LETTERS
https://doi.org/10.1038/s41562-018-0321-8
nature human behaviour

# Monophily in social networks introduces similarity among friends-of-friends

Kristen M. Altenburger and Johan Ugander*

**The observation that individuals tend to be friends with people who are similar to themselves, commonly known as homophily, is a prominent feature of social networks. While homophily describes a bias in attribute preferences for similar others, it gives limited attention to variability. Here, we observe that attribute preferences can exhibit variation beyond what can be explained by homophily. We call this excess variation monophily to describe the presence of individuals with extreme preferences for a particular attribute possibly unrelated to their own attribute. We observe that monophily can induce a similarity among friends-of-friends without requiring any similarity among friends. To simulate homophily and monophily in synthetic networks, we propose an overdispersed extension of the classical stochastic block model. We use this model to demonstrate how homophily-based methods for predicting attributes on social networks based on friends (that is, 'the company you keep') are fundamentally different from monophily-based methods based on friends-of-friends (that is, 'the company you're kept in'). We place particular focus on predicting gender, where homophily can be weak or non-existent in practice. These findings offer an alternative perspective on network structure and prediction, complicating the already difficult task of protecting privacy on social networks.**

Homophily is a commonly observed phenomenon in social networks whereby interactions occur frequently among similar individuals[1,2]. Homophily can originate from an individual's personal preference to become friends with similar others (choice homophily), structural opportunities to interact with similar others (induced homophily) or a combination of both[3]. The study of homophily focuses on aggregate patterns of interaction[4], whereas we highlight the need to jointly consider both the bias and excess variance in attribute preferences when studying social networks. To define excess variance, also called overdispersion, we first operationalize homophily as a bias parameter within a statistical model of interaction preferences in network data. Overdispersion then amounts to observing more variance in interaction preferences than expected under this homophily-only model. We refer to an overdispersion of preferences as monophily ('love of one') to indicate it as distinct from the preference bias introduced by homophily ('love of same'). Our analysis follows other advances in incorporating variance and overdispersion in social data analysis, such as estimating the size of subpopulations[5], documenting variations in the homophily of political ideologies[6], assessing gender variation in linguistic patterns[7], inferring social structure based on indirectly observed data[8] and leveraging link heterogeneity in label propagation[9].

An important consequence of homophily in a network—the typically assumed setting—is that even if an individual does not disclose private attribute information about themselves (such as their gender, age, race or political affiliation), methods for relational learning[10–15] can leverage attributes disclosed by that individual's similar friends to possibly predict their private attributes. However, when homophily is weak or non-existent, attribute prediction[16] is traditionally thought to be a difficult problem. We show that monophily in a network implies the existence of individuals with extreme preferences for a particular attribute possibly unrelated to their own attribute. The presence of these extreme preferences means that friends-of-friends are more likely to be similar. As a result, being friends with an individual with extreme attribute preferences is a strong predictive signal of one's own attributes. We motivate the empirical importance of monophily most strongly in the case of predicting gender on social networks (taken from the FB100 (ref. [17]) and Add Health[18] datasets, respectively; see Methods), where gender homophily can be weak in both online and offline settings[19–23]. We observe that monophily can still lead to accurate predictions in these weakly homophilous settings. We also observe the presence of monophily in settings known to exhibit strong homophily, specifically in the political affiliations of online blogs[24] and the contact network of terrorist group members and non-members in the Noordin Top Terrorist Network[25] (for dataset details, see Methods), demonstrating that there is additional structure to exploit for prediction beyond homophily.

This paper proceeds by first establishing how we choose to define the bias (homophily) and overdispersion (monophily) of attribute preferences. We then propose an extension of the stochastic block model—a classic model of biased preferences in networks[26]—that we call an overdispersed stochastic block model (oSBM). The oSBM can model homophily and monophily separately and allows us to compare our ability to predict missing attributes relative to the strength of homophily and monophily in a network. We show how the two-hop structural relationship induced by overdispersion (monophily) can exist in the complete absence of any one-hop bias (homophily). In terms of prediction, we thus find that overdispersed friendship preferences can drive successful classification even in the complete absence of any homophily. We conclude that friends-of-friends ('the company you're kept in') can disclose private attribute information that is otherwise undisclosed by friends ('the company you keep'). This finding extends the importance of privacy policies that protect networked data, while also proposing monophily as an intuitive structural property of social networks of independent interest. Finally, we highlight empirical results for predicting attributes where monophily-based prediction can perform well even in real-world networks with weak homophily.

The traditional homophily index of a graph[27,28] measures the aggregate pattern of individuals' biases in forming friendships with people of their own attribute class relative to people from other

Management Science and Engineering Department, Stanford University, Stanford, CA, USA. *e-mail: jugander@stanford.edu

classes. For a generic attribute class $r$ and assuming there are $k=2$ classes, the homophily index $\hat{h}_r$ with respect to class $r$ is defined as:

$$\hat{h}_r = \frac{\sum_{i\in r} d_{i,\mathrm{in}}}{\sum_{i\in r} d_{i,\mathrm{in}} + \sum_{i\in r} d_{i,\mathrm{out}}} = \frac{\sum_{i\in r} d_{i,\mathrm{in}}}{\sum_{i\in r} d_i} \tag{1}$$

where $d_i$ denotes its observed total degree, $d_{i,\mathrm{in}}$ denotes node $i$'s observed in-class degree and $d_{i,\mathrm{out}}$ denotes its observed out-class degree. For notational simplicity, we use $i\in r$ to index the set of nodes with attribute class $r$. Further, we let $n_r$ represent the total number of nodes with attribute $r$ and let $N$ denote the total population: $N=\sum_{r=1}^{k} n_r$. To describe the notation in the case of gender homophily among females ($r=F$), the homophily index for individuals $i\in F$ captures the relative number of interactions with other females ($d_{i,\mathrm{in}}$) relative to their total number of interactions ($d_i$).

We now show that the homophily index can be interpreted as the intercept term of a generalized linear model (GLM)[29]. For a comparison of the homophily index and the related measure (based on Pearson's correlation coefficient) known as assortativity[30], see Supplementary Note 1.4. This interpretation will later connect to a natural measure of monophily in terms of an overdispersed extension of that model. In measuring homophily for binary attributes (that is, between two attribute classes $r$ and $s$), we illustrate how to measure homophily for the first class, $r$. The set-up is analogous for the other class. We assume that each individual $i\in r$ forms in-class connections with the other $n_r$ individuals at a rate $p_{\mathrm{in},r}$ and out-class ties with the other $n_s$ individuals at a rate $p_{\mathrm{out}}$. We therefore expect for each individual $i\in r$ that their class-specific degrees obey the following distributions:

$$D_{i,\mathrm{in}} \mid p_{\mathrm{in},r} \sim \mathrm{Binom}(n_r, p_{\mathrm{in},r}) \tag{2}$$

$$D_{i,\mathrm{out}} \mid p_{\mathrm{out}} \sim \mathrm{Binom}(n_s, p_{\mathrm{out}}) \tag{3}$$

$$D_i \mid p_{\mathrm{in},r}, p_{\mathrm{out}} = D_{i,\mathrm{in}} \mid p_{\mathrm{in},r} + D_{i,\mathrm{out}} \mid p_{\mathrm{out}} \tag{4}$$

where $D_{i,\mathrm{in}}$ is a random variable describing the in-class degree, $D_{i,\mathrm{out}}$ describes the out-class degree and $D_i$ describes the total degree of node $i$ in class $r$. We explicitly condition these random variables on the parameters $p_{\mathrm{in},r}$ and $p_{\mathrm{out}}$ to make clear that these parameters are, for now, fixed and constant. Note that the random variables in equations (2)–(4) are approximately independent, but not completely: constraints on the joint distribution of the degrees corresponding to the constraints of the Erdős–Gallai theorem (since the degrees must correspond to a graph) create a dependence, but this dependence is small for graphs of modest size or larger[31] and we safely ignore it here. These distributions also correspond to situations where self-loops are allowed in the graph, which simplifies the derivation without any practical consequence.

We can model relative in-class preferences given the observed degree data using a GLM as follows. Let the observed degree data for class $r$ be $\{(d_{i,\mathrm{in}}, d_i), i\in r\}$. Among the individuals with attribute class $r$, their in-class degree distribution conditional on their total observed degree is approximately binomially distributed (Supplementary Note 1.1):

$$D_{i,\mathrm{in}} \mid d_i, p_{\mathrm{in},r}, p_{\mathrm{out}} \sim \mathrm{Binom}(d_i, n_r p_{\mathrm{in},r}/(n_r p_{\mathrm{in},r} + n_s p_{\mathrm{out}})) \tag{5}$$

We refer to the quantity $h_r = n_r p_{\mathrm{in},r}/(n_r p_{\mathrm{in},r} + n_s p_{\mathrm{out}})$ in the above expression as the 'homophily parameter', since it characterizes the bias for individuals to interact with similar others. By applying a logistic-binomial model[32,33] to the degree data, we can then obtain the maximum likelihood estimate (MLE) of this homophily parameter. The logistic link function is specified as $n_r p_{\mathrm{in},r}/(n_r p_{\mathrm{in},r} + n_s p_{\mathrm{out}}) = \mathrm{logit}^{-1}(\beta_{0r}) = e^{\beta_{0r}}/(1+e^{\beta_{0r}})$ assuming there are no additional covariates (which could otherwise be incorporated). For this model, the MLE of $\beta_{0r}$ (Supplementary Note 1.2) is then simply:

$$\widehat{\beta}_{0r}^{\mathrm{MLE}} = \mathrm{logit}\left(\sum_{i\in r} d_{i,\mathrm{in}} \Big/ \sum_{i\in r} d_i\right) \tag{6}$$

or equivalently $\widehat{\beta}_{0r}^{\mathrm{MLE}} = \mathrm{logit}(\hat{h}_r)$ and $\hat{h}_r = e^{\widehat{\beta}_{0r}^{\mathrm{MLE}}}/(1+e^{\widehat{\beta}_{0r}^{\mathrm{MLE}}})$, where $\hat{h}_r$ is exactly the homophily index specified in equation (1) above. Thus we see that the homophily index is precisely the MLE of the homophily parameter in this model, showing that it can also be computed from the intercept term estimated from a GLM applied to the observed degree data. An advantage of interpreting the homophily index within the GLM framework is that it provides a principled approach to computing statistical significance[34] using the $P$ value for the intercept term.

While this simple model captures the bias towards interactions occurring among similar individuals, it is a poor fit of the empirical variances in preferences that we observed in social data. That is, we observe that the empirical variance of attribute preferences can far exceed the variance expected under this homophily-only model (Supplementary Note 1.3). Fortunately, the GLM framework permits a straightforward way to test for overdispersion and then to extend the model in cases where overdispersion is statistically significant.

A variety of methods have been proposed to measure and model extra variation in count data[29,35–37]. We employ a quasi-likelihood approach[36], the least presumptive approach to modelling overdispersion compared with alternative methods. The quasi-likelihood set-up allows each node $i$ in class $r$ to have an individual latent preference for in-class friendships, $h_{i,r}$, such that $\mathbb{E}[h_{i,r}] = h_r$ and $\mathrm{Var}[h_{i,r}] = \phi_r h_r(1-h_r)$ for some $\phi_r \geq 0$. The parameter $\phi_r$ is introduced to incorporate the extra variation, and the variance is parameterized as such for notational convenience (Supplementary Note 1.3). This set-up can be thought of as loosely hierarchical, where $h_{i,r}$ is permitted to be random, but it does not specify a distribution on $h_{i,r}$. It instead uses $\phi_r$ to quantify how much nodes in class $r$ vary in allocating their in-class versus out-class friendships. When $\phi_r = 0$, there is no excess variation. $\phi_r > 0$ captures variation beyond the conventional model (Supplementary Note 1.3). Through an iterative procedure that maximizes a quasi-likelihood function (Supplementary Note 1.4), we jointly re-estimate the homophily index as well as the new monophily index captured by $\phi_r$.

We visually illustrate the distinction between homophily and monophily in one representative FB100 network, Amherst College, which has nearly balanced proportions of male and female students. Figure 1a shows histograms of individual relative proportions of same-gender friendships, illustrated separately for males and females. If this network was homophilous, we would expect to see the mean of these distributions deviate significantly from the relative class proportions, which is not the case for Amherst. We can see how the empirical distributions of preferences are more dispersed (less concentrated) than the homophily-only null distributions (for details of null model sampling, see Methods; for other schools, see Supplementary Note 3.2). Figure 1b provides an example network showing how similarity can emerge among friends-of-friends due to monophily, even in the complete absence of homophily or heterophily.

Next, we observe the existence of monophily across the networks we study. In Fig. 2a, we see that gender homophily indices are concentrated around the relative class proportions, and in Fig. 2b, we see that gender monophily is common across the full population of co-educational FB100 networks, shown as a function of class

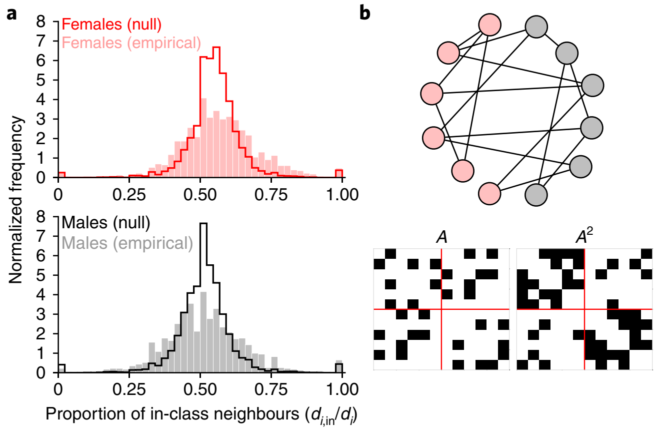

**Fig. 1 | Overdispersion in attribute preferences. a**, Amherst College Facebook network. Empirical distribution (filled bars) of in-class preferences for females and males compared with a null distribution (solid lines) based on preferences with binomial variation (see Methods). We observe overdispersion for females and males as the observed empirical variance is greater than under the null. **b**, A sample network without homophily or heterophily, but with monophily. We also show the link structure of the adjacency matrix ($A$) and the two-hop adjacency matrix ($A^2$). The matrices are grouped by attribute class where the red line separates classes. Monophily results in a block structure in the ties between friends-of-friends, but not between friends.

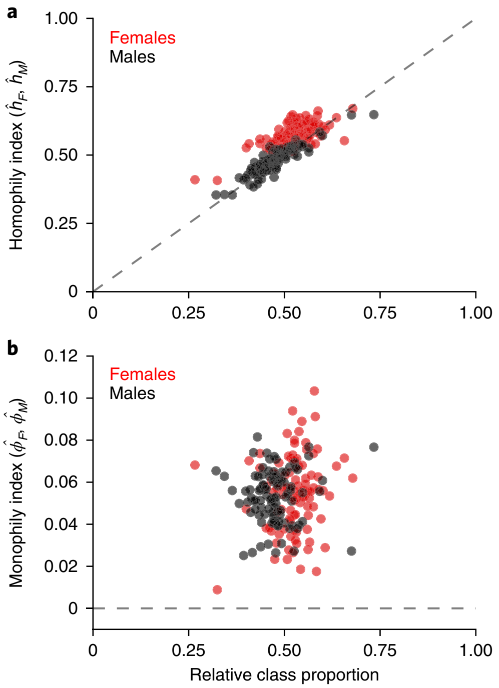

**Fig. 2 | Homophily and monophily across a population of friendship networks. a**,**b**, Gender homophily and monophily measured across the population of FB100 networks, showing the homophily index $\hat{h}_r$ (**a**) and the monophily index $\hat{\phi}_r$ (**b**) among both male (black) and female (red) students in each of the 97 co-educational college networks. The homophily indices are concentrated around relative class proportions (dashed line), while the monophily indices all show overdispersed preferences ($\hat{\phi}_r > 0$; $P < 10^{-3}$ for all networks) independent of the relative class proportions. Dashed lines indicate the lines of no homophily and no monophily, respectively.

proportion (proportion male or female). In Supplementary Table 3, we report the bias (homophily) and overdispersion (monophily) estimates for gender in the Amherst College network, political affiliations in a blog network and terrorist group membership in a communication network. We also test the statistical significance of these estimates[33,36] (Supplementary Note 1.4).

To examine the impact of monophily on network structure, we introduce a variation on the stochastic block model[26] (SBM) with overdispersed preferences so that we can independently simulate networks with known homophily and monophily. The stochastic block model, also known as the planted partition model[38], is a widely studied statistical distribution over graphs that is commonly used to model network association patterns. The SBM models preferences among $k$ classes of nodes by specifying a set of block sizes $n_1,\ldots,n_k$ and a preference matrix $\mathbf{P}$ where $\mathbf{P}_{a_i a_j}$ denotes the independent probability of an edge between nodes $i$ and $j$ in attribute classes $a_i$ and $a_j$. For modelling associations between two classes using an SBM, the matrix $\mathbf{P}$ is simply a $2\times 2$ matrix denoting the edge probabilities within and between classes. An assortative block structure is present when in-class probabilities are greater than out-class probabilities ($p_{\text{in}} > p_{\text{out}}$). We introduce overdispersion into the model by relaxing the usual restriction of fixed-class probabilities among all nodes in a given class. Instead, we assume a latent beta distribution on in- and out-attribute affinities[39]. Other latent distributions or other means of incorporating overdispersion[40,41] could be considered (for details of the oSBM, see Methods). The oSBM allows us to explore overdispersed preferences in a generative setting. In Fig. 3a we see that this model can capture the preference distributions we observe in empirical data. It also allows us to explore the relative performance of node inference methods on graphs with and without homophily and/or monophily.

We first provide a categorization of relational inference methods and then we examine the performance of these methods on oSBM graphs. Motivated by monophily, we observe that relational inference methods can be categorized based on the neighbourhood relationships they can exploit for classification, either using one-hop (friend) or two-hop (friend-of-friend) relations. This distinction amounts to more than just a difference in the number of relations: it is a direct analogue of a key distinction between the PageRank[42] and Hubs and Authorities[43] algorithms in graph ranking. PageRank is based on the principle that 'a node is important if it is linked to by other important nodes', while Hubs and Authorities is based on the principle that 'a node is important if it is linked to by nodes that link to important nodes'. These differing principles can extract very different notions of importance in graph ranking. Hubs and Authorities is motivated by web ranking problems where, for example, car companies don't link to other car companies but should still appear high in search results for 'cars'. Analogously, we observe that two-hop and one-hop methods are differently well-suited for different node classification problems.

Classification methods based on a node's one-hop (immediate) relations include:

- The one-hop majority vote (one-hop MV) classifier—also called the weighted-vote relational neighbour classifier[12]—builds directly on similarities between connected nodes. Unlabelled nodes are scored based on the proportion of labels among their neighbours. When a node does not have any labelled neighbours, the relative class proportions in the training data are used (Supplementary Note 2.1).
- The Zhu, Ghahramani and Lafferty (ZGL) method[44] scores unlabelled nodes by computing the relative probabilities of reaching each node in a graph under a random walk originating at the labelled node sets. The ZGL method can be characterized as an iterated/semi-supervised adaptation of one-hop MV[14].

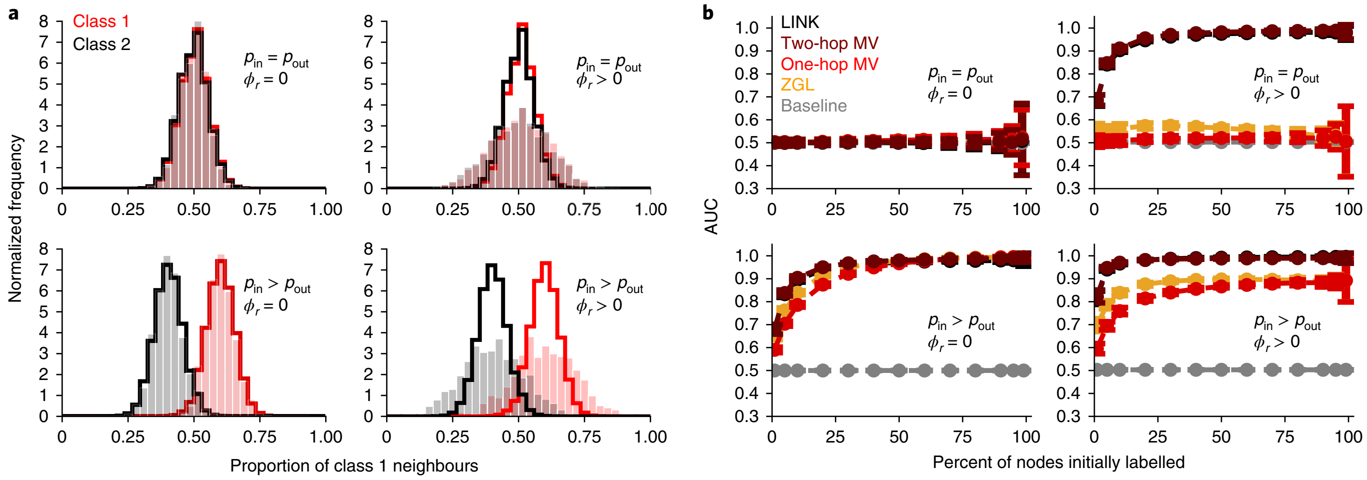

**Fig. 3 | Four different oSBMs and the associated performance of one-hop and two-hop classifiers. a**, Attribute preference distributions for four instances of oSBMs (filled bars) varying $p_{in}$, $p_{out}$ and $\phi_r$ parameters: no homophily and no monophily ($p_{in} = p_{out}$, $\phi_r = 0$); monophily but no homophily ($p_{in} = p_{out}$, $\phi_r > 0$); homophily but no monophily ($p_{in} > p_{out}$, $\phi_r = 0$); and both homophily and monophily ($p_{in} > p_{out}$, $\phi_r > 0$). A null distribution (solid line) is shown based on affinities with binomial variation (see Methods). **b**, Across the same corresponding oSBM settings, we compare the relative classification performance for the different inference methods described in the text. Points represent the mean AUC score and error bars denote s.d. across 100 cross-validated samples varying the percentage of initially labelled nodes in the network (see Methods). We observe a clear separation of performance in the case of monophily but no homophily.

Methods that exploit two-hop (neighbour-of-neighbour) relations include:

- The two-hop majority vote (two-hop MV) classifier uses the relationship between a node and its two-hop neighbours weighted by the number of length-2 paths. Unlabelled nodes are scored based on the weighted proportion of labels among their two-hop neighbours.
- LINK-logistic regression[45] uses labelled nodes to fit a regularized logistic regression model (Supplementary Note 2.2) that interprets rows of the adjacency matrix as sparse binary feature vectors, performing classification based on these features. The trained model is then applied to the feature vectors (adjacency matrix rows) of unlabelled nodes, which are scored based on the probability estimates from the model. Small variations that use the same feature set but employ, for example, support vector machines or random forests instead of logistic regression give qualitatively similar performance. We find that using the LINK feature set as part of a Naive Bayes classifier gives a clear view of LINK as a family of two-hop methods (Supplementary Note 2.3).

In Fig. 3b, we compare the relative performance of one-hop MV, ZGL, two-hop MV and LINK when attempting node classification on oSBM networks from each of four settings. We explore a typical node classification set-up where individuals reveal information completely at random[12,46–48] (that is, uniformly), meaning that the likelihood to be labelled or to provide public information does not depend on other attributes. The prediction task is then to infer private attributes using public attributes and the social network relationships. We compare these classification methods relative to a baseline model that assigns scores based on the relative class proportions observed in the training sample. We evaluate performance based on a weighted area under the curve (AUC) score. AUC is a typical metric for summarizing receiver operating characteristic curves across a range of decision thresholds[49] and is commonly employed for evaluating classifier performance in networked settings[50,51]. For a fuller discussion of alternative performance metrics such as accuracy, see Supplementary Note 2.4. We observe that in oSBM networks configured with only homophily ($p_{in} > p_{out}$, $\phi_r = 0$), all inference methods perform well. Meanwhile, in networks with only monophily ($p_{in} = p_{out}$, $\phi_r > 0$), one-hop MV and ZGL have no predictive power while LINK-logistic regression and two-hop MV show impressive performance despite the complete lack of homophily. We conclude that the presence of monophily can be sufficient, even in the complete absence of homophily, for accurate attribute inference in networks (for additional details, see Supplementary Note 2.5).

We examine predictive performance in applied settings where we have previously observed significant monophily and varying degrees of homophily. We focus on predicting gender, political affiliation and terrorist group membership in the networks previously introduced. In Fig. 4a, we observe limited performance using one-hop methods (one-hop MV and ZGL) to predict gender in Amherst College, our representative Facebook network (for additional FB100 networks, see Supplementary Note 3.2). Meanwhile, we see that the two-hop methods (two-hop MV and LINK) have higher performance, corroborating our intuition for two-hop methods being able to surface structural signals for classification in the presence of overdispersed preferences. In Fig. 4b, we observe that classifying gender on Facebook networks by two-hop MV consistently outperforms classification based on one-hop MV when evaluating classification performance across fully labelled networks, and two-hop MV in turn outperforms three-hop, four-hop and five-hop MV. We attribute the success of two-hop MV to monophily in this weakly homophilous setting.

Next, we evaluate prediction in homophilous networks—the typically assumed setting. For predicting political affiliation in a strongly homophilous network, Fig. 4c shows that all classification methods perform equivalently as long as a modest number of nodes are initially labelled. Since most individuals in this setting exhibit strong preferences for similar friends, similarity will also exist for friends-of-friends. That said, we observe the presence of overdispersed preferences in this network as well (Supplementary Note 3.4). The presence of monophily along with homophily in this setting raises additional considerations for monophily's impact on other social processes (for example, information diffusion processes that traditionally only consider homophily[52]). Finally, for predicting terrorist group membership, in Fig. 4d we

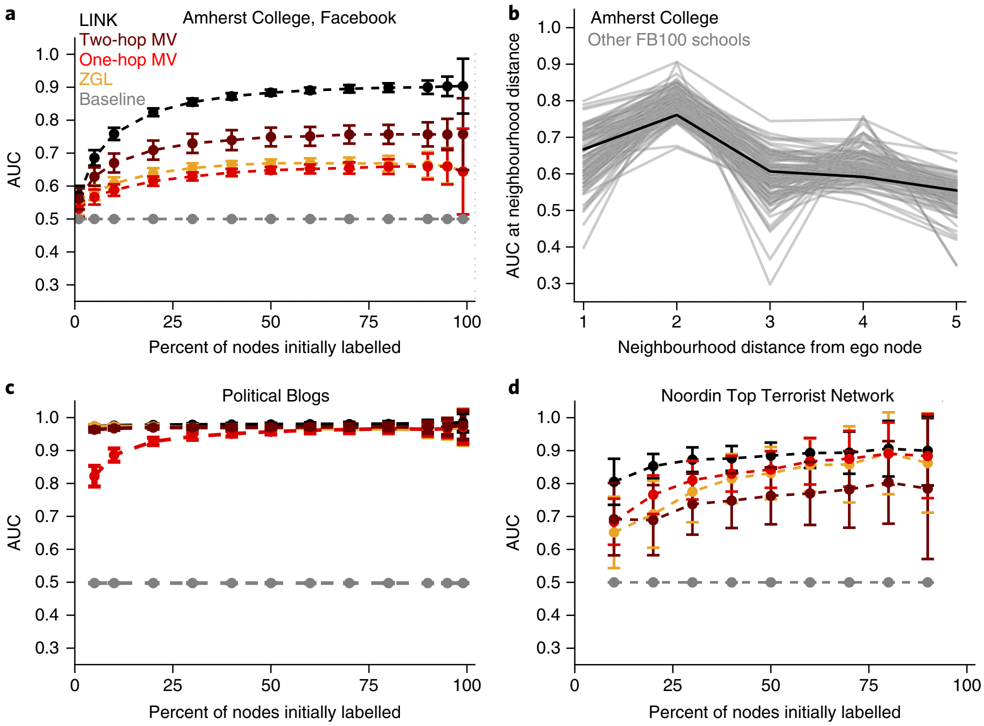

**Fig. 4 | Predicting gender, political affiliation and terrorist group affiliation.** Relative performance of one-hop (one-hop MV and ZGL) and two-hop (two-hop MV and LINK) relational learning classifiers across three demonstration networks: **a**, Amherst College, Facebook with 1,015 females (F) and 1,017 males (M), where $\hat{h}_F^{***}=0.55$, $\hat{h}_M^{***}=0.51$, $\hat{\phi}_F^{***}=0.04$ and $\hat{\phi}_M^{***}=0.04$. **c**, Political blogs (586 liberal (L) and 636 conservative (C)), where $\hat{h}_L^{***}=0.90$, $\hat{h}_C^{***}=0.91$, $\hat{\phi}_L^{***}=0.23$ and $\hat{\phi}_C^{***}=0.19$. **d**, Noordin Top Terrorist Network with 47 members (M) and 27 non-members (N), where $\hat{h}_M^{***}=0.89$, $\hat{h}_N=0.55$, $\hat{\phi}_M=0.02$ and $\hat{\phi}_N=0.00$. For the homophily and monophily indices, we indicate statistical significance ($^{*}P<0.1$, $^{**}P<0.05$, $^{***}P<0.01$). Points in the figures represent mean AUC scores and error bars denote s.d. across 100 cross-validated samples varying the percentage of initially labelled nodes in the network (see Methods). For additional Facebook colleges, see Supplementary Note 3.2. **b**, Across the 97 Facebook colleges (Amherst in black), the AUC for *k*-hop MV on fully labelled networks when predicting gender, varying the number of hops *k* at which *k*-hop MV is performed. The AUC at $k=2$ is highest for all 97 colleges.

observe higher performance for one-hop MV relative to two-hop MV. A critical operational assumption for terrorist networks is that these are homophilous networks[53] or that the 'company you keep' is the predominant network characteristic. We observe non-significant monophily among the membership class, although the lack of statistical significance may be due to the small network size (Supplementary Note 3.5). Our analysis emphasizes that additional consideration should be given to monophily in intelligence applications[54] since focusing only on homophily overlooks the friend-of-friend correlations caused by monophily that may still exist even after accounting for homophily.

This work introduces monophily as a fundamental property of social network preferences deserving broad consideration. In the spirit of a solution-oriented science[55,56] we have focused on the practical consequences of monophily for inferring missing attributes on social networks. These findings provide a new perspective on social network structure in general and attribute classification in particular, as well as further complicating the already difficult task of preserving privacy in social networks. By also introducing the oSBM as a modelling tool, we show how it is possible for overdispersed preferences to explain the 'predictability' of attributes in relational inference via two-hop similarity in settings with weak or even nonexistent homophily. The empirical overdispersion of preferences documented in this work motivates a re-examination of two-hop network structure in network analysis very broadly; for example, developing community detection methods[57] that engage with relations among friends-of-friends, or studying the evolution and dynamics of preference variance in temporal networks[58,59]. Methods for studying privacy in bipartite affiliation networks[45,60] should also be revisited. We believe that the overdispersion of preferences deserves study as a social structure in its own right and encourage investigations into correlates of extreme preferences. While preference biases have long been the predominant focus of social structure in networks, this work highlights the need to simultaneously give serious parallel consideration to variability.

## Methods

**Description of data.** When studying gender, we analysed populations of networks from two sources—the FB100 network dataset[17] (Supplementary Note 3.1) and the Add Health in-school friendship nomination dataset[18] (Supplementary Note 3.3). FB100, analysed in the main paper, consists of online friendship networks from Facebook collected in September 2005 from 100 US colleges primarily consisting of college-aged individuals[61]. Traud et al.[17,61] provide extensive documentation of the descriptive statistics of these networks. We excluded Wellesley College, Smith College and Simmons College from our analysis, all of which are single-sex institutions with >98% female nodes in the original network datasets. For political affiliation, we analysed the undirected version of the hyperlink network between US political blogs, where an edge exists as long as at least one weblog links to another[24]. For persons of interest, we analysed the communication network among members and non-members of the Noordin Top splinter group, where membership is defined by individuals who participated in a Noordin operation, were disclosed as being a member of Noordin's inner circle and/or were family or close friends of Noordin[25]. For all networks, we restricted the analysis to only nodes that disclose their attributes, completely removing those with missing labels. We also restricted the analyses to nodes in the largest (weakly) connected component to benchmark against classification methods[44] that assume a connected graph.

**Null distribution of preferences.** To visualize variation in attribute preferences in empirical networks across degrees, we compared the variance of the empirical distribution of $d_{i,\mathrm{in}}/d_i$ across all nodes $i$ in the same class $r$ with the variance of a binomial null distribution without overdispersion. Since the basic model assumes that $(D_{i,\mathrm{in}} \mid D_i = d_i) \sim \mathrm{Binom}(d_i, \hat{h}_r)$, we simulated draws from this distribution by repeatedly sampling from $\mathrm{Binom}(d_i, \hat{h}_r)$ for each node $i$ to produce a distribution of samples under the null.

**oSBM.** The proposed oSBM was defined by the block sizes $n_1,\ldots,n_k$, $k \times k$ preference matrix **P** and additional overdispersion parameters $\phi^{\star}_{\text{in}} \geq 0$ and $\phi^{\star}_{\text{out}} \geq 0$. Networks were generated from the model via a hierarchical approach. First, each node's in- and out-class degrees were generated by sampling class preference parameters ($p_{i,\text{in}}$ and $p_{i,\text{out}}$) from an appropriate latent beta distribution with specified means $p_{\text{in}}$ and $p_{\text{out}}$ for in- and out-class probabilities, respectively. We assumed the same mean across all attribute classes $r$, so we denoted this mean by $p_{\text{in}}$ instead of $p_{\text{in},r}$ for a given class $r$. Given the sampled individual preferences, a graph was generated analogously to how the degree-corrected SBM[62] attains prescribed degrees using a Chung–Lu construction[63], with expected in-degrees $d_{i,\text{in}} = n_r p_{i,\text{in}}$ and expected out-degrees $d_{i,\text{out}} = (N - n_r)p_{i,\text{out}}$ (Supplementary Note 4). We note that this oSBM complements related work on overdispersion in social network surveys[5] where an individual's relations within a class are taken to be distributed gamma–Poisson. The oSBM provides a full network model beyond mere counts. The number of connections a node has from a specific class will approximately follow a beta-binomial distribution in an oSBM, a close relative of the gamma–Poisson distribution[64].

**Description of cross-validation.** We varied the percentage of initially labelled nodes by selecting a labelled sample uniformly at random[12]. We trained our models varying the percentage of initially labelled nodes in the network. For a given fixed percent of labelled individuals (training dataset), we measure classification performance on the remaining unlabelled nodes (testing dataset), using the same train/test splits across the different inference methods. We evaluated performance for 100 different random samples of initially labelled nodes, reporting the mean weighted AUC for each given fixed percent of initially labelled nodes where the weights were based on the relative number of true class training labels. For predicting membership in the Noordin Top group, we enforced a stratified random sampling set-up due to the small dataset size. The vertical error bars denote the s.d. in AUC scores across the 100 samples.

**Code availability.** IPython notebooks and all corresponding code are available at https://github.com/kaltenburger/homophily_monophily_NHB, documenting all results and figures. All code was run using Python version 2.7.12.

**Data availability.** The FB100 dataset is publicly available at https://archive.org/details/oxford-2005-facebook-matrix and other public repositories. The Add Health dataset can be obtained from the Carolina Population Center at the University of North Carolina by contacting addhealth_contracts@unc.edu. The Political Blogs dataset used in this paper is available from http://www-personal.umich.edu/~mejn/netdata/ and elsewhere. The Noordin Top Terrorist Network is available from https://sites.google.com/site/sfeverton18/research/appendix-1.

Received: 7 April 2017; Accepted: 15 February 2018;
Published online: 19 March 2018

## References

1. Lazarsfeld, P. F. & Merton, R. K. Friendship as a social process: a substantive and methodological analysis. *Freedom Control Mod. Soc.* **18**, 18–66 (1954).
2. McPherson, M., Smith-Lovin, L. & Cook, J. M. Birds of a feather: homophily in social networks. *Annu. Rev. Sociol.* **27**, 415–444 (2001).
3. Kossinets, G. & Watts, D. J. Origins of homophily in an evolving social network. *Am. J. Sociol.* **115**, 405–450 (2009).
4. Raftery, A. E. Statistics in sociology, 1950–2000: a selective review. *Sociol. Methodol.* **31**, 1–45 (2001).
5. Zheng, T., Salganik, M. J. & Gelman, A. How many people do you know in prison? Using overdispersion in count data to estimate social structure in networks. *J. Am. Stat. Assoc.* **101** 409–423 (2006).
6. Boutyline, A & Willer, R. The social structure of political echo chambers: variation in ideological homophily in online networks. *Pol. Psychol.* **38** 551–569 (2017).
7. Bamman, D., Eisenstein, J. & Schnoebelen, T. Gender identity and lexical variation in social media. *J. Socioling.* **18**, 135–160 (2014).
8. McCormick, T. H. et al. A practical guide to measuring social structure using indirectly observed network data. *J. Stat. Theory Pract.* **7**, 120–132 (2013).
9. Peel, L. Graph-based semi-supervised learning for relational networks. In *Proc. 2017 SIAM Int. Conf. Data Mining* 435–443 (SIAM, 2017).
10. Neville, J. & Jensen, D. Supporting relational knowledge discovery: lessons in architecture and algorithm design. In *Proc. Data Mining Lessons Learned Workshop, 19th Int. Conf. Machine Learning* (JMLR, 2002).
11. Jensen, D., Neville, J. & Gallagher, B. Why collective inference improves relational classification. In *Proc. 10th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining* 593–598 (ACM, 2004).
12. Macskassy, S. A. & Provost, F. Classification in networked data: a toolkit and a univariate case study. *J. Mach. Learn. Res.* **8**, 935–983 (2007).
13. Sen, P. et al. Collective classification in network data. *AI Mag.* **29**, 93–106 (2008).
14. Bhagat, S., Cormode, G. & Muthukrishnan, S. in *Social Network Data Analytics* 115–148 (Springer, Boston, MA, 2011).
15. Taskar, B., Abbeel, P. & Koller, D. Discriminative probabilistic models for relational data. In *Proc. 18th Conf. Uncertainty in Artificial Intelligence* 485–492 (Morgan Kaufmann, 2002).
16. Duncan, G. T. & Lambert, D. Disclosure-limited data dissemination. *J. Am. Stat. Assoc.* **81**, 10–18 (1986).
17. Traud, A. L., Mucha, P. J. & Porter, M. A. Social structure of Facebook networks. *Physica A Stat. Mech. Appl.* **391**, 4165–4180 (2012).
18. Resnick, M. D. et al. Protecting adolescents from harm: findings from the national longitudinal study on adolescent health. *JAMA* **278**, 823–832 (1997).
19. Ugander, J., Karrer, B., Backstrom, L. & Marlow, C. The anatomy of the Facebook social graph. Preprint at https://arxiv.org/abs/1111.4503 (2011).
20. Thelwall, M. Homophily in MySpace. *J. Am. Soc. Inf. Sci. Technol.* **60**, 219–231 (2009).
21. Shrum, W., Cheek, N. H. & Hunter, S. Friendship in school: gender and racial homophily. *Sociol. Edu.* **61**, 227–239 (1988).
22. Neal, J. W. Hanging out: features of urban children's peer social networks. *J. Soc. Pers. Rel.* **27**, 982–1000 (2010).
23. Laniado, D., Volkovich, Y., Kappler, K. & Kaltenbrunner, A. Gender homophily in online dyadic and triadic relationships. *EPJ Data Sci.* **5**, 19 (2016).
24. Adamic, L. A. & Glance, N. The political blogosphere and the 2004 US election: divided they blog. In *Proc. 3rd Int. Workshop Link Discovery* 36–43 (ACM, 2005).
25. Roberts, N. & Everton, S. F. *Roberts and Everton Terrorist Data: Noordin Top Terrorist Network (Subset)* [Machine-readable data file] (2011).
26. Holland, P. W., Laskey, K. B. & Leinhardt, S. Stochastic blockmodels: first steps. *Social. Netw.* **5**, 109–137 (1983).
27. Coleman, J. Relational analysis: the study of social organizations with survey methods. *Human Organ.* **17**, 28–36 (1958).
28. Currarini, S., Jackson, M. O. & Pin, P. An economic model of friendship: homophily, minorities, and segregation. *Econometrica* **77**, 1003–1045 (2009).
29. McCullagh, P. & Nelder, J. A. *Generalized Linear Models* Vol. 37 (CRC Press, London, 1989).
30. Newman, M. E. J. Assortative mixing in networks. *Phys. Rev. Lett.* **89**, 208701 (2002).
31. Van Der Hofstad, R. *Random Graphs and Complex Networks* Vol. 1 (Cambridge Univ. Press, Cambridge, 2016).
32. Agresti, A. & Kateri, M. *Categorical Data Analysis* (Springer, Berlin, 2011).
33. Gelman, A. & Hill, J. *Data Analysis Using Regression and Multilevel/Hierarchical Models* (Cambridge Univ. Press, Cambridge, 2006).
34. Signorile, V. & O'Shea, R. M. A test of significance for the homophily index. *Am. J. Sociol.* **70**, 467–470 (1965).
35. Wedderburn, R. W. Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika* **61**, 439–447 (1974).
36. Williams, D. A. Extra-binomial variation in logistic linear models. *J. R. Stat. Soc. C Appl. Stat.* **31**, 144–148 (1982).
37. Morel, J. G. & Nagaraj, N. K. A finite mixture distribution for modelling multinomial extra variation. *Biometrika* **80**, 363–371 (1993).
38. Condon, A. & Karp, R. M. Algorithms for graph partitioning on the planted partition model. *Random Struct. Algor.* **18**, 116–140 (2001).
39. Crowder, M. J. Beta-binomial ANOVA for proportions. *J. R. Stat. Soc. C Appl. Stat.* **27**, 34–37 (1978).
40. DiPrete, T. A. & Forristal, J. D. Multilevel models: methods and substance. *Annu. Rev. Sociol.* **20**, 331–357 (1994).
41. Guo, G. & Zhao, H. Multilevel modeling for binary data. *Annu. Rev. Sociol.* **26**, 441–462 (2000).
42. Page, L., Brin, S., Motwani, R. & Winograd, T. *The PageRank Citation Ranking: Bringing Order to the Web* (Stanford Univ. InfoLab, 1999).
43. Kleinberg, J. M. Authoritative sources in a hyperlinked environment. *J. ACM* **46**, 604–632 (1999).
44. Zhu, X., Ghahramani, Z. & Lafferty, J. Semi-supervised learning using Gaussian fields and harmonic functions. In *Proc. 20th Int. Conf. Machine Learning* 912–919 (JMLR, 2003).
45. Zheleva, E. & Getoor, L. To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles. In *Proc. 18th Int. Conf. World Wide Web* 531–540 (IW3C2, 2009).
46. He, J., Chu, W. W. & Liu, Z. V. Inferring privacy information from social networks. In *Int. Conf. Intelligence and Security Informatics* 154–165 (Springer, 2006).
47. Rubin, D. B. Inference and missing data. *Biometrika* **63**, 581–592 (1976).
48. Heitjan, D. F. & Basu, S. Distinguishing "missing at random" and "missing completely at random". *Am. Stat.* **50**, 207–213 (1996).

49. Bradley, A. P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* **30**, 1145–1159 (1997).
50. Gallagher, B. & Eliassi-Rad, T. in *Advances in Social Network Mining and Analysis* 1–19 (Springer, Berlin, 2010).
51. Gong, N. Z. et al. Joint link prediction and attribute inference using a social-attribute network. *ACM Trans. Intell. Syst. Technol.* **5**, 27 (2014).
52. Golub, B. & Jackson, M. O. How homophily affects the speed of learning and best-response dynamics. *Q. J. Econ.* **127**, 1287–1338 (2012).
53. Stohl, C. & Stohl, M. Networks of terror: theoretical assumptions and pragmatic consequences. *Commun. Theory* **17**, 93–124 (2007).
54. Carrington, P. J. in *The SAGE Handbook of Social Network Analysis* 236–255 (SAGE, Los Angeles, CA, 2011).
55. Hofman, J. M., Sharma, A. & Watts, D. J. Prediction and explanation in social systems. *Science* **355**, 486–488 (2017).
56. Watts, D. J. Should social science be more solution-oriented? *Nat. Hum. Behav.* **1**, 0015 (2017).
57. Decelle, A., Krzakala, F., Moore, C. & Zdeborová, L. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys. Rev. E* **84**, 066106 (2011).
58. McPherson, J. M. & Ranger-Moore, J. R. Evolution on a dancing landscape: organizations and networks in dynamic Blau space. *Social. Forces* **70**, 19–42 (1991).
59. Yang, Y. et al. Gender differences in communication behaviors, spatial proximity patterns, and mobility habits. Preprint at https://arxiv.org/abs/1607.06740 (2016).
60. Kosinski, M., Stillwell, D. & Graepel, T. Private traits and attributes are predictable from digital records of human behavior. *Proc. Natl Acad. Sci. USA* **110**, 5802–5805 (2013).
61. Traud, A. L., Kelsic, E. D., Mucha, P. J. & Porter, M. A. Comparing community structure to characteristics in online collegiate social networks. *SIAM Rev.* **53**, 526–543 (2011).
62. Karrer, B. & Newman, M. E. J. Stochastic blockmodels and community structure in networks. *Phys. Rev. E* **83**, 016107 (2011).
63. Chung, F. & Lu, L. Connected components in random graphs with given expected degree sequences. *Ann. Comb.* **6**, 125–145 (2002).
64. Chatfield, C. & Goodhardt, G. J. in *Mathematical Models in Marketing* 53–57 (Springer, Berlin, 1976).

## Acknowledgements

We thank B. Fosdick, J. Kleinberg, I. Kloumann, D. Larremore, J. Nishimura, M. Porter, M. Salganik and S. Way for helpful comments. We thank attendees of the 2016 International Conference on Computational Social Science and the 2016 SIAM Workshop on Network Science for comments. This work was supported in part by the Department of Defense (DoD) through the National Defense Science and Engineering Graduate Fellowship (NDSEG) programme, the Akiko Yamazaki and Jerry Yang Engineering Fellowship and a David Morgenthaler II Faculty Fellowship. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

## Author contributions

K.M.A. and J.U. designed and performed the research and wrote the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** is available for this paper at https://doi.org/10.1038/s41562-018-0321-8.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Correspondence and requests for materials** should be addressed to J.U.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.