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. TODO: talk about appearance and disappearance of nodes

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 the recent years, significant developments have been made in the space of deep learning. Mainly in the development of new deep learning methods capable of learning graph-structured data [@bronsteinGeometricDeepLearning2017, @hamiltonRepresentationLearningGraphs2018, @kipfSemiSupervisedClassificationGraph2017] which is fundamental for SNA. Various of problems within the field have been revisited, including the community detection problems. The approaches have been expanded by incorporation of more complex features, solving the problems concerning multimodality and introduction of unsupervised learning.

Despite this resurgence, the DCD problem has received little attention. Though a few efforts have been made to incorporate the deep learning methods by introducing content-based similarity dynamic and usage of graph representation based CD algorithms within a temporal context, the current state of the art leaves a lot to be desired.

We structure this literature survey as follows: first we describe the various interpretations of community structure [], explore the current datasets and evaluation methods used for benchmarking of the current DCD methods []. Then, we dive in the current state of the art works on DCD by discussing both "classic" methods and novel deep learning based methods []. Finally, we discuss the current advances within the graph representation learning [] and community detection [] approaches.

3 COMMUNITY STRUCTURES

Communities in real-world networks can be of different kinds: disjoint (think of students belonging to different educational institutions), overlapping (person having membership in different social groups) and hierarchical (components of a car). One of the main reasons behind the complexity of CD is the unclear definition what a community actually is.

The classical methods intuitively describe communities as groups of nodes within a graph, such that the intra-group connections are denser than the intergroup ones. The more recent (usually deep learning) works define communities as a distribution over d-dimensional space which may incorporate both topological as well as content-based feature [@cavallariLearningCommunityEmbedding2017]. A more general definition is introduced in [@cosciaClassificationCommunityDiscovery2011] to create an underlying concept generalizing all variants found in the literature.

> Definition (Community). A community in a complex network is a set of entities that share some closely correlated sets of actions with the other entities of the community. A direct connection is considered as a particular and very important kind of action.

3.1 Dynamic Community

Similarly to how communities can be found in static networks, dynamic communities extends this definition by utilizing the temporal dimension to define its life cycle/evolution over a dynamic network. A dynamic community is defined as a collection of communities and set of transformations on thesis communities over time.

This persistence across time of communities subjected to progressive changes is an important problem to tackle. Since as noted by [@rossettiCommunityDiscoveryDynamic2018] it can be compared to the famous "the ship of Theseus" paradox. Because (verbatim), *deciding if an element*

composed of several entities at a given instant is the same or not as another one composed of some—or even none—of such entities at a later point in time is necessarily arbitrary and cannot be answered unambiguously.

Most of the works agree on two atomic transformations on the communities, including node/edge appearance and vanishing. While some such as [@pallaQuantifyingSocialGroup2007, @asurEventbasedFrameworkCharacterizing2009, @cazabetUsingDynamicCommunity2012] define a more extensive set of transformations (also referred to as events) which may be more interesting for analytical purposes:

- Birth, when a new community emerges at a given time.
- Death, when a community disappears. All nodes belonging to this community lose their membership.
- Growth, when a community acquires some new members (nodes).
- Contraction, when a community loses some of its members.
- Merging, when several communities merge to form a new community.
- Splitting, when a community is divided into several new ones.
- Resurgence, when a community disappears for a period and reappears.

These events / transformations are often not explicitly used during the definition and/or representation of dynamic communities. Nevertheless, most of the methods covered discussed in the following sections do defined a way in their algorithm to extract such event from the resulting data.

Finally, is important to note that dynamic networks can differ in representation. They can be represented as either a time-series of static networks (also referred to as snapshots), or as a real time stream of edges (referred to as temporal networks). Though, it should be noted that within the context of dynamic community detection they can be seen as equivalent as the conversion between the two representations can be done in a lossless way.

4 EVALUATION METHODS

As described in the previous section the definition for both community and dynamic community may be quite ambiguous. In this section we will cover various datasets which are used for the task of dynamic community detection and how detection and tracking results can be evaluated in a lesser ambiguous setting for comparison of various approaches. To disambiguate the process a little, during evaluation the resemblance/detection and matching/tracking tasks are evaluated separately.

4.1 Supervised Approaches

Evaluation of detected (dynamic) communities becomes much easier when the *ground truth communities* are provided. The evaluation is then done by comparing the difference between the produced communities and the effective ones. To perform this comparison, usually information theory based metric Normalized Mutual Information (NMI) is used which converts community sets to bit-strings and quantifies the "amount of information" can be obtained about one community by observing the other. [@lancichinettiDetectingOverlappingHierarchical2009]. Within the context of DCD these metrics are calculated for all snapshots and aggregated into one single measure.

Because in the real-world there are no planted communities, this approach is usually applied on synthetic datasets where the communities and their dynamicity is sampled from a distribution. Some approaches define ground truth communities using the metadata and node attributes which are present within the datasets which again allow for such supervised evaluation.

- TODO: should I talk about shortcomings of NMI and usage of NF1 by Rosetti et al ?

5 UNSUPERVISED APPROACHES

Another way to evaluate and compare different CD algorithms without knowing ground truth communities is using a quality function. Modularity is the most widely used measure [@newmanFastAlgorithmDetecting2004], since it measures the strength of division of a network into modules. Networks with high modularity have dense connections between the nodes within the modules, but sparse connections between nodes in different modules. Other measures are used as well including:

- Conductance: the percentage of edges that cross the cluster border
- Expansion: the number of edges that cross the community border
- Internal Density: the ratio of edges within the cluster with respect to all possible edges
- Cut Ratio and Normalized Cut: the fraction of all possible edges leaving the cluster
- Maximum/Average ODF: the maximum/average fraction of nodes' edges crossing the cluster border

5.1 Alternative Measures

In [@peelGroundTruthMetadata] the authors criticize these evaluation approaches by proving that they introduce severe theoretical and practical problems. For one, they prove the no free lunch therem for CD, ie. they prove that algorithmic biases that improve performance on one class of networks must reduce performance on others. Therefore, there can be no algorithm that is optimal for all possible community detection tasks as quality of communities may differ by the optimized metrics. Additionally, they demonstrate that when a CD algorithm fails, the poor performance is indistinguishable from any of the three alternative possibilities: (i) the metadata is irrelevant to the network structure, (ii) the metadata and communities capture different aspects of network structure, (iii) the network itself lacks structure. Therefore, which community is optimal should depend on it's subsequent use cases and not a single measure.

6 DATASETS

- Synthetic:
  - greene et al 2010
- Real world
  - Enron
  - Weibo
  - FB-wosn

7 DYNAMIC COMMUNITY DETECTION METHODS

7.1 Classical Methods

7.2 Deep Methods

8 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
-
  -