Analysis of Overlapping Communities in Signed Complex Networks

Mohsen Shahriari, Ying Li, Ralf Klamma
Advanced Community Information Systems (ACIS)
RWTH Aachen University
Ahornstr. 55, 52056 Aachen, Germany
{shahriari, yingli, klamma}@dbis.rwth-aachen.de

ABSTRACT

Networks with positive and negative connections have gained popularity in social media. Positive links show trust (friendship) and negative connections are sign of distrust(enmity). An instance of such networks is Wikipedia, in which some contributors are promoted to administrators by other users. Similar to unsigned networks, overlapping communities are pivotal structures, providing many opportunities to reveal useful information on signed networks. Detection of communities in this area is not well studied, which makes it almost impossible to select a good algorithm for recommender or analytic systems. In this paper, we perform a multifaceted analysis of overlapping community detection algorithms and compare them from different aspects; computing modularity, frustration, running time as for the metrics. This work assists social network analysis researchers in selecting a suitable algorithm for their systems, e.g. recommender systems.

Keywords

Signed social networks; overlapping community detection; community analysis

1. INTRODUCTION

Complex dynamical networks contain densely connected components named communities. Communities are important building blocks that reveal informative information about the network. Connections in signed graphs may represent people who are holding friendship or hostility relations or mapping of text meaning to being positive or negative. One may not be able to provide a unique definition of a community, however networks contain clusters which indicate traces of similarity and consolidation. Often density of connections is an important factor in unsigned networks, while balancing theory plays an important role in detection of communities in signed graphs; balancing theory requires edges within communities to be positive while those between communities to be negative [6]. Thus algorithms in unsigned networks cannot be applied directly on signed graphs. Yang et al. proposed an agent-based algorithm in which a node, with higher probability, walk inside its community than across the community border [12]. By examining localized aggregated transition probabilities, communities are detected. Anchuri and Magdon-Ismail devised a two-step approach to maximize modularity and minimize frustration [2]. In their algorithm, all nodes will be initially assigned to two communities according to the leading eigenvector of generalized modularity matrix and those nodes with low absolute value in the eigenvector will then be reassigned if their movement can lead to a higher overall value covering both modularity and frustration. This division is repeatedly carried out until the overall value cannot be improved any more. The multi-objective approach of Amelio and Pizzuti optimizes two objectives simultaneously: one is to raise the density of positive intra-connections and reduce the density of negative inter-connections while the other one is to minimize both negative intra-connections and positive inter-connections. A single best solution is then obtained by choosing either minimum frustration or maximum modularity [1]. All of the above mentioned approaches support community detection but do not detect overlapping structures. To the best of our knowledge, regarding Overlapping Community Detection (OCD), one may only refer to Signed Disassortative degree Mixing and Information Diffusion (SDMID), Signed Probabilistic Mixture model (SPM) and Multiobjective Evolutionary Algorithm for community detection from Signed social Networks (MEA$_s$-SN); these algorithms are demonstrated in the next section. One may notice that we know very little regarding the properties of OCD algorithms in signed networks, which may put scholars to choose appropriate algorithms to their data or recommendation engines. In this regard, we investigate these algorithms with metrics like modularity, frustration and execution times on real world and synthetic networks. Furthermore, they are analysed regarding community distributions, standalone nodes and fraction of overlapping nodes. In most of the experiments, SDMID wins over the other two algorithms.

2. SIGNED OCD ALGORITHMS

In this section, OCD algorithms in signed networks are described.

2.1 SDMID

SDMID as a leader-based technique, applies two fundamental social phenomena named disassortative degree mixing and signed information diffusion. It is a two-phase algorithm. In the first phase, influential nodes are identified through their effective degree and dissimilarity with their neighbours. In the second phase, a cascading process named signed information diffusion is initiated from leaders to calculate the membership of non-leader nodes to central nodes [11].

2.2 SPM

SPM algorithm can be categorized as a version of mixture models to generate connections with specific probabilities. It applies Expected- Maximization (EM) approach which maximizes the probability of latent variables. This algorithm requires an input parameter that needs to be set beforehand; number of communities. ω_{rs} is the probability of an edge e_{ij} choosing a community pair $\{r, s\}$ $(1 \leq r, s \leq k)$ with the constraint $\sum_{rs}\omega_{rs} = 1$. e_{ij} is located in one community if r=s and is between two communities if not. The probability of community r (s) choosing node i (j) is denoted as θ_{ri} (θ_{sj}). For any community r, given n nodes in the network, $\sum_i \theta_{ri} = 1$. As a result, the edge probability by SPM is:

$$P(e_{ij}|\omega,\theta) = \Big(\sum_{rr}\omega_{rr}\theta_{ri}\theta_{rj}\Big)^{A^+_{ij}}\Big(\sum_{rs(r\neq s)}\omega_{rs}\theta_{ri}\theta_{sj}\Big)^{A^-_{ij}}. \tag{1}$$

s.t. $\sum_{e_{ij}\in E} P(e_{ij}|\omega,\theta) = 1.$

Soft partition of a network is generated that is according to the probability of a node belonging to each community. SPM has some deficiencies such as its inability to tackle directed networks and the prime requirement regarding total number of communities [4].

2.3 MEA-SN

Structural similarity is the core part for MEA$_s$-SN as an evolutionary algorithm. The similarity between any two nodes can be calculated as follows:

$$s(u,v) = \frac{\sum_{x\in B(u)\cap B(v)} \psi(x)}{\sqrt{\sum_{x\in B(u)} w^2_{ux}} \cdot \sqrt{\sum_{x\in B(v)} w^2_{vx}}}. \tag{2}$$

where $B(u)$ $(B(v))$ is the set of node $u(v)$ and u's(v's) neighbours and $w_{ux}(w_{vx})$ is the weight of the edge connecting $u(v)$ and x. In this algorithm, two objective functions are employed to maximize positive similarities inside communities and negative similarities outside communities. MEA$_s$-SN is not able to handle directed networks.

3. DATASETS AND METRICS

One may apply both synthetic and real world networks to evaluate community detection algorithms. In our experiments, the evaluation protocol comprises both real world and synthetic networks. Synthetic networks provide the opportunity to test different parameter values and knowing of the ground truth communities is a plus. Moreover, networks can be potentially of large scale. On the contrary, real world networks cover the shortcoming of synthetic networks and they are naturally formed and thus this makes them highly suitable for tests regarding real life problems. To evaluate the algorithms with synthetic networks in signed graphs, we extended the original version of LFR synthetic networks based on a model proposed by Yang et. al [12] [8]. LFR networks comprise various parameters that need to be set. Liu et. al. integrated two new parameters with LFR networks: the fractions of negative connections within communities P_- and positive connections between communities P_+, which adjust the noise level of the concerned synthetic network [10]. The applied real world network is Wikipedia which is an adminship election (wiki-Elec) with 7,194 nodes and about 100,000 edges. Wikipedia is an online glossary that collaborative users contribute to add content to it. Contents are revised and accepted by managers of the Website. Managers or moderators are selected based on online polls, which include positive and negative voting toward the candidates [9].

Metrics applied to evaluate the OCD algorithms include Normalized Mutual Information (NMI), signed modularity, frustration and execution time.

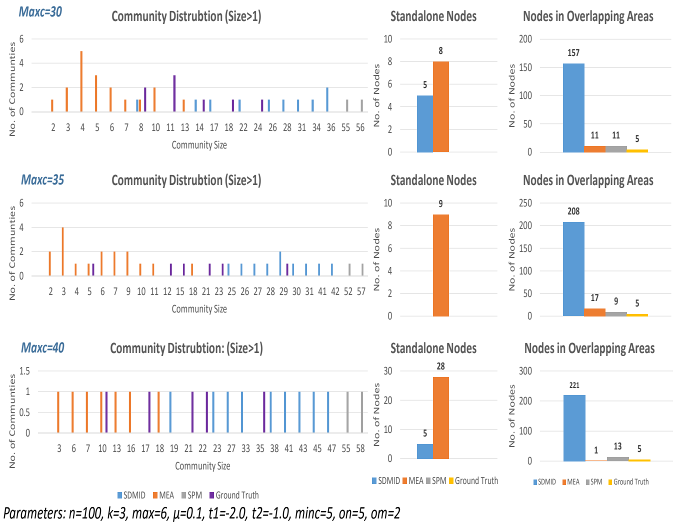

Figure 1: Community structure of detected covers for benchmark networks

3.1 Normalized Mutual Information (NMI)

NMI is a knowledge driven metric which is suitable when the ground truth information is available like experimenting on synthetic networks [8]. The node degrees and edge signs are not important, while it checks whether a node resides in the correct community. Extended NMI computes how much information is shared by two membership vectors; vector output of an OCD algorithm and vector containing the true community values. It ranges from 0 to 1 and a big value suggests a high similarity.

3.2 Signed Modularity

Modularity is a famous metric to evaluate OCD algorithms. We implemented a version of modularity which considers both signed networks and overlapping community structures. The formula for the modularity can be described as follows:

$$Q_{so} = \frac{1}{(2w^+)_e + 2|(w^-)_e|}\sum_i\sum_j\Big[w_{ij} - \Big(\frac{w^+_i w^+_j}{(2w^+)_e} - \frac{w^-_i w^-_j}{2|(w^-)_e|}\Big)\Big]\delta(C_i, C_j). \tag{3}$$

where $w^+(w^-)$ denotes the weighting of effective influence of positive (negative) edges while $w^+_i(w^-_i)$ denotes the total weight sum of all positive (negative) edges incident to node

i. $\delta(C_i, C_j)$ can take a value up to K, which represents the total number of belonging communities.

3.3 Frustration

Whenever three persons having relationships in a signed networks, two of the triangles are balanced and the other two are imbalanced. Or nodes inside communities may form positive relationships and nodes between communities may take negative ones. To keep this in mind, evaluating community detection algorithms can be performed by frsutration metric which computes the number of edges breaking the balancing theory [5]. Correspondingly in this paper, it is defined as the normalized weight sum of negative edges inside communities and positive edges between communities [3]. Similar to modularity measure, number of positive and negative edges refer to effective ones. The frustration metric (frus) can be best described as follows:

$$frus = \frac{\alpha \times |(w^{-}_{intra})_e| + (1-\alpha) \times (w^{+}_{inter})_e}{(w^{+})_e + |(w^{-})_e|} \qquad (4)$$

where α is the weighting parameter, $(w^{-}_{intra})_e$ and $(w^{+}_{inter})_e$ represent the effective weight sum of all negative edges residing within communities and of all positive edges located between communities respectively, and $(w^{+})_e$ and $(w^{-})_e$ are the effective weight sum of all positive edges and of all negative edges respectively. A higher frustration value suggests a higher extent of imbalance of the detected community structure while a lower value implies a better quality of the cover.

4. EVALUATION

We evaluated SDMID, SPM and MEAs-SN on two categories of data; synthetic networks and real world networks. As for synthetic networks, we run the experiments on different range of parameters and at each time, one of those parameters is modified. However, in this paper we contrast them over different properties. A network of 100 nodes is employed for the experiments and the default parameters are set to be as: k=3 (average degree), $maxk$=6 (maximum degree), μ=0.1 (the fraction of edges that each node shares with other nodes outside of its community), t_1= -2.0, t_2= -1.0 (exponents for degree and community size distributions), $minc$=5, $maxc$=30 (minimum and maximum community sizes), on=5 (the number of nodes in overlapping communities), om=2 (the number of communities they belong to), P_-=0.1 (the fraction of negative connections within communities), P_+=0.1 (the fraction of positive connections between communities).

Here we consider some synthetic networks with $maxc$ (30, 35 and 40) as the parameter. Distribution of community sizes, number of standalone nodes and number of nodes residing in overlapping areas are calculated and plotted. Figure 1 indicates that MEA$_\text{s}$-SN tends to identify small-sized communities. On the contrary, SDMID detects large-sized communities, because nodes even with weak connections to a leader are allowed to appear in the respective community. Furthermore, one may notice that SDMID has a large number of overlapping nodes; because of big communities with high overlaps. On the other hand, the benchmark network constitutes medium-sized communities. As for SPM number of communities are limited to two, each of its communities has way larger number of members than those detected by other two algorithms.

Additionally, Figure 1 demonstrates that MEA$_\text{s}$-SN specifies many nodes standing alone, which can be recognized as this algorithm emphasizes the tightness of communities and thus nodes unable to contribute to tightness are separated. Regarding SDMID, a small number of nodes are assigned to none of the communities. As for SPM, no conclusion can be made because it behaves throughout the experiments on benchmark networks as bisection algorithm and therefore all of the nodes are accommodated in either of the two communities.

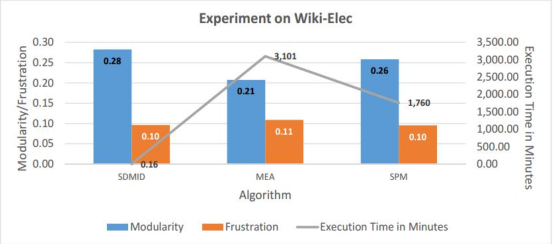

Figure 2: Performance result of the experiment on Wiki-Elec

To analyse a real world network, SPM requires a realistic number of communities, then 5 is selected based on the experiments of Javari and Jalili [7] that make sign predictions in large scale networks. The number of trials of EM algorithm which can be parametrized, is determined to be 3 in order to achieve realistic results. As for SDMID, we set the network coordination game to 3 iterations, leaving out loosely connected nodes for each community. As it can be observed in Figure 2, SDMID achieves superior performance in comparison with its peers in terms of modularity and execution time. SPM obtains the smallest frustration error, which is only slightly better than that of SDMID (0.0953 vs. 0.0962).

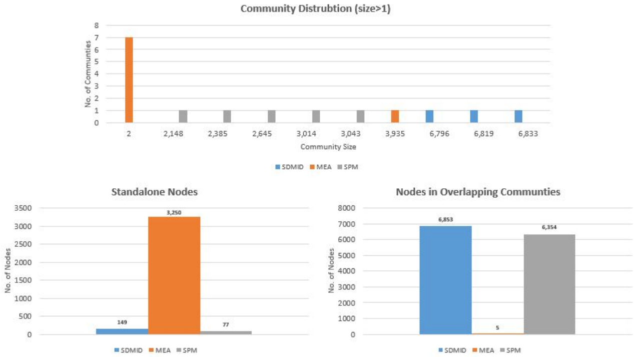

Figure 3: Community distribution in Wiki-Elec

Figure 3 indicates the community distribution of the found covers by the algorithms. One can figure out that MEA$_\text{s}$-SN detects a huge community while assigning large number of nodes to be stand alone. SPM generates the lowest standalone nodes while grouping the remainders in five clusters ranging from 2148 to 3043. SDMID identifies three clusters with approximately uniform sizes. MEA$_\text{s}$-SN only specifies 5 nodes in the overlapping communities, on the contrary, SD-

MID and SPM detect more than 6,000 overlapping nodes, which may reflect the realistic world.

Finally, experiments on synthetic networks are performed based on changing the value of one parameter and keeping the rest unchanged. We average over the results over all the runs and render the results through radar diagram. A radar chart shows the performance of a specific algorithm regarding a single evaluation metric through one of its poles. We normalized the values of the metrics at each poles and therefore higher values indicate better performance. As it can be observed in Figure 4, SDMID achieves the best execution time for both Wiki-Elec and in case of synthetic networks. MEA$_\text{s}$-SN and SPM behave somehow similar and much lower performance in comparison to SDMID. Furthermore regarding modularity, SDMID obtains the best result in Wiki-Elec and synthetic networks, which is followed by SPM and then by MEA$_\text{s}$-SN. Moreover considering frustration metric, one may figure out from radar chart, that SPM achieves the lowest frustration error in Wiki-Elec which is followed by SDMID and MEA$_\text{s}$-SN.

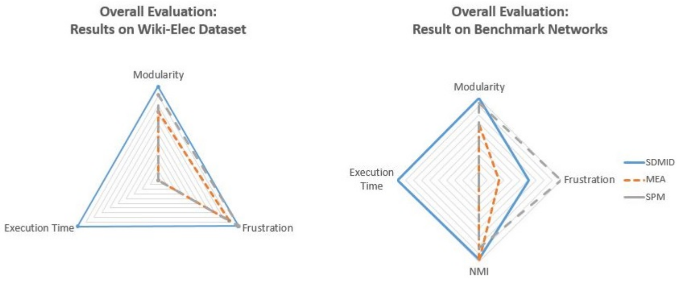

Figure 4: Evaluation of the algorithms with different metrics through radar diagrams.

5. DISCUSSION AND CONCLUSION

Signed social networks obtain increasing importance among researchers while many data resources can be mapped to network of both positive and negative connections. In signed networks, detection of overlapping communities is not well studied, in which it engenders the requirements for the analysis to compare the existing OCD algorithms. To evaluate the algorithms, several measures including modularity, frustration, execution time and NMI are employed. Moreover, fine grained analysis of detected communities comprising community distribution, stand alone nodes and number of overlapping members are enriched with the experiments. In all the experiments, SDMID achieved an obvious advantage regarding execution time. Furthermore, MEA$_\text{s}$-SN had the longest execution time, presumably due its evolutionary nature. Additionally, SDMID achieved the best performance in terms of modularity in most experiments, followed closely by SPM. SPM was superior to the other algorithms regarding frustration while no algorithm demonstrated a constant higher NMI values than the others. Regarding the community distribution, SDMID generated covers with big-sized communities with large areas of overlapping. On the contrary, MEA$_\text{s}$-SN detected small-sized communities with fewer overlaps. Considering the bisection property of SPM, it identifies many nodes assigned to overlapping areas of two large communities. As it can be recognized from this paper, research in signed networks may require OCD algorithms with better running time and precision performance. As for future work, we intend to devise further algorithms in this domain and test them on real world applications such as recommender systems. It also would be interesting to see results from other empirical signed networks.

Acknowledgment

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

6. REFERENCES

[1] A. Amelio and C. Pizzuti. Community mining in signed networks: A multiobjective approach. In *(ASONAM 2013)*, pages 95–99. IEEE/ACM, 2013.

[2] P. Anchuri and M. M. Ismail. Communities and balance in signed networks: A spectral approach. In *(ASONAM 2012)*, pages 235–242.

[3] D. Cartwright and F. Harary. Structural balance: a generalization of heider's theory. *Psychological Review*, 63(5):277–293, 1956.

[4] Y. Chen, X. L. Wang, B. Yuan, and B. Z. Tang. Overlapping community detection in networks with positive and negative links. *Journal of Statistical Mechanics: Theory and Experiment*, 2014(3):P03021, 2014.

[5] P. Doreian and A. Mrvar. A partitioning approach to structural balance. *Social Networks*, 18(2):149–168, 1996.

[6] P. Doriean. Evolution of human signed networks. *Metodološki zvezki - Advances in Methodology and Statistics*, 1(2):277–293, 2004.

[7] A. Javari and M. Jalili. Cluster-based collaborative filtering for sign prediction in social networks with positive and negative links. *ACM Transaction on Intelligent Systems and Technology*, 5(2):24:1–24:19, 2014.

[8] A. Lancichinetti, S. Fortunato, and J. Kertész. Detecting the overlapping and hierarchical community structure in complex networks. *New J Phys*, 11(3):033015, 2009.

[9] J. Leskovec, D. Huttenlocher, and J. M. Kleinberg. Signed networks in social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1361–1370, New York, NY, USA, 2010. ACM.

[10] C. Liu, J. Liu, and Z. Jiang. A multiobjective evolutionary algorithm based on similarity for community detection from signed social networks. *IEEE Transactions on Cybernetics*, 44(12), 2014.

[11] M. Shahriari and R. Klamma. Signed social networks: Link prediction and overlapping community detection. In *(ASONAM)*, Paris, France, 2015.

[12] B. Yang, W. K. Cheung, and J. Liu. Community mining from signed social networks. *IEEE Transactions on Knowledge and Data Engineering*, 19(10):1333–1348, 2007.