Mixing patterns and individual differences in networks

George T. Cantwell and M. E. J. Newman

Department of Physics, University of Michigan, Ann Arbor, Michigan, USA
Center for the Study of Complex Systems, University of Michigan, Ann Arbor, Michigan, USA

We study mixing patterns in networks, meaning the propensity for nodes of different kinds to connect to one another. The phenomenon of assortative mixing, whereby nodes prefer to connect to others that are similar to themselves, has been widely studied, but here we go further and examine how and to what extent nodes that are otherwise similar can have different preferences. Many individuals in a friendship network, for instance, may prefer friends who are roughly the same age as themselves, but some may display a preference for older or younger friends. We introduce a network model that captures this type of variance in assortativity along with an expectation-maximization algorithm for fitting it to observed network data. The fit allows us to make best estimates of the preferences of individual nodes, define metrics to quantify individual variation in assortativity, perform sensitive community detection even in the absence of traditional assortative structure, and accurately predict missing data in unlabeled or partially labeled networks.

I. INTRODUCTION

Networks are widely used to represent patterns of connections in complex systems, such as the Internet, the World Wide Web, and social and biological networks. A common feature of many networks is assortative mixing, the tendency of network nodes to be connected to others that are similar to themselves in some way. On the World Wide Web, for instance, one might expect web pages to link to others written in the same language more than they do to ones in different languages. In friendship networks (where the phenomenon is also known as homophily) many individuals have a preference for friends who are similar to themselves in terms of age, race, educational level, and other characteristics. Occasionally, one also encounters disassortative mixing, the tendency for nodes to connect to unlike others.

Assortative mixing has been widely studied. Researchers have examined and quantified assortativity as it occurs in a wide variety of real-world networks and created mathematical models such as the planted partition model and the stochastic block model that can mimic both assortative and disassortative behaviors. These methods and models, however, capture only the average mixing behavior of nodes, the average preference for members of one group to forge connections with another. There can be, and in many cases is, substantial variation about that average; all members of a group do not necessarily behave the same. For instance, networks of romantic interaction between individuals, which are widely studied in sociology, are mostly disassortative by gender: a majority of individuals have a preference for romantic engagements with members of the opposite sex. Some prefer romantic engagements with the same sex, however, and standard measures of overall assortative mixing would then say that the average individual has a small fraction of same-sex relationships and the rest are opposite-sex. But this is misleading: in fact, many individuals have strong preferences for one or the other, so the average measure does not, in this case, provide a good description of individual behaviors.

Furthermore, there can be interesting mixing patterns even when there is little or no average assortativity in a network. For example, a recent detailed study of Facebook showed little to no gender assortativity, but on the other hand Altenburger and Ugander point out that some people do have preferences—some individuals on Facebook strongly prefer either male or female friends. It is only when we average over the whole population that we see no effect.

Thus, traditional measures of average assortativity do not tell the whole story. In this paper we explore individual variation in mixing patterns using models and empirical data for real networks and present methods that can account for and quantify behaviors at the individual level. There is some previous literature on these issues. In their study of Facebook, mentioned above, Altenburger and Ugander introduced the concept of monophily, the extent to which people’s friends are similar to one another, while Peel et al. define a variant assortativity coefficient that characterizes assortativity within a local neighborhood in a network. In our work, we introduce a generative stochastic model of individual-level mixing and show how it can be used to model and analyze empirical network data. By fitting the model to data using maximum-likelihood methods, we extract best-fit values of the model parameters which have straightforward interpretations and can thus be used to characterize mixing patterns, in much the same way that the parameters of a normal distribution characterize mean and variance. The model is conceptually similar to one introduced previously in, although different in technical details.

As a demonstration of our methods, we apply them to two example networks, a friendship network of high school students and a linguistic network of word adjacencies in English text. The results show a number of interesting features. First, we find that there is indeed substantial individual variation in mixing patterns in both networks, implying that traditional average measures of mixing offer a substantially incomplete description of network structure. Second, in cases where the characteris-
tics of some nodes are unknown, our methods can be used to predict them, and moreover do so with high accuracy in our tests, even in cases where a substantial fraction of data are missing. In fact, if we have no data at all for any of the nodes—we are given only the network—our model can be used to label all nodes, effectively performing a kind of community detection on the network [11]. In some cases this calculation is roughly equivalent to traditional community detection but in others it is not. In particular, there can be cases where a network has no traditional community structure as discussed above, meaning that on average there is neither assortative nor disassortative mixing, and standard community detection methods fail in such cases. Our method, on the other hand, can still accurately divide the nodes into communities so long as individuals display preferences.

II. INDIVIDUAL PREFERENCES AND PATTERNS OF CONNECTION

Consider a network where nodes are divided into a number of discrete, non-overlapping groups, types, or categories, and where individual nodes have preferences about the types of the nodes with which they have network connections. For the moment we will assume that the type of every node is explicitly specified. (We will relax this assumption later, but even the case of known node types shows significant interesting behaviors.) Our network could be directed or undirected, but we will concentrate primarily on the directed case here, treating the undirected one as the special case when all edges are reciprocated.

Let $k_{is}$ be the number of edges from node $i$ to nodes of type $s$ and $k_i = \sum_s k_{is}$ be the total number of edges from node $i$—its out-degree, or total degree in the case of an undirected network. Then the ratio $k_{is}/k_i$, the fraction of edges from node $i$ to nodes of type $s$, gives us a simple measure of $i$'s preferences. For example, if we ask someone to list their friends and they list 19 men and 1 woman, it seems reasonable to assume they have a preference for male friends and the ratio 19/20 = 0.95 could be used to quantify the strength of that preference.

This measure is a rather crude one, however. In particular, it gives us little information when the degree $k_i$ is small. If someone lists only one friend, who happens to be a man, it would not be correct to conclude that this person has a 100% preference for male friends, or even that they have any preference at all. If you flip a coin once and it lands heads, you would be making a mistake to assume that the coin is biased in favor of heads. After all, the coin has to land some way up when you flip it.

To quantify preferences more accurately, we need to consider how the observed value of $k_{is}$ arises in probabilistic fashion from a node’s true underlying preference. Our goal is to define a sensible model by which $k_{is}$ is generated from the true preferences, then analyze our data in the framework of that model.

So how is $k_{is}$ generated? We could imagine that node $i$ considers every other node in turn and connects to those in group $s$ with a probability $x_{is}$, which measures $i$’s preference for group $s$. Then the elements of the adjacency matrix of the network would be Bernoulli random variables with means $x_{is}$, which in standard statistical notation would be written $A_{ij} \sim \text{Bernoulli}(x_{ij})$, where $g_j$ is the group or type label of node $j$.

This, however, is unsatisfactory for two reasons. First, as is often the case, it is simpler to use a Poisson rather than Bernoulli distribution: $A_{ij} \sim \text{Poisson}(x_{ij})$. In a sparse network where $x_{is} \ll 1$ the two distributions are nearly identical, but the Poisson distribution offers significant technical advantages. Second, and more importantly, many networks have broad degree distributions that are not well captured by either the simple Bernoulli or Poisson model. This issue can be dealt with by “degree-correction” [12], which introduces two additional sets of parameters $\phi_i$ and $\theta_i$ that control the in- and out-degrees of node $i$ respectively. (In an undirected network, the two would be equal $\phi_i = \theta_i$.) Then we let

$$A_{ij} \sim \text{Poisson}\left(\frac{\theta_i \phi_j x_{ij}}{\Phi_{g_j}}\right) \quad (1)$$

where

$$\Phi_s = \sum_{i \in s} \phi_i \quad (2)$$

is the sum of all $\phi_i$ for nodes in group $s$. This definition does not completely fix the values of the parameters, since one could increase any $\theta_i$ by a constant factor and decrease the corresponding $x_{is}$ by the same factor for all $s$ and it would have no effect on $A_{ij}$ or any property of the model. To fix the parameters, therefore—to make them “identifiable” in statistics parlance—one must also apply an overall normalizing condition, which we take to be

$$\sum_s x_{is} = 1, \quad (3)$$

so that $x_{is}$ effectively measures the fractional preferences of node $i$ for nodes in each other group. Similarly, we also need to choose a normalizing condition for the $\phi_i$, which we take to be

$$\sum_i \phi_i = \sum_i \theta_i. \quad (4)$$

Note that the choice of a Poisson rather than a Bernoulli distribution in Eq. (1) implies that the network may have multiedges—there may be two or more edges running between the same pair of nodes, so that $A_{ij} > 1$. On a sparse network, however, this happens vanishingly often and multiedges can normally be neglected [12].

With the choice [11], the distribution of $k_{is}$ is straightforward to compute, since $k_{is} = \sum_{j \in s} A_{ij}$, where the sum is over nodes in group $s$. Given that the $A_{ij}$ are independent Poisson random variables and that a sum of
Poisson variables is itself Poisson, the distribution of $k_{is}$ must be Poisson with mean
\[ \sum_{j \in s} \frac{\theta_i \phi_j x_{is}}{\Phi_s} = \theta_i x_{is}, \] (5)
where we have made use of Eq. (2). Hence we can write
\[ k_{is} \sim \text{Poisson}(\theta_i x_{is}), \] (6)
and
\[ k_i = \sum_s k_{is} \sim \text{Poisson}(\theta_i). \] (7)

Thus, with the model defined in this way, $\theta_i$ is equal to the expected (out-)degree at node $i$, independent of the node’s preferences, while $x_{is}$ controls the fraction of edges that lead to nodes in group $s$. We favor this model for the intuitive interpretation of its parameters along with the mathematical simplicity of the Poisson distribution.

Given the types of the nodes, we can now write down the probability of observing any given pattern of connections at node $i$:
\[ P(A_i | x_i, g, \theta, \phi) = \prod_j P(A_{ij} | x_i, g, \theta, \phi) \]
\[ = e^{-\theta_i} \prod_j \left( \frac{\theta_i \phi_j x_{ij}}{\Phi_{gj}} \right) A_{ij}^{\theta_i} \left( \Phi_{gj} \right)^{1 - A_{ij}}, \] (8)
where $A_i$ denotes the $i$th row of the adjacency matrix and $x_i$ is the vector with elements $x_{is}$.

Alternatively, if we are given both the types and the network structure, we could use the model to infer the preferences $x_i$. One tempting approach is to use maximum-likelihood estimation. However, maximization of Eq. (8) with the constraint that $\sum s x_{is} = 1$ simply leads to the trivial estimate $\hat{x}_{is} = k_{is}/k_i$, which, as we have said, gives little information when $k_i$ is small. A better approach is to infer the entire posterior probability distribution over $x_{is}$, as we now describe.

### III. DISTRIBUTION OF PREFERENCES

If we know the types $g_i$ of all nodes and the parameters $\phi_i, \theta_i$, we can compute the posterior distribution over the preferences $x_{is}$ from Eq. (8) via Bayes’ rule:
\[ P(x_i | A_i, \alpha, g, \theta, \phi) = \frac{P(A_i | x_i, g, \theta, \phi) P(x_i | \alpha)}{P(A_i | \alpha, g, \theta, \phi)}, \] (9)
where $P(x_i | \alpha)$ is the prior probability of $x_i$ and $\alpha$ is some set of parameters that characterizes this prior. Here we make the simple and common assumption that the prior is a Dirichlet distribution, which for a network with $c$ groups takes the form
\[ P(x | \alpha) = \frac{1}{B(\alpha)} \prod_{s=1}^c x_{is}^{\alpha_s - 1}, \] (10)
where $\alpha_s > 0$ for all $s$ and $B(\alpha)$ is the multi-dimensional beta function
\[ B(\alpha) = \frac{\prod \Gamma(\alpha_s)}{\Gamma(\alpha)}, \] (11)
with $\alpha_0 = \sum_s \alpha_s$ and $\Gamma(\alpha)$ being the standard gamma function. The Dirichlet distribution is defined and normalized over the regular simplex $\sum_s x_s = 1$. For the case $c = 2$ it is equivalent to the beta distribution. The expected value of $x$ within the distribution is $\alpha / \alpha_0$, and $\alpha_0$ also controls the width of the distribution. In the limit of large $\alpha_0$ the variance tends to zero and the distribution of $x$ is tightly clustered around the mean. Conversely, as $\alpha_0$ tends to zero almost all the probability density is in the corners of the simplex, as far away as possible from the mean.

We will allow each group or type $s$ to have a different distribution of preferences and hence a different set of Dirichlet parameters $\alpha_s$, so that the prior on $x_i$ is
\[ x_i \sim \text{Dirichlet}(\alpha_{g_i}). \] (12)

This is a natural choice: one can well imagine, for instance, that the men and women within a population have different distributions of preferences for male and female friends.

With this choice, we now write
\[ P(A_i | \alpha_{g_i}, g, \theta, \phi) = \int P(A_i | x, g, \theta, \phi) P(x | \alpha_{g_i}) \, dx \]
\[ = \frac{B(\alpha_{g_i} + k_i)}{B(\alpha_{g_i})} e^{-\theta_i} \prod_j \left( \frac{\theta_j \phi_j x_{ij}}{\Phi_{gj}} \right) A_{ij}^{\theta_i} \left( \Phi_{gj} \right)^{1 - A_{ij}}, \] (13)
where $k_i$ is the vector with elements $k_{is}$. Then the likelihood of the whole network is
\[ P(A | \alpha, g, \theta, \phi) = \prod_i P(A_i | \alpha_{g_i}, g, \theta, \phi). \] (14)

Given an observed network, we can now infer point estimates of the parameters by maximizing this likelihood. For the degree parameters $\theta_i$ and $\phi_i$, the maximum likelihood estimates are simply equal to the observed degrees, $\hat{\theta}_i = k_i$ (the out-degree) and $\hat{\phi}_i = k_i^{\text{in}}$ (the in-degree), or both are equal to the overall degree in the case of an undirected network. For $\alpha_r$ we have:
\[ \hat{\alpha}_r = \argmax_{\alpha_r} \prod_i P(A_i | \alpha_{g_i}, g, \theta, \phi) \]
\[ = \argmax_{\alpha_r} \prod_i \frac{B(\alpha_r + k_i)}{B(\alpha_r)} \]
\[ = \argmax_{\alpha_r} \sum_{\alpha} \left[ \ln B(\alpha_r + k_i) - \ln B(\alpha_r) \right]. \] (15)

This maximization does not have a closed form solution, but it can be performed straightforwardly using
In the context of the current model a sensible way to quantify this, if we know the preferences, is the average in-group preference

\[ a = \frac{1}{n} \sum_{i=1}^{n} x_{ig_i}. \]  

This is the average fraction of connections that fall within groups. We can estimate this quantity using Eq. (25). Setting \( f_r(x_i) = x_{ir} \), we have \( F(x) = a \) and hence

\[ a = \sum_{r} p_r \left( \int x_{ir} P(x|\alpha_r) \, dx \right) = \sum_{r} p_r \frac{\alpha_{rr}}{\alpha_r + \phi_r}. \]  

In a perfectly assortative network all nodes connect only to their own group and \( \alpha = 1 \), while in a perfectly disassortative network \( a = 0 \). For most real-world networks we expect the value to lie between these extremes, with higher values indicating more assortativity. A natural question to ask is what kinds of values we should expect to see. What constitutes a “high” value of \( a \)? One way to answer this question is to calculate the expected value \( \bar{a} \) within an appropriate null model.
A reasonable null model in this case is the model of Eq. (1) with just one type of node:

$$A_{ij} \sim \text{Poisson} \left( \frac{\theta_i \phi_j}{\sum_k \phi_k} \right).$$

This is also equivalent to Eq. (1) with more than one type of node and $x_{rs} = \Phi_s / \sum \Phi_i$ for all nodes, and hence $\alpha_{rr}/\alpha_{r0} = \Phi_r / \sum \Phi_i$ for this model, since $\alpha_{rr}/\alpha_{r0}$ is the mean of $x_{rs}$ over all nodes in group $r$. With $\phi_i$ set to the maximum-likelihood values $\phi_i = k_{in}^{\text{null}}$ we have $\Phi_r = \sum_i k_{in}^{\text{null}} \delta_{g_i r} = K_r$, the total number of edges incoming to group $r$, and $\sum_s \Phi_s = \sum_i k_{in}^{\text{null}} = m$, the total number of edges in the whole network. Then, using Eq. (27), we have

$$\bar{a} = \frac{1}{m} \sum_r p_r K_r.$$

Then the difference between the estimated value of $a$ and its expected value within the null model is

$$a - \bar{a} = \sum_r p_r \left( \frac{\alpha_{rr}}{\alpha_{r0}} - \frac{K_r}{m} \right).$$

When this quantity is greater than zero the preferences are more assortative than we would expect by chance. When it is less than zero the preferences are less assortative (or more disassortative) than expected. If we wish, we can normalize the difference so that it takes a maximum value of 1 at perfect assortativity, and thus define an assortativity coefficient

$$R = \frac{a - \bar{a}}{1 - \bar{a}} = \frac{\sum_r p_r (\alpha_{rr}/\alpha_{r0} - K_r/m)}{\sum_r p_r (1 - K_r/m)}.$$

This quantity, however, only measures traditional assortativity. As we have said, we are particularly interested in the variation of individual preferences about group means. We can express this variation in terms of the parameters $\alpha$ by computing how far preferences fall on average from their group’s mean:

$$\sigma^2 = \frac{1}{n} \sum_i \left( x_i - \frac{\alpha_{g_i}}{\alpha_{g_i,0}} \right)^2.$$

Applying Eq. (25) again we can estimate this quantity by setting $f_r(x_i) = (x_i - \alpha_r/\alpha_{r0})^2$, which gives

$$\sigma^2 = \sum_r p_r \sigma_r^2,$$

where

$$\sigma_r^2 = \int \left( x - \frac{\alpha_r}{\alpha_{r0}} \right)^2 P(x|\alpha_r) \, dx$$

$$= \sum_s \text{Var}(x_s) = 1 - \sum_s (\alpha_{rs}/\alpha_{r0})^2 / \alpha_{r0} + 1.$$

For a given value of the mean of $x_{rs}$ (which is $\alpha_{rs}/\alpha_{r0}$ for nodes in group $r$), the maximum value of $\sigma_r^2$ is $1 - \sum_s (\alpha_{rs}/\alpha_{r0})^2$, which occurs when the weight of the Dirichlet distribution is pushed as far as possible into the corners of the parameter space, i.e., when $\alpha_{r0} \to 0$. It is convenient to normalize by this maximum value to give a normalized variance

$$V_r = \frac{\sigma_r^2}{1 - \sum_s (\alpha_{rs}/\alpha_{r0})^2} = \frac{1}{\alpha_{r0} + 1},$$

which lies between zero and one and also has the nice property of being independent of the mean. Then we can define an overall normalized variance coefficient by

$$V = \sum_r p_r V_r = \sum_r \frac{p_r}{\alpha_{r0} + 1},$$

which also lies between zero and one.

This quantity represents the normalized mean-square distance between the preferences and their group means, averaged over all groups. When $V$ is close to zero every node in every group has preferences close to the group mean. If preferences are homogeneous in this way then the network is well described by the group average mixing rates and individuals’ preferences are well described by simply stating which group they belong to. Such a finding could be interesting and useful for instance in a social network: it would tell us a lot about a population if we found that their preferences were entirely determined by, say, gender or race.

At the other extreme, when $V$ approaches one, node preferences are as far away from the group mean as possible, and nodes, even within the same group, are very unlike each other in their preferences. In this scenario mixing is poorly described by average rates, since virtually no nodes behave according to the average for their group. Exactly what constitutes a “large” value of $V$ is somewhat arbitrary, but in our studies we have found that values of about 0.15 and above seem to correspond to notable variance.

### V. EXAMPLES

We have defined measures $R$ for preference assortativity and $V$ for variance. Table I shows results for these

| Network                      | $R$  | $V$  |
|------------------------------|------|------|
| College football             | 0.60 | 0.01 |
| Karate club                  | 0.74 | 0.05 |
| Political books              | 0.68 | 0.12 |
| Political blogs              | 0.80 | 0.15 |
| High school race/ethnicity   | 0.54 | 0.16 |
| Provisional IRA affiliation  | 0.60 | 0.23 |
| Word adjacency               | −0.21| 0.35 |

TABLE I: Normalized preference assortativity $R$ and preference variance $V$ for a selection of networks with known group assignments.
FIG. 1: Friendship preferences by race/ethnicity in a US high school. We show separate results for Asian, black, Hispanic, and white students. For each race/ethnicity the upper histogram (in green) shows the observed distribution of $k_{i\text{eth}}/k_i$, the naive estimate of preference, while the lower histogram (blue) shows point estimates of preferences from the model fit (i.e., the average of the posterior distribution for each node). The blue line on the same plot is the full inferred preference distribution. The dotted red line represents the average preference in the absence of assortativity.

quantities for a selection of previously studied networks with known group assignments, listed in order of increasing variance. As the table shows, all of the networks are highly assortative by our measure, except for the word adjacency network, which is disassortative.

The normalized variances $V$ take a range of values from virtually zero up to 0.35. Recall that low normalized variance indicates a network in which the members of a group have similar preferences; high variance indicates that they have widely varying preferences. Thus, for instance, the “karate club” network, which is a social network of university students, has relatively low variance, meaning it shows traditional community structure in which the members of a community are roughly alike in their preferences. The network of high school students, on the other hand, which one might expect to be similar, shows significantly higher variance. We discuss the high school and word networks in more detail below.

A. High school friendships and ethnicity

The network denoted “High school race/ethnicity” in Table 1 is a network of self-reported friendships between students in a US high school, taken from the National Longitudinal Study of Adolescent to Adult Health (commonly known as the “Add Health” study). The node labels in this case represent the (self-identified) ethnicities of the students, which take values “Asian,” “black,” “Hispanic,” “white,” “missing,” and “other.” In our analysis we discard the “missing” and “other” categories and focus on the remaining four. The particular school we look at is chosen for its diverse racial/ethnic composition.

The value of $R = 0.54$ for this network indicates that the school is strongly assortative by race, meaning that students had more within-group friendships than would be expected by chance. However, the groups also display
differences in the inferred distributions of their preferences, which are plotted in Fig. 1. Hispanic students, for instance, show a larger range of preferences than others. Note that this doesn’t necessarily imply that Hispanic students individually have diverse friendship groups—some of them do, but others show a strong preference for having mainly Hispanic friends, or for having few.

Also shown in Fig. 1 are the distributions of the naive preference estimates \( k_{is}/k_{i} \), which look quite different from the inferred distributions. As discussed in Section II, naive estimates are an unreliable indicator of true preferences, particularly when data counts are small.

**B. Word adjacencies**

The *Brown corpus* is a widely used data set consisting of samples of written English text compiled by researchers at Brown University in the 1960s [18]. Words in the data set are labeled with their part of speech—noun, adjective, verb, etc. Working from the fiction text contained in the corpus, we created a directed word adjacency network in which nodes represent words (limited to nouns, adjectives, and verbs) and there is a directed edge from word \( i \) to word \( j \) if word \( i \) is followed by word \( j \) at any point in the text.

Figure 2 shows the inferred distributions of preferences for nouns, verbs, and adjectives to be followed by nouns. For example, since adjectives normally come before nouns in English we would expect adjectives to have a preference for being followed by nouns. And indeed this is what we see—the red curve in Fig. 2 shows that most adjectives have a high preference for being followed by nouns. Nouns, on the other hand, aren’t usually followed by other nouns, although they can be, as shown by the blue curve: the distribution takes its most likely value around a preference of zero, but is spread across the whole range and there is still a relatively large density around preference 1, which is to say that some nouns strongly prefer to be followed by other nouns. Classic examples are titles such as “Mr.” and “Mrs.,” which are almost always followed by proper nouns. Likewise, although most verbs prefer to be followed by nouns, there are a handful that have a strong preference to be followed by another verb. These are the auxiliary verbs, such as “has” and “was”, in sentences like “He was sleeping.”

**VI. RELATION TO THE STOCHASTIC BLOCK MODEL**

The model employed in this paper is similar in some respects to the degree-corrected stochastic block model [12], a model widely used for the inference of community structure in networks, but there are some crucial differences, which we examine in this section.

In the degree-corrected stochastic block model each node is assigned to one of \( c \) groups. Let \( g_i \) be the group to which node \( i \) is assigned, and let \( p_{g} \) be the probability that a node is placed in group \( r \), so that

\[
g_i \sim \text{Categorical} \left(p \right). \tag{37}\]

Next, edges are placed between node pairs in a way that depends on the group assignments of the nodes in question. To make the analogy with our model closer we will consider a directed Poisson version of the block model. Then the number of edges from node \( i \) to node \( j \) is Poisson distributed with mean \( \theta_i \phi_j \omega_{rs} \), where \( r \) and \( s \) are the groups to which \( i \) and \( j \) belong, and \( \theta_i \), \( \phi_j \), and \( \omega_{rs} \) are all parameters that we choose. Formally,

\[
A_{ij} \sim \text{Poisson} \left( \theta_i \phi_j \omega_{g_i g_j} \right). \tag{38}\]

As with the model considered in this paper, the Poisson distribution in principle allows there to be more than one edge between the same pair of nodes, but in the common case of a sparse network the chances of this actually happening are small.

Comparing Eq. (38) to Eq. (1) for our model we see that the block model can be regarded as a special case of our preference model in which all nodes in a given group have identical preferences. Note that the block model and the preference model both have the same number of mixing parameters: the block model has parameters \( \omega_{rs} \) (a \( c \times c \) matrix) and the preference model has \( c \) vectors \( \alpha_s \), of \( c \) elements each.

If we assume that nodes in the preference model are placed in groups independently at random as in the stochastic block model, then we can summarize the two models as generative processes as follows:
[Block model]

1. Define the parameters $\theta_i$, $\phi_i$, $p_r$, and $\omega_{rs}$.
2. Assign each node to one of $c$ groups according to $g_i \sim \text{Categorical}(p)$. (39)
3. Generate edges independently with $A_{ij} \sim \text{Poisson}(\theta_i \phi_j \omega_{g_i g_j})$.

Under this version of the preference model, the joint likelihood for adjacency matrix $A$ and group assignment $g$ is given by

$$P(A,g|\alpha,\theta,\phi,p) = P(A|\alpha,g,\theta,\phi)P(g|p) = P(A|\alpha,g,\theta,\phi) \prod_r p_r^{n_r}$$

where $n_r$ is the number of nodes in group $r$ and $P(A|\alpha,g,\theta,\phi)$ is given by Eq. (14).

The stochastic block model is particularly useful as a tool for inference. Often we don’t actually know the underlying community structure and instead want to infer it from the network. If we assume that a network was generated by the stochastic block model then we can fit the model to the data and infer best-fit community assignments. Using Eq. (39) we can follow the same procedure for our preference model. When we have incomplete data for the nodes, we can fit the model and use it to “fill in” the missing data. Or, if we have no data at all, we can still fit the model and find the best divisions into disjoint sets. In the next section we describe how to do this.

Note that though this model generates directed networks, it is possible to use it for undirected networks as well: when fitting an undirected network one simply treats it as a directed network in which every edge is reciprocated. Resampling from the fitted model will not produce networks that look like the original, since most edges will not be reciprocated, but in other respects this approach appears to give reasonable results.

[Preference model]

1. Define the parameters $\theta_i$, $\phi_i$, $p_r$, and $\alpha_{rs}$.
2. Assign each node to one of $c$ groups according to $g_i \sim \text{Categorical}(p)$. (39)
3. Generate preferences independently with $x_i \sim \text{Dirichlet}(\alpha_{g_i})$.
4. Generate edges independently with $A_{ij} \sim \text{Poisson}(\theta_i \phi_j x_{g_i g_j})$.

Let us denote by $g'$ the set of known group assignments, which we will now assume to be a subset of the set $g$ of all group assignments. The remaining assignments are unknown. Then the joint likelihood of the network $A$ and the known assignments, given the model parameters, is

$$P(A,g'|\alpha,\theta,\phi,p) = \sum_{g\notin g'} P(A,g|\alpha,\theta,\phi,p),$$

where the sum is over all group labels $g$ that are not in $g'$, i.e., all unknown labels. The terms in this likelihood that involve $\theta$ and $\phi$ do not depend on $g$, so the result of maximizing with respect to these parameters will be the same as before, giving best estimates $\hat{\theta}_i = k_i$ and $\hat{\phi}_i = k_i^p$. To estimate the remaining parameters $\alpha$ and $p$, one needs to differentiate inside the sum, which leads to a complicated implicit equation that is not easy to solve, even numerically. Instead, therefore, we borrow a trick from the statistics toolbox and apply Jensen’s inequality, which states that for any distribution of a positive random variable $X$ we have $\ln(\mathbb{E}[X]) \geq \mathbb{E}[\ln X]$. Applying this inequality to the log of Eq. (40), we can show that

$$\ln \sum_{g\notin g'} P(A,g|\alpha,\theta,\phi,p) \geq \sum_{g\notin g'} q(g) \ln P(A,g|\alpha,\theta,\phi,p) q(g),$$

where $q(g)$ is any probability distribution over $g$ satisfying $\sum_g q(g) = 1$. One particularly useful choice of $q(g)$ is

$$q(g) = \frac{P(A,g|\alpha,\theta,\phi,p)}{\sum_{g\notin g'} P(A,g|\alpha,\theta,\phi,p)},$$

which, it’s straightforward to show, makes the left- and right-hand sides of (41) exactly equal, and hence also maximizes the right-hand side with respect to $q(g)$. A further maximization with respect to the parameters $\alpha$ and $p$ will then give us the answer we seek. To put that another way, a double maximization of the right-hand side with respect to both $q(g)$ and the parameters will give us our answer.

This leads us to an iterative algorithm for estimating the parameters, known as an expectation–maximization (or EM) algorithm, in which we perform the double maximization by simply maximizing alternately over $q(g)$ (using Eq. (42)) and over the parameters, repeating until the numbers converge.

In detail the algorithm is as follows:

1. Set $\theta_i = k_i$, $\phi_i = k_i^p$ and make an initial guess $\alpha^{(0)}$, $p^{(0)}$ for the other parameters (for instance the uniform choice $\alpha^{(0)}_{rs} = 1$, $p_r^{(0)} = 1/c$). Set $t = 1$.
2. Set

$$q(g) = \frac{P(A,g|\alpha^{(t-1)},\theta,\phi,p^{(t-1)})}{\sum_{g\notin g'} P(A,g|\alpha^{(t-1)},\theta,\phi,p^{(t-1)})}.$$  

3. Set

$$\alpha^{(t)} = \arg\max_{\alpha,p} \sum_{g\notin g'} q(g) \ln P(A,g|\alpha^{(t-1)},\theta,\phi,p^{(t-1)}).$$

VII. MISSING DATA AND COMMUNITY DETECTION

In our calculations so far we have looked at examples where the group, type, or category of every node in a network is known a priori. In other cases, however, some or all of the group assignments may be unknown, and in such cases we can use our preference model to simultaneously estimate both the model parameters and the missing data. To do this we make use of the joint data likelihood of Eq. (39), which we maximize while summing over the values of the unknown data.
4. Increase \( t \) by 1.

5. Repeat steps 2 to 4 until convergence is achieved.

The most difficult step of the algorithm is step 3, since the sum over groups often has too many terms to be evaluated exactly. Good approximations can however be made by Monte Carlo sampling using a standard Metropolis–Hastings algorithm. Details are given in Appendix B.

Once the EM algorithm converges it gives us not only estimates for the parameters but also the values of the unknown group labels \( g \), since, from Eq. (42):

\[
q(g) = \frac{P(A, g|\alpha, \theta, \phi, p)}{P(A, g'|\alpha, \theta, \phi, p)} = P(g|A, g', \alpha, \theta, \phi, p). \tag{43}
\]

In other words, \( q(g) \) is precisely the posterior distribution over \( g \), the unknown group labels.

Even in cases where all of the group assignments are unknown, the algorithm will still return best estimates of their posterior distribution, effectively functioning as a kind of community detection algorithm.

### A. Performance

To test the ability of the EM algorithm described above to deduce missing group labels we have applied it to the two example networks from Sections V A and V B—the high school friendship network and the word adjacency network. For each of these networks we repeatedly remove between 10 and 95% of the node labels and attempt to recover them using our EM algorithm. For comparison, we also compare the performance of the calculation to that of a standard degree-corrected stochastic block model. The results of these tests are shown in Fig. [3].

For the high school network, in which preference variance is moderate but not extreme (\( V = 0.16 \)), our model successfully recovers a significant fraction of the missing data, even in the limit where almost all data are discarded. On the other hand, it does not appear to perform any better than the block model, although we do learn more about the network since it gives us an estimate of individual preferences, which the block model does not.

For the word network, where variance is high (\( V = 0.35 \)), our model also successfully recovers a significant fraction of the missing data. Moreover it consistently outperforms the block model, which assumes all nodes in a community are alike in their preferences.

At this point a natural question to ask is, under what circumstances can this method be expected to succeed? Intuitively, the model attempts to label the nodes using both the average and variation of mixing patterns. If either of these are significantly different from what they would be in a random network we should be in a good position to recover missing data. On the other hand, since real-world data are presumably not actually generated from the model, it is hard to state general conditions for success. What we can do though is reverse the question, and ask: when will this method likely fail? A straightforward way to explore this question is by fitting the model to synthetic data—networks generated by the model—and trying to recover them.

The method described above finds maximum likelihood estimates of the model parameters, and it is well known that maximum likelihood estimates are consistent, meaning that if networks are generated from the model then in the limit of large and dense networks our method will uncover the underlying truth. This, however, is no guarantee that the method will work well for small and sparse networks. To test the performance in the latter case, we repeat the data recovery experiments above using synthetic data. We generate networks from the model with 200 nodes and two groups, degree parameters \( \theta = \phi = 5 \), assortativity \( R = 0.2 \), and two values for the variance, \( V = 0.001 \) and \( V = 0.25 \). We then repeatedly remove between 10 and 95% of the node labels and attempt to recover them.

The parameter values are chosen so that the task is relatively demanding. When the assortativity is high, for example, it is easy to recover the groups even when all labels are removed. Likewise, if the mean degree is high recovery becomes trivial. As assortativity is lowered for fixed small mean degree, however, recovery becomes more difficult. For the case of the stochastic block model, Decelle et al. [19] have shown that there is a sharp transition to a regime in which recovery is hard or in some cases impossible, but the details of this transition are not important for the present example.

Figure [4] shows our results for the two levels variance we study. In the low variance condition our preference model offers no improvement over the performance of the block model, both doing moderately well. In the high variance condition, however, the preference model considerably outperforms standard block modeling. These results mirror the situation observed for the high school friendship and word adjacency networks above.

Looking closely at Fig. [4] we see that when most of the data is removed, when assortativity and variance are fairly low, and when the network is fairly small and sparse, the fitting procedure fails to recover the data, even though we know it was generated from the model. In this regime we are practically guaranteed to fail. However, when there is moderate to high variation in individual preferences performance greatly improves, because variation implies that some nodes will systematically connect to other specific kinds. In a friendship network, for example, even if assortativity is weak and the members of any group are not much like their friends on average, high variation means that within the group there will be some individuals who are far from the average and have strong preferences to be either like or unlike their friends. We can use these strong preferences to accurately determine group membership.

Perhaps the main limitation of our approach is the assumption that the preferences are drawn from a Dirichlet
FIG. 3: Performance of two different methods for inferring missing data in cross-validation tests on two real-world networks: (a) a network of high school friendship and (b) a network of word adjacencies. The two methods studied are the approach defined in this paper and a traditional community analysis using the degree-corrected stochastic block model. The horizontal axis measures the fraction of nodes in the network with known group labels and the vertical axis measures the fraction of unknown labels correctly predicted by each algorithm.

FIG. 4: Performance of the two different methods for inferring missing data in cross-validation tests on synthetic networks. The networks were generated from the model defined in this paper with 200 nodes, two groups, $\theta = \phi = 5$, and $R = 0.2$. Two levels of variance were examined: (a) $V = 0.001$ and (b) $V = 0.25$. The horizontal axis measures the fraction of nodes in the network with known group labels and the vertical axis measures the fraction of unknown labels correctly predicted by each algorithm.

distribution, which rules out multimodal distributions for example. There is no reason, however, why this assumption could not be changed. Instead of a single Dirichlet distribution, for example, one could use a mixture of Dirichlet distributions, i.e., a linear combination of the density functions of different Dirichlet distributions. This would allow us to model more complicated behaviors, at the cost of introducing more parameters. However, we leave these developments for future work.

VIII. CONCLUSIONS

In this paper we have introduced a model of individual preferences in networks and a method for fitting it to observed network data. The fitted values of the parameters of the model have simple interpretations and we use them to define coefficients that represent the average similarity or assortativity of preferences between individuals, and the variance of preferences. For fully labeled networks in which the characteristics of every node are known, maximum likelihood estimation of the parameters is straightforward (comparable in its complexity to regression methods). When networks are only partly la-
beled, or completely unlabeled, the model can be used to recover unknown node characteristics and we present an expectation–maximization algorithm for doing so. We have demonstrated the effectiveness of our methods with applications to a selection of networks, including real-world examples and synthetically generated benchmarks. Among other things, we show that our methods can work well even when average mixing patterns are weak if there is still individual variation in preferences.

Acknowledgments

The authors thank Elizabeth Bruch, Alec Kirkley, and Maria Riolo for useful conversations. This work was funded in part by the US National Science Foundation under grants DMS–1407207 and DMS–1710848.

Appendix A: Calculating the value of $\alpha$

The maximum likelihood estimate for $\alpha_r$ is given by the maximum of

$$L_r(\alpha_r) = \sum_{i \in r} \left[ \ln B(\alpha_r + k_i) - \ln B(\alpha_r) \right].$$

(See Eq. [15].) Here $\ln B(x)$ is the log of the multivariate beta function,

$$\ln B(x) = -\ln \Gamma(x_0) + \sum_s \ln \Gamma(x_s),$$

(A2)

with $x_0 = \sum_s x_s$. Both the Jacobian and Hessian of $L_r$ are straightforward to compute, so in principle one could perform the maximization using optimizers such as Newton’s method that require second derivatives. In practice, however, we find this to be unnecessary—simpler methods that use only first derivatives (i.e., only the Jacobian) appear to perform well. The Jacobian is given by

$$\frac{\partial L_r}{\partial \alpha_{rs}} = \sum_{i \in r} \left[ \psi(\alpha_{rs} + k_i) - \psi(\alpha_{r0} + k_i) ight.$$

$$- \psi(\alpha_{rs}) + \psi(\alpha_{r0})],$$

(A3)
where \( \psi(x) = \Gamma(x)/\Gamma'(x) \) is the so-called digamma function. To enforce the constraint that \( \alpha_{rs} > 0 \), we let \( \alpha_{rs} = y_{rs}^2 \) and then perform an unconstrained maximization of \( f_r(y) = L_r(\alpha_{rs} = y_{rs}^2) \) with Jacobian
\[
\frac{\partial f_r}{\partial y_s} = 2 y_s \left( \frac{\partial L_r}{\partial \alpha_{rs}} \right)_{\alpha_{rs} = y_{rs}^2}.
\] (A4)

**Appendix B: EM algorithm**

To sample group assignments from the posterior distribution \( q(g) \propto P(A, g|\alpha, \theta, \phi, p) \) we use a standard Metropolis–Hastings Monte Carlo algorithm. Let \( U = \{i : q_i = \text{unknown}\} \) be the set of nodes for which we do not know \( q_i \). First we initialize \( g \) by choosing \( g_i \sim \text{Categorical}(p) \) for each \( i \in U \). Then we carry out the following steps:

1. Pick an \( i \) uniformly at random from \( U \).
2. Propose a new group \( g_i' \) for \( i \), uniformly at random from the set of all groups other than \( i \)'s current group.
3. Accept the move with probability \( q(g')/q(g) \), otherwise reject it. If the move is accepted \( i \) is moved to group \( g_i' \); if it is rejected \( i \) remains in its current group for this Monte Carlo step.
4. Repeat from step 1.

This process continues until a suitable number of independent samples have been drawn from the distribution of group assignments.

To update the estimates for \( \alpha \) and \( p \), we need to maximize the expected value of \( \ln P(A, g|\alpha, \theta, \phi, p) \). To do this, we compute the following three averages within the Monte Carlo sample:
\[
q_{ir} = \langle \delta_{g_i,r} \rangle,
\] (B1)
\[
X_{rk} = \left\langle \sum_i \delta_{g_i,r} \delta_{k,k} \right\rangle,
\] (B2)
\[
Y_{rsk} = \left\langle \sum_i \delta_{g_i,r} \delta_{k,s} \right\rangle,
\] (B3)

where \( i \) is a node label, \( r \) and \( s \) are group labels, and \( k \) is a node out-degree, with values running from 0 to the maximum out-degree for the network. Once we have estimates for these quantities, our estimates for \( p \) and \( \alpha \) are:
\[
\hat{p}_r = \frac{1}{n} \sum_i q_{ir},
\] (B4)
\[
\hat{\alpha}_r = \arg\max \{ E_r(\alpha_r) \},
\] (B5)

where \( E_r(\alpha_r) \) represents the terms in the expected log-likelihood that depend on \( \alpha_r \):
\[
E_r(\alpha_r) = \sum_{s=1}^c \sum_{k=0}^{k_{\max}} Y_{rsk} \ln(\alpha_{rs} + k) - \sum_{k=0}^{k_{\max}} X_{rk} \ln(\alpha_{r0} + k) - \sum_{i=1}^n q_{ir} \ln(B(\alpha_r)),
\] (B6)

and once again the Jacobian can be written in terms of the digamma function:
\[
\frac{\partial E_r}{\partial \alpha_{rs}} = \sum_{k=0}^{k_{\max}} Y_{rsk} \psi(\alpha_{rs} + k) - \sum_{k=0}^{k_{\max}} X_{rk} \psi(\alpha_{r0} + k) - \sum_{i=1}^n q_{ir} [\psi(\alpha_{rs}) - \psi(\alpha_{r0})].
\] (B7)

The full EM algorithm as used in this paper consists of initializing \( p \) and \( \alpha \) to uniform values \( p = 1/c, \alpha_{rs} = 1 \), then iteratively updating \( q_{ir}, X_{rk}, \) and \( Y_{rsk} \) by Monte-Carlo sampling and \( \hat{p}_r = \frac{1}{n} \sum_i q_{ir} \) and \( \hat{\alpha}_r \) by numerical optimization, using Eqs. (B6) and (B7). The process is repeated until \( L \) no longer increases.

To make predictions about a node’s missing data we look at \( q_{ir} \), which is the marginal probability that node \( i \) is in group \( r \). The group \( r \) for which \( q_{ir} \) is largest is then taken to be our best guess for the group to which node \( i \) belongs. When assessing the success of the method in our synthetic tests, we check whether this value is equal to \( g_i \), the true group assignment of the same node, and compute the total fraction of nodes with missing data for which we have recovered the correct assignment.