Predicting the Evolution of Communities in Social Networks Using Structural and Temporal Features

Maria Evangelia G. Pavlopoulou*, Grigorios Tzortzis †, Dimitrios Vogiatzis † and George Paliouras †
* National and Kapodistrian University of Athens, Athens, Greece
† Institute of Informatics and Telecommunications NCSR “Demokritos”, Athens, Greece

NATIONAL CENTRE FOR SCIENTIFIC RESEARCH “DEMOKRITOS”

SMAP 17 Bratislava
Collocated with UMAP 2017

9-10th July, 2017 at FIIT STU, Bratislava, Slovakia

Outline

1) Social networks

2) Predicting Community Evolution

3) Experiments

4) Conclusion

5) Future Work

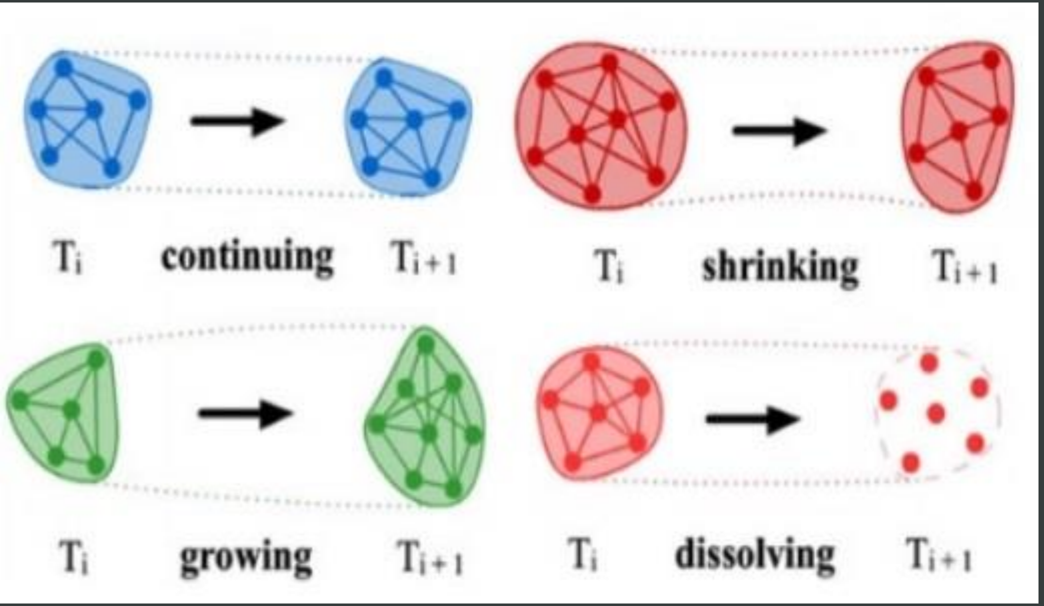
Social Networks

- Social networks are dynamic
- User relationships are intrinsically temporal and change over time
- Communities of users also change over time (i.e. evolve)

- Modeling and Predicting the evolution of a network community
- **Community evolution** : reduce or increase in size, appear or disappear from the network
- Communities are graphs
- Application in marketing, criminology, journalism

Our focus

- ❖ Focus on prediciting four popular evolutionary phenomena of communities



- ❖ We employ an extensive set of **structural** and **temporal** features
    - Capture various characteristics of the communities in order to get accurate predictions

Predicting Community Evolution

1) Segment the social network data into timeframes.

2) Detect the communities in each timeframe.

3) Track communities across time to identify their evolution and corresponding evolutionary events.

4) Compute structural and temporal community features.

5) Train a classifier to predict community evolution.

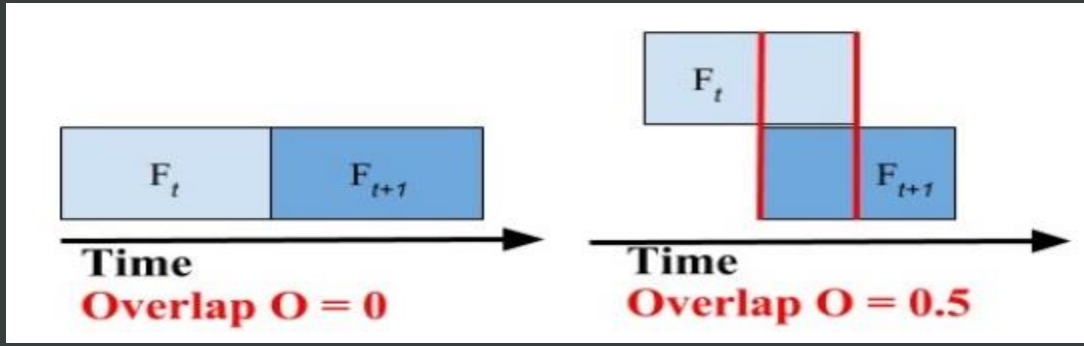
Mathematics Stack Exchange Dataset

- Q&A site for people studying math
- Users post questions, answer questions and comment on the users' posts
- All questions are **tagged** with their subject areas (**i.e. topics**)
- Answers and comments inherit the topics of the question they correspond to

Posts	Time Period
376030	2009 -2013

Segmentation into timeframes

- ❖ Data are timestamped. We discretize it into a predefined number of time-ordered timeframes Ft, t = 1, . . . , T.
- ❖ Each timeframe contains the same number of elements (user posts).
- ❖ Consecutive timeframes are allowed to overlap, with an overlap O ∈ [0, 1]
- ❖ The amount of overlap designates the percentage of the previous timeframe that is also part of the next timeframe.



Community Detection

- A community corresponds to a densely connected subset of users (i.e. a subgraph) of the timeframe graph that is loosely connected to the rest of the graph

- We take advantage of the topics associated with posts in Mathematics Stack Exchange

- Users belong in the same community if they make posts about the same topic.

Community ⟷ Mathematics Stack Exchange Topic

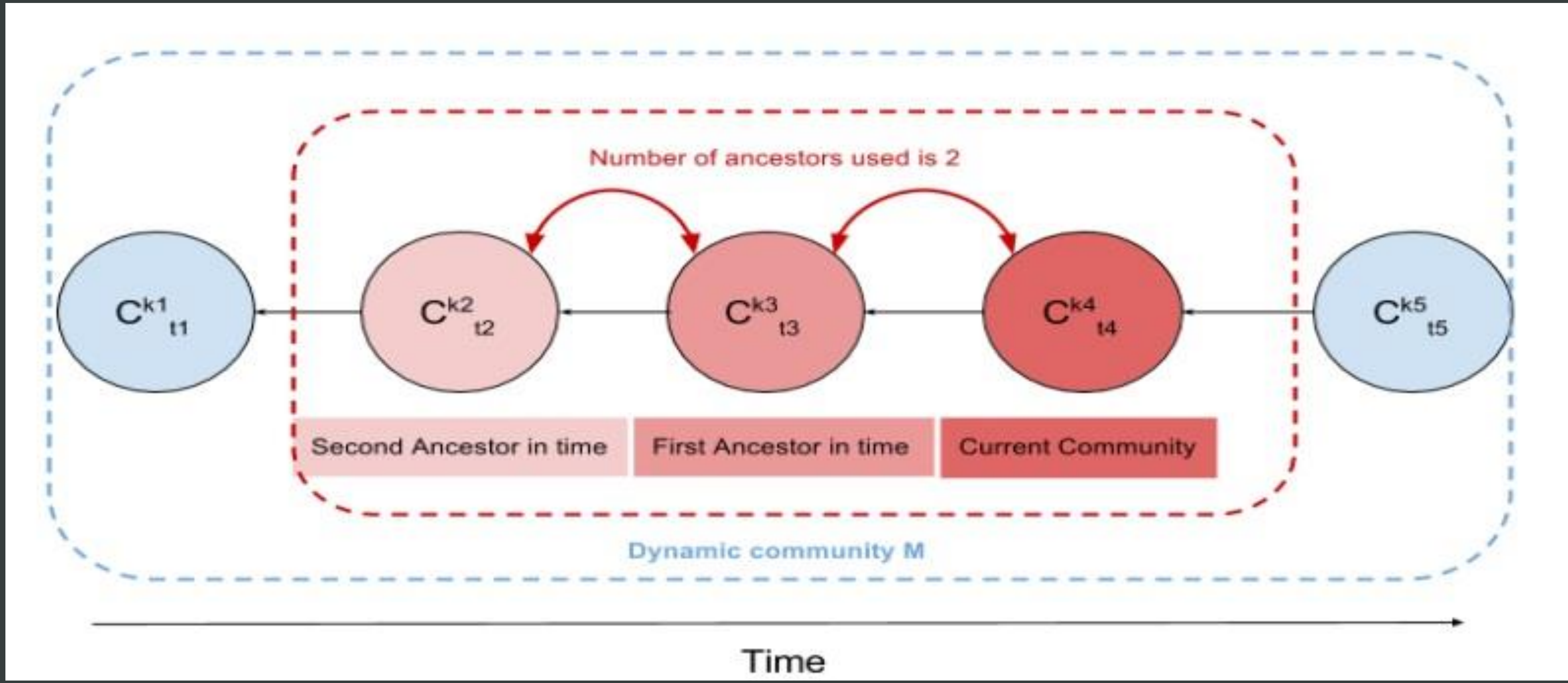
Community Tracking (1/2)

- ❖ Needed to build the ground truth of the data
- ❖ For each detected community in timeframe Ft we obtain its topic
- ❖ Then we look for a matching community with the same topic in a subsequent timeframe Ft' , t' > t

No matching community found	→	Dissolution
$n(C_t^k) - n(C_{t'}^{k'}) > th$	→	Shrinkage
$n(C_{t'}^{k'}) - n(C_t^k) > th$	→	Growth
Else	→	Continuation

Community Tracking (2/2)

- Matching communities do not necessarily belong to consecutive timeframes
- **Dynamic community :** a sequence of matched communities
- For a community, its past instances are referred to as the ancestors of the community.



Community Feature Engineering (1/3)

- **Structural Features**

1. **Relative Size :** normalized value of community's size in timeframe Ft
2. **Relative Edges Number :** normalized value of edges belonging to the community
3. **Density :** ratio of the actual edges of community to the maximum number of edges the community could have
4. **Cohesion :** product between the density and the inverse fraction of edges pointing outside of community
5. **Ratio Association** : average internal degree of a community's members
6. **Ratio Cut :** average external degree of a community's members
7. **Normalized Cut :** edge volume that points outside of the community

Community Feature Engineering (2/3)

8. **Average Path Length**

9. Diameter

10. Clustering Coefficient : how often, on average, the neighbours of a node of the community are also connected to each other

11. Centrality : how central (i.e. centre of importance) each node (i.e. user) of a community is

a) Closeness Centrality

b) Betweeness Centrality

c) Eigenvector Centrality

Community Feature Engineering (3/3)

- **Temporal Features**

1. **Structural features and evolutionary events of ancestors**
2. **Jaccard Coefficient :** members that are common in both instances of the community
3. **Join Nodes Ratio :** percentage of new members joining the community
4. **Left Nodes Ratio :** percentage of members leaving the community
5. **Activeness :** new edges per node that a community contains
6. **Lifespan :** ratio of the ancestors the community has based on the corresponding dynamic community, to the maximum number of ancestors it could have
7. **Aging :** average age of the community members, normalized by dividing with the maximum possible age of members

Learning A Predictive Model

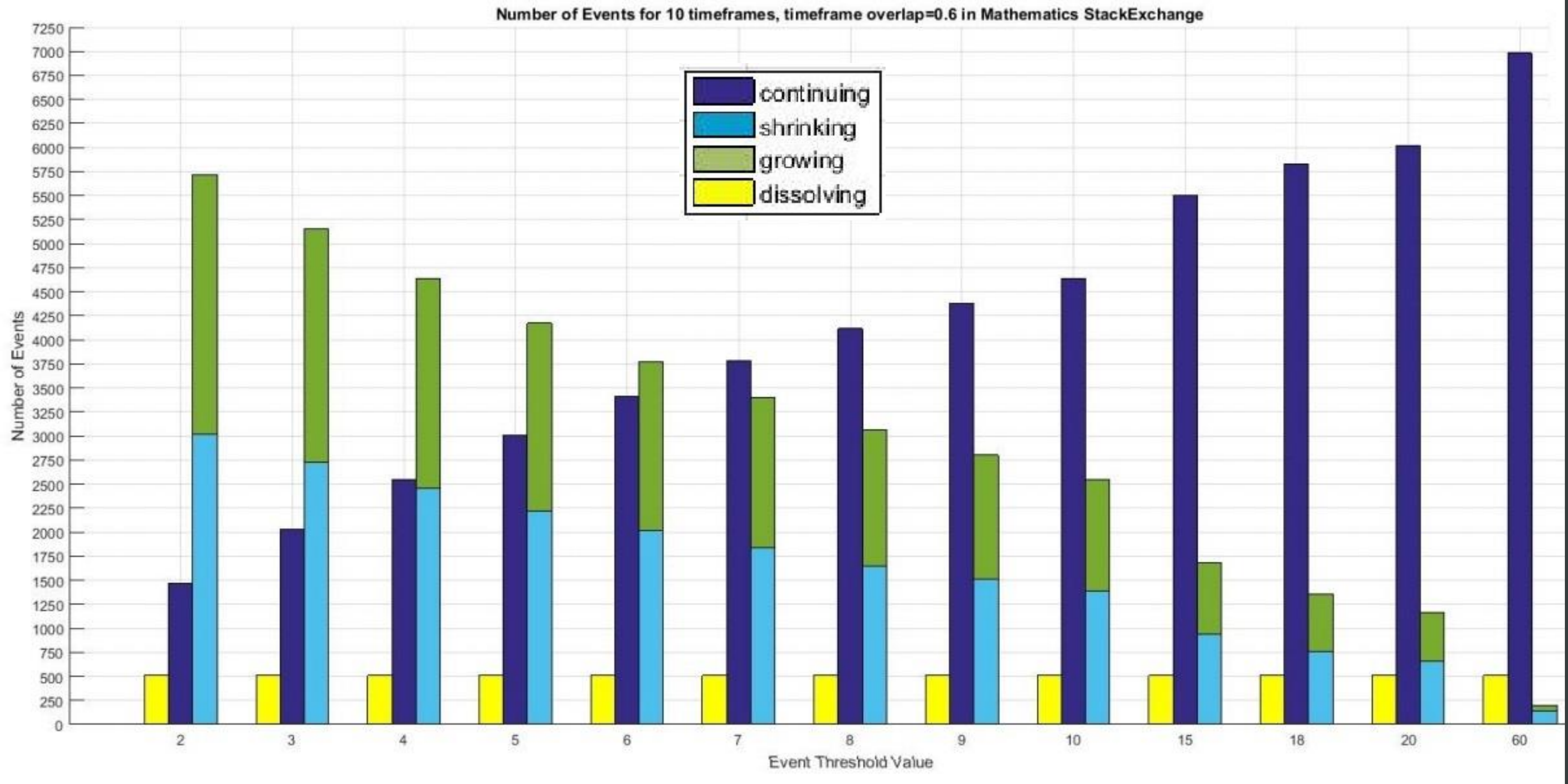
- ❖ Classifier used : Support Vector Machines, with RBF kernel

- ❖ Training – Testing : **Time Series Cross Validation**
  - variant of k-fold cross validation technique
  - respect the natural ordering of the timestamped data

- ❖ To counter the imbalance that exists in our dataset we apply WEKA's

- SMOTE oversampling technique
- Spreadsubsample undersampling technique

Imbalance in our dataset



Experimental Evaluation

- ❖ Computing temporal features : we use the **n** most recent ancestors in time

- ❖ We perform experiments for $n \in \{0,2,4,6\}$ and use

  i) only the structural features and the evolutionary events of the ancestors as temporal features

  ii) the complete set of temporal features.

- ❖ For each value of n we try $th \in \{2, 3, 4, 5, 6\}$ and optimize the internal parameters of the SVM classifier

- ❖ By changing **th** we actually change the ground truth

F1 score
structural features
temporal features : evolutionary events of ancestors

Ancestors	Continue	Shrink	Grow	Dissolve	Overall
0	0.4783	0.5644	0.1526	0.3587	0.4694
2	0.4683	0.5642	0.4282	0.4333	0.5072
4	0.4726	0.6238	0.3874	0.4263	**0.5105**
6	0.3783	0.6805	0.3257	0.4415	0.4785

F1 score
structural features
temporal features : complete set

Ancestors	Continue	Shrink	Grow	Dissolve	Overall
0	0.4783	0.5644	0.1526	0.3587	0.4694
2	0.5444	0.6652	0.4202	0.5762	**0.5720**
4	0.5403	0.7123	0.3884	0.5095	0.5475
6	0.4812	0.7152	0.3292	0.4857	0.5581

Experimenting with event threshold

- We experiment with a greater range of event threshold values
  th $\in$ {5, 10, 15, 20, 25, 30, 60}

- We become more strict while deciding whether a community has grown or shrunk

- The difference in size between two matched communities must be larger in order to assign these labels.

Results
all temporal features are included
extended set of event threshold values is used

Ancestors	Macro F1	Macro Recall	Macro Precision
0	0.6731	0.6545	0.6928
2	0.7662	0.7572	0.7754
4	**0.7719**	0.7608	0.7835
6	0.7194	0.7132	0.7259

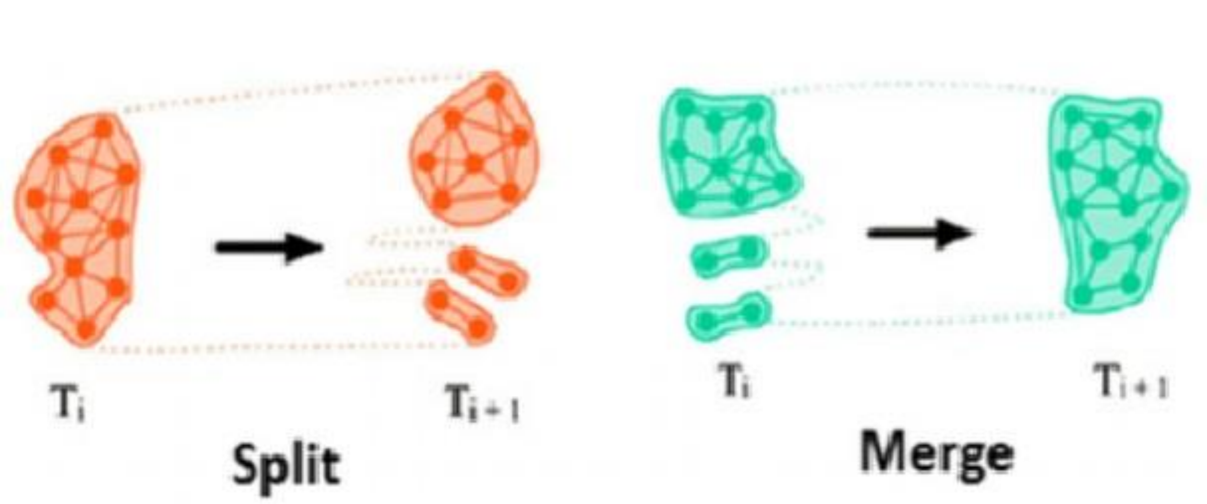
- For all number of ancestors tried , the best results were obtained for **th = 30**.

Conclusions

- ❖ Prediction accuracy improves when using temporal features on top of structural ones.
- ❖ The number of ancestors affects prediction results.
- ❖ The past of a community encapsulates information about its future.
- ❖ Using <u>four</u> ancestors gave the best results in our dataset.
- ❖ Evolution prediction results are improved if we do not go <u>too</u> far back in time

Future Work

- ❖ Predicting other types of community evolution



- ❖ Incorporating other types of features
  - Reputation in the Mathematics Stack Exchange site
  - Hashtags in Twitter
- ❖ Using other classifiers apart from SVMs
- ❖ Performing tests with more datasets
- ❖ Comparing our approach with existing ones

Thank you for your attention

Maria Evangelia Pavlopoulou

mary18pav@gmail.com