Physica A 391 (2012) 4165–4180

Contents lists available at SciVerse ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Social structure of Facebook networks

Amanda L. Traud [a,b,c,*], Peter J. Mucha [a,d], Mason A. Porter [e,f]

[a] Carolina Center for Interdisciplinary Applied Mathematics, Department of Mathematics, University of North Carolina, Chapel Hill, NC 27599-3250, USA
[b] Carolina Population Center, University of North Carolina, Chapel Hill, NC 27516-2524, USA
[c] Biomathematics, North Carolina State University, Raleigh, NC 27695, USA
[d] Institute for Advanced Materials, Nanoscience & Technology, University of North Carolina, Chapel Hill, NC 27599-3216, USA
[e] Oxford Centre for Industrial and Applied Mathematics, Mathematical Institute, University of Oxford, OX1 3LB, UK
[f] CABDyN Complexity Centre, University of Oxford, OX1 1HB, UK

ARTICLE INFO

Article history:
Received 7 October 2011
Received in revised form 3 December 2011
Available online 19 December 2011

Keywords:
Community detection
Clustering
Online social networks

ABSTRACT

We study the social structure of Facebook "friendship" networks at one hundred American colleges and universities at a single point in time, and we examine the roles of user attributes – gender, class year, major, high school, and residence – at these institutions. We investigate the influence of common attributes at the dyad level in terms of assortativity coefficients and regression models. We then examine larger-scale groupings by detecting communities algorithmically and comparing them to network partitions based on user characteristics. We thereby examine the relative importance of different characteristics at different institutions, finding for example that common high school is more important to the social organization of large institutions and that the importance of common major varies significantly between institutions. Our calculations illustrate how microscopic and macroscopic perspectives give complementary insights on the social organization at universities and suggest future studies to investigate such phenomena further.

Published by Elsevier B.V.

1. Introduction

Since their introduction, social networking sites (SNSs) such as Friendster, MySpace, Facebook, Orkut, LinkedIn, and myriad others have attracted hundreds of millions of users, many of whom have integrated SNSs into their daily lives to communicate with friends, send e-mails, solicit opinions or votes, organize events, spread ideas, find jobs, and more [1]. Facebook, an SNS launched in February 2004, now overwhelms numerous aspects of everyday life, and it has become an immensely popular societal obsession [1–4]. Facebook members can create self-descriptive profiles that include links to the profiles of their "friends", who may or may not be offline friends. Facebook requires that anybody who wants to be added as a friend have the relationship confirmed, so Facebook friendships define a network (graph) of reciprocated ties (undirected edges) that connect individual users. (In this article, we use the words "edge" and "link" interchangeably.)

The emergence of SNSs such as Facebook and MySpace has revolutionized the availability of social and demographic data, which has in turn had a significant impact on the study of social networks [1,5,6]. It is possible to acquire very large data sets from SNSs, though of course the population online and actively using SNSs is a biased sample of the broader population. Services like Facebook also contain large quantities of demographic data, as many users now voluntarily reveal voluminous amounts of detailed personal information. An especially exciting aspect of studying SNSs is that they provide an opportunity

[*] Corresponding author at: Biomathematics, North Carolina State University, Raleigh, NC 27695, USA. Tel.: +1 919 753 4123.
E-mail address: altraud@ncsu.edu (A.L. Traud).

0378-4371/$ – see front matter. Published by Elsevier B.V.
doi:10.1016/j.physa.2011.12.021

to examine social organization at unprecedented levels of size and detail, and they also provide new venues to test sampling effects [7]. One can investigate the structure of an SNS like Facebook to examine it as a network in its own right, and ideally one can also try to take one step further and infer interesting insights regarding the offline social networks that an SNS imperfectly parallels. Most people tend to draw their Facebook friends from their real-life social networks [1], so it is not entirely unreasonable to use a Facebook network as a proxy for an offline social network. (Of course, as noted by Hogan [8], one does need to be aware of significant limitations when taking such a leap of faith.)

Social scientists, information scientists, and physical scientists have all jumped on the SNS data bandwagon [9]. It would be impossible to exhaustively cite all of the research in this area, so we only highlight a few results; additional references can be found in the review by Boyd and Ellison [1]. Boyd [10,11] also conducted an empirical study of Facebook and MySpace, concluding that Facebook tends to appeal to a more elite and educated cross section than MySpace. The company RapLeaf [12] has compiled global demographics on the age and gender usage of numerous SNSs. Other recent studies have investigated the manifestation on SNSs of race and ethnicity [13], religion [14], gender [15,16], and national identity [17]. Other research has illustrated that online friendship networks can be exploited to improve shopper recommendation systems on websites such as Amazon [18]. (Presumably, this is becoming increasingly prominent in practice.)

Several papers have attempted to increase understanding of how SNS friendships form. For example, Kumar et al. [19] examined preferential attachment models of SNS growth, concluding that it is important to consider different classes of users. Lampe et al. [20] explored the relationship between profile elements and number of Facebook friends, and other scholars have examined the importance of geography [21] and online message activity [22] to online friendship formation. Other papers have established the existence of strong correlations between network participation and website activity, including the motivation of people to join particular groups [23], the recommendations of online groups [24], online messages and friendship formation [22], interaction activity versus sense of belonging [25], and the role of explicit ideological relationship designations in affecting voting behavior [26,27]. Lewis et al. [3] used Facebook data for an entire class of freshmen at an unnamed, private American university to conduct a quantitative study of social networks and cultural preferences. The same data set was also used to examine user privacy settings on Facebook [28].

In the present paper, we study the complete Facebook networks of 100 American colleges and universities from a single-day snapshot in September 2005. This paper is a sequel to our previous research on 5 of these institutions [29], in which we developed some of the methodology that we employ here. In September 2005, one needed a `.edu` e-mail address to become a member of Facebook. We thus ignore links between nodes at different institutions and study the Facebook networks of the 100 institutions as 100 separate networks. For each network, we have categorical data encompassing the gender, major, class year, high school, and residence (e.g., dormitory, House, fraternity, etc.) of the users. We examine homophily and community structure (network partitions that are obtained algorithmically) for each of the networks and compare the community structure to partitions based on the given categorical data. We thereby compare and contrast the organizations of the 100 different Facebook networks, which arguably allows us to compare and contrast the organizations of the underlying university social networks to which they provide an imperfect counterpart. In addition to the inherent interest of these Facebook networks, our investigation is important for subsequent use of these networks – which were formed via ostensibly the same generative mechanism – as benchmark examples for numerous types of computations, such as new community detection methods.

The remainder of this paper is organized as follows. We first discuss the Facebook data and present the methods that we used for testing homophily at the dyad level and demographic organization at the community level. We then present and discuss results on the largest connected components of the networks, student-only subnetworks, and single-gender subnetworks. Finally, we summarize and discuss our findings.

2. Data

The data, sent directly to us by Adam D'Angelo of Facebook, consists of the complete set of users (nodes) from the Facebook networks at each of 100 American institutions (which we enumerate in Table A.1) and all of the "friendship" links between those users' pages as they existed on one particular day in September 2005. Each institution in the data is additionally identified by a number appearing as part of its name that appears to correspond to the order in which each institution "joined" Facebook. Apart from preparing the network representation of friendships, we employed only the first two digits of the user ID numbers. This enabled us to identify the institutional affiliation of each user in the provided list of institutions; we otherwise ignored the additional digits in each ID number. Most of the institutions on the provided list are clearly identified, and there are only a small number of disambiguation problems. For instance, 4 different "UC" institutions plus "Cal" are in the data, and there are 2 "Texas" listings. One could presumably identify these institutions using the complete ID numbers of affiliated users and their corresponding Facebook pages, but we have not used the ID numbers in this way.

Similar snapshots of Facebook data from 10 Texas institutions were analyzed recently by Mayer and Puller [4], and a snapshot from "a diverse private college in the Northeast US" was studied by Lewis et al. [3]. Other studies of Facebook have typically obtained data either through surveys [1] or through various forms of automated sampling [30], thereby missing nodes and links that can impact the resulting graph structures and analyses. We only consider ties between people at the same institution, yielding 100 separate realizations of university social networks and allowing us to compare the structures at different institutions.

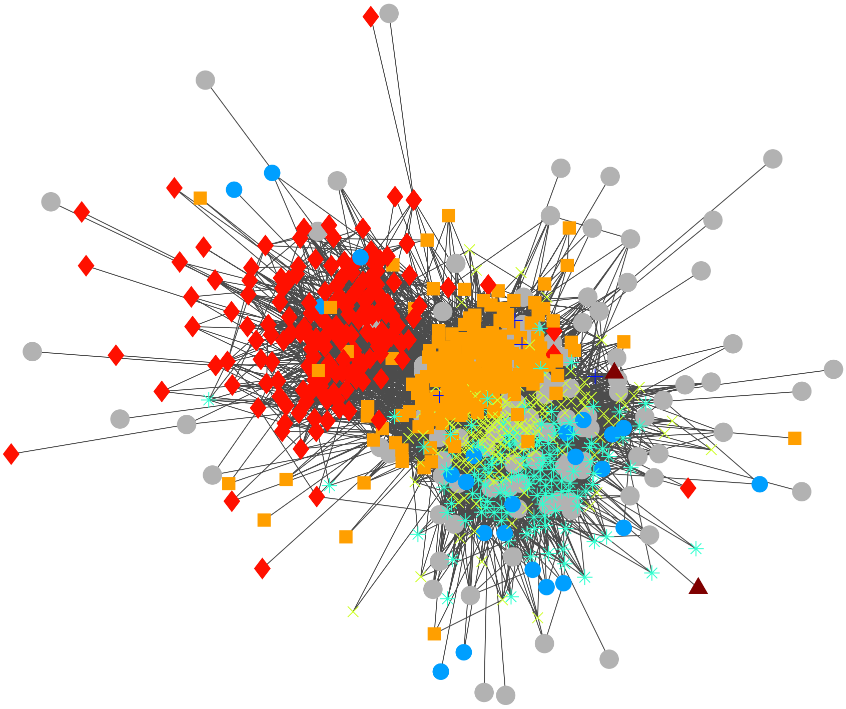


Fig. 1. (Color online) Largest connected component of the student-only subset of the Reed College Facebook network. (We used a Fruchterman–Reingold visualization [31].) Different node shapes and colors indicate different class years (gray circles denote users who did not identify an affiliation), and the edges are randomly shaded for easy viewing. Clusters of nodes with the same color/shape suggest that common class year has an important effect on the aggregate structure.

We consider four networks for each of the 100 Facebook data sets: the largest connected component of the full network (which we hereafter identify as "Full"), the largest connected component of the student-only network ("Student"), the largest connected component of the female-only network ("Female"), and the largest connected component of the male-only network ("Male"). The Male and Female networks are each subsets of the Full network rather than of the Student network. Each network has a single type of unweighted, undirected connection between nodes and can thus be represented as an adjacency matrix **A** with elements $A_{ij} = A_{ji}$ indicating the presence ($A_{ij} = 1$) or absence ($A_{ij} = 0$) of a tie between nodes i and j. The resulting tangle of nodes and links, which we illustrate for the Reed College Student Facebook network in Fig. 1, can obfuscate any organizational structure that might be present.

The data also includes limited demographic (categorical) information that is volunteered by users on their individual pages: gender, class year, and (using numerical identifiers) high school, major, and residence. We use a "Missing" label for situations in which individuals did not volunteer a particular characteristic. The different characteristics allow us to make comparisons between institutions, under the assumption (see the discussion by Boyd and Ellison [1]) that the communities and other elements of structural organization in Facebook networks reflect (even if imperfectly) the social communities and organization of the offline networks on which they are based. It is an important research issue to determine just how imperfect this might be [8], but this is far beyond the scope of the present paper (though we hope that others will take on this particular challenge). The conclusions that we draw in this paper apply directly to the university Facebook networks from a single-day snapshot in September 2005, and we expect that they can provide insight about the real-world social networks at the institutions as well.

3. Methods

We study each network at both the dyad level and the community level. We first consider homophily [32–34], which we quantify by assortativity coefficients using the available categorical data. For some of the smaller networks, we additionally perform independent logistic regression on node pairs to obtain the log odds contributions to edge presence between two nodes that have the same categorical-data value. We similarly fit exponential random graph models (ERGMs) [35–40] with triangle terms to these smaller networks. Finally, we partition the networks by algorithmically detecting communities [41, 42], which we compare to the given categorical data using the technique in this paper's prequel [29]. Calculating assortativity values and log odds contributions allows us to examine "microscopic" features of the networks, and comparing algorithmic partitions of the networks to the categorical data allows us to examine their "macroscopic" features. As we illustrate below, both perspectives are important because they provide complementary insights.

3.1. Assortativity

A general measure of scalar assortativity r relative to a categorical variable is given by Newman [34,43]:

$$r = \frac{\mathrm{tr}(\mathbf{e}) - \|\mathbf{e}^2\|}{1 - \|\mathbf{e}^2\|} \in [-1, 1], \tag{1}$$

where $\mathbf{e} = \mathbf{E}/\|\mathbf{E}\|$ is the normalized mixing matrix, the elements E_{ij} indicate the number of edges in the network that connect a node of type i (e.g., a person with a given major) to a node of type j, and the entry-wise matrix 1-norm $\|\mathbf{E}\|$ is equal to the sum of all entries of E. By construction, this formula yields $r = 0$ when the amount of assortative mixing is the same as that expected independently at random (i.e., e_{ij} is simply the product of the fraction of nodes of type i and the fraction of nodes of type j), and it yields $r = 1$ when the mixing is perfectly assortative.

3.2. Logistic regression and exponential random graphs

We further measure the influence of the available user characteristics on the likelihood of a "friendship" tie via a fit by logistic regression (under an assumption of independent dyads) and by an ERGM specification that includes triangle terms. Our focus is on trying to calculate the propensity for two nodes with the same categorical value to form a tie. We consider each of the four categorical variables (major, residence, year, and high school) and use the `ERGM` package in R [35] for both models (treating each network as undirected). We used R 2.11.1 and the `statnet` package version 2.1–1, and we note that different versions of R and `statnet` caused different degrees of convergence with the structural elements in the model. We obtained results for the 16 smallest institutions. (We did these calculations on a 32-bit operating system, which restricts the network sizes that can be processed.) Both models that we consider are based on a standard ERGM parametrization $P_\theta\{\mathbf{Y} = \mathbf{A}\} = \exp\{\theta \cdot \mathbf{g}(\mathbf{A})\}/\kappa(\theta)$ describing the distribution of graphs with model coefficients θ corresponding to statistics calculated from the adjacency matrix $\mathbf{A}$ (with a normalizing factor κ to ensure that the formula yields a probability distribution) [35–39]. The vector-valued function $\mathbf{g}$ is associated with the corresponding ERGM.

In the first model (logistic regression), we include five statistics (with five corresponding θ coefficients): the total density of ties (`edges`) and the common classifications (`nodematch`) from each of four node/user characteristics: residence, class year, major, and high school. For example, the $\theta_{\text{highschool}}$ contribution describes the additional log-odds predisposition for a "friendship" tie when two users are from the same high school. In all cases, we ignore possible contributions from missing characteristic data: two nodes with the same missing data field are not treated as having the same value for the characteristic. Rather than include gender explicitly in the model, we instead additionally fit the model to the single-gender subnetworks in order to be consistent with the treatment of gender in the community-level comparisons below. In the second model (an ERGM), we add a `triangle` statistic to account for the observed amount of transitivity in the network data. This gives a total of six θ coefficients: edges, common residence, common class year, common major, common high school, and the triangle coefficient.

3.3. Community detection

The global organization of social networks often includes coexisting modular (horizontal) and hierarchical (vertical) organizational structures, and myriad papers have attempted to interpret such organization through the computational identification of "community structure". Communities are defined in terms of cohesive groups of nodes with more internal connections (between nodes in the same group) than external connections (between nodes in the group and nodes in other groups). As discussed at length in two recent review articles [41,42] and in references therein, the ensemble of techniques available to detect communities is both numerous and diverse. Existing techniques include hierarchical clustering methods such as single linkage clustering, centrality-based methods, local methods, optimization of quality functions such as modularity and similar quantities, spectral partitioning, likelihood-based methods, and more. Communities are considered to not be merely structural modules but are also expected to have functional importance because of the large number of common ties among nodes in a community. For example, communities in social networks might correspond to circles of friends or business associates, and communities in the World Wide Web might encompass pages on closely-related topics. In addition to remarkable successes on benchmark problems, investigations of community structure have observed correspondence between communities and "ground truth" groups in diverse application areas—including the reconstruction of college football conferences [44] and the investigation of such structures in algorithmic rankings [45]; the investigation of committee assignments [46], legislation cosponsorship [47], and voting blocs [48,49] in the United States Congress; the examination of functional groups in metabolic networks [50]; the study of ethnic preferences in school friendship networks [51]; and the study of social structures in mobile-phone conversation networks [52].

In the present paper, we investigate the community structures of the Facebook networks from each of the 100 colleges and universities. (See the visualization of the community structure for Reed College in Fig. 2.) For each institution, we

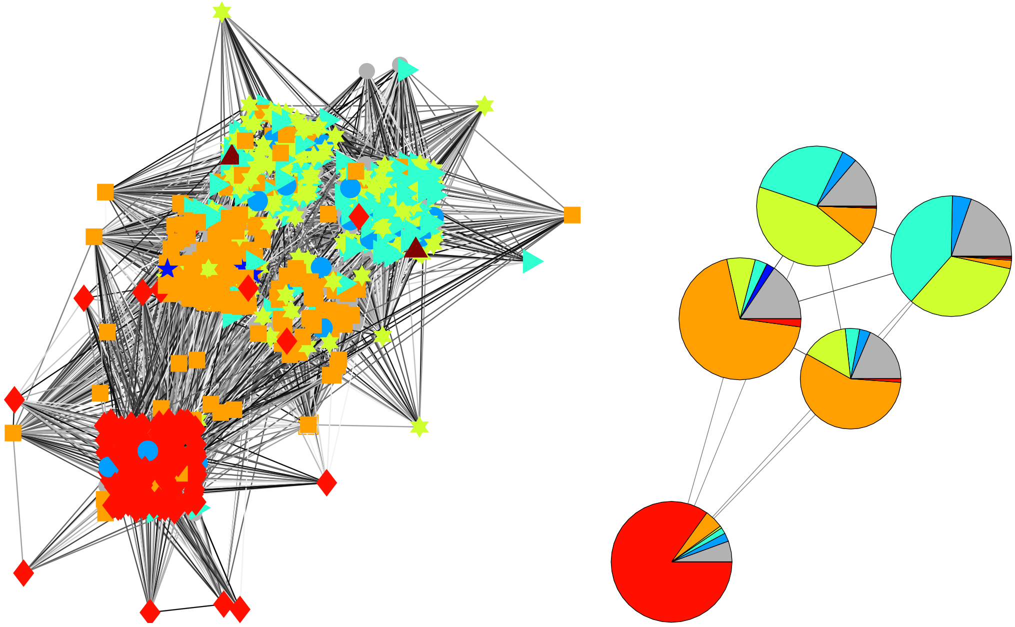


Fig. 2. (Color online) (Left) Visualization of community structure of the Reed College Student Facebook network shown in Fig. 1. Node shapes and colors indicate class year (gray dots denote users who did not identify an affiliation), and the edges are randomly shaded for easy viewing. We place the communities using a Fruchterman–Reingold [31] layout, and we use a Kamada–Kawaii [53] layout to position the nodes within communities [54]. (Right) The same network layout but with each community depicted as a pie. Larger pies represent communities with larger numbers of nodes. Darker edges indicate the presence of more connections between the associated communities.

consider the Full, Student, Female, and Male networks. We seek to determine how well the demographic labels included in the data correspond to algorithmically computed communities. Assortativity provides a local measure of homophily, but that does not provide sufficient information to draw conclusions about the global organization of the Facebook networks. For example, two students who attended the same high school are typically more likely to be friends with each other than are two students who attended different high schools, but this will not necessarily have a meaningful community-level effect unless enough of the students went to common high schools. As we will see below, high school tends to be a much more dominant organizing characteristic of the social structure at the large institutions than at small institutions, presumably because of a significant frequency of common high-school pairs at the large institutions.

We identify communities by optimizing the "modularity" quality function $Q = \sum_i (e_{ii} - b_i^2)$, where e_{ij} denotes the fraction of ends of edges in group i for which the other end of the edge lies in group j and $b_i = \sum_j e_{ij}$ is the fraction of all ends of edges that lie in group i. High values of modularity correspond to community assignments with greater numbers of intra-community links than expected at random (with respect to a particular null model [41,42,55]). Although numerous other community detection methods are also available, modularity optimization is perhaps the most popular way to detect communities and it has been successfully applied to many applications [41,42]. One might also consider using a method that includes a resolution parameter [56] to avoid issues with resolution limits [57]. However, our primary focus is on global organization of the networks, so we limit our attention to the default resolution of modularity. This focus arguably biases our study of communities to large structures, such as those influenced by common class year, making the observed correlations with other demographic characteristics even more striking.

To try to ensure that the communities we detect are properties of the data rather than of the algorithms that we used, we optimize modularity (with default resolution) using 6 different combinations of spectral optimization, greedy optimization, and Kernighan–Lin (KL) node-swapping steps [58] (in the manner discussed by Newman [59]). Specifically, we use (1) recursive partitioning by the leading eigenvector of a modularity matrix [55], (2) recursive partitioning by the leading pair of eigenvectors (including an extension [60] of the method in Ref. [55]), (3) the Louvain greedy method [61], and each of these three choices supplemented with small increases in the quality Q that can be obtained using KL node swaps. Each of these 6 methods yields a partition into disjoint communities, and we obtain our comparisons (described in Section 3.4) by considering each of these 6 partitions.

Modularity optimization is NP-hard [62], so one must be cautious about the large number of near-degenerate partitions in the modularity landscape [63]. However, by detecting coarse observables – in particular, the global organization of a Facebook network based on the given categorical data – and considering results that are averaged over multiple optimization methods, one can obtain interesting insights. The specific "best" partition will vary from one method to another, but some of the predicted coarse organizational structure of the networks (see below) is robust to the choice of community detection algorithm.

3.4. Comparing communities to node data

Once we have detected communities for each institution, we will compare the algorithmically-obtained community structure to the available categorical data for the nodes. We recently developed a methodology to accomplish this goal in Ref. [29] (where we considered only 5 institutions among the 100 in order to illustrate the techniques). This method of comparison can be applied to the output of any "hard partitioning" algorithm, in which each node is assigned to precisely one community (cf. "soft partitioning" methods, in which communities can overlap). We briefly review that methodology here.

To compare a network partition to the categorical demographic data, we standardize (using a z-score) the Rand coefficient of the communities in that partition compared to partitioning based purely on each of the four categorical variables (one at a time). For each comparison, we calculate the Rand z-score z in terms of the total number of pairs of nodes in the network M, the number of pairs that are in the same community M_1, the number of pairs that have the same categorical value M_2, and the number of pairs of nodes that are both in the same community and have the same categorical value w [29]. The Rand coefficient is given in term of these quantities by $S = [w + (M - M_1 - M_2 + w)]/M$ [64]. We then calculate the z-score for the Rand coefficient [29,65]:

$$z = \frac{1}{\sigma_w}\left(w - \frac{M_1 M_2}{M}\right), \tag{2}$$

where

$$\sigma_w^2 = \frac{M}{16} - \frac{(4M_1 - 2M)^2(4M_2 - 2M)^2}{256M^2} + \frac{C_1 C_2}{16n(n-1)(n-2)} + \frac{[(4M_1 - 2M)^2 - 4C_1 - 4M][(4M_2 - 2M)^2 - 4C_2 - 4M]}{64n(n-1)(n-2)(n-3)}, \tag{3}$$

n is the number of nodes in the network, the coefficients C_1 and C_2 are given by

$$C_1 = n(n^2 - 3n - 2) - 8(n+1)M_1 + 4\sum_i n_{i\cdot}^3,$$
$$C_2 = n(n^2 - 3n - 2) - 8(n+1)M_2 + 4\sum_j n_{\cdot j}^3, \tag{4}$$

n_{ij} denotes an element of a contingency table and indicates the number of nodes that are classified into the ith group of the first partition and the jth group of the second partition, $n_{i\cdot} = \sum_j n_{ij}$ is a row sum, and $n_{\cdot j} = \sum_i n_{ij}$ is a column sum. Each z-score indicates the deviation from randomness in comparing the community structure with the partitioning based purely on that single demographic characteristic. One needs to be cautious when interpreting such deviations from randomness as strengths of correlation. In particular, given the dependence on system size inherent in this measure, one should not overinterpret the relative values of z-scores from different institutions. Nevertheless, the z-scores provide a reasonable proxy quantity both for the statistical significance of correlation and for the relative strengths of correlation in a specified network.

4. Results

We now use the methods outlined in the previous section to study the Facebook networks. We first follow the order of presentation above and then make some observations in combinations. Complete results are available in the tables in the Supplementary Data.

4.1. Assortativity

We tabulate the assortativities based on gender, major, residence, class year, and high school for all networks (and subsets thereof) in Table A.2.

For almost all of the institutions and each of the 4 network subsets, the class year attribute produces higher assortativity values than the other available demographic characteristics. However, Rice University (31), California Institute of Technology (36), University of Georgia (50), Mich (67), Auburn University (71), and University of Oklahoma (97) are each examples in which residence provides the highest assortativity values (again, for each of the 4 network subsets). We discussed Caltech (i.e., California Institute of Technology) as a focal example in Ref. [29], in which we introduced the community comparison methods that we employ below.

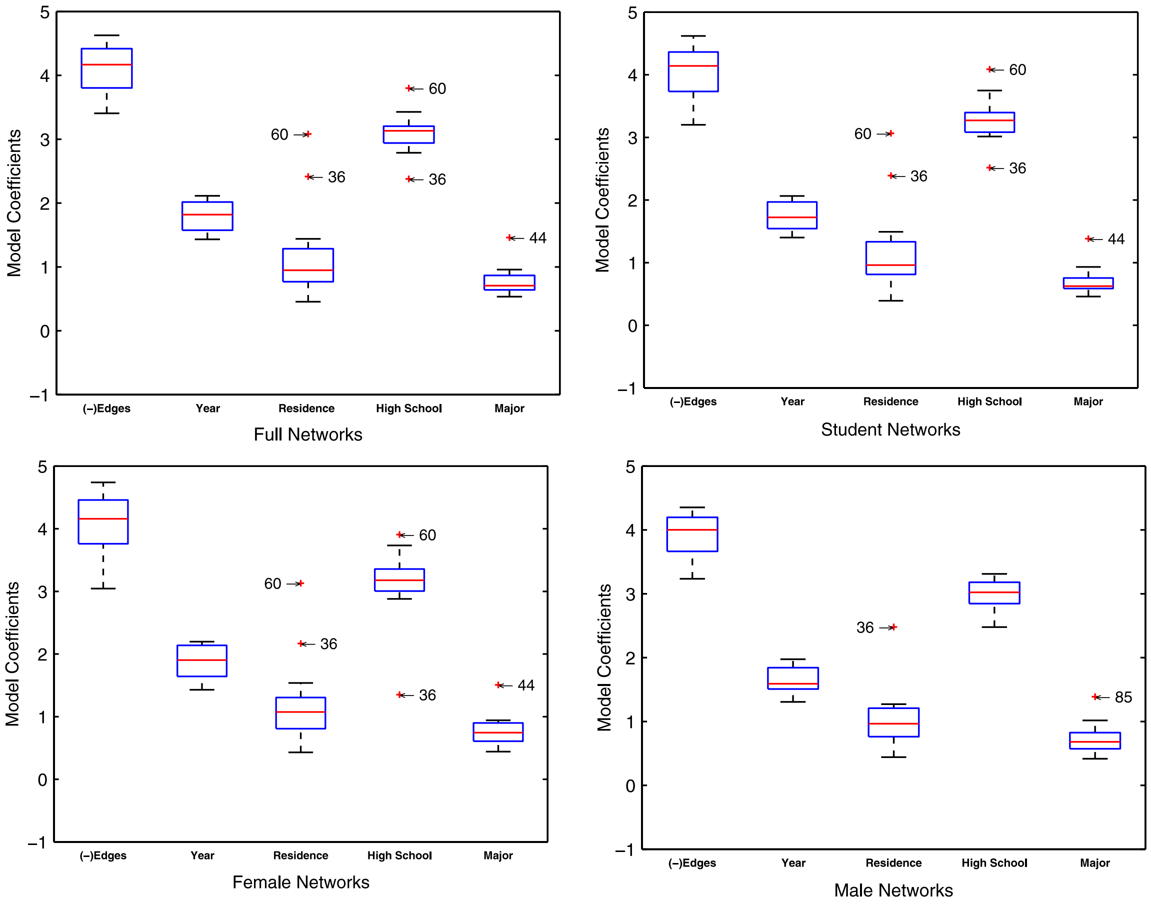


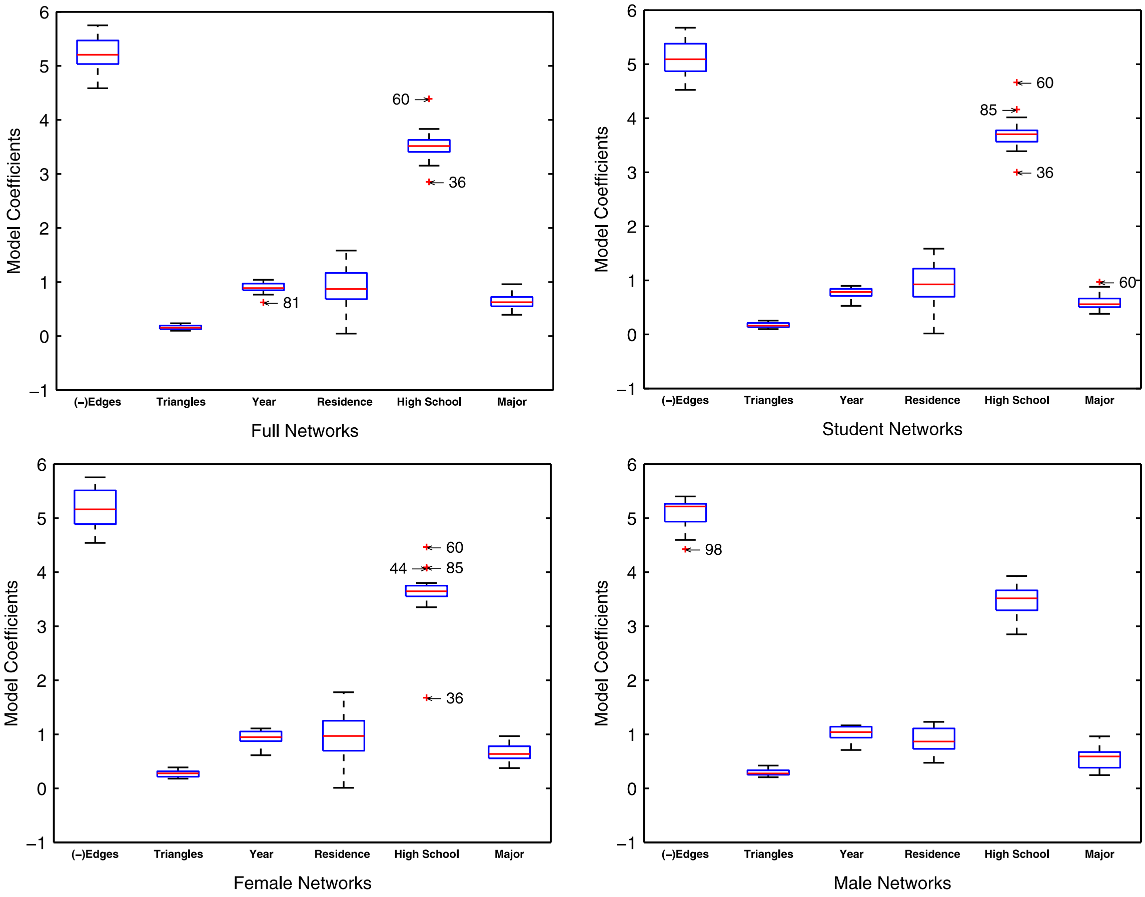
Fig. 3. (Color online) Box plots (indicating median, quartiles, extent, and outliers of the distribution) of the logistic regression `nodematch` coefficients for the 16 smallest institutions in the data for the model described in the main text. We plot the $-\theta_{\text{edges}}$ values to present results with greater resolution. We separately present our results for the Full, Student, Female, and Male networks.

Other institutions have varying orderings of class year and residence assortativity among the 4 network subsets. At MIT (8), USF (51), Notre Dame (57), University of Maine (59), UC (61), UC (64), and MU (78), residence gives the highest assortativity in the Male networks. The UCF (52) Female network has its highest assortativity with residence. The Full network and the Male network for University of California at Santa Cruz (68) have their highest assortativity values with residence. Both the Male and Female networks at UIllinois (20), Tulane (29), UC (33), Florida State University (53), Cal (65), University of Mississippi (66), University of Indiana (69), Texas (80), Texas (84), University of Wisconsin (87), Baylor (93), University of Pennsylvania (94), and University of Tennessee (95) have their highest assortativity values with residence; all other networks from these institutions have their highest assortativity values with class year.

Some outlying observations can be tied directly to small samples. For example, Simmons (81) is a female-only college. It has only four males in the Full network; none of the males had any connections with another male, so the gender assortativity values for both the Full and Student networks are very close to 0. Similar gender numbers are also present in the data from Wellesley (22) and Smith (60).

4.2. Dyad-level regression and exponential random graphs

We use the two statistical models described in Section 3.2 to study the 16 smallest institutions. The (dyad-independent) logistic regression model includes contributions from edges (network density) and matched user (node) characteristics for each of four demographic variables. We present the results for this model in Table A.3. The second model that we consider is an ERGM, which supplements the first model with a structural `triangle` contribution. We present the results for this model in Table A.4. These calculations give views of the networks at the microscopic (dyad-level) scale that supplement the results that we obtained using the assortativity statistics.


Full Networks

Student Networks

Female Networks

Male Networks

Fig. 4. (Color online) Box plots (indicating median, quartiles, extent, and outliers of the distribution) of the exponential random graph model coefficients described in the main text for the 16 smallest institutions in the data. We plot the $-\theta_{\text{edges}}$ values to present results with greater resolution. We separately present our results for the Full, Student, Female, and Male networks.

We consider the results from the 16 smallest institutions by fitting the models to each of their Full, Student, Female, and Male networks. Because each of the resulting model coefficients appears to be statistically significant at a p-value of less than 10^{-4}, we interpret the importance of node matching on the different demographic characteristics directly from the magnitude of the corresponding model coefficients. We summarize the results for these 16 institutions using the box plots in Figs. 3 and 4. The box plots identify the outliers by institution number: Caltech (36), Oberlin (44), Smith (60), Simmons (81), Vassar (85), and Reed (98). (As we have only performed this regression for the 16 smallest institutions in the data, one should not jump to conclusions from this list of outliers.) For all institutions and all 4 types of networks for each institution, the highest coefficient in the employed ERGM model (with `triangle` terms) is given for matching the high school category, and the value of this coefficient is significantly higher than those for the other node-matching coefficients. Only the Caltech (36) Female network has ERGM coefficients for year, residence, and high school that are very close to each other. For each network, both of these models reported convergence after three iterations [35].

4.3. Comparison of communities

We now discuss community-level results for each network using z-scores of the Rand coefficient to compare partitions obtained via algorithmic community detection to partitions based on each characteristic. That is, each community detection result identifies a group assignment for each node, thereby producing a network partition (called a "hard" partition) in which each node is assigned to exactly one community. One can also obtain a hard partition for each network by selecting a single characteristic and grouping nodes according to that characteristic. Every network that we study (including the subnetworks) has at least one z-score in the set $\{z_{\text{Major}}, z_{\text{Year}}, z_{\text{HS}}, z_{\text{Residence}}\}$ with a value greater than 5. Although the distribution of Rand coefficients is decidedly not Gaussian, particularly in the tails of the distributions [29,66,67], this $z = 5$ threshold indicates that at least one characteristic in each network exhibits strong statistical significance. Moreover, the vast majority of our comparisons (see Table A.5) exceed the $z = 2$ threshold. (That is, they essentially lie outside 95% confidence intervals.)

To visualize and compare the varied strengths of organization according to the different demographic characteristics, we represent the four z-scores obtained for each network (Full, Student, Female, and Male) of an institution using 3-dimensional barycentric (tetrahedral) coordinates [68,69]. We start by setting all negative z-scores to 0, as all observed negative z-score values are small enough to be statistically insignificant. We then normalize by the sum of the z-scores to obtain

$$\begin{aligned} z_1 &= \frac{z_{\text{Major}}}{z_{\text{Major}} + z_{\text{Year}} + z_{\text{HS}} + z_{\text{Residence}}}, \\ z_2 &= \frac{z_{\text{Residence}}}{z_{\text{Major}} + z_{\text{Year}} + z_{\text{HS}} + z_{\text{Residence}}}, \\ z_3 &= \frac{z_{\text{Year}}}{z_{\text{Major}} + z_{\text{Year}} + z_{\text{HS}} + z_{\text{Residence}}}, \\ z_4 &= \frac{z_{\text{HS}}}{z_{\text{Major}} + z_{\text{Year}} + z_{\text{HS}} + z_{\text{Residence}}}. \end{aligned} \tag{5}$$

From these four z-score values, we calculate coordinates $X = (x_1, x_2, x_3)$ located inside a tetrahedron. For example, one can obtain a tetrahedron whose vertices are $p_1 = (1, 0, 0)$, $p_2 = (\cos(2\pi/3), \sin(2\pi/3), 0)$, $p_3 = (\cos(4\pi/3), \sin(4\pi/3), 0)$, and $p_4 = (0, 0, \sqrt{2})$ with the transformation

$$\begin{aligned} X &= (T \times Z) + p_4, \\ T &= \begin{bmatrix} p_1 - p_4 & p_2 - p_4 & p_3 - p_4 \end{bmatrix}, \\ Z &= \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix}. \end{aligned} \tag{6}$$

The information from $z_4 = 1 - (z_1 + z_2 + z_3)$ is implicitly included in (6) because of the normalization. Each of the 4 vertices of the tetrahedron corresponds to a limit in which the corresponding z-score completely dominates the other three z-scores. That is, at a vertex, the entire z-score sum arises from the corresponding component.

Because of the strong role of class year, we visualize the tetrahedra from a perspective located above the vertex corresponding to class year and project the result into the opposing face of the tetrahedron. We calculate the point X for each of the 6 algorithmic partitions of each network (i.e., using the aforementioned 6 different community detection methods). For each institution, we plot a disk whose center lies at the midpoint of these 6 sets of X coordinates. The width of each disk is proportional to the maximum difference between a pair of these 6 sets of coordinates (with these distances separated into bins of width 0.1, as indicated in the legends of Figs. 5–8). For example, in Fig. 5, the Pepperdine (86) results have a maximum distance of 0.0141 between partitions, so Pepperdine (86) is represented by one of the smallest disks. Harvard (1) has a maximum distance of 0.1581 between partitions; this lies in [0.1, 0.2), so Harvard (1) is represented by one of the disks of second smallest size. We emphasize that the computed differences are much larger than what one sees using the depicted disks, whose sizes allow one to discern the results from different institutions.

In Figs. 5–8, we show each of the 100 institutions, identified by number (see Table A.1), using a disk that we have color-coded according to the Cartesian distance of its center from the Year vertex. Class year is the predominant organizing category among the ones present in the data, so most of the institutions are located very close to the Year vertex. We zoom in on the Year vertex for each figure in order to better discern the relative importance of class year at the institutions. Importantly, the social organizations of a few institutions differ considerably from those of the majority. Each of these institutions lies close to the Residence vertex, so their community structures are organized predominantly according to dormitory residence. Foremost among these institutions are Rice (31) and California Institute of Technology (36). As we discussed in Ref. [29], California Institute of Technology (Caltech) is well known to be organized almost exclusively according to its undergraduate "House" system [70].

Because we repeatedly observe a strong correlation of class year with community structure, it is relevant to recall that the community detection method that we have employed optimizes modularity at the default resolution. Because of the resolution limit of modularity [57], it might be interesting to explore individual networks at different scales using resolution parameters [41,42,56]. We reiterate, however, that our focus in the present paper is on large-scale features of network partitions rather than on the precise community affiliations of nodes in such partitions.

In Fig. 5, we show the social organization tetrahedron for the Full networks (i.e., for the largest connected components of the complete networks) for all institutions. Although the community structures of nearly all of the Full networks are organized overwhelmingly by class year, a few of them are also heavily influenced by dormitory residence. (We already mentioned above that Rice (31) and Caltech (36) are organized predominantly by residence.) For example, dormitory residence also dominates the community structure at UC Santa Cruz [UCSC] (68), though to a lesser extent than at Rice and Caltech. We also observe relatively high residence z-scores at Smith (60), Auburn (71), and University of Oklahoma (97). Major seems to be most important relative to the other available characteristics at Oberlin (44) and Maine (59), though in both cases its relative importance pales in comparison to that of class year. High School seems to have its largest importance

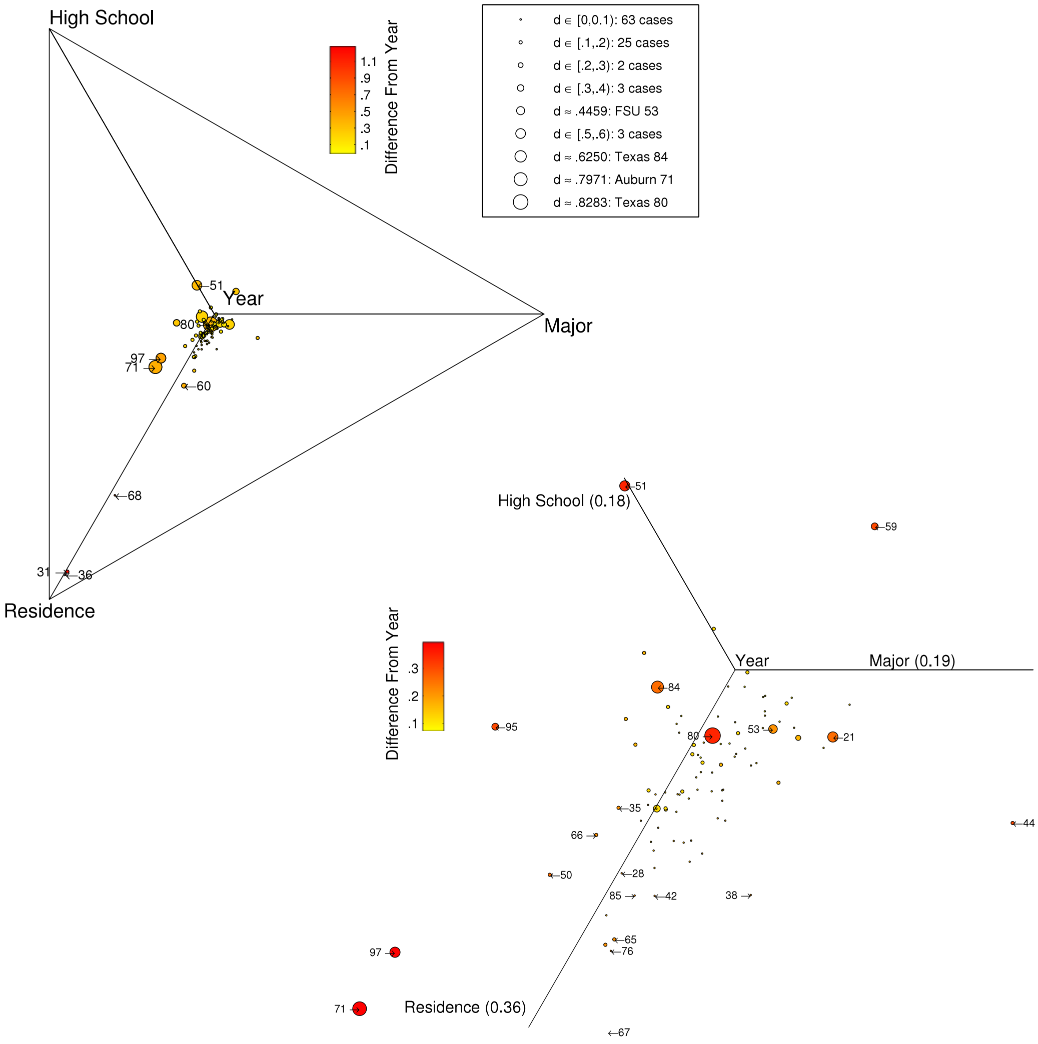


Fig. 5. (Color online) (Upper Left) Social organization tetrahedron for the community structures of the Full component (i.e., largest connected component) of the networks for each of the 100 institutions. Lighter disks indicate an organization that is based more predominantly on class year. See the main text for a description of this figure. (Lower Right) Magnification near the Year vertex. The legend illustrates the disk size as a function of the maximum distance d between a pair of the 6 different partitions of the network. Most cases (88 out of 100 institutions) have $d < 0.2$.

at USF (51) and Tennessee (95), though class year is again even more important. Most of the institutions are clustered tightly near the Year vertex, but Residence can often be rather important (and is sometimes even the most important category, as we have seen in three cases).

In Fig. 6, we show the social organization tetrahedron for the Student networks (i.e., for the largest connected components of the student-only subnetworks) for all institutions. As we saw with the Full networks, most of the institutions have community structures that are organized overwhelmingly according to class year. Rice, Caltech, Smith, UCSC, Auburn, and Oklahoma are again exceptions, as dormitory residence also exerts considerable (or even primary) influence at these institutions. Additionally, considering the Student networks reduces the relative dominance of the Year vertex, although it clearly still dominates the social organization. This feature is illustrated by institutions such as UC (64), UF (21), and Rutgers (89).

In Fig. 7, we show the social organization tetrahedron for the Female networks (i.e., for the largest connected components of the female-only subnetworks) for all institutions. Class year is once again the overwhelmingly dominant organizing characteristic, and dormitory residence is again important at institutions such as Rice, Caltech, Smith, UCSC, Auburn, and Oklahoma. However, we now observe an increased importance of the High School vertex. USF (51), Tennessee (95), UF (21), FSU (53), and GWU (54) all lie closer to the High School vertex than was the case in the Full and Student networks.

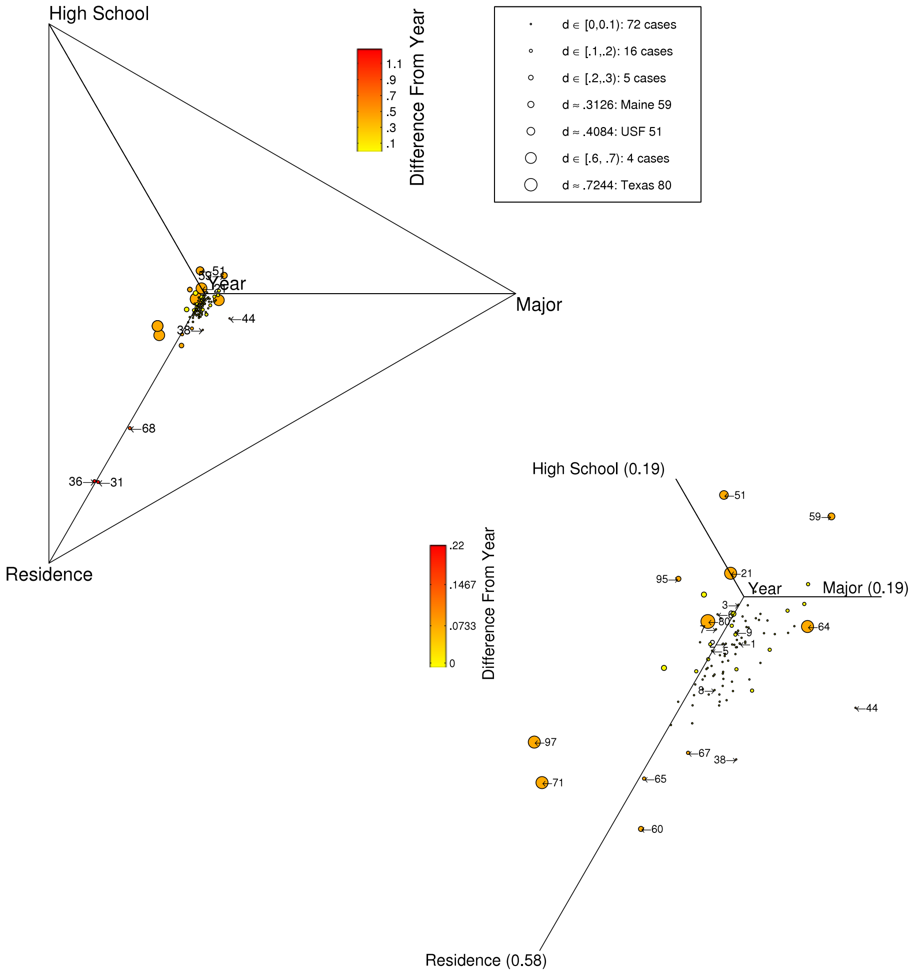


Fig. 6. (Color online) (Upper Left) Social organization tetrahedron for the community structures of the Student component of the networks for each of the 100 institutions. Lighter disks indicate an organization that is based more predominantly on class year. See the main text for a description of this figure. (Lower Right) Magnification near the Year vertex. As in Fig. 5, the disk sizes correspond to the maximum distances between partitions.

In Fig. 8, we show the social organization tetrahedron for the Male networks (i.e., for the largest connected components of the male-only subnetworks) for all institutions. Class year is once again the overwhelmingly dominant organizing characteristic, and dormitory residence is again the most important category at institutions such as Rice, Caltech, and UCSC. Interestingly, considering the Male network suggests that residence is the most important factor for the social organization for the males at Notre Dame (57). Residence also exerts an important influence on the males at Mich (67). This is starkly different from what we observed for these institutions in the Full, Student, and Female networks (and would seem to be something interesting to investigate more thoroughly in the future using other data and methods). The Male UCF (52), MSU (24), USF (51), Auburn (71), and Maine (59) networks are strongly influenced by High School. The Male networks at Texas (80), Rutgers (89), and UIllinois (20) stand out from other universities because of their proximity to the Major vertex. This is true for Oberlin (44) as well, though one observes this for all 4 networks for this institution.

4.4. Discussion

As described above, we see using the z-scores of the Rand coefficients for demographic characteristics versus algorithmic community assignments that class year is the strongest organizing factor at most institutions and that residence is much more important for the community organization at some institutions than at others. The importance of residence is especially prominent at Rice (31) and Caltech (36). We also observe that the Male networks tend to be more scattered around the Year vertex, as some institutions exhibit a stronger correlation with major, whereas others have a stronger correlation with high school. This suggests that there are potential differences in the gender patterns of friendships, which would be interesting to investigate in future studies with different data. We do not explore this general issue further and instead

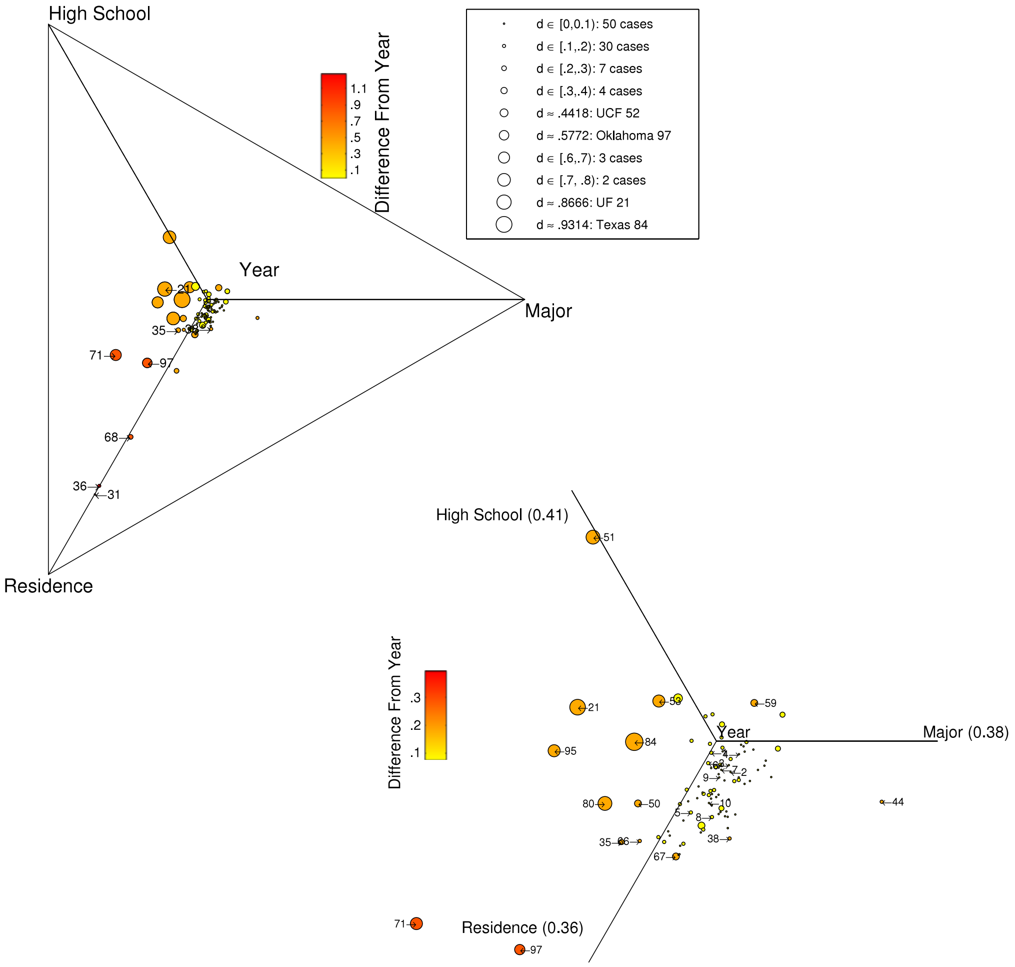


Fig. 7. (Color online) (Upper Left) Social organization tetrahedron for the community structures of the Female component of the networks for each of the 100 institutions. Lighter disks indicate an organization that is based more predominantly on class year. See the main text for a description of this figure. (Lower Right) Magnification near the Year vertex. As in the two previous figures, the disk sizes indicate the maximum distances between partitions.

attempt to identify interesting comparisons with the results that we obtained above. Although it is of course impossible to be exhaustive in our observations, we present all of our assortativity values, regression-model coefficients, and community-comparing *z*-scores in the tables in Supplementary Data part A. We also highlight some interesting facets of our results.

Of particular interest is the comparison of results from the dyad-level regression models to those from community-level correlations. We note, in particular, that the logistic regression and exponential random graph model that we employed for the smallest 16 institutions specify that almost all institutions and all of their subnetworks give the highest model-coefficient contribution toward the presence of edges between nodes from common High Schools. However, as we have seen – and which is particularly evident using the visualizations with tetrahedra – at the community level, most institutions are organized by class year and have a relatively small correlation with high school.

Even in the rare cases in which the rank ordering of the four categories (year, residence, major, and high school) at the community level matches that obtained via dyad-level model coefficients, such as with the logistic regression model for the Full and Female networks from Caltech (36), the relative sizes of the contributions at the dyad level are completely different from those observed at the community level. Caltech supplies an illustrative example of the different insights obtained from community detection versus logistic regression and exponential random graph models both because of its small size and because of its outlying correlation with dormitory residence at the community level. A simple interpretation of the apparent dichotomy between the dyad-level model coefficients and the correlations at the community level is that the presence of two students from the same high school at a small institution like Caltech yields a significant increase in the likelihood of a tie between those students. Even though the corresponding model coefficient is smaller than in any of the other of the 16 smallest institutions, it is comparable to that for common residence (called "Houses" at Caltech). Nevertheless, the very small number of node pairs (relative to the total number of such pairs) at Caltech that have matching high schools has a

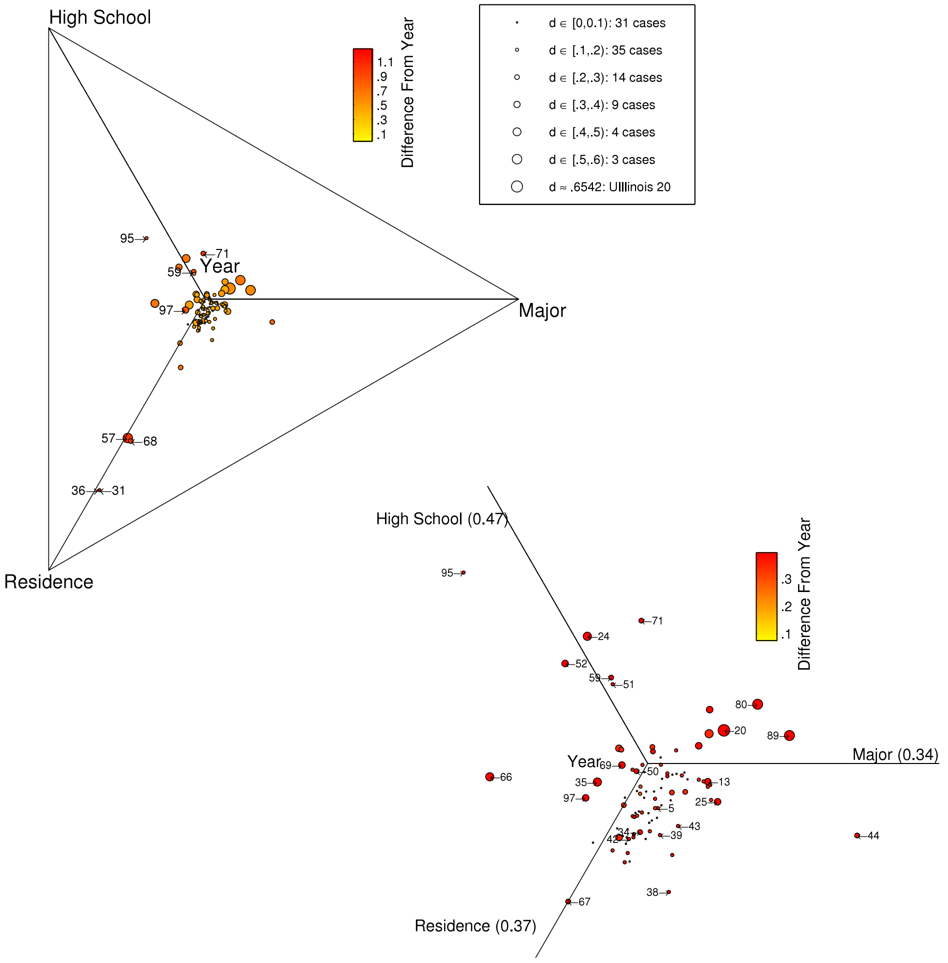


Fig. 8. (Color online) (Upper Left) Social organization tetrahedron for the community structures of the Male component of the networks for each of the 100 institutions. Lighter disks indicate an organization that is based more predominantly on class year. See the main text for a description of this figure. (Lower Right) Magnification near the Year vertex. As in the three previous figures, disk size indicates the maximum distance between partitions. We note that there are more $d > 0.2$ cases here than in the previous figures. This illustrates the greater variability in the relative positions of the z-scores in the different Male networks than was the case for the Full, Student, and Female networks.

very small effect at the community level, as the algorthmically-obtained communities are correlated overwhelmingly with House affiliation. The ERGM result with triangle contributions makes this distinction even more striking, as the common high-school coefficient is actually larger than the coefficient from common House.

We also observe other features that might be worthy of future investigation using other data sets and methodologies. We report the results of our calculations in depth in Tables A.1–A.5. Here we highlight only a few potentially interesting examples in which different methods or different subnetworks yield apparently different qualitative conclusions. For example, we found that major is the second most important factor for the organization of the communities in all of the Oberlin (44) networks, but only for the Full and Male networks does the logistic regression give the second highest coefficient for major. We also observed that the relative ordering of major at the same institution is sometimes gender-dependent. For example, major gives the second largest z-score in the Female and Male networks of Stanford (3), but it gives the fourth largest z-score in Stanford's Full network. Even more interesting, major gives the second largest z-score for the Female network at UVA (16), the third largest z-score for UVA's Male network, and the fourth largest z-score for its Full network. The communities in the Auburn (71) Female network are dominated by residence, but those in the other Auburn networks are not. Similarly, the communities in the MIT (8) Male network are dominated by residence, but those in the other MIT networks are not. Another interesting disparity based on gender occurs in the communities in the Tennessee (95) networks. High school is the primary organizing factor for the Male network, the secondary organizing factor for the Student network, and the tertiary organizing factor for the Female and Full networks.

5. Conclusions

We have studied the social structure of Facebook "friendship" networks at one hundred American institutions at a single point in time (using data from September 2005). To compare the organizations of the 100 institutions using categorical data, we considered both microscopic and macroscopic perspectives. In particular, calculating assortativity coefficients and regression-model coefficients based on observed ties allows one to examine homophily at the local level, and algorithmic community detection allows a complementary macroscopic picture. These approaches complement each other, providing different perspectives on investigations of these Facebook networks. Such complementary calculations are particularly valuable when the microscopic and macroscopic perspectives identify different dominant contributions. For example, in the Caltech networks, the assumed ground truth of the importance of the House system is captured better by computing community structure.

This "real-world ensemble" of 100 networks formed by ostensibly similar mechanisms has the potential to provide a testing ground for various models of network formation. Because of the useful comparisons such an ensemble can facilitate, this data will similarly be useful for studies of dynamic processes on networks, algorithmic community detection, and so on. Because of the different rates of initial Facebook adoption at different institutions, the single point in time represented by the data might usefully describe different stages in the formation of an online social network. In order to pursue such ideas further, one needs to start by studying the networks for their own sake and comparing their structures. This was the goal of the present paper. In particular, we have identified some of the key differences across these 100 realizations of online social networks.

Some of our observations confirm conventional wisdom or are intuitively clear, providing soft verification of our investigation via expected results. For example, we found that class year is often important, Houses are important at Caltech, and high school plays a greater role in the social organization of large universities than it does at smaller institutions (where there are typically fewer pairs of people from the same high school). Other results are quite fascinating and merit further investigation. In particular, the differences in the community structures of the female-only and male-only networks would be interesting to investigate in both offline and online settings. The Facebook data suggests that women are typically more likely to have friends within their common residence (among the demographic data to which we have access) but that the characteristics in the communities in the male-only networks exhibit a wider variation. Investigating this thoroughly would require different data sets and methodologies, especially if one wishes to discern the causes of such friendships from observed correlations.

The Facebook networks that we study offer imperfect counterparts of corresponding real-life social networks, which have different properties from online social networks. It is thus crucial that our results are complemented by studies of the corresponding real networks in order to quantify the extent of such differences.

Acknowledgements

We thank Adam D'Angelo and Facebook for providing the data used in this study. We also acknowledge Sandra González-Bailón, Eric Kelsic, and Fred Stutzman for useful discussions. We thank Christina Frost for her help with developing some of the employed graph visualization code (available at http://netwiki.amath.unc.edu/VisComms) [54]. We would also like to thank James Fowler, Jim Moody, and Tom Snijders for their helpful insight. ALT was funded by the NSF through the UNC AGEP (NSF HRD-0450099) and by the UNC ECHO program. PJM was funded by the NSF (DMS-0645369) and the UNC ECHO program. MAP acknowledges a research award (#220020177) from the James S. McDonnell Foundation.

Appendix. Supplementary data

Supplementary material related to this article can be found online at doi:10.1016/j.physa.2011.12.021.

References

[1] D.M. Boyd, N.B. Ellison, Social network sites: definition, history, and scholarship, Journal of Computer-Mediated Communication 13 (1) (2007) article 11.
[2] D.M. Boyd, Why youth (heart) social network sites: the role of networked publics in teenage social life, in: D. Buckingham (Ed.), MacArthur Foundation Series on Digital Learning—Youth, Identity, and Digital Media Volume, MIT Press, Cambridge, MA, 2007, pp. 119–142.
[3] K. Lewis, J. Kaufman, M. Gonzalez, M. Wimmer, N.A. Christakis, Tastes, ties, and time: a new (cultural, multiplex, and longitudinal) social network dataset using Facebook.com, Social Networks 30 (4) (2008) 330–342.
[4] A. Mayer, S.L. Puller, The old boy (and girl) network: social network formation on university campuses, Journal of Public Economics 92 (2008) 329–347.
[5] V. Krebs, Social network analysis software & services for organizations, communities, and their consultants, 2008., http://www.orgnet.com.
[6] L.A. Lievrouw, S. Livingstone (Eds.), The Handbook of New Media, updated student ed., Sage Publications Ltd., London, UK, 2005.
[7] M. Kurant, M. Gjoka, C.T. Butts, A. Markopoulou, Walking on a graph with a magnifying glass: stratified sampling via weighted random walks, in: Proceedings of ACM SIGMETRICS'11, San Jose, CA, June 2011.
[8] B. Hogan, A comparison of on and offline networks through the Facebook API, Working Paper, 2009. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1331029.
[9] S. Rosenbloom, On Facebook, scholars link up with data, New York Times, 17 December 2007.
[10] D.M. Boyd, Viewing American class divisions through Facebook and Myspace, Apophenia Blog Essay, June 24, 2007. http://www.danah.org/papers/essays/ClassDivisions.html.

[11] D.M. Boyd, White flight in networked publics? How race and class shaped american teen engagement with Myspace and Facebook, in: L. Nakamura, P. Chow-White (Eds.), Digital Race Anthology, in: Race After the Internet, Routledge Press, 2011, pp. 203–222.

[12] V. Sodera, Rapleaf study reveals gender and age data of social network users (press release). Available at: http://business.rapleaf.com/company_press_2008_07_29.html, 2008.

[13] R. Gajjala, Shifting frames: race, ethnicity, and intercultural communication in online social networking and virtual work, in: M.B. Hinner (Ed.), The Role of Communication in Business Transactions and Relationships, Peter Lang, New York, NY, 2007, pp. 257–276.

[14] R. Nyland, C. Near, Jesus is my friend: religiosity as a mediating factor in internet social networking use, Paper Presented at AEJMC Midwinter Conference, Reno, NV, 2007.

[15] N.W. Geidner, C.A. Fook, M.W. Bell, Masculinity and online social networks: male self-identification on Facebook.com, Paper Presented at Eastern Communication Association 98th Annual Meeting, Providence, RI, 2007.

[16] L. Hjorth, H. Kim, Being there and being here: gendered customising of mobile 3G practices through a case study in Seoul, Convergence 11 (2) (2005) 49–55.

[17] S. Fragoso, WTF a crazy Brazilian invasion, in: F. Sudweeks, H. Hrachovec (Eds.), Proceedings of CATaC 2006, Murdoch University, Murdoch, Australia, 2006.

[18] R. Zheng, F. Provost, A. Ghose, Social network collaborative filtering, Working paper CeDER-8-08. Center for Digital Economy Research, Stern School of Business, New York University, 2008.

[19] R. Kumar, J. Novak, A. Tomkins, Structure and evolution of online social networks, in: KDD'06: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, 2006, pp. 611–617.

[20] C. Lampe, N.B. Ellison, C. Steinfeld, A familiar Face(book): profile elements as signals in an online social network, in: Proceedings of Conference on Human Factors in Computing Systems, ACM Press, New York, NY, 2007, pp. 435–444.

[21] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, A. Tomkins, Geographic routing in social networks, Proceedings of the National Academy of Sciences 102 (33) (2005) 11623–11628.

[22] S.A. Golder, D. Wilkinson, B.A. Huberman, Rhythms of social interaction: messaging within a massive online network, in: C. Steinfield, B.T. Pentland, M. Ackerman, N. Contractor (Eds.), Proceedings of Third International Conference on Communities and Technologies, Springer, London, UK, 2007, pp. 41–66.

[23] L. Backstrom, D. Huttenlocher, J. Kleinberg, X. Lan, Group formation in large social networks: membership, growth, and evolution, in: Proceedings of 12th International Conference on Knowledge Discovery in Data Mining, ACM Press, New York, NY, 2006, pp. 44–54.

[24] E. Spertus, M. Sahami, O. Büyükkökten, Evaluating similarity measures: a large-scale study in the Orkut social network, in: Proceedings of 11th International Conference on Knowledge Discovery in Data Mining, ACM Press, New York, NY, 2005, pp. 678–684.

[25] A. Chin, M. Chignell, Identifying active subgroups within online communities, in: Proceedings of the Centre for Advanced Studies, CASCON, Conference, Toronto, Canada, 2007.

[26] M. Brzozowski, T. Hogg, G. Szabo, Friends and foes: ideological social networking, in: Proceedings of the SIGCHI Conference on Human Factors in Computing, ACM Press, New York, NY, 2008.

[27] T. Hogg, D. Wilkinson, G. Szabo, M. Brzozowski, Multiple relationship types in online communities and social networks, in: Proceedings of the AAAI Spring Symposium on Social Information Processing, AAAI Press, 2008.

[28] K. Lewis, J. Kaufman, N.A. Christakis, The taste for privacy: an analysis of college student privacy settings in an online social network, Journal of Computer-Mediated Communication 14 (1) (2008) 79–100.

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

[30] M. Gjoka, M. Kurant, C.T. Butts, A. Markopoulou, Walking in Facebook: a case study of unbiased sampling of OSNs, in: Proceedings of IEEE INFOCOM'10, San Diego, CA, 2010.

[31] T.M.J. Fruchterman, E.M. Reingold, Graph drawing by force-directed placement, Software—Practice and Experience 21 (1991) 1129–1164.

[32] S. Wasserman, K. Faust, Social Network Analysis: Methods and Applications, Structural Analysis in the Social Sciences, Cambridge University Press, Cambridge, UK, 1994.

[33] M. McPherson, L. Smith-Lovin, J.M. Cook, Birds of a feather: homophily in social networks, Annual Review of Sociology 27 (2001) 415–444.

[34] M.E.J. Newman, Networks: An Introduction, Oxford University Press, Oxford, UK, 2010.

[35] M.S. Handcock, D.R. Hunter, C.T. Butts, S.M. Goodreau, M. Morris, ERGM: a package to fit, simulate and diagnose exponential-family models for networks, Journal of Statistical Software 24 (3) (2008) 1–29.

[36] G. Robins, P. Pattison, Y. Kalish, D. Lusher, An introduction to exponential random graph (p^*) models for social networks, Social Networks 29 (2007) 173–191.

[37] O. Frank, D. Strauss, Markov graphs, Journal of the American Statistical Association 81 (1986) 832–842.

[38] S. Wasserman, P. Pattison, Logit models and logistic regressions for social networks. I: an introduction to Markov graphs and p^*, Psychometrika 61 (1996) 401–425.

[39] M.J. Lubbers, T.A.B. Snijders, A comparison of various approaches to the exponential random graph model: a reanalysis of 102 student networks in school classes, Social Networks 29 (2007) 489–507.

[40] S.J. Cranmer, B.A. Desmarais, Inferential network analysis with exponential random graph models, Political Analysis 19 (1) (2011) 66–86.

[41] M.A. Porter, J.-P. Onnela, P.J. Mucha, Communities in networks, Notices of the American Mathematical Society 56 (9) (2009) 1082–1097, 1164–1166.

[42] S. Fortunato, Community detection in graphs, Physics Reports 486 (3–5) (2010) 75–174.

[43] M.E.J. Newman, Mixing patterns in networks, Physical Review E 67 (2) (2003) 026126.

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

[45] T. Callaghan, P.J. Mucha, M.A. Porter, Random walker ranking for NCAA division I-A football, The American Mathematical Monthly 114 (9) (2007) 761–777.

[46] M.A. Porter, P.J. Mucha, M.E.J. Newman, C.M. Warmbrand, A network analysis of committees in the United States House of Representatives, Proceedings of the National Academy of Sciences 102 (20) (2005) 7057–7062.

[47] Y. Zhang, A.J. Friend, A.L. Traud, M.A. Porter, J.H. Fowler, P.J. Mucha, Community structure in Congressional cosponsorship networks, Physica A 387 (2008) 1705–1712.

[48] A.S. Waugh, L. Pei, J.H. Fowler, P.J. Mucha, M.A. Porter, Party polarization in congress: a network science approach, 2009. arXiv:0907.3509.

[49] P.J. Mucha, T. Richardson, K. Macon, M.A. Porter, J.-P. Onnela, Community structure in time-dependent, multiscale, and multiplex networks, Science 328 (5980) (2010) 876–878.

[50] R. Guimerà, L.A.N. Amaral, Functional cartography of complex metabolic networks, Nature 433 (2005) 895–900.

[51] M.C. González, H.J. Herrmann, J. Kertész, T. Vicsek, Community structure and ethnic preferences in school friendship networks, Physica A 379 (2007) 307–316.

[52] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, A.-L. Barabási, Structure and tie strengths in mobile communication networks, Proceedings of the National Academy of Sciences 104 (18) (2007) 7332–7336.

[53] T. Kamada, S. Kawai, An algorithm for drawing general undirected graphs, Information Processing Letters 31 (1989) 7–15.

[54] A.L. Traud, C. Frost, P.J. Mucha, M.A. Porter, Visualization of communities in networks, Chaos 19 (4) (2009) 041104.

[55] M.E.J. Newman, Finding community structure in networks using the eigenvectors of matrices, Physical Review E 74 (2006) 036104.

[56] J. Reichardt, S. Bornholdt, Statistical mechanics of community detection, Physical Review E 74 (2006) 016110.

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

[58] B.W. Kernighan, S. Lin, An efficient heuristic procedure for partitioning graphs, The Bell System Technical Journal 49 (1970) 291–307.
[59] M.E.J. Newman, Modularity and community structure in networks, Proceedings of the National Academy of Sciences 103 (23) (2006) 8577–8582.
[60] T. Richardson, P.J. Mucha, M.A. Porter, Spectral tripartitioning of networks, Physical Review E 80 (3) (2009) 036111.
[61] V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, Journal of Statistical Mechanics 2008 (10) (2008) P10008.
[62] U. Brandes, D. Delling, M. Gaertler, R. Goerke, M. Hoefer, Z. Nikoloski, D. Wagner, On modularity clustering, IEEE Transactions on Knowledge and Data Engineering 20 (2) (2008) 172–188.
[63] B.H. Good, Y.-A. de Montjoye, A. Clauset, Performance of modularity maximization in practical contexts, Physical Review E 81 (4) (2010) 046106.
[64] W.M. Rand, Objective criteria for the evaluation of clustering methods, Journal of the American Statistical Association 66 (336) (1971) 846–850.
[65] L. Hubert, Nominal scale response agreement as a generalized correlation, British Journal of Mathematical and Statistical Psychology 30 (1977) 98–103.
[66] R.J. Brook, W.D. Stirling, Agreement between observers when the categories are not specified in advance, British Journal of Mathematical and Statistical Psychology 37 (1984) 271–282.
[67] E. Kulisnkaya, Large sample results for permutation tests of association, Communications in Statistics—Theory and Methods 23 (1994) 2939–2963.
[68] E.W. Weisstein, Barycentric coordinates, in: Wolfram Mathworld, 2011. Available at: http://mathworld.wolfram.com/BarycentricCoordinates.html.
[69] J.N. Franklin, Methods of Mathematical Economics: Linear and Nonlinear Programming, Fixed-Point Theorems, SIAM, Philadelphia, PA, 2002.
[70] A.H. Looijen, M.A. Porter, Legends of Caltech III: Techer in the Dark, Caltech Alumni Association, Pasadena, CA, 2007.