Predictive Analysis of Temporal and Overlapping Community Structures in Social Media

Mohsen Shahriari, Stephen Gunashekar, Marven von Domarus, Ralf Klamma
Advanced Community Information Systems (ACIS)
RWTH Aachen University
Ahornstr. 55, 52056 Aachen, Germany
{shahriari, sguna, domarus, klamma}@dbis.rwth-aachen.de

ABSTRACT

Digital media has some observable traces named communities. Several events such as split, merge, dissolve and survive happen to communities in social media. But what are significant features to predict these events? And to which extent a feature is relevant in a social media? To answer these questions, we perform a study on community evolution analysis and prediction. We employ three overlapping community detection (OCD) algorithms from literature to the case of time-evolving networks including social, email communication and co-authorship networks. Group evolution discovery (GED) technique is applied to track the identified communities. We compare structural properties of OCD algorithms and investigate most persistent communities over time. Furthermore, static and temporal features of a community are applied to build a logistic classifier for community evolution prediction (CEP). Results reveal important features to predict events happening to a community.

Keywords

Overlapping community detection; community evolution tracking; community evolution prediction

1. INTRODUCTION

People, agents and their interactions are the smallest part of social media and bigger components can be well described by (overlapping) communities [18, 16]. As a matter of fact, communities confront various transformations such as grow, shrink, merge, split, birth and death which are caused by inscribed and circumscribed effects [2, 17]. The internal and external effects imposed to a community are pretty much significant while their meso-scopic study provides the opportunity to seize a better understanding of their alterations and purposely control their evolution [2, 9]. In this regard, we aim to know more about significant features for community evolution prediction. Moreover, how the importance varies in the context of different social networks?

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://10.1145/2872518.2889292 .

To find an answer for the questions, we apply three OCD algorithms named SLPA [19], DMID [16] and AFOCS [14] on Enron email communication, Facebook and DBLP networks in approximately ten-year temporal period. We track and extract the community events such as split, merge, dissolve and survive via comparing the evolution of overlapping communities over consecutive time steps. Furthermore, we build a logistic regression model based on static and dynamic structural properties of communities like centralities, density, (dis)assortative degree mixing and influential nodes. Moreover, we analyse properties of the most persistent community over all domains to observe how community features are changing over time. Results reveal prediction accuracy of the model along with the significance of features in the prediction task. Mainly the impact of community features very much depends on how to detect communities and the context of social media.

Contribution: Community Evolution Prediction (CEP) and Extracting Significant Structural Features We investigate the evolution of the most persistent community over time by calculating various measures. Moreover, three categories of node level, community level and selective features are considered for the prediction task. In fact, we identify the most influential properties of communities and reveal their importance in CEP. For instance, size of a community is the most significant feature for CEP for all the events (merge, split, dissolve, survive) over all social media domains. Finally, relation between how to detect communities and prediction accuracies are investigated.

Contribution: Overlapping Community Detection and Graph Structures For the first time, we apply OCD for CEP because they behave more realistic. In order to comprehend the relationship between the structural properties of community definition and prediction model, three OCD algorithms are used. Number of overlapping nodes, number of communities, average community sizes are compared. Moreover, their corresponding prediction accuracies are computed. Results indicate that significance of features for CEP highly depends on how to define and detect communities.

Contribution: CEP and the Relation to Various Social Media The case study uses multiple datasets from different domains to understand their effect. Social, email communication and co-authorship networks are among them. The purpose of using different data resources from various domains is to observe how the importance of features depend on the context of the communities. Results indicate

Features	**Definition and Description**
LeaderRatio	ratio of leaders
LDegCen	leaders average Degree
LClosenessCentrality	leaders average closeness
LEigenCentrality	leaders average eigenvector
SizeRatio	$S_i^P = \frac{\lvert V_i^P \rvert}{N}$: community size
Density	$D_i^P = \left[\frac{2\lvert E_i^P \rvert}{\lvert V_i^P \rvert(\lvert V_i^P \rvert - 1)}\right]_{2\lvert E}$
Cohesion	$C_i^P =$

Table 2: Comparisons of number (C), average size (AvgSize) of communities and number of overlapping members (Ovl) for SLPA, DMID and AFOCS on some datasets.

	SLPA C	*SLPA AvgSize*	*SLPA Ovl*	*AFOCS C*	*AFOCS AvgSize*	*AFOCS Ovl*	*DMID C*	*DMID AvgSize*	*DMID Ovl*
Karate Club	4	10.5	0.23	0	0	0	2	15	0.43
Dolphins	3	216	0.048	34	3.3	0.008	6	25.2	0.65
PolBlogs	5	245.4	0.002	823	148	0	15	1026.2	0.839
NetScience	407	4.1	0.1656	658	4.2	0.1	61	177.31	0.55
Power	737	9.9	0.40376	4256	2.1	0.076	569	462.8	0.96
CA-GrQc	954	6.9	0.24475	2683	3.9	0.038	263	654.7	0.8051
p2p-Gnutella08	230	18.24	0.50436	5274	2.39	0.04	614	573.4	0.6803

3.1 Community Mapping

GED method [4] is used to detect possible events occurring in a network. GED interprets the temporal network as a list of time frames with graphs leading to a successive event such as continuing, splitting, merging, dissolving, forming. GED applies an *Inclusion* measure that evaluates the inclusion of one group (or community) in another which is formulated using two quantitative and qualitative factors. The inclusion factor of a community C_i in snapshot t and another community C_j in snapshot $t+1$ is represented as $I(C_i^t, C_j^{t+1})$ and is calculated as,

$$I(C_i^t, C_j^{t+1}) = \frac{|C_i^t \cap C_j^{t+1}|}{|C_i^t|} \cdot \frac{\sum_{x \in C_i^t \cap C_j^{t+1}} NI(C_i^t(u))}{\sum_{u \in C_i^t} NI(C_i^t(u))} \quad (1)$$

where $NI(C^t(u))$ is the node indicator which is any statistical metric that evaluates the node's importance within a community.

3.2 Features for Prediction

As for the prediction task, we apply different set of structural features shown in the Table 1. We consider three sets of features including node level, community level and temporal features. Node level features are computed for most influential nodes named leaders. We consider nodes with the highest 20 percent eigenvector centrality values as leaders. Node level features consider static and dynamic centrality values of leader members. We also compute static and dynamic community level features such as density, cohesion, size, average clustering coefficient, degree and closeness centrality. Dynamic features indicate change in properties regarding two consecutive snapshot of a community. Previous state of a community is taken into account as a binary variable e.g. Previous-Dissolve checks if previous state of a community is dissolve. Prediction accuracy $\frac{(TP+TN)}{(P+N)}$ is applied to evaluate the test results. TP is true positive rate and TN is true negative rates in test runs.

Regarding preprocessing of datasets, the ground truth of the classes are extracted based on the GED mapping. In other words, we extract events including survive, merge, split and dissolve based on successive network snapshots. For instance, when we observe same community at time t and $t+1$ then we consider it as a survive event. As the result of the preprocessing, GED method detected 21 survive, 31 merge, 12 split and 24 dissolve events in the Enron dataset. GED also identified 51 survive, 288 merge, 156 split and 8870 dissolve events for Facebook communities. Finally, 146 survive, 200 merge, 187 split and 3319 dissolve were labelled for communities in DBLP. We use existing filter in WEKA like SMOTE to synthetically balance the classes.

4. RESULTS

4.1 OCD Algorithm Properties

In this section, we discuss some properties of the algorithms DMID, SLPA and AFOCS on different datasets [11, 1] including Zachary karate club (34 nodes, 78 edges, social network), Dolphins (62 nodes, 159 edges, biological network), PolBlogs (1224 nodes, 19022 edges, Internet topology network), NetScience (1461 nodes, 2742 edges, collaboration network), Power Grid (4941 nodes, 6594 edges, Technical network), General Relativity (CA-GrQ) (5242 nodes, 28968 edges, collaboration network) and Gnutella (6301 nodes, 20777 edges, peer-to-peer network). Table 2 indicates number of found communities (C), average community size (AvgSize) and percentage of overlapping nodes (Ovl). As it can be observed, percentage of overlapping nodes for DMID algorithm for almost all of the datasets including Karate Club (0.43), Dolphins (0.65), PolBlogs (0.839), Netscience (0.55), PowerGrid (0.96), CA-GrQc (0.8051) and Gnutella (0.55) is more than other two algorithms; SLPA with respectively 0.23, 0.048, 0.002, 0.1656, 0.40376, 0,24475 and 0.504 and AFOCS with respectively 0, 0.008, 0, 0.1, 0.076, 0.038, 0.04. This indicates that DMID detects communities with high overlapping part without tending to merge them. Hence, DMID has higher average community sizes than SLPA and AFOCS. For Dolphins, PolBlogs NetScience, Power, CA-GrQc and p2p-Gnutella08, AFOCS detects respectively 34, 823, 658, 4256, 2683 and 5274 communities which are more than number of communities in comparison to DMID with respectively 6, 15, 61, 569, 263 and 614 and SLPA respectively with 3, 5, 407, 737, 954 and 230 communities on these datasets. This as well confirms that AFOCS prefers to form more overlapping communities than SLPA and DMID.

4.2 Properties of the persistent community

Moreover, we analyse the largest persistent community for different datasets and algorithms. Structural features mentioned in Table 1 are considered for this. Each of the algorithms in Figure 1 has its own legend. Regarding email network, DMID, AFOCS and SLPA experience the most persistent community of longevity 5. Stabilities of detected communities are somehow similar for the longest persistent community in the Enron email network. Moreover, com-

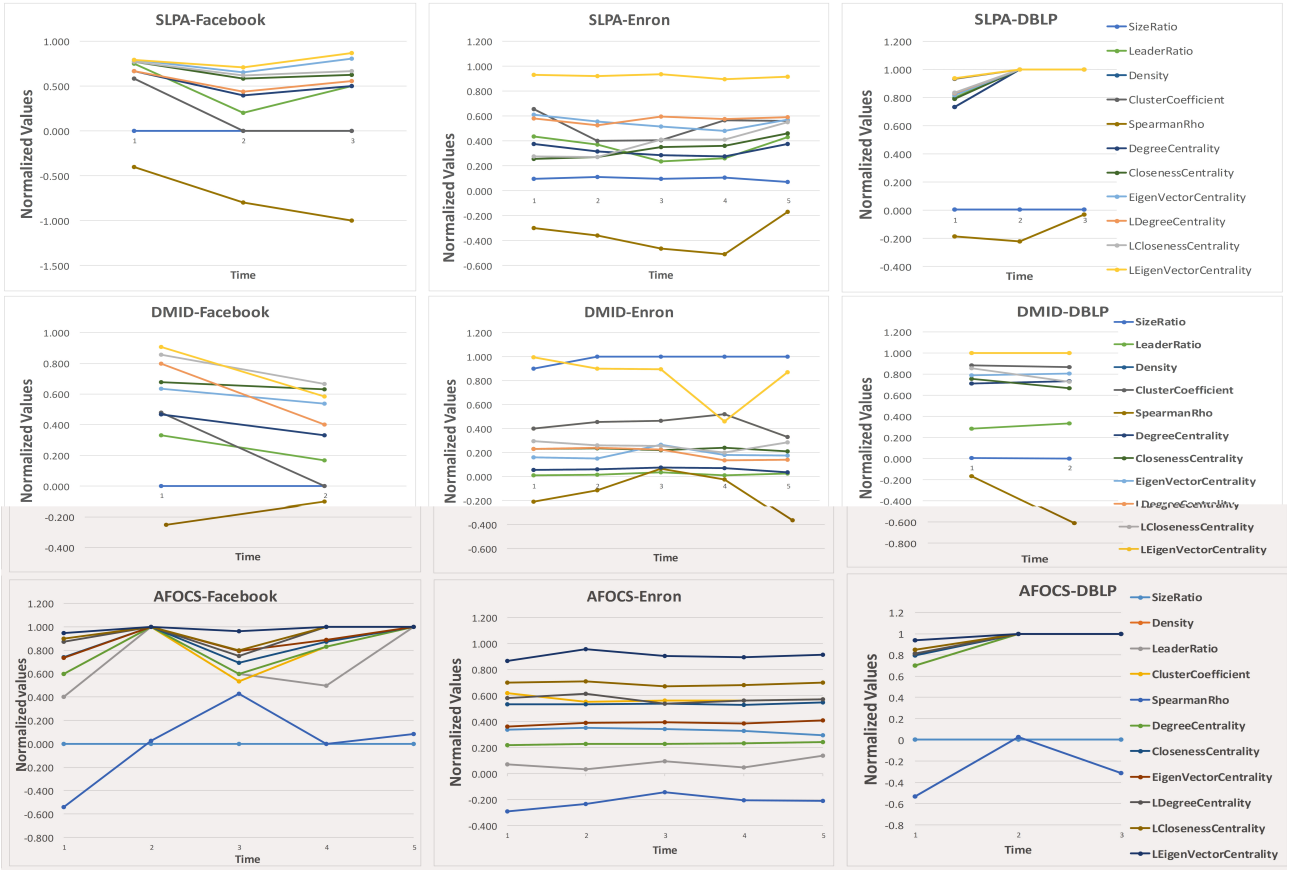

Figure 1: Different properties of the longest persistent community on different datasets based on the algorithms

munity properties are quite similar. Only SpearmanRho measure is low and somehow fluctuating in all three cases in the Enron network. This indicates that the most persistent community has disassortative degree mixing property. Therefore, disassortative degree mixing is significant for community continuation (disappearing).

Regarding the Facebook dataset, DMID, SLPA and AFOCS resulted in 2, 3 and 5 most persistent communities. This is somehow compliant with the results in Table 2 in which DMID detects bigger communities with more overlap. Due to smaller community sizes detected by AFOCS, it leads to the longest persistent community of length 5. Range of property values for DMID, SLPA and AFOCS are quite different because their persistent communities are different and would possess divergent property values.

Regarding DMID and SLPA, properties are more divergent and fluctuating in comparison to AFOCS. All of them experience high disassortative degree mixing with AFOCS having highest fluctuation of disassortative degree mixing. For DBLP also all the algorithms have smaller longevity for most persistent community with SLPA (3), DMID (2) and AFOCS (3) while most of people co-author with few people and smaller communities form. The property values for DMID are quite divergent in comparison to SLPA and AFOCS. The community has negative assortative degree mixing with abrupt fluctuations before disappearing.

4.3 Community Evolution Prediction

4.3.1 Community Prediction Accuracy

In this section, results of the prediction task are shown in Table 3 and we enumerate the important features in CEP for each of the events happening to the community. Prediction accuracies are only shown for selective features based on a wrapper method available in WEKA data mining tool library. Because they lead to better prediction accuracies so we only consider them in our discussion. Regarding Facebook, the highest prediction value for the survive event can be observed for AFOCS (82.35). This is also compliant with the results in Table 2 in which AFOCS detects smaller communities. Smaller communities are more likely to survive than bigger ones. Regarding dissolve (dead) event in Facebook, DMID has a prediction accuracy of 78.57 which is higher than selective features of SLPA-Facebook (72.82) and AFOCS-Facebook (63.28). This is because DMID detects bigger communities with more overlap than AFOCS and SLPA. So bigger communities have higher tendency to dissolve (die) than small communities. Regarding merge, SLPA-Facebook (67.83), DMID-Facebook (67.22) and AFOCS-Facebook (65.02) have approximately similar prediction accuracies regarding the merge event. Because small and big communities have similar tendency for the merge event. Finally for split, SLPA-Facebook has the prediction accuracy

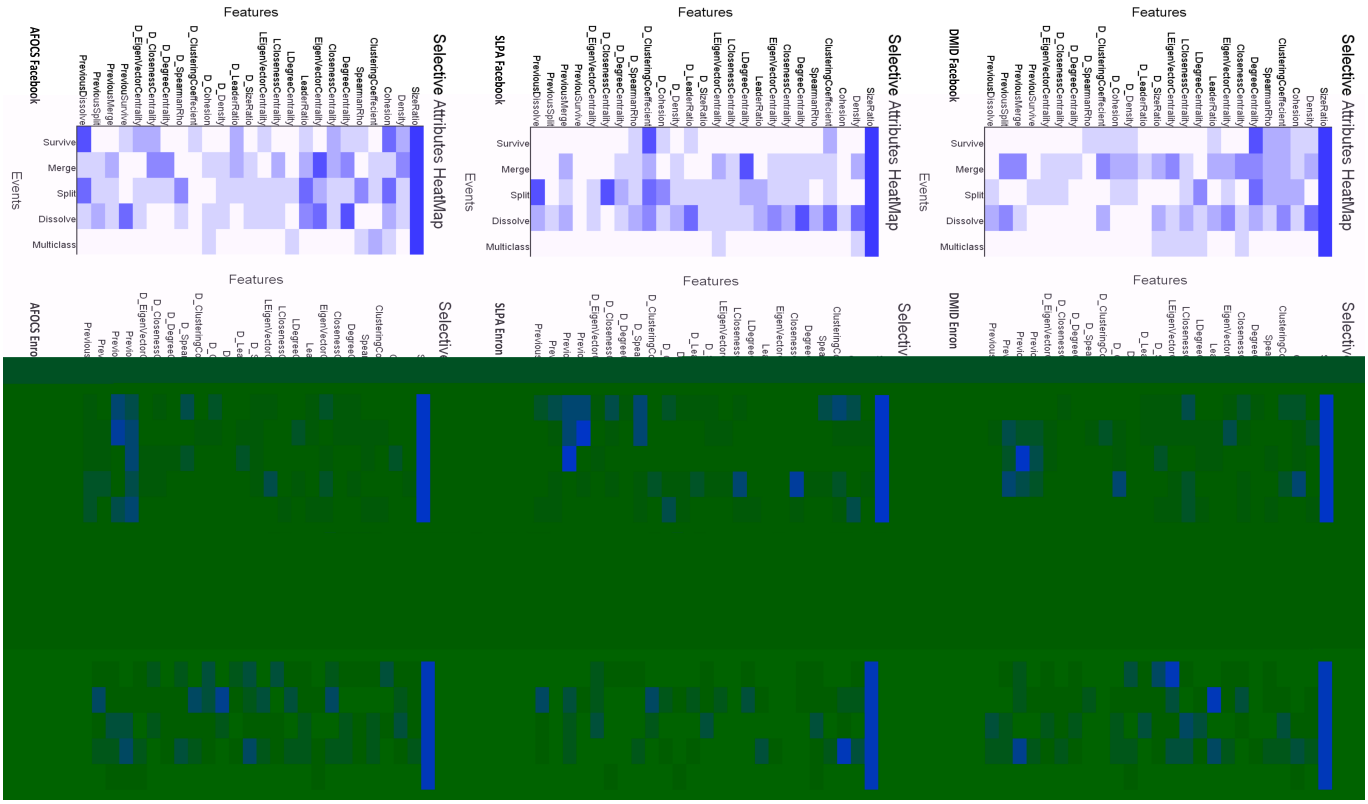

Figure 2: Comparison of important features.

of 77.94 which is higher than DMID-Facebook (63.75) and AFOCS-Facebook (66.03).

As for DBLP and the survive event, DMID (82.14) leads to higher accuracy in comparison to SLPA (67.16) and AFOCS (55.99). As for dissolve, SLPA (67.2) and DMID (66.42) are higher than AFOCS (55.99). As for merge, DMID (66.76) can better predict happening of merge event in comparison to SLPA (64.78) and AFOCS (63.62). Finally, DMID (74.29) has higher prediction accuracy in comparison to SLPA (65.56) and AFOCS (67.78) for split event. Communities in the DBLP dataset are inherently small and componentized, therefore, DMID might detect more realistic size communities and lead to higher prediction in all the events.

Regarding Enron dataset and the survive event, we have a little different pattern with DMID (88.71) and SLPA (76.92) and AFOCS (82.14). As for dissolve, SLPA (93.75) has higher prediction accuracy than DMID (88.64) and AFOCS (79.17). As for merge, DMID (95.59) has higher prediction accuracy in comparison to SLPA (78.79) and AFOCS (74.56). As for split, DMID (88.71) as well yields better prediction accuracy in comparison to SLPA (83.33) and AFOCS (81.25). In general DMID algorithm has higher prediction accuracy for Enron and DBLP datasets in all events. Moreover, its prediction accuracy for Facebook regarding dissolve and merge events is high and competitive. On the other end, use of AFOCS leads to smaller and more stable communities over time. Results indicate that prediction of community structures highly depends on how to detect communities. In other words, the same set of features lead to different prediction accuracies when considering different ways of identifying communities.

4.3.2 *Comparison of features*

Figure 2 shows heat map of the features which are used for the prediction task. It can be observed for all of the events and all of the algorithms, size of the community is an important feature for CEP. Regarding Facebook and DMID algorithm, in most of the events centrality measures and temporal features play important role. Change in leader ratio is also important in dissolve event relating to SLPA. With AFOCS algorithm, one can observe that not only centrality measures but also the ratio of influential nodes and assortative degree mixing are more important. Assortative degree mixing provides good indicators in survive and merge of communities. For Facebook which is a social network, AFOCS provides a various range of features for the prediction task. As for Enron, only eigenvector centrality in dissolve is important. For SLPA with Enron, measures like cohesion, closensss centrality, size and density of a community are important in survive and merge events. Regarding split and dissolve only centrality and change in size of community play important role in prediction. Here, AFOCS is more similar to the SLPA and approximately the same set of features play important role.

As for DBLP and DMID algorithm, centrality measures and temporal features play important role for survive. Regarding merge, leader ratio is important. For split centrality measures and for dissolve leader ratio and temporal features are more important. As for SLPA, the situation is a little bit different and cohesion, previous merge and centrality

Table 3: Community prediction accuracy

	Survive	Dissolve	Merge	Split
SLPA-Facebook	75	72.82	67.83	77.94
DMID-Facebook	78.57	78.57	67.22	63.75
AFOCS-Facebook	82.35	63.28	65.02	66.03
SLPA-Enron	76.92	93.75	78.79	83.33
DMID-Enron	88.71	88.64	95.59	88.71
AFOCS-Enron	82.14	79.17	74.56	81.25
SLPA-DBLP	67.16	67.2	64.78	65.56
DMID-DBLP	82.14	66.42	66.76	74.29
AFOCS-DBLP	55.99	55.99	63.62	67.78

measures are important for survive. Regarding merge, centrality measure, clustering coefficient and previous merge play important role. As for split, cohesion, eigenvector centrality and leader closeness centrality are important. Centrality measures and temporal features play important role for dissolve. Regarding AFOCS in merge and survive, temporal features like change in cohesion, previous survive and previous merge are important. Regarding split, density and previous merge and for dissolve temporal features are important. Altogether, the importance of features are different in various social media and even with different algorithms. In other words, the importance of the features for the prediction task of each event and dataset very much depends on how to detect communities.

5. CONCLUSION

In this paper, we apply static and dynamic community and node level features to the case of community evolution prediction problem. First, significant features are identified for each separate event happening to the community. Size ratio is the most important feature to predict events happening to communities. Second, results indicate that community fate prediction depends on how to detect communities and dynamics of OCD algorithm. We would like to further explore the results with other algorithms and other social media e.g. open source developer and learning networks. Finally, further fine-grained investigation needs to be done for each of social media to justify observed effects.

6. ACKNOWLEDGMENTS

The work has received funding from the European Commission's FP7 IP Learning Layers under grant agreement no 318209.

7. REFERENCES

[1] The Koblenz network collection. http://konect.uni-koblenz.de/, May 2015.

[2] L. Backstrom, D. Huttenlocher, J. M. Kleinberg, and X. Lan. Group formation in large social networks: Membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD*, pages 44–54. ACM Press, 2006.

[3] P. Bródka, P. Kazienko, and B. Kołoszczyk. Predicting group evolution in the social network. In *Social Informatics*, volume 7710 of *Lecture Notes in Computer Science*, pages 54–67. Springer Berlin Heidelberg, 2012.

[4] P. Bródka, S. Saganowski, and P. Kazienko. GED: the method for group evolution discovery in social networks. *Soc. Netw. Anal. Min.*, 3(1):1–14, 2013.

[5] I. Derényi, G. Palla, and T. Vicsek. Clique percolation in random networks. *Physical Review Letters*, 94(16):160202, 2005.

[6] B. Gliwa, P. Bródka, A. Zygmunt, P. Kazienko, S. Saganowski, and J. Kozlak. Different approaches to community evolution prediction in blogosphere. *CoRR*, abs/1306.3517, 2013.

[7] M. Goldberg, M. Magdon-Ismail, S. Nambirajan, and J. Thompson. Tracking and predicting evolution of social communities. *PASSAT and SocialCom*, pages 780–783, 2011.

[8] D. Jin, B. Yang, C. Baquero, D. Liu, D. He, and J. Liu. Markov random walk under constraint for discovering overlapping communities in complex networks. *CoRR*, abs/1303.5675, 2013.

[9] S. R. Kairam, D. J. Wang, and J. Leskovec. The life and death of online groups: Predicting group growth and longevity. In *WSDM '12*, pages 673–682, New York, NY, 2012. ACM.

[10] K. Konstantinidis, S. Papadopoulos, and Y. Kompatsiaris. *Community Structure and Evolution Analysis of OSN Interactions Around Real-world Social Phenomena*. PCI '13. ACM, Thessaloniki, Greece, 2013.

[11] J. Leskovec and A. Krevl. Stanford large network dataset collection. http://snap.stanford.edu/data, June 2015.

[12] Y.-R. Lin, Y. Chi, S. Zhu, H. Sundaram, and B. L. Tseng. Facetnet: A framework for analyzing communities and their evolutions in dynamic networks. In *WWW08*, pages 685–694, New York, NY, USA, 2008. ACM.

[13] Newman, M E J. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA*, 103(23):8577–8582, 2006.

[14] N. P. Nguyen, T. N. Dinh, S. Tokala, and M. T. Thai. Overlapping communities in dynamic networks: Their detection and mobile applications. *MobiCom*, pages 85–96, 2011.

[15] N. P. Nguyen, T. N. Dinh, Ying Xuan, and M. T. Thai. Adaptive algorithms for detecting community structure in dynamic social networks. In *INFOCOM, 2011 Proceedings IEEE*, pages 2282–2290, 2011.

[16] M. Shahriari, S. Krott, and R. Klamma. Disassortative degree mixing and information diffusion for overlapping community detection in social networks DMID. In *WWW 2015 - Companion Volume*. 2015.

[17] M. Takaffoli, R. Rabbany, and O. R. Zaiane. Community evolution prediction in dynamic social networks. In *ASONAM 2014*, pages 9–16, 2014.

[18] J. Xie and B. K. Szymanski. Towards linear time overlapping community detection in social networks. In *PAKDD 2012*, volume 7302 of *Lecture Notes in Computer Science*, pages 25–36. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[19] J. Xie, B. K. Szymanski, and X. Liu. Slpa: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process. *CoRR*, abs/1109.5720, 2011.