COMMUNITY STRUCTURE IN ONLINE COLLEGIATE SOCIAL NETWORKS

AMANDA L. TRAUD[1], ERIC D. KELSIC[2], PETER J. MUCHA[1,3], AND MASON A. PORTER[4]
[1]CAROLINA CENTER FOR INTERDISCIPLINARY APPLIED MATHEMATICS, DEPARTMENT OF MATHEMATICS, UNIVERSITY OF NORTH CAROLINA, CHAPEL HILL, NC 27599-3250, USA
[2]DEPARTMENT OF SYSTEMS BIOLOGY, HARVARD MEDICAL SCHOOL, HARVARD UNIVERSITY, BOSTON, MA 02115, USA
[3]INSTITUTE FOR ADVANCED MATERIALS, NANOSCIENCE & TECHNOLOGY, UNIVERSITY OF NORTH CAROLINA, CHAPEL HILL, NC 27599, USA
[4]OXFORD CENTRE FOR INDUSTRIAL AND APPLIED MATHEMATICS, MATHEMATICAL INSTITUTE, UNIVERSITY OF OXFORD, OX1 3LB, UK

Abstract. *We apply the tools of network analysis to study the roles of university organizations and affiliations in structuring the social networks of students by examining the graphs of Facebook "friendships" at five American universities at a single point in time. In particular, we investigate each single-institution network's community structure, which we obtain by partitioning the graphs using an eigenvector method. We employ both graphical and quantitative tools, including pair-counting methods that we interpret through statistical analysis and permutation tests, to measure the correlations between the network communities and a set of self-identified user characteristics (residence, class year, major, and high school). We additionally investigate single-gender subsets of the university networks and also examine the impact of incomplete demographic information in the data. Our study across five universities allows one to make comparative observations about the online social lives at the different institutions, which can in turn be used to infer differences in offline lives. It also illustrates how to examine different instances of social networks constructed in similar environments, while emphasizing the array of social forces that combine to form simplified "communities" obtainable by the consideration of the friendship links. In an appendix, we review the basic properties and statistics of the employed pair-counting similarity coefficients and recall, in simplified notation, a useful analytical formula for the z-score of the Rand coefficient.*

1. Introduction. Social networks are a pervasive part of everyday life. Although they have long been studied by social scientists [101], the mainstream awareness of their ubiquity has arisen only recently, in part because of the rise of social networking sites (SNSs) on the World Wide Web. Since their introduction, SNSs such as Friendster, MySpace, Facebook, Orkut, LinkedIn, and hundreds of others have attracted 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 [9]. As recent work has demonstrated, the implications and importance of online social networks are diverse and signicant [60], in part because of their role as a reflection of offline society [9,30,34,35,43,79]. Meanwhile, the scientic study of real-world networks, including the study of SNSs, has expanded in recent years, and there is strong optimism that formal statistical and graph-theoretic analysis can not only help achieve a better understanding of the structure and dynamics of online social networks but also inform useful additions, modications, and uses of such services.

The plethora of available activities in SNSs underscores the fact that online social networks include multiple types of interacting informational and social structures:

Who is "friends" with whom? Who has e-mailed whom? Who are members of the same organizations or other groups? Who is attending which events? The interaction between social and other types of connections is of utmost importance, with people benefitting from their connections and role in the social network every day (see, e.g., [14, 15]) and advertisers obviously keen to utilize the information inherent in such networks. Accordingly, one of the goals of the study of online social networks is to better understand their structures and dynamics in order to be able to exploit them in a systematic manner. One of the especially exciting aspects of massive SNSs is that they not only give researchers the chance to test theories of regularity in social behavior at unprecedented levels of extent and detail, but that they also offer the opportunity to develop applications that use the knowledge of numerous people from one's social network to provide important services (such as effective recommendation systems [105]).

1.1. Network Science. Networks (graphs) provide a powerful representation for analyzing complex systems of interacting agents, such as the users of SNSs. Accordingly, the study of networks has now become pervasive in biology, information science, sociology, and many other disciplines [70, 95]. The simplest type of network is an unweighted, undirected, unipartite graph, which consists of a collection of nodes (representing agents) connected by edges (representing ties/connections). Important generalizations include the consideration of different types of edges—which can, e.g., be weighted, directed, or signed (with agents disposed towards or against other agents)—and the study of bipartite networks, in which one type of node (e.g., an individual) must be connected to a second type of node (such as an attended event).

The scientific study of networks has its origin in the social sciences, with edges typically defining a specific relation between two individuals or organizations, which can be embedded in numerous types of social networks whose structure plays an important role in explaining their behavior [70, 101, 103]. Such relations can be characterized by the existence of friendship, support, kinship, contact, communication, presence at a common event, or membership in a common organization. Studies of interorganizational networks have yielded insights into how alliances and other ties are formed, how they affect organizational performance, and how various organizational practices spread in such networks [14, 15, 22, 42, 87, 99]. Although the study of networks has a long history, their study intensified in the late 1990s because of interest in the Internet and an increase in readily-available, large-scale data. This motivated the further development of tools for studying social, biological, and technological networks [3, 24, 70, 95, 103]. Such research has generated numerous important insights on the effects of network topology on individuals' behavior, including collaborations [41, 69], community formation [21, 28], and hierarchical and modular organization [88, 92].

1.2. Social Networking Site Research. Facebook, an SNS launched in February 2004, has overwhelmed numerous aspects of everyday life [8, 9], becoming an especially popular obsession among college and high school students (and, increasingly, among others members of society). 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 one wants to add as a friend confirms the relationship, so the friendship network is undirected. Recent sociological research has shown that most people typically draw their Facebook friends from their real-life social networks [9], implying that a Facebook network can be used as an approximate proxy for an offline social network. Accordingly, important features of SNSs include

the ability of users to articulate and make visible their social networks, and to interact with their networks in automatic ways, such as through "news feed" broadcast mechanisms that would be difficult to replicate offline.

The emergence of SNSs such as Facebook and MySpace has revolutionized the availability of social and demographic data, which has in turn impacted the study of social networks [9, 52, 60]. Traditionally, social network data has been gathered using surveys, which typically limits the sizes of the graphs one could consider, biases the types of people captured in the network, and introduces numerous sources of data error. Now, however, one can easily acquire very large and accurate data sets from SNSs, though of course the population online and actively using SNSs remains a biased sample of the broader population (e.g., individuals have different propensities to interact online). Services like Facebook also allow one to obtain better demographic data, as many users now give out voluminous amounts of personal detail voluntarily. This newfound wealth of available information also raises questions about balancing the desire to proclaim identify with public disclosure of that information [96–98], and one might also reasonably wonder if such exhibitionism has any measurable effects on network structure or collective behavior.

Social scientists, information scientists, and physical scientists have all been quick to jump on the data bandwagon that has resulted from this demographic revolution [91]. It would be impossible for us to exhaustively cite all of the germane research, so we only highlight a few results here; additional references can be found in the review by Boyd and Ellison on SNS history and research [9]. Boyd also wrote a popular essay about her empirical study of Facebook and MySpace, concluding that Facebook tends to appeal to a more elite and educated cross-section than does MySpace [7]. Very recently, the company RapLeaf has compiled global demographics on the age and gender usage of numerous SNSs (including Facebook) [93]. Other recent studies have investigated the manifestation on SNSs of race and ethnicity [34], religion [79], gender [35, 43], and national identity [30]. Preliminary research has also suggested that online friendship networks can be exploited to improve shopper recommendation systems on websites such as Amazon [109].

A number of papers have attempted to better understand how SNS friendships form. For example, Kumar *et al.* [54] examined preferential attachment models of SNS growth, concluding that it is important to consider different classes of users (including passive members, inviters, and linkers). Lampe *et al.* [55] explored the relationship between profile elements and number of Facebook friends, and other scholars have examined the importance of geography [59] and online message activity [37] to online friendship formation. Several other papers have established strong correlations between network participation and website activity, including the motivation of people to join particular groups [5], the recommendations of online groups [94], online messages and friendship formation [37], interaction activity versus sense of belonging [18], and the role of explicit ideological relationship designations (on Essembly) in affecting voting behavior [13, 44]. An especially intriguing recent paper uses Facebook data for an entire class of freshmen at an unnamed, private American university to provide a quantitative study of social networks and cultural preferences [58]. The same data set has also been used to examine user privacy settings on Facebook [57].

1.3. Community Structure in Networks. The global organization of real-world networks typically includes coexisting modular (horizontal) and hierarchical (vertical) organizational structures [21, 28]. Myriad papers in the recent network science literature have attempted to interpret such organization through the compu-

tation of structural modules or communities, defined in terms of mesoscopic groups of nodes with more internal connections (between nodes in the group) than external connections (between nodes in the group and nodes in other groups) [21, 28]. Such communities, which are not typically identified in advance, are often considered to be not merely structural modules but are also expected to have functional importance in network dynamics. For example, communities in social networks ("cohesive groups" in the sociology literature [66, 67]) might correspond to circles of friends or business associates, communities in the World Wide Web might encompass pages on closely-related topics, and some communities in biological networks have been shown to be related to functional modules [40].

With such motivation, the identification and investigation of community structure has become its own cottage industry in network science since the 2002 paper of Girvan and Newman [36] made seminal contributions and helped turn questions about community structure into a playground for statistical physicists and mathematicians. The number of methods since published to detect communities in various types of networks is now both enormous and continuously expanding [28]. The classes of available techniques broadly include hierarchical clustering methods such as single linkage clustering [49], betweenness-based methods [36,72], local methods [6,19,56], maximization of quality functions such as modularity and similar quantities [71, 74–76, 89], spectral partitioning [73], likelihood-based methods [20], and more. In addition to remarkable successes on benchmark examples, such as the infamous Zachary Karate Club [106], community structure investigations have led to success stories in diverse application areas—including the reconstruction of college football conferences [36] and the investigation of such structures in algorithmic rankings [16]; the analysis of committee assignments [82–84], legislation cosponsorship [108], and voting blocs [104] in the U.S. Congress; the examination of motifs and other functional groups in genetic [65] and metabolic [40] networks; and the study of ethnic preferences in school friendship networks [38] and social structures in mobile-phone conversation networks [80].

1.4. Overview of the Facebook Data and the Present Analysis. In this paper, we investigate the community structure of single-institution Facebook networks representing the full set of user pages (nodes) from each of five American universities and all of the links between those pages, representing reciprocated "friendship," as they existed in June 2005. We specifically consider only ties between students at the same institution, yielding five separate realizations of university social networks and allowing us to comparatively examine the structures at different institutions. Our study includes a small technical institute, a pair of private universities, and a pair of large state universities. The data includes limited demographic information provided by users on their individual pages—including gender and data fields that represent (by anonymous numerical identifiers) high school, class year, major, and dormitory residence/"House" (depending on the specific housing mechanism of the institution). This allows us to make interesting comparisons between different universities, under the assumption (per the discussion in [9]) that the communities and other elements of structural organization in Facebook networks are a reflection of the social communities and organization of the offline networks on which they're based. By examining these different social networks constructed in similar environments, we can also explore the varied and complicated dependencies of community formation on the individual attributes and local cultural details. This complicated, real-world situation contrasts with graphs in which community formation depends completely on structural properties, such as link topology and weights, as often needs to be assumed [21, 28].

Starting from the Facebook data, we focus our attention in Section 2 on the notion of community detection using a simple, well-studied benchmark example (the Zachary Karate Club) to illustrate these ideas, briefly discuss the wealth of methods available in the literature, and present the eigenvector algorithm developed by Newman that we utilize in the present paper to obtain the community structures in the networks of each university. We then proceed to compare these algorithmically-identified communities with the demographic data of the users in each community, using graphical tools in Section 3 to investigate the correlation between the communities and demographic information at Caltech (the smallest of the five institutions we study) and the University of North Carolina at Chapel Hill. In Section 4, we use the Karate Club benchmark example to discuss various quantitative means to compare network partitions, focusing on the use of z-scores of pair-counting indices. Because these z-scores depend only very weakly on the specific pair-counting coefficient employed and are indeed identical for many of the standard pair-counting indices (as we show in the Appendix), they provide a simpler interpretation of the strengths of the observed correlations than the raw pair-counting values, allowing correlations to be compared more readily. We subsequently apply these methods in Section 5 to compare the algorithmically-identified communities with the user characteristics in the Facebook networks. For each of our five university data sets, we consider the full network and both the male and female subnetworks. In Section 6, we study the role of unreported demographic data on our results using two different protocols for addressing this missing information. We conclude in Section 7 with a discussion of how our findings might inform us about the social networks of these universities. We also consider the present paper in the context of research on network community structure. In the Appendix, we review essential properties of the employed pair-counting similarity coefficients, identify their common statistical elements, and recall (in simplified notation) a useful analytical formula for the z-score of the Rand coefficient. The primary contributions of the present work thereby include the novel investigation of the correlations between the community structures and personal characteristics for such collegiate SNSs, the quantitative measurement of such correlations in terms of clearly elaborated statistical properties, and the conclusions obtained regarding the different strengths of community and organizational correlations at different universities.

2. Preliminaries: Community Detection in Networks. A social network with a single type of connection between nodes can be represented as an adjacency matrix A with elements A_{ij} that give the weight of the tie between nodes i and j. The Facebook networks we study are unweighted, with A_{ij} taking the values 1 and 0, indicating the presence and absence of a connection, respectively. The resulting tangle of links, which we show for the Caltech Facebook network in Fig. 2.1, often obfuscates the presence of organizational structure in the network. Our goal in this section is to discuss how to identify groups of friends in the form of structural network "communities"—groups of nodes with more internal connections between nodes in the group than external connections between nodes of the group and nodes in other groups [21, 28]—so that we can later compare the composition of the identified communities to groups formed from common user characteristics. Because of its relatively small size, we will use Caltech as our first illustrative example in Section 3 to help motivate the general problem of how to compare algorithmically-obtained network communities.

One of the earliest ideas for clustering nodes in a graph to find communities and hierarchies is known as single linkage clustering [49]. Another classical approach can

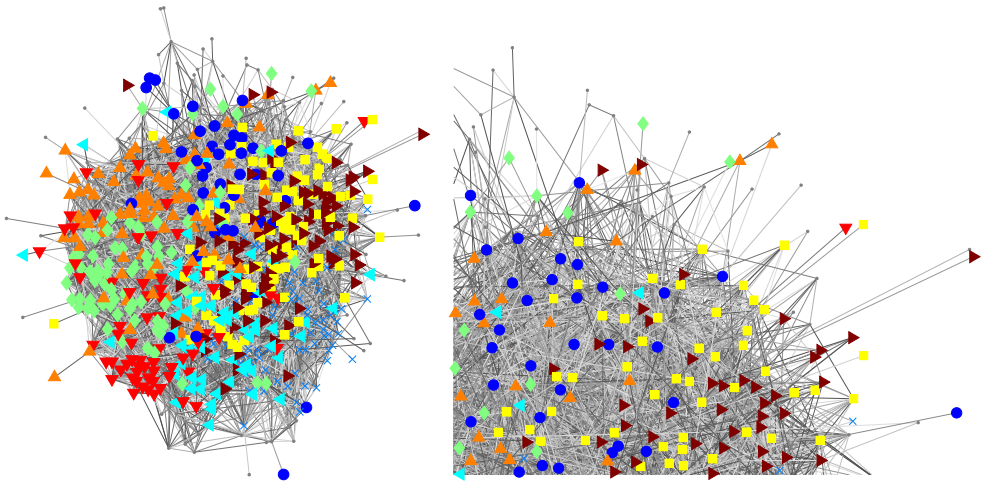

FIG. 2.1. *[Color] (Left) A Fruchterman-Reingold visualization [33] of the largest connected component of the Caltech Facebook network. Node shapes and colors indicate House affiliation, with gray dots denoting individuals who did not identify a House affiliation. The edges are colored using random shades of gray for visualization purposes. (Right) Magnification of a portion of the network. Observe the clusters of nodes with the same color/shape, suggesting that House affiliation has a significant effect on the existence of friendships (edges) in this graph.*

be found in the graph partitioning literature in computer science [26,85], although the traditional methods there typically require one to specify the sizes of communities in advance, and are thus problematic for studying social networks [73]. More recently, community detection has become one of the most prominent ideas investigated in the network science community, where statistical physicists, computer scientists, and applied mathematicians have all made fundamental contributions.

The earliest algorithms employed by statistical physicists used the ideas of betweenness to iteratively pick out and remove high-traffic edges (or other network components) that lie on a large number of paths between vertices. The repeated application of such a procedure eventually fragments a network into components [36,70,75,76]. Such methods are thus examples of 'divisive' algorithms because they start from the full network and divide it into smaller subnetworks. On the other hand, 'agglomerative' algorithms such as single linkage clustering start with individual nodes and form communities by joining them. Either way, the order in which an algorithm either joins or splits groups can be subsequently visualized using a dendrogram (tree), though the specific details of such dendrograms can sometimes indicate more about the details of the employed algorithm than those of the underlying network structure [28].

One of the presently dominant approaches in community detection involves the optimization of a quality function known as *modularity*, which counts the total edge weight of all intra-community connections compared to that which might be expected at random (under some null model) [27, 71, 73, 74, 76], though there are also numerous other popular approaches to determining network community structure [6, 19, 20, 28, 56, 81, 89]. In the context of the big-picture goal of finding functional or cohesive groups in networks, all of these techniques implicitly make strong assumptions about the relation between structural and functional properties of the network. That is, given the available data and methods, many researchers (ourselves included) are typically forced to limit attention to the available information about the links in

deriving structures, with little assurance that the behaviors of and on the network are best represented by this limited information. Moreover, one often assumes that there is a single "best" clustering or level of organization, and that such a preferred clustering is actually meaningful. The questionable nature of this last point was discussed recently in [2]. The comparative investigation of communities (after they have been obtained algorithmically) that we conduct in the present paper provides some insight into the other assumptions. Moving beyond these assumptions is a very difficult problem that deserves serious attention.

A high modularity value indicates a strong or significant structural split in the network and has been found to be a good indicator of functional network divisions in many cases [73, 74], though again we stress that structural modules might provide only an approximation of the associated functional modules. Modularity, computed for selected partitions of the network, thereby measures the extent to which the identified interactions between nodes take place within the identified community partitions rather than across them. An appealing feature (shown very recently) of detecting communities by maximizing modularity is that determining a network's community structure in this way is equivalent to visualizing it using particular parameter values in a force-directed layout [78].

While identifying the partitioning of a network into communities that maximize modularity is known to be NP-complete [10], approximate optima can be found using eigenvector methods [73, 74]. These methods have the additional benefit of being computationally efficient for large, sparse networks such as those we study in our Facebook examples (which can have up to several tens of thousands of nodes in a single-institution network), so that relatively rapid computations yield network partitions with high modularities. We have accordingly chosen to utilize Newman's leading-eigenvector method in the present work; our investigation can be repeated for any desired community-detection method. In Section 5.3, we briefly examine the effects of algorithm choice by doing an example comparison using the results of the eigenvector method both with and without Kernighan-Lin-Newman (KLN) iterations. We have also checked some of the qualitative results using recursive bisection with leading pairs of eigenvectors (also described in [73]), though we do not discuss these computations here.

To provide context for our investigation, we find it useful to summarize the main idea in the original formulation of the eigenvector method for community detection, following the presentation of [73], where this spectral method was detailed. Specifically, we consider the method based on taking the eigenvector corresponding to the largest positive eigenvalue (the so-called "leading eigenvector") of the modularity matrix B, whose components are given in terms of the adjacency matrix A by [73, 74]

$$B_{ij} = A_{ij} - \frac{k_i k_j}{2m}\,, \tag{2.1}$$

where $k_i = \sum_j A_{ij}$ is the degree of the ith node and $m = \frac{1}{2}\sum_i k_i$ is the number (or total weight) of edges in the network. Subtracting the fraction of expected edges in (2.1) corresponds to a specific choice of null model, suggested by Newman and Girvan [73, 75], with average degree equal to that of the corresponding configuration model [77] (a random graph with an arbitrary, specified degree distribution). More generally, the choice of null model specifies the baseline for counting the number of intra-community edges beyond that expected "at random." The components of the leading eigenvector $\mathbf{v}$ of B are used to bisect the network according to the sign of

its components. Subsequent bisections are obtained recursively, keeping track of the fact that each subnetwork one considers is actually part of a larger network, until the modularity can no longer be maximized with further subdivisions [73, 74].

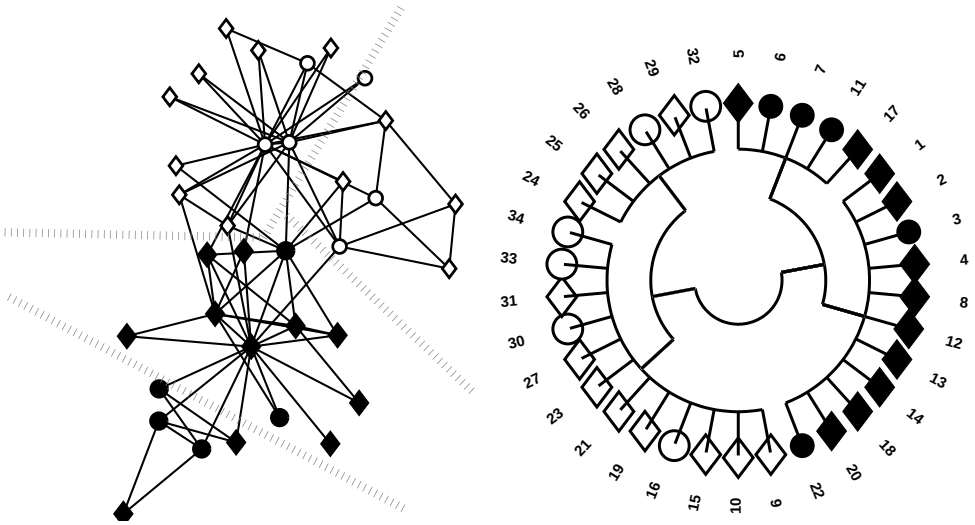

FIG. 2.2. *(Left) A visualization of the Zachary Karate Club network [106], using a Kamada-Kawai spring-embedding algorithm [50], in which nodes are solid or open depending on their later club affiliation (after a disagreement prompted the breakup of the organization) and have been randomly assigned a shape (circle or diamond). The dashed lines indicate the boundaries between the communities determined using the leading-eigenvector method with the KLN iterations of [73]. We use the random assignment of nodes in Section 4 to discuss the quantitative comparison of partitions. (Right) Dendrogram of the Zachary Karate Club network to group nodes into the same communities as in the left panel (with the shadings and shapes of the nodes also the same). Note that the initial bisection into two groups is identical to the observed split of the club.*

Perhaps the most famous benchmark example used to illustrate community-detection algorithms is the Zachary Karate Club, in which an internal dispute led to the schism of a karate club into two smaller clubs [106]. Figure 2.2 (Left) depicts the connections between members in the original club. The Karate Club network provides an instructive example for community-finding algorithms because we expect any calculated communities to be very similar to the memberships of the two post-dispute clubs, indicated by the open and closed symbols in the figure. In Fig. 2.2 (Right), we show the communities that result from applying the leading-eigenvector approach. The initial bisection into two branches identified by the algorithm is observed starting from the center of the ring in Fig. 2.2 and moving outward (i.e., we show a dendrogram drawn in polar coordinates). The success of the algorithm is apparent, as the initial bisection reflects the actual membership of the new clubs. This particular algorithm subsequently splits the network into 4 communities, as indicated near the outside of the ring in Fig. 2.2, the identification of which has been improved using KLN iterations, described and recommended in [73, 74]. The partition indicated in the figure has a slightly higher modularity ($Q = 0.4198$) than that obtained using the leading-eigenvector method in isolation (with $Q = 0.3934$), the latter differing from the former (in the figure) by assignment of nodes #1 and #12 to the community that includes nodes #5–7. In particular, both partitions give the same initial bisection into the two post-dispute clubs, though such correspondence need not occur in general.

3. Comparing Communities Visually. We now demonstrate some of the advantages and limitations of visually comparing communities to demographic information, focusing on two of our five Facebook networks: Caltech and UNC. As the smallest network in our data set, and which two of the present authors attended as students, Caltech provides an illustrative example that we also know very well from personal experience. The undergraduate "House" system at Caltech, appearing in lieu of dormitory residence in our data, is modeled after the Oxbridge college system. Caltech's Housing system impacts student life enormously, both socially and academically [61], and is even used by the university as one of its primary selling points in attracting new undergraduates. At the beginning of their first year at Caltech, undergraduate students choose one of the eight Houses and usually remain a member of it throughout their collegiate career. At the time of our data set (June 2005), students could only select seven of these Houses: Blacker, Dabney, Fleming, Lloyd, Page, Ricketts, and Ruddock. The residents of the eighth House (Avery) included graduate students, faculty, postdocs, and undergraduate students affiliated with each of the other seven Houses.

Accordingly, we have assigned to each node in the visualization of the Caltech network in Fig. 2.1 a shape (and color) that designates undergraduate "House," with small dots denoting individuals who did not identify a House affiliation. The full Caltech Facebook network at the time of our data included 1099 users; the largest connected component included 762 nodes and 16651 edges, giving a mean degree of 21.66. Most of the other users were singletons without any specified links. We show these statistics for all five of the Facebook networks we study in Table 5.1.

From visual inspection of Fig. 2.1, it should not be surprising that Caltech's community structure is strongly correlated with the House structure. We illustrate this in Fig. 3.1 using a pie-chart dendrogram of Caltech, grouping nodes into the communities at the maximum modularity found using the leading-eigenvector method. (One can make a similar plot with other divisive community-detection methods.) To obtain this partition, with 7 communities (pies) and modularity $Q = 0.3594$, we recursively bisected the Caltech network down to the highest modularity that the algorithm could obtain (i.e., when subsequent bisections reduced the modularity). The area of each pie in the figure is proportional to the number of nodes in the community it represents, and each color-coded pie slice (with size proportional to the number of nodes) indicates the individuals of one House affiliation. White slices signify individuals who did not identify any House affiliation. The central portion of Fig. 3.1 indicates the divisive bisections made by the leading-eigenvector algorithm, and the order of these splits is represented by the radius of their associated arcs (moving outward from the center).

Unlike other universities (see the discussion in Section 5), we find that House affiliation is the Caltech network's primary organizing principle, which is what we expected. This provides a reality check for the community-detection methods we employ, as Caltech's House structure is so dominant socially that the partition produced by any reasonable community-detection method should exhibit a strong correlation with House affiliation. Indeed, each pie in Fig. 3.1 is dominated by members of one House. Blacker, Fleming, Lloyd, Page, Ricketts, and Ruddock each dominate one of the larger communites. There is only one House that the obtained community structure does not similary respect: Dabney House does not have its own community (aside from a very tiny one) but rather has a significant presence in the communities dominated by Ricketts and (to a lesser extent) Lloyd and Blacker. Dabney, Rick-

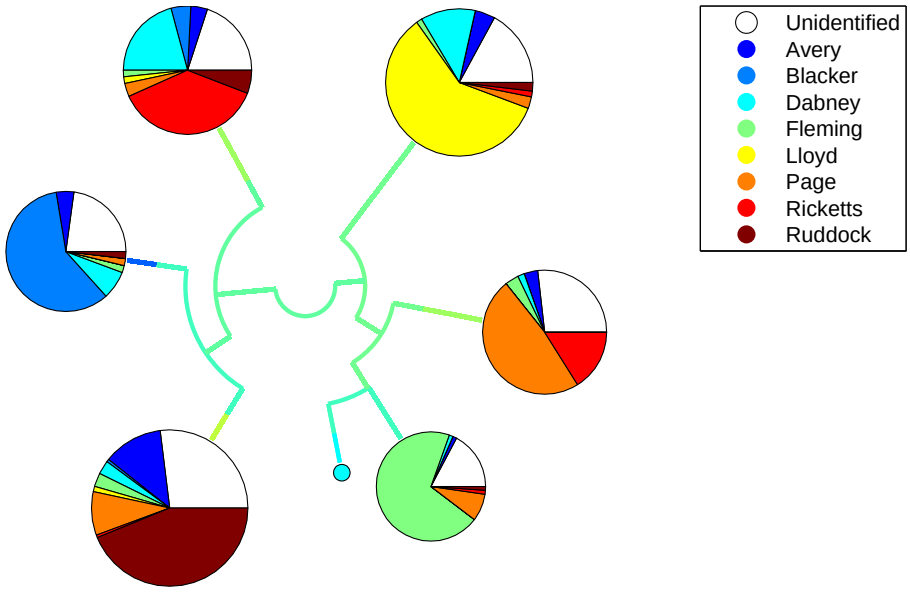

FIG. 3.1. *[Color] Pie-chart dendrogram of the Caltech network, colored according to House affiliation (the analog of dormitory residence). To produce this plot, we recursively bisected the Caltech Facebook network using the leading-eigenvector method until modularity could not be increased by subsequent divisions. Each of the communities in the final partition is represented by a pie whose area is proportional to the number of nodes it encompasses. The central (small radius) portion of the dendrogram shows the partitions that have been made, with smaller-radius partitions occurring at earlier stages in the algorithm. White portions correspond to people who didn't identify a House affiliation. As one can see in this figure—and unlike the other universities (see the discussion in the text)—the Caltech friendship network is organized predominantly by House.*

etts, Fleming, and Blacker are geographically proximate, constituting Caltech's four "South Houses," with Dabney, Ricketts, and Blacker known to be especially closely associated socially with each other. As expected based on its different residency rules, almost all of the pies include a number of people who identify Avery as their affiliation. Members of all of the Houses live in Avery, so the wide dispersal of Avery nodes in most of the network's main communities—rather than its domination of its own pie—was also expected.

To give more examples, consider the three "North Houses" (Lloyd, Page, and Ruddock).[1] There are a significant number of Page residents in the Ruddock community as well as in the community that contains almost every student affiliated with Fleming. Page and Fleming have both long been known at Caltech for being particularly popular House choices among students interested in athletics, so we conjecture that many of these connections have arisen through this particular common interest. (The fact that Fleming seems to be more closely associated with Page House than its fellow South Houses is accordingly not surprising, as Fleming has long been culturally rather different than the other South Houses.) One very interesting observation is the geographical isolation that seems to exist in the Caltech communities even though its campus is extremely small. In fact, this isolation is a known feature to Caltech

[1] Avery, located at the north edge of campus, is geographically isolated from all the other Houses.

students and alumni, who frequently discuss the apparent social divisions between individual Houses, between the North and South Houses, and between Avery and all of the other Houses (see, e.g., [1]). It is natural to wonder if community detection on current data would find a community dominated by Avery, since its promotion to official House status. Examining the formation of such a community using longitudinal data would be even more interesting, but is beyond the scope of the present study.

In principle, one can also make limited predictions about the possible House affiliations of the white nodes (who did not identify any such affiliations) based on the composition of the communities in which they are placed. At minimum, one can conclude that there is a good chance that they would be interested in, e.g., alumni events organized around the dominant House of that community whether or not they were officially affiliated with that House. As we discuss later, the organization of the Caltech Facebook network by House differs completely from that in other universities. Accordingly, the present study is useful for making comparative observations between different universities, which constitute individual "instances" of social networks constructed in different environments, under the assumption that the communities and structural organization in Facebook networks is a reflection of that in their corresponding real-life networks. Our outside knowledge of Caltech allows us to provide strong qualitative support for the validity of our computational analysis, providing an important reality check for the methods. Meanwhile, the observed heterogeneity in the communities, even at an institution like Caltech whose social structure seems to be mostly dominated by a single feature (House affiliation), underscores the important point that networks typically have multiple levels of organization rather than a single best one, as has also been discussed recently (in more abstract contexts) [4, 20, 56, 89].

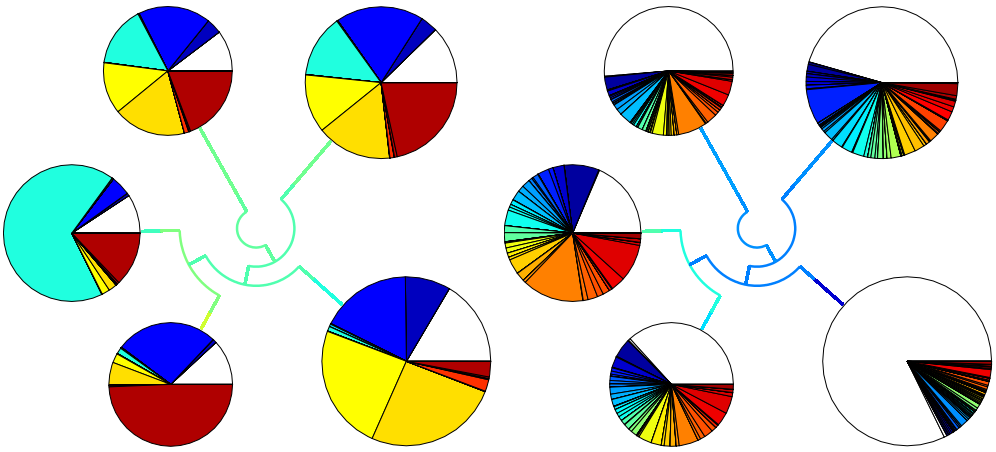

FIG. 3.2. *[Color] Pie-chart dendrograms of UNC, colored by (Left) class year and (Right) residence. White slices correspond to individuals who didn't self-identify the relevant characteristic.*

Despite the successes above, it is necessary to go beyond visual comparisons, as is illustrated by consideration of the Facebook data for UNC (the present affiliation of two of this paper's authors). In Fig. 3.2, we show two different pie-chart dendrograms of the identified communities in this network; one is colored by class year, and the other is colored by dormitory residence. In contrast to the Caltech network, it is much harder to make definitive conclusions from such visual information. To investigate the social organization of UNC and most other universities, it is essential to quantitatively

compare the detected communities with the available demographic groups.

4. Comparing Communities Quantitatively. As discussed in Section 3, a visual comparison of different network partitions can in some cases indicate significant correlations. In other cases, however, it can be extremely difficult to make definitive conclusions with visual tools alone. It is thus desirable to also employ quantitative comparisons, for which a variety of approaches can be used. For instance, there is a common desire to compare two different sets of clusters of a network [64]—whether the comparison is between groups obtained from different community-detection algorithms, between a specified clustering method and a known "correct" split [21], or between a set of communities and a second obtained by detected communities in some perturbation of the network's adjacency matrix [51]. In the present application, we desire to separately compare the algorithmically-identified communities to groupings of the nodes according to each of the self-identified characteristics, in order to determine which demographic characteristics best correlate with the network community structures of each university.

As explored by Meila [64] and concisely reviewed in [51], the many methods available in the partition-comparison literature can be classified roughly into three groups: (1) pair counting, (2) cluster matching, and (3) information-theoretic methods. The last category includes "variation of information" [64] and "normalized mutual information" [21]. One can also employ statistical analysis that uses exponential random graph models (ERGMs) [31, 62, 102], although this requires one to use an underlying model of how network links arise in the first place. We choose to focus in the present paper on a collection of pair-counting methods, in part because of their relative simplicity. That same simplicity has both strengths and weaknesses: On one hand, it has led to a plethora of specialized definitions; on the other, most pair-counting measures suffer from a serious interpretation difficulty because of the unclear range of good scores for a given setting. However, in the present study we make a powerful, unifying observation that standardized rescaling of a variety of these pair-counting scores yields quantitatively similar z-score values, providing a clearer intuition and alleviating the need to worry about the specificity of a single pair-counting measure. In the Appendix, we discuss (1) the algebraic form of the transformation to z-scores for some of the pair-counting scores and (2) details of the statistical similarities to other pair-counting indices, whose z-scores can also be obtained by permutation tests if necessary [39]. This standardization of the pair-counting coefficients provides a very useful clarifying interpretation that we will utilize in later sections.

4.1. Pair-Counting Methods. The essential idea in using pair-counting methods to compare two network partitions is to define a similarity coefficient by examining the placements of all possible node pairs in the two partitions. Each pair drawn from the n nodes of a network can be classified according to whether they fall in same or different groups in the partitions. Specifically, we denote the counts of the number of node pairs in each classification as w_{11} (in the same groups in both partitions), w_{10} (same in the first but different in the second), w_{01} (different in the first, but the same in the second), and w_{00} (different in both). The sum of these quantities is, by definition, equal to the total number M of node pairs, so that $M = w_{11} + w_{10} + w_{01} + w_{00} = \binom{n}{2} = n(n-1)/2$.

Given two partitions of a network (e.g., one given by community detection and another by grouping users according to a specified characteristic), one can obtain many different pair-counting similarity coefficients using different algebraic combinations of the $w_{\alpha\beta}$ counts. Throughout this paper, we use the notation S_i to refer generally to

such coefficients, where a specific choice of i indicates a particular choice. The pair-counting coefficients we employ are Rand ($i = \mathrm{R}$), Jaccard ($i = \mathrm{J}$), Fowlkes-Mallow ($i = \mathrm{FM}$), Minkowski ($i = \mathrm{M}$), Γ ($i = \Gamma$), and Adjusted Rand ($i = \mathrm{AR}$). See [17,48,64] for a wider range of measures, further discussion, and additional references. We remark that pair-counting methods comprise only a subset of a more general class of association measures that can used for studying *contingency tables*. In this context, the element n_{ij} of such a table indicates the number of elements (nodes) in the ith group of the first partition and the jth group of the second one [47, 53, 64]. See the Appendix for an additional discussion.

We focus first on the Rand similarity coefficient, $S_{\mathrm{R}} = (w_{11} + w_{00})/M$, one of the earliest and simplest pair-counting methods [86]. The Rand index counts the fraction of node pairs identified the same way by both partitions—either together in communities in both or placed in different communities in both. It is bounded between 0 (when no pair placements are the same) and 1 (identical partitions). While the Rand coefficient is extremely intuitive, and can be used fruitfully in many settings, even a brief consideration can reveal its deficiencies. In particular, the Rand coefficient for comparing two network partitions that each contain large numbers of communities tends to be skewed closer to the value 1 simply because of the large fraction of node pairs that are placed in different communities (i.e., w_{00} is a large fraction of M), even when comparing two partitions with little in common. Given this behavior, the notion of a "good" value of the Rand index for identifying similar partitions clearly depends on other details.

A simple proposal for trying to fix this problem with S_{R} is to remove the explicit role of w_{00}, such as in the Jaccard index, $S_{\mathrm{J}} = w_{11}/(w_{11}+w_{10}+w_{01})$, or the Fowlkes-Mallow similarity coefficient, $S_{\mathrm{FM}} = w_{11}/\sqrt{(w_{11}+w_{10})(w_{11}+w_{01})}$. As with the Rand index, these coefficients are necessarily bounded between 0 and 1, and the latter value is again obtained when comparing identical partitions. While both S_{J} and S_{FM} clearly avoid the problematic effects of large w_{00}, their complete ignorance of node pairs classified similarly into different communities skews the comparison unfairly in the opposite direction. This yields unnaturally high values of S_{J} and S_{FM} when comparing network partitions with very few communities (or when one partition consists of a single community).

We also use the Minkowski and Γ similarity coefficients. The Minkowski coefficient, given by $S_{\mathrm{M}} = \sqrt{(w_{10} + w_{01})/(w_{10} + w_{11})}$, is notably asymmetric in its consideration of the two partitions. The first of the two partitions serves as a distinguished reference, and the Minkowski coefficient provides a measurement of the number of mismatches relative to the number of similarly-grouped pairs in that reference (though we will see in the Appendix that it is again symmetric after standardization). Hence, S_{M} values closer to 0 are considered better. The Γ similarity coefficient, defined as

$$S_{\Gamma} = \frac{M w_{11} - (w_{11} + w_{10})(w_{11} + w_{01})}{\sqrt{(w_{11} + w_{10})(w_{11} + w_{01})(M - (w_{11} + w_{10}))(M - (w_{11} + w_{01}))}} \,,$$

has the most complicated algebraic form of the similarity coefficients we employ. We again refer the reader to [17, 48, 64] for further details and additional measures.

While each of the aforementioned similarity coefficients—Rand, Jaccard, Fowlkes-Mallow, Minkowski, and Γ—has obvious apparent strengths, most notably their definitions in relatively simple algebraic forms, no one of them is clearly better than the others. Moreover, each of these coefficients suffers from the problem that one does not know *a priori* what values constitute "good" ones. To illustrate this point, we

	$S_{\rm FM}$	S_Γ	$S_{\rm J}$	$S_{\rm M}$	$S_{\rm R}$	$S_{\rm AR}$
"Observed"	0.7313	0.6092	0.5348	0.9327	0.7736	0.5414
"Random"	0.3867	0.0150	0.2204	1.4094	0.4831	0.0126

TABLE 4.1

Similarity coefficients (Fowlkes-Mallow, Γ, Jaccard, Minkowski, Rand, and Adjusted Rand) for comparing the 4-community partition of the Zachary Karate Club identified algorithmically versus the "observed" split of the club into 2 new clubs (indicated by the open/closed symbols in Fig. 2.2) and a "random" split into 2 groups (indicated by the node shapes in the same figure).

consider their application to the (ubiquitous) Zachary Karate Club network depicted in Fig. 2.2. In Table 4.1, we collect these similarity coefficient values for comparing the maximum-modularity partition ("maximum" insofar as identified in Section 2) into 4 communities with the observed split of the club into 2 groups (indicated by open/closed node symbols in Figure 2.2). Recall that the observed split into 2 clubs is the same as that given by combining pairs of the algorithmically-identified communities. Beyond the trivial recognition that the partitions into 2 and 4 groups are different, the similarity values themselves are not immediately enlightening. Accordingly, we also calculate the similarity coefficients (also shown in Table 4.1) that compare the same 4-community partition with a randomly-generated partition into 2 groups (indicated by node shapes in the figure) that clearly disagree with the community structure. Recalling that larger values indicate closer agreement (except for $S_{\rm M}$, for which values closer to 0 are better), the good correlation between the algorithmically-identified communities and the observed split becomes more evident by the comparison.

Despite this seemingly successful application using the Zachary Karate Club, additional pitfalls remain. In particular, as we observe with the Facebook networks (see the discussion below) and explore further in the Appendix, the various S_i values depend intimately on the size of the network and the numbers and sizes of the groups in each partition. Therefore, it is not always clear whether the ordering of the coefficient values allows one to ascertain the relative correlations between different partitions. For example, given two partitions of one network that have, say, a Rand value of 0.6 between them and two partitions of a second network that have a Rand value of 0.8 between them, it is not at all guaranteed that the second pair should necessarily be classified as closer partitions (to each other) than the first pair. Consequently, the general problem of knowing what values indicate a good correlation remains.

More complicated attempts to alleviate the problem of identifying "good" coefficient values include the introduction of various "adjusted" indices that attempt to parametrically define null models corresponding to independent partitions. The motivation for such procedures is so that the comparisons might be reported as a similarity relative to that which might be obtained "at random." For instance, one can construct adjusted indices by subtracting the expected value (under some null model and typically conditional on maintaining the numbers and sizes of groups in the two partitions) and subsequently rescaling the result by the difference between the maximum allowed value and the mean value [47]. One such example index, which uses a simple bound on the maximum allowed value, is the Adjusted Rand coefficient [47]

$$S_{\rm AR} = \frac{w_{11} - \frac{1}{M}(w_{11}+w_{10})(w_{11}+w_{01})}{\frac{1}{2}\left[(w_{11}+w_{10})+(w_{11}+w_{01})\right] - \frac{1}{M}(w_{11}+w_{10})(w_{11}+w_{01})}\,.$$

We remark that constructing a sharp upper bound subject to the constraint of fixed row and column sums in a given contingency table is known to be a very difficult

combinatorial optimization problem. Indeed, even the question of counting how many total ways there are to fill such a table under constrained row and column sum margins is a difficult one [23]. Hence, it is typically necessary to employ a weaker bound such as that used in the Adjusted Rand of [47], where the maximum value of w_{11} (which we recall denotes the number of pairs placed in same groups in both partitions) is given by the average of the number of same-group pairs for each partition individually.

As described in [64], adjusted indices can be problematic because using only the expected and maximum values in no way guarantees the accuracy of comparisons between similarity coefficients in different settings—for example, when comparing different pairs of partitions, which is what we do when we compare algorithmically-obtained communities to groupings constructed using self-identified demographics. In this paper, we approach the problem of identifying "good" coefficient values by directly standardizing the similarity coefficients and providing a context for the obtained results by computing coefficient z-scores (i.e., by examining for each coefficient value its number of standard deviations away from the mean). When necessary, such z-scores can be obtained non-parametrically using permutation tests [39]. Our decision to standardize comparisons using z-scores is motivated by the goals of the adjusted indices, our interest in demographic correlations, and our observations above including the random partition in the Zachary Karate Club network.

4.2. Permutation Tests and z-Scores. The aforementioned quantitative methods, including the "raw" S_i values of the pair-counting scores, might be appropriate for comparing partitions that are exceptionally close to one another, as it might be desirable in such cases to have some notion of distance between partitions. However, as we will demonstrate using the Caltech network in Section 5.3, it is apparent that two partitions can be very highly correlated with one another even if there are a large number of differences in individual node assignments. Accordingly, it seems more appropriate to report the correlation strength relative to that obtained at random, with a scaling dictated by the width of the distribution, rather than by raw similarity scores or their adjusted-by-maximum values. To provide a standardization of the pair-counting coefficients, we therefore calculate z-scores, $z_i = (S_i - \mu_i)/\sigma_i$, equal to the number of standard deviations σ_i that the S_i-value is better (more correlated) than the mean μ_i, for each of the S_i values ($i \in \{\mathrm{FM}, \Gamma, \mathrm{J}, \mathrm{M}, \mathrm{R}, \mathrm{AR}\}$). (For the Minkowski coefficient S_M, we also multiply by -1 to account for smaller S_M values corresponding to higher correlations.) Positive z-scores thus indicate (positive) correlations, whereas negative z-scores indicate anti-correlations.

In the Appendix, we recall (in simplified notation) the formulas for the mean and variance of the Rand coefficient under the "hypergeometric distribution" of equally likely assignments subject to maintaining the numbers and sizes of groups in each partition. We additionally show that z_R and a number of the other z_i-scores are indeed identical to each other upon standardization of their underlying S_i values. In situations for which simple formulas for the necessary moments do not appear to be available (i.e., for the Jaccard and Minkowski indices), we resort to the computationally straightforward (albeit intensive if one desires high accuracy) method of interpreting the calculated S_i values in terms of their distributions that we obtain using permutation tests [39], again under the same typical null model of equally-likely node assignments conditional on the constancy of the numbers and sizes of groups. Specifically, starting from two network partitions whose correlation we want to measure, we first calculate the similarity values S_i and then obtain a context for these values by repeatedly computing S_i under random permutation of the node assign-

ments in one of the partitions. (Performing additional random permutation in the second partition is redundant.) We thereby aim to compare the similarity coefficients between the two partitions to the distributions of such coefficients from the appropriate ensemble of partition pairs, while automatically preserving the numbers and sizes of groups in both partitions.

The permutation test strategy might seem problematic because the number of random permutations becomes computationally intractable for all but the smallest networks. In practice, however, the moments μ_i and σ_i converge rapidly. While numerical computation of the cumulative distribution values for specified S_i (the "p-value") indeed requires sampling a large fraction of the total ensemble, calculating z-scores only requires one to sample the first two moments of the distribution. In our computations, we typically use 10000 permutations for the Facebook networks (even for the larger ones, where the number of nodes is actually larger than the number of permutations considered) and observe the expected apparent convergence of the sample mean and variance, giving z-scores typically accurate to two significant figures. Because of the small number of samples that we consider, we do not include the S_i values for the non-permuted comparison in the calculation of the moments because this is a distinguished value not obtained under truly random conditions. We observe that the distributions of S_i values obtained using permutation tests between algorithmically-identified communities and self-identified demographic groupings appear to be roughly Gaussian, though certainly not precisely Gaussian. For instance, the observed skewnesses (third central moments normalized by σ_i^3) and kurtoses (fourth central moments normalized by σ_i^4) are always near 0 and 3, respectively. We stress, however, that because of significant non-Gaussian behavior in the distribution tails, the z-scores do not yield accurate p-values using a Gaussian assumption. Hence, while one should of course prefer to directly examine the p-values from the cumulative distribution, one cannot hope to calculate them without the full, computationally-unobtainable distribution. We thus focus our discussion on the z-scores themselves.

We return to the Zachary Karate Club network to illustrate these ideas. We consider a sequence of random permutations of the node assignments and calculate the similarity coefficients in order to compare the maximum-modularity partition (4 communities of 11, 5, 12, and 6 nodes) with the partition according to the observed split into two groups (equivalent to combining the 11-node and 5-node communities into one group and the 12-node and 6-node communities into another). After obtaining the means and standard deviations of the similarity coefficients from these permutations (see Table 4.2), which are in good agreement with the analytical formulas (when available; see the Appendix), we calculate the z-scores to compare these two partitions for each of the aforementioned pair-counting similarity measures. We similarly run permutation tests to compare the 4-community partition with the "random" partition (2 groups of 23 and 11 nodes), identified by circles and diamonds in Fig. 2.2. We note, in particular, that because the group sizes in the "random" partition are different from those in the "observed" split, the μ_i and σ_i values are also different in Table 4.2, as is expected because the ensemble of partitions obtained by random permutation must be different as a result of the constraint to preserve group sizes.

The z-scores in Table 4.3 reveal a remarkable simplification, as some of them appear to be identical to each other up to the number of significant figures in the table. Indeed, the Fowlkes-Mallows, Γ, Rand, and Adjusted Rand z-scores are provably identical, as their corresponding similarity indices are linear functions of one another (see, e.g., [48] and the discussion in the Appendix). Because of this equivalence, we

	FM	Γ	J	M	R	AR
"Observed" μ	0.3559	-2×10^{-5}	0.2045	1.3769	0.5064	3×10^{-5}
σ	0.0258	0.0419	0.0183	0.0261	0.0184	0.0371
"Random" μ	0.3780	8×10^{-5}	0.2147	1.4180	0.4765	-3×10^{-5}
σ	0.0252	0.0437	0.0178	0.0262	0.0191	0.0367

TABLE 4.2

Calculated means (μ) and standard deviations (σ), obtained through permutation tests (10^6 realizations), for each of the similarity coefficients shown in Table 4.1 (Fowlkes-Mallow, Γ, Jaccard, Minkowski, Rand, and Adjusted Rand) for partitions of the Zachary Karate Club network. In particular, we compare the 4-community, maximum-modularity partition to the "observed" split of the club into 2 new clubs (indicated by the open/closed symbols in Fig. 2.2) and a "random" split into 2 groups (identified by the node shapes in the same figure).

	$z_{\rm FM}$	z_Γ	$z_{\rm J}$	$z_{\rm M}$	$z_{\rm R}$	$z_{\rm AR}$
"Observed"	14.6	14.6	18.0	17.1	14.6	14.6
"Random"	0.343	0.343	0.322	0.329	0.343	0.343

TABLE 4.3

Calculated z-scores (i.e., number of standard deviations away from the mean), obtained using permutation tests, for each of the similarity coefficients in Table 4.1 (Fowlkes-Mallow, Γ, Jaccard, Minkowski, Rand, and Adjusted Rand) for partitions of the Zachary Karate Club network. In particular, we compare the 4-community, maximum-modularity partition to the "observed" split of the club into 2 new clubs (indicated by the open/closed symbols in Fig. 2.2) and a "random" split into 2 groups (identified by the node shapes in the same figure).

henceforth restrict our attention among these metrics with identical z-scores to the Rand coefficient z-score, which we denote by $z_{\rm R}$ (to respect the seminal role of the Rand coefficient in the pair-counting literature). We stress, however, that the linear transformations between the S_i values include information about the number of pairs M_c that are classified in same groups in the cth partition ($c \in \{1, 2\}$). Hence, when the numbers and sizes of groups in the partitions change, the raw similarity values S_i accordingly also change. As we demonstrate in Section 5, this implies that the values of the similarity indices can have different orderings in different comparisons even when their z-scores are identical, further supporting our preference to standardize such pair-counting using z-scores.

Another interesting observation is that the other z-scores in Table 4.3 (Minkowski and Jaccard) are also reasonably close to $z_{\rm R}$. This relative similarity in z-scores follows from the fact, shown in the Appendix, that their p-values for a specified comparison between two partitions are each necessarily identical to that for Rand (and its linearly-transformed equivalents) because of the imposed constraints on the numbers and sizes of groups in those partitions. However, the resulting $z_{\rm J}$ and $z_{\rm M}$ are *not* equivalent to $z_{\rm R}$ because the associated transformations from $S_{\rm R}$ to $S_{\rm J}$ and $S_{\rm M}$ are nonlinear. Consequently, because the p-values of a comparison are identical in each of these cases, the variations in the derived z-scores provide an (admittedly crude) indication of the variability in the shapes of the distributions.

In Fig. 4.1 (particularly in the left panel), one can see that the distributions of the similarity indices are clearly not Gaussian. Their distributions are not even known in general, except in the large-sample asymptotic limit [53]. Therefore, as emphasized in [48], there is no known theoretical threshold for deciding when these similarity measures indicate an unusually high correlation between two partitions. Nevertheless, as observed for instance in the right panel of Fig. 4.1, the distributions

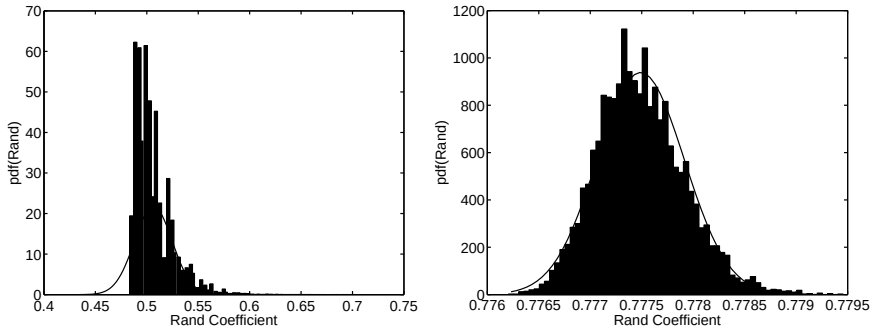

FIG. 4.1. *Example permutation distributions of the Rand coefficient obtained by comparing (Left) the 4-community, maximum-modularity partition identified for the Zachary Karate Club to random permutations of the observed 2-group split of the community (see Fig. 2.2) and (Right) the communities obtained by recursive application of the leading-eigenvector method to the largest connected component of the Caltech data (containing 762 nodes in 7 communities) with random permutations of the network partitioned according to identified House affiliation (9 categories, including one for those who did not identify an affiliation). For comparison, we have also plotted the Gaussian distributions with the same means (0.506 for the Zachary Karate Club and 0.777 for Caltech) and standard deviations (0.0184 and* 4.25×10^{-4}*) as the permutation distributions (which, respectively, have skewnesses of 1.67 and 0.421 and kurtoses of 7.22 and 3.37).*

of the pair-counting indices appear to become more Gaussian for even the smallest Facebook network we consider, so the z-scores of the traditional two-sided 95% and 99% confidence intervals should not deviate significantly from their Gaussian values of 1.96 and 2.58, respectively.

The Zachary Karate Club example, in conjunction with the discussion in the Appendix, illustrates our assertion that it is typically more useful to adjust similarity indices by subtracting the mean behavior and rescaling by the size of the standard deviation than to adjust them using the maximum possible deviation from the mean—obtaining "standardized" indices, cf. "adjusted" indices. The known equivalence for each similarity coefficient of the (albeit unknown) p-values and the relative similarity in the z-scores indicates that the z-scores provide more detailed information than any of the raw or "adjusted" (by maximum) indices by themselves. Moreover, the z_{R}-score can be calculated relatively easily from available formulas, such as in [46] or (in simplified form) in the Appendix.

Of course, as we will see with the Facebook examples in Section 5, calculating z-scores of the pair-counting indices is not a panacea, particularly when comparing networks of different sizes. Nevertheless, we find them exceptionally useful for examining the correlations between communities and the groupings according to the available demographics in our Facebook data. Before we concentrate on using these z-scores to measure correlations, it is also instructive to compare our results (discussed in Section 5) versus what might have been available using other methods, such as variation of information (VI) [64] and the (non-standardized) Adjusted Rand [47]. To do this, we show a scatter plot of z_{R} versus various relevant quantities in Fig. 4.2. One immediate observation is that while the Adjusted Rand S_{AR} values trend positively with z_{R} (recall that $z_{\mathrm{R}} = z_{\mathrm{AR}}$), there are situations with very small S_{AR} that have much larger z_{R} values than should be expected at random.

Because VI has a genuine metric structure [64], it has recently been used to examine networks whose community structure evolves gradually either in time [25] or

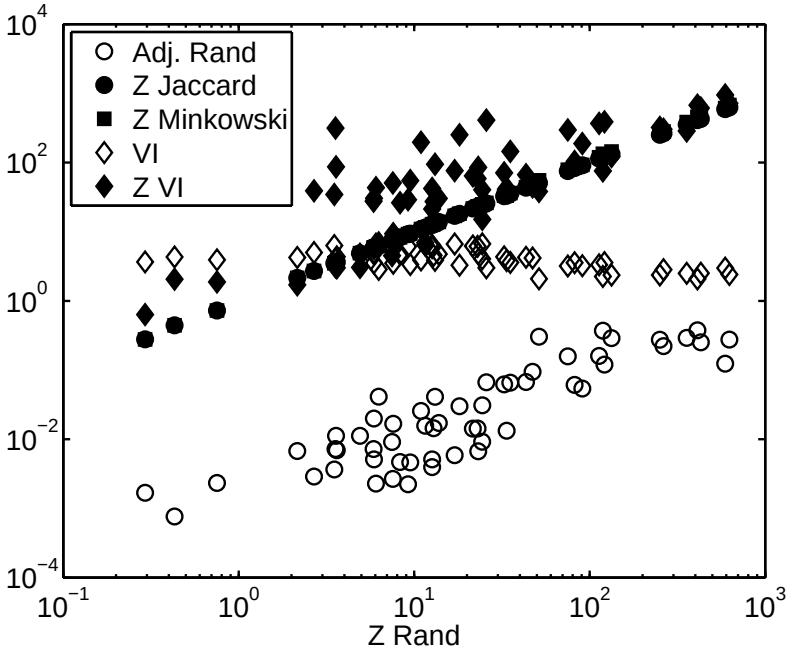

FIG. 4.2. *Scatter plot of z_{R} (i.e., the z-score of the Rand coefficient) on the horizontal axis versus (on the vertical axis) the other pair-counting z-scores (z_{J} and z_{M}), variation of information (VI), a z-score for VI obtained using permutation tests, and the adjusted Rand similarity index S_{AR}. The depicted data points are drawn from the 60 situations examined in Section 5: 5 universities compared with each of 4 demographic groupings for each of 3 networks per university (full, women-only, and men-only). We discuss the empirical observations in this plot at the end of Section 4.2.*

with respect to some system parameter [90]. However, while VI is an excellent way to compare two *nearby* partitions [64], even situations that are expected to give the most demographically-homogeneous communities—such as our comparison in Section 3 of the algorithmically-determined Caltech communities versus the House-affiliation groupings—give a pair of partitions with many detailed differences in individual node assignments. Such partition pairs are consequently not very close to each other according to a metric such as VI, yet they are obviously far more similar to each other than should be expected at random. Accordingly, raw VI scores appear to not be very useful here and are indeed only weakly correlated with z_{R} in Fig. 4.2. Indeed, we observe (not shown) that the mutual information between two such partitions is often a very small fraction of the total information. However, one can use permutation tests to process VI into a corresponding z_{VI}-score, which correlate strongly with the z_{R} values (again see Fig. 4.2). Hence, while the raw VI values do not indicate that the partitions are themselves close to each other in distance, the z_{VI}-scores show that they are nevertheless much closer than might be expected at random. Unfortunately, there does not appear to be a clear correspondence between z_{VI} and z_{R}, so we maintain our focus on the latter, in part due to the simplicity of its calculation (appealing again to the formulas in the Appendix).

5. Facebook Communities. Armed with the lessons from Section 4 that one can provide a better context for comparing network partitions using the z_i-scores of the different pair-counting similarity coefficients, we turn our attention to the Face-

Institution	Caltech	Georgetown	Oklahoma	Princeton	UNC
Nodes	1099	12195	24110	8555	24780
Connected Nodes	762	9388	17420	6575	18158
Connected Edges	16651	425619	892524	293307	766796
Average Degree	21.85	45.34	51.24	44.61	42.23
# Communities	7	48	5	5	5
Modularity	0.359	0.393	0.285	0.382	0.369

TABLE 5.1

Basic characteristics of our five Facebook networks: total number of users (nodes) in the data, number of nodes in the largest connected component, number of edges in the largest connected component, and average degree in the largest connected component. We additionally indicate the number of communities and modularity obtained using the leading-eigenvector method.

book data. We algorithmically generate one set of communities for the largest connected component of each institution's network, using Newman's leading-eigenvector method for simplicity. We then compare those communities to the partitions obtained by grouping users according to each of the self-identified characteristics: major, class year, high school, and dormitory/House. We examine five universities—California Institute of Technology (Caltech), Georgetown University, Princeton University, University of North Carolina at Chapel Hill (UNC), and University of Oklahoma—in order to concisely illustrate our methodology and findings. These institutions are of different sizes and different presumed predispositions towards organization according to the four available characteristics. In Table 5.1, we present their total node number, the number of nodes in their largest connected component, the number of edges in that component, and the average degree (i.e., average number of friends per user) of that component. In the same table, we also show the number of detected communities and modularity of the computed community structure.

For each institution and each comparison, we calculate the z-scores of the selected pair-counting similarity coefficients, which we obtained by generating distributions of the same coefficients from random permutations of the node assignments. We confirm that the obtained z-scores have reasonably converged by comparing with those obtained using half of the generated permutations and with the analytical formulas (A.1)–(A.3) for z_{R}. We typically report three significant figures in each case, although as discussed in Section 4.2 one should not expect to obtain more than two significant figures from 10000 random permutations.

Because there are situations in which individuals elected not to disclose some subset of the self-identifying characteristics (major, year, high school, and dorm), we were forced to create a separate "Missing" label for each of the demographics and group relevant users into that artificial group. In the present section, we ignore the artificial nature of this group. Accordingly, in the pie-chart dendrograms, we have consistently indicated such missing fields using white wedges in order to properly convey the extent of the missing data. We will revisit this issue in Section 6.

As an example of other productive ways to study Facebook networks, we also consider their single-gender subsets (i.e., subgraphs that include only same-gender nodes and links) and again investigate the correlation between communities and characteristics as one way of examining the similarities and differences between the "typical" behavior of women and men in these networks. We thus identify the largest connected components of the women-only and men-only subnetworks for each university, find the communities of these subnetworks (using the leading-eigenvector method), and

Institution	Caltech	Georgetown	Oklahoma	Princeton	UNC
Connected Nodes	217:459	4379:3937	8164:7870	2701:3095	9616:6996
Connected Edges	2349:6266	102398:82406	284279:170890	69195:64679	240130:131304
Average Degree	10.82:13.65	23.38:20.93	34.82:21.71	25.62:20.90	24.97:18.77
# Communities	4:7	6:8	11:10	5:17	5:10
Modularity	0.321:0.376	0.448:0.355	0.349:0.393	0.399:0.379	0.336:0.296

TABLE 5.2

Basic characteristics of the largest connected components in the single-gender Facebook subnetworks of each university (number of nodes, number of edges, and average degree in the largest connected component) and of the community structure (as obtained by the leading-eigenvector method). In each case, we give the number for the women-only network followed by that for the men-only network.

compare them to the available user demographics.

Examining the basic characteristics of each subnetwork and its communities listed in Table 5.2, we note that the sizes of the women-only and men-only networks do not sum to those of the largest components of each university. That is, ignoring cross-gender links has (unsurprisingly) reduced the total numbers of nodes in these largest components. We also comment that there are some potentially interesting changes in the numbers of communities listed in Table 5.2 (compare to Table 5.1), but further analysis of these possible changes would require careful study of the identified communities themselves (including a comparison with different community-detection algorithms) and is thus beyond the scope of the current discussion. Instead, we focus our attention on measuring the correlation between these communities of the single-gender subnetworks (as obtained from the leading-eigenvector method) and the other user demographics.

5.1. California Institute of Technology. We now revisit Caltech's community structure, which we previously examined by visual inspection in Section 3. In its most basic form, the leading-eigenvector method identifies 7 communities (with modularity $Q = 0.3594$), which can be seen in Fig. 3.1. As we discussed (and one can see clearly in the figure, colored according to self-identified House affiliation), the Caltech community structure has a strong correlation with House affiliation.

To investigate this quantitatively, we calculate the similarity coefficients of the 7-community partition versus the four available user characteristics (presented in Table 5.3). In so doing, we note that the raw similarity-coefficient values appear to be insufficient to the task of comparing these communities, which is unsurprising after the discussion of Section 4. The rank ordering of the correlation strengths of the communities with the different demographics is not consistent for different pair-counting indices—even for the ones that we already know are simple linear transformations of one another—because the changes due to the constants in the linear transformations between S_i-values and z_i-scores (see below and the Appendix) mask the relative order indicated by the z-scores (shown in Table 5.4). For instance, the raw Fowlkes-Mallows value (S_{FM}) appears to order the categories House, year, major, and high school (in descending order of correlation with the communities); whereas Rand (S_{R}) and Minkowski (S_{M}) order them House, high school, major, and year (recalling that smaller S_{M} values indicate better agreement). Meanwhile, although they each agree that the correlation with House is strongest, the raw S_i values differ wildly in how much they set apart the House correlation. In particular, the S_{R} and S_{M} values might lead one to interpret that the correlation with House is only marginally stronger than that with high school, even though Caltech is so tiny that it contains very few students

	S_{AR}	S_{FM}	S_{Γ}	S_{J}	S_{M}	S_{R}	VI
"Major"	0.0069	0.12	0.0078	0.0579	1.124	0.779	4.332
"House"	0.2892	0.3989	0.2943	0.2452	1.0231	0.8169	2.3579
"Year"	0.0167	0.1836	0.0168	0.1011	1.2591	0.7227	3.5089
"High School"	0.0112	0.0868	0.0169	0.0314	1.0473	0.8082	4.7316

TABLE 5.3

Similarity coefficients (Adjusted Rand, Fowlkes-Mallow, Γ, Jaccard, Minkowski, and Rand) and variation of information for comparing the 7-community partition of the Caltech data versus each of the four self-identified user characteristics.

	Full (z_{J},z_{M},z_{R})	Women (z_{J},z_{M},z_{R})	Men (z_{J},z_{M},z_{R})
Caltech: "Major"	3.64, 3.63, 3.63	0.276, 0.277, 0.279	2.15, 2.15, 2.15
"House"	153, 144, 133	58.8, 55.3, 50.8	145, 134, 120
"Year"	7.63, 7.6, 7.57	6.36, 6.32, 6.26	5.96, 5.94, 5.91
"High School"	4.9, 4.89, 4.88	0.725, 0.726, 0.727	0.442, 0.443, 0.443
Georgetown: "Major"	2.71, 2.71, 2.71	5.75, 5.74, 5.73	13.9, 13.9, 13.8
"Dorm"	123, 119, 114	81.1, 78.5, 75.3	36.8, 36.3, 35.7
"Year"	717, 677, 627	494, 458, 410	294, 281, 264
"High School"	25, 24.8, 24.6	11.6, 11.6, 11.5	-0.515, -0.515, -0.515
Oklahoma: "Major"	7.5, 7.49, 7.49	12.6, 12.6, 12.6	9.23, 9.22, 9.22
"Dorm"	26.4, 26.1, 25.7	3.61, 3.61, 3.6	18.3, 18.2, 18.1
"Year"	9.41, 9.4, 9.39	33.3, 33.2, 33.1	24.5, 24.5, 24.4
"High School"	21.8, 21.7, 21.6	12.8, 12.8, 12.8	23.3, 23.2, 23.2
Princeton: "Major"	44.8, 44.1, 43.4	49.3, 48.3, 47.1	12.9, 12.8, 12.8
"Dorm"	11, 10.9, 10.9	13.3, 13.2, 13.1	33.6, 33.1, 32.6
"Year"	483, 459, 428	288, 272, 252	408, 384, 353
"High School"	-4.41, -4.42, -4.43	-1.68, -1.68, -1.68	7.44, 7.42, 7.41
UNC: "Major"	23, 22.9, 22.9	8.32, 8.31, 8.3	5.95, 5.94, 5.93
"Dorm"	128, 125, 122	-3.97, -3.98, -3.99	3.59, 3.58, 3.57
"Year"	628, 612, 593	93.4, 92.4, 91.1	85.5, 84.4, 83.1
"High School"	17, 17, 16.9	6.15, 6.15, 6.15	3.51, 3.51, 3.5

TABLE 5.4

Permutation-test-obtained z-scores of the pair-counting similarity indices for comparing the algorithmically-identified communities of Facebook networks for each university versus self-identified user characteristics. The "Full" network includes all users in the largest connected component of the institution (see the statistics in Table 5.1). We also consider single-gender subgraphs, in which the networks labeled "Women" and "Men" only include links between individuals of the same gender. We include each of the three different pair-counting z-scores, despite the minor nature of their quantitative differences. (Recall that $z_{\mathrm{AR}} = z_{\mathrm{FM}} = z_{\Gamma} = z_{\mathrm{R}}$; see the discussion in the Appendix.) Large z-scores, particularly relative to others in the same data set, indicate significant organizational demographics.

at one time that come from the same high school.

These apparent disagreements across the S_i values occur even though we know that their corresponding p-values in the (unobtained) random distributions are identical. While we cannot directly calculate those p-values, we can obtain the z-scores for each. These z-scores (see Table 5.4) differ slightly quantitatively while maintaining a consistent interpretation of the roles of the four characteristics at Caltech: House is most important, followed distantly by year, high school, and major (again, in descending order of correlation with the communities).

To further emphasize the insufficiency of the raw similarity-coefficient values by themselves (again see Table 5.3), we highlight that only the Adjusted Rand coefficient

achieves the same ordering for the correlation of the communities with the four characteristics as the consistent z-score interpretation in Table 5.4. However, there is no guarantee for even this agreement. Indeed, we have observed examples in which the z-scores for comparing two different pairs of partitions are ordered differently from their S_{AR} values. Such disagreements result from the manners in which the numbers and sizes of groups in the different characteristic partitions impact the values of the coefficients. That is, while they are each functions of the number (w_{11}) of pairs common to both partitions , those functions include factors based on the total number of pairs (M) as well as the number of pairs (M_j) appearing in the same group in the jth individual partition ($j \in \{1, 2\}$). Given such dependence, coupled with the different sizes of the groups in the self-identified partitions, we should no longer be surprised by the poor performance of the raw S_i values and will only consider z_i-scores for the remainder of our discussion.

Applying the leading-eigenvector method to the single-gender Caltech networks yields 4 communities for the women and 7 for the men. Accordingly, while we again see the unsurprising extreme importance of House affiliation, the z-scores suggest that the effect is somewhat stronger for men than it is for women. The slightly positive association with class years is the same for both genders. Additionally, men seem to have a very slightly positive association by major, whereas women have essentially none. Neither single-gender network seems to positively associate according to their high school affiliation, which is sensible given how few people from the same high school are present at any one time at a small university like Caltech.

5.2. Other Universities. We now briefly present our observations for the other universities. For simplicity, we will often refer to a single z-score in our discussion, though the same conclusions hold for each of the z-score choices in Table 5.4.

5.2.1. University of North Carolina at Chapel Hill. As we saw in Section 3, visual inspection of the correlation between the community structure and demographic groups in the UNC network is not particularly fruitful. As shown in Fig. 3.2, the leading-eigenvector method finds 5 communities in the largest connected component. In contrast to what we observed in the Caltech data, the z-scores applied to the full UNC network suggest that class year is the primary organizing characteristic (of the four available to us) and that dormitory residence is also prominent. Major and high school have smaller but noticeable positive effects. Interestingly, the single-gender networks seem to remove the effect of dorm entirely, with small negative z-scores for women and small positive ones for men. The z-scores for major and high school become smaller but remain positive.

We highlight the large values of the z-scores for UNC, especially as compared to Caltech. In particular, while we strongly advocate the use of z-scores to measure the strengths of correlations (relative to the other available quantitative alternatives), it is nevertheless clear that such a statistical statement remains imperfect when comparing the visually very strong House correlation at Caltech ($z_{\mathrm{R}} = 133$) versus the strength of the year correlation at UNC ($z_{\mathrm{R}} = 593$). The simple explanation for this discrepancy is the different sizes of the two data sets (762 nodes versus 18158 nodes), which causes the seemingly weaker correlation in the latter to be statistically much further in the tail of the random distribution than the former. We consider further study of this size effect to be a potentially valuable avenue of future work.

5.2.2. University of Oklahoma. The other large state university we consider is the University of Oklahoma, which has 5 communities in the partition obtained

using the leading-eigenvector method. Based on the z-scores in Table 5.4, the communities seem to break up primarily according to a combination of dormitory residence and high school, with year and major as somewhat (but not appreciably) smaller determining factors. Interestingly, all four z-scores have the same order of magnitude for Oklahoma, which is very different from what we observed for the other universities. Moreover, none of these values is even close to as large as the largest for UNC, despite the similar sizes of their respective networks. Accordingly, it is more difficult to confidently indicate how the full communities are organized at University of Oklahoma. Class year seems to be the most important characteristic in both single-gender networks at Oklahoma, and dormitory residence seems to be more important to men than it is to women.

However, a visual inspection of pie-chart dendrograms (not included here) indicates that significantly fewer people at Oklahoma self-identified their dormitory residence than their high school. We thus postpone further conclusions here to Section 6, in which we consider different ways of handling missing demographic data.

5.2.3. Princeton University. We show the community structure for Princeton University (with 5 groups, obtained using the leading-eigenvector method) in Fig. 5.1 (colored by class year and by major). Note that the size of the Princeton data set (with over 8500 nodes, including 6575 of them in the largest connected component) is disproportionately large relative to the institution's size; this is a result of (relatively) early Facebook adoption at Princeton. The z-scores in Table 5.4 reveal that Princeton students break up into communities predominantly according to class year, with a reasonably large organization by major, a small positive organization by dormitory, and an organization that appears to be correlated (slightly) negatively with high school affiliation. The z-scores in the single-gender networks suggest that class year is very important for both men and women, major is more important for women than for men, dorm is more important for men than for women, and high school has a small positive effect for men but a small negative one for women.

Princeton has the smallest dorm z-score among all the universities we examine. This may be due in part to an ambiguous definition of dorm at Princeton, as different individuals might reasonably use this field to indicate their residential college, eating club, or current living location. It may also be an effect of missing data.

5.2.4. Georgetown University. We show the community structure for Georgetown University in Fig. 5.2 (with 48 communities identified by the leading-eigenvector method and colored by class year/dormitory). Similar to UNC, the z-scores in Table 5.4 indicate that the communities break up primarily according to class year and secondarily by dormitory residence. There also seems to be some importance to high school affiliation but very little to major (which, by contrast, seems to have more relevance at UNC). Indeed, major seems to be less important for Georgetown's social network than for any of the other four universities.

The z-scores of the single-gender networks indicate that year and dorm remain the primary and secondary organizing characteristics. High school seems to have some positive organizing effect for the communities in the network of men, whereas essentially no effect (the z-score is slightly negative) in the network of women. The z-score for major indicates a small positive effect for both genders.

5.3. Additional Considerations. Before continuing further, we strongly emphasize that any conclusions derived from the numbers in Table 5.4 should be interpreted with questioning caution, as one should of course be careful about how they

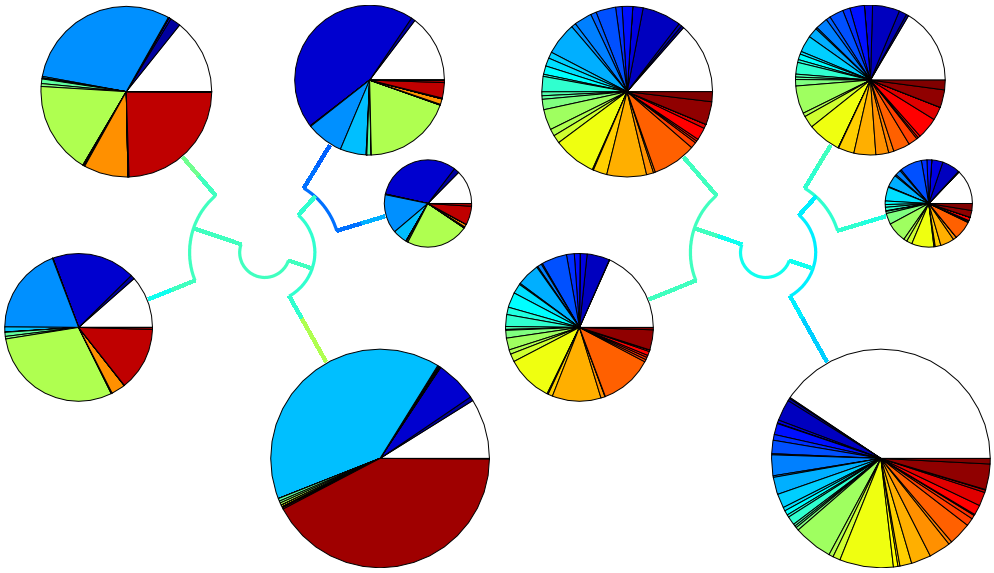

FIG. 5.1. *[Color] Pie-chart dendrograms of Princeton, colored by (Left) class year and (Right) major. (As before, white slices correspond to people who didn't identify the relevant characteristic.)*

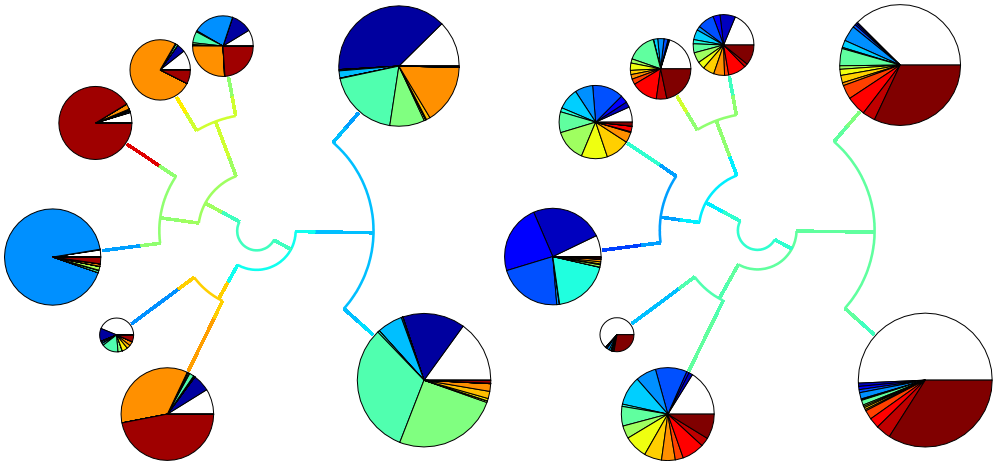

FIG. 5.2. *[Color] Pie-chart dendrograms of Georgetown, colored by (Left) class year and (Right) dormitory residence. (As before, white slices correspond to people who didn't identify the relevant characteristic.) The smallest pie near the 8 o'clock position is actually a collection of 38 very small communities that we combined in the plot for visual simplicity.*

might be influenced by our chosen methodologies. For instance, one should be curious about the possible role of the missing user characteristics in the calculations of the correlation measures. In particular, the sometimes large number of unidentified characteristics, shown as white slices in the pie-chart dendrograms, might skew the results through their large number. We attempt to better understand the role of this missing data in Section 6. One should also be wary of the possible influence of the selected definition of "community" and the method of its detection.

As an example of the effect of the community identification, we consider the pos-

	7 communities from eigenvector method alone Full (z_J,z_M,z_R)	6 communities from eigenvector + KLN Full (z_J,z_M,z_R)
Caltech: "Major"	3.64, 3.63, 3.63	3.14, 3.14, 3.13
"House"	153, 144, 133	202, 187, 168
"Year"	7.63, 7.6, 7.57	3.62, 3.61, 3.60
"High School"	4.9, 4.89, 4.88	5.02, 5.01, 5.00

TABLE 5.5

Permutation-test-obtained z-scores of the pair-counting similarity indices for comparing the 6-community partition of the Caltech data, obtained by the leading-eigenvector method plus KLN iterations, versus each of the self-identified user characteristics. For convenience, we also show the z-scores from Table 5.4 for the 7 communities obtained by the leading-eigenvector method alone.

sible role of the specific community-identification algorithm in the results obtained on the full Caltech network by computing its community structure in a different manner: starting from the same leading-eigenvector method, we now post-process the resulting communities with KLN iterations that can be used to improve the partitioning further [73, 74]. This refined method returns 6 communities (we previously obtained 7) and a slightly higher modularity ($Q = 0.3987$). Over 20% of the nodes are assigned differently by the KLN iterations, so it is certainly possible that this could affect the correlations with the demographics. Comparing the results of Table 5.5 with the earlier results in Table 5.4, we see that while the z-scores change slightly, the main qualitative conclusions remain almost exactly the same. Namely, the Caltech community structure is very heavily correlated with the House structure and is additionally correlated, though much more weakly, with (in descending order) high school, year, and major. Contrasting this with the results in Table 5.4, we see that the ordering of high school and year have been reversed, with the class year z-scores now about half as large as they were before. This comparison cautions against the over-dependence on any of the specific quantitative values. At the same time, the qualitative agreement between the two community-detection methods is excellent and is especially evident in the identification of the very strong correlation with House.

We stress again that one should not attempt to interpret the exceedingly large z-scores in Table 5.4 in terms of the p-values one would obtain with those scores in a Gaussian distribution. We know virtually nothing about the tails of our distributions, and there is no *a priori* reason to expect them to be Gaussian. We additionally caution that we have significant amounts of missing data affecting the results in Table 5.4, which we attempt to address in the next Section.

One should also stress the difference between causation and correlation; we have examined *correlations* in this paper. As discussed in the sociological literature on SNSs (see [9] and references therein), it is obviously very interesting and important to attempt to discern which common characteristics have resulted from friendships and which ones might perhaps influence the formation of friendships. It might also be interesting to employ ERGMs using various models of link formation in addition to the fully random one that we implicitly consider through the selected standardizations. In terms of the individual characteristics discussed above, high school and class year are outside an individual's control, so one would expect those particular correlations to also indicate how some friendships might have formed. Common residences, on the other hand, can both encourage new friendships and arise because of them. We note, finally, that SNS friendships provide only a surrogate for offline ones, so that one can

	Connected Users	Not Missing Major	Not Missing Dorm/House	Not Missing Year	Not Missing High School	Not Missing Any
Caltech	762	692	597	655	635	501
Georgetown	9388	7510	6594	8374	7562	4774
Oklahoma	17420	15779	7203	13732	14998	5510
Princeton	6575	4940	4355	5801	5214	2920
UNC	18158	15492	8989	15883	15414	6719

TABLE 6.1
Sizes of each of the data sets used in the different procedures for handling missing data.

also expect to find differences between the community structure of, e.g., Facebook networks and the real-life networks they imperfectly represent [9].

6. Missing Demographic Data. In the previous section, we examined the correlations between the algorithmically-identified communities of the Facebook networks and the available node characteristics. Unfortunately, as noted above, because these characteristics rely on self-identification on individuals' Facebook pages, many nodes are missing demographic data. For simplicity, we previously grouped such incompleteness in the node demographics into separate "Missing" groups for each characteristic and proceeded to calculate z-scores without any further attention to this problem. However, as the large white areas (which we use to indicate the "Missing" values in our visualizations) in some of the pies and the total size of the missing data problem enumerated in Table 6.1 both evince, one needs to worry about the role of missing data in our conclusions. One could approach the issue of missing data using sophisticated tools such as multiple imputation, likelihood, or weighting methods [45]. For the purposes of the present study, however, we modestly address only the aggregate size of the effect of missing data on our measured correlations using various restrictions of our data to users with more complete demographic information.

Because the missing user characteristics do not affect the topologies of the friendship graphs, we start with the algorithmically-determined community assignments for each individual that we obtained above for the largest connected components of the Facebook networks. We remark that while such online networks have no missing information in the nodes and links (and hence in the community assignments), there is of course missing sociological information resulting from the fact that online friendships are not equivalent to offline ones, as the online social network constitutes an approximate but imperfect proxy for the offline one. [9].

For each demographic characteristic, we are then faced with a contingency table of community assignments and groupings made by the characteristic, with one of those groups specifically identifying the missing values. In order to examine a single user characteristic, it is natural to simply ignore the users who left that characteristic field empty. This is equivalent to studying the same contingency table but ignoring the column corresponding to the missing data. That is, because the underlying network has not changed at all, we neither restrict information from this network nor recompute the communities; instead, we take the identified communities and calculate correlations with a specified characteristic using only those users who specified a value for that characteristic. We show the z_{R}-scores from this data processing in Table 6.2 as the "Removed Missing" results, obtained here from the analytical formulas (A.1)–(A.3) rather than from permutation tests. For easy comparison, we also include the original "Included Missing" results from Table 5.4 in Table 6.2 (though we have now recomputed these numbers analytically from the formula, for comparison).

	Included Missing			Removed Missing			Removed Any Missing		
	F	W	M	F	W	M	F	W	M
Caltech: "Major"	3.62	0.293	2.15	3.48	0.294	2.22	3.85	-0.0625	0.916
"House"	133	51.4	119	195	56.4	132	175	48.9	122
"Year"	7.59	6.28	5.89	8.14	5.96	5.88	6.84	4.62	4.98
"High School"	4.91	0.752	0.431	1.72	2.27	2.51	1.1	1.57	1.45
Georgetown: "Major"	2.69	5.87	13.8	15.8	7.43	3.72	15.7	6.76	1.46
"Dorm"	113	75	35.3	172	99.2	27	142	84.7	28.1
"Year"	626	411	264	712	491	286	661	458	198
"High School"	24.5	11.5	-0.522	12.7	2.39	6.5	8.29	2.08	4.45
Oklahoma: "Major"	7.56	12.6	9.22	8.45	14	12.5	15.8	8.5	7.32
"Dorm"	25.8	3.56	18.1	63.8	170	164	57.8	156	147
"Year"	9.5	33.6	24.4	3.65	33.9	25.1	12.7	12	37.7
"High School"	21.6	12.7	23	22.6	29.1	47.6	27.1	18	29.9
Princeton: "Major"	43.4	47.2	12.9	18.6	9.6	6.74	8.76	4.65	5.46
"Dorm"	10.9	13.2	32.6	54.6	34.2	49.5	54.9	32.5	37.9
"Year"	429	251	357	407	276	403	191	122	280
"High School"	-4.43	-1.68	7.48	7.03	2.75	2.32	4.19	2.79	2.09
UNC: "Major"	23.2	8.3	5.9	22.2	3.57	2.17	23.7	2.42	2.36
"Dorm"	121	-3.95	3.59	96.7	48	21.7	99.5	45.7	18.7
"Year"	592	90.8	82.1	691	100	87.9	307	37.5	30.8
"High School"	17	6.06	3.51	82.2	28	27.8	58.9	13.9	14.7

TABLE 6.2

Analytically-obtained z_R-scores (using the formulas in the Appendix) for comparing the algorithmically-identified communities of each university's Facebook network versus the self-identified user characteristics. In each situation ("Included Missing", "Removed Missing", and "Removed Any Missing"), the first subcolumn gives the results for the full network (F), the second gives the result for the women-only subnetwork (W), and the third gives the result for the men-only subnetwork (M). The "Included Missing" column corresponds to the results previously shown in Table 5.4 that treat missing data as a separate identifying group for each characteristic. (The slight difference in numbers, which we show for comparative purposes to illustrate the formulas in the Appendix, arises from the fact that we obtained them from an analytical formula in this table but from numerical computations using permutation tests in Table 5.4.) We obtain the "Removed Missing" results for each characteristic by removing individuals who did not disclose that specific characteristic from the corresponding similarity-index calculation. The "Removed Any Missing" results include only individuals who self-identified each of the four characteristics.

While the removal of missing data seems reasonable enough, such a procedure introduces the potential problem that the total sizes of the partitions for a given university are no longer equal (as one can see in Table 6.1). For instance, examining the pie-chart dendrograms of the UNC data in Figure 3.2 clearly reveals that there are significantly more missing dormitory identifications than missing year identifications. The procedure outlined above thus includes the community assignment of a far smaller number of users at UNC for dealing with missing dorm data than it does in the correlation calculated after removal of the missing year data. Given the expected trends in z-scores with system size, this difference in the data sizes should be a concern. As such, we also consider a more aggressive procedure for handling the missing data, labeled "Removed Any Missing" in Table 6.2, in which we remove a user from consideration in the correlation calculation if *any* of their four characteristics (major, House/dorm, year, and high school) are missing, to ensure that the sizes of the data are the same for a given university. In so doing, however, we caution that a significant fraction of the total available data can be removed (again see Table 6.2).

The main results in Table 6.2 agree qualitatively with those in Section 5, when we treated the "missing" designation as if it were any other characteristic identifier. In particular, the previously-discussed roles of the dominant correlations are mostly reinforced by these additional results. The dominant correlation in the Caltech community structure remains with House assignment. The Georgetown communities are most strongly correlated with class year and are also very strongly related to dormitory. While Oklahoma's correlations are also similar to those reported previously, this analysis of missing data suggests that there might be a stronger organization according to dormitory residence (especially for its single-gender subnetworks). The communities in the Princeton network retain their strong correlation with class year, but once again the processing of the missing data suggests that there is a stronger correlation with dormitory than asserted above (though one does not see the same large effect with the single-gender networks in this case). The effect of major also seems to be somewhat smaller here, and the effect of common high school now appears to be somewhat positive. We similarly again see the strong correlations with year and dorm in the UNC network, but examining the missing field suggests a stronger correlation with high school than what we observed in Section 5. These additional considerations of the missing data also indicate that some of the possible gender differences suggested by Table 5.4 may not be real effects.

7. Discussion and Conclusions. We have shown that the tools of network science—and of community detection in particular—are demonstrably useful for studying the online social networks of universities and inferring interesting insights about the prominent driving forces of community development in their corresponding offline social networks. In particular, we used an eigenvector-based network partitioning algorithm (due to Newman [73]) to detect communities in the Facebook networks of individual universities. We then investigated measures of comparing network partitions (specifically, the algorithmically-identified communities) with those obtained by grouping individuals according to self-identified characteristics such as class year, dormitory/House, major, and high school. While numerous pair-counting measures using different algebraic combinations of terms are available in the literature, we observed that most of the variability in the resulting values disappears when such similarity scores are interpreted statistically using z-score values. We found that z-scores provide an immediate (though not quantitatively perfect) interpretation about the likelihood that such values might arise at random, indicating significant correlations between the algorithmically-identified communities and multiple self-identified characteristics. Additionally, considering the missing data in the Facebook networks strongly reinforced the above qualitative conclusions concerning the major organizing principles of each university's Facebook network as a whole, while also revealing some likely strong correlations that the inclusion of the "missing" field in user characteristics seemed to wash out in the full networks.

Our z-score computations allowed us to go beyond visual examination of community structure to make additional interesting insights about the organizational structure of university Facebook networks and, in principle, of the offline social structures they imperfectly represent. We found that the organizational structure at Caltech, which depends very strongly on House affiliation, is starkly different from those of the other universities we studied. We also observed that Georgetown, Princeton, and UNC communities are organized predominantly by class year, with secondary effects due to dormitory residence (UNC and Georgetown) or major (at Princeton, though different ways of processing missing data indicate dormitory residence as the secondary influ-

ence there). Oklahoma's communities showed the strongest positive effect of common high school, at a level roughly as important as dormitory residence. Different ways of processing the missing demographic data at UNC also indicated a strong correlation with high school, but not quite as large as with dormitory. Examining each university's single-gender subnetworks (including only links between individuals of the same gender) suggested some possible differences between the communities of women and men, though some of the quantitative details varied with the different means of treating the missing data.

The above sociological conclusions have multiple facets. First, some of our observations confirm conventional wisdom or are intuitively clear, providing soft verification of our analysis via expected results—e.g., that Houses are important at Caltech, class year is often important at large universities, and that high school plays a larger role at state universities. Second, the resulting confidence in the methodology lends support for the other observations, such as the relative importance of dormitory and high school at Oklahoma or possible differences observed in the gender subnetworks. Additionally, the heterogeneity of the demographic data inside each community (as visualized by the differently-colored slices in each community in the pie-chart dendrograms), along with the calculated important roles of multiple types of demographics in the Facebook networks, indicates that no single attribute can entirely explain the community structure. Otherwise, one would expect the community pies to each be dominated mostly by one color. While this is not surprising sociologically, it demonstrates quite poignantly that one should not simply attempt algorithmically to find a single "best" network partition that suffices to explain any sort of unique structure or organization of a network. (This was also pointed out recently in [20, 56].) The observed community heterogeneity also hints that the real structure of the Facebook networks might involve important factors beyond those encoded in the links themselves. While such complexity in interpersonal relationships is of course expected, declared links are the only information used in the available community-detection algorithms because of constraints in the available data. Nevertheless, as we illustrate using the Facebook examples, it is important to highlight that new methods need to be developed for situations in which more data can be incorporated.

In the future, it would be extremely interesting to systematically investigate similar observed features of Facebook networks by extending this investigation to other universities, as such a wide comparative study might allow for increased understanding about the factors that drive their social organization. To conduct such an investigation, it would also be desirable to incorporate data for additional demographic characteristics (such as fraternity/sorority affiliation, ethnicity, and religion) that one would expect to lead to the formation of cohesive communities. It might also be interesting to investigate whether the newest students on campus organize differently than their older peers; for instance, one might hypothesize that first-year student communities would more highly correlate with dormitory residence. The present paper attempts to provide foundational steps for such a desired comparative investigation in order to construct and demonstrate a meaningful methodology.

We hope that the present study is instructive not only in outlining the concepts of community detection, but also in demonstrating the use of statistically-interpreted pair-counting similarity scores. The latter contributions include the presentation of an easy-to-implement analytical formula for the z-score of the Rand index in the Appendix. We also demonstrated the use of permutation tests for indices for which such a simple formula appears to be unknown.

Numerous interesting open questions remain. Research on community detection has thus far focused on the development of methods for *structural* community detection. Although structural communities, constructed using only link topologies and weights, have been shown to correspond closely to functional communities in some situations [21, 28], such correspondence is in general only approximate. To better determine the desired functional groups of networks, one should in principle explicitly incorporate appropriate system knowledge directly into graph-partitioning algorithms, but it is not typically clear how to do so. This is related to a simple, very important, and underdeveloped idea that we have attempted to explore in the present paper: How does one examine the features of network communities after they have been obtained and (ultimately) what does one do with this information? Our goal in this paper has been to use example networks that are familiar from everyday experience to present one manner in which one can investigate the features of algorithmically-constructed communities. To do this, we primarily employed pair-counting similarity scores and an associated statistical interpretation. Although we found that z-scores virtually eliminated the quantitative differences between different pair-counting similarity scores for a given institution-characteristic combination, the comparison with the statistics of other correlative measures (specifically, with variation of information) holds only qualitatively. Moreover, while the tendency for z-scores to increase with network size is intuitively clear, this trend obscures the ability to make quantitative comparisons between different institutions. Simultaneously, the typically positive and sometimes extremely large values of the z-scores we observed in comparing partitions point to unsurprisingly complex dependencies of the community structures of the online social networks on multiple characteristics of the individuals involved.

The continuing study and refinement of community-detection techniques have the potential to make a significant impact not only scientifically but also in everyday life. For example, alumni associations from several universities can exploit network structures to improve the services that they offer [60, 107]. Indeed, one of the potentially very useful practical applications of community detection is to suggest the actual groups of individuals who should be invited to events, solicited for funding, etc., as one would expect these groups to be determined approximately but not exactly by known demographic labels such as dorm/House affiliation. More generally, community detection might be used to make intelligent predictions about unknown or withheld demographic information. As discussed in a recent paper by Clauset, et al. [20], the tools of community detection can also be used to infer missing edges (and hence missing social ties). Finally, because of the established and accelerating importance and prominence of community detection, we hope that the present paper offers not only an analysis of a fascinating example but also provides an opening for other mathematical scientists to make contributions to this field.

Acknowledgements. We thank Adam D'Angelo and Facebook for providing the data used in this study. We also acknowledge Danah Boyd, Barry Cipra, Aaron Clauset, Barbara Entwisle, Katie Faust, Avi Feller, Dan Fenn, James Fowler, Justin Howell, Nick Jones, Tom Maccarone, Jim Moody, Mark Newman, Andy Shaindlin, and Ashton Verdery for useful discussions at various stages of this project. We are especially indebted to Aaron Clauset and James Fowler for thorough readings of a draft of this manuscript and to Christina Frost for developing some of the graph visualizations that we used (among others discussed in [32]). ALT was funded by the NSF through the Alliance for Graduate Education and Professoriate program at UNC (NSF HRD-0450099). EDK's primary contributions to this project were made

while he was a student at California Institute of Technology, funded by Caltech's Summer Undergraduate Research Fellowship (SURF) program. PJM was funded by the NSF (DMS-0645369) and by start-up funds provided by the Institute for Advanced Materials, Nanoscience, and Technology and the Department of Mathematics at the University of North Carolina at Chapel Hill. MAP did some of his work on this project while a member of the Center for the Physics of Information and the Department of Physics at California Institute of Technology.

REFERENCES

[1] *House system at the California Institute of Technology*. `http://en.wikipedia.org/wiki/House_System_at_Caltech`.

[2] G. Agarwal and D. Kempe, *Modularity-maximizing network communities via mathematical programming*. arXiv:0710.2533, 2008.

[3] R. Albert and A.-L. Barabási, *Statistical mechanics of complex networks*, Reviews of Modern Physics, 74 (2002), pp. 47–97.

[4] A. Arenas, A. Fernández, and S. Gómez, *Analysis of the structure of complex networks at different resolution levels*, New Journal of Physics, 10 (2008), 053039.

[5] L. Backstrom, D. Huttenlocher, J. Kleinberg, and 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.

[6] J. P. Bagrow and E. M. Bollt, *A local method for detecting communities*, Physical Review E, 72 (2005), 046108.

[7] D. Boyd, *Viewing american class divisions through Facebook and MySpace. Apophenia* blog essay. `http://www.danah.org/papers/essays/ClassDivisions.html`. June 24, 2007.

[8] ———, *Why youth (heart) social network sites: The role of networked publics in teenage social life*, in MacArthur Foundation Series on Digital Learning - Youth, Identity, and Digital Media Volume, D. Buckingham, ed., MIT Press, Cambridge, MA, 2007, pp. 119–142.

[9] D. M. Boyd and N. B. Ellison, *Social network sites: Definition, history, and scholarship*, Journal of Computer-Mediated Communication, 13 (2007), art. 11.

[10] U. Brandes, D. Delling, M. Gaertler, R. Goerke, M. Hoefer, Z. Nikoloski, and D. Wagner, *Maximizing modularity is hard*. physics/0608255 (2006).

[11] R. L. Brennan and R. J. Light, *Measuring agreement when two observers classify people into categories not defined in advance*, British Journal of Mathematical and Statistical Psychology, 27 (1974), pp. 154–163.

[12] R. J. Brook and W. D. Stirling, *Agreement between observers when the categories are not specified in advance*, British Journal of Mathematical and Statistical Psychology, 37 (1984), pp. 271–282.

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

[14] R. S. Burt, *Structural holes and good ideas*, American Journal of Sociology, 110 (2004), pp. 349–399.

[15] ———, *Brokerage and Closure: An Introduction to Social Capital*, Oxford University Press, Oxford, 2005.

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

[17] R. J. G. B. Campello, *A fuzzy extension of the Rand index and other related indexes for clustering and classification assessment*, Pattern Recognition Letters, 28 (2007), pp. 833–841.

[18] A. Chin and M. Chignell, *Identifying active subgroups within online communities*, in Proceedings of the Centre for Advanced Studies (CASCON) Conference, Toronto, Canada, 2007.

[19] A. Clauset, *Finding local community structure in networks*, Physical Review E, 72 (2005), 026132.

[20] A. Clauset, C. Moore, and M. E. J. Newman, *Hierarchical structure and the prediction of missing links in networks*, Nature, 453 (2008), pp. 98–101.

[21] L. Danon, A. Diaz-Guilera, J. Duch, and A. Arenas, *Comparing community structure identification*, Journal of Statistical Mechanics: Theory and Experiment, (2005), P09008.

[22] G. F. Davis, M. Yoo, and W. E. Baker, *The small world of the American corporate elite, 1982-2001*, Strategic Organization, 1 (2003), pp. 301–326.
[23] J. A. De Loera and B. Sturmfels, *Algebraic unimodular counting*, Mathematical Programming, 96 (2003), pp. 183–203.
[24] S. N. Dorogovtsev, A. V. Goltzev, and J. F. F. Mendes, *Critical phenomena in complex networks*. 0705.0010 (2007).
[25] D. Fenn, M. A. Porter, N. S. Jones, and N. F. Johnson, *Temporally evolving communities in multi-channel data*. In preparation (2008).
[26] M. Fiedler, *Algebraic connectivity of graphs*, Czech. Math. Journal, 23 (1973), pp. 298–305.
[27] S. Fortunato and M. Barthelemy, *Resolution limit in community detection*, Proceedings of the National Academy of Sciences, 104 (2007), pp. 36–41.
[28] S. Fortunato and C. Castellano, *Community structure in graphs*, in Encyclopedia of Complexity and System Science, B. Meyers, ed., Springer Verlag, Heidelberg, Germany, 2008 (arXiv:0712.2716).
[29] E. B. Fowlkes and C. L. Mallows, *A method for comparing two hierarchical clusterings*, Journal of the American Statistical Association, 78 (1983), pp. 553–569.
[30] S. Fragoso, *WTF a crazy brazilian invasion*, in Proceedings of CATaC 2006, F. Sudweeks and H. Hrachovec, eds., Murdoch University, Murdoch, Australia (2006).
[31] O. Frank and D. Strauss, *Markov graphs*, Journal of the American Statistical Association, 81 (1986), pp. 832–842.
[32] C. Frost, *Network and community visualizations*. In preparation (2008).
[33] T. M. J. Fruchterman and E. M. Reingold, *Graph drawing by force-directed placement*, Software—Practice and Experience, 21 (1991), pp. 1129–1164.
[34] R. Gajjala, *Shifting frames: Race, ethnicity, and intercultural communication in online social networking and virtual work*, in The Role of Communication in Business Transactions and Relationships, M. B. Hinner, ed., Peter Lang, New York, NY, 2007, pp. 257–276.
[35] N. W. Geidner, C. A. Fook, and M. W. Bell, *Masculinity and online social networks: Male self-identification on Facebook.com*, in Paper presented at Eastern Communication Association 98th Annual Meeting, Providence, RI, 2007.
[36] M. Girvan and M. E. J. Newman, *Community structure in social and biological networks*, Proceedings of the National Academy of Sciences, 99 (2002), pp. 7821–7826.
[37] S. A. Golder, D. Wilkinson, and B. A. Huberman, *Rhythms of social interaction: Messaging within a massive online network*, in Proceedings of Third International Conference on Communities and Technologies, C. Steinfield, B. Pentland, M. Ackerman, and N. Contractor, eds., Springer, London, U.K., 2007, pp. 41–66.
[38] M. C. González, H. J. Herrmann, J. Kertész, and T. Vicsek, *Community structure and ethnic preferences in school friendship networks*, Physica A, 379 (2007), pp. 307–316.
[39] P. Good, *Permutation, Parametric, and Bootstrap Tests of Hypotheses*, Springer-Verlag, New York, NY, 2005.
[40] R. Guimerà and L. A. N. Amaral, *Functional cartography of complex metabolic networks*, Nature, 433 (2005), pp. 895–900.
[41] R. Guimerà, B. Uzzi, J. Spiro, and L. A. N. Amaral, *Team assembly mechanisms determine collaboration network structure and team performance*, Science, 308 (2005), pp. 697–702.
[42] P. R. Haunschild, *Interorganizational imitation: The impact of interlocks on corporate acquisition activity*, Administrative Science Quarterly, 38 (1993), pp. 564–592.
[43] L. Hjorth and H. Kim, *Being there and being here: Gendered customising of mobile 3G practices through a case study in Seoul*, Convergence, 11 (2005), pp. 49–55.
[44] T. Hogg, D. Wilkinson, G. Szabo, and M. Brzozowski, *Multiple relationship types in online communities and social networks*, in Proc. of the AAAI Spring Symposium on Social Information Processing, AAAI Press, 2008.
[45] Nicholas J. Horton and Ken P. Kleinman, *Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models*, The American Statistician, 61 (2007), pp. 79–90.
[46] L. Hubert, *Nominal scale response agreement as a generalized correlation*, British Journal of Mathematical and Statistical Psychology, 30 (1977), pp. 98–103.
[47] L. Hubert and P. Arabie, *Comparing partitions*, Journal of Classification, 2 (1985), pp. 193–218.
[48] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*, Prentice Hall, Englewood Cliffs, NJ, 1988.
[49] S. C. Johnson, *Hierarchical clustering schemes*, Psychometrica, 32 (1967), pp. 241–254.
[50] T. Kamada and S. Kawai, *An algorithm for drawing general undirected graphs*, Information

Processing Letters, 31 (1989), pp. 7–15.

[51] B. Karrer, E. Levina, and M. E. J. Newman, *Robustness of community structure in networks*, Physical Review E, 77 (2008), 046119.

[52] V. Krebs, *Orgnet.com: Social network analysis software & services for organizations, communities, and their consultants.* `http://www.orgnet.com` (2008).

[53] E. Kulisnkaya, *Large sample results for permutation tests of association*, Communications in Statistics – Theory and Methods, 23 (1994), pp. 2939–2963.

[54] R. Kumar, J. Novak, and A. Tomkins, *Structure and evolutoon of online social networks.* 12th International Conference on Knowledge Discovery and Data Mining, 2006.

[55] C. Lampe, N. Ellison, and 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.

[56] A. Lancichinetti, S. Fortunato, and J. Kertesz, *Detecting the overlapping and hierarchical community structure of complex networks.* arXiv:0802.1218 (2008).

[57] K. Lewis, J. Kaufman, and N. A. Christakis, *An analysis of college student privacy settings in the Facebook.com social network*, Information, Communication, and Society, (2008). In press.

[58] K. Lewis, J. Kaufman, M. Gonzalez, M. Wimmer, and N. A. Christakis, *Tastes, ties, and time: A new (cultural, multiplex, and longitudinal) social network dataset using Facebook.com*, Social Networks, (2008). In press.

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

[60] L. A. Lievrouw and S. Livingstone, eds., *The Handbook of New Media*, Sage Publications Ltd., London, UK, updated student edition ed., 2005.

[61] A. H. Looijen and M. A. Porter, *Legends of Caltech III: Techer in the Dark*, Caltech Alumni Association, Pasadena, CA, 2007.

[62] M. J. Lubbers and 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), pp. 489–507.

[63] N. Mantel, *The detection of disease clustering and a generalized regression approach*, Cancer Research, 27 (1967), pp. 209–220.

[64] M Meila, *Comparing clusterings — an information based distance*, J. Multivariate Analysis, 98 (2007), pp. 873–895.

[65] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, *Network motifs: Simple building blocks of complex networks*, Science, 298 (2002), pp. 824–827.

[66] J. Moody and D. R. White, *Structural cohesion and embeddedness: A hierarchical concept of social groups*, American Sociological Review, 68 (2003), pp. 103–127.

[67] J. Moody and R. White, D., *The cohesiveness of blocks in social networks*, Sociological Methodology, 31 (2001), pp. 305–359.

[68] J. W. Van Ness, *A method for comparing two hierarchical clusterings: Comment*, Journal of the American Statistical Association, 78 (1983), pp. 576–579.

[69] M. E. J. Newman, *Scientific collaboration networks: I. network construction and fundamental results*, Physical Review E, 64 (2001).

[70] ———, *The structure and function of complex networks*, SIAM Review, 45 (2003), pp. 167–256.

[71] ———, *Fast algorithm for detecting community structure in networks*, Physical Review E, 69 (2004), 066133.

[72] ———, *A measure of betweeenness centrality based on random walks*, Social Networks, 27 (2005), pp. 39–54.

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

[74] M. E. J. Newman, *Modularity and community structure in networks*, Proceedings of the National Academy of Sciences, 103 (2006), pp. 8577–8582.

[75] M. E. J. Newman and M. Girvan, *Mixing patterns and community structure in networks*, in Statistical Mechanics of Complex Networks, R. Pastor-Satorras, J. Rubi, and A. Diaz-Guilera, eds., Springer-Verlag, Berlin, Germany, 2003.

[76] ———, *Finding and evaluating community structure in networks*, Physical Review E, 69 (2004), 026113.

[77] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, *Random graphs with arbitrary degree distributions and their applications*, Physical Review E, 64 (2001), 026118.

[78] A. Noack, *Modularity clustering is force-directed layout.* arXiv:0807.4052 (2008).

[79] R. Nyland and C. Near, *Jesus is my friend: Religiosity as a mediating factor in Internet social networking use*, in Paper presented at AEJMC Midwinter Conference, Reno, NV, 2007.

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

[81] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, *Uncovering the overlapping community structure of complex networks in nature and society*, Nature, 435 (2005), pp. 814–818.

[82] M. A. Porter, A. J. Friend, P. J. Mucha, and M. E. J. Newman, *Community structure in the U. S. House of Representatives*, Chaos, 16 (2006), 041106.

[83] M. A. Porter, P. J. Mucha, M. E. J. Newman, and A. J. Friend, *Community structure in the United States House of Representatives*, Physica A, 386 (2007), pp. 414–438.

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

[85] A. Pothen, H. Simon, and K.-P. Liou, *Partitioning sparse matrices with eigenvectors of graphs.*, SIAM Journal of Matrix Analysis and Applications, 11 (1990), pp. 430–452.

[86] W. M. Rand, *Objective criteria for the evaluation of clustering methods*, Journal of the American Statistical Association, 66 (1971), pp. 846–850.

[87] H. Rao, G. F. Davis, and A. Ward, *Embeddedness, social identity and mobility: Why firms leave the NASDAQ and join the New York Stock Exchange*, Administrative Science Quarterly, 45 (2000), pp. 268–292.

[88] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabasi, *Hierarchical organization of modularity in metabolic networks*, Science, 297 (2002), pp. 1551–1555.

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

[90] S. Reid, J.-P. Onnela, M. A. Porter, N. S. Jones, and P. J. Mucha, *Comparing the community structures of complex networks*. In preparation (2008).

[91] Stephanie Rosenbloom, *On Facebook, scholars link up with data*. December 17, 2007.

[92] M. Sales-Pardo, R. Guimerà, A. A. Moreira, and Luís A. N. Amaral, *Extracting the hierarchical organization of complex systems*, Proceedings of the National Academy of Sciences, 104 (2007), pp. 15224–15229.

[93] V. Sodera, *Rapleaf study reveals gender and age data of social network users*, (2008). Press release, available at `http://business.rapleaf.com/company_press_2008_07_29.html`.

[94] E. Spertus, M. Sahami, and 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.

[95] S. H. Strogatz, *Exploring complex networks*, Nature, 410 (2001), pp. 268–276.

[96] F. Stutzman, *Adopting the Facebook: A comparative analysis*, (2006). Ph.D. research report, available at `http://www.ibiblio.org/fred/pubs/stutzman_wp5.pdf`.

[97] ———, *An evaluation of identity-sharing behavior in social network communities*, International Digital and Media Arts Journal, 3 (2006).

[98] ———, *The vibrancy of online social spac*, in Mobilizing Generation 2.0: A Practical Guide to Using Web 2.0 Technologies to Recruit, Engage & Activate Youth, Ben Rigby, ed., Jossey-Bass, New York, NY, 2008.

[99] M. Useem, *The Inner Circle: Large corporations and the rise of business political activity in the US and UK*, Oxford University Press, Oxford, 1984.

[100] D. L. Wallace, *A method for comparing two hierarchical clusterings: Comment*, Journal of the American Statistical Association, 78 (1983), pp. 569–576.

[101] Stanley Wasserman and Katherine Faust, *Social Network Analysis: Methods and Applications*, Structural Analysis in the Social Sciences, Cambridge University Press, Cambridge, UK, 1994.

[102] S. Wasserman and P. Pattison, *Logit models and logistic regressions for social networks. i: An introduction to markov graphs and p^**, Psychometrika, 61 (1996), pp. 401–425.

[103] D. J. Watts, *The "new" science of networks*, Annual Review of Sociology, 30 (2004), pp. 243–270.

[104] A. S. Waugh, L. Pei, J. H. Fowler, M. A. Porter, and P. J. Mucha, *Party polarization in congress: A social network approach*. In preparation (2008).

[105] L.-S. Wu, R. Akavipat, and F. Menczer, *6S: P2P Web index collecting and sharing application*. Available at `http://riao.free.fr/applications/6S\%20P2P\%20Web.pdf`, 2007.

[106] W. W. Zachary, *An information flow model for conflict and fission in small groups*, Journal of Anthropological Research, 33 (1977), pp. 452–473.

[107] A. S. Zagier, *Graduates tap online alumni networks for job leads, lost friends.* November 27, 2006.
[108] Y. Zhang, A. J. Friend, L. Traud, A., M. A. Porter, J. H. Fowler, and P. J. Mucha, *Community structure in Congressional cosponsorship networks*, Physica A, 387 (2008), pp. 1705–1712.
[109] R. Zheng, F. Provost, and A. Ghose, *Social network collaborative filtering.* Preprint (CeDED working paper), 2007.

Appendix A. Statistics of Pair-Counting Indices. The starting point for pair-counting comparisons between partitions is with contingency table entries n_{ij}, which denote the number of nodes assigned to group i in the first partition and to group j in the second partition. Contingency table entries can be used to calculate the number of node pairs that are assigned to the same group in both partitions (w_{11}), to different groups in both partitions (w_{00}), to the same groups in the first but different ones in the second (w_{10}), and to different groups in the first partition but the same one in the second (w_{01}). One can then specify pair-counting indices in terms of these quantities. For example, the Rand index is given by $S_{\rm R} = (w_{11} + w_{00})/M$, where $M = w_{11} + w_{10} + w_{01} + w_{00} = \binom{n}{2}$ is the total number of pairs.

In the present paper, we have emphasized the utility of interpreting pair-counting similarity coefficients in terms of their values relative to an ensemble of random partitions (e.g., as might be obtained by permuting the original assignments). As long as the randomly-selected partitions are constrained to have the same numbers and sizes of groups as the original partitions—i.e., as long as the row and column sums, $n_{i\cdot} = \sum_j n_{ij}$ and $n_{\cdot j} = \sum_i n_{ij}$, and total number of elements $n = \sum_{ij} n_{ij}$ remain constant—then the total number of pairs $M = \binom{n}{2}$, the number of pairs classified the same way in the first partition, $M_1 = \sum_i \binom{n_{i\cdot}}{2}$, and the analogous quantity for the second partition, $M_2 = \sum_i \binom{n_{\cdot j}}{2}$, likewise remain constant. These constraints then imply, regardless of the detailed method of random generation, that any pair-counting index specified as a function of the $w_{\alpha\beta}$ counts can be equivalently specified by the single variable $w = w_{11} = \sum_{ij} \binom{n_{ij}}{2}$ because $w_{10} = M_1 - w$, $w_{01} = M_2 - w$, and $w_{00} = M - M_1 - M_2 + w$. It follows immediately that the Fowlkes-Mallows, Γ, Rand, and Adjusted Rand coefficients are each linear functions of w:

$$
\begin{aligned}
S_{\rm FM} &= w/\sqrt{M_1 M_2}\,, \\
S_{\Gamma} &= (Mw - M_1 M_2)/\sqrt{M_1 M_2 (M - M_1)(M - M_2)}\,, \\
S_{\rm R} &= \frac{1}{M}(w + M - M_1 - M_2)\,, \\
S_{\rm AR} &= \left(a - \frac{M_1 M_2}{M}\right) \Big/ \left(\frac{1}{2}(M_1 + M_2) - \frac{M_1 M_2}{M}\right)\,.
\end{aligned}
$$

These seemingly different indices are hence also each linear functions of each other [48]. Other indices (not used in the present paper) that can also be expressed as linear functions of w include Wallace's two asymmetric indices [100], the Γ coefficient in [46], and the agreement measure suggested in [11].

Numerous studies have attempted to assess the utility of various similarity measures in their raw forms, but we find it most useful to return to a classical statistical approach, advocated in [11, 29] (and presumably also by others), wherein such measures are used in the context of testing significance levels of the obtained values versus those expected at random. One then needs to select a specific null model (and deal with the limitations inherent therein [68, 100]), regardless of whether one wishes to quantify the randomness analytically or computationally. If the null model is an inde-

pendent random model, then the resulting significance-level test to reject the hypothesis of independence must be interpreted correctly: A result that is less likely under the independent hypothesis need not be more likely in the alternative scenario [39]. In the present context, this implies that a stronger rejection of independence does not allow one to conclude that two partitions are "closer" in all senses. Indeed, we would recommend using a proper distance metric such as variation of information (VI) [64] for comparing partitions that are close to one another. In contrast, in most of our Facebook examples the mutual information of a pair of partitions is small compared to the total information in each. In such cases, two partitions can be relatively far from each other according to a distance measure but might nevertheless be very far in the tail of the distribution of what can be expected at random.

Any similarity index S_i that is a linear function of the single variable w must be statistically equivalent in any null model that generates ensembles of partition pairs in which M, M_1, and M_2 are constrained to remain constant. Specifically, because the domains of each of the S_i are linear transformations of the domain of w (and of each other), the z-score and p-value associated with a specified w (for given M, M_1, and M_2) must be the same for every similarity index that depends linearly on w. In deference to the seminal presentation of the Rand index [86], we refer to the resulting z-score as z_{R}, although it is equivalent by linearity to the z-score advocated explicitly by Brennan and Light [11], with $z_{\mathrm{R}} = (w - \mu_w)/\sigma_w$, where μ_w and σ_w are the mean and standard deviation, respectively, of $w = w_{11} = \sum_{ij} \binom{n_{ij}}{2}$. In the absence of another compelling null model, we adopt the fully random hypergeometric distribution with fixed row and column sum marginals. The expected value then becomes $\mu_w = M_1 M_2 / M$, as for the adjusted Rand index [47]. The calculation of higher-order moments is more involved [11, 12, 46, 63].

In order to make the Rand index z-score z_{R} as simple as possible to calculate, we concisely present the formulas of [46] in a slightly simplified (to our eyes) form using the present notation:

$$z_{\mathrm{R}} = \frac{1}{\sigma_w} \left(w - \frac{M_1 M_2}{M} \right) , \tag{A.1}$$

where

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

with C_1 and C_2 obtained from third powers of the row and column sum marginals,

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

and M, M_1, M_2, n, $n_{i\cdot}$, and $n_{\cdot j}$ are as defined above.

While we advocate the use of z_{R}, we caution that the significance levels (equivalently, the p-values of the cumulative distribution) associated with them are not equal to those for a Gaussian distribution. The distribution for large samples is asymptotically Gaussian [53], but the distribution associated with comparing a particular

pair of partitions need not be so. Indeed, as shown in [12], the tails of the distribution can be quite heavy, so that the probability of obtaining extreme z-scores can be orders-of-magnitude higher than that given by the normal distribution. Nevertheless, the Gaussian approximation is reasonable up to at least two standard deviations (i.e., past the 95% confidence interval) for all but the most extreme cases (see, e.g., Fig. 4.1). Given the straightforward calculation of (A.1)–(A.3), we prefer to use z_{R} directly, with the caveat that the Rand indices do not translate directly to p-values.

Even without ever calculating the p-values themselves, it is instructive to note the similarity of the linear-in-w similarity coefficients to the Jaccard and Minkowski indices, which are not linear in w:

$$S_{\mathrm{M}}^2 = \frac{M_1 + M_2 - 2w}{M_1}\,, \qquad \frac{1}{S_{\mathrm{J}}} = \frac{M_1 + M_2}{w} - 1\,.$$

The asymmetry in the Minkowski index is clearly limited; switching which partition is the reference changes the coefficient by a multiplicative factor. Finally, because the square root and multiplicative inverse are both monotonic operations in the domains of these indices ($S_{\mathrm{M}} > 0$, $0 \leq S_{\mathrm{J}} \leq 1$), it follows that the p-values of the cumulative distributions of each are identical to the p-value of w itself even though the z-scores can be different from z_{R}. Therefore, every one of the pair-counting indices considered in the present paper are actually identical to each other for the purpose of testing the significance levels of the (null) independent hypothesis. Consequently, even though we do not ever directly calculate the corresponding p-values, the associated z_{R}-scores of the linear-in-w indices provide an easy measure with which to order the strength of rejection of the null hypothesis.