SBM Based Community Detection: School Friendship Network

Ziya Nazım Perdahcı[1*], Mehmet Nafız Aydın[2], Kenan Kafkas[2]

1 Mimar Sinan Fine Arts University, **2** Kadir Has University, * Corresponding author, nazim.ziya.perdahci@msgsu.edu.tr

Abstract

Many networks of interest to Information Systems researchers exhibit community structure. 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. In the present work, we aim at introducing this elegant approach as the state-of the-art knowledge to the IS research community, manage to extent the method to multi-edge networks, and apply successfully the extended method to a real-world school best friendship context.

Keywords: SBM, Community detection, Best friends network; School management.

Citation: Perdahcı, 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.

Editor: H. Kemal İlter, Ankara Yıldırım Beyazıt University, Turkey

Received: August 19, 2018, **Accepted:** October 18, 2018, **Published:** November 10, 2018

IMISC 2018 PAPER

SBM Based Community Detection: School Friendship Network

Ziya Nazım Perdahçı[a], Mehmet Nafiz Aydın[b], Kenan Kafkas[b]

[a]Mimar Sinan Fine Arts University
[b]Kadir Has University

Abstract

Many networks of interest to Information Systems researchers exhibit community structure. 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. In the present work, we aim at introducing this elegant approach as the state-of the-art knowledge to the IS research community, manage to extend the method to multi-edge networks, and apply successfully the extended method to a real-world school best friendship context.

Keywords

SBM; Community Detection; Best Friends Network; School Management.

Stokastik Blok Modelleme Tabanlı Topluluk Tespiti: Okul Arkadaşlık Ağı

Özet

Yönetim 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. Bu çalışmada, topluluk tespiti yöntemi çok-kenarlı ağlara genişletilerek, aynı yöntem okulda yakın arkadaşlık ağına başarılı bir şekilde uygulanmış ve bu yeni yaklaşım Bilişim Sistemleri araştırmacıları topluluğuna sunulmuştur.

Anahtar Kelimeler
Stokastik Blok Modelleme, Topluluk Tespiti, Yakın Arkadaşlık Ağı, Okul Yönetimi

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. 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 (Bui & Jones, 1992). The class of community detection algorithms relying on the modularity maximization (Newman, 2006) is not without issues as (Good at al., 2010) successfully demonstrated that in some cases the optimal partition detected may not correspond to the intuitive one. Yet another widely discussed issue is the resolution limit of modularity optimization (Fortunato & Barthelemy, 2007). Modularity offers, for all practical purposes, an initial understanding of a network macro-scale structure, however, we believe that it is time to embrace a different approach.

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 Possion degree distribution. Stochastically equivalent means the nodes in the same block indicate their equivalent roles in generating network structure (Aicher et al., 2015). The idea of blocks is attributed to such terms as “groups”, “communities”, “clusters” in various research domains. Given that the connectivity structure of the nodes in the real-world networks follow power law degree distributions (Barabasi, 2009), 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 using Bayesian inference along with a Monte Carlo sampling scheme.
In the present work our contributions are three-fold. First, we aim at introducing this elegant approach as the state-of the-art knowledge to the IS research community. Second, we manage to extent the method to multi-edge networks. Third, we successfully apply the extended method to a real-world school best friendship context.

Background

Lack of SBM in IS Research

Essentially the very idea of community detection has been adopted in two ways: contributes to establishment of theoretical accounts in connection with other reference disciplines, and serves as a means to solve information management related real-world problems. For the former, to give an example, one can see that the community issue takes place in the context of organizational issue that eventually leads to community effect on institutionalized cognitive structures (Miranda et al., 2015). Even though the attempt to elegantly use community detection for theory construction is worth noticing in (Miranda et al., 2015), its validity leaves one in doubt as numerical Newman modularity maximization reaches a plateau of 4-to-14 cluster solutions, from which the authors arbitrarily select 4 and discard single-node partitions arbitrarily.
For the latter, numerous exemplary studies can be given to see how researchers tried to employ modularity for community detection despite its limitations. To give two of the recent works that are of interest to this paper we can refer to (Zhang et al. 2016; Golbeck et al., 2017). Zhang et al. (2016) propose a hierarchical community detection method for customer segmentation in an undirected brand–brand network. Golbeck et al. (2017), on the other hand, employ Gephi (Bastian et al., 2009) to detect communities in a big data context where one can set the resolution to lower values to get more (smaller in size) communities.
To sum up, in IS Research where the community detection is the focused research issue, modularity maximization is widely used, 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 SMB is yet to be introduced.

School Friendship Context

To be aware of perceived school experience of students is of paramount importance to school managers, which is an essential part of school climate (Simon et al., 1996; Pashiardis, 2000) or school culture (Marcoulides at al., 2005). It has been argued that measures of classroom and friendship may enhance managers' ability to draw conclusions about the relationship between school belonging and educational achievement (Goodenow & Grady, 1993). It is tempting to probe for this information by surveying students' perception with questions such as "How was the school?" or "How did you feel about the school?" The scope of the survey, however, would not reflect beyond individual experiences. School managers naturally desire to be aware of an extent to which school-wide friendship exists and can be sustained. Of particular importance among the types of friendships is the "best friends" networks as such choices are much more stable (Leenders, 1996) and asymmetric (Ball and Newman, 2013). In the present work, we conduct community detection in a best friendship network on a real-world (anonymized) dataset collected from a high-school 10th grade students by employing a standard name generator technique (Kudaravalli et al., 2017).

Methods

The Monte Carlo scheme of Riolo et al. (2017) and the C code thereof are for undirected networks where ties represent symmetric relations. Name generator processes are, however, inherently asymmetric; best friend choices may not always be symmetric. The network model to represent asymmetric relations best is to use a directed network. The next best model would be a multi-edge undirected network where each link between nodes represent a choice of a best friend. To test this idea we prepared two toy networks and fed them into the program (see Table 1, the first two columns). Clearly, The DCSBM community detection along with the Monte Carlo scheme gives different results for the toy networks. Armed with these findings, we prepared a multi-edge undirected best friendship network.
Knowing the fact that SBM community detection is limited to connected components (Abbe & Sandon, 2015), we deployed the Monte Carlo scheme separately on the two larger components of the network. The rest of the four network components are four-cliques which are deemed to comprise closed communities of best friends.

Practical Application of the Riolo et al. Method for Community Detection

We propose the following practical strategy for identifying communities
Step 1. For directed networks only: Convert the network component(s) into multi-edge undirected network component(s)
Step 2. Prepare a Graph Modelling Language File (GML) for the component(s) beyond the cliques.
Step 3. Add a subroutine to the original C source code provided by Riolo et al. (2017) to print out communities.
Step 4. Compile the C code with the following constants:
Number of Monte Carlo sweeps performed: 1,100,000
Rate at which sweeps sampled: 100
Maximum number of groups: (To be decided) Depends on the network at hand, explore the number of groups by initially setting it to a large value like 100, plot the histogram of the number of groups, note where the histogram peaks, as a rule of thumb set the maximum number of groups to around twice the peak value of groups.
Step 5. Compile the C code again after setting the maximum number of groups in 4th step.
Step 6. Run the C code successively 100 times, delaying the next run by 10-15 sec to allow for the Monte Carlo simulations to start from a considerably different seed.
Step 7. Plot the histogram of the number of groups for the aggregated list of partitions obtained in 6th step.
Step 8. Spot the peak bin on the histogram, search for the maximum likelihood community structure within the largest bin.

INSERT FIGURE 1 HERE

Findings and Discussion

Table 1 summarizes the findings for DCSBM community detection.

INSERT TABLE 1 HERE

For the largest network component of 177 students, DCSBM community detection method identifies eight stochastically equivalent blocks which vary in size in such a way that lets us categorize them into three groups: small, medium, and large. Four of them are of medium size having 18, 21, 23, and 26 students, two of them are of small size having 10 and 14 students, and another two of them are of large size having 30 and 35 students.
For the second largest network component of 20 students, DCSBM community detection identifies four stochastically equivalent blocks. Three of these communities are equally sized having four students each while one of them is relatively larger having seven students.
In the literature, it is a standard practice to treat node metadata as though they were ground-truth communities and use them to make sense of the communities detected. For this purpose, we have obtained two metadata for each node from the School Information System which are the gender and the classroom of each student.
According to the classroom metadata, the best friends network exhibits a pattern that we can call classmate community. Ten of the twelve communities contain classmates. That number includes community no 2, because it has only one nonclassmate. The other pattern the network exhibits is cross-class community. The rest of the communities belong to this group, one having students from two classes, the other having students from four classes.
According to gender metadata the network exhibits three remarkable patterns. The large size communities have equal number of the two gender types, in the communities of medium size male students outnumber female students with the exception of community no 3 and in the small size communities female students outnumber male students to the extent that we can call them female-clubs.
As far as school-wide interactions are concerned, we contend that out of all these stochastically equivalent blocks the one that should appeal to the school managers more are the two cross-class communities, because they are formed beyond the physical limitations of class boundaries.

INSERT FIGURE 2 HERE

Conclusion

In this paper, we have described how Riolo et al. work (2017) can be put into practice in a real-world scenario where the method is extended to multi-edge undirected networks. The primary limitation of DCSBM community detection is that it does not incorporate metadata into the fitting procedure. To this end, our research group is aware of Peel et al.'s (2017) extended model, which the authors call neoSBM, and the work is underway to fit the network data into their model for a better understanding of the relation between SBM communities and the students' metadata.

References

Abbe, E., & Sandon, C. (2015, October). Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on* (pp. 670-688). IEEE.

Ball, B., & Newman, M. E. (2013). Friendship networks and social status. *Network Science*, *1*(1), 16-30.

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. *Icwsm*, *8*(2009), 361-362.

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.

Goodenow, C., & Grady, K. E. (1993). The relationship of school belonging and friends' values to academic motivation among urban adolescent students. The Journal of Experimental Education, 62(1), 60-71.

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

Kudaravalli, S., Faraj, S., & Johnson, S. L. (2017). A Configural Approach to Coordinating Expertise in Software Development Teams. *MIS Quarterly*, *41*(1).

Leenders, R. T. A. (1996). Evolution of friendship and best friendship choices. *Journal of Mathematical Sociology*, *21*(1-2), 133-148.

Marcoulides, G. A., Heck, R. H., & Papanastasiou, C. (2005). Student perceptions of school culture and achievement: Testing the invariance of a model. *International Journal of Educational Management*, *19*(2), 140-152.

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., & 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.

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

Pashiardis, G. (2000). School climate in elementary and secondary schools: Views of Cypriot principals and teachers. *International Journal of Educational Management*, *14*(5), 224-237.

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

Simon, S. J., Grover, V., Teng, J. T., & Whitcomb, K. (1996). The relationship of information system training methods and cognitive ability to end-user satisfaction, comprehension, and skill transfer: A longitudinal field study. *Information Systems Research*, *7*(4), 466-490.

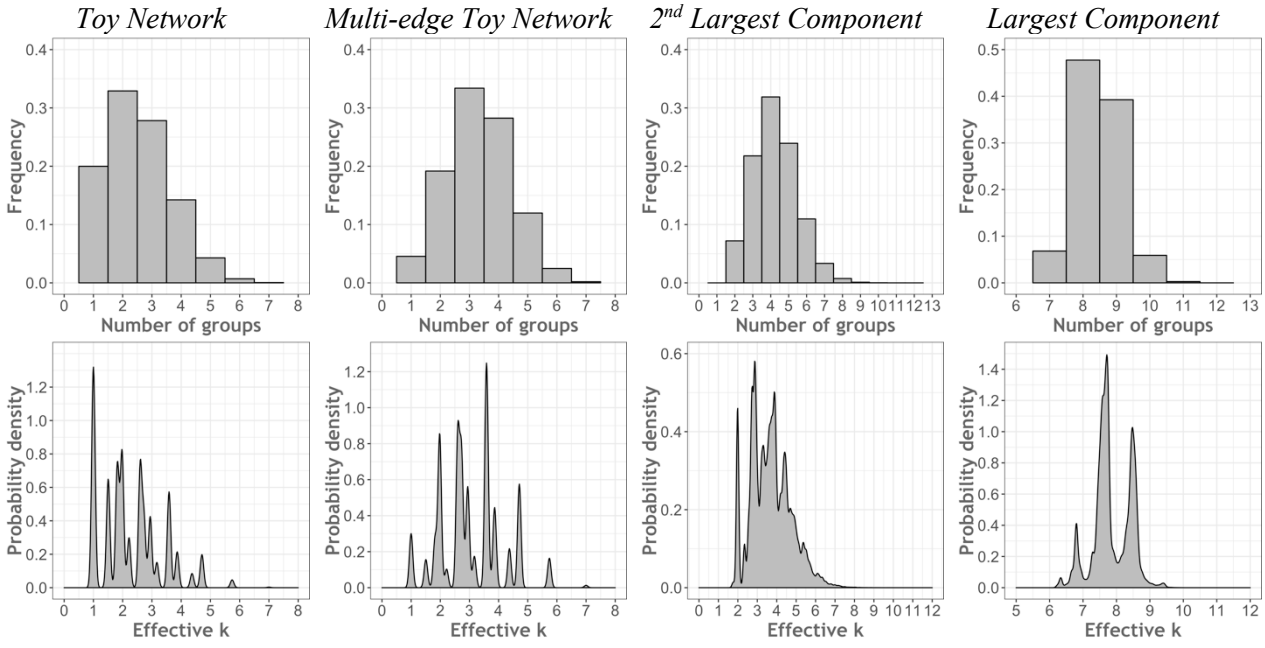
Zhang, K., Bhattacharyya, S., & Ram, S. (2016). Large-Scale Network Analysis for Online Social Brand Advertising. *Mis Quarterly*, *40*(4).

Figures and Tables

Table 1

SBM Communities and Ground Truth Communities

Network Component	Community No.	Size (N)	Gender		Class					
			Male	Female	10A	10B	10C	10D	10E	10F
Largest Component	1	10	1	9					10	
	2	14	1	13				13		1
	3	18	8	10	6	9		1	1	
	4	21	16	5				21		
	5	23	17	6					23	
	6	26	17	9		21	5			
	7	35	18	17						35
	8	30	15	15			30			
	Total	177	93	84	6	30	35	35	34	36
2nd Largest Component	1	4	4		4					
	2	4	3	1	4					
	3	5	1	4	5					
	4	7	1	6	7					
	Total	20	9	11	20					



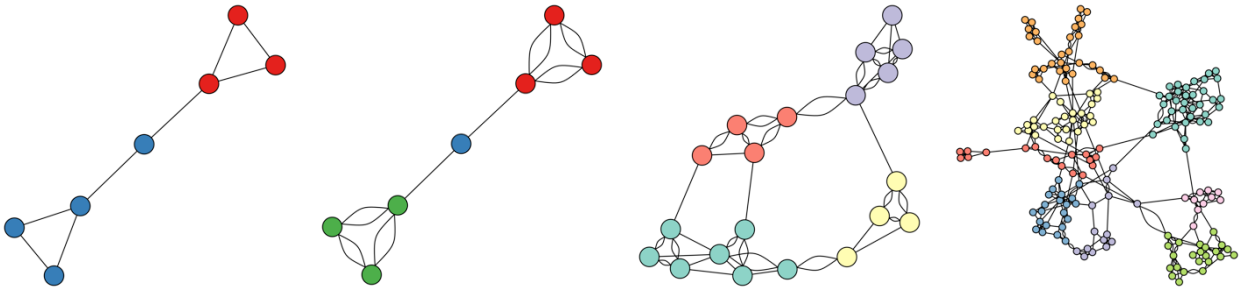


Figure 1. Results for two toy model networks and two largest network componens of the real-world best friends school network. From top to bottom the results shown are the number of groups calculated using RCRN method, the maximum likelihood DCSBM community structure.

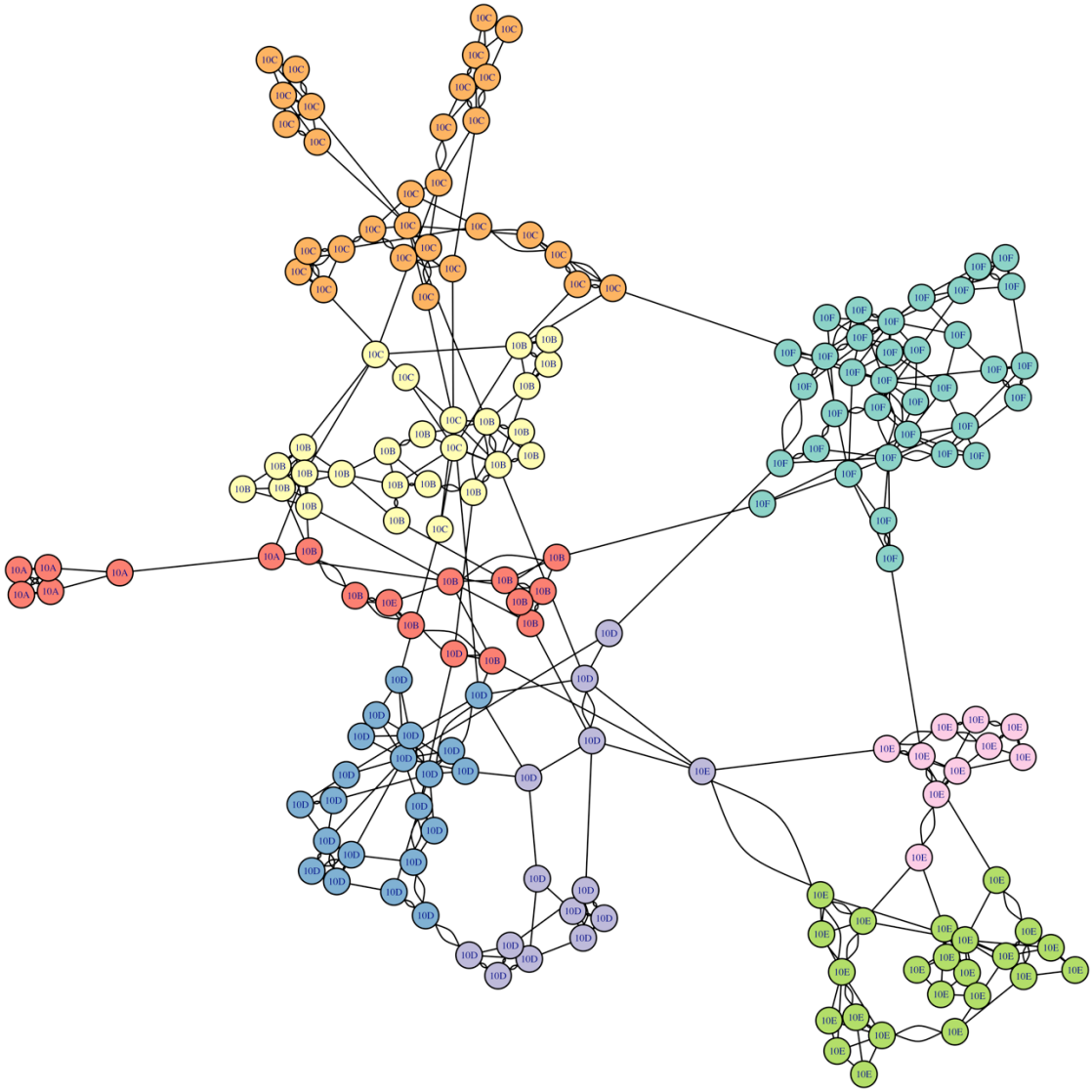


Figure 2. The network map of the largest component with the colors representing the maximum likelihood DCSBM community structure. The node metadata indicate physical classes.