Community Detection through Representation learning in Evolving Heterogenous Networks

By Egor Dmitriev (6100120)

EGOR DMITRIEV, Utrecht Univeristy, The Netherlands

Recent developments in big data and graph representation learning have allowed researchers to make breakthroughs in social network analysis and the identification of communities. While opening a lot of research opportunities, such approaches are highly limited to snapshots of rapidly evolving social networks. This, in fact, is a great simplification of the real-world situation which is always evolving and expanding by the user and/or machine interactions.

Relying on novel research of dynamic graph representation learning, the goal of my thesis project is to build a framework for community detection and representation in evolving heterogeneous networks. To verify the merit of the proposed framework, it will be evaluated against baselines on static heterogeneous graphs, and analyzed against gathered twitter dataset on covid measures.

Social Network Analysis (SNA) is a huge part of the Network Science field and is concerned with the process of investigating social structures that occur in real-world using Network and Graph Theory. These social structures usually include social media networks, economic transaction networks, knowledge networks and disease transmission networks. One main issue to address while studying this type of real-world events lies in identification of meaningful substructures hidden within the overall complex system. The SNA is therefore applied to extract patterns from the data usually in form of information flow, identification of high throughput nodes and paths, and discovery of communities and clusters. In this thesis we are going to focus on the problem of community discovery.

This thesis proposal is structured as follows: in this sections we are going to introduce basic concepts and challenges of Dynamic Community Detection. In section [] we will describe the problem we are trying to solve as well as formulate the research questions. In section [] a brief literature survey is conducted on identifying current state of the art and approaches to Community Detections. In section [] we will elaborate on our methodology for solving the posed problem and answering the research questions. Finally, in section [] the concrete planning for th research project is laid out.

1 COMMUNITY DETECTION

Problem of partitioning a complex network into *communities* which represent groups of individuals with high interaction density while individuals from different communities have comparatively low interaction density is known as Community Discovery (CD). CD is a task of fundamental importance within CD as it discloses deeper properties of networks. It provides insight into networks' internal structure and its organizational principles.

Many useful applications of CD have been studied by researchers including identification of criminal groups [@sarvariConstructingAnalyzingCriminal2014], social bot detection [@karatasReviewSocialBot2017], targeted marketing [@mosadeghUsingSocialNetwork2011], and public health / disease control [@salatheDynamicsControlDiseases2010].

With the explosion of human- and machine-generated data, often collected by social platforms, more datasets are emerging having rich temporal information that can be studied. CD operates only on static networks. Meaning that their temporal dimension is often omitted, which often

does not yield a good representation of the real-world where networks constantly evolve. Such networks are often referred to as dynamic networks as their components such as nodes and edges may appear and fade from existence. Accordingly community detection on such dynamic networks is called Dynamic Community Detection (DCD).

DCD algorithms, by incorporating additional temporal data are often able to both outperform their counterpart CD algorithms [@faniUserCommunityDetection2020], as well as providing additional information about communities for analysis [@pallaQuantifyingSocialGroup2007]. This additional information comes in form of community events such as (birth, growth, split, merging, and death) or in form of ability to track membership of certain individuals over time.

2 CHALLENGES IN COMMUNITY DETECTION

DCD is seen as the hardest problem within Social Network Analysis. Reason for this is mainly because DCD, unlike CD, also involves tracking the found communities over time. This tracking relies on consistency of the detected communities as usually slight changes to the network may cause a different community membership assignment.

Additionally, increasing richness of the data is not only limited to temporal data. The real-world data often connects entities of different modalities together. This multi-modality occurs through the fact that the entities and relations themselves may be of different kinds (meta topology-based features). For example users, topics and user-produced documents in a social network, or vehicles and landmarks in a traffic network. Another example of multi-modality in networks comes in form of node and relation features (content-based features). These features may come in form of structured (numerical, categorical or vector data) or unstructured data such as images and text. It is of high importance to explore this multi-modal data as it may not always be possible to explain the formation of communities using network structural information alone.

As noted earlier, meta topological features may be used to differentiate between different kind of nodes or edges to encode additional information.

Finally, a more common issue is that there is no common definition for a community structure. Within networks it is usually described in terms of membership assignment, while in more content-based settings communities are described in terms of distributions over topics which usually represent interest areas. The first definition only accounts for disjoint communities, while second is too vague as there may also be overlapping and hierarchical communities.

- RQ1. Does consideration of temporal evolution of meta topology-based and content-based features lead to higher consistency and quality communities within dynamic community detection?
- RQ2. Does a density based representation for communities the more general representation of temporal communities than the current state of the art (leading nodes or membership assignment)?
- RQ3. What deep representation learning architecture works best for twitter dynamic social network?
- Improve exisiting methods by
  - Incoporating metapath info
  - Metapath can replace topic modelling
  - RQ:
    * Is approach better than graph based
    * Is the representation more generalizable

The problem of dynamic community detection was noticed quite early on in within the SNA community and a considerable amount of research have been made in order to provide a comprehensive analysis of the network. While the said research was mostly focused on discovery of communities using topologically-based features and node connectivity, the covered methods did research the limitations and challenges posed by a temporal context.

In recent years, with the emergence of

3 DATASETS

4 EVALUATION

5 DYNAMIC COMMUNITY DETECTION METHODS

5.1 Classical Methods

5.2 Deep Methods

6 GRAPH REPRESENTATION LEARNING

- components
- problems with current solutions
- datasets [@rossettiCommunityDiscoveryDynamic2018]
- DCD is seen as the hardest problem within Social Network Analysis. Reason for this is mostly because DCD, unlike CD, also involves tracking the found communities over time which brings
- –