Predicting node memberships in evolving clusters

Evangelia Tsoukanara†
Department of Applied Informatics
University of Macedonia Thessaloniki, Greece
etsoukanara@uom.edu.gr

Georgia Koloniari
Department of Applied Informatics
University of Macedonia Thessaloniki, Greece
gkoloniari@uom.edu.gr

Evaggelia Pitoura
Computer Science & Engineering,
University of Ioannina
Ioannina, Greece
pitoura@cse.uoi.gr

ABSTRACT

Most previous research considers community evolution prediction as the prediction of future occurring events at community level. In our work, we focus on individual nodes and define a novel problem of predicting the future state of a node. Motivated by the extensive utilization of embeddings in machine learning tasks, we introduce embeddings-based features and compare them to the typical features used in community evolution prediction. In addition, taking advantage of the broad information the history of a node provides, we deploy chains of features.

INTRODUCTION

Community evolution is about tracking changes of identified groups of nodes within a dynamic network. Most works focus on simple transformations that involve communities, usually called *events* [1]. Several structural features are used to predict the future state of a community, while historical chains of features are utilized to model the history of a community and improve prediction accuracy [2]. In our work, we focus on nodes and aim to predict changes in the cluster membership of a node, modeling our problem as a classification problem setting three node events / classes: *stay*, *move,* and *drop*.

PROBLEM DEFINITION

A temporal social network is often represented as $\{G_1, G_2, \dots, G_n\}$, where $G_i = (V_i, E_i)$ represents a snapshot of graph G, and V_i, E_i are the set of nodes and edges at time i. A sequence of clusterings $C_1, C_2, \dots, C_i$ corresponds to a consecutive set of timeframes of graph G, where $C_i = \{C_i^1, C_i^2, \dots, C_i^m\}$ represents the clustering of G_i in m clusters. We define our problem as follows.

Given a sequence of clusterings $C_1, C_2, \dots, C_i$ *and a node* $v \in C_i^j, 1 \le j \le m$, *predict the next state of node* v *in the next timeframe* $i + 1$ *as*:

- $stay, S$: *node* v *stays in the same cluster in* $i + 1, v \in C_{i+1}^j$
- $move, M$: *node* v *moves to another cluster in* $i + 1, v \in C_{i+1}^k, k \neq j$, *and*
- $drop, D$: *node* v *drops out of the network,* $v \notin V_{i+1}$.

METHODOLOGY

We employ the ComE [3] framework to get our node and cluster-median embeddings, and the cluster membership of the node for each snapshot, which next, will form the classification features. ComE jointly solves the tasks of community detection, community embedding and node embedding introducing a community-aware proximity.

Given that similar nodes have similar embeddings, we use as features to our classifier the Euclidean *distance* between pairs of embeddings, that is, the distance of the node from the most or least similar node in the cluster, the average distance from all cluster nodes, and the distance from the median of the cluster the node belongs to. Similarly, we exploit the distance from the closest other median, and the relevant distances regarding the nodes outside the cluster.

Our baseline method includes classic features, i.e. structural node features typically used in community evolution prediction tasks, that is, *degree*, *betweenness*, *closeness*, and *eigenvector* centrality at cluster and network level.

Next, our clusters are mapped across consecutive snapshots based on the majority of common nodes. Various lengths of evolution chains are deployed to capture the history of each node. Each chain consisting of node features from each snapshot along with the state of the node at the next snapshot sets up the final block of features for our classifier.

RESULTS

We use three datasets for our evaluation, the *DBLP*[1] citation network extracted from DBLP, the *Email-eu*[2] communication network and *Syntgen* produced using the Syntgen [4] generator. We use stratified 5-fold cross-validation that preserves class distribution to tackle the problem of imbalanced data and employ a Random Forest classifier to conduct our experiments.

Fig. 1a depicts class performance (denoted as S\M\D) for each dataset w.r.t. each problem using the default chain length (5). We observe that both ComE-based (denoted as CM) and classic (denoted as CL) features perform in a similar way, while needing 5493s and 34618s respectively to get computed. Also, we notice that class *move* seems to be the most difficult to predict.

Fig. 1b illustrates performance with increasing chain length. Although, we notice an upward trend as length of chains increase, we must point out that as a chain gets longer the nodes available for prediction are reduced.

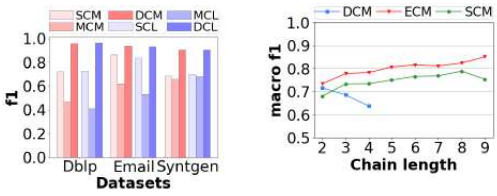

Fig. 1: Macro f1 score (a) per class and, (b) per chain length

CONCLUSION

In this work, we introduced a novel problem on community evolution prediction, focusing on individual nodes instead of communities. We studied the prediction of the future state of a node w.r.t. the community the node belongs, taking advantage of the available history. Comparing to the baseline method, our method performs similarly or in some cases slightly better, while requiring far less computation time.

REFERENCES

[1] Palla, G., Barabasi, A.L., Vicsek, T.: Quantifying social group evolution. Nature 446, 664-667 (2007)
[2] Cavallari, S., Zheng, V.W., Cai, H., Chang, K.C.C., Cambria, E.: Learning community embedding with community detection and node embedding on graphs. In: Proceedings of the 2017 ACM CIKM. p. 377{386 (2017)
[3] Saganowski, S.: Predicting community evolution in social networks. In: Proceedings of the 2015 IEEE/ACM ASONAM. p. 924{925 (2015)
[4] Pereira, L.R., Lopes, R.J., Louc~a, J.: Syntgen: a system to generate temporal networks with user-specifed topology. Journal of Complex Networks 4(0), 1{26 (2019)

[1] https://www.aminer.org/citation

[2] https://snap.stanford.edu/data/email-Eu-core-temporal.html