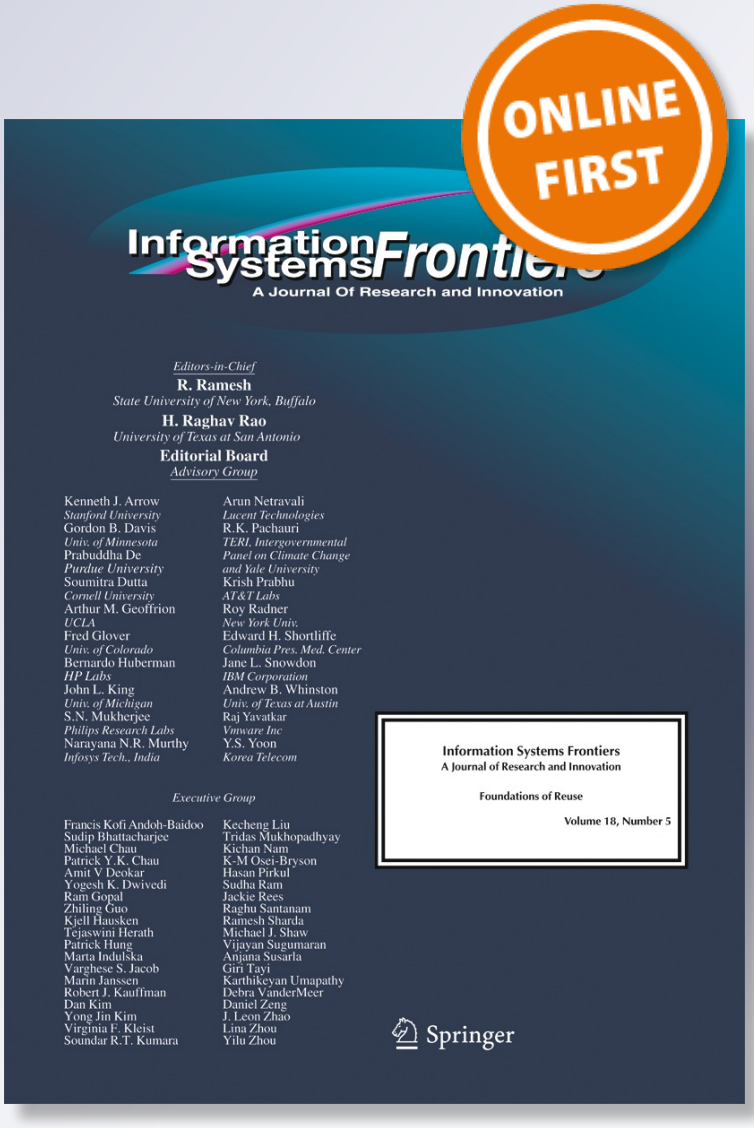
Uncovering the effect of dominant attributes on community topology: A case of facebook networks

Yi-Shan Sung, Dashun Wang & Soundar Kumara

Information Systems Frontiers
A Journal of Research and Innovation

ISSN 1387-3326

Inf Syst Front
DOI 10.1007/s10796-016-9696-0



Springer

Inf Syst Front
DOI 10.1007/s10796-016-9696-0

Uncovering the effect of dominant attributes on community topology: A case of facebook networks

Yi-Shan Sung[1] · **Dashun Wang**[2] · **Soundar Kumara**[1]

Abstract Community structure points to structural patterns and reflects organizational or functional associations of networks. In real networks, each node usually contains multiple attributes representing the node's characteristics. It is difficult to identify the dominant attributes, which have definitive effects on community formation. In this paper, we obtain the overlapping communities using game-theoretic clustering and focus on identifying the dominant attributes in terms of each community. We uncover the association of attributes to the community topology by defining dominance ratio and applying Pearson correlation. We test our method on Facebook data of 100 universities and colleges in the U.S. The study enables an integrating observation on how the offline lives infer online consequences. The results showed that people in class year 2010 and people studying in the same major tend to form denser and smaller groups on Facebook. Such information helps e-marketing campaigns target right customers based on demographic information and without the knowledge of underlying social networks.

✉ Soundar Kumara
skumara@psu.edu

Yi-Shan Sung
yqs5097@psu.edu

Dashun Wang
dwang@ist.psu.edu

[1] Department of Industrial & Manufacturing Engineering, The Pennsylvania State University, University Park, PA 16802, USA

[2] Kellogg School of Management and Northwestern Institute on Complex Systems, Northwestern University, Evanston, IL 60208, USA

Keywords Dominant attribute · Community detection · Facebook · Game-theoretic clustering · Dominance ratio · Community topology

1 Introduction

The interactions of components in most of the real systems can be captured by a network structure. The analysis of this network topology can lead to discovery of various interesting properties. This has resulted in immense benefits in various fields of research (Albert and Barabási 2002; Newman 2003; Cavdur and Kumara 2014a, b). Community structures - exemplified by dense connections within a group of nodes and sparse connections between groups - is a property worth exploring deeper to make inferences on the properties of networks. Because of the existence of more links within the groups than between the groups, communities usually have functional or organizational significance (Fortunato 2010). Therefore, after a community detection algorithm is applied to a network, the important question we ask is: "Given a community are there any dominant attributes in this community?" An answer to this question will possibly lead us to conjecture that these dominant attributes are responsible for community formation. Each node in the network is associated with several attributes and by looking at the commonality of these attributes in a community we will be able to identify dominant attributes.

Dominant attributes can be interpreted into two different ways: 1) in terms of the community structure, and 2) in terms of a specific community. With respect to the whole community structure, we can answer which categorical attributes can define the groups that quite correspond to the network-structural communities. To do so, the grouping based on a given class of node attributes can be regarded as one set of communities. Then the similarity of this grouping to a set of

Published online: 28 September 2016

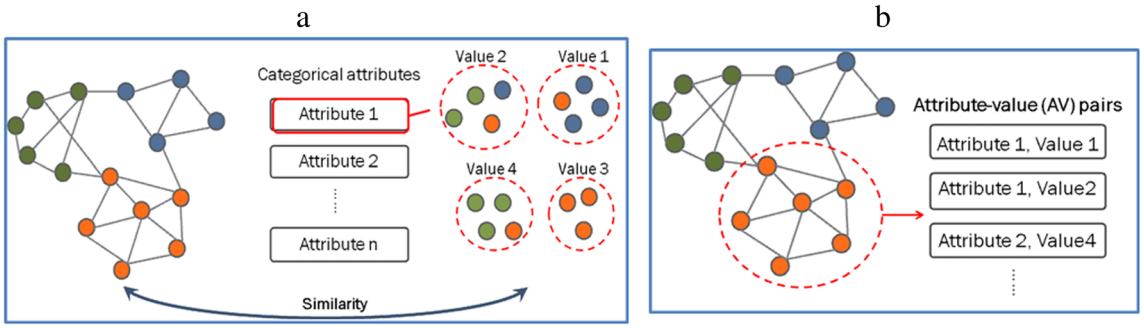


Fig. 1 Illustration of dominant attributes (**a**) in terms of the whole community structure and (**b**) in terms of a specific community

algorithmically-detected communities is measured to quantify the effect of this attribute to community formation (see Fig. 1a). Traud et al. (2011, 2012) adopted pair counting and Rand coefficient as the similarity measure to Facebook networks and report that class years are the dominant attributes to the community formation. However, by using this method, we may identify the dominant attributes of the community structure in the global view but know little about which attributes dominate a specific local community and which local values of a specific attribute contributes to the local community. The other way is to identify dominant attributes in terms of a specific local community. For one categorical attribute, there are several values to represent different characteristics of a node in terms of that attribute. For example, major is one of the attributes for people in the university, and the values stand for different majors. In order to identify what values of attribute are dominant to a given community, we quantify the dominance of each Attribute-Value pair (AV pair) (see Fig. 1b). Our focus in this paper is to develop a methodology to find which of the AV pairs dominate a given cluster.

The proliferation of the Internet led to the rapid development of social networking sites such as Facebook, Twitter, LinkedIn, etc. By using these sites, users can seek out their offline friends or others with similar interests, backgrounds and disposition. They can have some online interactions with each other, and these online social activities can also be brought to offline world and strengthen their social networks. The ubiquity of online social networks makes e-marketing much easier than before (Constantinides and Fountain 2008; Trusov et al. 2009). To promote a new product and target the right consumers in an effective and efficient way, sales would like to know which kinds of customers can accelerate the promotion and who have the preferences to which kinds of products. E-marketing applications rely heavily on the ability to understand the structure of social networks.

Applying the analysis of community structures to social networks can decompose these networks into several partitions in which people in the same partition have common interests or similar properties. In order to identify the commonalities within the communities, our primary objective in this paper is to investigate how to measure the dominance of attributes in a given cluster and which dominant attributes have the effect on community characteristics, such as size and density. Once this information is obtained, marketing campaigns can target individuals using information about their attributes and they need not have knowledge of the underlying social network. For this purpose, the related previous works are reviewed, including works regarding to community detection algorithms, methods for identifying dominant attributes and applications of community detection algorithms to online social networks. Then we demonstrate how to identify dominance of attributes in community structure and evaluate the effect of attributes on community topology. The data used for proving the efficacy of methods are from Facebook in 100 universities of varying size, which are used in Traud et al. (2011, 2012). The data were posted by Porter (2011).

2 Literature review

Based on the underlying methodological principles, community detection algorithms are classified into five categories (Papadopoulos et al. 2012). These five categories are vertex clustering, divisive, cohesive subgraph discovery, community quality optimization and model-based methods. Algorithms in the first two categories, such as k-means (Hartigan and Wong 1979) and inter-community edge removal (Girvan and Newman 2002), show both time and memory complexities higher than quadratic number of network nodes and not applicable to large scale networks. Methods focusing on achieving subgraph internal cohesiveness include finding n-cliques (Luce 1950), ρ-quasi cliques (Matsuda et al. 1999) and k-cores (Seidman 1983). One of the most efficient methods belonging to this category is Scan (Xu et al. 2007), which defined structural similarity to quantify the extent of common neighbors of node pair and based on this to detect communities in networks. Algorithms seeking to optimize community quality always involve an objective function to stand for the quality of the whole community structure. Modularity (Newman 2004), comparing the actual density within subgraphs to the expected density in corresponding random graphs, is the most popular and common used quality function of community structure. Many algorithms have been

developed based on modularity (Clauset et al. 2004; Newman and Girvan 2004; Guimera and Amaral 2005; Blondel et al. 2008). Algorithms belonging to model-based methods detect communities either by a dynamic process or modeling clustering problems based on statistical nature. Label propagation (Raghavan et al. 2007) is a near-linear-time algorithm that assigns a label to each node to represent its cluster membership. Cluster-based compression method (Rosvall and Bergstrom 2007, 2008) finds the cluster structure by encoding network topology and a good clustering is when the cost of encoding achieves the minimum.

Though most of the community detection algorithms focus on non-overlapping clustering, some are constructed to allow communities to overlap. Clique percolation method (Palla et al. 2005) defines a community as a series of adjacent *k*-cliques, where two *k*-cliques are adjacent if they share *k-1* nodes. A recent line of enquiry focuses on defining communities based on links instead of nodes (Evans and Lambiotte 2009; Ahn et al. 2010). In contrast to node-based community detection algorithms, grouping links naturally leads to communities with pervasive overlap, while preserving the hierarchical organization in networks. Some community detection techniques use the logic of seed expansion to detect community structure. Lancichinetti et al. (2009) define a fitness function for detecting communities, and a node is randomly selected to expand its group members by that function. This process continues until each node in a network is assigned to at least one community. Iterative scan method (Lancichinetti et al. 2009) starts to find a community with a seed candidate cluster and then adds or deletes one node at each round according to a weight function until no improvement on the function by one change. Rank removal method (Baumes et al. 2005) identifies core clusters by removing high-page-ranking nodes from a network, and then these removed nodes are re-added into network to join one or more core clusters based on the weight function.

Some approaches extend the existing non-overlapping communities detection methods to the overlapping subgraphs. Wei et al. (2009) use the spectral partitioning to find the seed clusters and expand the seeds to the overlapping clusters by lazy random walks. After label propagation method was introduced in 2007 (Raghavan et al. 2007) several modifications of the algorithm appeared till today and the main argument in favor of the algorithm is its simplicity and speed. For example, Gregory (2010) uses label propagation to detect overlapping clusters. For each propagation step, each node copies all the labels of its neighbors into its label set and each label is assigned a coefficient which represents the belonging strength, such that all coefficient for each node sum to one. Xie et al. (2011) use label propagation to detect overlapping communities. Unlike the method proposed by Gregory (2010), where each node forgets the labels gained in the previous iterations, Xie et al. (2011) design an algorithm in which each node has a memory to store all the labels received in the past and the occurrence frequency of labels represents the "belonging" strength. Fuzzy c-means modularity optimization (Zhang et al. 2007) projects nodes into d-dimensional Euclidean space and a new modularity function considering the fuzziness in belonging to different clusters is used to detect the overlapping clusters.

Once a suitable community detection algorithm is applied to networks, the next step is to find out if there are any common attributes that could possibly have contributed to the community formation. For example, in protein-protein interaction networks, proteins with the same functions tend to form communities (Chen and Yuan 2006), and in a World Wide Web network, groups of web pages correspond to those with related topics (Eckmann and Moses 2002; Flake et al. 2002). In Belgian mobile phone network, language is the key attribute to form groups since people speak in the same languages tend to have much more frequent communication than people speak in different languages (Blondel et al. 2008). In complex systems, an element usually has multiple attributes to represent its identity and to dig out which attributes have significant impact on community structure is one of the issues of big data analysis. By measuring the similarity between the algorithmically-detected clusters and the attribute-based clusters, it was found that class year is critical to community formation in Facebook networks at universities (Traud et al. 2011, 2012).

In addition to Traud et al. (2011, 2012), several studies have analyzed online social networks by using community detection algorithms, though most of them are based on non-overlapping clustering algorithms. Pujol et al. (2009) found that the proportion of edges within communities is more consistent across various community sizes by using modularity optimization to detect communities in Twitter and Orkut data. Bonneau et al. (2009) applied modularity optimization to the Facebook network which was constructed by nodes with degree not higher than 8 and found that the modularity to be almost the same as that of the complete network. Mislove et al. (2010) applied modularity optimization to clustering the users with revealed attributes. Then the similarity between the detected clusters and the attribute-based clusters was identified by normalized mutual information. The results showed that merely 20 % knowledge about the user attributes can infer the attributes of the remaining users with 80 % accuracy.

3 Methodology

Identifying dominant attributes is one of the ways to interpret detected community structures. Once the suitable method is chosen and applied to a given problem, the next task is to investigate what caused these communities to evolve. Since usually each node contains a number of attributes to represent its characteristic, it is difficult to identify which attribute(s) have the definitive effects on community formation.

Therefore, we need quantitative methods to identify dominant attributes.

3.1 Identification of dominant attributes

In this paper, the dominance of each AV pair was quantified to see which of the AV pairs dominate a given cluster. Since different AV pairs in the whole population vary in number, simply using the quantity of an AV pair in a given cluster cannot represent the attribute(s) dominance. The dominance of each AV pair should be defined based on the number of distinct AV pairs in the population. Suppose we randomly choose some nodes into a group, there may not be any attribute of significance in that group. In such a selection, the percentage of each AV pair in this group will be the same or not significantly different from the percentage of that AV pair in the population. On the other hand, if a group of nodes with some AV pair whose percentage in the group is significantly larger than that in the population, that AV pair is dominant in this group. By comparing the percentage of an AV pair in a given cluster to the percentage of this AV pair in the population, we can recognize whether the AV pair dominates the cluster. We denote the percentage of AV pair (a, v) in the population by $f_{a,v}$ and the percentage of AV pair (a, v) in community c by $f^c_{a,v}$. Then the dominance ratio of AV pair (a, v) in community c is defined as follows.

$$r^c_{a,v} = \frac{f^c_{a,v}}{f_{a,v}}$$

If $r^c_{a,v} < 1$, it means that the percentage of AV pair (a, v) in the selected node group (cluster) is smaller than the percentage in population, which indicates that (a, v) does not dominate community c. As the dominance ratio of (a, v) goes higher, this AV pair is more dominant in the community.

3.2 Effects of node attributes on community topology

By using dominance ratio, the properties of each community can be inferred in terms of the dominant attributes. Then we can identify the effect of the attributes on the community structure. In specific, it will be interesting to know which kinds of AV pairs have a strong relationship to the community topology. Several network metrics such as centralities and density enable to characterize the properties of community. In this paper, community size and community density were selected to represent the community topology. Then the effect of an AV pair on community size or density is defined by the relation between the dominance ratio of an AV pair and the corresponding community size or density. Several methods can be used to quantify the dependence of two variables. In this study, we assume that the correlation between the dominance ratio and corresponding community properties is near linear and thus Pearson correlation (Lee Rodgers and Nicewander 1988) is chosen to test the dependence between the dominance ratio and the community topology. We define d^c as community density in community c and s^c as the community size in community c. The dependence between the dominance ratio of AV pair (a, v) and the community density, denoted as $corr_d$ is as follows.

$$corr_d = \frac{\sum_c \left(r^c_{a,v} - \overline{r}_{a,v}\right)\left(d^c - \overline{d}\right)}{\sqrt{\left(r^c_{a,v} - \overline{r}_{a,v}\right)^2}\sqrt{\left(d^c - \overline{d}\right)^2}}$$

where $\overline{r}_{a,v}$ is the average dominance ratio of AV pair (a, v) over the observed communities in a network, and $\overline{d}$ is the average community density of the observed communities in a network. Similarly, the dependence between the dominance ratio of AV pair (a, v) and the community size, denoted as $corr_s$ is:

$$corr_s = \frac{\sum_c \left(r^c_{a,v} - \overline{r}_{a,v}\right)\left(s^c - \overline{s}\right)}{\sqrt{\left(r^c_{a,v} - \overline{r}_{a,v}\right)^2}\sqrt{\left(s^c - \overline{s}\right)^2}}$$

where $\overline{s}$ is the average community size of the observed communities in a network. The two-tailed t-test was used to determine whether the correlation coefficient is statistically significant.

4 Facebook data

Data used to study dominant attributes are Facebook data of 100 universities and colleges in the U.S. from a single-day snapshot in 2005. In the data, users are recorded by numbers to protect privacy. To build Facebook networks, one user is regarded as one node, and if two users were friends on Facebook on the day the data was extracted, we put an undirected link between these two nodes. Since friendship between different schools is not considered in the data, there are exactly 100 independent networks.

The data also encompass demographic information which was provided by users on their pages of Facebook. Seven categorical attributes stand for this information: 1) gender, 2) major, 3) second major, 4) class year (year supposed to graduate), 5) dormitory, 6) high school, and 7) whether student or faculty. Here, the values of all categorical attributes except class year are described by anonymous numerical identifier. If individuals did not provide the information of some attributes, the value 0 was used to represent it. To see if the results of analysis are affected by the size of universities, the100 universities were classified as small (less than 10,000 users), medium (between 10, 000 and 20,000) and large with more than 20,000 users. Following these criteria Facebook

networks were categorized into 50 small, 33 medium and 17 large networks. The maximum size is 41,554 (The Pennsylvania State University), and the minimum size is 769 (CalTech).

For the sake of clarifying the effect of attributes on network topology, two Facebook networks with small size among the 100 American universities were visualized. In Fig. 2a, different colors of nodes indicates different class years of users, and any missing value is represented by grey color. In Fig. 2b, different colors represent different dormitories of users, and nodes with unknown dormitories are hidden for simplicity. Both graphs show that nodes with the same colors have the tendency to get together, which implies users living in the same dormitory or graduate in the same year (senior, junior etc.) are more likely to be Facebook friends. However, we can also see that many nodes with different colors are mixed together. It indicates that a single attribute cannot partition graphs very well, and for one cluster, there may be multiple dominant attributes to represent its properties. It also implies the existence of overlapping communities on Facebook networks.

5 Analysis and results

5.1 Game theoretic approach to clustering

Since Facebook networks contain overlapping community structure, we need suitable overlapping community detection algorithms. In this paper, we use game-theoretic clustering (Mandala et al. 2014) to detect overlapping communities in 100 Facebook networks. Game theory allows us to study the strategic interactions among players, and once a strategy reaches Nash equilibrium, no player can gain more reward by purely changing its own strategy. A Game theoretic approach to clustering is a community detection algorithm for undirected and unweighted graphs. It is an algorithm that outperforms several other overlapping community detection algorithms such as clique percolation method (Palla et al. 2005) and local expansion algorithm (Lancichinetti et al. 2009) by testing on artificial networks. Game-theoretic clustering also has an advantage of its computational complexity in number of edges. Some agent-based clustering algorithms such as SLPA (Xie et al. 2011) also have near-linear running time and similar performance to game-theoretic clustering. However, game-theoretic clustering has an explicit function to define clusters and a clear stopping criterion.

In game-theoretic clustering, each node in the network is regarded as a player and the set of cluster labels it chooses is its strategy. The reward of each player comes from the neighbors belonging to the same cluster. Every player can choose more than one cluster label but encounters a penalty, and more the labels a player picks more the penalty the player pays. Thus in order to maximize the rewards, each player will choose the labels which the maximum number of its neighbors belong to and minimize the number of clusters the player joins at the same time. The reward function of player v (Mandala et al. 2014) is as follows.

$$r_v(l; sl_{-v}) = \sum_{w \in N_v} |l \cap sl_w| - \rho \sum_{w \in V \setminus \{v\}} |l \cap sl_w|$$

where l is one of the cluster labels chosen by player v, sl_v is a set of labels selected by player v, and N_v is a set of players who are the neighbors of player v. $V\setminus\{v\}$ represents a set of all players in the observed network other than player v. The first term of the equation counts the number of its neighbors selecting the same label, and the second term is the discounted number of the players choosing label l, so the cluster density is maintained no less than ρ. In order to limit the number of labels player v can have, the cost function is given as:

$$c_v(sl_v) = \frac{1}{2}\lambda_v |s_v|(|sl_v| - 1)$$

where λ_v is the penalty coefficient representing how much player v will lose if player v joins one more cluster. In order

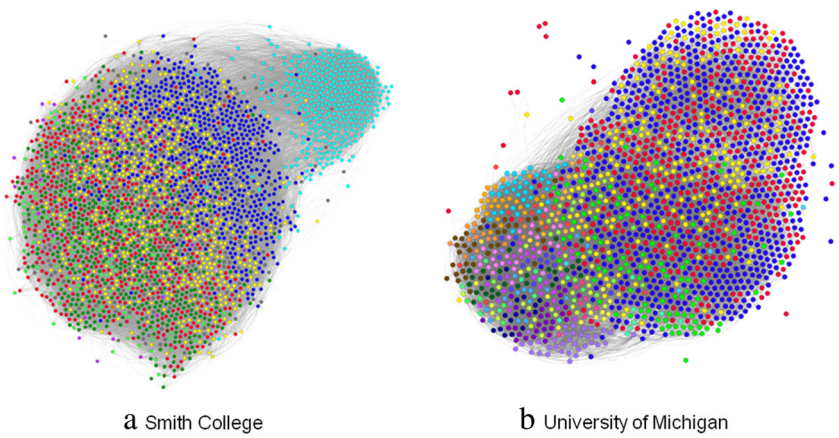


Fig. 2 Largest connected component of Facebook networks at (**a**) Smith College and (**b**) University of Michigan

to let all players result in the same cost of joining multiple clusters, $\lambda_v = \lambda(1-\rho)$. Based on the experiments, $\lambda = 2$ works well in practice. The utility of player v (Mandala et al. 2014) obtained by reward and cost function is given below:

$$U_v(sl_v; sl_{-v}) = \sum_{w \in N_v} r_v(l; sl_{-v}) - c_v(sl_v)$$

Based on the utility function, each player in game-theoretic clustering is able to find the best strategy which generates the highest utility under the condition that the strategies of the other nodes are known. The clustering solution is found when game-theoretic clustering reaches equilibrium. The best strategy of players can be updated sequentially and simultaneously. In this paper, the sequential way of updating was applied to find out communities in 100 Facebook networks. To apply game-theoretic clustering to networks, we need to give the minimum density ρ of each community and penalty coefficient λ representing how much each player will lose if the player joins one more community. In this paper, we set ρ to be 0.1 and 0.2, so λ is 1.8 and 1.6, respectively, based on the suggestion that $\lambda = 2(1-\rho)$ (Mandala et al. 2014). Since it would be interesting to see if the analysis results change if non-overlapping community structure is detected in Facebook networks, the third set of (ρ, λ) is set to (0.1, M). Here M is a sufficiently large number to make penalty very high if one player wants to join more than one cluster. In this paper, we set M=100000.

The detected communities with size smaller than 20 were discarded due to less community-like structure. In order to identify the effect of the specific AV pairs on the community size and density, the dominance ratio of the AV pairs in each community and the corresponding community size and density were first computed. Then, the correlations of the dominance ratio of the AV pairs to the community size and to the community density were calculated. In this paper we focus on

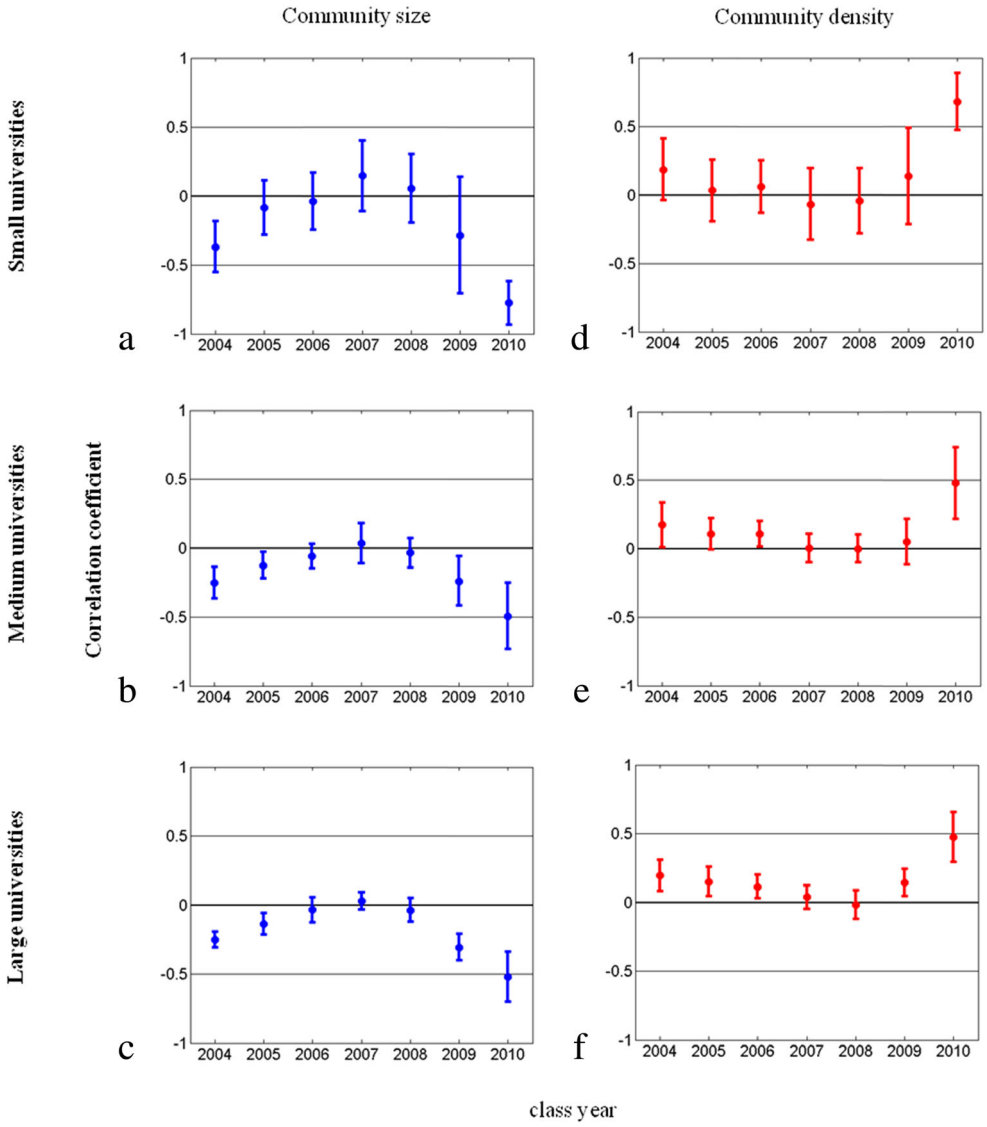


Fig. 3 Mean and standard deviation of correlation between dominance ratio and cluster size/density across class years when using game-theoretic clustering with $(\rho, \lambda) = (0.1, 1.8)$ to detect community structure

class year, major and dorm, helping us answer two questions: (1) whether people in different class years form different sizes or different density groups, and (2) whether people in the same major/ in the same dormitory, tend to form the denser groups.

5.2 Effect of class year on online community topology

To see the effect of class year on community topology, we first computed the dominance ratio of each class year in each community detected by game-theoretic clustering, and then we identified the correlation of the dominance ratio to the cluster size and density. Since Facebook was launched in 2004, users graduating earlier than 2004 were ignored.

First we focus on the results by game-theoretic clustering with $(\rho, \lambda) = (0.1, 1.8)$. Figures 3 and 4 are two sets of error bar plots regarding the mean correlation and the corresponding standard deviation with respect to universities of various sizes. Note that the correlation coefficients here are all with p-value smaller than 0.05. Figure 3 demonstrates the correlation of dominance ratio of each class year to cluster size and density. The more the correlation deviates from zero, the more significant result is. It shows that the dominance ratio of class year 2010 has relative significant correlation to cluster size and density compared to other class years. This correlation is more significant when we analyzed the data from small universities. We also observed that the standard deviation of the correlation coefficient becomes smaller when the size of universities is larger no matter what class years are considered.

These results lead to three conclusions: (1) that cluster size tends to have negative correlation to the dominance ratio with respect to class year while cluster density tends to have positive correlation to the dominance ratio; (2) that the dominance

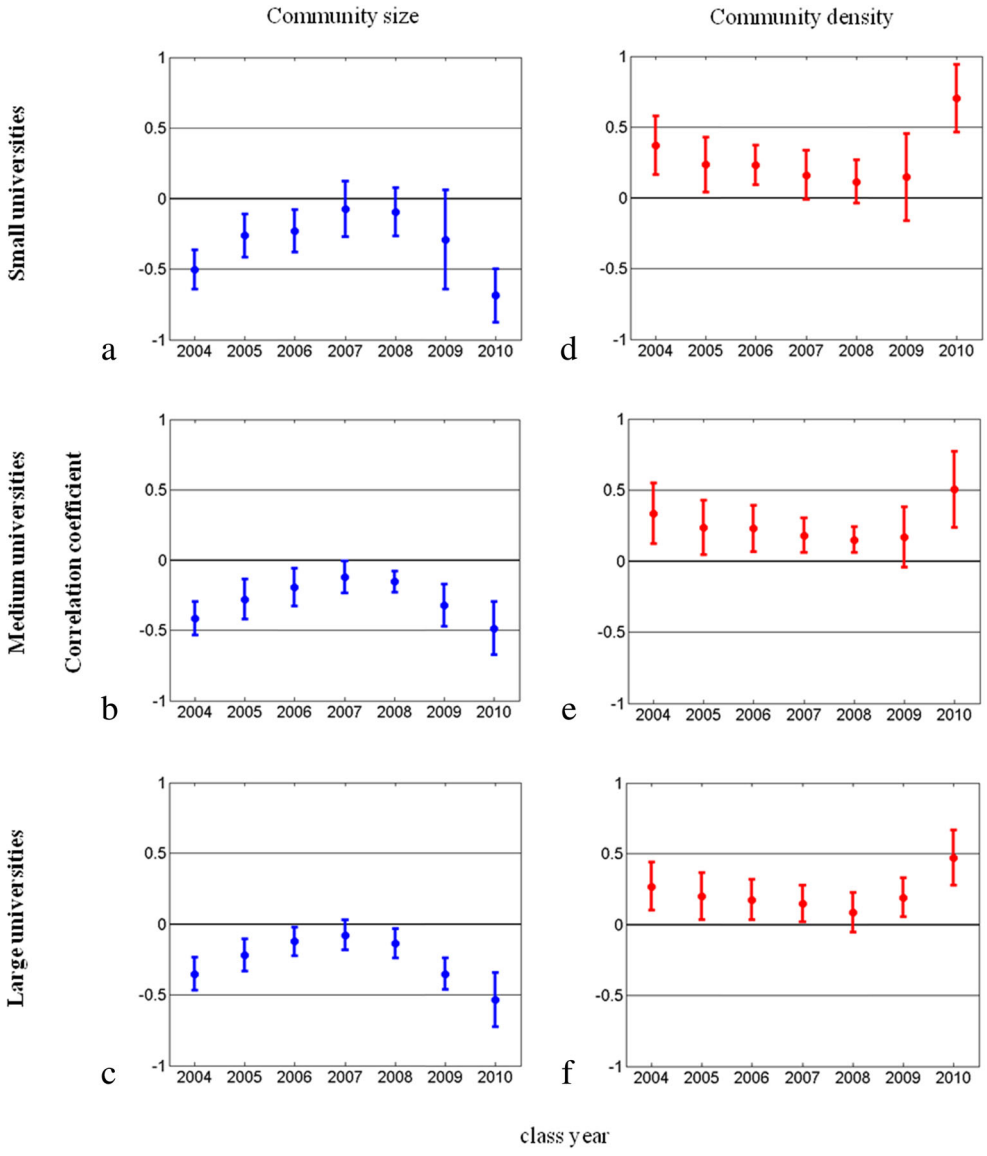


Fig. 4 Mean and standard deviation of correlation between dominance ratio and cluster size/density across class years when using game-theoretic clustering with $(\rho, \lambda) = (0.2, 1.6)$ to detect community structure

ratio of class year 2010 has significant correlation to cluster size and density irrespective of the size of universities selected; and (3) that the dependence between dominance ratio of years and cluster structure in large universities is more consistent since the standard deviation is small. From the second conclusion, we can infer that in general people in class year 2010 have higher possibility to form groups with small size and high density than people graduated in earlier years. Note that there is a high standard deviation of the correlation between the dominance ratio of class year 2009. It implies that at some small universities, people in class year 2009 also tend to form small size groups on Facebook.

We tried to detect denser community structure in Facebook networks– using game-theoretic clustering with $(\rho, \lambda) = (0.2, 1.6)$. Figure 4 shows the corresponding correlation between the dominance ratio and cluster size/density across class years. The results are mostly consistent with those shown in Fig. 3. The only difference is that the standard deviation of the correlation coefficient did not necessarily drop down as the size of universities becomes larger.

So far we used game-theoretic clustering to detect overlapping communities in Facebook networks and did the analysis that we believed that the networks contain overlapping community structure. It would be interesting to see the results if we assume a non-overlapping community structure. Therefore, game-theoretic clustering with $(\rho, \lambda) = (0.1, 100000)$ was applied to identify the non-overlapping clusters and the effect of class year was shown in Fig. 5. Compared to Fig. 3, the results regarding the correlation between the dominance ratio and the cluster size are very similar, but the dominance ratio of class year 2010 does not have strong positive correlation to cluster density at medium and large universities. It means no

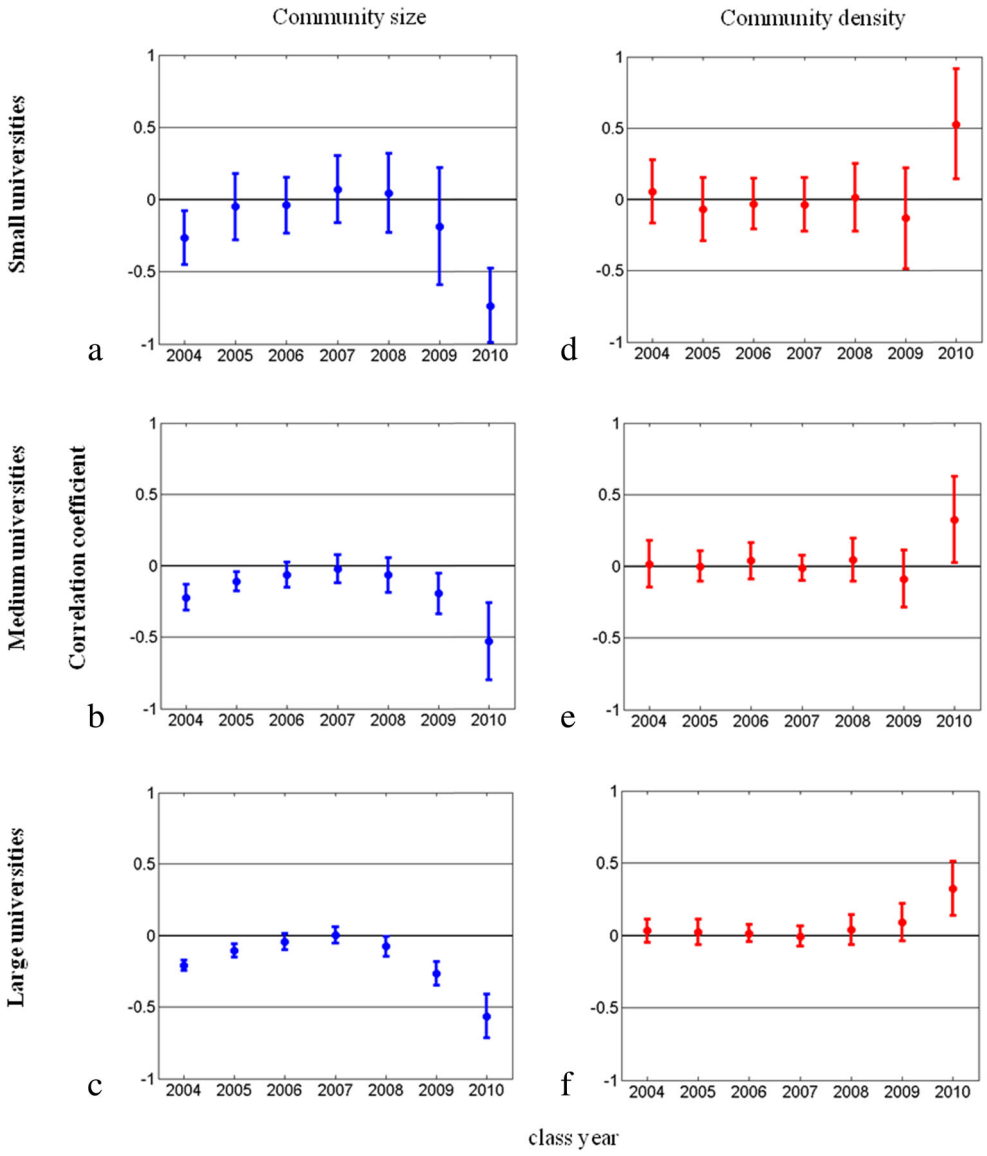


Fig. 5 Mean and standard deviation of correlation between dominance ratio and community size/density across class years when using game-theoretic clustering with $(\rho, \lambda) = (0.1, 100000)$ to detect community structure

significant association between the dominance ratio of class year 2010 and the cluster density exists if we consider the community structure in Facebook networks is non-overlapping.

5.3 Geographical effects on online community topology

Within a university, people studying in the same major (due to similarities in classes) and living in the same dorm (living in the same neighborhood) can be considered geographically co-located. We computed the correlations of the dominance ratio of different majors and dorms to the community size and density as community topology, and then we investigated what fraction of majors and dorms are with these strong correlations. To do so, a correlation higher than |0.5| with confidence level 95 % is defined as strong negative or strong positive correlation depending whether the value is negative or positive.

Figures 6 and 7 show the mean and standard deviation of percentage of majors and dorms with strong correlation between dominance ratio and community topology. In terms of community size (see Fig. 6), the results were similar when game-theoretic clustering with $(\rho, \lambda) = (0.1, 1.8)$ and with $(\rho, \lambda) = (0.2, 1.6)$. No matter what size of universities is considered, on an average more than 50 % of majors with a strong negative correlation between the dominance ratio and the community size exists. However, in terms of dormitories, less than 30 % have this property except for the parameter $(\rho, \lambda) = (0.2, 1.6)$ at small universities. From these results, we can conclude that people in the same major tend to form smaller size communities, while the communities formed by people in the same dormitory do not have this property.

As the non-overlapping community structure was detected, it would be interesting to check if the percentage of majors or dormitories having strong correlation to community size is similar to that in overlapping communities. As shown in Fig. 6, on an average, about 40 % of majors with a strong negative correlation between the dominance ratio and the community size exist, and around only 10 % dormitories

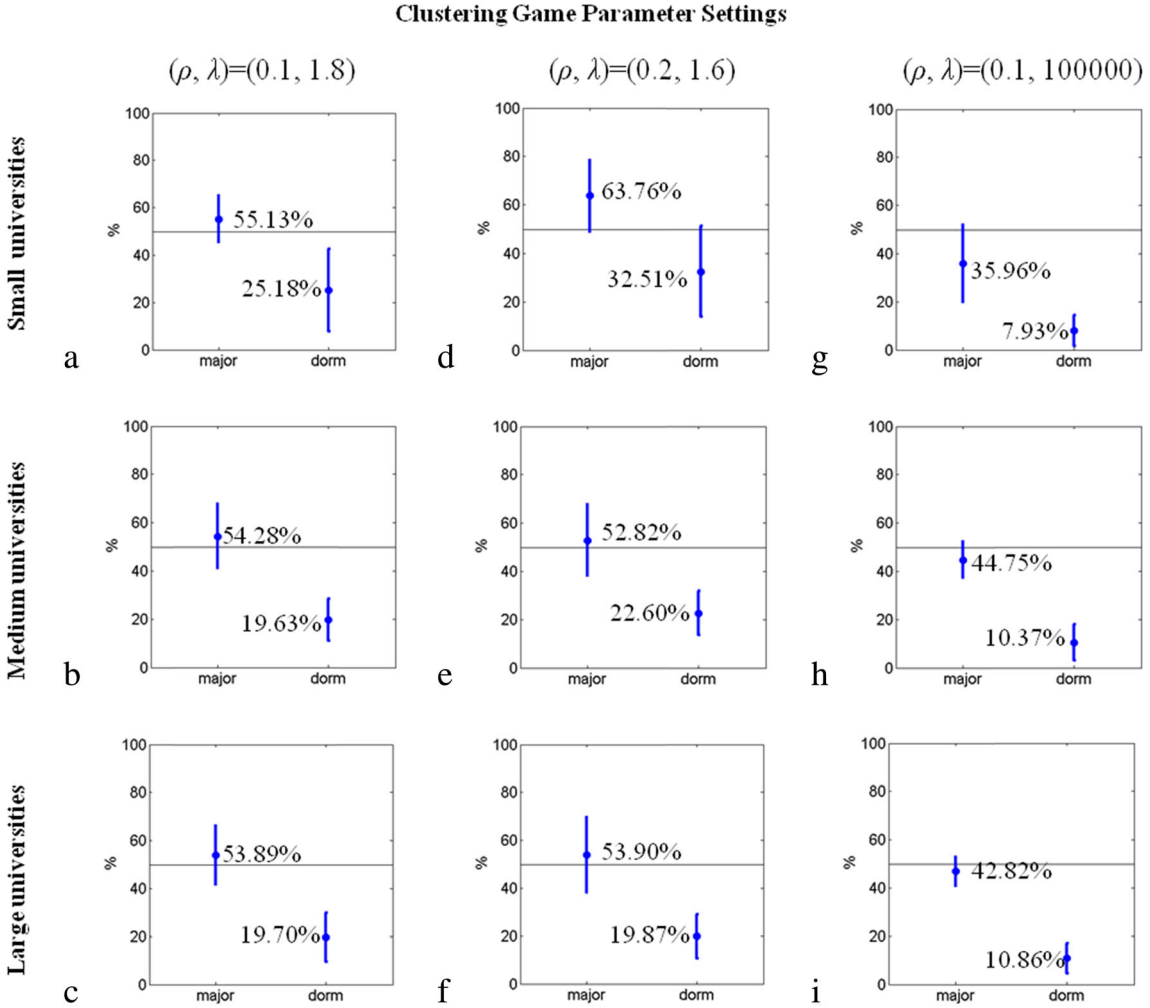


Fig. 6 Percentage of majors/dorms with strong correlation between dominance ratio and community size under different parameter settings of game-theoretic clustering

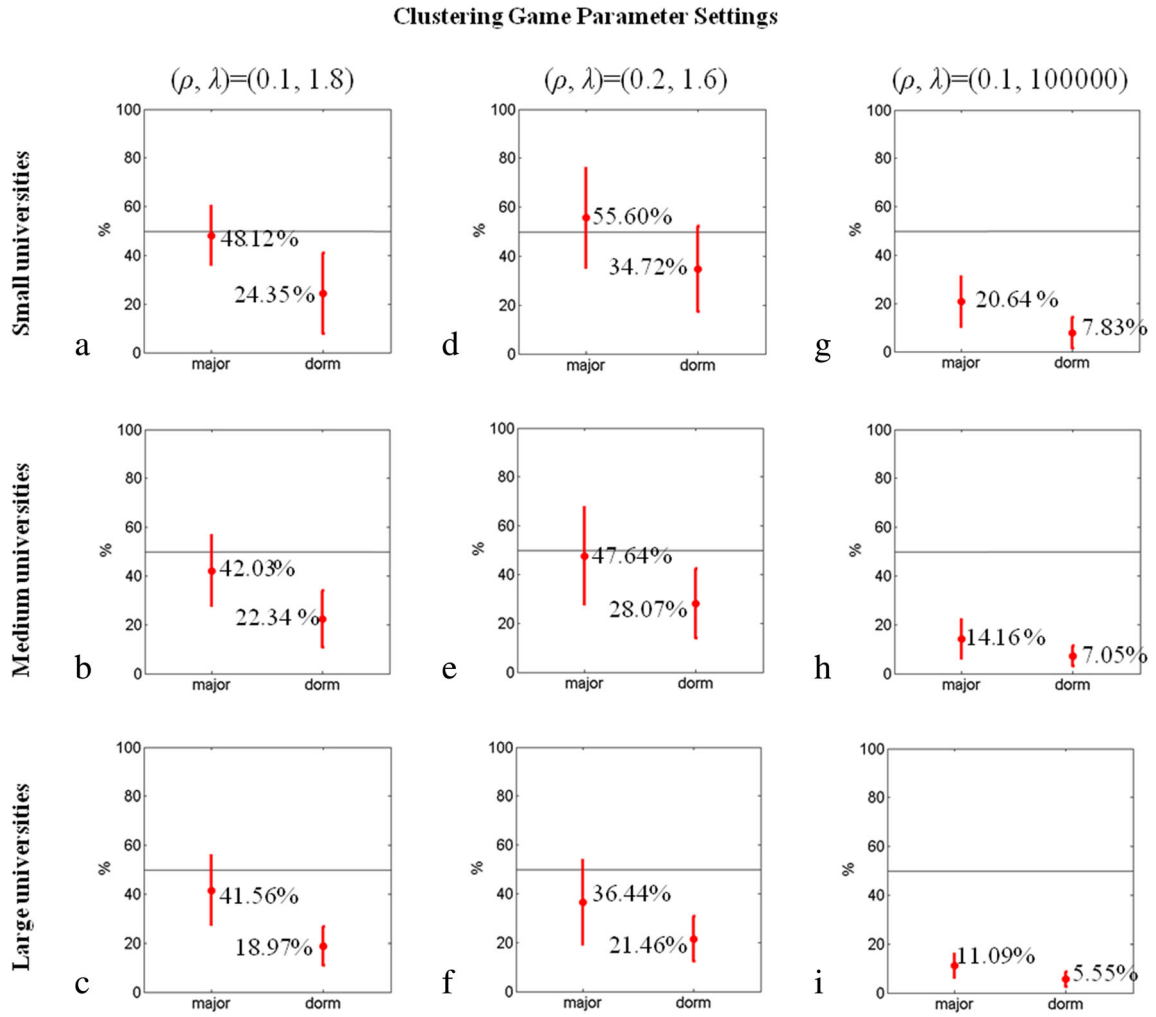


Fig. 7 Percentage of majors/dorms with strong correlation between dominance ratio and community density under different parameter settings of game-theoretic clustering

have this property. The percentage of both majors and dormitories has a big drop. This implies that the dependence of dominance ratio of majors to community size is not that significant in non-overlapping community structures.

From Fig. 7, we observe that the percentage of majors whose dominance ratio has strong positive correlation to community density is higher than that of dormitories. On average, when overlapping community structure is considered, the difference is about 20 % or more, except for the results by game-theoretic clustering with $(\rho, \lambda) = (0.2, 1.6)$ at large universities. The significant difference between majors and dormitories implies that people in the same majors tend to form denser groups compared to people in the same dormitories. On the other hand, when we analyzed non-overlapping communities, we observe that this percentage of both majors and dorms is much smaller. It indicates that it is not common to see the dominance ratio of majors or dormitories has strong correlation to community density in non-overlapping community structure.

6 Conclusions and future work

In this paper, we defined the dominant attributes as node attributes that have significant effect on community formation. We also defined that the dominant attributes can be identified in terms of the whole community structure or in terms of a specific cluster. To answer the question, what attributes have the significant effects on detected communities, this paper was dedicated to identifying dominant attributes in terms of each local community. We developed dominance ratio to quantify the dominance degree of an AV pair in a given community. The effect of an attribute on community topology is defined as the dependence of the dominance ratio of this attribute to the corresponding cluster size and density, and the dependence is quantified by Pearson correlation.

To demonstrate the feasibility of the methods, we applied game-theoretic clustering to identify overlapping communities in Facebook networks. This task aimed at analyzing how the offline characteristics of people affect the topology

of the communities they build in online social networks. The results indicated that people in class (graduating) year 2010 tend to form dense and small communities, but the communities formed by people in other class years do not favor any values of size or density. We also used game-theoretic clustering to identify non-overlapping communities, and the results only indicated that people in class year 2010 tend to form small communities but no significant effect on the density. We further identified the effect of major and dormitory on community topology. When game-theoretic clustering was used to detect the overlapping communities, we found that people in the same major tend to form small and dense communities, yet the communities formed by people living in the same dorms do not have this property. When we identified non-overlapping communities on Facebook data, we found that the community size and density are not significantly affected by the dominance of either majors or dormitories.

Though these findings can be an artifact of the data, we have laid out a methodology using game-theoretic clustering and AV pair dominance. Our methodology can be applied to diverse domains such as retail sales data to study co-purchased items to generate dominant attributes of products and customers, and patient data to find commonalities of clinical parameters and diseases.

In this work we have used Pearson correlation to identify the dependence between the dominance ratio and the cluster topology. Pearson correlation can only identify linear associations. In the future, other correlation metrics can be used to observe how the effect of attributes on cluster structure, such as Spearman rank correlation (Sarmanov 1962), maximal correlation (Traud et al. 2011) and maximal information coefficient (MIC) (Reshef et al. 2011). Furthermore, in this paper, we only used game-theoretic clustering to find out the communities in Facebook networks. It will be interesting to apply other overlapping community detection algorithms to study how the results behave.

References

Ahn, Y. Y., Bagrow, J. P., & Lehmann, S. (2010). Link communities reveal multiscale complexity in networks. *Nature, 466*(7307), 761–764.

Albert, R., & Barabási, A. L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics, 74*(1), 47.

Baumes, J., Goldberg, M. K., Krishnamoorthy, M. S., Magdon-Ismail, M., & Preston, N. (2005). Finding communities by clustering a graph into overlapping subgraphs. *IADIS AC, 5*, 97–104.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 2008*(10), P10008.

Bonneau, J., Anderson, J., Anderson, R., & Stajano, F. (2009). Eight friends are enough: social graph approximation via public listings. In *Proceedings of the Second ACM EuroSys Workshop on Social Network Systems* (pp. 13–18). ACM.

Cavdur, F., & Kumara, S. (2014a). A network view of business systems. *Information Systems Frontiers, 16*(1), 153–162.

Cavdur, F., & Kumara, S. (2014b). Network mining: applications to business data. *Information Systems Frontiers, 16*(3), 473–490.

Chen, J., & Yuan, B. (2006). Detecting functional modules in the yeast protein–protein interaction network. *Bioinformatics, 22*(18), 2283–2290.

Clauset, A., Newman, M. E., & Moore, C. (2004). Finding community structure in very large networks. *Physical Review E, 70*(6), 066111.

Constantinides, E., & Fountain, S. J. (2008). Web 2.0: conceptual foundations and marketing issues. *Journal of Direct, Data and Digital Marketing Practice, 9*(3), 231–244.

Eckmann, J. P., & Moses, E. (2002). Curvature of co-links uncovers hidden thematic layers in the world wide web. *Proceedings of the National Academy of Sciences, 99*(9), 5825–5829.

Evans, T. S., & Lambiotte, R. (2009). Line graphs, link partitions, and overlapping communities. *Physical Review E, 80*(1), 016105.

Flake, G. W., Lawrence, S., Giles, C. L., & Coetzee, F. M. (2002). Self-organization and identification of web communities. *Computer, 35*(3), 66–70.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports, 486*(3), 75–174.

Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences, 99*(12), 7821–7826.

Gregory, S. (2010). Finding overlapping communities in networks by label propagation. *New Journal of Physics, 12*(10), 103018.

Guimera, R., & Amaral, L. A. N. (2005). Functional cartography of complex metabolic networks. *Nature, 433*(7028), 895–900.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Applied Statistics*, 100–108.

Lancichinetti, A., Fortunato, S., & Kertész, J. (2009). Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics, 11*(3), 033015.

Lee Rodgers, J., & Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician, 42*(1), 59–66.

Luce, R. D. (1950). Connectivity and generalized cliques in sociometric group structure. *Psychometrika, 15*(2), 169–190.

Mandala, S., Kumara, S., & Chatterjee, K. (2014). A Game-Theoretic Approach to Graph Clustering. *INFORMS Journal on Computing*.

Matsuda, H., Ishihara, T., & Hashimoto, A. (1999). Classifying molecular sequences using a linkage graph with their pairwise similarities. *Theoretical Computer Science, 210*(2), 305–325.

Mislove, A., Viswanath, B., Gummadi, K. P., & Druschel, P. (2010). You are who you know: inferring user profiles in online social networks. In *Proceedings of the third ACM international conference on Web search and data mining* (pp. 251–260). ACM.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM Review, 45*(2), 167–256.

Newman, M. E. (2004). Fast algorithm for detecting community structure in networks. *Physical Review E, 69*(6), 066133.

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

Palla, G., Derényi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature, 435*(7043), 814–8.

Papadopoulos, S., Kompatsiaris, Y., Vakali, A., & Spyridonos, P. (2012). Community detection in social media. *Data Mining and Knowledge Discovery, 24*(3), 515–554.

Porter, M. A. (2011). Facebook 100 Data Set in Quantum Chaotic Thoughts. Retrieved from http://masonporter.blogspot.com/2011/02/facebook100-data-set.html. Accessed May 2011.

Pujol, J. M., Erramilli, V., & Rodriguez, P. (2009). Divide and conquer: Partitioning online social networks. *arXiv preprint arXiv: 0905.4918*.

Raghavan, U. N., Albert, R., & Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E, 76*(3), 036106.

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M., & Sabeti, P. C. (2011). Detecting novel associations in large data sets. *Science, 334*(6062), 1518–1524.

Rosvall, M., & Bergstrom, C. T. (2007). An information-theoretic framework for resolving community structure in complex networks. *Proceedings of the National Academy of Sciences, 104*(18), 7327–7331.

Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences, 105*(4), 1118–1123.

Sarmanov, O. V. (1962). Maximum correlation coefficient (nonsymmetric case). *Selected Translations in Mathematical Statistics and Probability, 2*, 207–210.

Seidman, S. B. (1983). Network structure and minimum degree. *Social Networks, 5*(3), 269–287.

Traud, A. L., Kelsic, E. D., Mucha, P. J., & Porter, M. A. (2011). Comparing community structure to characteristics in online collegiate social networks. *SIAM Review, 53*(3), 526–543.

Traud, A. L., Mucha, P. J., & Porter, M. A. (2012). Social structure of Facebook networks. *Physica A: Statistical Mechanics and its Applications, 391*(16), 4165–4180.

Trusov, M., Bucklin, R. E., & Pauwels, K. (2009). Effects of word-of-mouth versus traditional marketing: findings from an internet social networking site. *Journal of Marketing, 73*(5), 90–102.

Wei, F., Qian, W., Wang, C., & Zhou, A. (2009). Detecting overlapping community structures in networks. *World Wide Web, 12*(2), 235–261.

Xie, J., Szymanski, B. K., & Liu, X. (2011). Slpa: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process. *In Data Mining Workshops (ICDMW), 2011 I.E. 11th International Conference on* (pp. 344–349). *IEEE*.

Xu, X., Yuruk, N., Feng, Z., & Schweiger, T. A. (2007). Scan: a structural clustering algorithm for networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 824–833). ACM.

Zhang, S., Wang, R. S., & Zhang, X. S. (2007). Identification of overlapping community structure in complex networks using fuzzy c-means clustering. *Physica A: Statistical Mechanics and its Applications, 374*(1), 483–490.

Yi-Shan Sung is a PhD candidate in Industrial Engineering at Penn State. She graduated from National Taiwan University with master degree and National Tsing Hua University with undergraduate degree. During her stay at Penn State, she has worked as an intern at Geisinger Health System, Danville, PA. Her research interests are mainly in analytics and in particular graph analytics and health analytics. She is a student member of The Institute of Industrial Engineers (IIE) and The Institute for Operations Research and the Management Sciences (INFORMS).

Dashun Wang is Associate Professor of Management and Organizations at the Kellogg School of Management, and (by courtesy) Industrial Engineering & Management Sciences at the McCormick School of Engineering. At Northwestern, He is also associated with NICO, the Northwestern Institute on Complex Systems. Prior to joining Kellogg, he was Assistant Professor of Information Sciences and Technology at the Pennsylvania State University and a Research Staff Member at the IBM T.J. Watson Research Center. Dashun received his PhD in Physics in 2013 from Northeastern University, where he was a member of the Center for Complex Network Research. From 2009 to 2013, he had also held an affiliation with Dana-Farber Cancer Institute, Harvard University as a Research Associate. Dashun received my B.S. degree in Physics from Fudan University in 2007. He is a recipient of the AFOSR Young Investigator Award (2016).

Soundar Kumara is the Allen, E., and Allen, M., Pearce Professor of Industrial Engineering at Penn State. He also holds a joint appointment with the Department of Computer Science. Has an affiliate appointment with the school of Information Sciences and Technology. His research interests are in Manufacturing Process Monitoring, IT in Manufacturing and Service Sectors, Health Analytics, Graph Analytics and Large Scale Complex Networks. He is a Fellow of Institute of Industrial Engineers (IIE), Fellow of the International Academy of Production Engineering (CIRP), and Fellow of American Association for Advancement of Science (AAAS), and American Association of Mechanical Engineers (ASME).