Community Detection of Knowledge Communities with Augmented Graph Convolution Networks

Carringtone Kinyanjui [1]   Simone Turchetti [1]   Roberto Lalli [2]

[1]University of Manchester, CHSTM   [2]Polytechnic University of Turin

Community Detection

Community detection is the process of automating the discovery of communities in networks.

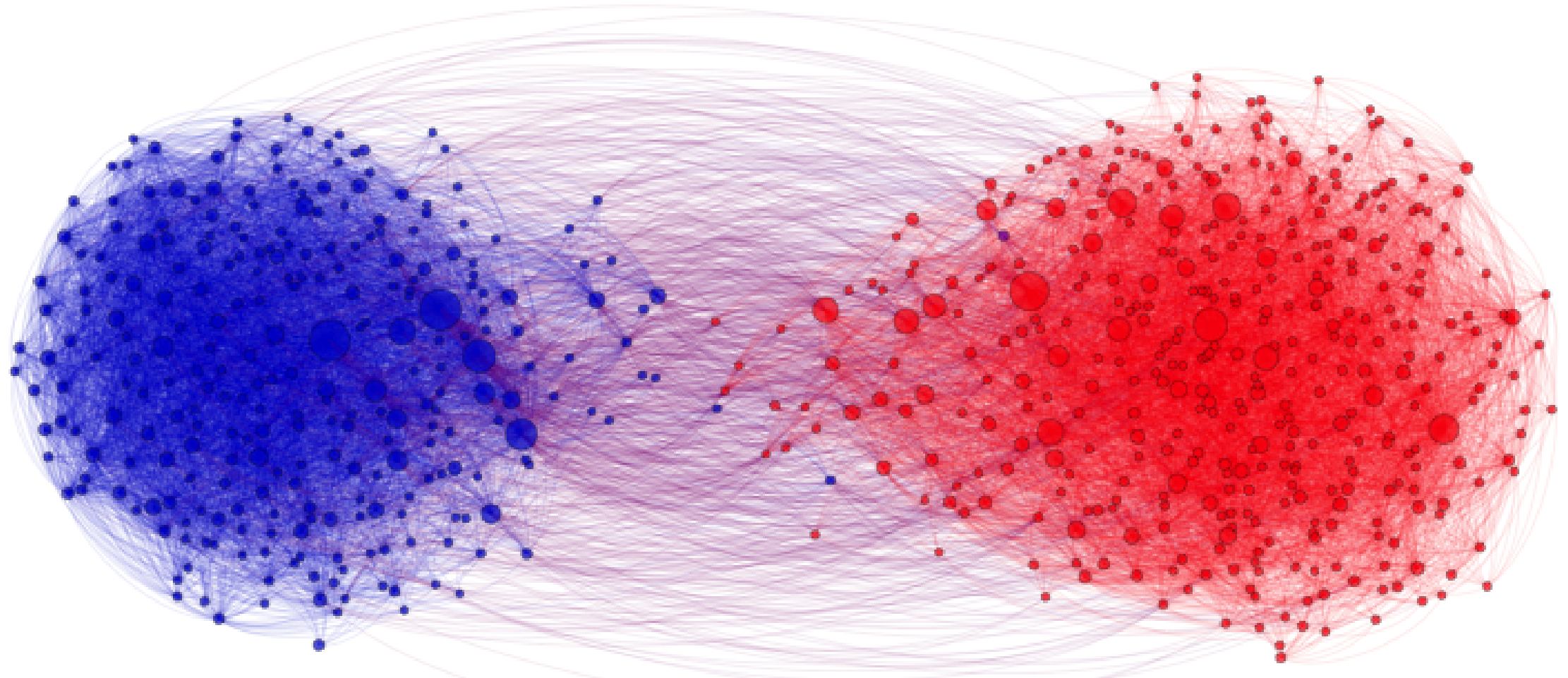

Figure 1. Community detection used to infer political polarisation in the United States. From: allthingsgraphed.com

A social network is the product of interaction of multiple agents within a particular setting. The analysis of the structure, sub-structure and invariances within social networks is known as social network analysis (SNA). This is typically operationalised by applying the methods of graph theory in understanding social to Social networks. SNA has emerged as a versatile discipline, with applications to economics, political science, physics and scientometrics.

Community Detection

A community is a substructure of social networks where links between the nodes are stronger within the nodes than without. This is the notion of metaphorical closeness communicated in the figure above. Communities yield a wealth of information depending on the context of the social network studied. Communities could be a sign of polarisation, formation of new interests in social media, and institutionalisation of a scientific research program. Community detection thus is a critical process of understanding the evolution of social networks. Given the size, diversity and evolving nature of social networks, it is necessary to automate this process. There are means of achieving this:

- **Matrix Factorisation** Social networks are adequately represented using an adjacency matrix. A factorisation of this matrix could yield information on community structure
- **Kernel Methods** Once an embedding of the nodes is obtained, one could cluster the node embeddings, yielding information on community structure.
- **Deep Learning** This represents SOTA in community detection because of the structure preserving ability of deep learning algorithms especially of high dimensional structure of networks.

Examples of deep learning algorithms used for community detection are Generative Adversarial Networks (GAN) an Graph Convolution Networks (GCN).

Abstract

We present preliminary work developing an NLP augmented community detection algorithm. We show why we expect theoretical improvements by use of simple embedding layers on a sigmoid activation function. We then present the results of experimental tests. The data for experimental test is obtained from scientometric data, specifically of nuclear physics research articles published from 1980-2010. First an LDA clustering of the work is performed, followed by a one-hot encoding of the nodes. Secondly, the nodes are encoded using NLP information.

Previous Work

Deep learning algorithms used for community detection have established different forms of performance accross different domains. Examples of SOTA and near SOTA deep learning community detection algorithms below:

Model	Accuracy	Overlapping Communities?
CNN (*Xin et. al.*)	80%	No
GCN		No
Graph Attention	80%	yes
GAN(*Jia et. al.*)	0.904	Yes

Table 1. Performance of Deep Learning Algorithms

Majoriity of the Learning algorithms studied above abstract the graph from the Neural networks. In this case the social network is extracted from data, node embeddings are then learnt on the nodes themselves. The content, or purpose of the social network is generally not taken into account.

We propose that the content of the communication channels of social networks should be taken into account. This could augment node embeddings and could help in clustering of the nodes according to node properties. We present here theoretical reasons why we hold this to be true. We present the data we will used to test this hypothesis and preliminary results from node embedding experiments.

Data

Our data contains a source of scientometric data across different domains. This involves the the primary research papers, with Author institution metadata extracted from these papers. We do not follow the standard scientometric scheme of singularly obtaining metadata because it is important to also obtain the papers themselves, ie the **content** of the social networks. We follow the steps below to extract the data and feed it into a Graph Convolution Network

1. **Data Extraction:**Once scraped, the data Spacy was used to extract the names of the countries from which the papers were published. This was then stored. Next an adjacency matrix was extracted from the affiliation matrix. Latent Dirichlet allocation was then used to created simulated specialised communities within the Nuclear Physics research community.
2. **Training** The data was then passed through a Graph Convolution Network. One-hot encoded nodes together with the adjacency matrices were passed through the GCN. From this, the GCN learns node embeddings which it could then use for community clustering. We present below the results of preliminary experiments.

Results

Below are the results of the baseline study. We trained two variations of GCNs. The first was trained on two layers and the second was trained on six layers. We obtained better results which are shown below. We also show that the GCN is stable by training the network on 10,000 epochs.

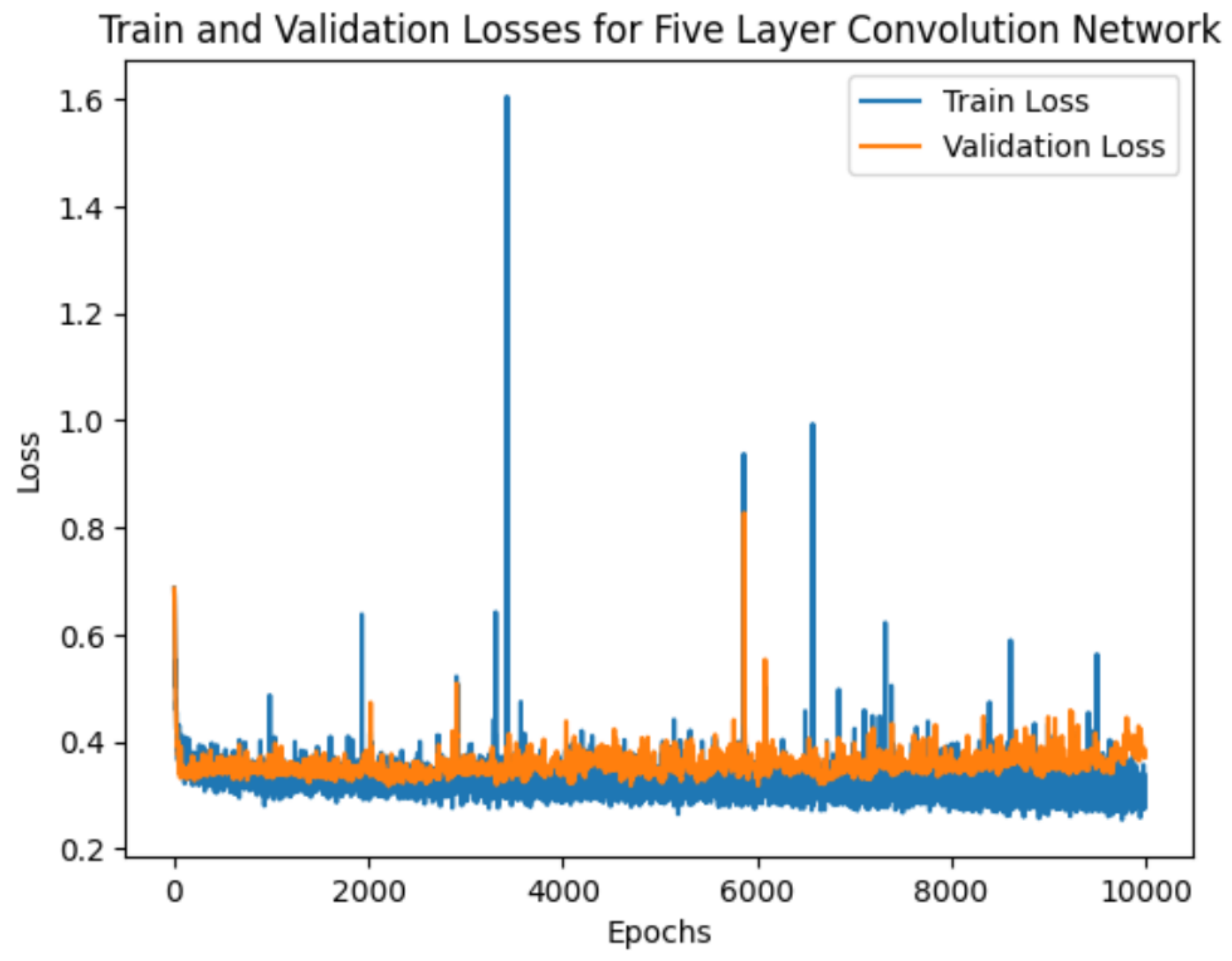

Figure 2. Train and Test Losses for Five Layers and Ten Thousand Epochs

Discussion and Conclusion

The GCN is able to converge efficiently, obtaining reasonable performance. However we were only able to obtain marginal improvements in node classification. The next step in our work is to augment the node embeddings using pretrained NLP embeddings to see whether the community detection algorithm will register marginal improvements. Theoretically, we would expect that the addition of the NLP embeddings will improve the ability of the network to carry out community detection, if the bottlenecks of NLP saturation and curse of dimensionality are avoided.

Theoretical Considerations

We consider an extension of Yuting et. al. Of interest is the Discriminator which is just a product of sigma activation functions

$$D(S) = \Pi\sigma(d_u^t \cdot d_v)$$

where d_v and d_u are vectors representing node embeddings. We can extend and decompose the representation vectors to

$$= \frac{e^{T_G + T_{NLP}}}{e^{T_G + T_{NLP}} + 1}$$

It can be seen that when $T_G \gg T_{NLP}$, this is the standard network embedding task in network analysis. On reverse when $T_{NLP} \gg T_G$, the problem reduces to an NLP clustering. Thus, a balance ought to be sought when carrying out NLP-augmented graph learning.

References

[1] Xing, S., Shan, X., Fanzhen, L., Jia, W., Jian, Y., Chuan, Z., ... Yu, P. S. (2022). A comprehensive survey on community detection with deep learning. IEEE Trans. Neural Netw. Learn. Syst.

[2] Jia, Y., Zhang, Q., Zhang, W., Wang, X. (2019, May). Communitygan: Community detection with generative adversarial nets. In The World Wide Web Conference (pp. 784-794).