A Review on Predicting Evolution of Communities

A. KARATAŞ[1] and S. ŞAHİN[1]

[1] İzmir Institute of Technology, İzmir/Turkey, arzum.karatas@iyte.edu.tr
[1] İzmir Institute of Technology, İzmir/Turkey, serapsahin@iyte.edu.tr

***Abstract* – In recent years, research on dynamic networks has increased as the availability of data has grown tremendously. Understanding the dynamic behavior of networks can be studied at the mezzo-scale (e.g., at the community level), as communities are the most informative structure in nonrandom networks and also evolve over time. Tracking the evolution of communities can provide evolution patterns to predict their future development. For example, a community may either grow into a larger community, remain stable, shrink into a smaller community, split into several smaller communities, or merge with another community. Predicting these evolutions is one of the most difficult problems in social networks. Better predictions of community evolution can provide useful information for decision support systems, especially for group-level tasks. So far, this problem has been studied by some researchers. However, there is a lack of a survey/review of existing work. This has prompted us to conduct this study. In this paper, we first categorize the existing works according to their methodological principles. Then, we focus on the works that use machine learning classifiers for prediction in this decade as they are in majority. We then highlight open problems for future research. In this way, this paper provides an up-to-date overview and a quick start for researchers and developers in the field of community evolution prediction.**

***Keywords* – Community, Evolving Communities, Predicting Evolution of Communities.**

I. INTRODUCTION

With the advances in computerization and technology in dynamic networks such as social networks, mobile networks, collaborative networks, etc., huge amounts of data have been created. Thus, the availability of data has increased tremendously, consequently the research on dynamic networks has also increased. Understanding the dynamic behavior of networks can be elaborated at the community (mezzo-scale) level, as communities are the most meaningful structure in nonrandom networks.

A community is a subgroup of at least three members that are more closely connected than the rest of the network. Communities in a dynamic network evolve over time. Therefore, the community may go through some evolutionary events. For example, a community may be stable, become larger or smaller in terms of the size of its members, split into several communities, or merge into a new community. Tracking the evolution of communities means observing the evolutionary behavior of communities over a period of time. Predicting community evolution is the task of predicting the most likely evolutionary event for communities based on their history of tracked communities. Tracking and predicting these evolutionary events can provide valuable information for decision support systems, especially for group-level tasks. For example, tracking and predicting the evolution of groups of criminals can be very useful in criminology.

In the literature, many researchers have addressed the prediction of community evolution. Unfortunately, there is no work that classifies these works. However, such a classification would both organize the diversity of existing work and promote development in the field. This motivated us to write this paper.

The main contributions of this paper are (i) a categorization of existing community evolution prediction methods according to their techniques, (ii) an overview of supervised learning based methods due to the majority in this field, (iii) highlighting the open research areas for future researchers.

The rest of this paper is organized as follows. Section II introduces the basic concepts of community evolution analysis and the problem of community evolution prediction. Our proposed classification of existing methods for predicting community evolution is presented in Section III. Then, a recent overview of supervised learning based methods is presented in Section IV. Then, open problems for supervised learning based methods are presented in Section V to motivate potential researchers. Finally, the paper concludes with a brief discussion of current research directions in Section VI.

II. PRELIMINARIES

A. Basic Concepts

Nonrandom networks contain an inherent community structure. The widely accepted definition of a community is that a subgroup within the network is strongly connected to each other and has a loose connection to the rest of the network. Communities in the network may be overlapping (multiple communities may have members in common) or disjoint (communities have no members in common). Detecting the hidden communities in the network is called community discovery. In the literature, there are many methods for detecting disjoint or overlapping communities, such as Louvain [1], Leiden [2], and CPM [3].

When networks are dynamic, communities exhibit evolutionary behavior over time. The possible evolutionary events for an existing community are "grow", "shrink", "continue", "merge", "split", and "dissolve". That is, a

community becomes larger/smaller/stable in terms of its membership, it may merge into a new community, split into smaller communities, or disappear. Observing the evolutionary behavior of communities in a given time interval is called tracking the evolution of communities. First, a dynamic network is represented as a series of snapshots. Then, a two-step methodology (e.g., (i) independent detection of communities in each snapshot and (ii) matching of detected communities) is usually applied for tracking. Some communities appear in each snapshot (e.g., consecutive evolution), while others do not (e.g., nonconsecutive evolution). Figure 1 illustrates the types of community evolution, with communities represented by circles. In the figure, the community in the blue circles evolves consecutively, while the community in the purple circles evolve nonconsecutively.

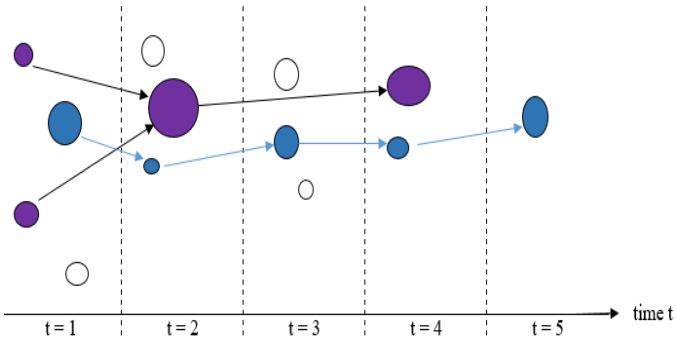

Figure 1: Illustration of evolution types of communities

B. *Problem Definition*

Let $G_t = (V_t, E_t)$ be a static graph representing a snapshot of a static network at time t, where V_t represents the vertices of the network and E_t represents the set of edges, and let D be a dynamic network represented by a time-ordered sequence of static networks such as $D = \{G_1, G_2, \cdots, G_t\}$ where t is the total number of static networks.

We define a partition $\{C_t^1, C_t^2, \cdots, C_t^k\}$ representing the discovered communities on each G_t, where a community is a subset of densely connected members instead of the rest of the G_t using an existing community detection method.

For each community C, a sequence of communities reflecting evolution over time is discovered at each time step using an existing method for tracking community evolution. This task requires that the communities in ascending time steps be matched to represent the evolution of the communities. Therefore, matching communities must be similar in terms of their nodes. Jaccard similarity (the ratio of the number of common members to the number of total members of two compared communities) is most commonly used to determine their similarity.

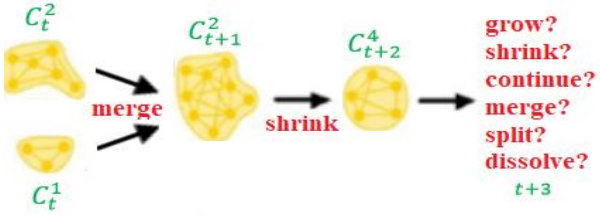

Figure 2: Predicting evolution of communities

Figure 2 illustrates the process of predicting community evolution. In the figure, C^1 and C^2 merge at time step *t* to C^2 at time step *t+1*, and C^2 has lost some of its members at time step *t+1* and shrinks consecutively at time step *t+2*. The prediction process takes this sequence and generates the most likely outcome as an evolution event (e.g., shrink, grow, merge, split, continue or dissolve).

The problem is informally defined as predicting future evolution events for matching communities based on their alignments over time.

III. CLASSIFICATION OF THE EXISTING PREDICTION METHODS FOR COMMUNITY EVOLUTION

There is no classification of predictive methods for community evolution in the literature. However, such a classification is helpful and organizes the variety of existing methods in this field. Thus, such a classification helps the developments in this field. Therefore, we propose a classification of existing prediction methods for community evolution according to the techniques used. There are three main classes, which are schematically shown in Figure 3:

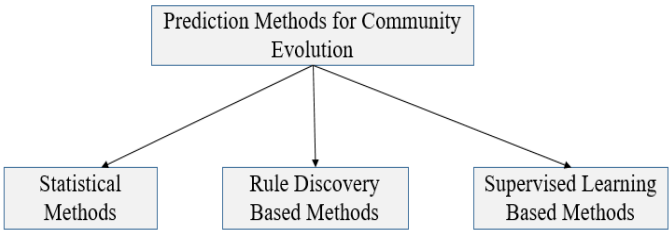

Figure 3: Classification chart of the existing prediction methods for community evolution

- Statistical Methods
- Rule Discovery Based Methods
- Supervised Learning Methods

Statistical methods provide a formal approach to modeling and predicting the evolution events of communities based on their history (e.g., aligned/matching communities) [4]. Rule discovery based methods first look for patterns in the time series representation of communities, then create rules for evolutionary events based on these patterns, and then make predictions based on these rules [5]. Supervised learning based methods first determine features to represent communities and then train supervised classifiers based on these features and the evolutionary history of the communities.

IV. SUPERVISED LEARNING BASED PREDICTION METHODS

The most common strategy for predicting community evolution is to use supervised classifiers. Methods using this strategy follow a two-step methodology: (i) analyzing the

evolution of communities and (ii) applying supervised classifiers based on selected features of communities for predetermined evolution events. Work in the literature over the last decade (between 2011 and 2021) is reviewed. Related work is summarized below.

In their work, Brodka et al [6] present and evaluate a supervised learning method for predicting the evolution of communities with respect to six events of community evolution such as growing, shrinking, continuing, merging, splitting and dissolving. They use the Group Evolution Detection method (GED) [7] to detect events between successive time steps and construct event sequences to describe the evolution of a given community. Each event sequence consists of the member sizes and events of all three previous communities. These sequences serve as input to the classifiers such as Naïve Bayes, Decision Tree, Random Forest and others provided by the data mining software WEKA [8] to predict the next event for a given community.

İlhan and Öğüdücü [9] propose a new approach to predict the next event of a community using a time series ARIMA model. In their study, community events are predicted by predicting community characteristics. The feature values are used to classify the possible events. The authors [10] propose a framework for detection of most prominent subset of community features to predict evolution of communities. They assert that their framework requires extraction of minimal number of community features.

Takaffoli et al. [11] use a two-step technique to predict the near future of a community through supervised learning. In this technique, they first decide whether the community survives, and then make the prediction whether the community survives. They diversify the type of features by using not only structural features of the community, but also features of influential members, temporal changes in features, contextual attributes, and features of past events. They consider only evolutionary sequences that have only two lengths.

Saganowski et al. [12] present two methods for predicting the following evolution event of a community. The first method uses the Stable Group Changes Identification (SGCI) method [13] and the other uses the GED method [7]. They use the CPM method [3] for community detection. The authors use evolution chain lengths, group features (e.g., size, density, leadership, etc.), node features (e.g., total degree, in-degree, etc.), and group aggregation (e.g., sum, average, minimum, and maximum). They then perform feature selection using ordinary (J48 and Random Forest) and ensemble classifiers (AdaBoost and Bagging). They conclude that longer group history leads to better prediction and the most recent group history has the largest impact on the next community change.

Diakidis et al. [14] address the problem of predictability of community evolution as a task of supervised learning. However, they predict four events of community evolution, namely continuation, shrinking, growth and dissolution. They use both sequential (e.g., Conditional Random Fields with Linear Chain and with Skip Chain) and ordinary classifiers (e.g., Naïve Bayes, Bayes Net, Logistic Regression, SVM, etc.) for the prediction task and compare the performance of the classifiers. These classifiers were trained on structural (e.g., size, density, etc.), content (topic diversity with TF-IDF), and contextual features (e.g., number of hashtags, size of tweets, and number of tweets with promotional URLs, etc.), as well as the previous state of a community as features for Twitter. They conclude that the sequential features are better than the ordinary ones because they also capture the past information first.

Pavlopoulou et al [15] present a framework for predicting community evolution. They study how past evolutions of a community affect the prediction of four evolution events such as growth, continuation, shrink, and dissolution. They use some structural (e.g. density, cohesion, diameter, etc.) and temporal features (e.g. lifespan, aging, join nodes ratio and left nodes ratio, etc.) to predict through supervised learning. They also specify the number of ancestors to be used for computing the temporal features, e.g., two or four ancestors according to their dataset from Mathematics Stack Exchange. They used the GED method [4] to track the evolution of the community, and Support Vector Machine (SVM) with RBF kernel (an exponential kernel) as a classifier for predicting the next evolution event. However, they did not consider merge and split events.

Dakiche et al [16] proposed a method for predicting the evolution of communities by capturing the interdependence of the rates of change of characteristics describing a community over time instead of the actual values. They considered only the rates of change in the structural and characteristic traits of influential members of a community. They examined the length of evolution sequences and concluded that the length of the sequences directly affects the amount and quality of information obtained. However, the quality of information may decrease with long sequences.

Dakiche et al. [17] propose a new framework for studying the distribution of activities over time to enable proper partitioning of the network. They claim that a properly partitioned network enables more accurate prediction of community events. After applying their novel network partitioning method, they proceed with a simple prediction method. That is, they apply the method GED [7] to detect group evolutions. Then they proceed to the prediction part. For this task, they specify characteristics. In their study, structural (e.g., density, cohesion, size ratio, etc.) and influential member characteristics (e.g., average leadership degree, average leadership closeness, and average leadership eigenvector) are used. Later, well-known supervised learning classifiers such as J48, Random Forest, Bagging and SVM were used.

Table 1 summarizes related work, with some important criteria listed in the first column. For tracking, they mainly use GED [7], SCGI [13] and some special methods developed by them. In the *Prediction Manner* row, the studies make predictions for consecutively or nonconsecutively evolving networks or both, where *CE* is for consecutively evolving communities and *NE* is for nonconsecutively evolving communities. Only the ML model of Takaffoli et al. [11] can predict the next stage of a community either at the next time step or at later time steps. While the method of Brodka et al. [6], Saganowski et al. [12], the method of İlhan and Ögüdücü [9], and the two methods of Dakiche et al. [16, 17] makes

prediction for all possible events of community evolution, others cannot characterize all events for prediction. For software attributes, Weka [8] is used for developing, training and testing ML models, CFinder is used for applying CPM and MODEC for community tracking, and CRFSuite is used for sequential classifiers such as Conditional Random Fields (CRF) [18].

Table 1: Related works using supervised learning in last decade

	Related Work							
Attributes	Brodka et al. [6]	İlhan& Öğüdücü [9]	Takaffoli et al. [11]	Saganowski et al. [12]	Diakidis et al. [14]	Pavlopoulou et al. [15]	Dakiche et al. [16]	Dakiche et al. [17]
Year	2012	2013	2014	2015	2015	2017	2019	2021
Tracking Method	GED [7]	A specific method	MODEC [19]	SCGI [13]	GED[7]	GED[7]	GED[7]	GED[7]
Prediction Manner	CE	CE	CE &NE	CE	CE	CE	CE	CE
Unpredicted events	None	None	Continue	None	Merge Split	Merge Split	None	None
Software	CFinder[1] Weka[2]	Weka	MODEC [19] Weka	CFinder Weka	CFinder Weka CRFsuite [3]	Weka	CFinder Weka	Weka

V. OPEN PROBLEMS FOR SUPERVISED LEARNING BASED PREDICTION METHODS

Predicting the evolution of communities is a challenging subject. In this section, we only address the research opportunities that arise from prediction methods that use supervised classifiers, as these are the most commonly used in this area.

Public datasets containing the ground truth evolution events for predicting community evolution are not available. The works in Table 1 use the results of tracking methods as ground truth. However, there is no tracking method that works with 100% accuracy for datasets. Therefore, we need datasets that tell the truth as a benchmark.

In addition, it is necessary to develop a *methodology* for creating the ground truth dataset for a given dataset.

Machine learning is a very powerful tool when the data set is large. As we have more and more data to analyze, we should take advantage of it. That is, there is a need to find evolutionary patterns of communities in the *data without labeling them*.

With the emergence and proliferation of the mobile web and consequently mobile networks, we propose a model/method for predicting the evolution of communities in dynamic *mobile networks*. This will provide the ability to analyze these datasets in a limited-memory environment, which will contribute to the areas of mobile networks, tracking, and community evolution prediction.

VI. CONCLUDING DISCUSSION

Communities are the most meaningful structure in real networks. Knowing their evolution and making predictions about their future provides very useful information for decision support systems in many domains. Therefore, many researchers are concerned with these issues.

Table 2 summarizes our proposed classes, some related works, their limitations and possible research areas. Since existing statistical methods for predicting community evolution only consider topological features, possible future solutions should diversify the types of features, such as temporal features, content-based features, and the features of influential members, rather than only structural features. In addition, solutions must also cover the prediction of dissolution events.

As it is seen from the table, existing rule discovery based methods discover prediction patterns/rules in time series. However, unsupervised learning based methods are better at discovering patterns/rules in data, even if they are not time series. Therefore, unsupervised machine learning methods can be used to develop possible solutions for features. Since existing methods only consider structural features, the solutions can also consider content-based features.

As mentioned in Section V, supervised learning methods require public benchmark datasets that contain evolution events with evolving communities, or a tool to generate these benchmark datasets. Therefore, such datasets or tools can be developed. Since supervised learning methods require labeled data to be trained, potential feature solutions can also be developed with unsupervised learning. In addition, mobile networks are an active research area, and developing a

[1] http://www.cfinder.org/
[2] https://www.cs.waikato.ac.nz/ml/weka/
[3] http://www.chokkan.org/software/crfsuite/

model/method that works with dynamic mobile networks can be considered as a research direction.

Since there is no taxonomy of existing methods for predicting community evolution, we propose a classification in this paper. Since most of the works belong to supervised learning based methods, we focus on them by reviewing related works and giving an outlook on open research problems. Finally, we give a discussion. Thus, this paper provides an up-to-date overview and a quick start for researchers and developers in the field of community evolution prediction.

Table 2: Summary of prediction method classes and research directions

Prediction Class	Some Example works	Limitations	Research Directions
Statistical Methods	Tajeuna et al. [4]	• Disregard dissolution of communities • Regard only topological features	• Diversifying feature types such as influential members, temporal changes in features, contextual attributes etc. • Analysis of dissolution event
Rule Discovery Based Methods	Koloniari et al. [5]	• Discovery of rules on time series • Regard only topological features	• Developing unsupervised machine learning classifiers to uncover prediction patterns/rules • Regarding content based features
Supervised Learning Based Methods	Brodka et al. [6] İlhan&Öğüdücü [9] Takaffoli et al. [11] Saganowski et al. [12] Diakidis et al. [14] Pavlopoulou et al. [15] Dakiche et al. [16] Dakiche et al. [17]	• No available public benchmark datasets • No method to generate benchmark datasets • Need to labeled data • Mobile network datasets are not considered.	• Developing benchmark datasets • Developing a methodology for generating the benchmark datasets • Developing unsupervised machine learning classifiers • Developing a model/method working on dynamic mobile networks.

REFERENCES

[1] V. D. Blondel, J. –L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks", *J. Stat. Mech. Theory Exp.*, vol. 2008, Oct. 2008. Art. no. 10008. DOI: 10.1088/1742-5468/2008/10/P10008

[2] V. A. Traag, L. Waltman, and N. J. Van Eck, "From Louvain to Leiden: guaranteeing well-connected communities," *Sci. Rep.*, vol. 9, Mar. 2019. Art. no. 5233. DOI: 10.1038/s41598-019-41695-z

[3] G. Palla, A. L. Barabási, and T. Vicsek, "Quantifying social group evolution," *Nature*, vol. *446, no.* 7136, pp. 664-667, Apr. 2007. DOI: 10.1038/nature05670.

[4] E. G. Tajeuna, M. Bouguessa, M., and S. Wang, "Modeling and predicting community structure changes in time-evolving social networks," *IEEE Trans. Knowl. Data Enq*, vol.31, no. 6, pp. 1166-1180, June 2019. DOI: 10.1109/TKDE.2018.2851586.

[5] G. Koloniari, G. Evangelidis, N. Sachpenderis, and I. Milonas, "A Framework for Predicting Community Behavior in Evolving Social Networks," in Proc. ACM the 9th Balkan Conf. Inform., Sofia, Bulgaria 2019, pp. 1-4. DOI: 10.1145/3351556.3351583

[6] P. Bródka, P. Kazienko, and B. Bartosz, "Predicting group evolution in the social network," in Proc. Int. Conf. Soc. Inf., Lausanne, Switzerland, 2012, pp. 54-67. DOI: 10.1007/978-3-642-35386-4_5

[7] P. Bródka, S. Saganowski, and P. Kazienko, "GED: the method for group evolution discovery in social networks." Soc. Netw. Anal. Min., vol. 3, no. 1, pp. 1-14, Mar. 2013. DOI: 10.1007/s13278-012-0058-8

[8] E. Frank, M. A. Hall, and I. H. Witten, "The WEKA Workbench". Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

[9] N. Ilhan, and Ş. G. Öğüdücü, "Community event prediction in dynamic social networks," in 12th Int. Conf. Mach. Learn. Appl., Miami, USA, 2013, pp. 191–196. DOI:10.1109/ICMLA.2013.40

[10] N. İlhan, Ş. G. Öğüdücü, "Feature identification for predicting community evolution in dynamic social networks," *Eng. Appl. Artif. Intell.*, vol. 55, pp. 202-218, Oct. 2016. DOI: 10.1016/j.engappai.2016.06.003.

[11] M. Takaffoli, R. Rabbany, and O. R. Zaïane, "Community evolution prediction in dynamic social networks," in 2014 IEEE/ACM Int. Conf. Adv. Soc. Netw. Anal. Min., Beijing, China, 2014, pp. 9-16. DOI:10.1109/ASONAM.2014.6921553

[12] S. Saganowski, B. Gliwa, P. Bródka, A. Zygmunt, P. Kazienko, and J. Koźlak, "Predicting community evolution in social networks," *Entropy*, vol. 17, no. 5, pp. 3053-3096, May 2015. DOI: 10.3390/e17053053

[13] B. Gliwa, S. Saganowski, A. Zygmunt, P. Bródka, K. Przemyslaw, and J. Kozak, "Identification of group changes in blogosphere," in 2012 IEEE/ACM Int. Conf. Adv. Soc. Netw. Anal. Min., Istanbul, Turkey, 2021, pp. 1201-126. DOI: 10.1109/ASONAM.2012.207

[14] G. Diakidis, D. Karna, D. Fasarakis-Hilliard, D. Vogiatzis, and G. Paliouras, "Predicting the evolution of communities in social networks," in the 5th Int. Conf. Web Intell., Min. Semant., Larnaca, Cyprus, 2015, pp. 1-6. DOI: 10.1145/2797115.2797119

[15] M. E. G. Pavlopoulou, G. Tzortzis, D. Vogiatzis, and G. Paliouras, "Predicting the evolution of communities in social networks using structural and temporal features," in the 12th Int. Workshop Semant. Soc. Media Adapt. Pers., Bratislava, Slovakia, 2017, pp. 40-45. DOI: 10.1109/SMAP.2017.8022665

[16] N. Dakiche, F. T. Benbouzid-Si, Y. Slimani, and K. Benatchba, "Community evolution prediction in dynamic social networks using community features' change rates," in the 34th ACM/SIGAPP Symp. Appl. Comput., Limassol, Cyprus, 2019, pp. 2078–2085. DOI: 10.1145/3297280.3297484

[17] N. Dakiche, F. T. Benbouzid-Si, K. Benatchba, and Y. Slimani, "Tailored network splitting for community evolution prediction in dynamic social networks," *New Gener. Comput.*, vol. 39, pp. 303–340, April 2021. DOI: 10.1007/s00354-021-00122-6

[18] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labelling sequence data," in Proc. Int. Conf. Mach. Learn., Massachusetts, USA, 2001, pp. 282-289.

[19] M. Takaffoli, F. Sangi, J. Fagnan, and O. R. Zaïane, "MODEC — Modeling and Detecting Evolutions of Communities," in 5th Int. AAAI Conf. Web Soc. Media, Catalonia, Spain, 2011, pp. 626-629.