COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Veridical Network Embedding

Tian Zheng
Department of Statistics & Data Science Institute,
LEAP (Learning Earth with AI and Physics), an NSF STC
Columbia University

Joint work with Owen Ward, Zhen Huang, and Andrew Davison

Ward, O. G., Huang, Z., Davison, A., & Zheng, T. (2021). Next waves in veridical network embedding. *Statistical Analysis and Data Mining*: The ASA Data Science Journal, 14(1), 5-17.

- 77 nodes and 254 edges
- Aim to discover both homophily and structural equivalence.
- *node2vec* learns embedding, which is followed by *k-means* clustering.

Grover, A., & Leskovec, J. (2016, August). node2vec: Scalable feature learning for networks. In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 855-864).

- Protein-Protein Interactions for *Homo Sapiens*.
- 3890 nodes, 76584 edges.
- 50 *labels* for the nodes.
- Multi-label classification based on the network.
  - Training data: a fraction of nodes and their labels.
  - Learning goal: the labels of the remaining nodes.
  - Workflow:

$$\text{Network} \longrightarrow \text{Embedding} \longrightarrow \text{Predictive modeling of nodal labels}$$

Grover, A., & Leskovec, J. (2016, August). node2vec: Scalable feature learning for networks. In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 855-864).

- Network data $\rightarrow$ vectors (features).
- Represents network structures with geometry.
- Network embedding is rarely the *end goal.*
- Challenges
  - Different notations.
  - Lack of evaluation metrics. "How well does it work?"
- How would the *Veridical Data Science* PCS principles apply?
  - Predictability, Computability, and Stability

- Network: $G = (V, E)$.
- V is the set of n vertices or nodes.
- E is the set of edges, which can also be represented by the adjacency matrix A.
- The edges can be weighted/unweighted, directed/undirected. No self-edge.
- Nodal attributes/covariates: X
- Network embedding:

$$\Phi : G \times \mathcal{X} \to (\mathcal{Z}, \rho).$$

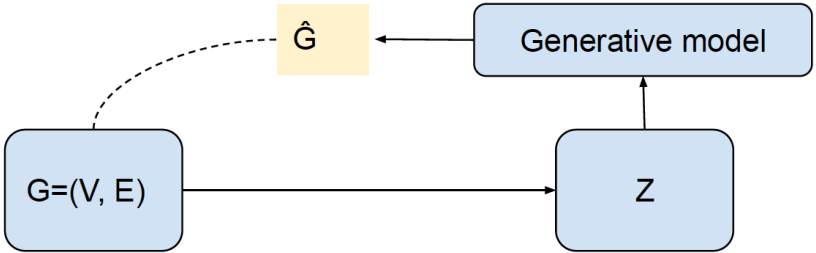
- $\mathcal{Z}$ is a matrix, with vectors in a metric space with associated metric ρ.

- space to represent the network in?
- features of the network to preserve?
  - how do we evaluate whether these features are preserved?
- combine nodal covariates with the learned representation?
- tasks that will use the embedding?

- Representation space
  - Most popular: lower-dimensional Euclidean spaces
  - Unit spheres, hyperbolic spaces
- Features of network to be preserved
  - Connectivity
  - Network structures such as communities
  - *First-order proximity*, e.g., presence of edges
  - *Second-order proximity*, k-step transition probability, homophily, etc
  - Similarity in nodal attributes.
- *Similarity* in the representation space
- Subsequent modeling and learning
  - Clustering and classification of nodes

- Spectral clustering
- Latent space models
- Random-walk based network embedding
- Neural network models for supervised network embedding

- Assume there is a generating process.

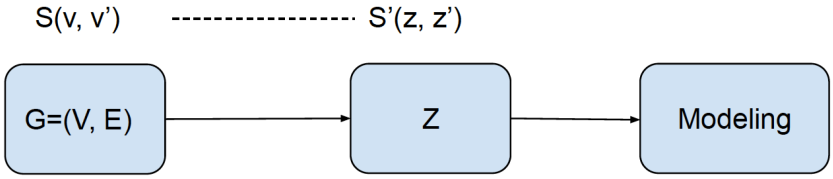


- Structural Deep Network Embedding. Wang, D., Cui, P., & Zhu, W. (2016, August). Structural deep network embedding. In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1225-1234).

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_2 + \alpha\mathcal{L}_1 + \nu\mathcal{L}_{reg} \\ &= \|(\hat{A} - A) \odot B\|_F^2 + \alpha \sum_{i,j=1}^{n} A_{i,j}\|z_i - z_j\|_2^2 + \frac{\nu}{2}\sum_{k=1}^{K}(\|W^{(k)}\|_F^2 + \|\hat{W}^{(k)}\|_F^2),\end{aligned}$$

- A is the adjacency matrix and $\hat{A}$ is the reconstruction based on z.
- The error is weighted by B: $B_{ij} = 1$ if $A_{ij} = 0$ and $B_{ij} = \beta > 1$ if $A_{ij} > 0$.
- $\|\cdot\|_F$ is the Frobenius norm of matrix.
- $W^{(k)}$ and $\hat{W}^{(k)}$ are the weights of the encoding and decoding neural networks.

- "matching the similarities"



- $$L(S, S') = \sum_{i,j} S(v_i, v_j) S'(z_i, z_j) + R(z),$$

TABLE 1 Some unsupervised methods, their similarity measures and the corresponding loss function used to learn the representation

Methods	$S(v_i, v_j)$	$S'(z_i, z_j)$	**Loss function**
Laplacian eigenmap [10]	W_{ij}	$\parallel z_i - z_j \parallel_2^2$	$\sum_{i,j} S(v_i, v_j) S'(z_i, z_j), s.t. z^T D z = I.$
Graph factorization [11]	W_{ij}	$z_i^T z_j$	$\frac{1}{2}\sum_{(i,j)\in E}(S(v_i, v_j) - S'(z_i, z_j))^2 + \frac{\lambda}{2}\sum_i \parallel z_i \parallel_2^2$
Line (1st-Order) [12]	W_{ij}	$\log \sigma(z_i^T z_j)$	$-\sum_{i,j} S(v_i, v_j) S'(z_i, z_j)$
Line (2nd-Order) [12]	W_{ij}	$\frac{\exp(z_j'^T z_i)}{\sum_k \exp(z_k'^T z_i)}$	$-\sum_{i,j} S(v_i, v_j) S'(z_i, z_j)$
GraRep [14]	$p_k(v_j \mid v_i)$	$\log \sigma(z_j'^T z_i)$	$\sum_{i,j}^{\lvert V\rvert}\left[S(v_i, v_j)\log \sigma(z_j'^T z_i) + \frac{\lambda}{\lvert V\rvert}\mathbb{E}_{v_k\sim p_k}\log[1-\sigma(z_k'^T z_i)]\right]$
Deepwalk [3]	co-occurrence	$\log \sigma(z_j'^T z_i)$	$\log \sigma(z_j'^T z_i) + \sum_{l=1}^{k}\mathbb{E}_{v_l\sim P_n}\log(1-\sigma(z_l'^T z_i))$
Node2vec [2]	in RWs		

Note: Here $S(v_i, v_j)$ is the similarity on the network; $S'(z_i, z_j)$ is the similarity in the representation space; (z_i, z_i') is the representation of v_i, and W is the weight matrix. We have used a slight abuse of notation: when the representation $\Phi(v_i) = (z_i, z_i')$ $S'(z_i, z_j)$ should be understood as $S'(\Phi(v_i), \Phi(v_j))$.

- Email network of all students at a US university during one semester in the academic year 2003-2004.
  Kossinets, G., & Watts, D. J. (2006). Empirical analysis of an evolving social network. science, 311(5757), 88-90.
- Nodal information
- Treated as unweighted and undirected.
- An edge: at least one email during the semester.
- Consider the largest connected component.
- 18492 nodes and 260048 edges.
- Methods considered: Spectral embedding, Deepwalk
- Learning tasks: prediction of academic fields ($k = 12$) and student status ($k = 5$) using random forests and logistic regression.
- Implementation: scikit-learn; stratified 5-fold cross-validation; 100 repetitions.

Spectral Clustering of the Graph Laplacian

- **Model Description.**
- **Inputs.** Adjacency matrix A of a graph $G(V, E)$
- **Output.** Cluster assignments of the nodes in the network.
- **Procedure.** Given a Laplacian matrix L of A, construct the bottom k eigenvectors of L, $u_1, \ldots, u_k$ which give the matrix $U = (u_1, \ldots, u_k) \in \mathbb{R}^{n \times k}$. Perform a distance clustering using the rows of U, to cluster each node.
- **Task.** Clustering of the nodes.
- **Source Code.** Implemented in most software packages.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019, January). Model cards for model reporting. In Proceedings of the conference on fairness, accountability, and transparency (pp. 220-229).

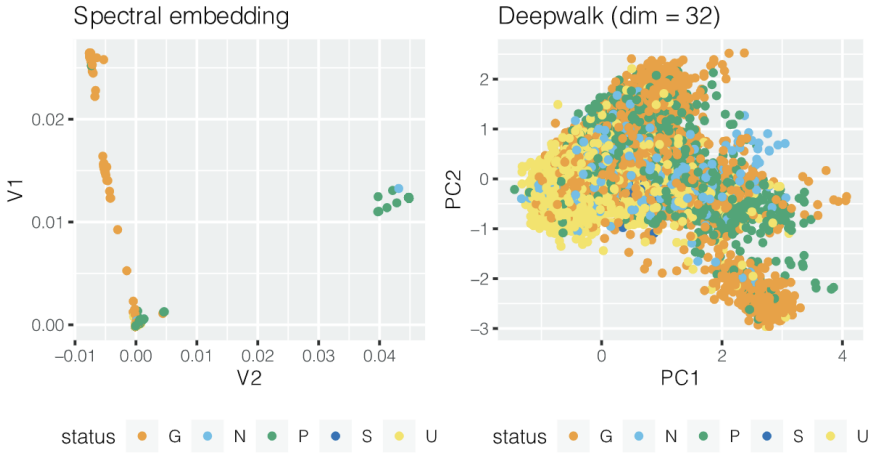
DeepWalk [38]

- **Model Description.** Apply SkipGram to learn network embedding with random walks.
- **Inputs.** Graph $G(V, E)$, window size w, embedding dimension d, number of walks per vertex γ, walk length t.
- **Output.** The vertex representations $\Phi(V) \in \mathbb{R}^{|V| \times d}$.
- **Loss Function.**

$$\sum_{\mathcal{W}_k} \sum_{v_i \in \mathcal{W}_k} -\log \Pr(\{v_{i-w}, \cdots, v_{i-1}, v_{i+1}, \cdots, v_{i+w}\} | \Phi(v_i)).$$

- **Task.** Multi-Label Classification.
  - **Evaluation Data.** BlogCatalog, Flickr, YouTube.
  - **Classification Model.** One-vs-rest logistic regression.
  - **Metrics.** Macro-F_1 and Micro-F_1 scores.
- **Source Code.** https://github.com/phanein/deepwalk.

Email network: Embedding



- Yu, B., & Kumbier, K. (2020). Veridical data science. Proceedings of the National Academy of Sciences, 117(8), 3920-3929.
- **Predictability**: how well a model represents *relationships* in the original data.
- **Computability**: how well does the algorithm scale?
- **Stability**: how much will the results change when the data and/or model are perturbed?

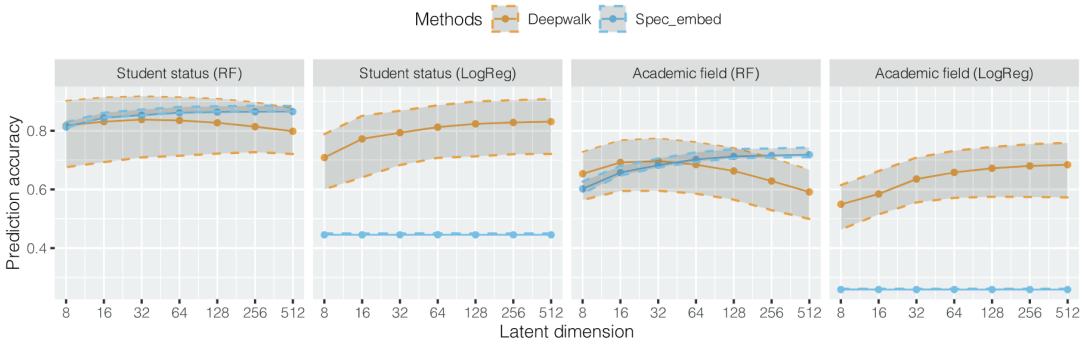
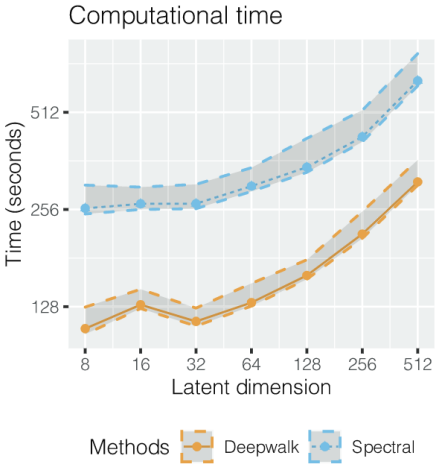


FIGURE 2 The mean accuracy of prediction (averaged over 100 replications, each with 5 scores from stratified cross-validation) for one of the 12 academic fields and one of the five students status with different machine learning models (random forest and logistic regression). The confidence bands are plotted using the 2.5% and 97.5% sample quantiles

Email network: Computability



Factors need to be considered for stability.

- Choice of representation space.
- Features of network to be preserved.
- Similarity considered.
- Learning algorithms.
- Measures of performance (e.g., subsequent learning tasks).
- Perturbation of data (i.e., network).

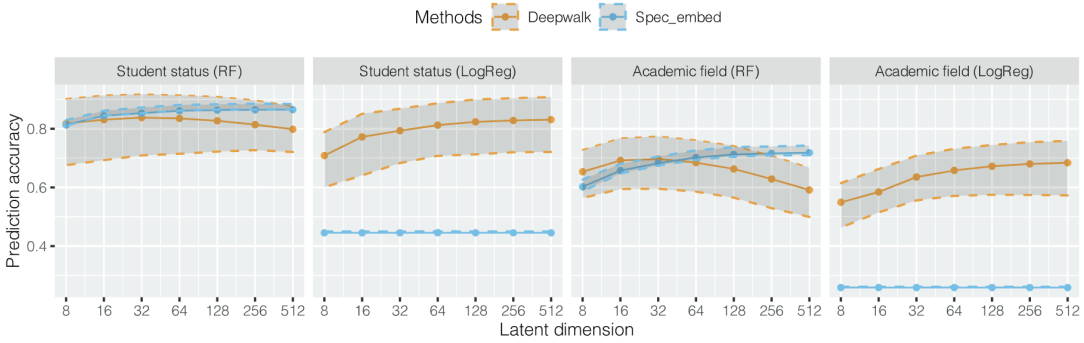


FIGURE 2 The mean accuracy of prediction (averaged over 100 replications, each with 5 scores from stratified cross-validation) for one of the 12 academic fields and one of the five students status with different machine learning models (random forest and logistic regression). The confidence bands are plotted using the 2.5% and 97.5% sample quantiles

- Network embedding is an exciting and important area of research.
- Towards a PCS framework for problem setup and model assessment.
- Many unsolved questions:
  - Richer network features to address domain specific questions.
  - Richer geometry of the embedding space.
  - Further research on understanding the relation between random walks on networks and network structures.
  - Benchmark data sets and ...
  - Metrics, metrics, metrics.

Questions?