

# Community Detection through Representation learning in Evolving Heterogenous Networks

A Master's Thesis proposal

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.

## CONTENTS

|  |   |
|--|---|
| Abstract                                   | 1 |
| Contents                                   | 1 |
| 1 Introduction and Backgrounds             | 2 |
| 1.1 Community Detection                    | 2 |
| 1.2 Challenges in Community Detection      | 2 |
| 2 Research questions                       | 3 |
| 3 Literature review                        | 3 |
| 3.1 Community Structures                   | 4 |
| 3.1.1 Dynamic Community                    | 5 |
| 3.2 Evaluation methods                     | 5 |
| 3.2.1 Evaluation of Link-based methods     | 6 |
| Annotation Based                           | 6 |
| Metric Based                               | 6 |
| Alternative Measures                       | 6 |
| 3.2.2 Evaluation of Representation methods | 7 |
| 3.3 Datasets                               | 7 |
| 3.3.1 Synthetic                            | 7 |
| 3.3.2 Real World                           | 7 |
| 3.4 Dynamic Community Detection Methods    | 7 |
| 3.4.1 Classical Methods                    | 7 |
| 3.4.2 Deep Methods                         | 8 |
| 3.5 Graph Representation Learning          | 8 |
| 4 Approach                                 | 8 |
| 5 Planning                                 | 8 |
| References                                 | 8 |

## 1 INTRODUCTION AND BACKGROUNDS

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 2 we will describe the problem we are trying to solve as well as formulate the research questions. In section 3 a brief literature survey is conducted on identifying current state of the art and approaches to Community Detections. In section 4 we will elaborate on our methodology for solving the posed problem and answering the research questions. Finally, in section 5 the concrete planning for th research project is laid out.

### 1.1 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 [21], social bot detection [11], targeted marketing [14], and public health / disease control [20].

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 Fani et al. [7], as well as providing additional information about communities for analysis [16]. 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.

### 1.2 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, asymmetric edges, etc

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.

## 2 RESEARCH QUESTIONS

- ~ 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?
- WIP 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)?
- WIP RQ3. What deep representation learning architecture works best for twitter dynamic social network?
- Improve existing methods by
  - Incorporating metapath info
  - Metapath can replace topic modelling
  - RQ:
    - \* Is approach better than graph based
    - \* Is the representation more generalizable

## 3 LITERATURE REVIEW

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 [2, 10, 12] 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 multi-modality 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 review as follows: first we describe the various interpretations of the community structure 3.1, explore the current datasets and evaluation methods used for benchmarking of the current DCD methods 3.2. Then, we dive in the current state of the art works on DCD by discussing both “classic” methods and novel deep learning based methods section 3.3. Finally, we discuss the current advances within the graph representation learning [] and community detection [] approaches.

### 3.1 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 *link-based* (referred to as classic) community detection methods intuitively describe communities as groups of nodes within a graph, such that the intra-group connections are denser than the intergroup ones. This definition is primarily based on the *homophily* principle, which refers to the assumption that similar individuals are those that are densely connected together. Therefore, these kind of methods look for sub-graph structures such as cliques and components that identify connectedness within the graph structure to represent the communities.

Unfortunately, in most cases link-based methods fall short to identity communities of similar individuals. This is mainly due to two facts: (i) many similar individuals in a social network are not explicitly connected together, (ii) an explicit connection does not necessarily indicate similarity, but may explained by sociological processes such as conformity, friendship or kinship [6, 7].

Because of this, and with success of (deep learning) representation based methods, the more recent works define communities as a distribution over  $d$ -dimensional space which may span over both topological as well as content-based features [3]. Similarly, also hybrid methods are introduced which combine link-based approach with content representation learning methods [7].

A more general definition is introduced in [5] to create an underlying concept generalizing all variants found in the literature. In link-based methods, a direct connection is considered as a particular and very important kind of action, while newer methods also consider content or interest overlap.

*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.

- TODO: Content Based Methods: (similarity idea)

- Modeled communities **based on topics of interest** through a community-user-topic generative process
- **Communities** are **formed around multiple correlated topics** where each topic can be reused in several different communities
- User can be a member of different communities but with varying degrees of membership
- **Distributional semantics** states that words that occur in similar contexts are semantically similar
- **Recommendation Point of view**

**3.1.1 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 these communities over time.

This persistence across time of communities subjected to progressive changes is an important problem to tackle. Since as noted by [18] 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 [1, 16, Cazabet et al. [4]] 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.

## 3.2 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 how detection and tracking results can be evaluated in a lesser ambiguous setting to compare various approaches.

To disambiguate the process a little, during evaluation, the resemblance/detection and matching/tracking tasks are evaluated separately.

### 3.2.1 Evaluation of Link-based methods.

*Annotation Based.* 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, 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 [13].

A possible drawback of this measure is that its complexity is quadratic in terms of identified communities. In [19] alternative measure (NF1) with linear complexity is introduced which similarly to F1 score uses the trade-off between precision and recall (of the average of harmonic means) of the matched communities. In the follow-up work [17] the authors describe a way to apply this measure within the context of DCD by calculating this score for all the snapshots and aggregating the results into one single measure.

In real-world there are usually no ground truth communities. Therefore this approach is usually applied on synthetic datasets where the communities and their dynamicity is sampled from a distribution. Alternative approach some papers take is by defining ground truth communities using the metadata and node attributes present within the datasets. Some datasets may include annotated communities, but this is not common within DCD datasets.

*Metric Based.* 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 [15], 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, and 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

*Alternative Measures.* In [?] the authors criticize these evaluation approaches by proving that they introduce severe theoretical and practical problems. For one, they prove the no free lunch theorem 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.

- Todo: lead into “subsequent” use cases and representation based methods

3.2.2 Evaluation of Representation methods.

- Explore Recommendation / Link prediction task

–

3.3 Datasets

3.3.1 Synthetic.

| Paper                | Description  |
|----------------------|--|
| Greene et al. [9]    | Generate Graphs based on Modularity measure  |
| Granell et al. [8]   |  |
| Hamilton et al. [10] |  |
|                      | Generate Time dependent Heterogeneous graphs using modularity optimization and multi-dependency sampling |

3.3.2 Real World.

| Dataset               | Description  |
|-----------------------|--|
| Enron                 | Includes: Persons, Email Categories, Sentiment, Email Content  |
| KIT (dead)            | Includes: Persons, Tweets, Followers;<br><b>Excludes: Tweet Content</b>                                  |
| Weibo                 |  |
| Digg                  | Includes: Persons, Stores, Followers, Votes;<br><b>Excludes: Content</b>                                 |
| Slashdot              | Includes: Persons, Votes; <b>Excludes: Content</b><br>Actor movie network; Content is implicitly defined |
| IMDB                  |  |
| WIKI-RFA              | Network of Voters and Votees. Links are votes and vote comments  |
| FB-wosn               | User friendship links and User posts on users walls; <b>Excludes: Content</b>                            |
| TweetUM (dead)        | Twitter Tweets, User Profiles and Followers; Includes: Content   |
| Reddit Pushift        | User Submissions and Posts on Subreddits; With timestamps  |
| Bitcoin Trust Network | Network Nodes and peer Ratings; With timestamps  |
| LastFM1k              | User - Song Listen histories; With timestamps  |
| MovieLens25M          | Users and Movie Ratings; With timestamps   |

3.4 Dynamic Community Detection Methods

3.4.1 Classical Methods.

### 3.4.2 Deep Methods.

## 3.5 Graph Representation Learning

- components
- problems with current solutions
- datasets [18]
- 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
- 

## 4 APPROACH

## 5 PLANNING

abc

## REFERENCES

- [1] Sitaram Asur, Srinivasan Parthasarathy, and Duygu Ucar. An event-based framework for characterizing the evolutionary behavior of interaction graphs. *ACM Transactions on Knowledge Discovery from Data*, 3(4):16:1–16:36, December 2009. ISSN 1556-4681. doi: 10.1145/1631162.1631164.
- [2] Michael M. Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: Going beyond Euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, July 2017. ISSN 1053-5888, 1558-0792. doi: 10.1109/MSP.2017.2693418.
- [3] Sandro Cavallari, Vincent W. Zheng, Hongyun Cai, Kevin Chen-Chuan Chang, and Erik Cambria. Learning Community Embedding with Community Detection and Node Embedding on Graphs. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 377–386, Singapore Singapore, November 2017. ACM. ISBN 978-1-4503-4918-5. doi: 10.1145/3132847.3132925.
- [4] Rémy Cazabet, Hideaki Takeda, Masahiro Hamasaki, and F. Amblard. Using dynamic community detection to identify trends in user-generated content. *Social Network Analysis and Mining*, 2012. doi: 10.1007/s13278-012-0074-8.
- [5] Michele Coscia, Fosca Giannotti, and Dino Pedreschi. A Classification for Community Discovery Methods in Complex Networks. *Statistical Analysis and Data Mining*, 4(5):512–546, October 2011. ISSN 19321864. doi: 10.1002/sam.10133.
- [6] Chris Diehl, Galileo Namata, and Lise Getoor. Relationship Identification for Social Network Discovery. pages 546–552, January 2007.
- [7] Hossein Fani, Eric Jiang, Ebrahim Bagheri, Feras Al-Obeidat, Weichang Du, and Mehdi Kargar. User community detection via embedding of social network structure and temporal content. *Information Processing & Management*, 57(2):102056, March 2020. ISSN 03064573. doi: 10.1016/j.ipm.2019.102056.
- [8] Clara Granell, Richard K. Darst, Alex Arenas, Santo Fortunato, and Sergio Gómez. Benchmark model to assess community structure in evolving networks. *Physical Review E*, 92(1):012805, July 2015. ISSN 1539-3755, 1550-2376. doi: 10.1103/PhysRevE.92.012805.
- [9] Derek Greene, Dónal Doyle, and Pádraig Cunningham. Tracking the Evolution of Communities in Dynamic Social Networks. In *2010 International Conference on Advances in Social Networks Analysis and Mining*, pages 176–183, August 2010. doi: 10.1109/ASONAM.2010.17.
- [10] William L. Hamilton, Rex Ying, and Jure Leskovec. Representation Learning on Graphs: Methods and Applications. *arXiv:1709.05584 [cs]*, April 2018.
- [11] Arzum Karataş and Serap Şahin. A Review on Social Bot Detection Techniques and Research Directions. October 2017.
- [12] Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv:1609.02907 [cs, stat]*, February 2017.



- [13] Andrea Lancichinetti, Santo Fortunato, and Janos Kertesz. Detecting the overlapping and hierarchical community structure of complex networks. *New Journal of Physics*, 11(3):033015, March 2009. ISSN 1367-2630. doi: 10.1088/1367-2630/11/3/033015.
- [14] Mohammad Mosadegh and Mehdi Behboudi. Using Social Network Paradigm for Developing a Conceptual Framework in CRM. *Australian Journal of Business and Management Research*, 1:63–71, August 2011. doi: 10.52283/NSWRCA.AJBMR.20110104A06.
- [15] M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69(6):066133, June 2004. ISSN 1539-3755, 1550-2376. doi: 10.1103/PhysRevE.69.066133.
- [16] Gergely Palla, Albert-László Barabási, and Tamás Vicsek. Quantifying social group evolution. *Nature*, 446(7136):664–667, April 2007. ISSN 1476-4687. doi: 10.1038/nature05670.
- [17] Giulio Rossetti. ANGEL: Efficient, and effective, node-centric community discovery in static and dynamic networks. *Applied Network Science*, 5(1):26, June 2020. ISSN 2364-8228. doi: 10.1007/s41109-020-00270-6.
- [18] Giulio Rossetti and Rémy Cazabet. Community Discovery in Dynamic Networks: A Survey. *ACM Computing Surveys*, 51(2):35:1–35:37, February 2018. ISSN 0360-0300. doi: 10.1145/3172867.
- [19] Giulio Rossetti, Luca Pappalardo, and Salvatore Rinzivillo. A Novel Approach to Evaluate Community Detection Algorithms on Ground Truth. In Hocine Cherifi, Bruno Gonçalves, Ronaldo Menezes, and Roberta Sinatra, editors, *Complex Networks VII: Proceedings of the 7th Workshop on Complex Networks CompleNet 2016*, Studies in Computational Intelligence, pages 133–144. Springer International Publishing, Cham, 2016. ISBN 978-3-319-30569-1. doi: 10.1007/978-3-319-30569-1\_10.
- [20] Marcel Salathé and James H. Jones. Dynamics and Control of Diseases in Networks with Community Structure. *PLOS Computational Biology*, 6(4):e1000736, April 2010. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1000736.
- [21] Hamed Sarvari, Ehab Abozinadah, Alex Mbaziira, and Damon Mccoy. Constructing and Analyzing Criminal Networks. In *2014 IEEE Security and Privacy Workshops*, pages 84–91, May 2014. doi: 10.1109/SPW.2014.22.