Community Detection through Representation learning in Evolving Heterogenous Networks

A Master's Thesis proposal

EGOR DMITRIEV, Utrecht University, 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.

CONTENTS

Author's address: Egor Dmitriev, e.dmitriev@students.uu.nl, Utrecht University, The Netherlands.

1 INTRODUCTION AND BACKGROUND

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 the 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 the 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 section, we are going to introduce basic concepts and challenges of Dynamic Community Detection. In section 2 a brief literature survey is conducted on identifying the current state of the art and approaches to Dynamic Community Detection. In section 3 we will describe the problem we are trying to solve as well as formulate the research questions. In section 4 we will elaborate on our proposed methodology for solving the posed problem and answering the research questions. Finally, in **??** the concrete planning for the research project is laid out.

1.1 Community Detection

The 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 SNA 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 [35], social bot detection [16], targeted marketing [25], and public health/disease control [34].

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 incorporate additional temporal data are often able to both outperform their counterpart CD algorithms [9, 11, 23, 31], as well as provide additional information about communities for analysis [27]. This additional information comes in form of community events such as (birth, growth, split, merging, and death) or in form of the ability to track the membership of certain individuals over time.

1.2 Challenges in Community Detection

DCD is seen as the hardest problem within Social Network Analysis. The reason for this is mainly because DCD, unlike CD, also involves tracking the found communities over time. This tracking relies on the consistency of the detected communities, as usually slight changes to the network may cause a different community membership assignment. Not properly accounting for this uncertainty may cause community and result drift [6].

Additionally, the increasing richness of the data is not only limited to temporal data. The real-world data often connects entities of different modalities. This multi-modality occurs through the fact that the entities and relations themselves may be of different types (meta topology-based features). For example users, topics, and documents in a social network (or vehicles and landmarks in a traffic network). More complex networks may include asymmetric relationships, and temporal networks may include appearing, disappearing, or streaming edges/nodes.

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.

Finally, a more systematic 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 modeled topics (that usually represent interest areas) or distributions over latent similarity space. Both definitions have their shortcomings as they often fail to account for more complex community structures (such as overlapping and hierarchical communities) and non-linearity of structures often found in the real world.

Task of community detection is often compared to clustering and graph clustering, which not always may be a fair comparison as a main focus point in many CD algorithms is the fact that the amount of communities is unknown a priori. Communities are not are never planted in the real world and the algorithms should detect them in an unsupervised manner.

2 LITERATURE REVIEW

The problem of dynamic community detection was noticed quite early on within the SNA community and a considerable amount of research has been made in order to provide a comprehensive analysis of the network. While the said research was mostly focused on the 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, 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 [3, 15, 18] which is fundamental for SNA. Because of this, various problems within the field have been revisited, including community detection problems. The approaches have been expanded by incorporation of more complex features, solving the problems concerning multi-modality, and the 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, the definition of unified constraints for end-to-end learning, 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 the literature as follows: first, we describe the various interpretations of the Community Structure in section 2.1. Next, we explore various approaches and techniques related to Graph Representation Learning in section 2.2. Then, we provide an overview of the current state-of-the-art approaches for Community Detection and Dynamic Community Detection tasks in section 2.3 and section 2.4. Finally, we discuss the ways to evaluate the said algorithms in section 2.5 and the datasets available in section 2.6.

2.1 Community Structures

The goal of this section is to introduce fundamental structures for the Dynamic Community Detection task. We do this by combining various definitions used in the relevant literature as well as establishing the purpose for these structures, before proceeding into approaches for detecting communities in the following sections.

2.1.1 Communities

Communities in real-world networks can be of different kinds: disjoint (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 that there is not one unique definition what a community actually is.

The *link-based* (also 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 inter-group 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 identify 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 [7, 9].

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.

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.

2.1.2 Dynamic Communities

Similar to how communities can be found in static networks, dynamic communities extend this definition by utilizing the temporal dimension to define their life cycle/evolution over a dynamic network. A dynamic community is characterized by a collection of communities and a set of transformations on these communities over time.

This persistence of communities across time subjected to progressive changes is an important problem to tackle. Though, as noted by [32] the problem 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, 27, 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 in the following sections do define a way in their algorithm to extract such events from the resulting data.

Finally, it 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). Within the global 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. The latter, temporal networks are often used to handle incremental changes to the graph and are most commonly applied within real-time community detection settings.

2.2 Graph Representation Learning

The representation-based approaches stem from the field of computational linguistics which relies heavily on the notion of *distributional semantics* stating that words occurring in similar contexts are semantically similar. Therefore the word representations are learned as dense low-dimensional representation vectors (embeddings) of a word in a latent similarity space by predicting words based on their context or vice versa [24, 29]. Using the learned representations similarity, clustering and other analytical metrics can be computed.

The success of these representation learning approaches has spread much farther than just linguistics as similar ideas are also applied to other fields including graph representation learning. Methods such as deepwalk [30], LINE [36], and node2vec [13] use random walks to sample the neighborhood/context in a graph (analogous to sentences in linguistic methods) and output vector representations (embeddings) that maximize the likelihood of preserving the topological structure of the nodes within the graph.

Whereas previously the structural information features of graph entities had to be hand-engineered, these new approaches are data-driven, save a lot of time labeling the data, and yield superior feature/representation vectors. The methods can be trained to optimize for *homophily* on label prediction or in an unsupervised manner on link prediction tasks.

Newer approaches introduce the possibility for the fusion of different data types. GraphSAGE [14] and Author2Vec [37] introduce a methodology to use node and edge features during the representation learning process. Other approaches explore ways to leverage heterogeneous information present within the network by using *metapath* based random walks (path defined by a series of node/link types) [8] or by representing and learning relations as translations within the embedding space [2]. In Nguyen et al. [26] the authors introduce a way to encode temporal information by adding chronological order constraints to various random walk algorithms. Other relevant advancements within the field include Graph Convolutional Networks (GCN) [19] and (Variational) Graph Auto-Encoders (GAE) [17] which present more effective ways to summarize and represent larger topological neighborhoods or whole networks.

2.3 Link-based Approaches

2.3.1 Community Detection

Modularity.

Louvain Method.

Label Propagation algorithm.

2.3.2 Dynamic Community Detection

Independent Community Detection and Matching.

Dependent Community Detection.

Simultaneous community detection.

Dynamic Community Detection on Temporal Networks (Evolution).

2.4 Representation-based Approaches

2.4.1 Community Detection

Affiliation Graph Networks.

2.4.2 Dynamic Community Detection

2.5 Evaluation

As described in the previous sections, 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.

2.5.1 Annotated

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 [22].

A possible drawback of this measure is that its complexity is quadratic in terms of identified communities. In [33] 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 [31] 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.

2.5.2 Metric 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 [22].

A possible drawback of this measure is that its complexity is quadratic in terms of identified communities. In [33] 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 [31] 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.

2.5.3 Task specific

In [28] 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.

2.6 Datasets

2.6.1 Synthetic Datasets

Paper	Description
Lancichinetti et al. [21]	Static networks (widely used)
Greene et al. [12]	Generate Graphs based on Modularity measure
Granell et al. [11]	
Hamilton et al. [15]	Generate Time dependent Heterogeneous graphs using modularity optimization and multi-dependency sampling
SYN - Ghalebi et al. [10]	
SBM - Lancichinetti and Fortunato [20]	extracted from the dynamic Stochastic Block Model

2.6.2 Real World Datasets

Dataset	Description
Enron	Includes: Persons, Email Categories, Sentiment, Email Content
KIT (dead)	
Weibo	Includes: Persons, Tweets, Followers; **Excludes: Tweet Content**
Digg	Includes: Persons, Stores, Followers, Votes; **Excludes: Content**
Slashdot	Includes: Persons, Votes; **Excludes: Content**
IMDB	Actor movie network; Content is implicitly defined
WIKI-RFA	Wikipedia Adminitrator Election; Network of Voters and Votees. Links are votes and vote comments
FB-wosn	User friendship links and User posts on users walls; **Excludes: Content**
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
Memetracker	
Rumor Detection	Rumor Detection over Varying Time Windows; Twitter data; With timestamps

3 RESEARCH QUESTIONS

4 APPROACH

PLANNING

[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] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating Embeddings for Modeling Multi-relational Data. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[3] 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.

[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] Narimene Dakiche, Fatima Benbouzid-Si Tayeb, Yahya Slimani, and Karima Benatchba. Tracking community evolution in social networks: A survey. *Information Processing & Management*, 56(3): 1084–1102, May 2019. ISSN 0306-4573. doi: 10.1016/j.ipm.2018.03.005.

[7] Chris Diehl, Galileo Namata, and Lise Getoor. Relationship Identification for Social Network Discovery. pages 546–552, January 2007.

[8] Yuxiao Dong, Nitesh V. Chawla, and Ananthram Swami. Metapath2vec: Scalable Representation Learning for Heterogeneous Networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pages 135–144, New York, NY, USA, August 2017. Association for Computing Machinery. ISBN 978-1-4503-4887-4. doi: 10.1145/3097983.3098036.

[9] 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.

[10] Elahe Ghalebi, Baharan Mirzasoleiman, Radu Grosu, and Jure Leskovec. Dynamic Network Model from Partial Observations. *arXiv:1805.10616 [cs, stat]*, February 2019.

[11] 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.

[12] 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.

[13] Aditya Grover and Jure Leskovec. Node2vec: Scalable Feature Learning for Networks. *arXiv:1607.00653 [cs, stat]*, July 2016.

[14] William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive Representation Learning on Large Graphs. *arXiv:1706.02216 [cs, stat]*, September 2018.

[15] William L. Hamilton, Rex Ying, and Jure Leskovec. Representation Learning on Graphs: Methods and Applications. *arXiv:1709.05584 [cs]*, April 2018.

[16] Arzum Karataş and Serap Şahin. A Review on Social Bot Detection Techniques and Research Directions. October 2017.

[17] Thomas N. Kipf and Max Welling. Variational Graph Auto-Encoders. *arXiv:1611.07308 [cs, stat]*, November 2016.

[18] Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv:1609.02907 [cs, stat]*, February 2017.

[19] Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv:1609.02907 [cs, stat]*, February 2017.

[20] Andrea Lancichinetti and Santo Fortunato. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. *Physical Review E*, 80(1):016118, July 2009. ISSN 1539-3755, 1550-2376. doi: 10.1103/PhysRevE.80.016118.

[21] Andrea Lancichinetti, Santo Fortunato, and Filippo Radicchi. Benchmark graphs for testing community detection algorithms. *Physical Review E*, 78(4):046110, October 2008. ISSN 1539-3755, 1550-2376. doi: 10.1103/PhysRevE.78.046110.

[22] 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.

[23] Xin Liu, Weichu Liu, Tsuyoshi Murata, and Ken Wakita. Community Detection in Multi-Partite Multi-Relational Networks Based on Information Compression. *New Generation Computing*, 34(1): 153–176, March 2016. ISSN 1882-7055. doi: 10.1007/s00354-016-0206-1.

[24] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*, September 2013.

[25] 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.

[26] Giang Hoang Nguyen, John Boaz Lee, Ryan A. Rossi, Nesreen K. Ahmed, Eunyee Koh, and Sungchul Kim. Continuous-Time Dynamic Network Embeddings. In *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18*, pages 969–976, Lyon, France, 2018. ACM Press. ISBN 978-1-4503-5640-4. doi: 10.1145/3184558.3191526.

[27] 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.

[28] Leto Peel, Daniel B. Larremore, and Aaron Clauset. The ground truth about metadata and community detection in networks. *Science Advances*, 3(5):e1602548, 2017. doi: 10.1126/sciadv.1602548.

[29] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global Vectors for Word Representation. In *EMNLP*, volume 14, pages 1532–1543, January 2014. doi: 10.3115/v1/D14-1162.

[30] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. DeepWalk: Online Learning of Social Representations. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710, August 2014. doi: 10.1145/2623330.2623732.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. LINE: Large-scale Information Network Embedding. *Proceedings of the 24th International Conference on World Wide Web*, pages 1067–1077, May 2015. doi: 10.1145/2736277.2741093.

[37] Xiaodong Wu, Weizhe Lin, Zhilin Wang, and Elena Rastorgueva. Author2Vec: A Framework for Generating User Embedding. *arXiv:2003.11627 [cs, stat]*, March 2020.