Community Identification among Video Social Platform Youtube

Network Science Analytics – Final Project

Florian BARRAL – Maria BOSCH – Rémi CANARD – Hugo FERNANDEZ – Clarice HAYRABEDIAN

Our Team

Florian BARRAL

Maria BOSCH

Rémi CANARD

Hugo FERNANDEZ

Clarice HAYRABEDIAN

Outline

1. Introduction and motivation – ***What are we working on ?***
2. Problem definition and Related work
3. Methodology
4. Evaluation
5. Conclusion

Introduction and Motivation

- Leading question: ***how to detect communities in social networks ?***
- Main applications:
  - Optimization of:

Product recommendations

Targeted marketing

- Increase global traffic

Outline

1. Introduction and motivation
2. Problem definition and Related work – ***Main goals and Challenges***
3. Methodology
4. Evaluation
5. Conclusion

Defining the problem

OBJECTIVE:

Detect communities in the graph that matches the Ground Truth Communities already present in the dataset

DATASET DESCRIPTION

- From a selection of communities containing **at least 4 members**
- We have a **connected undirected graph** containing:
  - More than **1.1 million nodes** and almost **3 million edges**
  - An **average degree of 5.2 nodes**

Main Challenges

- Reduce the **computation time** of the classic community detection algorithms given the huge size dataset

- Given the **scarce literature** and freely available **code implementations** or libraries:
  - difficulty to find **algorithms** to suit more than a **specific dataset**
  - difficulty to assess which **scoring metrics** are going to be best

Related Work

Andersen et al. (2006)

- In order to improve computation time in community detection, the latter can be done **starting from a node and without analyzing the full graph**

Yang and Leskovec (2012):

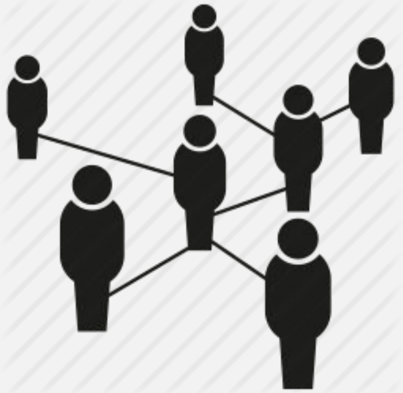
- They work shows that scoring functions such as **the conductance score** well capture the structure of ground-truth communities
- They explore **detecting communities from a single seed node**

Outline

1. Introduction and motivation
2. Problem definition and Related work
3. Methodology – ***How we address our problem***
4. Evaluation
5. Conclusion

Data Preparation

Graph studied:
Nodes: 1 134 890
Edges: 2 987 624



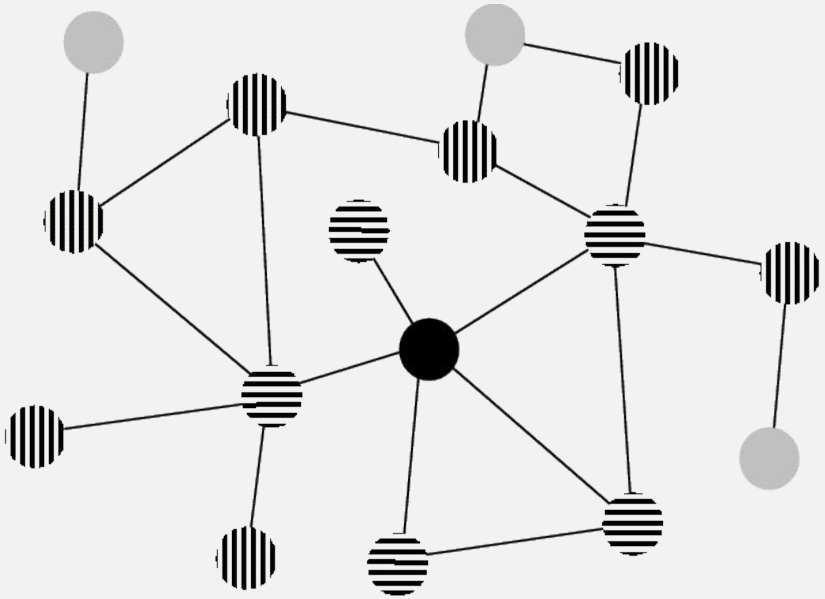
⚠ Classical algorithms applied to complete graph ⟶ Very high computational time

💡 Smart algorithm exploring graph from a seed node

Original approach from a seed node

Discovering communities from a seed node ...

⟶ Automatic detection of the community of a node and its other members



Benefits of this approach:

1. *No specific input data:* no hyperparameters
2. *Scalability*: Computational time proportional to the size of the detected community (NOT the size of the network)

Original approach from a seed node

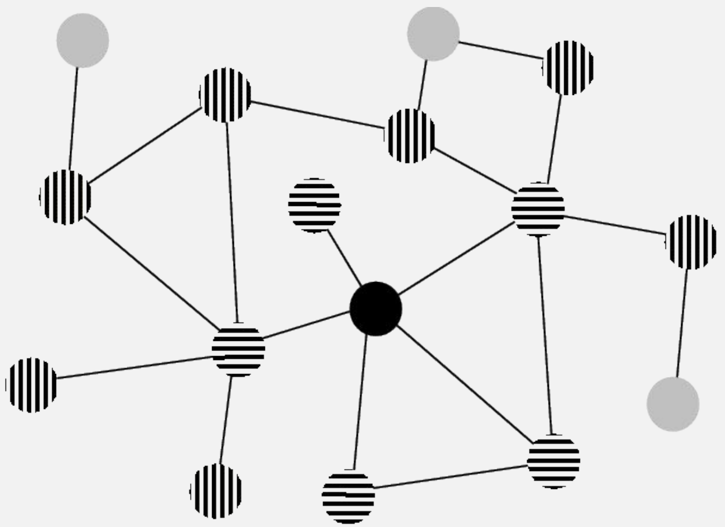
How does it work ?

Personalized PageRank

- Start from a seed node

- random walk from the seed node

⇒ Rank the node



Conductance

- measure of the quality of the community

- Add node to a community and stop when conductance is in a local minima

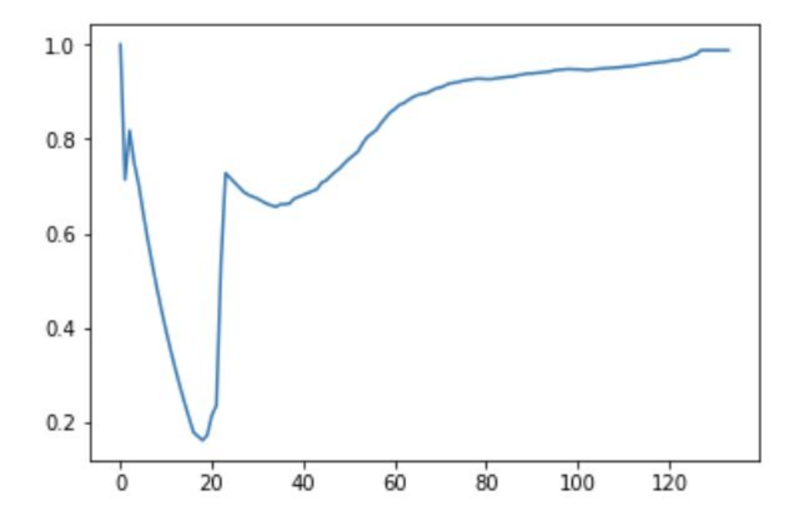
Outline

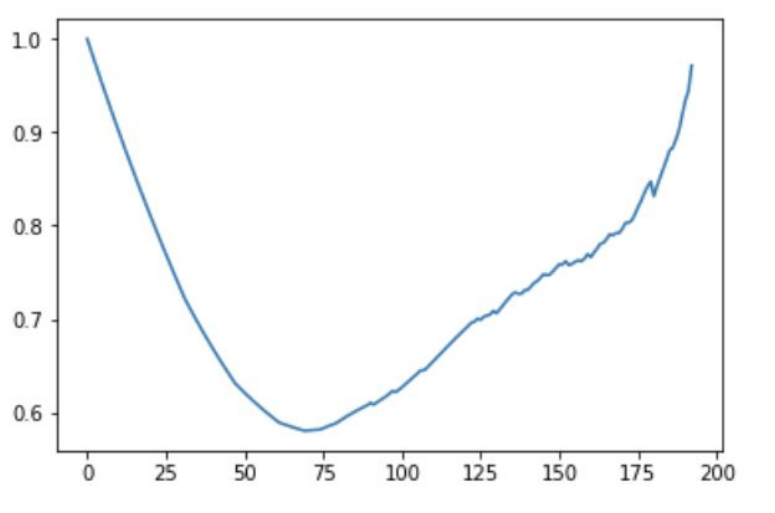
1. Introduction and motivation
2. Problem definition and Related work
3. Methodology
4. Evaluation – ***Assessing the results of our analysis***
5. Conclusion

Model Evaluation

- Evaluating the community of the seed node detected and the real one
- With **F1 Scoring**: $$FI = 2\frac{pr}{p+r} \text{ where } p = \frac{TP}{TP+FP}, \ r = \frac{TP}{TP+FN}$$

- Examples of the **evolution of the conductance of a seed node** depending on the number of neighbors taken: we assess with ground truth the “communities” that emerge from each minimum





Outline

1. Introduction and motivation
2. Problem definition and Related work
3. Methodology
4. Evaluation
5. Conclusion – ***What we learned from our study***

Conclusion

- Reasonable computational time for communities detection from seed nodes

  → impact for recommendation for big networks

- Work perspectives:
  - average different metrics (i.e. triad participation ratio)
  - combine with machine learning & reinforcement learning

THANK YOU !