YÖNETİM BİLİŞİM SİSTEMLERİ DERGİSİ
http://dergipark.gov.tr/ybs

Yayın Geliş Tarihi: 25.10.2019
Yayına Kabul Tarihi: 03.01.2020
Online Yayın Tarihi: 15.05.2020

Cilt:6, Sayı:1, Yıl:2020, Sayfa: 79-85
e-ISSN: 2630-550X

SBM TOPLULUK TESPİTİNDE TEMEL GERÇEK VE ÜST VERİ İLİŞKİSİ: OKUL ARKADAŞLIK AĞI

Kenan KAFKAS[1], Nazım Ziya PERDAHÇI[2], Mehmet Nafiz AYDIN[3]

[1] Yönetim Bilişim Sistemleri, İşletme Fakültesi, Kadir Has Üniversitesi, İstanbul Türkiye

[2] Enformatik, Mimar Sinan Üniversitesi, İstanbul Türkiye

[3] Yönetim Bilişim Sistemleri, İşletme Fakültesi, Kadir Has Üniversitesi, İstanbul Türkiye

ÖZET

Bilişim Sistemleri araştırmacılarının ilgi alanına giren ağların birçoğunda topluluk yapısına rastlanır. Bu makro ölçekli yapılarda doğal olarak ortaya çıkan toplulukların tespit edilmesi büyük veri kümelerinin yönetilebilir gruplara ayrılması açısından gereklidir. Böylece bu sistemlerin orta ölçekte anlaşılabilir hale gelmesi mümkün olur. Önceki çalışmamızda, Stokastik Blok Modelleme yaklaşımını kullanarak üst veri ve temel gerçeği karşılaştırdık. Bu çalışmamızda, üst veri ile topluluk yapısının ilişkisini ölçebilen bir istatistiksel yöntem olan neoSBM'i bir gerçek dünya arkadaşlık ağı veri seti üzerinde uygulayarak sunuyoruz.

Anahtar Kelimeler: SBM, neoSBM, Community Detection, Best Friends Network

GROUND TRUTH AND METADATA RELATIONSHIP IN SBM COMMUNITY DETECTION: SCHOOL FRIENDSHIP NETWORK

ABSTRACT

Many data sets which are studied by Information Systems researchers involve networks that exhibits community structure. Dividing the large networks into manageable groups (communities) is a crucial first step to understand the network in macro scale. Which then enables the researchers to analyze the data in meso-scale. In our previous work we presented Stochastic Block Model approach and compared the metadata with the ground truth. In present study we introduce a statistical technique called neoSBM that can reveal the relationship between metadata and the community structure on the same real-world school best friendship data set.

Keywords: SBM, neoSBM, Community Detection, Best Friends Network

**Bu çalışma, birinci yazarın ikinci ve üçüncü yazar danışmanlığında hazırladığı doktora çalışmalarından üretilmiş ve 6. Uluslararası Yönetim Bilişim Sistemleri Konferansında (IMISC2019, 9-12 Ekim 2019, Kadir Has Üniversitesi, İstanbul) sunulmuştur.*

**This manuscript is produced from the PHD studies prepared by the first author under the supervision of the second and the third authors and was presented at the 6thInternational Management Information Systems Conference (IMISC2019, 9-12 October 2019, Kadir Has University, İstanbul).*

INTRODUCTION

Many networks of interest to Information Systems researchers exhibit community structure (Chen et al., 2012; Chau & Xu 2012). That is, the structure of the network is such that the nodes in the same blocks are more connected than the nodes in different blocks. "This macro-scale structure is so natural that community detection is an essential task to divide large networked data sets into manageable groups to enable an understanding of a system at the meso-scale" (Perdahci et al., 2017). Among the IS research groups, the Newman modularity criterion (Newman & Girvan 2004) has been the primary tool used for uncovering the community structure of large networked systems (Miranda et al., 2015; Zhang et al., 2016; Perdahci et al., 2017; Golbeck et al., 2017) so far.

Modularity was originally proposed by Newman (2002) as a quantitative measure of network correlation but later on promoted as a panacea for the long-standing graph bisection problem by Bui and Jones (1992). Due to issues such as resolution limit or non-intuitive partitions Good at al. (2010), Fortunato and Barthelemy (2007) different approaches are embraced. One of the prominent methods is Stochastic Block Modelling (SBM). The pioneering work of Holland et al. (1983) about the stochastic block model (SBM), which is coined as classic SBM, takes a completely different approach to the community detection task. In this approach, a dataset is fit into stochastically equivalent blocks based on a Poisson degree distribution. Stochastically equivalent means the nodes in the same block indicate their equivalent roles in generating network structure (Aicher et al., 2015).

Newman suggested that the classic SBM needs to be extended to a slightly more sophisticated model, coined the term Degree Corrected SBM (DCSBM) and demonstrated that this correction successfully fits the real-world datasets into intuitive partition (Karrer & Newman, 2010). A fundamental shortcoming of SBM is that the model requires us to know in advance how many blocks a network contains. To get around this limitation, Riolo et al. (2017) presented a method for estimating the number of blocks in an undirected network. Our previous work (Perdahci et al., 2018) introduced an approach to employ degree-corrected SBM method to a real-world school best friendship network by translating directed nature of connections to multi-edge network. We could not include the second part of our work "relationship between ground truth and metadata" due to conference paper restrictions and concluded the paper by mentioning this situation as a limitation and future work. In this paper we examine the relevance of metadata with the detected communities using SBM on the same school friendship dataset. This time we incorporate the metadata (class information) into the SBM to inspect its relationship with the network structure (ground truth).

The algorithms that detect communities are often evaluated by how well they detect ground truth communities. The ground truth is the connections between nodes, in other words the network itself and the metadata is the attributes of the nodes. In this study, we examine a school friendship network. The ground truth here is the friendship links between students and metadata we use is the class attribute of the students. Treating node attributes or metadata as ground truth is standard practice. However, Peel, Larremore and Clauset (2017) shows that "the metadata are not the same as ground truth. Treating them as such induces severe theoretical problems. For instance, if we assume generating a network that contains a certain community structure is a function, its inverse function i.e. Community detection is not unique. To put differently, it is impossible to uniquely solve an inverse problem when the function to be inverted is not a bijection." This is one of the theoretical problems. Yet they acknowledge that "Community detection remains a powerful tool and node metadata still have value so a careful exploration of their relationship with network structure can yield insights of genuine worth.". Their statistical method called neoSBM help diagnose the relationship between metadata and network structure.

Community detection is a widely employed approach in IS studies for purposes such as market segmentation, recommender systems, product promotion, social media analytics. However, it is worth noticing that the phrase stochastic block has not been used explicitly in flagship IS research publications acknowledged by the Association for Information Systems, including MISQ, Management Science, IS Frontiers, Journal of MIS, Journal of AIS, and Journal of Information Technology. It is likely that the class of community detection methods based on SBM is not used at all and the state-of-the art knowledge of community detection with SBM is yet to be introduced. In the

present work, we employ the neoDCSBM algorithm (a degree corrected extension of neoSBM) to find the relationship between metadata and ground truth using the same real-world best friendship network and compare the new findings with the previous ones. To put in other words, the previous work introduces a novel community detection method to IS community and the present work is an effort to validate and to evaluate the performance of the method by inspecting the relevance of the metadata and the ground truth. Our aim is to present solutions to IS problems with community understanding to establish research capacity for IS community.

METHOD

NeoDCSBM method extends the standard SBM by starting with a given community structure which is the metadata partitions in this case. Then the algorithm gradually changes the community assignments of the nodes and apply standard SBM. A cost function is introduced for varying the communities of the nodes. "As the cost of freeing nodes is reduced, the algorithm creates a path through the space of partitions from metadata to the optimal community partition and, as it does so, we monitor the improvement of the partition by the increase in SBM log likelihood. Beyond direct comparison of the partitions, this method shows how the metadata and inferred community partitions are related" (Peel et al., 2017).

In our real-world best friendship network case, the metadata is the class attribute of the 10th grade students. NeoDCSBM algorithm requires two inputs: the edge list (network itself, the ground truth) and community memberships (metadata). We feed the algorithm with the edge list of the largest component (177 students and 388 edges) and classes (six classes from A to F) of students as the metadata.

NeoDCSBM accepts the class metadata as the initial community structure therefore, each class is accepted as a separate community at the beginning. In the next step the algorithm assigns two states to each node as either "fixed" or "free". Initially all nodes are "fixed" to their classes afterwards, a number of nodes q are assigned as "free" meaning that the community of those nodes This value is used to form a penalty function which will be the cost of freeing a node. Which will keep the number of free nodes q in check while the maximization process continues. Finally, we plot the number of free nodes and neoDCSBM log likelihoods as a function of θ along with the detected community structure to inspect the results and compare with previous findings.

Examining the Results

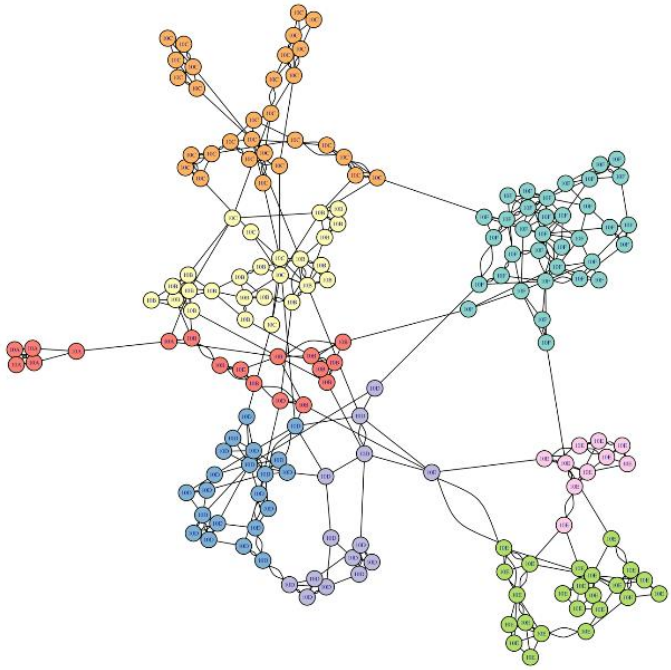
There are two behavior types to be examined in the produced plots:

- "A steady increase indicates neoDCSBM is incrementally refining the metadata partition until it matches the globally optimal SBM communities. This behavior implies that the metadata and community partitions represent related aspects of the network structure.
- A constant log likelihood for a substantial range of θ, followed by a sharp increase or jump indicates that the neoDCSBM has moved from one local optimum to another. Multiple plateaus and jumps indicate that several local optima have been traversed, revealing that the partitions are capturing different aspects of the network's structure." (Peel et al., 2017)

FINDINGS

In our previous work, we found that optimum number of communities was eight (figure 1). However, in this study, the number of communities is constrained by the class metadata therefore, neoDCSBM finds six different communities based on the six 10th grade classes.

Figure 1. Largest component from the previous study.



The minimum number of free nodes required to reach the maximum SBM likelihood is shown in figure 2 as a function of θ. This figure shows how many nodes are let free to maximize the log likelihood (figure 3 below) that indicates a better community fit. As seen on the figure, only one node changes to reach the 2nd local optimal partition in the next figure. Other 2 nodes are let free to reach plateau 3. However, it takes 90 free nodes to reach the global optimum.

Figure 3 shows the log likelihood values as the θ increases. The log likelihood Indicates the relationship between metadata and the inferred community. Each plateau indicates a local optimum partition (community) followed by a global optimum. As Peel et. al. explains; "A steady increase indicates that the neoSBM is incrementally refining the metadata partition until it matches the globally optimal SBM communities. On the other hand, the SBM likelihood remains constant for a substantial range of θ, followed by a sharp increase or jump. Multiple plateaus and jumps indicate that several local optima have been traversed, revealing that the partitions are capturing different aspects of the network's structure." In this case this means that there are dynamics other than class affiliation in the network structure.

Figure 2. Bernoulli prior probability of a node being free q.

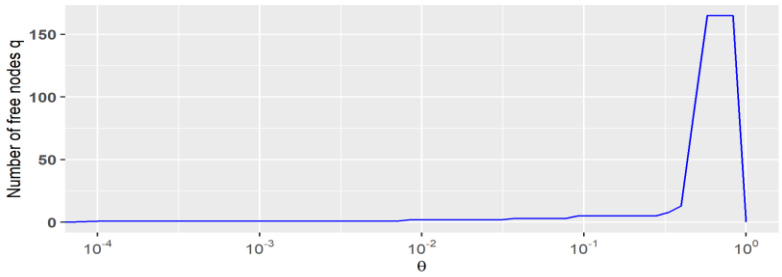
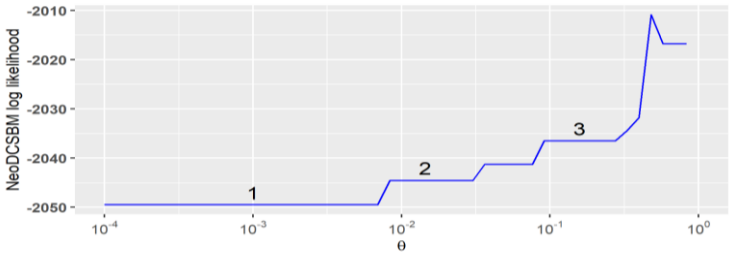


Figure 3. Log likelihood values as a function of θ.



There are three local optimums which can be noticed by the constant SBM likelihood values that remain for a range of θ followed by a peak value indicating the global optimum. Local optimums (plateaus) are indicated with numbers 1,2 and 3. In the plot. The plateau 1 is reached by changing membership of only one student from 10E to 10B. The plateau 2 changes the membership of one more student from 10C to 10B. Plateau 3 adds a student from 10E to a 10D. The other increments of likelihood do not show constant behavior meaning that the search for the optimum is underway.

At the final stage the log likelihood reaches a global optimum after a sharp increase which is achieved by freeing 90 nodes that ends up with 21 students assigned to a different community from the initial assignment. Figure 4 and 5 shows the network maps of the class metadata and the neoDCSBM global optimum respectively which can be interpreted as before and after snapshots of the network community structure. The neoDCSBM algorithm starts from this prior and tries to find stochastically equivalent groups by freeing minimal number of nodes. The algorithm changes the community assignment of a 10E student (1). The second local optimum (2) changes the assignment of a 10C student. And the third local optimum another 10E. However, global optimum changes the communities of 22 students indicated by orange and dark green nodes hinting that the algorithm is capturing a different aspect of network structure.

Figure 4. The network map where each class is accepted as community.

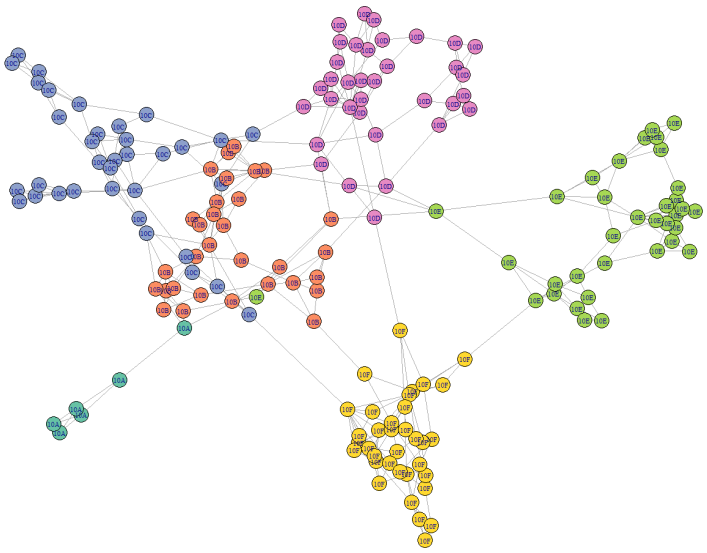
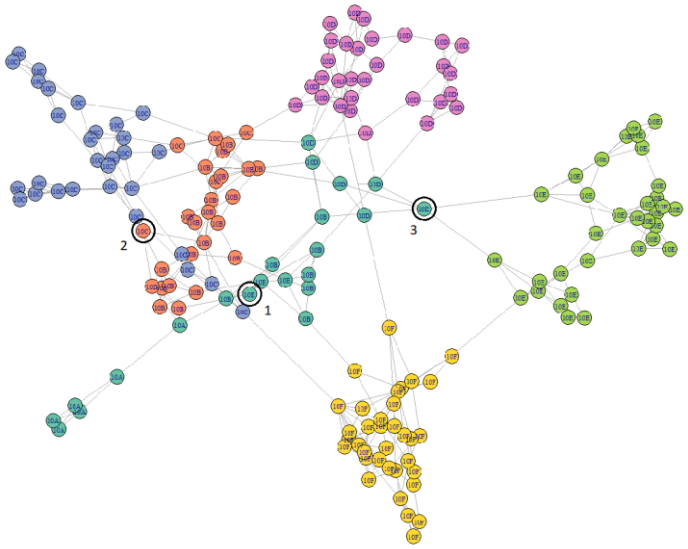


Figure 5. The network map of the neoDCSBM global optimum communities.



DISCUSSION AND CONCLUSIONS

Upon close inspection on the neoDCSBM result (Figure 3 and figure 5), we see that the first local optimum, plateau 1 on figure 3, only takes one student from 10E and puts the student to a community of class 10B which is an intuitive move. This node is indicated as 1 on figure 5. The second optimum also takes only one student, this time from 10C to the same community. The third optimum changes again only one student from 10E to the community of class 10D.

As for the global optimum, we see that the new overall community structure is not far from the initial communities (class metadata). However, there are small yet significant changes implying that the algorithm detects a different aspect of the network structure compared to the metadata. In other words, it detects dynamics that cannot be explained only by the class affiliation in this part of the network. In the beginning for instance, there is a small community which consists of only six students from class 10A (dark green). The small community of this six class 10A students begins to grow as the algorithm searches for optimums. The optimum community structure joins ten students from 10B, four students from 10D and two students from 10E to this community forming a medium sized, highly mixed group. This tells us that the community detection method detects a different aspect of the network structure other than the class metadata. The fact that 10B students mixing with other communities while most of the network remains intact is not a surprise since, this class staged a Shakespearean play that year, making them more popular and social among 10th grades. A limitation for this study is that it involves only class metadata however, in a school friendship context, there are several other student attributes such as gender or test achievement scores. The rest of the class 10A forms the second largest component of the friendship network. Thus, isolated from the largest component. In our previous work (Perdahci et al., 2018), we examined this component which consists of 20 students all from the same class. Since this study is limited to only class metadata and the second largest component involves only a single class, we omit this component from present study hopefully applying the gender metadata as a future work.

These findings agree with the previous paper's findings except that the previous work had higher resolution with eight communities which divided class 10E and 10D to two subgroups. Nevertheless, we see that the friendship network involves a slightly different community structure than class metadata can explain. We can say that neoDCSBM method can be used to statistically diagnose the relationship between metadata and the ground truth. With this in mind, we need to quantify this relationship with a sound statistical method and our research group is working on Blockmodel Entropy

Significance Test (BESTest) which computes the entropy of the SBM that describes the detected partitions (Peel et al., 2017).

REFERENCES

Bui, T. N., & Jones, C. (1992). Finding good approximate vertex and edge partitions is NP-hard. Information Processing Letters, 42(3), 153-159

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: from big data to big impact. MIS quarterly, 1165-1188.

Chau, M., & Xu, J. (2012). Business intelligence in blogs: Understanding consumer interactions and communities. MIS quarterly, 1189-1216.

Fortunato, S., & Barthelemy, M. (2007). Resolution limit in community detection. Proceedings of the National Academy of Sciences, 104(1), 36-41.

Golbeck, J., Gerhard, J., O'Colman, F., & O'Colman, R. (2017). Scaling Up Integrated Structural and Content-Based Network Analysis. Information Systems Frontiers, 1-12.

Karrer, B., & Newman, M. E. (2011). Stochastic blockmodels and community structure in networks. Physical review E, 83(1), 016107.

Miranda, S. M., Kim, I., & Summers, J. D. (2015). Jamming with Social Media: How Cognitive Structuring of Organizing Vision Facets Affects IT Innovation Diffusion. Mis Quarterly, 39(3).

Newman, M. E. (2002). Assortative mixing in networks. Physical review letters, 89(20), 208701.

Newman, M. E., & Girvan, M. (2004). Finding and evaluating community structure in networks. Physical review E, 69(2), 026113.

Newman, M. E. (2006). Modularity and community structure in networks. Proceedings of the national academy of sciences, 103(23), 8577-8582.

Peel, L., Larremore, D. B., & Clauset, A. (2017). The ground truth about metadata and community detection in networks. Science advances, 3(5), e1602548.

Perdahci, Z. N., Aydin, M. N., & Kariniauskaitė, D. (2017). Dynamic Loyal Customer Behavior for Community Formation: A Network Science Perspective.

Perdahci, Z. N., Aydın, M. N., Kafkas, K. (2018) SBM Based Community Detection: School Friendship Network. IMISC2018:Fifth International Management Information Systems Conference.

Zhang, K., Bhattacharyya, S., & Ram, S. (2016). Large-Scale Network Analysis for Online Social Brand Advertising. Mis Quarterly, 40(4).