



Received: February 14, 2020
Accepted: May 06, 2021
Published Online: June 30, 2021

AJ ID: 2021.09.01.MIS.02
DOI: 10.17093/alphanumeric.688660
Research Article

Ground Truth in Network Communities and Metadata-Aware Community Detection: A Case of School Friendship Network

Kenan Kafkas *



Kadir Has University, Istanbul, Turkey, kenankafkas@gmail.com

Nazım Ziya Perdahçı, Ph.D.



Assist. Prof., Department of Informatics, Mimar Sinan Fine Arts University, Istanbul, Turkey, nz.perdahci@msgsu.edu.tr

Mehmet Nazif Aydın, Ph.D.



Assoc. Prof., Faculty of Economics, Administrative and Social Sciences, Kadir Has University, Istanbul, Turkey, mehmet.aydin@khas.edu.tr

* Kadir Has Üniversitesi, Cibali Mah. Kadir Has Cad., 34083, Fatih, İstanbul, Türkiye.

ABSTRACT

Real-world networks are everywhere and can represent biological, technological, and social interactions. They constitute complicated structures in terms of type of things and their relations. Understanding the network requires better examination of the network structure that can be achieved at various scales including macro, meso, and micro. This research is concerned with meso scale for a student best friendship network where sub-structures in which groups of entities (students) take different functions. In this study we address the following research questions: To what extent would NeoSBM as a stochastic process underlie best friendship interaction and in turn ground truth interactions (i.e. reported best friendship)? Do metadata such as gender or class contribute to this understanding? How can one support school managers from a meta-data aware community detection perspective? Our findings suggest that metadata aware community detection can be an effective method in supporting decision-making for class formation and group formation for in and out school activities.

Keywords:

SBM, neoSBM, Community Detection, Best Friends Network.



1. Introduction

Real-world networks are everywhere and can represent biological, technological, and social interactions (Barabási, 2009). They constitute complicated structures in terms of type of things and their relations. Understanding the network requires better examination of the network structure that can be achieved at various scales including macro, meso, and micro. Macro scale is concerned with network measures focusing on structural characteristics at the global scale. For instance, dissemination of ideas between people should be examined if and how the overall structure exhibit typical random or real-world characteristics. Network exhibits common characteristics at meso scale and one of them is that they are composed of sub-structures in which groups of entities take different functions (Chen et al., 2012; Chau & Xu, 2012). For instance, customers of different segments exhibit different shopping behavior from the products or services of a company.

Technically speaking, a network which consists of various communities has a structure such that the nodes in the same blocks are more connected than the nodes in different blocks. "This meso-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., 2018). 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.

Finding out the building groups, or so-called blocks, of a network is an essential step in understanding it. In addition to determining these larger pieces of the system, we want to know the answer to the questions "What processes shape this network? Are there underlying patterns?" The definition of community detection is determining these large-scale structures and the aim of this paper is to make a contribution by finding the answers to above questions. Today, most managers in various areas confront the problems that involve complex systems and community detection is an important part in their toolbox. Although the methods utilized in our research can be applied universally, we demonstrate the methods on a high school best-friendship network. In school context, the environment is a highly interconnected complex system that involves numerous overlapping communities in various sizes. There are many types of actors such as students, instructors, administrators. With each innovation in communications technologies the system becomes more complicated and this poses new threats as well as new opportunities for school managers. Analysis of the communities in the network and combining metadata obtained from Learning Management Systems (LMS) with the communities, can help managers to get insight and to make better decisions. From instructor's perspective, communities are an essential part of teaching. Communities of practice and communities of inquiry are increasingly being applied by researchers and practitioners in higher education online learning (Hopkins, 2017).

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. Our previous work Perdahci et al. (2018) 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.

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 (Bui & Jones, 1992). Due to issues such as resolution limit or non-intuitive partitions (Good et al., 2010; Fortunato & 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 about the students (class and gender) into the SBM to inspect its relationship with the network structure (ground truth).

It is a common practice to evaluate the performance of a community detection algorithm by its accuracy in finding the "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 are the class and the gender attribute of the students. Treating node attributes or metadata as ground truth is standard practice. However, Peel,

Larremore and Clauset (2017) claim 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 helps us diagnose the relationship between metadata and network structure.

To put it differently, the ground truth that lies underneath the network structure is not directly connected with the metadata, however, metadata are not completely irrelevant. NeoSBM helps us to find out the relationship if it exists. In this study we address the following research questions: To what extent NeoSMB as a stochastic process would underlie best friendship interaction and in turn ground truth interactions (i.e. reported best friendship)? Do metadata such as gender or enrolled class contribute to this understanding? How can one support school managers from a metadata aware community detection perspective? It is shown that inspecting the relationship between ground truth and metadata can give managers a competitive advantage by providing them with further insight into understanding their environment.

2. Metadata-Aware Community Detection: NeoSBM

The dataset is a best friendship network of 10th graders in a high school. We conducted a simple survey asking students to self-report three of their best friends. By connecting best friends together, we formed the friendship network. We also obtained the class, gender and GPA grades of the students from the Learning Management System. There are 209 10th graders in six classes in the school. The friendship network comprises of 620 best friendship ties in total however, we are focusing on the two largest connected components of the network which comprise of 177 students containing 388 friendship connections and 20 students containing 36 friendship connections, respectively. The remaining four connected components are all four-student groups, and each is accepted as a community by themselves. In our real-world best friendship network case, the metadata are the class and gender 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 and classes (six classes from A to F) of students as the metadata.

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 (Eq. 1) 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” (Peel

et al., 2017). Beyond direct comparison of the partitions, this method shows how the metadata and inferred community partitions are related.

Peel et al. (2017) provided the detailed explanation of the algorithm in the supplementary materials section of the paper along with the Python code. To efficiently examine the results, we have modified the code keeping the algorithm section intact. Visually inspecting the resulting communities is an essential part of our process therefore, we have developed a shiny application using the RStudio platform (Allaire, 2012) that plots the network map of selected likelihoods that is determined by the algorithm. That way we were able to visually examine the communities following the path that algorithm took to get to the global optimum.

NeoDCSBM accepts the class metadata as the initial community structure therefore, each class is accepted as a separate community at the beginning. The same is applied to the gender metadata afterwards. 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 can be chosen by the neoDCSBM model. The model calculates the number of free nodes by limiting its value with a cost function using the θ parameter which is a Bernoulli prior. This value is used to form a penalty function which will be the cost of freeing a node and 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 (Perdahci et al., 2018).

$$L_{neo}(G; \pi, z) = \prod_{ij} W_{\pi_i \pi_j}^{A_{ij}} (1 - W_{\pi_i \pi_j})^{(1-A_{ij})} \prod_i \theta^{\delta_{z_i, r}} (1 - \theta)^{\delta_{z_i, b}} \quad (1)$$

Variable	Definition
G	A network
A_{ij}	The number of edges between nodes i and j , $A_{ij} \in \{0,1\}$
W_{rs}	The probability of an edge between nodes in groups r and s
π	A partition of nodes into groups
z	Neo-state indicator variable $z_i \in \{b, r\}$
θ	Bernoulli prior probability parameter
L_X	Log likelihood L of model X
$\delta_{a,b}$	The Kronecker delta $\delta_{a,b} = 1$ for $a = b$; $\delta_{a,b} = 0$ for $a \neq b$;

Equation 1. Likelihood function of NeoSBM, first product is the standard SBM, second product is the penalty function.

In general, there are two behavior types to be examined in the produced plots (Peel et al., 2017):

- “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”.

3. Findings

3.1 Class Metadata (the Largest Component)

(Perdahci, et al., 2018)'s simulations on estimating the number of communities based on (Riolo et al., 2017)'s algorithm for the largest network component results in eight communities (Figure 1). However, constrained by the class metadata, neoDCSBM finds six different communities based on the six 10th grade classes.

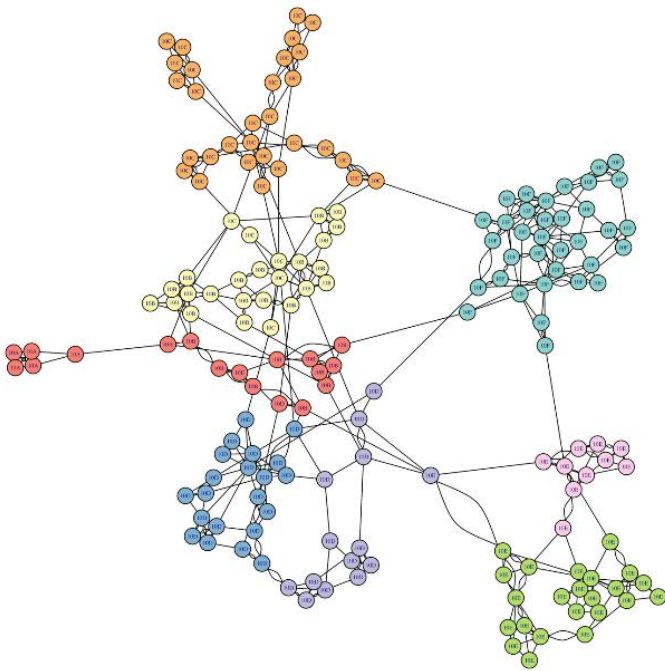


Figure 1. The network map of the largest component from our previous study. The DCSBM detects eight different communities in the largest component mostly explained by class affiliation.

The minimum number of free nodes required to reach the maximum SBM likelihood is shown in Figure 2 as a function of θ . Figure 3 shows the log likelihood values as a function of θ . The log likelihood is an indication of relationship between detected community partitions and the metadata.

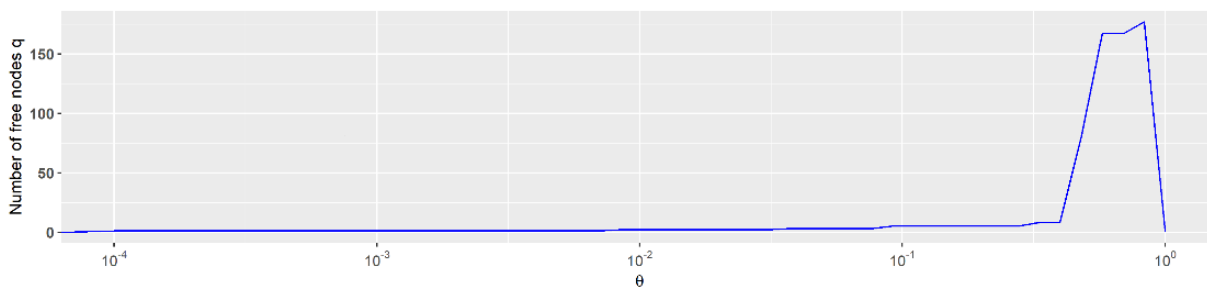


Figure 2. Number of free nodes q as a function of θ (Bernoulli prior probability of a node being free)

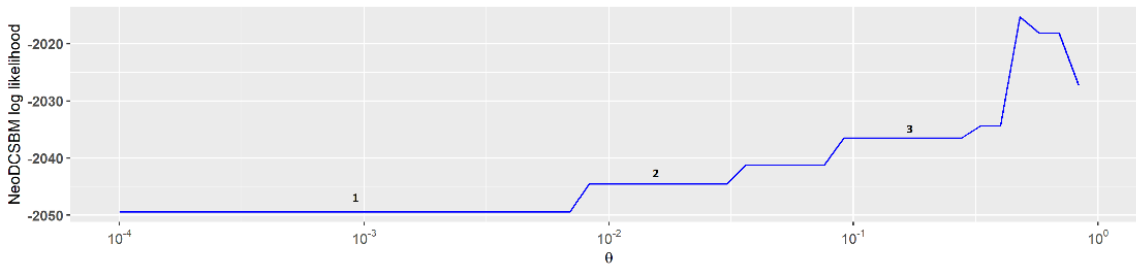


Figure 3. Log likelihood values as a function of θ (Bernoulli prior probability parameter)

There are three local optimums which can be noticed by the constant SBM likelihood values that remain for a substantial 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

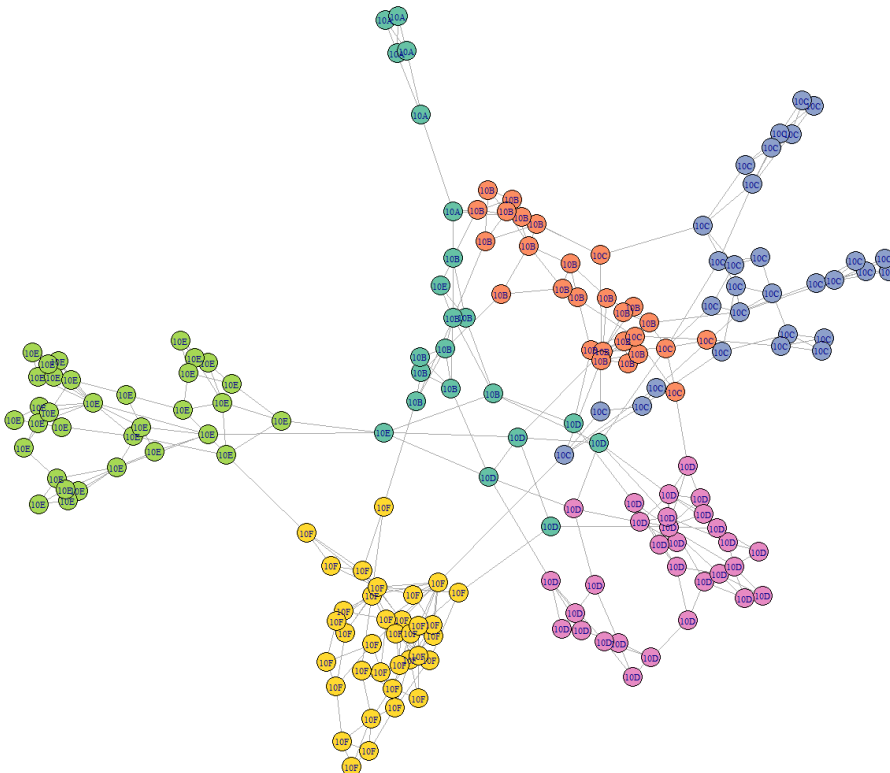


Figure 4. The network map where each class is accepted as community

The neoDCSBM algorithm start from this prior and tries to find stochastically equivalent groups by freeing minimal number of nodes.

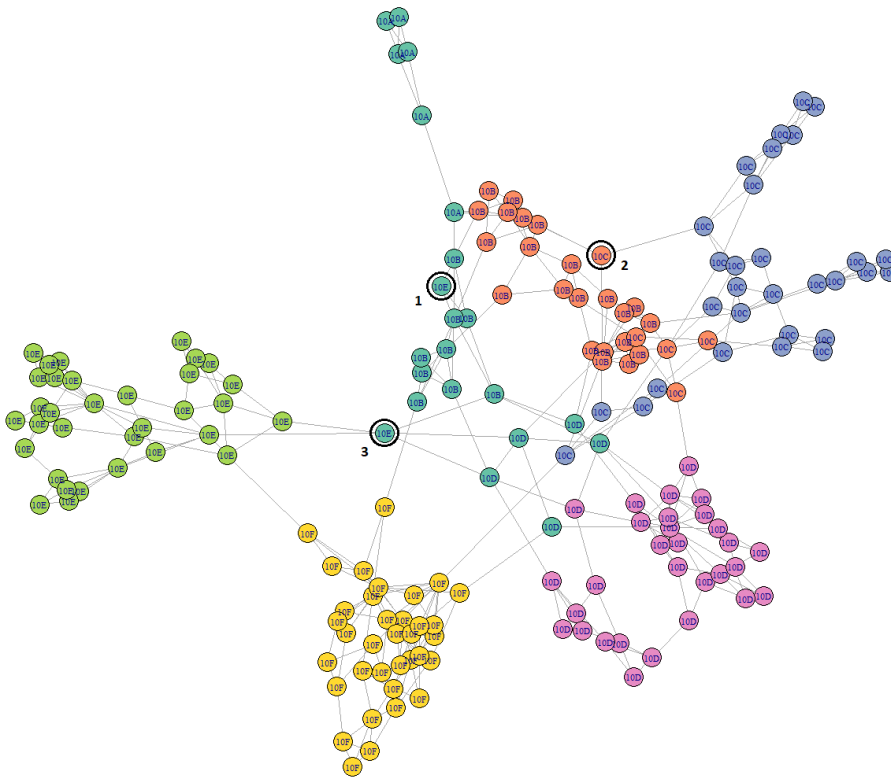


Figure 5. The network map of the neoDCSBM global optimum communities.

3.2 Gender Metadata (the Largest Component)

Gender is another metadata we have examined. Naturally, there are only two community partitions regarding the metadata as male and female. The likelihood plot in figure 6 shows that the NeoDCSBM algorithm finds two local optimums represented by plateaus 1 and 2. Although the increases in gender path is not sharp as the ones in class plot, once the local optimum is reached, the likelihood remains constant for a long interval of theta. This behavior indicates that the metadata and the community partitions are capturing different aspects of the network structure.

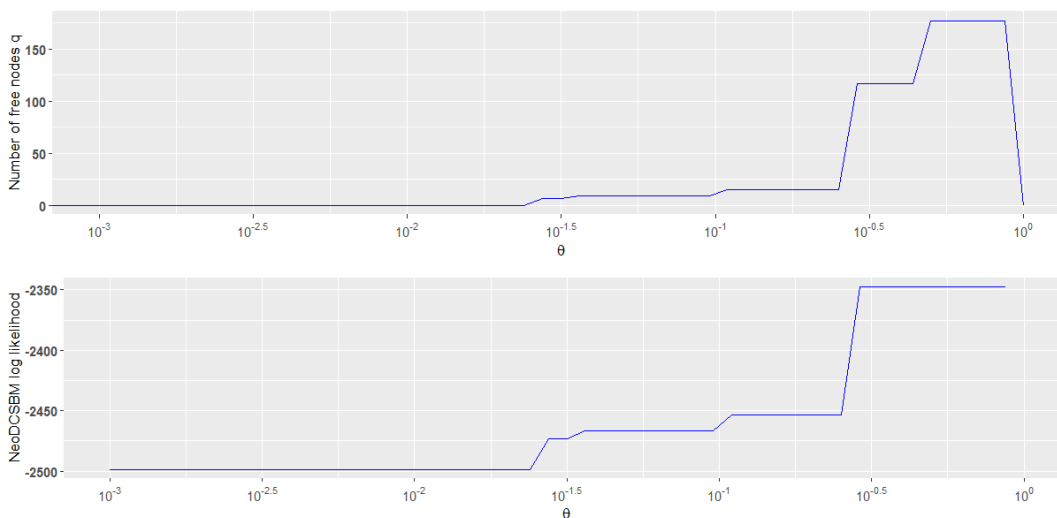


Figure 6. Number of free nodes and the log likelihood values of gender metadata of largest component as a function of θ (Bernoulli prior probability parameter)

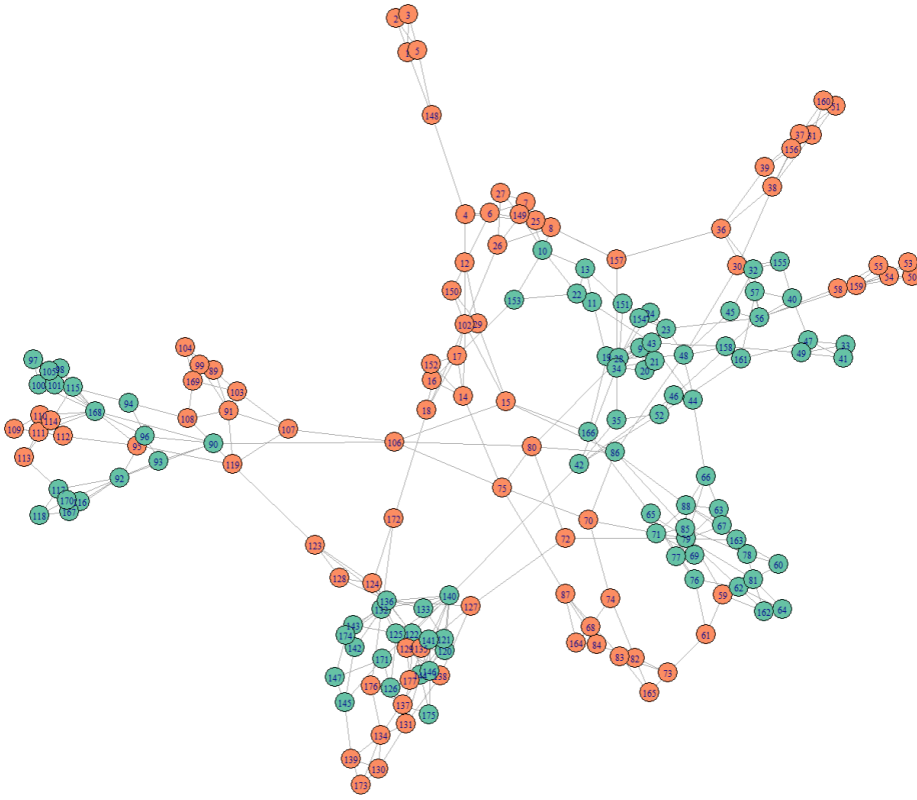


Figure 7. NeoDCSBM second local optimum for gender metadata

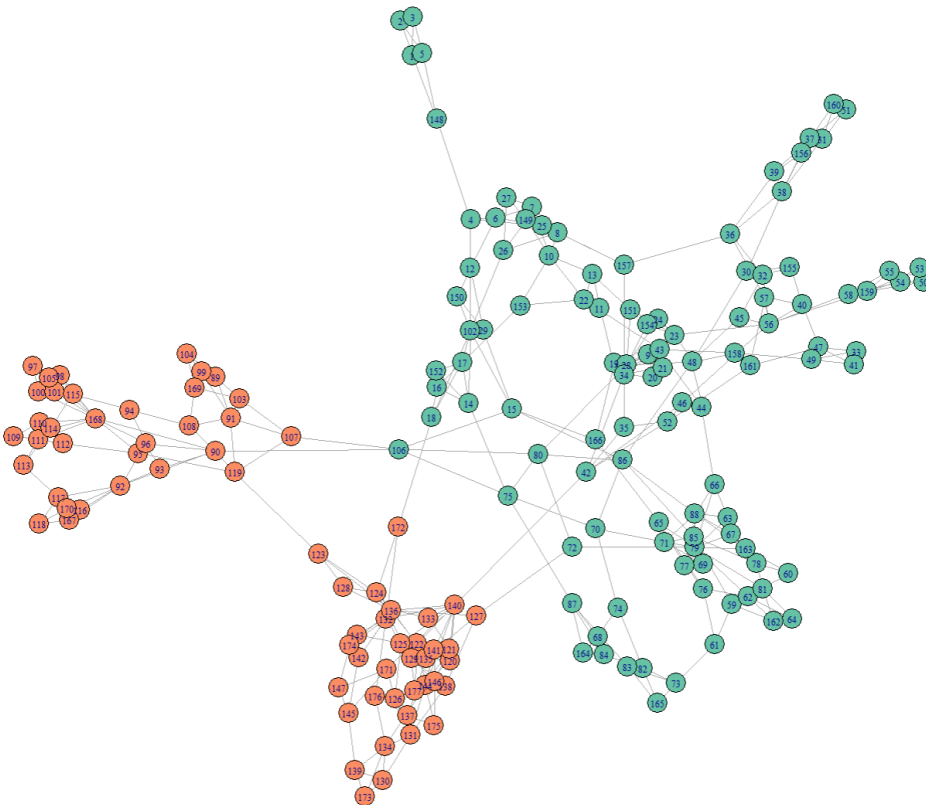


Figure 8. NeoDCSBM global optimum for gender metadata

Figure 7 is a network map of the second local optimum of the largest component based on gender metadata. Despite being highly mixed in terms of gender, when it comes to separating network into two partitions based on connections NeoDCSBM here divides mostly the classes into two groups. Figure 8 is a network map of the global optimum of the largest component based on gender metadata. Here algorithm draws a clear line between two communities where only five of the nodes of the community on the left (orange nodes) and four nodes of the community on the right (green nodes) have ties outside the community they belong to.

3.3. Gender Metadata (Second Largest Component)

Figure 9 shows the likelihood plot of the second largest component. There is only one local optimum before the global optimum. The algorithm performs community detection in three stages; metadata where the gender determines the communities (figure 10-a, orange nodes are female and green nodes are male), local optimum which is a contribution of NeoSBM and lastly, the global optimum which is the result of the standard degree corrected SBM. As seen in table1, local optimum detects a new community partitioning by swapping membership of two students who have more connections to opposite gender. (figure 10-b). As the algorithm reaches the global optimum membership of two groups each consisting of four students is swapped between two communities (figure 10-c).

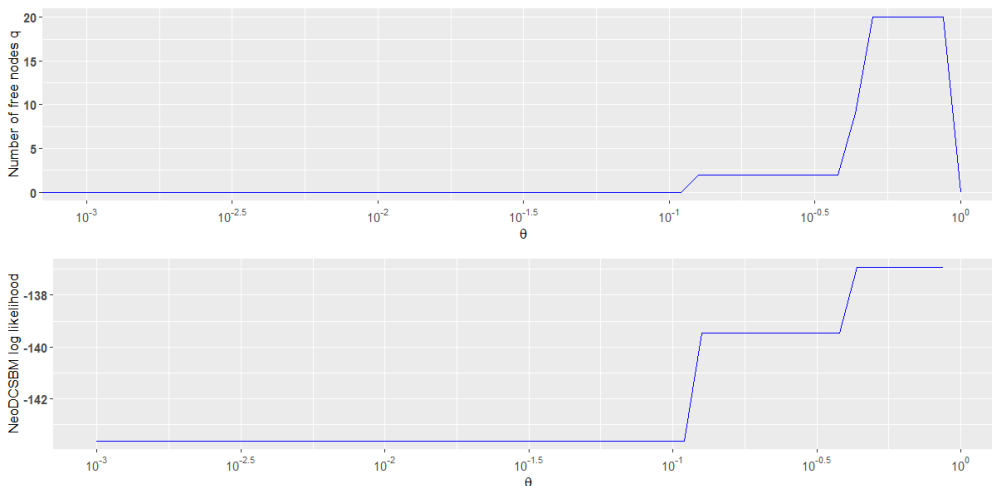
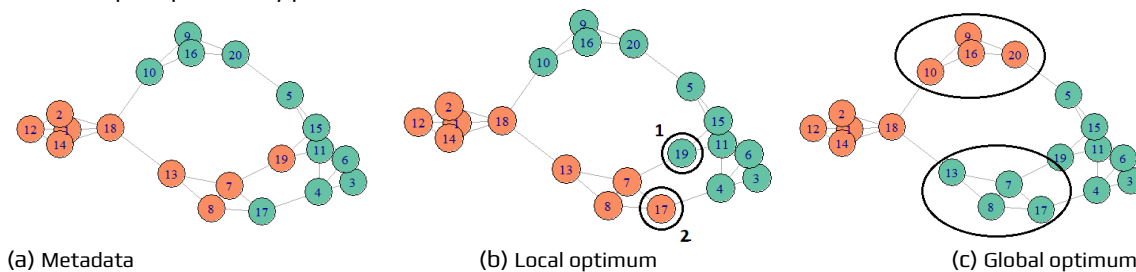


Figure 9. Number of free nodes and log likelihood values of gender metadata of the second largest component as a function of θ (Bernoulli prior probability parameter)



(a) Metadata

(b) Local optimum

(c) Global optimum

Figure 10. Metadata, local optimum and global optimum community network maps of second largest component based on gender metadata

To further examine the second largest component, NeoSBM detects two optimal partitions. In what follows, C1 will denote the optimal partition, that is the partition with maximal NeoDCSBM likelihood, and C2 will denote the runner-up optimal partition.

	C1		C2		Remark on the partitions observed
	F	M	F	M	
Metadata	9	0	0	11	Four nodes within the two partitions (ground-truth communities) have links to the rest of the network
Local Optimum	8	1	1	10	Three nodes within the two partitions (communities) have links to the rest of the network
Global Optimum	5	4	4	7	Only two nodes within the two partitions (communities) has links to the rest of the network

Table 1. Community membership distributions based on gender metadata of the students

Considering the gender metadata, the second best NeoDCSBM partition seems to make a better job; however, the best partition is stronger in the sense that fewer of the nodes within communities have links to the rest of the network. As the algorithm progresses number of nodes that connect communities decreases.

A closer inspection reveals that local optimum and global optimum partitions are more alike than meets the eye: one can assemble these two partitions from the same “building blocks” (Riolo & Newman, 2019) which are $B_1(4, 0)$, $B_2(0, 5)$, $B_3(1, 3)$, $B_4(6, 1)$ where the notation used is $B_i(f, m)$. B stands for “building block”, f: number of females, m: number of males, i: building block index which runs from 1 to 4. With this notation:

$$\text{Global Optimum} = \{C_1 = \{B_1, B_2\}, C_2 = \{B_3, B_4\}\}$$

$$\text{Local Optimum} = \{C_1 = \{B_1, B_4\}, C_2 = \{B_2, B_3\}\}$$

where Cs denote communities. One can easily recognize that although not as strong as class metadata, gender metadata can be considered the ground truth for these four building blocks.

4. Discussion and Conclusion

Upon close inspection on the NeoDCSBM result for the class metadata (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) where the rest of the class is an isolated component which is the second largest in the network with 20 students. 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. This activity might have helped them break the boundaries of their class (metadata) and form out-group friendship relations (ground truth). Such activities can be noticed by many instructors or managers however, there are several dynamics that might not be noticed so easily. Metadata-aware community detection enables such discoveries supporting decision makers.

(Van Geel et al., 2016) state that data-driven decision making by school managers can improve student achievements. Metadata-aware community detection offers an understanding of complex systems data to support school managers' decision-making efforts. Managerial decisions in a school involves various group activities such as forming study groups, arranging field trips, assigning class memberships and lab partnerships and so on. Often, managers or instructors make these decisions intuitively or based on simple classifications such as gender, academic success, even sometimes by choosing students randomly. With metadata-aware community detection, a relevant metadata can be chosen from LMS and can be used to arrange appropriate student groups for intended activities. For instance, when assigning students to classes, class metadata with a friendship network can be used. Alternatively, while deciding lab session members or study groups, a collaboration network (students who took same class, or worked on the same project) can be utilized by the administrators.

Looking at the findings of our case study in terms of class, administrators of this particular school can see that while classes 10F completely and 10E almost completely can be defined by their class membership, classes 10B and 10A are not only defined by their classes. If school aims to encourage social interactions among students, school managers and instructors can target the classes 10F and 10E. Aytac (2015) state that school managers' lack of Talent Management (TM) skills result in low level of commitment by the teachers. Instead of a student friendship network, examining a teacher's collaboration network can provide valuable insights to managers to improve their TM skills.

As for the managerial Implications of second largest component, Newman et al. argues that the building blocks are largely invariant with respect to a select community detection algorithm. If that is the case, investigating the building blocks should be as important, if not more important, as community detection.

These findings agree with the literature e.g., Perdahci et al. (2018) 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. One 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

- Allaire, J. (2012). RStudio: integrated development environment for R. Boston, MA, 537, 538.
- Aytac, T. (2015). The relationship between teachers' perception about school managers' talent management leadership and the level of organizational commitment. *Eurasian Journal of Educational Research*, 15(59), 165-180.
- Bui, T. N., & Jones, C. (1992). Finding good approximate vertex and edge partitions is NP-hard. *Information Processing Letters*, 42(3), 153-159
- Barabási, Albert-László (2009) "Scale-free networks: a decade and beyond." *science* 325, no. 5939 412-413.
- 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.
- Hopkins, M. (2017). A Review of social network analysis and education: Theory, methods, and applications. Karrer, B., & Newman, M. E. (2011). Stochastic block models 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.
- Riolo, M. A., & Newman, M. E. J. (2019). Consistency of community structure in complex networks. *arXiv preprint arXiv:1908.09867*.
- 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., Aydın, M. N., Kafkas, K. (2018, October) SBM Based Community Detection: School Friendship Network. Paper presented at the Fifth International Management Information Systems Conference.
- Perdahci, Z. N., Aydın, M. N., & Kariniauskaitė, D. (2017). Dynamic Loyal Customer Behavior for Community Formation: A Network Science Perspective.
- Van Geel, M., Keuning, T., Visscher, A. J., & Fox, J. P. (2016). Assessing the effects of a school-wide data-based decision-making intervention on student achievement growth in primary schools. *American Educational Research Journal*, 53(2), 360-394.
- Zhang, K., Bhattacharyya, S., & Ram, S. (2016). Large-Scale Network Analysis for Online Social Brand Advertising. *Mis Quarterly*, 40(4).

