Representation Learning and its Applications in Heterogenous Networks

A Saikia
M.S. Data Science
Indiana University - Bloomington
arsaikia@iu.edu

***Abstract*— In this work, we implement metapath2vec, a meta-path based representation learning technique that uses a modified skip-gram model to learn the latent k-dimensional representation of nodes in a user-artist heterogeneous interactions network. We will show that metapath2vec embeddings can be used for heterogeneous network mining tasks like similarity search and node clustering.**

I. INTRODUCTION

Representation learning can be leveraged to encode the structure and semantics of the rich and complex data present in social, biological, and information networks. Advances in natural language processing (NLP), specifically a group of models called word2vec [1] has led to word2vec inspired network representation learning frameworks such as DeepWalk [2], LINE [3], and node2vec [4]. These methods enable automated discovery of low dimensional meaningful features from raw networks by using a random walk based neighborhood sampling strategy followed by a skip-gram model to encode structurally and semantically similar nodes. While these frameworks work well for homogeneous networks, most social, biological, and information networks are heterogeneous with diverse node types and edges between them. An example of such a network is an interactions graph representing users, songs they listen to, and artists/bands who wrote these songs. metapath2vec[5] is a scalable representation learning framework that tries to solve the issues arising from the homogeneous treatment of diverse node types and edges in heterogeneous networks. It learns latent feature representation of nodes by generating meta-paths based biased neighborhoods and leveraging the skip-gram model that maximizes the probability of having a heterogeneous context. The learned features can be then used for network mining tasks such as node classification, community detection, and similarity search.

II. DATA

The dataset [6] has been obtained from Last.fm online music system. It has 1892 users listening to 17,632 artists. Each artist has one-to-many mapping with 11,946 unique tags. The users are also interconnected in a social network generated from Last.fm 'friend' relations. From a network standpoint, the heterogeneous nodes in this network are users (U), artists (A) ,and tags (T). The different experiments we plan to run to test the efficacy of the metapath2vec embeddings are as follows -

- What is the right meta-path for this network?

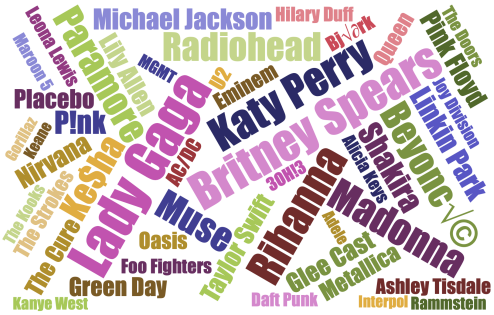

Fig. 1: Wordcloud for top 100 most listened artists

Fig. 2: Wordcloud for top 150 most common tags

- Use embeddings to perform similarity search and recommend similar artists to users.
- Perform clustering on artist nodes. Is metapath2vec able to cluster artists of the same genre?
- Look at the sensitivity of these results by varying hyper-parameters of the algorithm like walks per node, walk length, dimensions and neighborhood size
- Use TensorFlow Embedding Projector [7] to visualize if the embedding vectors can implicitly learn semantic similarities between nodes.

III. EXPERIMENTS

A. Exploratory Analysis

We first explored the user-artist interactions graph. Figure 3 shows the bipartite structure of the graph where the nodes on the left are the users and the nodes on the right are the artists. The node size is proportional to the degree of the nodes.

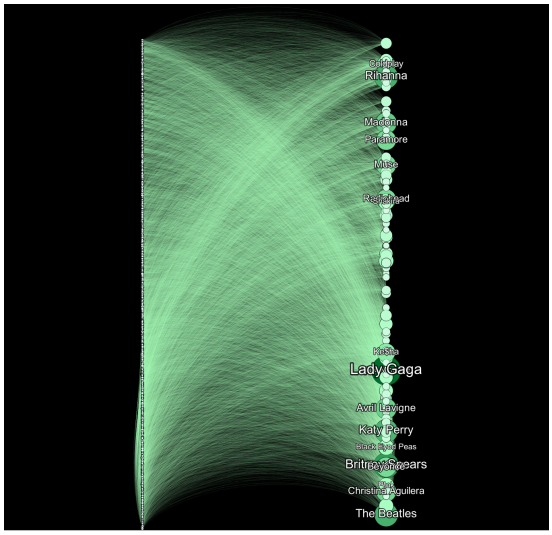

Fig. 3: User-Artist interactions graph for top 125 most listened artists.

We also looked at the friendship network (figure 5) to identify influential nodes in this friendship network. Nodes are sized by eigenvector centrality.

We also used the associated tags for each artist to construct the influence network (figure 6) , where an edge exists between two artists if they have at least 20 tags in common. The different colors represent the different communities detected by the Louvain method with modularity optimization. The nodes are sized by the betweenness centrality. We can see artists like 'The Beatles' and 'Madonna' are very prominent who have been historical very popular in influencing many artists.

B. Metapath2vec Embedding

Choosing the right metapath scheme is very crucial in creating the perfect embeddings. The metapaths are similar to those generated by random walkers, except in this case the walkers are biased as per the metapath scheme. The different metapath schemes that were identified are as follows -

- Artist – User – Artist (AUA)
- User – Artist – User (UAU)
- User – Artist – Tag – Artist – User (UATAU)

Generating the actual metapaths depends on two hyper-parameters - Walk Length (l) and Walks per node (n). From a computational standpoint, generating the metapaths is a complex task and takes an exorbitant amount of time. For this project - two sets of metapaths were created with the 'UAU' scheme having $l = 10$ and $n = \{5, 10\}$. The meta-path scheme 'UAU' represents the music co-preference relationships on an artist (A) between two users (U).

The skip-gram model then uses these metapaths to create the target and context node pairs by sliding a window over each metapath. The skip-gram model itself has two hyper-parameters the neighborhood size (d) which represents the size of the window and embedding size (k) which represents the size of the latent vector/embedding. For each

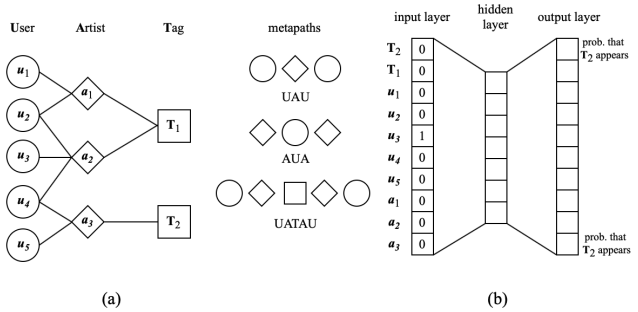

Fig. 4: An illustrative example of (a) heterogeneous user-music interactions network and (b) skip-gram architecture of metapath2vec for embedding this network.

of the two sets of metapaths, 16 sets of embeddings were generated with $k = \{16, 32, 64, 128\}$ and $d = \{3, 5, 7, 10\}$. Next, we evaluate the quality of the latent representations over two classical heterogeneous network mining tasks - similarity search and node clustering. In addition, we also use the TensorFlow Embedding Projector [7] to visualize the nodes in 3-dimensional space.

IV. Results

A. Similarity Search

To test the efficacy of the embeddings, we performed similarity search by calculating cosine similarity between the node embeddings. Since there was no ground truth we had to manually inspect the results and decide which hyper-parameters i.e. l, n, d, k did better than the others.

TABLE I: Similarity search with l = 10, n = 5, k = 16, d = 3

Rank	Lady Gaga	Iron Maiden	Linkin Park
1	Rihanna	Megadeth	Green Day
2	Glee Cast	AC/DC	My Chemical Romance
3	Kylie Minogue	Metallica	30 Seconds to Mars
4	Beyonce	Black Sabbath	Nickelback
5	Britney Spears	Guns N' Roses	Paramore
6	Shakira	Rammstein	Evanescence
7	Nelly Furtado	Queen	Fall Out Boy
8	Alicia Keys	735	Panic! At the Disco
9	David Guetta	Led Zepplin	Marilyn Manson
10	Christina Aguilera	827	Avenged Sevenfold

Table I lists the top 10 similar results for querying 3 very popular artists - Lady Gaga, Iron Maiden and Linkin Park. One can see that for the query 'Iron Maiden', we get artists of the similar heavy metal/hard rock genre like Megadeth(1st), AC/DC(2nd), Metallica(3rd). Similarly for the query 'Linkin Park' we get artists of the similar punk/alternative rock genre like Green Day(1st), My Chemical Romance(2nd), 30 Seconds to Mars(3rd). We observed that for a few cases, the query returned users (designated by numbers) in the top 10 similar nodes for artists. This indicates that the embeddings are not perfect and the hyper-parameters require further refinement.

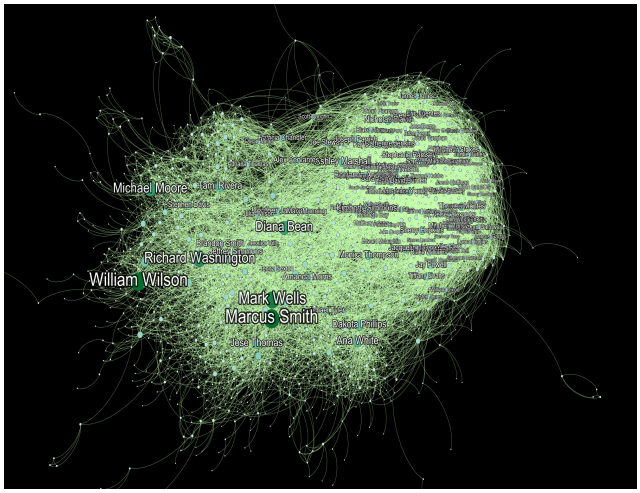

Fig. 5: User-User friendship network

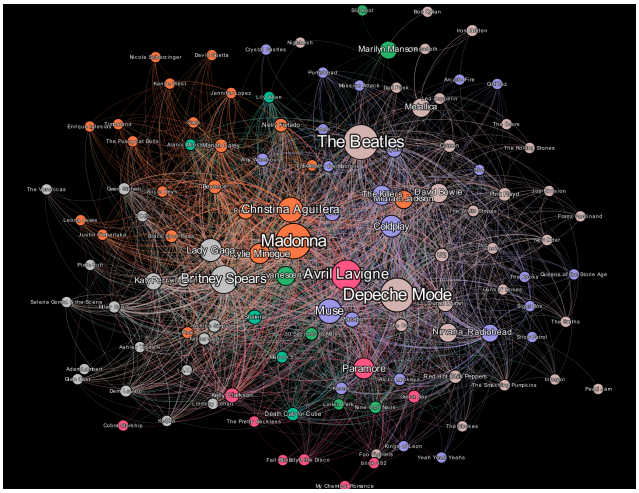

Fig. 6: Artist-Artist influence network

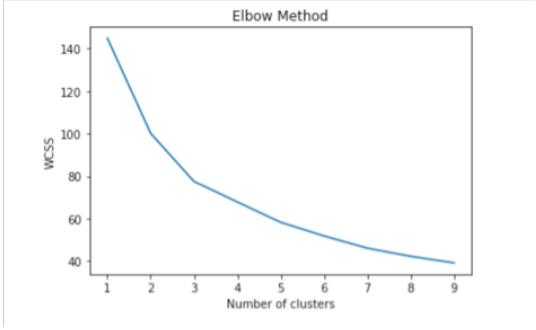

Fig. 7: Elbow curve after fitting k-means. Optimum value of number of clusters - k is inconclusive

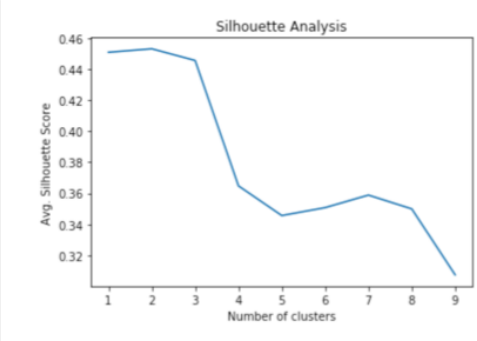

Fig. 8: Average silhouette score as a function of number of clusters after fitting k-means. Optimum value of k is 2

B. Node Clustering

Next, we checked the performance of the embeddings in clustering the artist nodes. We leveraged two different clustering algorithms - k-means and DBSCAN to get meaningful clusters of artists. While k-means tries to build even sized clusters by minimizing the sum of squared distance of all points from the respective cluster centers, DBSCAN tries to grow each cluster by combining points belonging to the same neighborhood.

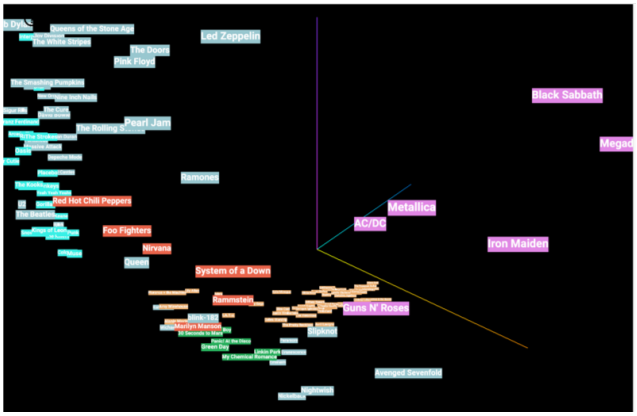

Fig. 9: Results of DBSCAN projected onto 3D space after PCA. Artists are colored based on cluster assignment.

We looked at the elbow curve (figure 7) and silhouette analysis (figure 8) to check the performance of k-means. Both these techniques are commonly used to get the optimum number of clusters (k) in a completely unsupervised setting. We observed that the highest value of average silhouette score was for $k = 2$. But this did not seem right as there were certainly more two categories of artists in the data.

Next, we tried DBSCAN to find groups of similar artists. In the absence of any ground truth, we had to visually inspect the clusters to evaluate the performance of the algorithm. We used TensorFlow Embedding Projector [7] to visualize the results in lower dimensional (2D/3D) space. Figure 9 shows DBSCAN was able to find six group of artists. It was successfully able to cluster artists belonging to niche genres like pop (in yellow), heavy-metal (in purple). But the cluster assignments were far from perfect and in the absence of an evaluation metric one could easily argue against these assignments.

V. Conclusion

Representation learning is a very powerful and effective technique for mining large volumes of heterogeneous

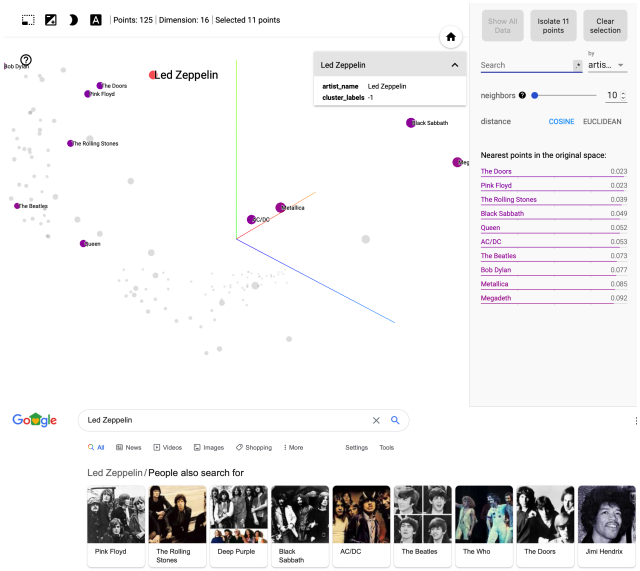

Fig. 10: Comparison of results from metapath2vec based similarity (top) and Google search (bottom) for the query 'Led Zeppelin'. We can see similar bands as the top results like 'Pink Floyd', 'The Doors', 'The Rolling Stones' etc.

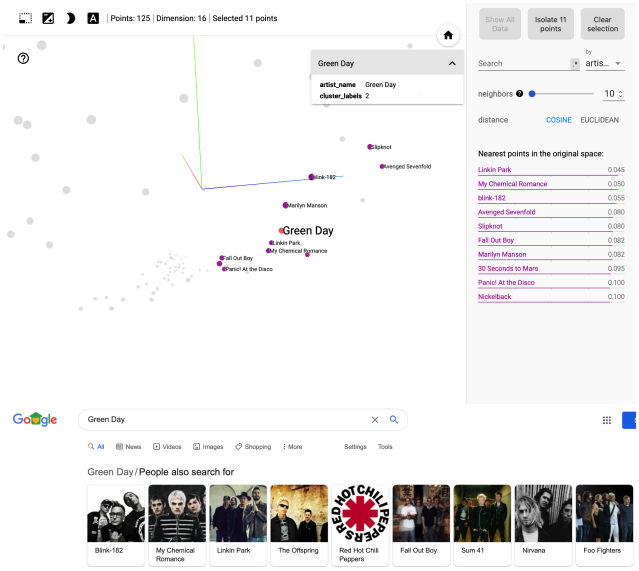

Fig. 11: Comparison of results from metapath2vec based similarity (top) and Google search (bottom) for the query 'Green Day'. We can see similar bands as the top results like 'Blink-182', 'My Chemical Romance', 'Linkin Park' etc.

information networks. In this work, we show its efficacy in similarity search and clustering. We were able to find similar artists using cosine similarity very accurately and the results were even comparable with those returned by Google search engine (see Figure 10 and Figure 11). For node clustering, in the absence of ground truth, it is challenging to find the best clusters, and test for parameter sensitivity.

It would be interesting to see how these results compare with those of homogeneous network embedding methods like Node2vec [4], Deepwalk [2] etc. Also, since the latent embeddings are learned representation of the node's content from its structural and semantic relationships in the network, one could try to augment the generated embeddings with metadata and gauge the effect of additional metadata on the prediction performance. One interesting application of these embeddings could be for artist recommendation, where historical user listening behavior is used to create a taste vector by averaging the embeddings of the artists the user has listened to. New artists can be recommended by querying artists similar to the taste vector.

VI. Source Code

Source code for all experiments and results can be found here https://github.com/arunavsk/Network-Science-I606.

VII. Acknowledgements

The author wishes to thank Prof YY Ahn for his instructions and support. This work was part of Network Science (INFO-I606) course for Spring 2020 at Indiana University, Bloomington.

References

[1] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. CoRR abs/1301.3781 (2013). http://arxiv.org/abs/1301.3781

[2] Perozzi, Bryan, Rami Al-Rfou, and Steven Skiena. "Deepwalk: Online learning of social representations." In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 701-710. 2014

[3] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. LINE: Large-scale Information Network Embedding.. In WWW '15. ACM.

[4] Aditya Grover and Jure Leskovec. 2016. Node2Vec: Scalable Feature Learning for Networks. In KDD '16. ACM, 855–864.

[5] Dong, Yuxiao, Nitesh V. Chawla, and Ananthram Swami. "metapath2vec: Scalable representation learning for heterogeneous networks." In Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 135-144. 2017.

[6] http://ir.ii.uam.es/hetrec2011/datasets.html

[7] Smilkov, Daniel, Nikhil Thorat, Charles Nicholson, Emily Reif, Fernanda B. Viégas, and Martin Wattenberg."Embedding projector: Interactive visualization and interpretation of embeddings." arXiv preprint arXiv:1611.05469 (2016).