Overlapping community detection with graph neural networks

Introduction

Network science is a field that has seen significant development in the last years due to the increment in quantity and data availability. With the increasing communication mechanisms and the great amount of data those interactions produce, the academic and productive sectors have seen the importance of complex network analysis. Many of the daily processes or systems that we use today can be represented as a network, such as optimization for transportation networks (Danila et al., 2006; Zanin and Lillo, 2013), recommendation systems for e-commerce (Zanin et al., 2008; Fiasconaro et al., 2015), information diffusion in social networks (Wasserman and Faust, 1994), etc.

A **complex network** is a mathematical representation of a discrete real-world complex system (El-moussaoui et al., 2019). They are frequently studied through graph theory, in which a network *G* is represented as a set of nodes or vertices *V* connected by *E* edges, denoted by ***G****(V, E)*. A graph is usually represented as a binary squared matrix which indicates whether a node is connected to another. If the graph is undirected, the matrix is symmetric, and if the edges are weighted, the matrix is no longer binary, i.e., each cell contains the weight that relates the two nodes. For instance, nodes can be airports or stop lights in transportation networks, and edges can represent following relations or friendships in social networks. Though graph theory is a purely mathematical and structural area, complex network science deals with the actual concepts the nodes and edges represent.

A **community** in a complex network is a sub-graph in which the nodes are densely connected between them and sparsely connected with the rest of the network nodes (Fortunato, 2010), as shown in Figure 1. Detecting communities allows us to analyze social and biological networks (Girvan & Newman, 2002), detect fraud (Pinheiro, 2012), or find affine researchers to a certain set of topics (Rebhi et al., 2016).

This proposal is largely based on Shchur and Günnemann (2019) work.

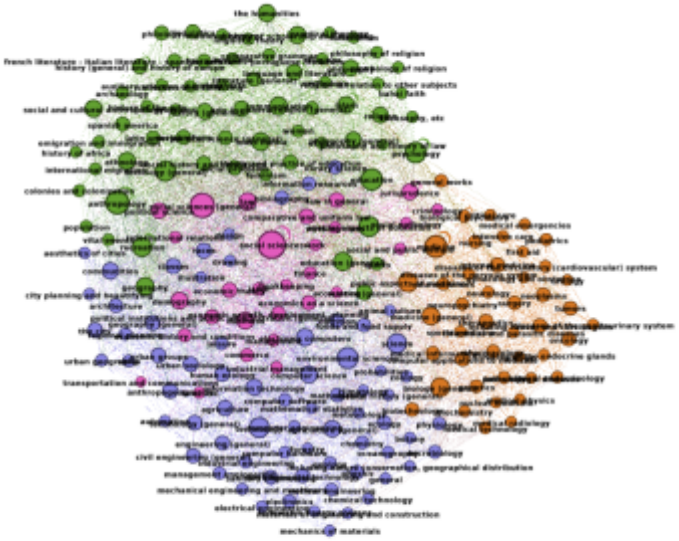

Fig 1. Communities in the network of related concepts in OpenAlex database. Each color represents a community of closely related concepts. Source: Own elaboration.

Objective

Assign each node to C communities. Such assignments can be represented as an affiliation matrix $F \in \mathbb{R}^{NxC}$, where F_{uc} denotes the strength of the node u's membership in community c. The main idea is to generate F with a Graph Neural Network (GNN):

$$F := GNN_\theta(A, X)$$

Where A is the adjacency matrix of the network, and X is the matrix of characteristics of the nodes. The strategy is to construct a 2-layer graph convolutional neural network with the hidden size 128 and the output of size C (number of communities to detect). The activation function must be a ReLu to ensure the non-negativity of F. The likelihood function is:

$$\mathcal{L}(F) = -\mathbb{E}_{(u,v)}[log(1 - exp(-F_u F_v^T))] + \mathbb{E}_{u,v}[F_u F_v^T]$$

Dataset

The OpenAlex dataset has five scholarly entities:
Works: papers, books, proceedings, datasets, etc.
Authors: are the people who created a work.
Venues: The journal or event that hosts a work.
Institutions: Universities or other organizations that have an affiliation with authors.
Concepts: A set of subjects that tag every work.

These entities are organized in a documental database where each register of entities is enclosed in a list called a collection. Each register of each collection is related to multiple registers of other collections as depicted in the schema in Fig 2 and a sample of the coauthorsip network it spans in Fig 3.

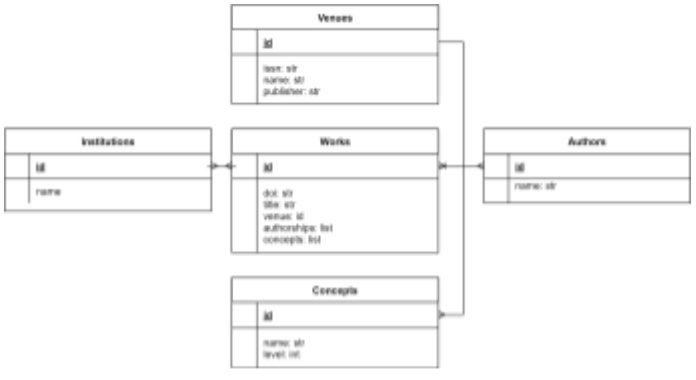

Fig 2. Relational schema for OpenAlex database entities. Source: Own elaboration.

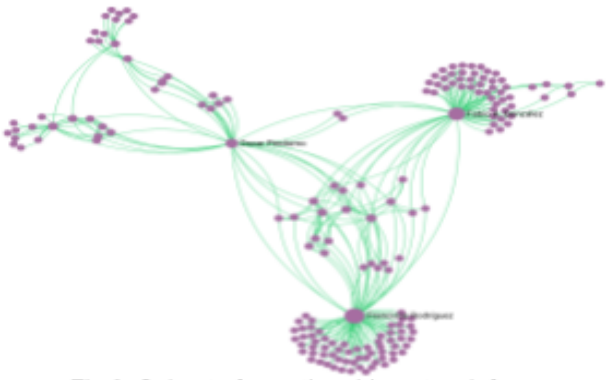

Fig 3. Subset of coauthorship network from colombian researchers. Source: Own elaboration.

From the database, I sampled all the works with at least one author from a Colombian university. From that set, I filled out the other entities related to each work, that is, coauthors, their affiliations, the concepts, and the venues.
From that subset of the database, I constructed a coauthorship graph, in which the nodes are authors that are connected if they publish at least a paper together, and the edges contain the number of collaborations the authors have. The other variables from the author constitute the matrix of characteristics.
Anyone can access the dataset at https://zenodo.org/record/7186309

Performance metrics

Overlapping normalized mutual information (McDaid et al., 2011) with ground truth communities assigned by the concepts labels in OpenAlex database.

References

Danila, B., Yu, Y., Marsh, J. A., & Bassler, K. E. (2006). Optimal transport on complex networks. *Physical Review E*, *74*(4), 046106

Fiasconaro, A., Tumminello, M., Nicosia, V., Latora, V., & Mantegna, R. N. (2015). Hybrid recommendation methods in complex networks. *Physical Review E*, *92*(1), 012811..

McDaid, A. F., Greene, D., & Hurley, N. (2011). Normalized mutual information to evaluate overlapping community finding algorithms. *arXiv preprint arXiv:1110.2515*.

El-moussaoui, M., Agouti, T., Tikniouine, A., & El Adnani, M. (2019). A comprehensive literature review on community detection: Approaches and applications. *Procedia Computer Science*, *151*, 295-302.

Fortunato, S. (2010). Community detection in graphs. *Physics reports*, *486*(3-5), 75-174.

Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. Proceedings of the national academy of sciences, 99(12), 7821-7826.

Pinheiro, C. A. R. (2012). Community detection to identify fraud events in telecommunications networks. *SAS SUGI proceedings: customer intelligence*.

Rebhi, W., Yahia, N. B., & Saoud, N. B. B. (2016, November). Hybrid community detection approach in multilayer social network: scientific collaboration recommendation case study. In *2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA)* (pp. 1-8). IEEE.

Shchur, O., & Günnemann, S. (2019). Overlapping community detection with graph neural networks. *arXiv preprint arXiv:1909.12201*.

Wasserman, S., & Faust, K. (1994). Social network analysis: Methods and applications.

Zanin, M., Cano, P., Buldú, J. M., & Celma, O. (2008, January). Complex networks in recommendation systems. In *Proc. 2nd WSEAS Int. Conf. on Computer Engineering and Applications, World Scientific Advanced Series In Electrical And Computer Engineering. Acapulco, Mexico: World Scientific Advanced Series In Electrical And Computer Engineering*.

Zanin, M., & Lillo, F. (2013). Modelling the air transport with complex networks: A short review. *The European Physical Journal Special Topics*, *215*(1), 5-21.