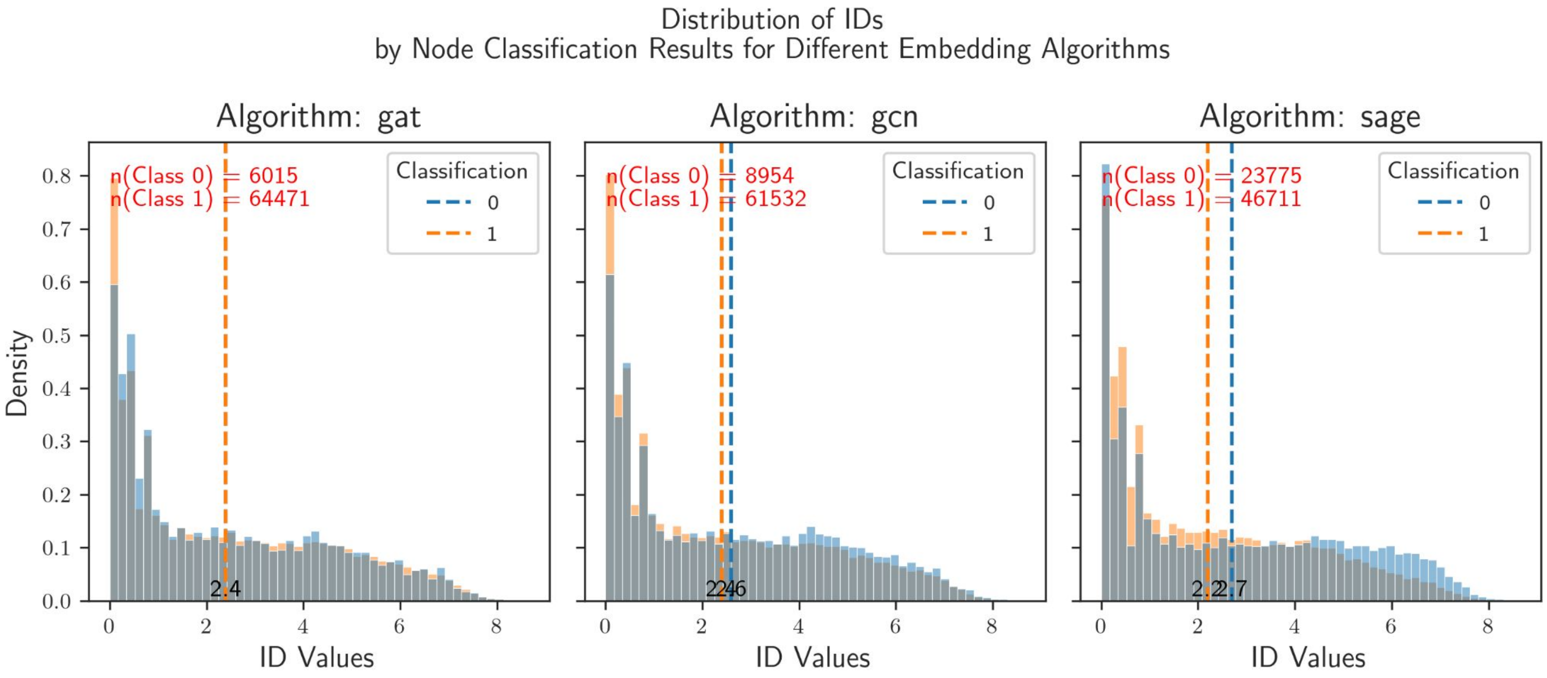
Analysis	Variable 1	Variable 2	Correlation	p-value
Embed. IDs & Graph Met.	gat	node_homophily_ratio	-0.14	0.02
Graph IDs & Embed. IDs	dim_graph_geol	gcn	-0.15	0.02
Graph IDs & Node Class.	avg_nclid_graph	gcn_acc	-0.40	0.00
Graph IDs & Node Class.	dim_graph_geol	sage_acc	-0.35	0.00
Graph IDs & Node Class.	graph_nclen	gcn_acc	-0.31	0.00
Graph IDs & Node Class.	graph_nclen	sage_acc	0.31	0.00
Graph IDs & Node Class.	avg_nclid_graph	gat_acc	-0.25	0.00
Graph IDs & Node Class.	avg_nclid_graph	sage_acc	0.19	0.00
Graph IDs & Node Class.	graph_nclen	gat_acc	-0.18	0.00
Graph IDs & Node Class.	dim_graph_geol	gcn_acc	-0.14	0.02

Distribution of embedding IDs





Real graphs



Both Welch's t-test and Mann-Whitney U test have p-values < 0.05 for GCN and SAGE. Cohen's D: -0.006, -0.11, -0.24

Inference: Statistically significant Pearson and Spearman correlations

	Graph ID	Graph measures	Embedding ID
Embedding ID	YES (GEOL ID with GCN's, embeddings ID), (RPSBM set) (-)	YES (node homophily with GAT's, SAGE's embeddings ID) (RPSBM set) (-)	NA
Downstream task	YES (GEOL and NCLID with node classification and anomaly detection) (-)	YES (degree centrality with GCN's node classification and link prediction) (-), (+)	NO
Graph measures	YES (NCLID with closeness centrality and avg degree) (+)	NA	NA

Discussion: research questions

graph's ID with embeddings' ID: one weak negative correlation between GEOL ID with GCN's node embeddings' ID

Different distributions of embeddings IDs from embedding algorithms:
– GCN has lowest avg Embedding IDs, SAGE has notably spread out distribution

Graph's ID with downstream tasks: multiple significant negative correlations

graph metrics with a node's embeddings' ID: a few weak negative correlations with node homophily

torch's SBM and *RPSBM* classes enable generation of a large number of graphs with help of graph parameters, we have ensured variety in data generation

Conclusions

Performance of node classification, anomaly detection and link prediction suffer due to higher graph IDs

Graph IDs have stronger correlation with downstream tasks compared to graph measures

Embedding techniques influence the ID of resulting representations

GCN shows the strongest relationship between ID and classification accuracy

No significant correlations found between embedding IDs and downstream tasks

Contrasting results from related work discussed earlier

- Related work experiments only on Conv nets
- Related work top-5 score used to evaluate classification

Limitations & future work

Limitations

- Pearson and spearman correlations only examine linear or monotonic relationships
- More experiments required to make stronger arguments where few weak correlations observed

Future work

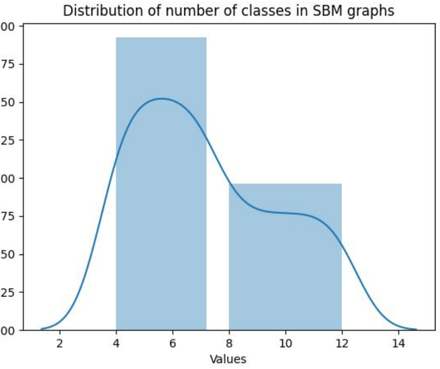
- Examine LID-elastic node2vec embeddings in downstream tasks
- ID measures could be incorporated into the loss function of GNNs to obtain graph embeddings by LID-aware deep learning techniques
- Bringing in edge embeddings into the investigation
- Exploring other downstream tasks, such as graph classification, edge outlier detection, to evaluate broader applicability.
- Adapting advanced non-linear and complex network ID estimators for graph datasets

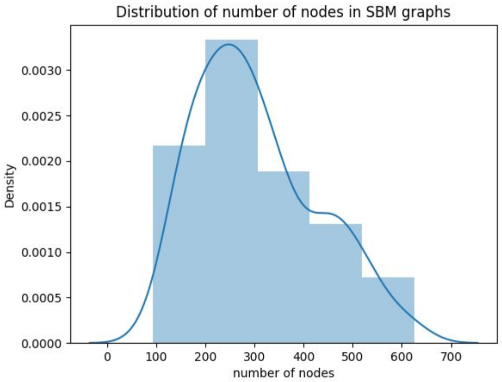
Thank you

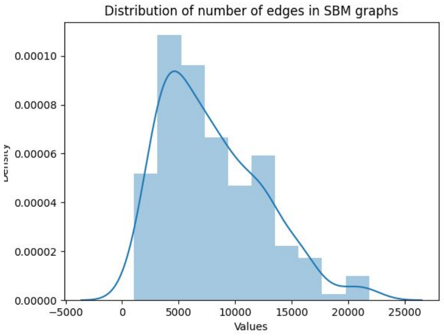
References

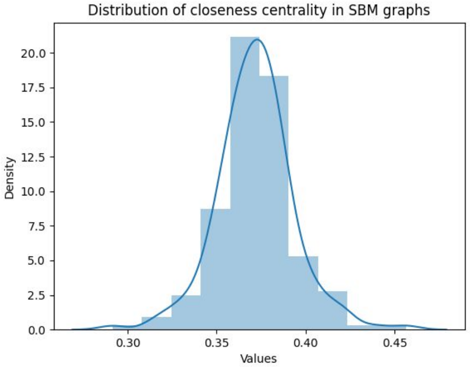
- X. Ma, Y. Wang, M. E. Houle, S. Zhou, S. Erfani, S. Xia, S. Wijewickrema, and J. Bailey, “Dimensionality-driven learning with noisy labels,”
- L. Amsaleg, J. Bailey, D. Barbe, S. Erfani, M. E. Houle, V. Nguyen, and M. Radovanović, “The vulnerability of learning to adversarial perturbation increases with intrinsic dimensionality,”
- “Intrinsic dimension of data representations in deep neural networks”

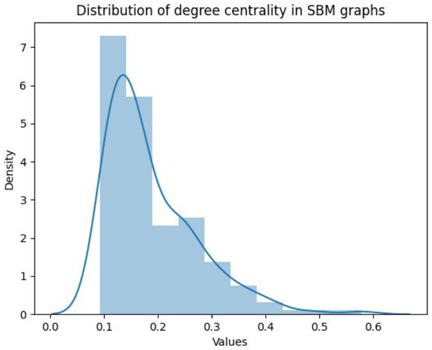
Appendix

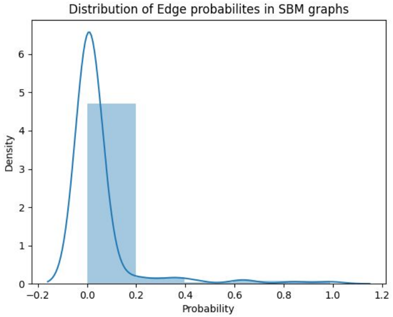

Distribution of number of classes in SBM graphs
Values

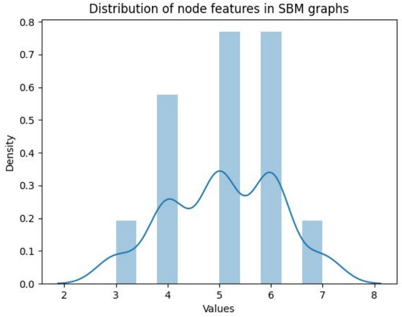

Distribution of number of nodes in SBM graphs
Density
number of nodes


Distribution of number of edges in SBM graphs
Values

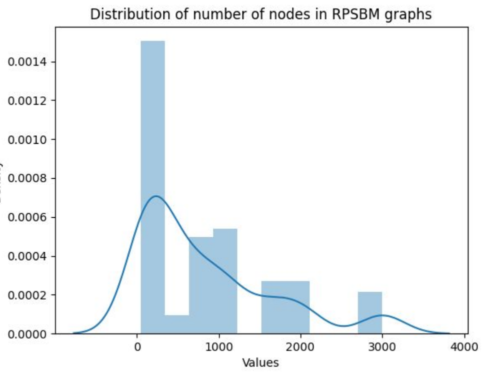

Distribution of closeness centrality in SBM graphs
Density
Values

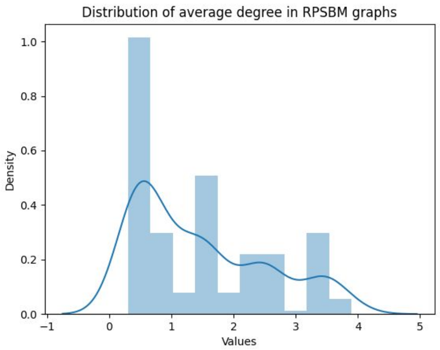

Distribution of degree centrality in SBM graphs
Density
Values

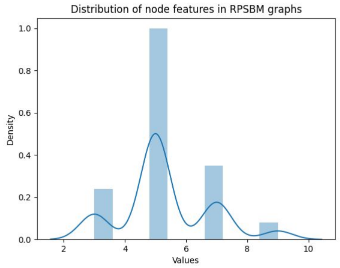

Distribution of Edge probabilites in SBM graphs
Density
Probability

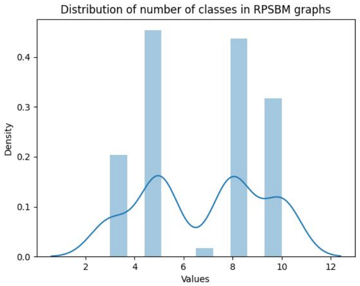

Distribution of node features in SBM graphs
Density
Values

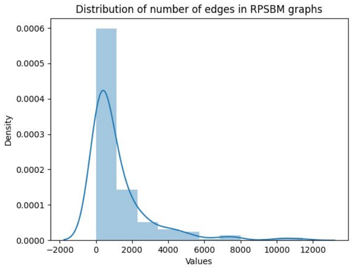
SBM Graphs

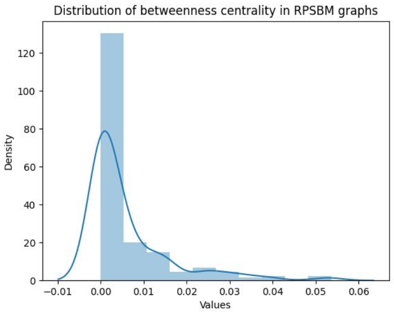


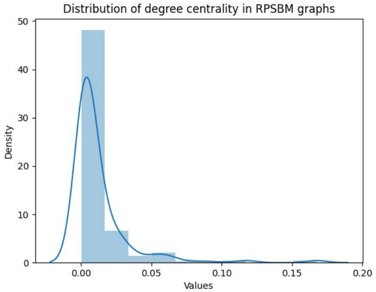


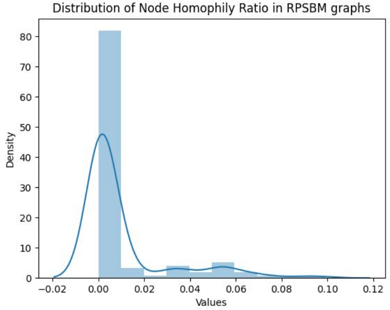












RPSBM Graphs

LID-aware node2vec embeddings

- Node2vec parameters: NRW, LRW, return param *p* and in-out param *q*

$$\mathrm{NRW}(v) = \lfloor (1 + \text{NC-LID}(v)) \cdot B \rfloor \qquad \mathrm{LRW}(v) = \lfloor W / (1 + \text{NC-LID}(v)) \rfloor$$

- Increase the frequency of high NC-LID nodes in sampled random walks in order to better preserve their close neighborhood in formed embeddings
- In the second algorithm, they adjust *p* and *q* using NCLID values
- NCLID better indicator of nodes with low F_1 scores in node2vec embeddings than node centrality metrics
- *lid-n2v-rw* and *lid-n2v-rwpq* significantly outperform node2vec by between 19% - 47% on different datasets

SBM node level results

Metric	gat	gcn	sage
Length of Accurate Preds	59086	58412	58373
Length of False Preds	227	901	940
Variance of Accurate Preds	0.6186	0.6209	0.6201
Variance of False Preds	0.4529	0.3983	0.4721
Welch's T-test T-statistic	-8.3314	-21.1825	-13.6252
Welch's T-test P-value	7.5097e-15	1.2000e-81	8.1677e-39
Mann-Whitney U test U-statistic	4556896.5	16204639.5	20273861.5
Mann-Whitney U test P-value	6.9604e-17	1.9215e-87	5.0038e-43
Cohen's d	-0.5432	-0.6397	-0.4334

Metric	gat	gcn	sage
Length of Accurate Preds	64471	61532	46711
Length of False Preds	6015	8954	23775
Variance of Accurate Preds	0.8871	0.8950	0.9029
Variance of False Preds	0.8736	0.8243	0.8383
Welch's T-test T-statistic	-0.4677	-10.4229	-29.7466
Welch's T-test P-value	0.6400	2.5211e-25	1.8242e-192
Mann-Whitney U test U-statistic	191410817.0	255702777.0	492493906.5
Mann-Whitney U test P-value	0.0993	3.7321e-28	1.2021e-133
Cohen's d	-0.0063	-0.1196	-0.2477

Real graphs node level results

SBM correlations

Pearson

Analysis	Variable 1	Variable 2	Correlation	p-value
Graph IDs vs Node Classification Metrics	avg_nclid_graph_nc_LID	sage_test_acc_node_classification	-0.33	0
Graph IDs vs Anomaly Detection Metrics	dim_graph_geol	anomaly_avg_precision_score_anomaly_prediction	-0.31	0
Graph IDs vs Node Classification Metrics	avg_nclen_graph_nclen	sage_test_acc_node_classification	0.31	0
Graph IDs vs Anomaly Detection Metrics	avg_nclid_graph_nc_LID	anomaly_avg_precision_score_anomaly_prediction	-0.28	0
Graph IDs vs Node Classification Metrics	avg_nclid_graph_nc_LID	gcn_test_acc_node_classification	-0.25	0
Graph IDs vs Node Classification Metrics	avg_nclid_graph_nc_LID	gat_test_acc_node_classification	-0.22	0
Graph IDs vs Node Classification Metrics	avg_nclen_graph_nclen	gat_test_acc_node_classification	0.21	0
Graph IDs vs Node Classification Metrics	dim_graph_geol	gcn_test_acc_node_classification	-0.21	0
Graph IDs vs Anomaly Detection Metrics	dim_graph_geol	anomaly_roc_auc_anomaly_prediction	-0.2	0.01
Graph IDs vs Anomaly Detection Metrics	avg_nclid_graph_nc_LID	anomaly_roc_auc_anomaly_prediction	-0.19	0.01
Graph IDs vs Node Classification Metrics	dim_graph_geol	sage_test_acc_node_classification	-0.18	0.01
Graph IDs vs Node Classification Metrics	avg_nclen_graph_nclen	gcn_test_acc_node_classification	0.17	0.02

statistically significant correlations: 12

Spearman

Analysis	Variable 1	Variable 2	Correlation	p-value
Graph IDs vs Node Classification Metrics	avg_nclid_graph_nc_LID	sage_test_acc_node_classification	-0.41	0.00
Graph IDs vs Node Classification Metrics	avg_nclid_graph_nc_LID	gcn_test_acc_node_classification	-0.36	0.00
Graph IDs vs Node Classification Metrics	avg_nclid_graph_nc_LID	gat_test_acc_node_classification	-0.32	0.00
Graph IDs vs Node Classification Metrics	avg_nclen_graph_nclen	sage_test_acc_node_classification	0.29	0.00
Graph IDs vs Node Classification Metrics	dim_graph_geol	sage_test_acc_node_classification	-0.29	0.00
Graph IDs vs Anomaly Detection Metrics	dim_graph_geol	anomaly_avg_precision_score_anomaly_prediction	-0.28	0.00
Graph IDs vs Node Classification Metrics	dim_graph_geol	gcn_test_acc_node_classification	-0.26	0.00
Graph IDs vs Anomaly Detection Metrics	avg_nclid_graph_nc_LID	anomaly_avg_precision_score_anomaly_prediction	-0.24	0.00
Graph IDs vs Anomaly Detection Metrics	dim_graph_geol	anomaly_roc_auc_anomaly_prediction	-0.21	0.00
Graph IDs vs Anomaly Detection Metrics	avg_nclid_graph_nc_LID	anomaly_roc_auc_anomaly_prediction	-0.19	0.01
Graph IDs vs Node Classification Metrics	avg_nclen_graph_nclen	gat_test_acc_node_classification	0.19	0.01
Graph IDs vs Node Classification Metrics	avg_nclen_graph_nclen	gcn_test_acc_node_classification	0.19	0.01
Graph Metrics vs Node Classification Metrics	degree_cent_graph_metrics	gcn_test_acc_node_classification	-0.16	0.03
Graph Metrics vs Node Classification Metrics	num_classes_graph_metrics	sage_test_acc_node_classification	0.15	0.04

statistically significant correlations: 14

RPSBM correlations

Pearson

Analysis	Variable 1	Variable 2	Correlation	p-value
Graph IDs vs Node Classification Metrics	avg_nclid_graph_nc_LID	gcn_acc_node_classification	-0.4005393645	0.00
Graph IDs vs Node Classification Metrics	dim_graph_geol	sage_acc_node_classification	-0.3453903566	0.00
Graph IDs vs Node Classification Metrics	avg_nclen_graph_nclen	gcn_acc_node_classification	-0.3110456329	0.00
Graph IDs vs Node Classification Metrics	avg_nclen_graph_nclen	sage_acc_node_classification	0.3094486752	0.00
Graph IDs vs Node Classification Metrics	avg_nclid_graph_nc_LID	gat_acc_node_classification	-0.2542027844	0.00
Graph IDs vs Node Classification Metrics	avg_nclid_graph_nc_LID	sage_acc_node_classification	0.1902028843	0.00
Graph IDs vs Node Classification Metrics	avg_nclen_graph_nclen	gat_acc_node_classification	-0.181412288	0.00
Graph IDs vs Embedding IDs (mind_ml)	dim_graph_geol	mind_ml_gcn_embeddings	-0.1504609706	0.02
Graph Metrics vs Link Prediction Metrics	node_features_graph_metrics	link_pred_test_auc_link_prediction	-0.1473289298	0.02
Graph IDs vs Node Classification Metrics	dim_graph_geol	gcn_acc_node_classification	-0.1434102198	0.02
Embedding IDs vs Graph Metrics (mind_ml)	mind_ml_gat_embeddings	node_homophily_ratio_graph_metrics	-0.1415364599	0.02
Graph Metrics vs Link Prediction Metrics	num_classes_graph_metrics	link_pred_test_auc_link_prediction	-0.1299922955	0.05

statistically significant correlations: 12

Spearman

Analysis	Variable 1	Variable 2	Correlation	p-value
Graph IDs vs Node Classification Metrics	avg_nclid_graph_nc_LID	gcn_acc_node_classification	-0.49	0.00
Graph IDs vs Node Classification Metrics	avg_nclen_graph_nclen	gcn_acc_node_classification	-0.45	0.00
Graph IDs vs Node Classification Metrics	avg_nclid_graph_nc_LID	gat_acc_node_classification	-0.41	0.00
Graph IDs vs Node Classification Metrics	dim_graph_geol	sage_acc_node_classification	-0.38	0.00
Graph IDs vs Node Classification Metrics	avg_nclen_graph_nclen	gat_acc_node_classification	-0.38	0.00
Graph IDs vs Node Classification Metrics	avg_nclen_graph_nclen	sage_acc_node_classification	0.29	0.00
Graph IDs vs Node Classification Metrics	avg_nclid_graph_nc_LID	sage_acc_node_classification	0.23	0.00
Embedding IDs vs Graph Metrics (mind_ml)	mind_ml_sage_embeddings	node_homophily_ratio_graph_metrics	-0.18	0.00
Graph Metrics vs Link Prediction Metrics	num_classes_graph_metrics	link_pred_test_auc_link_prediction	-0.15	0.02
Graph IDs vs Graph Metrics	avg_nclid_graph_nc_LID	avg_degree_graph_metrics	0.14	0.03
Graph Metrics vs Link Prediction Metrics	degree_cent_graph_metrics	link_pred_test_auc_link_prediction	0.13	0.05
Graph IDs vs Graph Metrics	avg_nclid_graph_nc_LID	close_cent_graph_metrics	0.13	0.04

statistically significant correlations: 14

Real graphs' correlations

Pearson

Analysis	Variable 1	Variable 2	Correlation	p-value
Graph IDs vs Embedding IDs (mind_ml)	dim_graph_geol	mind_ml_sage_embeddings	1	0.00
Graph Metrics vs Node Classification Metrics	node_features_graph_metrics	gat_test_acc_node_classification	-0.85	0.01
Graph Metrics vs Node Classification Metrics	node_features_graph_metrics	gcn_test_acc_node_classification	-0.83	0.02
Graph Metrics vs Link Prediction Metrics	num_nodes_graph_metrics	link_pred_test_auc_link_prediction	0.82	0.03

statistically significant correlations: 4

Spearman

Analysis	Variable 1	Variable 2	Correlation	p-value
Graph IDs vs Graph Metrics	dim_graph_geol	num_classes_graph_metrics	-0.87	0.01
Graph Metrics vs Link Prediction Metrics	num_edges_graph_metrics	link_pred_test_auc_link_prediction	0.86	0.01

statistically significant correlations: 2

Deep Anomaly Detection on Attributed Networks

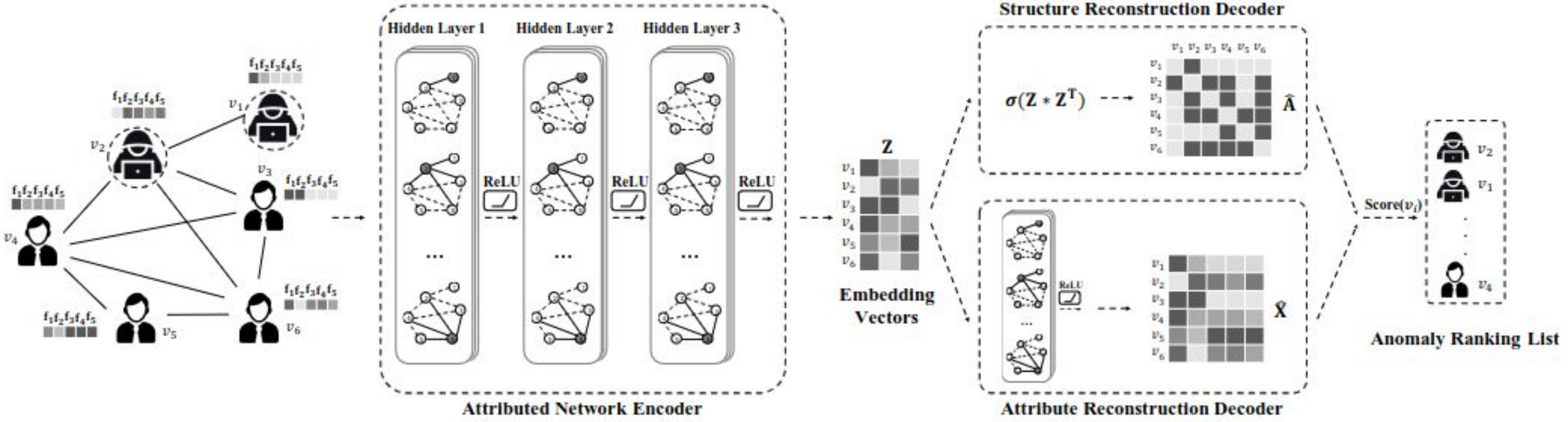


Fig: Encoder-Decoder SOTA architecture for link prediction