Community Detection in Networks with Node Features

Yuan Zhang, Elizaveta Levina, and Ji Zhu
Department of Statistics
University of Michigan
Ann Arbor, MI 48109
{yzhanghf,elevina,jizhu}@umich.edu

Abstract

Many methods have been proposed for community detection in networks, but most of them do not take into account additional information on the nodes that is often available in practice. In this paper, we propose a new joint community detection criterion that use both the network and the features to detect community structure. One advantage our method has over existing joint detection approaches is the flexibility of learning the impact of different features which may differ across communities. Another advantage is the flexibility of choosing the amount of influence the feature information has on communities. The method is asymptotically consistent under the block model with additional assumptions on the feature distributions, and performs well on simulated and real networks.

1 Introduction

Community detection is a fundamental problem in network analysis, which has been extensively studied in a number of domains – see [18, 4, 20] for some examples of applications. A number of approaches to community detection are based on probabilistic models for networks with communities, such as the stochastic block model [10], the degree-corrected stochastic block model [12], and the latent factor model [9]. Other approaches work by optimizing a criterion measuring the strength of community structure in some sense and spectral approximations to such criteria. Examples include normalized cuts [21], modularity [17, 16], and many variants of spectral clustering, for example [19].

Many of the existing community detection methods focus on analyzing the network based on its adjacency matrix only. However, we often have access to additional information on the nodes, which we will refer to as node features, and sometimes edges as well, e.g., [24, 23, 11]. In many networks the distribution of node features is correlated with community structure [15], and thus a natural question is whether we can improve community detection by using the node features. Several approaches have been proposed that assume generative models for the network and its features, including the network random effect model [8], the embedding feature model [27], the latent variable model [7], the discriminative approach [26], the latent multigroup membership graph model [14], and the social circles model for ego networks [15]. Most of these approaches were designed to fit specific feature types, and their effectiveness depends heavily on the correctness of model specification. Approaches that are not model-based include edge weighing by node feature similarity [25], attribute-structure mining [22] and SA-clustering [2]. Most methods in this category use all features the same way without considering which ones influence the community structure, and lack the flexibility of balancing the information coming from network adjacency matrix and its node features. Including irrelevant node features may potentially lead to worse community detection, while selecting features that cluster strongly, which is in itself a difficult problem in clustering, may not correspond to features that correlate with community structure.

In this paper, we propose a new joint community detection criterion that combines network edge information and node features. The idea is that by properly weighing edges according to feature similarities on their terminal nodes, the community structure in the network is enhanced and thus the detection is improved. We learn which features are most helpful in identifying community structure from data, and allow for the possibility that having similar features may make nodes more or less likely to connect, thus allowing for both assortative- and disassortative-type behavior for each individual feature. On an intuitive level, our method looks for an agreement between clustering structures suggested by the two data sources, the adjacency matrix and the node features. Once the community structure is estimated, we can formally select relevant node features via a permutation test, or we can alternatively achieve feature selection by using an ℓ_1 penalty on the feature coefficients. Numerical experiments on simulated and real examples show that our method performs well compared to methods that use either the network alone or the features alone for clustering.

2 The joint community detection criterion

Our method is based on the intuition that communities are characterized by having more edges within themselves than between. While this is certainly not the only possible type of community structure, it is a very common one, and this intuition underlies many other methods for community detection, e.g., modularity [16]. Our goal is to use such a community detection criterion based on the adjacency matrix alone, and then weigh edges according to their feature similarities to improve detection. While many such criteria have been proposed, having a simple criterion linear in the adjacency matrix makes optimization much more feasible in our particular situation, as will become clear below. Let A denote the adjacency matrix with $A_{ij} = 1$ if there is an edge between nodes i and j and $A_{ij} = 0$ otherwise, or else an edge weight (our methods work the same way for weighted and unweighted networks). Let $\boldsymbol{f}_i$ denote the p-dimensional feature vector of node i. Let K be the number of communities we are looking for, and let $\boldsymbol{e}$ be a vector of label assignment, with $e_i = k$, $k = 1, \ldots, K$, if node i belongs to community k. Let $E_k = \{i : e_i = k\}$, and let $|E_k|$ be the number of nodes in community k. As a starting point for community detection, we use a very simple analogue of modularity,

$$R(\boldsymbol{e}) = \sum_{k=1}^{K} \frac{1}{|E_k|^{\alpha}} \sum_{i,j \in E_k} A_{ij} \,, \tag{1}$$

which is then maximized over all possible assignments $\boldsymbol{e}$. Here $\alpha > 0$ is a tuning parameter, and rescaling by $|E_k|^{\alpha}$ is designed to rule out trivial solutions that put all nodes in the same community. This criterion can be shown to be consistent under the stochastic block model by checking the conditions of the general theorem in [1]. Note also that when $\alpha = 2$, the criterion is approximately the sum of edge densities within communities, and when $\alpha = 1$, the criterion is the sum of average "within community" degrees, which both intuitively represent community structure.

Next, we introduce feature-based edge weights which, ideally, should upweigh edges within communities and downweigh edges between them, thus enhancing the community structure in the observed network. However, node features may not be perfectly correlated with community structure, different communities may be driven by different features, as pointed out by [15], and features themselves may be noisy. Thus we need to learn the impact of different features on communities as well as balance the role of network information and node features in community detection. This leads us to propose a **joint community detection criterion (JCDC)**, defined by

$$\max_{\boldsymbol{e},\boldsymbol{\beta}} R(\boldsymbol{e};\boldsymbol{\beta}) := \sum_{k=1}^{K} \frac{1}{|E_k|^{\alpha}} \sum_{i,j \in E_k} A_{ij} w(\boldsymbol{f}_i, \boldsymbol{f}_j; \boldsymbol{\beta}_k) \tag{2}$$

where α is a tuning parameter chosen by the user, and $\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K\}$ are the unknown vector of coefficients that controls the impact of each feature for communities $1, \ldots, K$. Note that each community has its own vector of coefficients, which allows for features playing different roles in different communities.

For the sake of simplicity, we model the edge weight w_{ij} as a function of the node features $\boldsymbol{f}_i$ and $\boldsymbol{f}_j$ via a p-dimensional vector of their similarity measures $\boldsymbol{\phi}_{ij} = \boldsymbol{\phi}(\boldsymbol{f}_i, \boldsymbol{f}_j)$, setting

$$w(\boldsymbol{f}_i, \boldsymbol{f}_j; \boldsymbol{\beta}_k) = w(\boldsymbol{\phi}_{ij}; \boldsymbol{\beta}_k) \tag{3}$$

The choice of similarity measures in $\boldsymbol{\phi}$ depends on the type of $\boldsymbol{f}_i$ (for example, on whether the features are numerical or categorical) and is determined on a case by case basis; the only important property of similarity is that it assigns higher values to features that are more similar. To eliminate potential differences in units and scales, we standardize all $\boldsymbol{\phi}_{ij}$ along each feature dimension.

Finally, we choose a functional form for w. This function should be increasing in "overall similarity" between nodes, and choosing a concave function facilitates optimization. In this paper, we use the exponential function

$$w(\boldsymbol{\phi}_{ij}, \boldsymbol{\beta}) := w_{\max} - e^{-\langle \boldsymbol{\phi}_{ij}, \boldsymbol{\beta} \rangle} \tag{4}$$

One can use other functions of similar shapes, for example, the logit exponential function, which we found empirically to perform similarly. Note that (4) depends on an additional tuning parameter $w_{\max} > 1$, whose role is to balance the roles of A and $F := \{\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n\}$ in community detection; this will be discussed in detail in Sections 3.1 and 4.2.

3 Optimizing the joint community detection criterion

The joint community detection criterion needs to be optimized over both the community assignment $\boldsymbol{e}$ and the parameters $\boldsymbol{\beta}$, which determine the edge weights w_{ij}. Using block coordinate descent, we optimize the joint criterion by fixing one variable and optimizing over the other one, and iterating until convergence.

3.1 Optimizing over label assignments with fixed weights

In this step we treat all edge weights w_{ij}'s as fixed constants. It is infeasible to search over all n^K possible label assignments, and, like many other community detection methods, we rely on a greedy label switching algorithm to optimize over $\boldsymbol{e}$, specifically, the tabu search [6], which updates the label of one node at a time. When the target community sizes are large, our method allows for a simple local update which does not require recalculating the entire criterion. For a node i, the condition for the algorithm to prefer to assign it to community k rather than l is,

$$\frac{S_{kk} + 2S_{i \leftrightarrow k}}{|E_k + 1|^\alpha} + \frac{S_{ll}}{|E_l|^\alpha} > \frac{S_{kk}}{|E_k|^\alpha} + \frac{S_{ll} + 2S_{i \leftrightarrow l}}{|E_l + 1|^\alpha} \,, \tag{5}$$

where S_{kk} and S_{ll} denote twice the total edge weights in communities k and l, respectively, and $S_{i \leftrightarrow k}$ and $S_{i \leftrightarrow l}$ denote the sum of edge weights between node i and all other nodes in E_k and E_l, respectively. When $|E_k|$ and $|E_l|$ are large, (5) is approximately equivalent to

$$\frac{S_{i \leftrightarrow k}}{|E_k|} \cdot \frac{|E_k|^{1-\alpha}}{|E_l|^{1-\alpha}} > \frac{S_{i \leftrightarrow l}}{|E_l|} \,. \tag{6}$$

The simplified condition (6) allows for a "local" update for the label of i without calculating the entire criterion. It also illustrates the impact of the tuning parameter α: when $\alpha = 1$, both sides of (6) can be viewed as averaged weights of all edges connecting node i to communities E_k and E_l, respectively. Then our method assigns node i to the community with which it has the strongest connection. When $\alpha \neq 1$, the LHS is multiplied by a factor $(|E_k|/|E_l|)^{1-\alpha}$. Suppose E_k is larger than E_l; then choosing $0 < \alpha < 1$ increases the preference for assigning a node to the larger community, while $\alpha > 1$ favors smaller communities.

Note that all the S's in (6) are functions of edge weights and thus depend on the tuning parameter $w_{\max}$. When $\boldsymbol{\beta} = \mathbf{0}$, all weights are equal to the constant $w_{\max} - 1$. On the other hand, for all values of $\boldsymbol{\beta}$ the sharp upper bound on any edge weight is $w_{\max}$. Therefore, the ratio $r_w = w_{\max}/(w_{\max} - 1)$ is the maximum amount by which our method can reweigh an edge. When $w_{\max}$ is large, $r_w \approx 1$, and thus the information from the network structure dominates. When $w_{\max}$ is close to 1, r_w is large and the feature-driven edge weights have a large impact. However, even if $w_{\max}$ is close to 1, the features will not necessarily dominate the network information, especially in a sparse graph, since we only consider and reweigh observed edges.

While the tuning parameter $w_{\max}$ controls the amount of influence features can have on community detection, it does not affect the estimated parameters $\boldsymbol{\beta}$ for a fixed community assignment. This is

easy to see from rearranging terms in (2):

$$R(\boldsymbol{e}, \boldsymbol{\beta}) = \sum_{k=1}^{K} \frac{1}{|E_k|^{\alpha}} \sum_{(i,j)\in E_k} A_{ij} w_{\max} + \sum_{k=1}^{K} \frac{1}{|E_k|^{\alpha}} \sum_{(i,j)\in E_k} A_{ij} g(\boldsymbol{f}_i, \boldsymbol{f}_j; \boldsymbol{\beta}_k) \tag{7}$$

where $g(\boldsymbol{f}_i, \boldsymbol{f}_j; \boldsymbol{\beta}_k) := w(\boldsymbol{f}_i, \boldsymbol{f}_j; \boldsymbol{\beta}_k) - w_{\max}$. The tuning parameter $w_{\max}$ is only involved in the term that does not depend on $\boldsymbol{\beta}$.

Asymptotically, if the feature weights satisfy some conditions, the optimal community assignment is consistent under the stochastic block model. More precisely, let $\boldsymbol{c}$ be the true labels, and let $\mathbb{P}(A_{ij} = 1 | c_i = k \text{ and } c_j = l) = \rho_n P_{kl}$, where $\rho_n := \mathbb{P}(A_{ij} = 1) \to 0$ is the global edge probability scaling parameter. Let $\tilde{P}_{kl} := P_{kl}\mathbb{E}[w_{ij}|(i,j) \in (\boldsymbol{c}_k, \boldsymbol{c}_l)]$. If for any $a \neq b$, we have $\alpha(\tilde{P}_{aa} + \tilde{P}_{bb}) > 2\tilde{P}_{ab}$ for $\alpha \geq 1$ and $(2^{\alpha} - 1)\min(\tilde{P}_{aa}, \tilde{P}_{bb}) > \tilde{P}_{ab}$, then $\mathbb{P}(\hat{\boldsymbol{e}} = \boldsymbol{c}) \to 1$ if $n\rho_n / \log n \to +\infty$ and $\|\hat{\boldsymbol{e}} - \boldsymbol{c}\| \xrightarrow{p} 0$ if $n\rho_n \to +\infty$. The proof follows the reasoning of [1] and [28] and is omitted here due to length constraints.

3.2 Optimizing over weights with fixed label assignments

Since we chose a concave edge weight function (4), for a given community assignment $\boldsymbol{e}$ the joint criteiron is a concave function of $\boldsymbol{\beta}_k$, and it is straightforward to optimize over $\boldsymbol{\beta}_k$ by gradient ascent. The role of $\boldsymbol{\beta}_k$ is to control the impact of different features on each community. One can show by a Taylor-series type expansion around the maximum (details omitted) and also observe empirically that for our method, the estimated $\hat{\boldsymbol{\beta}}_k$ is correlated with the feature similarity levels between nodes in community k. Specifically, we found that in practice the estimated $\hat{\boldsymbol{\beta}}_k^{(\ell)}$ is an increasing function of the sample mean of the corresponding similarity $\hat{\mathbb{E}}[\phi_{ij}^{(\ell)}|(i,j) \in E_k]$. In other words, our method produces large estimated $\hat{\boldsymbol{\beta}}_k^{(\ell)}$'s for a feature d if it has high similarity values $\boldsymbol{\phi}_{ij}^{(\ell)}$'s for $i, j \in E_k$. However, in the extreme case, the optimal $\hat{\boldsymbol{\beta}}_k^{(\ell)}$ can be $+\infty$ if all $\boldsymbol{\phi}_{ij}^{(\ell)}$'s are positive in community k or $-\infty$ if all $\boldsymbol{\phi}_{ij}^{(\ell)}$'s are negative. To avoid extreme solutions like this, we subtract a penalty term $\lambda\|\boldsymbol{\beta}\|$ from JCDC while optimizing over $\boldsymbol{\beta}$. We use a small value of λ that serves as a safeguard against extreme solutions and does not much affect moderate values of estimated $\boldsymbol{\beta}_k^{(\ell)}$.

In order to formally assess the significance of each feature for a particular community, we can perform a permutation test once we have estimated $\boldsymbol{\beta}_k$ and the label assignments $\boldsymbol{e}$. For each feature $\ell \in \{1, \ldots, p\}$, we generate a random permutation $(n_1, \ldots, n_N)$ of $(1, \ldots, N)$. Then we permute the dth dimension of the node similarity measure $\boldsymbol{\phi}_{ij}^{(\ell)}$ into $\boldsymbol{\phi}_{n_i n_j}^{(\ell)}$ and optimize the joint community detection criterion over $\boldsymbol{\beta}^{(\ell)}$ while fixing all $\boldsymbol{\phi}_{ij}^{(\ell')}$ for $\ell' \neq \ell$. Repeating this many times gives us the null distribution for $\boldsymbol{\beta}^{(\ell)}$ under the hypothesis that the ℓ-th feature has no impact on communities (since permuting the similarity values at random destroys any such relationship even if there was one). This null distribution can be then used to compute a p-value for the estimated $\boldsymbol{\beta}_k^{(\ell)}$.

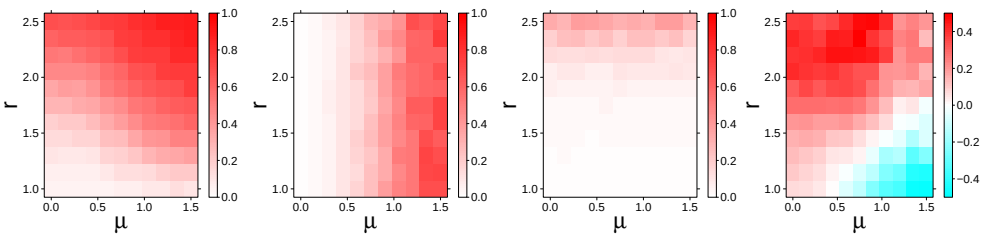
4 Simulation studies

In this section, we compare the performance of the JCDC method with methods that use network information or node features only on simulated data. We also investigate the impact of the tuning parameters α and $w_{\max}$ on the results of the JCDC method.

In all cases below, we generate networks with $n = 100$ nodes and $K = 2$ non-overlapping communities. The edges are generated independently as $A_{ij} \sim \text{Bernoulli}\left(\frac{2dr}{1+r}\right)$ if nodes i and j are in the same community and $\text{Bernoulli}\left(\frac{2d}{1+r}\right)$ if nodes i and j are in different communities. Here d is a parameter that controls the expected overall edge density of the network, and r denotes the edge probability ratio for within and between communities. For each node i, the features are generated from a bivariate normal distribution, with $\boldsymbol{f}_i \sim N((\mu, 0)^T, \sigma^2 I)$ if node i is in community 1 and $\boldsymbol{f}_i \sim N((-\mu, 0)^T, \sigma^2 I)$ if node i is in community 2.

4.1 Comparison to methods that use network structure or node features only

For this experiment, each community consists of 50 nodes, and $d = 0.1$. We vary the values of μ and r and compare the JCDC method with Newman-Girvan modularity, which only uses network information, and the K-means algorithm, which only uses node features. Agreement between the estimated communities and the true community labels is measured using normalized mutual information (NMI), a quantity commonly used in the network literature, with higher values indicating better agreement.



(a) JCDC (b) K-means (c) Modularity (d) Difference in NMI

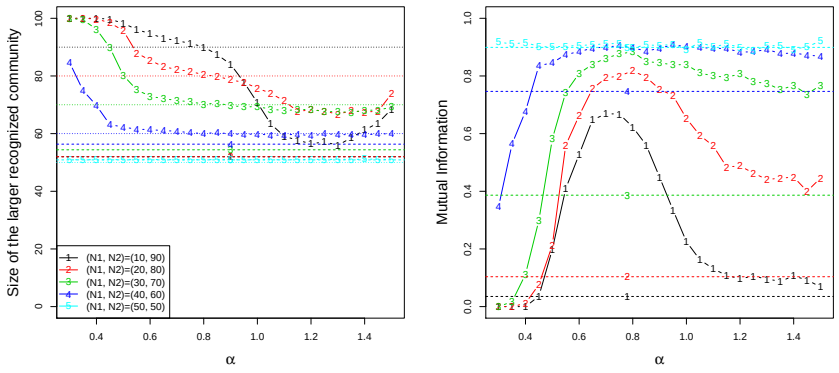
Figure 1: (a),(b),(c): Performance of different methods measured by NMI as a function of μ and r. (d) Difference in NMI between JCDC and the better of modularity and K-means.

The performance for all methods is shown in heatmaps in Figure 1 (a)-(c). Figure 1d plots the difference in NMI between JCDC and the better of modularity and K-means for each pair of values of μ and r. As one would expect, as μ increases, the performance of K-means improves and that of modularity stays constant, while as r increases, the performance of modularity improves and that of K-means stays constant. In most cases, the JCDC method outperforms both modularity and K-means, except when r is small and μ is relatively large. This is the scenario where the within community edge probability is similar to that of between community (a low signal high noise setting), while the feature distributions of the two communities are very different. The inferiority of JCDC to K-means in this scenario is understandable as JCDC always uses both the network structure and the features, and in this case only the features are informative.

4.2 The impact of tuning parameters

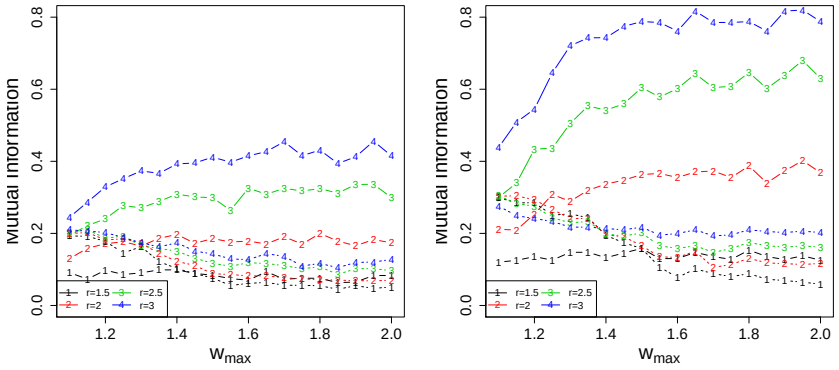
Two user-selected parameters need to be fixed ahead of time, α and $w_{\max}$. We first evaluate the impact of α on the estimated community size and detection accuracy. The values of d and r are set to 0.3 and 2, respectively. We vary the number of nodes in the smaller community from 10 to 50 and the value of α from 0.3 to 1.5. Figure 2 records the number of nodes in the estimated larger community by JCDC and the corresponding NMI. For comparison, we also record the results for the Newman-Girvan modularity, which tends to produce communities of similar size, no matter what the truth is. For JCDC, as α increases, the sizes of estimated communities become more balanced. This observation agrees with the intuition discussed in Section 3.1, i.e., a small α favors larger communities while a large α favors smaller communities when choosing where to assign a node. In terms of community detection accuracy, Figure 2b shows that the JCDC method outperforms modularity over a wide range of values of α. However, the more unbalanced the true communities are, the narrower the range of α over which JCDC achieves the best performance. This is expected since unbalanced community sizes make the problem more challenging.

Next, we investigate the impact of $w_{\max}$, which controls the amount of influence features can have on community detection. To serve the goal, we generate the edges and node features based on two different community structures. Specifically, we consider two community assignments, $\boldsymbol{c}^A$ and $\boldsymbol{c}^F$. In $\boldsymbol{c}^A$, we set $\boldsymbol{c}^A_i = 1$ for $i = 1, \ldots, 30$ and $\boldsymbol{c}^A_i = 2$ for $i = 31, \ldots, 100$, while in $\boldsymbol{c}^F$, we set $\boldsymbol{c}^F_i = 1$ for $i = 1, \ldots, 70$ and $\boldsymbol{c}^F_i = 2$ for $i = 71, \ldots, 100$. Then the edges are generated based on $\boldsymbol{c}^A$ with the node features generated based on $\boldsymbol{c}^F$. The values of μ and α are set to 1.5 and 1, respectively. We vary the values of $w_{\max}$ and r and inspect the agreement between the estimated communities $\hat{\boldsymbol{e}}$ and $\boldsymbol{c}_A$ and $\boldsymbol{c}_F$, respectively. The results are shown in Figure 3.



(a) Size of the larger estimated community (b) Detection accuracy (NMI)

Figure 2: Solid lines correspond to JCDC, horizontal dashed lines to modularity; horizontal dotted lines in (a) show the true size of the larger community. Figures (a) and (b) share the same legend.



(a) $d = 0.08$ (b) $d = 0.15$

Figure 3: Solid lines correspond to NMI between $\hat{\boldsymbol{e}}$ and $\boldsymbol{c}_A$, dashed lines to NMI between $\hat{\boldsymbol{e}}$ and $\boldsymbol{c}_F$.

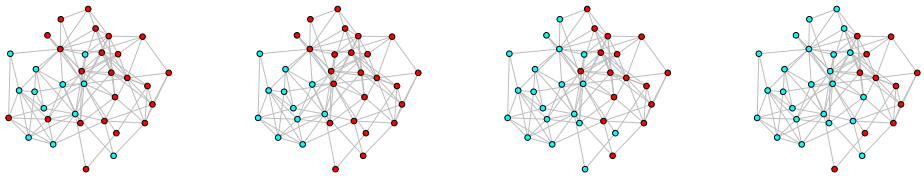
As we can see, in general, when $w_{\max}$ is small and r is not very large, the estimated community structure agrees more with $\boldsymbol{c}^F$ than with $\boldsymbol{c}^A$. When $w_{\max}$ increases, the estimated $\hat{\boldsymbol{e}}$ becomes more similar to $\boldsymbol{c}^A$. This again agrees with the discussion in Section 3.1, i.e., the first term in (7) dominates the second one when $w_{\max}$ is large enough. Further, as r increases, the community structure in the network becomes more prominent, thus $\hat{\boldsymbol{e}}$ becomes more similar to $\boldsymbol{c}^A$. It is also interesting to note that the rate at which the NMI between $\hat{\boldsymbol{e}}$ and $\boldsymbol{c}^F$ decreases is lower than that at which the NMI between $\hat{\boldsymbol{e}}$ and $\boldsymbol{c}^A$ increases. This suggests that in practice, one may consider to use a relatively large value of $w_{\max}$.

5 Data applications

5.1 The Mexican political elite network

The Mexican political elite network [5, 3] consists of 35 Mexican politicians including presidents and their close associates. The edge between two politicians indicates a significant tie of any type between them (political, business, friendship, etc.). There is one available continuous node feature, the year in which the person first assumed significant power in the government. We also know the politician's backgrounds – military or civilian – which we can compare to detected communities. In early years of the time period under consideration, the military dominated the government, and

over the years civilian politicians gradually took over. The background partition and the community detection results by different methods – by our method (JCDC), modularity on the network alone, and K-means on the year – are shown in Figure 4.



(a) Political background (b) JCDC (c) Modularity (d) K-means

Figure 4: The background partition (red: military, blue:civilian) and community detection results by different methods.

The absorption of civilian politicians into the government was a gradual process and there is no single cut-off threshold on year that separates the backgrounds of politicians well. This explains why K-means does not agree well with political background. On the other hand, politicians stepping into power in years close to each other tend to have similar backgrounds, which is accounted for by higher edge weights in our method. As measured by normalized mutual information (NMI) between estimated communities and the background partition, our method (NMI=0.37) outperforms both modularity (NMI=0.20) and K-means (NMI=0.26). The same holds for the Jaccard Index(JI): our method achieves JI=0.85, modularity JI=0.74, and K-means JI=0.66. The estimated $\boldsymbol{\beta}$ is $(1.59, 1.04)$ with permutation test based p-values of $(0.02, 0.00)$. This indicates that our method recognizes the year as a significant feature in formation of both estimated communities.

5.2 The lawyer friendship network

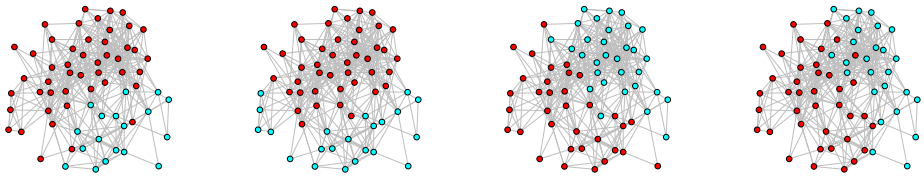
The Lawyer friendship network [13] consists of 71 lawyers in a Northeastern US corporate law firm. The edges indicate friendship ties between lawyers. There are 7 features available on each node: status (partner or associate), gender, office location (Boston or Hartford), years with the firm, age, practice (litigation or corporate) and law school attended (Harvard, Yale, University of Connecticut, or other). We eliminated 6 isolated nodes with zero degrees. Figure 5 shows heat plots of the adjacency matrix rearranged by sorting the columns and the rows of the original adjacency matrix in ascending order by each feature. Partition by office location (Figure 5(c)) shows the clearest community structure, and thus we will use partition by office location to compare to community detection on the network using all the other variables as node features. In this comparison, we omit the very small Providence office, which has only two non-isolated nodes and only two edges to other nodes. Communities estimated by different methods and the office location partition are

(a) status (b) gender (c) office (d) year (e) age (f) practice (g) school

Figure 5: Adjacency matrices aligned along each marginal feature

shown in Figure 6. Comparing the estimated communities to the office location partition, our JCDC method, modularity, and K-means achieve normalized mutual information of 0.34, 0.02, and 0.00, respectively, and the corresponding Jaccard indexes are 0.85, 0.56, and 0.54. Clearly, using the additional features improves community detection in this case.

The coefficients $\boldsymbol{\beta}_k$ and their p-values estimated in our method are given in Table 1. Our method identifies status and practice as features important in the first community and finds no significant



(a) Office location (b) JCDC (c) Modularity (d) K-means

Figure 6: The office location partition and community detection results from different methods

features in the second community. This suggests that status and practice help in detecting the first community, while the second community is determined primarily from the information on the network itself.

Table 1: Estimated $\boldsymbol{\beta}_k$ and their p-values

	status	gender	year	age	practice	school
Community 1	0.38(0.00)	0.02(0.31)	0.15(0.12)	0.18(0.10)	0.15(0.048)	0.00(0.37)
Community 2	0.05(0.10)	0.00(0.32)	0.07(0.12)	0.03(0.17)	0.00(0.31)	0.00(0.28)

6 Discussion

The JCDC method we proposed has the ability to incorporate feature information into community detection and improve results compared to using the network information alone or the feature information alone. It is designed for community structures manifesting themselves by more connections within communities, and benefits the most from features that are correlated to the community structure. It also has the ability to identify relevant features and allows for features playing different roles in different communities, which is key to good performance in realistic scenarios.

There are several aspects of our method that can be improved upon or extended in future work. One is accounting for variations in node degrees, which are usually regarded as independent of community structure, but may in some cases be correlated with features. Another direction for future work is extending our method to the case of overlapping community cases. The JCDC criterion (2) can be decomposed into a sum over separate "quality measures" on each estimated community, which can in principle be optimized in parallel to allow for overlaps. The global maximum of each quality measure will correspond to the same community, which is not a desirable solution, and thus initializing the search from distinct and possibly non-overlapping communities is crucial. These issues are a topic for future work.

Acknowledgments

E.L. is partially supported by NSF grants DMS-01106772 and DMS-1159005. J.Z. is partially supported by a NSF grant DMS-0748389 and a NIH grant R01GM096194.

References

[1] P. J. Bickel and A. Chen. A nonparametric view of network models and Newman-Girvan and other modularities. *Proc. Natl. Acad. Sci. USA*, 106:21068–21073, 2009.

[2] H Cheng, Y Zhou, and J. X. Yu. Clustering large attributed graphs: A balance between structural and attribute similarities. *ACM Trans. Knowl. Discov. Data*, 5(2):12:1–12:33, February 2011.

[3] W. de Nooy, A. Mrvar, and V. Batagelj. *Exploratory social network analysis with Pajek*, volume 27. Cambridge University Press, 2005.

[4] B. S. Dohleman. Exploratory social network analysis with pajek. *Psychometrika*, 71(3):605–606, 2006.

[5] J. Gil-Mendieta and S. Schmidt. The political network in mexico.

[6] F. Glover. Future paths for integer programming and links to artificial intelligence. *Comput. Oper. Res.*, 13(5):533–549, May 1986.

[7] M. S. Handcock, A. E. Raftery, and J. M. Tantrum. Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):301–354, 2007.

[8] P. D. Hoff. Random effects models for network data. In *Dynamic social network modeling and analysis: Workshop summary and papers*, pages 303–312. National Academies Press Washington, DC, 2003.

[9] P. D. Hoff. Modeling homophily and stochastic equivalence in symmetric relational data. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press, Cambridge, MA, 2007.

[10] P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: first steps. *Social Networks*, 5 (2):109–137, 1983.

[11] N. P. Hummon, P. Doreian, and L. C. Freeman. Analyzing the structure of the centrality-productivity literature created between 1948 and 1979. *Science Communication*, 11(4):459–480, 1990.

[12] B. Karrer and M. E. J. Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83:016107, 2011.

[13] E. Lazega. *The collegial phenomenon: the social mechanisms of co-operation among peers in a corporate law partnership*. Oxford University Press, 2001.

[14] J. Leskovec M. Kim. Latent multi-group membership graph model. *International Conference on Machine Learning*, 2012.

[15] J. McAuley and J. Leskovec. Learning to discover social circles in ego networks. In *Advances in Neural Information Processing Systems 25*, pages 548–556, 2012.

[16] M. E. J. Newman. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA*, 103 (23):8577–8582, 2006.

[17] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69(2):026113, Feb 2004.

[18] E. M. Rogers and D. L. Kincaid. *Communication networks: Toward a new paradigm for research*. Free Press New York, 1981.

[19] J. Ruan and W. Zhang. An efficient spectral algorithm for network community discovery and its applications to biological and social networks. In *Seventh IEEE International Conference*, pages 643–648, 2007.

[20] T. Schlitt and A. Brazma. Current approaches to gene regulatory network modelling. *BMC bioinformatics*, 8(Suppl 6):S9, 2007.

[21] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[22] A. Silva, W. Meira, Jr., and M. J. Zaki. Mining attribute-structure correlated patterns in large attributed graphs. *Proc. VLDB Endow.*, 5(5):466–477, 2012.

[23] Tom AB Snijders, Philippa E Pattison, Garry L Robins, and Mark S Handcock. New specifications for exponential random graph models. *Sociological methodology*, 36(1):99–153, 2006.

[24] C. Steglich, T. A. B. Snijders, and P. West. Applying siena: An illustrative analysis of the coevolution of adolescents' friendship networks, taste in music, and alcohol consumption. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 2(1):48, 2006.

[25] E. Viennet et al. Community detection based on structural and attribute similarities. In *ICDS 2012, The Sixth International Conference on Digital Society*, pages 7–12, 2012.

[26] T. Yang, R. Jin, Y. Chi, and S. Zhu. Combining link and content for community detection: a discriminative approach. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 927–936. ACM, 2009.

[27] H. Zanghi, S. Volant, and C. Ambroise. Clustering based on random graph model embedding vertex features. *Pattern Recogn. Lett.*, 31(9):830–836, July 2010.

[28] Y. Zhao, E. Levina, and J. Zhu. Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics*, 40(4):2266–2292, 2012.