MLNS Project proposal
Given Youtube data use Community detection algorithms to detect groups of users

Gladys Roch / Aurélien Houdbert

March 2021

1 Motivation and Problem Definition

In social networks (and any real-world network in general), nodes organize into densely-connected groups : network communities. In this project we chose to work on a Youtube network dataset provided by Stanford snap platform along with a paper presenting results of community detection on various graph networks [8]. The objective will be to identify communities among the network using different algorithms and approaches and compare their performances. Community detection is an already very covered subject. Originally approached with statistics [4] recent papers proposed deep learning approaches to deal with community detection [6, 2, 9].

2 Data

We are using the Youtube-online-social-network database with ground-truth communities from Stanford Large Network Dataset Collection [1].

Network statistics	
Nodes	1,134,890
Edges	2,987,624
Average clustering coefficient	0.0808
Diameter (longest shortest path)	20
90-percentile effective diameter	6.5

Community statistics	
Number of communities	8,385
Average community size	13.50
Average membership size	0.10

3 Methodology

Community detection methods can be broadly categorized into two types ; Agglomerative Methods and Divisive Methods. In Agglomerative methods, edges are added one by one to a graph which only contains nodes. Edges are added from the stronger edge to the weaker edge. Divisive methods follow the opposite of agglomerative methods. In there, edges are removed one by one from a complete graph.

Four popular community detection algorithms are :

- Louvain Community Detection
- Surprise Community Detection
- Leiden Community Detection
- Walktrap Community Detection

The objective is to implement, study and improve on these algorithms.

3.1 Louvain Community Detection

The Louvain method is an heuristic algorithm to detect communities in large networks. It maximizes a modularity score for each community, where the modularity quantifies the quality of an assignment of nodes to communities. This means evaluating how much more densely connected the nodes within a community are, compared to how connected they would be in a random network. [5] It is widely seen as one of the best algorithms for detecting communities, but can lead to arbitrarily badly connected communities.

3.2 Surprise Community Detection

Surprise is another metric that evaluates the quality of a partition of a network into communities. The algorithm based on the Surprise metric is similar to the Louvain community detection algorithm except that it uses surprises instead of modularity. [1]

[1] http ://snap.stanford.edu/data/com-Youtube.html

3.3 Leiden Community Detection

This algorithm improves on the Louvain algorithm. It guarantees to find well-connected clusters and is faster. [7]

3.4 Walktrap Community Detection

Walktrap is another approach for community detection based on random walks in which distance between vertices are measured through random walks in the network. The idea of the algorithm is that random walks on a graph/network tend to get trapped into densely connected parts corresponding to communities. Walktrap uses the result of random walks to merge separate communities in a bottom-up manner. Quality of the partitions can be evaluated using any available quality criterion. It can be either modularity as in Louvain community detection or any other measure. [**walktrap**]

4 Evaluation

There exist many metrics to evaluate the quality of communities. Four of them are well adapted as they quantify desirable properties we want to see in our predicted communities.

– ***Separability*** characterises how well a community is separated from the rest of the network.
– ***Density*** will characterise how well nodes from a community are connected to each other.
– ***Cohesiveness*** characterises the structure of the communities. Indeed it is expected from communities to be homogeneously connected and one should not be able to split the community into smaller ones.
– ***Clustering coefficient*** represents how well neighbors of a node are connected with each other.

Intuitively, communities are expected to follow certain characteristics and these four metrics are well adapted to represent the quality of communities when combined together.

As we haven't started working on the project yet, these metrics seem to be the most interesting ones to compare and optimize when building our communities, but the structure of the network or the algorithms we decide to implement might have an impact on those metrics and we might reconsider choosing other better adapted metrics.

Références

[1] Marín I Aldecoa R. "Deciphering Network Community Structure by Surprise." In : (2011). url : `https://doi.org/10.1371/journal.pone.0024195`.

[2] Sandro Cavallari et al. "Learning Community Embedding with Community Detection and Node Embedding on Graphs." In : (2017). url : `https://sentic.net/community-embedding.pdf`.

[3] Javier Del Ser et al. "Community detection in graphs based on surprise maximization using firefly heuristics". In : (juil. 2016), p. 2233-2239. doi : `10.1109/CEC.2016.7744064`.

[4] M. Girvan et M. E. J. Newman. "Community structure in social and biological networks." In : (2002). url : `https://www.pnas.org/content/pnas/99/12/7821.full.pdf`.

[5] Mahantesh Halappanavar Hao Lu et Ananth Kalyanaraman. "Parallel Heuristics for Scalable Community Detection". In : (2014). url : `https://arxiv.org/pdf/1410.1237.pdf`.

[6] Fanzhen Liua et al. "Deep Learning for Community Detection : Progress, Challenges and Opportunities." In : (2020). url : `https://arxiv.org/abs/2005.08225`.

[7] L. Waltman V. A. Traag et N. J. van Eck. "From Louvain to Leiden : guaranteeing well-connected communities". In : (2019). url : `https://arxiv.org/pdf/1810.08473.pdf`.

[8] Jaewon Yang et Jure Leskovec. "Defining and Evaluating Network Communities based on Groundtruth." In : (2012). url : `https://arxiv.org/abs/1205.6233`.

[9] Liang Yang et al. "Modularity Based Community Detection with Deep Learning." In : (2016). url : `https://www.ijcai.org/Proceedings/16/Papers/321.pdf`.