# Section 2: Overview of graph embedding methods

## Graph Embedding

1. Random-walk
2. Neural Network
3. Matrix-Factorization Based
4. Node proximity for signed networks

A definition that is generalized from the second-order node proximity for unsigned networks (e.g. LINE) [1st condition]

nSNE introduced a 2nd condition, that is, if two nodes in a signed network are similar, they not only should satisfy the second-order node proximity but also have similar sign context.

[Demonstrated by tradeoff parameter Beta]

Provides a unified objective function that can preserve both the node and edge pattern of the network

nSNE: Embeddings of nodes and mapping functions learned via back-propagation algorithm

## Protein Embedding

## Choosing edges

Reason for choosing edges {Song, 2018 #50}

Problem with converting node to edge embeddings

Table 1: Element-wise operators for computation of representation for pairs of nodes, where u and v represent the source and target nodes respectively.

|  |  |
| --- | --- |
| Operator | Definition |
| Sum |  |
| Average |  |
| Hadamard |  |
| Concatenation |  |

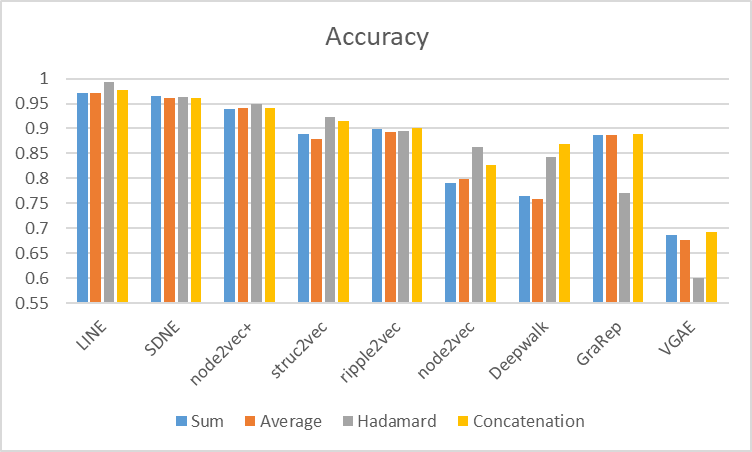


Figure 1: Comparison of element-wise operators used across NRL methods on the network reconstruction dataset

For five out of the nine methods, hadamard product is highly stable and consistently yeilds the best performance.

In contrast, for GraRep and VGAE, sum, average and concatenation all outperformed hadamard product.

For ripple2vec and Deepwalk, results obtained using hadamard and concatenation operators are comparable. Therefore, hadamard was selected as the number of dimensions is half that of concatenation. Concatenation was chosen for GraRep and VGAE.

**Network Reconstruction Problem**

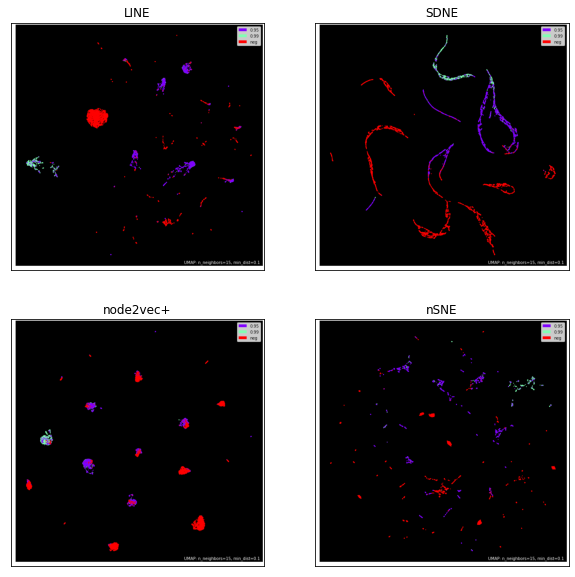


Figure 2(a): Plot of edges used in the network reconstruction (and experimentally verified) training set

*Full caption: Plot of edges used in the network reconstruction (and experimentally verified) training set, split into 3 classes. Green, purple and red data points denotes the interacting positive samples within 0.99 threshold, 0.95 threshold and the non-interacting negative samples respectively.*

These 4 methods were able to clearly distinguish highly interacting protein pairs (in green) from remaining protein pairs (with lower interaction probability) and the negative pairs.

However, only SDNE performance (w/ tune hyperparamters) remained consistent, even on other datasets (without having to retune hyperparameters).

Due to Homogeneity, completeness and v-measure score?

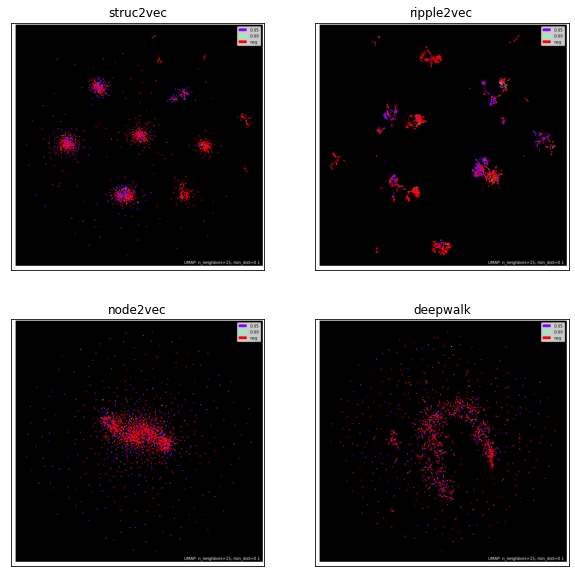


Figure 2(b): Plot of edges used in the network reconstruction (and experimentally verified) training set

[Random-walk methods]

For node2vec and deepwalk, there is only 1 distinct cluster, with no clear separation between the positive and negative protein pairs.

Although clusters were formed, the positive and negative pairs overlap and are not able to distinguish the highly interacting protein pairs.

Methods that did not perform well in the network reconstruction problem was “discarded”.

A good network embedding method should ensure that the learned embeddings can preserve the original network structure {Wang, 2016 #10} \cite{RN10}.

Therefore, in subsequent experiments, (i.e., unbalanced dataset and experimentally verified), we only tested on these 4 methods (i.e., LINE, SDNE, node2vec+ and nSNE).

##### Higher (2nd) – order proximity

###### LINE

(Importance of having 2nd order proximity) + results

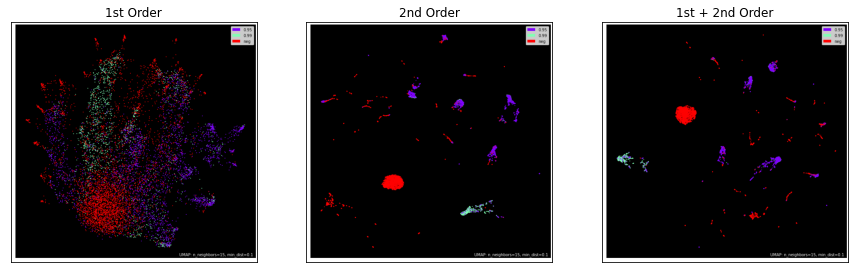


Figure 3: placeholder

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Accuracy | Sensitivity | Specificity | Precision | F1-Score | ROC-AUC | PR-AUC |
| 1st Order | 0.8345 | 0.8561 | 0.813 | 0.8247 | 0.8395 | 0.9079 | 0.895 |
| 2nd Order | 0.9883 | 0.9906 | 0.9861 | 0.9862 | 0.9884 | 0.9991 | 0.9991 |
| 1st + 2nd Order | 0.9922 | 0.9978 | 0.9865 | 0.9867 | 0.9922 | 0.9994 | 0.9992 |

\* Number of epochs was set to 10 for all.

##### Preserving edge attributes

###### Node2vec VS node2vec+

How having edges weights instead of binary (1/0) improve performance

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Accuracy | Sensitivity | Specificity | Precision | F1-Score | ROC-AUC | PR-AUC |
| node2vec | 0.8626 | 0.8964 | 0.8287 | 0.8394 | 0.8668 | 0.9124 | 0.8718 |
| node2vec + | 0.9493 | 0.9713 | 0.9274 | 0.9305 | 0.9504 | 0.9872 | 0.986 |

\* The same hyper-parameters were set for both (i.e, number of walks = , walk length = , p= , q = )

###### nSNE

signed VS unsigned network

Satisfying the 2nd condition (If two nodes in a signed network are similar, they should satisfy the condition of having similar sign context) {Song, 2018 #50}

When β = 1, only the first condition for node proximity of signed networks was used (i.e., they only satisfy the second-order node proximity)

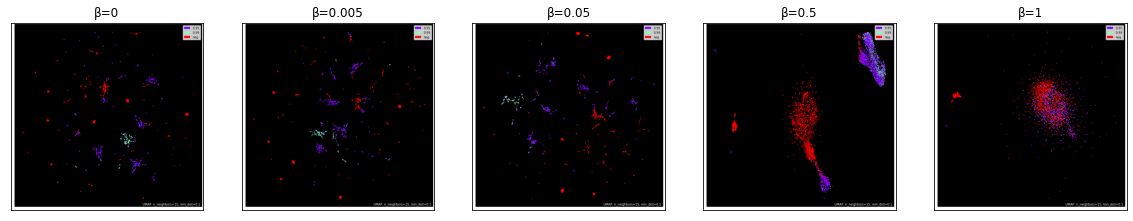


Figure 4: placeholder

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| β | Accuracy | Sensitivity | Specificity | Precision | F1-Score | ROC-AUC | PR-AUC |
| 0 | 0.9957 | 0.9964 | 0.9951 | 0.9951 | 0.9957 | 0.9999 | 0.9999 |
| 0.005 | 0.9964 | 0.9969 | 0.996 | 0.996 | 0.9964 | 0.9999 | 0.9999 |
| 0.05 | 0.9942 | 0.9955 | 0.9928 | 0.9929 | 0.9942 | 0.9999 | 0.9999 |
| 0.5 | 0.9791 | 0.9865 | 0.9717 | 0.9722 | 0.9793 | 0.9984 | 0.9984 |
| 1 | 0.8652 | 0.8664 | 0.8641 | 0.8646 | 0.8654 | 0.9348 | 0.9244 |

\* Vector dimension, K was set to 128 for all.

# Section 3: Experiments

## Chosen IAV strains

**[Insert Table]**

## Datasets constructed/used

* **Explored 3 problems:**

1. Network reconstruction problem
2. Testing on an experimentally verified dataset
3. Link prediction on an unbalanced dataset problem

* Formally introduce HVPPI

Threshold Score range (between 0 and 1)

0.99 >= 0.375

0.95 >= 0.212 and < 0.375

0.90 >= 0.143 and < 0.212

Negative < 0.143

* Overview of complete network graph

|  |  |  |
| --- | --- | --- |
| Data | Complete | Edgelist1 |
|
| # nodes | 15, 685  (41 IAV, 15, 644 Human) | |
| # nodes involved in interaction | 12, 438  (41 IAV,  12, 397 Human) | 12, 437  (41 IAV,  12, 396 Human) |
| # edges | 91, 217  (3738 0.99,  26, 044 0.95, 61, 435 0.90) | 48, 8822  (2872 0.99,  15, 782 0.95,  30,228 0.90) |
| # non-interacting pairs | 550, 187 | |
| HVPPI Score (MIN) | 0.000 | 0.001 |
| HVPPI Score (MAX) | 0.99 | 0.99 |
| Average Degree | 14.6675 | 7.8607 |

1 All experiments conducted were based on this edgelist.

2 As HVPPI is a prediction tool, only edges with consistent prediction scores on 2 runs were considered. **Therefore, only 53.5% of the interacting edges were used to construct the edgelist as input to the different graph embedding models.**

* Overview of constructed datasets

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Dataset |  | | | | | Link Prediction  (Unbalanced dataset)6 | |
| Train1 | Test 1  (Network Reconstruction)1 | Test 2  (Experimentally Verified) | Test 3  (Experimentally Verified) | Test 4 | Train | Test |
| # nodes involved in interaction | 1098  (19 IAV, 1079 Human) | 376  (16 IAV, 360 Human) | 361  (14 IAV, 347 Human) | 355  (10 IAV, 345 Human) | 346  (14 IAV, 332 Human) | 6522  (30 IAV,  6492 Human) | 2681  (28 IAV,  2653 Human) |
| # Positive samples  (i.e., interacting pairs) | 4014 | 446 | 4462 | 4494  (402 PR8 + 47 Aichi) | 4465 | 91, 217 - 48, 882 = 42, 335  42, 335 \* 0.9 = 38, 101 | 42, 335 \* 0.1 = 4234 |
| # Negative samples  (i.e., non-interacting pairs) | 4014 | 446 | 4463 | 4463 | 4463 | 38, 101 \* 4 = 152, 404 | 4234 \* 9 =  38, 106 |
| HVPPI Score (MIN) | 0.000 | 0.001 | 0.023  (excluding negative samples) | 0.009 (PR8)  0.047 (Aichi) | 0.286 | 0.004 | 0.004 |
| HVPPI Score (MAX) | 0.99 | 0.931 | 0.981 | 0.100 (PR8)  0.139 (Aichi) | 0.988 | 0.845 | 0.743 |
| Average Degree | 7.3115 | 2.3723 | 2.4709 |  |  | 11.6838 | 3.1585 |

**1 Edges are found in edgelist.**

2 Positive samples from HPIDB 3.0 \cite{}

3 Negative samples from negatome 2.0 \cite{} database. Human-Human protein interactions only. Involving 482 Human Protein Nodes

1. Positive samples from HPIDB 3.0 checked againt HVPPI (i.e., indicated by HVPPI to be non- interacting).
2. Compared against HVPPI probability score and obtained the “Top” 446 interactions

6 All remaining **interacting** protein pairs not in edgelist made up the positive samples in this dataset.

Ratio (Positive: Negative)

Training + Validation 1 (20%): 4 (80%)

Test 1 (10%): 9 (90%)

To (further) demonstrate robustness of different (graph embedding/NRL) methods on incomplete interaction network. [Does not need all edges to achieve “good quality” graph embeddings].

## 3.3 Experiments

### **Models**

#### Feedforward neural network (FNN)

FNN with early stopping

#### Skip-GNN

Use same code and hyperparameters as original paper

Experiment using different NRL methods to train input features (embedding) for the original graph

### **Network Representation Learning (NRL) methods**

Review of network representation learning (NRL) methods to learn node embedding from an (unsigned) network

* Weighted network (using HVPPI scores (score ranges from 0 to 1) as the edge weights): node2vec+

[Mention that edge weights were changed to -1 and 1 based on threshold of 0.143]

* Signed network (-1 and 1): Neural network signed network embedding (nSNE) {Song, 2018 #50}

## 3.4 Present results for each method + dataset

1. Mention that hyperparameters were selected following guidelines provided by {Yue, 2020 #3}

Table 2: Optimized hyper-parameters (\*based on network reconstruction dataset)

|  |  |
| --- | --- |
| Method | Hyper-parameters |
| deepwalk | num\_walks = 16, walk\_length = 32 |
| Node2vec | num\_walks = 8, walk\_length = 32  p = 0.25 , q= 0.5 |
| Node2vec+ | num\_walks = 8, walk\_length = 32  p = 0.25 , q= 0.5 |
| Struc2vec | num\_walks = 128, walk\_length = 16 |
| Ripple2vec | num\_walks = 8, walk\_length = 8 |
| LINE | 1st + 2nd order, epochs = 10 |
| SDNE | α = 0, β = 10 |
| VGAE | hidden1 = 256, hidden2 = 128 |
| GraRep | ksteps=2 |
| nSNE | K = 128, β = 0.005 |

Comparison of Feed-forward neural network VS Skip-GNN (Network Reconstruction Dataset)

* Why I decided to “abandon” Skip-GNN

2. Choosing of mapping functions to compute edge embedding

* Concatenation, Sum, Average, Hadamard Product
* No choosing required for nSNE

3. Comparison with protein embedding

* For experimentally verified dataset

### Clustering evaluation

* Metrics

1. Adjusted rand index

Function that measures the similarity of two assignments

Takes into account that random chance will cause some (data points) to occupy the same clusters

1. Adjusted mutual information score

Measures the agreement of two assignments (split between clusters and split between the ground truth class labels)

Normalized against chance

1. Homogeneity, completeness and V-measure {Rosenberg, 2007 #53}

Homogeneity: Each cluster contains only members of a single class

Completeness: All members of a given class are assigned to the same cluster

V-measure: Harmonic mean of homogeneity and completeness

1. Fowlkes-Mallows score

Geometric mean of the pairwise precision and recall

# Section 4: Discussion

Comment on overall performance of graph embedding and its performance against protein embedding

* Visualizations (UMAP) and evaluation metrics to support discussion of graph embedding methodologies
* Clustering PPI networks
* Proteins interacting with each other often participates in the same biological processes or can be associated with specific biological functions being strongly related {Tornow, 2003 #52} (i.e., more similar, shorter distance between proteins)
* Segregate data points into groups such that data points placed in the same group are more similar to each other than to those in other groups
* Discussion points:
* Node2vec VS node2vec +
* Weights would help the random walk to focus more on the relevant nodes in the graph
* LINE (1st order proximity VS 2nd and 1st + 2nd order proximity)

In biological networks, similarity between nodes that do not directly interact has proven incredibly useful in the last decade across a variety of interaction networks. {Huang, 2020 #6}

* Neural network methods (i.e., LINE, SDNE and nSNE which uses a MLP model) able to perform better on link prediction task, even when dataset is unbalanced (Looking at PR-AUC scores)
* Better clustering evaluation scores = Better performance on a “new” unseen dataset (i.e., experimentally verified)

[SDNE & nSNE performs better than LINE and node2vec+]

* From **Specificity** scores, SDNE and nSNE have lower false positive rate (0 predicted as 1). Even though testing was conducted on human-human interactions instead of iav-human interactions

## Limitations

Ultimately, HVPPI is a prediction tool. Therefore, it is not 100% accurate.

With reference to (Test 2 Experimentally verified), there exists protein pairs verified by experimental techniques to be interacting but indicated as non-interacting by HVPPI. (i.e., below score of 0.143)

In our dataset, 176 of such samples were included, which hvppi scores ranging from 0.023 to 0.141.

\*\* Need to combine computational method with high-throughput experimental techniques for even more reliable results {Sarkar, 2019 #51}

# Section 5: Future work

* This work can be extended to include other IAV strains.
* Probability instead of binary

# Section 6: Conclusion