sociological science

Estimating Homophily in Social Networks Using Dyadic Predictions

George Berry,[a] Antonio Sirianni,[b] Ingmar Weber,[c] Jisun An,[d] Michael Macy[a]

a) Cornell University; b) Dartmouth College; c) Qatar Computing Research Institute; d) Singapore Management University

Abstract: Predictions of node categories are commonly used to estimate homophily and other relational properties in networks. However, little is known about the validity of using predictions for this task. We show that estimating homophily in a network is a problem of predicting categories of dyads (edges) in the graph. Homophily estimates are unbiased when predictions of dyad categories are unbiased. Node-level prediction models, such as the use of names to classify ethnicity or gender, do not generally produce unbiased predictions of dyad categories and therefore produce biased homophily estimates. Bias comes from three sources: sampling bias, correlation between model errors and node degree, and correlation between node-level model errors along dyads. We examine three methods for estimating homophily: predicting node categories, predicting dyad categories, and a hybrid "ego–alter" approach. This analysis indicates that only the dyadic prediction approach is unbiased, whereas the node-level approach produces both high bias and high overall error. We find that node-level classification performance is not a reliable indicator of accuracy for homophily. Although this article focuses on a particular version of homophily, results generalize to heterophilous cases and other dyadic measures. We conclude with suggestions for research design. Code for this article is available at https://github.com/georgeberry/autocorr.

Keywords: homophily; networks; machine learning; quantitative methodology

Citation: Berry, George, Antonio Sirianni, Ingmar Weber, Jisun An, and Michael Macy. 2021. "Estimating Homophily in Social Networks Using Dyadic Predictions." Sociological Science 8: 285-307.

Received: January 24, 2021

Accepted: April 4, 2021

Published: August 2, 2021

Editor(s): Jesper Sørensen, Filiz Garip

DOI: 10.15195/v8.a14

RESEARCHERS have long sought to understand the pattern, causes, and consequences of *homophily*, or the tendency for like to associate with like (Coleman 1958; Blau 1977; Marsden 1987; McPherson, Smith-Lovin, and Brashears 2006; Kossinets and Watts 2009). Measuring the similarity of nodes along along racial, ethnic, gender, social status, cultural, emotional, political, and socioeconomic lines is a core area of research in network science (Marsden 1987; McPherson, Smith-Lovin, and Cook 2001; Mollica, Gray, and Treviño 2003; McPherson et al. 2006; Kossinets and Watts 2009; Thelwall 2009; Wimmer and Lewis 2010; De Choudhury 2011; DiPrete et al. 2011; Lewis, Gonzalez, and Kaufman 2012; Colleoni, Rozza, and Arvidsson 2014; Smith, McPherson, and Smith-Lovin 2014; Bakshy, Messing, and Adamic 2015; Halberstam and Knight 2016; Himelboim et al. 2016; Cesare et al. 2017b; Hofstra et al. 2017; Messias, Vikatos, and Benevenuto 2017).

There are cases when we are interested in understanding homophily with respect to a certain node attribute, but we must rely on predictions (or imputations) of that attribute for analysis. For instance, when studying social media platforms such as Twitter, detailed network and behavioral information is available via an API, but demographic and attitudinal characteristics (race, gender, class, political ideology, etc.) are not. Collecting the absent information for all nodes would be prohibitively expensive, necessitating that researchers turn to predictions. Usually, predictions are generated using a combination of readily available information (the

network, text) and a small set of nodes where the attribute of interest is collected. Results presented in this article apply to this context. More generally, results extend to evaluating homophily with respect to any partially observed values that are predicted or imputed from other information. This can include survey or administrative data analyzed from a network perspective.

We study the use of such predictions or imputations to estimate homophily from a methodological point of view. The goal is to determine when predictions can be used to provide an unbiased and low-error estimate of homophily. Despite the widespread usage of predictions to estimate homophily and other relational measures in a variety of empirical settings (De Choudhury 2011; Colleoni et al. 2014; Bakshy et al. 2015; Himelboim et al. 2016; Hobbs et al. 2016; Boutyline and Willer 2017; Cesare et al. 2017b; Messias et al. 2017), little work investigates when such predictions provide reasonable estimates (Berry et al. 2018).

A standard process for studying relational measures with predictions proceeds as follows: the attribute of interest for a small set of *ground truth* nodes is coded by humans, and then a supervised learning model is used to predict the attribute of interest for all nodes in the network (Molina and Garip 2019). In the case of online social network data, publicly available information (names, profile photos, or text) along with the ground truth labels are used to generate predictions (De Choudhury 2011; Al Zamal, Liu, and Ruths 2012; Barberá 2016; Cesare et al. 2017b; Messias et al. 2017; Hofstra and de Schipper 2018). Such node-level prediction models are evaluated using metrics like accuracy, precision, recall, and area under the curve (AUC).

We find that estimating homophily is a problem of predicting dyad categories, rather than node categories. The article proceeds from this basic point. We should not expect models designed to provide unbiased (or low-error) predictions of node categories to also provide unbiased (or low-error) predictions of *dyad* categories. Methods that provide node-level predictions produce bias when estimating homophily for three reasons: (1) correlation of model errors with node degree, (2) network autocorrelation of node-level classification errors along dyads, and (3) sampling bias: a model designed to generalize to the population of nodes does not necessarily also generalize to the population of edges. Beyond bias, we find that node-level models produce the highest overall error of methods we consider and that node-level performance metrics (e.g., accuracy, AUC) are not reliable indicators of homophily estimate accuracy.

Although node-level models are inappropriate for a relational measure such as homophily, predicting dyad categories directly can provide an unbiased estimate of homophily. When direct dyadic prediction is not feasible, we show that an intermediate approach, which we term *ego–alter*, can reduce both bias and overall error. These three methods (node, dyad, and ego–alter) are formally studied and evaluated with simulation.

The payoff to improving homophily estimates is obvious. Although these findings can augment approaches to homophily estimation in multiple research contexts, the results presented here are particularly important for those who study online interaction. Understanding rates of interaction between groups online both builds on prior studies of homophily grounded in observed offline interaction

and helps us understand how online spaces are more or less conducive to social integration. In online social networks, understanding the structure of homophily is crucial for understanding echo chambers, differential access to information, network integration, and the sources of biased perceptions of networks (Barberá et al. 2015; Karimi et al. 2018; Lee et al. 2019).

Contributions and Roadmap

We formalize the homophily estimation problem as a dyadic prediction problem (also referred to as "edge classification"; see Aggarwal, He, and Zhao [2016]). In this case, the goal is to predict whether nodes at both ends of a dyad have a certain attribute. This contrasts with the more common node prediction approach, where the goal is to predict whether an individual node has a certain attribute.

Formalization clarifies sources of bias when estimating homophily. First, there may be sampling bias: the sample of ground truth nodes may be collected in a way that does not generalize to the target population (the population of edges in this case).

Second, there may be correlation between model residuals and node degree (the node's total number of connections in the network). Sociologically, high degree nodes may be particular in ways not captured by the available covariates, leading to concentrated prediction errors for high degree nodes.

Finally, the residuals of adjacent nodes may be correlated. This is the most fundamental problem studied, and to our knowledge it can only be avoided by employing a dyadic prediction model. Sociologically, this may occur because node-level classifiers may be disproportionately biased in certain parts of the larger social network. Put differently, attribute prediction error may be correlated with network position.

Statistically, residual correlation along dyads also occurs because node error is correlated with the unexplained part of neighbor outcomes. This means that, as long as true homophily levels are nonzero, any estimate of homophily based purely on the aggregation of node predictions can have network autocorrelated errors. We provide tools to deal with these three problems and improve homophily estimates.

The findings here underscore the importance of incorporating predictive methods at the dyadic level. We show that randomly sampling edges and predicting dyad categories directly produces an unbiased estimate of homophily in the network. Although this dyad-level estimator is unbiased, it can produce high absolute error estimates in some cases and may be infeasible in practice for data collection reasons.

We then study the more common strategy of using node-level predictions to estimate homophily. This demonstrates that the standard practice of developing a node-level classification model to maximize node-level classification performance does not provide reliable homophily estimates. We explore a variety of node-level estimation strategies, which show both high bias and high overall error. These differences in error for homophily can be unrelated to differences in node-level classifier performance. For instance, we show that it is possible to develop two models that perform nearly identically on node-level performance measures like

accuracy and AUC, yet have average biases for estimating homophily that differ by a factor of three (2.5 percent and 7.5 percent respectively).

Careful analysis indicates that a specific node-level modeling strategy, which we term the *ego–alter model*, can substantially reduce error. This approach makes a node-level prediction for each neighbor the node has, and can be considered a middle ground between a node and dyadic approach. It also indicates that bias and error can be reduced by including specific network features as variables. We study the relative performance of different models model by simulation, and we conclude with practical suggestions.

The article proceeds as follows. First, we set up the problem and formalize our measure of homophily. Then, we show that, in the case of random edge sampling, a model predicting dyads directly produces an unbiased estimate of homophily. Next, we formally show that node-level models should not be expected to produce unbiased homophily estimates. To address this problem, we then introduce the ego–alter model. These three models (dyad, node, ego–alter) are evaluated using simulations. Using these simulations, we demonstrate that classification performance at the node level (e.g., accuracy, precision, recall, AUC) is not informative about bias and error in homophily estimates. Finally, we conclude with suggestions for research design. The online supplement contains additional simulations and examples of how researchers can extend findings here to other relational measures.

Problem Statement

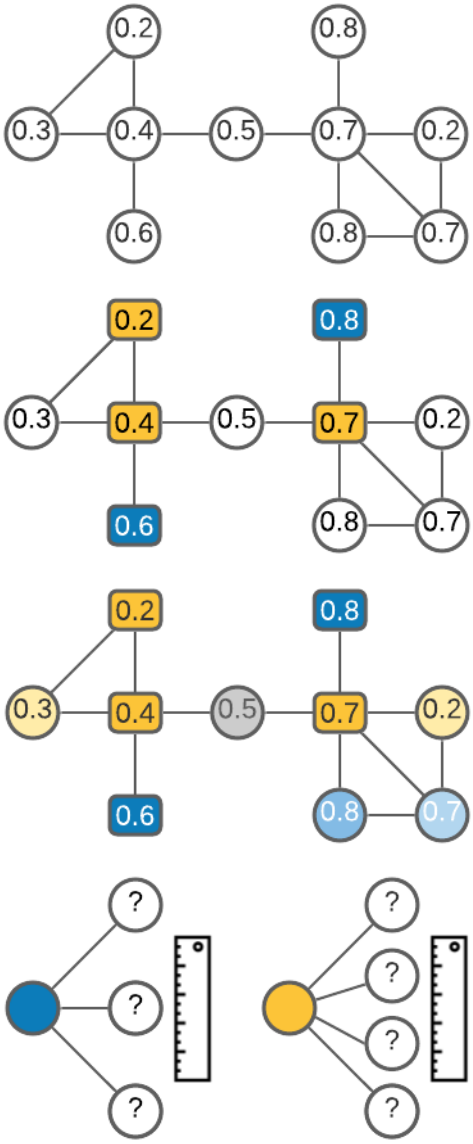
The process of homophily estimation on a graph (network) involves ascertaining the true category of a small set of nodes or edges for the attribute of interest, using this ground truth set to predict the attribute for the full set of nodes or edges. These predictions are used to estimate homophily. This process is depicted visually in Figure 1.

More formally, consider a set of nodes (vertices) V and edges E that comprise a graph, $G = (V, E)$. The edges in the graph are bidirected (symmetric), meaning that an edge $(i, j) \in E \implies (j, i) \in E$. This can be thought of as an undirected graph with explicit consideration of each end of each dyad. Nodes have an attribute of interest Y_i and a set of covariates X_i.

The researcher is unaware of the main attribute of interest Y_i for most nodes. However, we assume the structure of the graph is fixed and known by the researcher and that covariates X_i are observable for each node. This means that network statistics, such as the degree (number of connections) of each node, can be computed. By assumption, the attribute of interest is binary, $Y_i \in \{0, 1\}$. Results generalize to node categories that take on more than two values.

The node categories Y_i are known only for a small set of nodes, termed the *ground truth set* or the *labeled set*. A model must be used to predict Y_j for all j not in the ground truth set, using information in the network G and in the covariates X_j. Predictions $\hat{Y}_j$ are generated for all unlabeled nodes, and these predictions are used to estimate the aggregate outcome in the network.

The particular network-level outcome we study is homophily, although the method is general to any task that requires estimating dyad-level outcomes. A



We start with a graph of known edges and nodes, G. We also have a feature, X, that is known for each node, in this case numbers between 0 and 1. Each node also has an unobservable color value, Y, of blue or yellow. We are interested in homophily along color, but we cannot readily observe color.

We are able to get the true color of a subset of nodes, L, which are now shown as (blue or yellow) rectangles instead of circles. We see that nodes with higher values of X are more likely to have a color value (Y) of blue instead of yellow.

We use information from our labeled nodes to build a classifier that predicts the color of the node based on its value X. These nodes are now shaded according to their predicted likelihood of being blue or yellow, but their true color remains unknown.

Using our network data and our predictions of group membership (color), we seek to measure homophily, which we operationalize as the average proportions of connections that a node has to other members of the same group. This is nontrivial because the predictions of neighbors (and the residuals of neighbors) will be correlated if the network is at all homophilous.

Figure 1: A summary of the homophily estimation process along a classified variable (Y; represented by color) with a known predictor (X; represented by a number), a subset of labeled nodes (L; represented by rectangles), and a known graph structure (G).

combination of sampling strategy and modeling strategy is *unbiased* when, in expectation, aggregated model predictions $\hat{Y}_j$ constructed from a sample of labeled nodes equal the true value in the network. We will often indicate that an expectation holds for a particular strategy S of sampling the ground truth nodes, notated $\mathbb{E}_S$. The three types of sampling we study are random edge sampling, random node sampling, and proportional-to-degree node sampling.

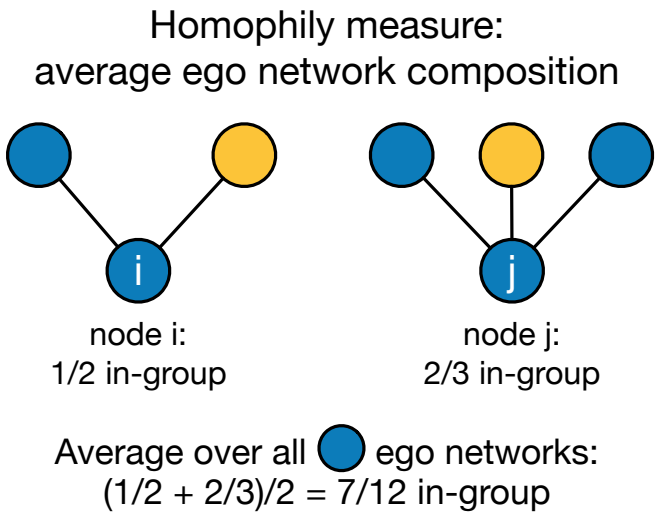


Figure 2: We use average ego network composition to capture homophily from the perspective of the blue nodes. We assume node categories (blue and yellow) must be predicted with a model. This estimand is expressed analytically in Equation (2).

Homophily Measure

We operationalize homophily as the average fraction of ego networks composed of in-group members (visualized in Figure 2). Average ego network composition has been extensively studied in sociology, primarily in research concerning the General Social Survey (GSS) network module (Marsden 1987; McPherson et al. 2006). The average ego network composition measures what the network tends to look like from the perspective of members of a given group. For instance, black respondents to the GSS have been found to have higher average racial heterogeneity in their core discussion networks than white respondents[1] (Marsden 1987).

Average ego network composition can be written as a sum over ego networks, taking into account the category of both ego and alter. Let Y indicate the category of a node; for instance, in the case of racial homophily, category a may indicate black, category b may indicate white, and so on. Assume that we are studying homophily for group a, with $Y_i = 1$ if i is in category a and 0 otherwise. Then the average fraction of group a's ego networks that are composed of alters in group a (Figure 2) can be written as follows:

$$H = T[Y_i]^{-1} \sum_i Y_i \frac{1}{d_i} \sum_{j \in \mathrm{N}(i)} Y_j, \tag{1}$$

where d_i indicates the degree of node i, $\mathrm{N}(i)$ is a function that returns the neighbors of i, and $T[Y_i]$ indicates the size of group a, $\sum_{i \in V} Y_i$. For example, if $H = 0.7$, it means that an average ego network for group a is composed of 70 percent in-group members.

Note that Equation (1) can be rewritten as a sum over all network dyads by rearranging the summation:

$$H = T[Y_i]^{-1} \sum_{(i,j)\in E} \frac{1}{d_i} Y_i Y_j = T[Y_i]^{-1} \sum_{(i,j)\in E} \frac{1}{d_i} Y_{ij}, \tag{2}$$

where E are edges in graph G. Rewriting the edge-level outcome $Y_i Y_j$ as a single random variable Y_{ij} provides an expression of homophily in terms of edge categories. Estimating H can therefore be considered either a node or dyadic prediction task.

Methods of Homophily Estimation

In this section, we study three methods to estimate homophily from predictions. Each method combines a sampling strategy with a modeling strategy. The goal is providing both an unbiased and low-error estimate of Equation (2) using predicted values for Y. The three methods studied are as follows (see Figure 3 for a visualization):

1. **Dyadic**: random dyad sampling with a dyad-level model predicting edge categories directly
2. **Node**: Random node sampling with a node-level model predicting node categories directly
3. **Ego–alter**: Sampling nodes proportional to degree combined with an ego–alter modeling strategy that produces a prediction for ego from the perspective of each alter

Of the three methods, only the dyadic method is unbiased. The node method is closest to what is done in practice and serves as a useful benchmark. The ego–alter method is novel and provides a strategy for homophily estimation in cases where it is infeasible to use the dyadic method directly: it is biased but lower-error than node methods (and sometimes dyadic methods) in simulations. In all three cases, formalization clarifies the sources of bias: sampling inappropriate for generalization to the population of edges, model residual correlation with node degree, and model residual correlation between nodes in a dyad (for instance, adjacent nodes may tend to have positively or negatively correlated residuals).

Throughout, we focus on estimating the part of Equation (2) that concerns dyads ($\sum \frac{1}{D_i} Y_i Y_j$). In other words, we assume the total number of nodes in a group $T[Y_i]$ is known in advance. Estimating the overall number of nodes in a particular group in the network is a quantification problem that has been extensively studied in prior work, from both machine learning and respondent driven sampling perspectives (Salganik and Heckathorn 2004; Forman 2005, 2008; Gao and Sebastiani 2016; González et al. 2017). The consequences of the need to estimate the overall number of nodes in the graph are discussed in the Practical Advice section.

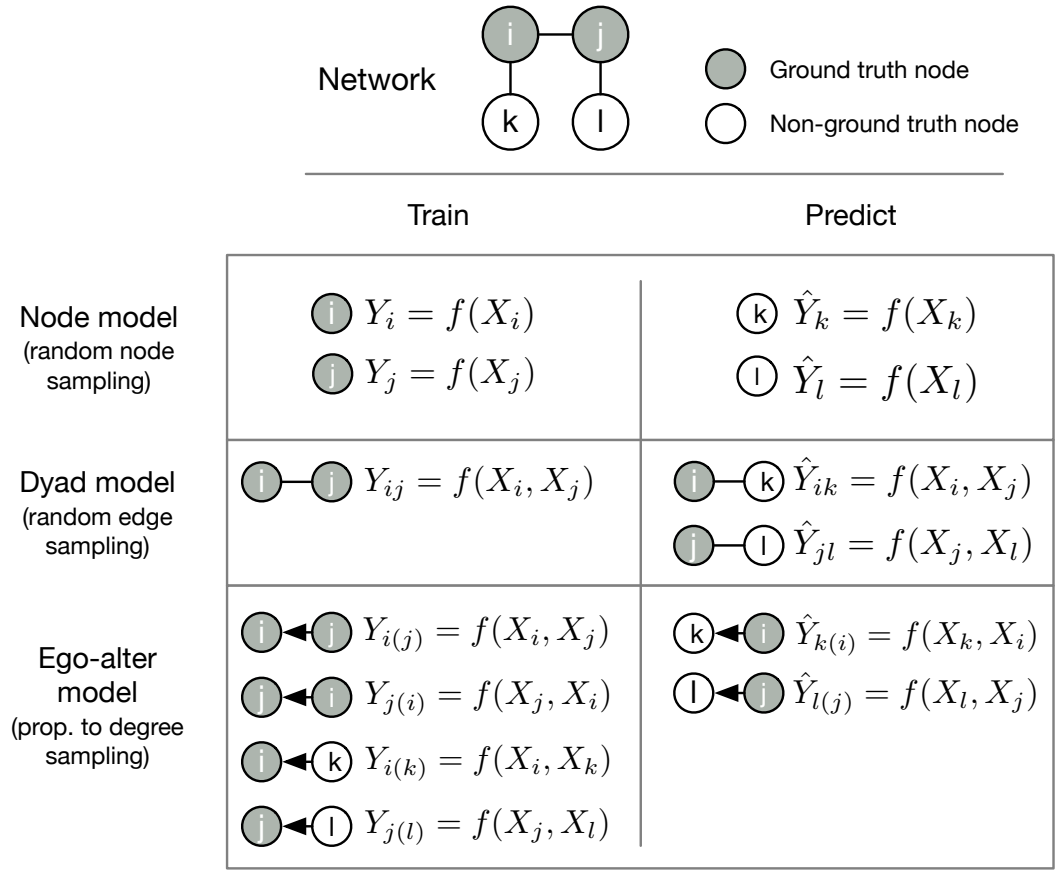


Figure 3: A depiction of how the models studied in this article (node, dyad, and ego–alter) turn a simple network and a ground truth set into a prediction task. The node model predicts each node's category using node covariates X_i. The dyad model predicts dyad categories directly using information on both ego and alter X_i, X_j. The ego–alter model predicts each node category for each neighbor of the focal node, using ego and alter covariates X_i, X_j.

Dyadic Regression as an Unbiased Estimator

A random sample of edges drawn from a graph can be used to produce an unbiased estimate of homophily, although in some cases this estimate can have high overall error. Equation (2) shows how homophily can be estimated with knowledge of edge categories Y_{ij}, where $Y_{ij} = 1$ if $Y_i = Y_j = 1$ and 0 otherwise. Assume a set of edges are drawn randomly for labeling, and predictions are made at the edge level, $\hat{Y}_{ij} = \mathbb{E}_D[Y_{ij}|\frac{1}{d_i}, X_i, X_j]$, where D indicates the expectation is taken with respect to labeling a set of random dyads. $\frac{1}{d_i}$ is the inverse of the degree of node i, and X_i, X_j are the covariates used to predict Y for i and j, respectively. This yields the following estimator for homophily:

$$\hat{H}_{\text{dyadic}} = T[Y_i]^{-1} \sum_{(i,j)\in E} \frac{1}{d_i} \hat{\mathbb{E}}_D \left[Y_{ij} | \frac{1}{d_i}, X_i, X_j \right]. \tag{3}$$

Estimating the expectation with an unbiased model such as ordinary least squares (OLS) then provides an unbiased estimate of homophily $\hat{H}_{\text{dyadic}}$.

To see this, substitute, using the conditional expectation function decomposition (Angrist and Pischke 2009),

$$\hat{\mathbb{E}}_D\left[Y_{ij}|\frac{1}{d_i}, X_i, X_j\right] = \hat{Y}_{ij} = Y_{ij} - e_{ij} \tag{4}$$

into Equation (3) to obtain

$$\hat{H}_{\text{dyadic}} = T[Y_i]^{-1} \sum_{(i,j)\in E} \Bigg(\underbrace{\frac{1}{d_i} Y_{ij}}_{\text{Truevalue}} - \underbrace{\frac{1}{d_i} e_{ij}}_{\text{Biasterm}} \Bigg). \tag{5}$$

This leads to the following condition for unbiasedness.

Condition 1. *When* $\sum_{(i,j)\in E} \mathbb{E}_D[\frac{1}{d_i} e_{ij}] = 0$, *the expectation of the estimate equals the true value:* $\mathbb{E}_D[\hat{H}_{\text{dyadic}}] = H_{\text{dyadic}}$.

The first portion of Condition 1 holds when we use $\frac{1}{d_i}$ as a predictor in the OLS estimation of Y_{ij}. By the principle of mean independence, the residual e_{ij} is uncorrelated with any function of an included regressor.[2] By the definition of uncorrelated random variables, $\mathbb{E}_D[\frac{1}{d_i} e_{ij}] = \mathbb{E}_D[\frac{1}{d_i}]\mathbb{E}_D[e_{ij}]$. $\mathbb{E}_D[e_{ij}] = 0$ also follows from mean independence of the residuals and the fact that an edge sample was used. Therefore $\mathbb{E}_D[\frac{1}{d_i} e_{ij}] = 0$. In sum, the homophily estimate is unbiased when $\frac{1}{d_i}$ is used as a predictor, the ground truth set is a random edge sample, and dyad categories are predicted directly.

This argument concerns model *residuals*, not error terms. No assumptions have been made about causality, true functional form, or predictive accuracy. With a random edge sample and a fitting procedure such as OLS, $\frac{1}{d_i}$ is the only required variable in for an unbiased estimate. However, this can produce a high variance estimate with high average error, which means that including covariates is therefore still important for both variance reduction (along with and addressing cases of nonrandom sampling).

Node Predictions and Bias

Researchers commonly use node predictions to estimate homophily. We examine this approach in this section, assuming that a random node sample is drawn and a model predicting node categories is estimated. In general, node-level models are both biased and in simulations have higher overall error than the other strategies we consider. Bias may be positive or negative, but in the intuitive case where node-level models make similar errors for adjacent nodes, homophily is underestimated. Researchers should therefore expect bias when using node-level models except in rare cases where observable covariates completely explain the pattern of homophily.

Using the identity $Y_{ij} = Y_i Y_j$, homophily can be written in terms of node categories as

$$H_{\text{node}} = T[Y_i]^{-1} \sum_{(i,j)\in E} \frac{1}{d_i} Y_i Y_j,$$

and the estimator using predictions for Y_i, Y_j is written as

$$\hat{H}_{\text{node}} = T[Y_i]^{-1} \sum_{(i,j)\in E} \frac{1}{d_i} \hat{Y}_i \hat{Y}_j. \tag{6}$$

Equation (6) clarifies that when using node predictions to estimate homophily, predictions of adjacent nodes are multiplied. Substituting $\hat{Y}_i = Y_i - e_i$ and $\hat{Y}_j = Y_j - e_j$ yields an expression of the true value plus error:

$$\hat{H}_{\text{node}} = T[Y_i]^{-1} \sum_{(i,j)\in E} \Bigg(\underbrace{\frac{1}{d_i} Y_i Y_j}_{\text{Truevalue}} \underbrace{- \frac{1}{d_i} \hat{Y}_i e_j - \frac{1}{d_i} \hat{Y}_j e_i - \frac{1}{d_i} e_i e_j}_{\text{Biasterms}} \Bigg). \tag{7}$$

The middle two terms $\frac{1}{d_i}\hat{Y}_i e_j$, $\frac{1}{d_i}\hat{Y}_j e_i$ concern predictions correlated with neighbor residuals, whereas the final term $\frac{1}{d_i} e_i e_j$ is the residual correlation term. These terms multiply errors along edges.

Equation (7) provides an unbiasedness condition in the case of random node sampling N for including nodes in the ground truth set.

Condition 2. *When* $\sum_{(i,j)\in E} \mathbb{E}_N[-\frac{1}{d_i}\hat{Y}_i e_j - \frac{1}{d_i}\hat{Y}_j e_i - \frac{1}{d_i} e_i e_j] = 0$, *the expectation of the estimate equals the true value:* $\mathbb{E}_N[\hat{H}_{\text{node}}] = H_{\text{node}}$.

Condition 2 is substantially more complex than Condition 1 for the dyadic model. Bias arises from at least two sources here. First, the nodes included in the ground truth sample are drawn from the population of nodes, but the target population is the edges of the graph. Straightforward sampling bias can occur in this case.

Second, the residual correlation term of Condition 2, $\frac{1}{d_i} e_i e_j$, indicates that the average product of the *unexplained* part of Y for adjacent nodes must be zero. In a standard node-level regression of form $Y_i = \beta X_i + e_i$, there is no guarantee that on average adjacent nodes will have $\mathbb{E}_N[\frac{1}{d_i} e_i e_j] = 0$. This is because it's impossible to control for the neighbor category Y_j or the neighbor residual e_j directly. This information is not available in general, and even if it were it's not clear how to incorporate it into a model that makes a single prediction for each node.

These issues indicate that node sampling and models making a single prediction for each node are not well suited for the homophily estimation task. The intuitive reason is that a method designed to generalize to the population of nodes has trouble generalizing instead to the population of edges.

Ego–Alter Method

In some cases it is infeasible to sample and label dyads directly, and the node model has both bias and potentially large overall error. An intermediate approach can be taken where predictions are made at the edge level, but only for the ego end of the dyad. In other words, for each ground truth node i, a prediction for i is made for

each neighbor j, notated $\hat{Y}_{i(j)}$. These predictions are made for i whether or not j is a ground truth node. (Figure 3 depicts this.)

In this ego–alter method, available information on each neighbor j is included in the model in addition to ego i's information. The ego–alter strategy can produce lower-error estimates than both node and (in some cases) dyadic models while remaining more flexible on the required ground truth data than dyadic models. Because it makes node-level predictions, it is still biased due to residual correlation (the final term in Condition 2).

As with dyad and node methods, the ego–alter method requires thinking carefully about sampling: the natural sampling strategy is sampling nodes proportional to degree. If egos i are chosen for labeling proportional to degree, the total set of edges (i, j) included will approximate an edge sample.[3] Note that in this case, the j's are not labeled, which distinguishes this approach from edge sampling where both ends of each dyad are labeled. How to best account for the correlation within each i is a separate modeling problem that we leave to future work.

Simulation

We now assess the performance of these three strategies for estimating homophily using simulation. We introduce the network generation and evaluation process and then proceed to simulation results. A subset of all results is presented in this section, with full model results in the online supplement.

Network Generation

We simulate graphs where edge creation depends on two factors: the category Y_j of a potential alter j and the degree of the potential alter d_j. This graph generation process simulates the case where social group and network prominence are important factors. Before edge generation, the proportion of each group is chosen. A relationship between Y_i and a single covariate X_i is also fixed in advance. This choice means that correlation of (X_i, X_j) along edges happens only through the category Y and that the category Y has predictive power for association net of the information in X. This matches empirical cases where the category Y is itself something people choose to associate on: political affiliation is an example (Mosleh et al. 2021).

Graph simulations are generated as follows: 4,000 nodes are generated, with 70 percent having $Y = 1$ and 30 percent having $Y = 0$. We study homophily for the $Y = 1$ (majority) group. We then generate X conditional on the Y value, with $X|Y = 1 \sim \mathrm{N}(1, 1)$ and $X|Y = 0 \sim \mathrm{N}(-1, 1)$. This creates an X variable that is highly informative about Y, predicting about 86 percent of the Y values correctly with an AUC of around 0.92.

Given this information on nodes, edges are generated using a modified Barabási–Albert procedure drawing on the graph growing literature (Barabási and Albert 1999; Overgoor, Benson, and Ugander 2019). Each node in the graph gets five bidirected links and decides how to allocate those based on a combination of node

degree and node category. When each new node "arrives" to create its links, it chooses according to the following conditional logit (Overgoor et al. 2019):

$$\mathbb{P}((i,j)=1)=\frac{\exp(\alpha\log d_j+\beta h_{ij})}{\sum_k \exp(\alpha\log d_k+\beta h_{ik})}, \tag{8}$$

where $h_{ij}=1$ if i and j have the same Y value. The $\alpha\log d_j$ term creates preferential attachment, whereas the βh_{ij} term generates homophily when $\beta>0$. We fix $\alpha=0.8$ and study a variety of β values.

Network Sampling and Modeling Methods

Using a ground truth sample of 500 (out of 4,000) nodes to estimate a model, we classify all edges and estimate homophily across 100 simulation runs. Because the focus of this article is on addressing the relational part of the homophily estimation problem, we assume the total fraction of nodes in the majority group is known in advance.

Nodes are sampled according to three separate sampling strategies: random edge sampling, random node sampling, and node sampling proportional to degree. For random edge sampling, random dyads are sampled until 500 distinct nodes have been labeled.

Model performance is evaluated in two ways: bias and absolute error. Bias is the average of $(\hat{H}-H)/H$ across all simulation runs and represents the systematic deviation from the true value. Absolute error is the average absolute error relative to the true underlying value, or the average of $|\hat{H}-H|/H$ across all simulation runs. It captures how far estimates tend to be from the true value. Note that both bias and absolute error are normalized, giving the interpretation of "percent error."

All models are implemented as logistic regression estimating main effects terms for the included covariates, which is described in Table 1.

Main Simulation Results

In this section, we examine simulations with the homophily parameter $\beta=0.7$, which corresponds to a moderate homophily value of around 0.4.[4]

As shown in Figure 4, the default approach of using a node-level classifier with no network features performs poorly. Homophily is underestimated by between five percent and 10 percent, with average absolute error of about the same magnitude. Even when accounting for the inverse degree term $\frac{1}{d_i}$, the node-level approach underestimates homophily by around 7.5 percent.

The dyad model produces an unbiased estimate of homophily in the case of dyad sampling. Overall error is also among the lowest of the models considered, meaning the dyad model is the preferred choice for this particular simulation. The ego–alter model also produces low overall error but tends to underestimate homophily by about 2.5 percent in this particular simulation. Surprisingly, the lowest-error version of the ego–alter model comes with dyad sampling rather than proportional-to-degree sampling, and proportional-to-degree sampling does not produce gains in the case examined here.

Table 1: Models evaluated.

Model name	Dependent variable	Included information
Node (no network)	Y_i	X_i
Node	Y_i	$X_i, \frac{1}{d_i}$
Node (more covariates)	Y_i	$X_i, \frac{1}{d_i}, \log(d_i)$
Dyad (no network)	Y_{ij}	X_i, X_j
Dyad	Y_{ij}	$X_i, X_j, \frac{1}{d_i}$
Dyad (more covariates)	Y_{ij}	$X_i, X_j, \frac{1}{d_i}, \frac{1}{d_j}, \log(d_i), \log(d_j)$
Ego–alter	$Y_{i(j)}$	$X_i, X_j, \frac{1}{d_i}$
Ego–alter (more covariates)	$Y_{i(j)}$	$X_i, X_j, \frac{1}{d_i}, \frac{1}{d_j}, \log(d_i), \log(d_j)$

All models are standard main effects regressions; for instance, the "Node" model is $Y_i = \beta_0 + \beta_1 X_i + \beta_2 \frac{1}{d_i} + e_i$. Ego is indexed by i and alter by j. All models are run for each simulation condition, although only certain models are presented in the main text. The "Node (no network)" model is the baseline approach of a node-level model with no network information included.

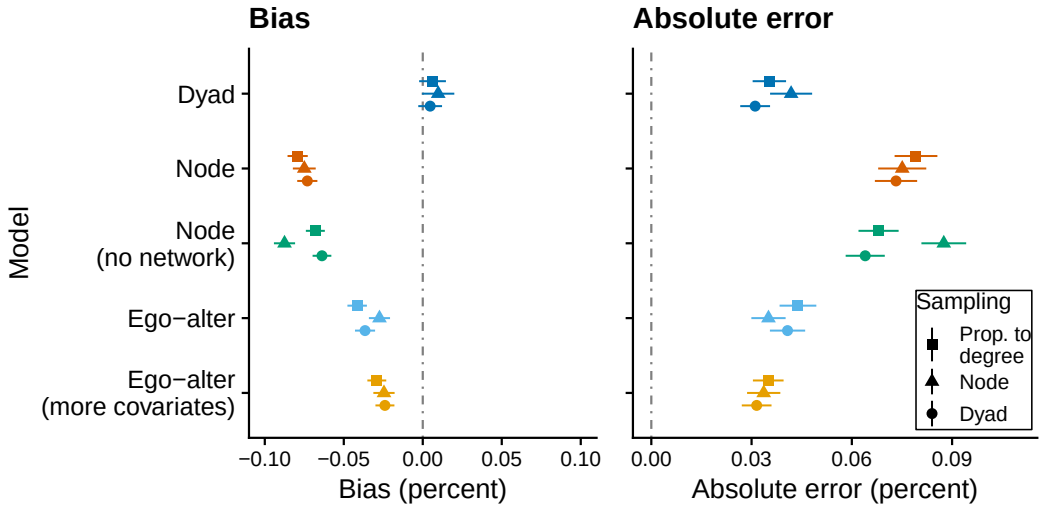


Figure 4: The bias and absolute error of homophily estimates using five different models, across sampling types. Node-level models without network variables display large biases in the presence of network-correlated unobserved features. Including network information reduces this bias, and using edge or ego–alter models reduces this bias further. Note that although dyadic regression is unbiased, it does not provide the lowest-error estimates. Because the same number of nodes is sampled in both edge and node sampling, edge sampling is more efficient for this choice of simulation parameters.

Varying Homophily Strength

Examining the performance of different models while varying simulation parameters provides insight into the sources of both bias and variance when estimating homophily.

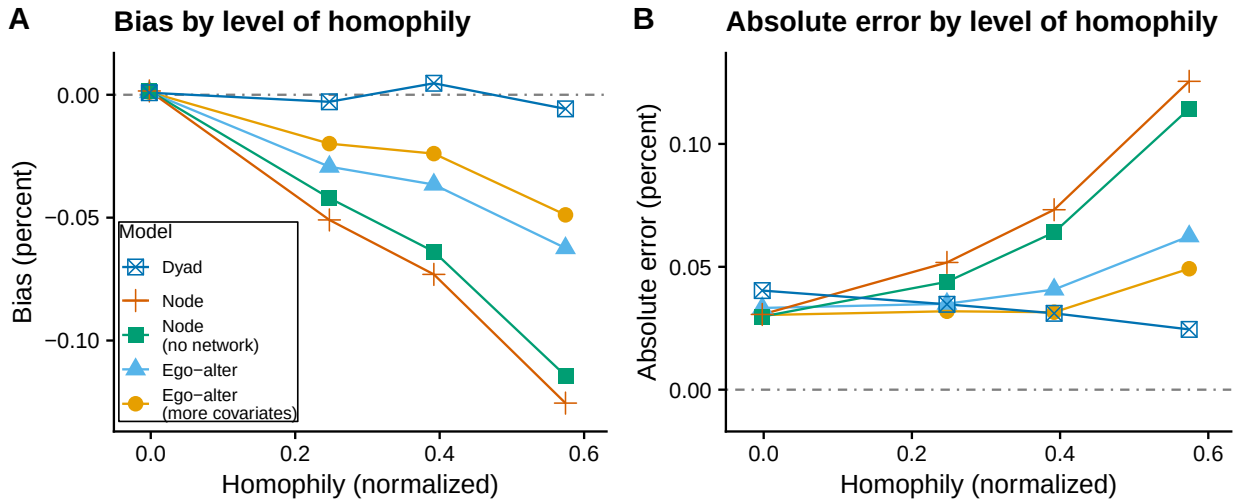


Figure 5: As homophily increases (increasing β in Eq. [8]), bias and absolute error increase for all non–edge-level models. At high homophily values, the edge model becomes the lowest overall error approach while maintaining unbiasedness. This plot displays results for edge sampling, but results are similar for node and proportional-to-degree sampling.

We vary β in Equation (8), which controls the strength of homophily, presenting results in Figure 5. This reveals an important dynamic: as homophily increases, the edge-level model improves in terms of absolute error. In low and moderate homophily environments, the ego–alter model outperforms the edge model in overall error, but as homophily increases the edge model becomes more efficient. Results are similar in the heterophilous case and are presented in Tables 1 and 2 of the online supplement.

As cross-group ties decline in a high homophily environment, it becomes easier for an edge-level model to distinguish between edges that belong to one group or the other. But when there are relatively many cross-group edges, a strategy of predicting dyad categories directly can be noisy.

As homophily increases, the bias of all methods aside from the dyad model show increasing bias. However, the ego–alter models substantially reduce bias over the node models across homophily values.

Node-Level Performance and Network-Level Estimands

Models designed for prediction are usually evaluated on observation-level performance metrics such as accuracy, precision, recall, and AUC. When using predictions to estimate an aggregate such as homophily, strong observation-level performance is encouraging but not sufficient for high-quality estimates of the aggregate. An error-free model will by definition produce a perfect estimate of homophily, but even models with strong out-of-sample observation-level performance can make errors correlated along dyads that bias homophily estimates.

Ego–alter models offer only a slight increase in nodal accuracy compared with node-level models but offer a far more substantial improvement in homophily

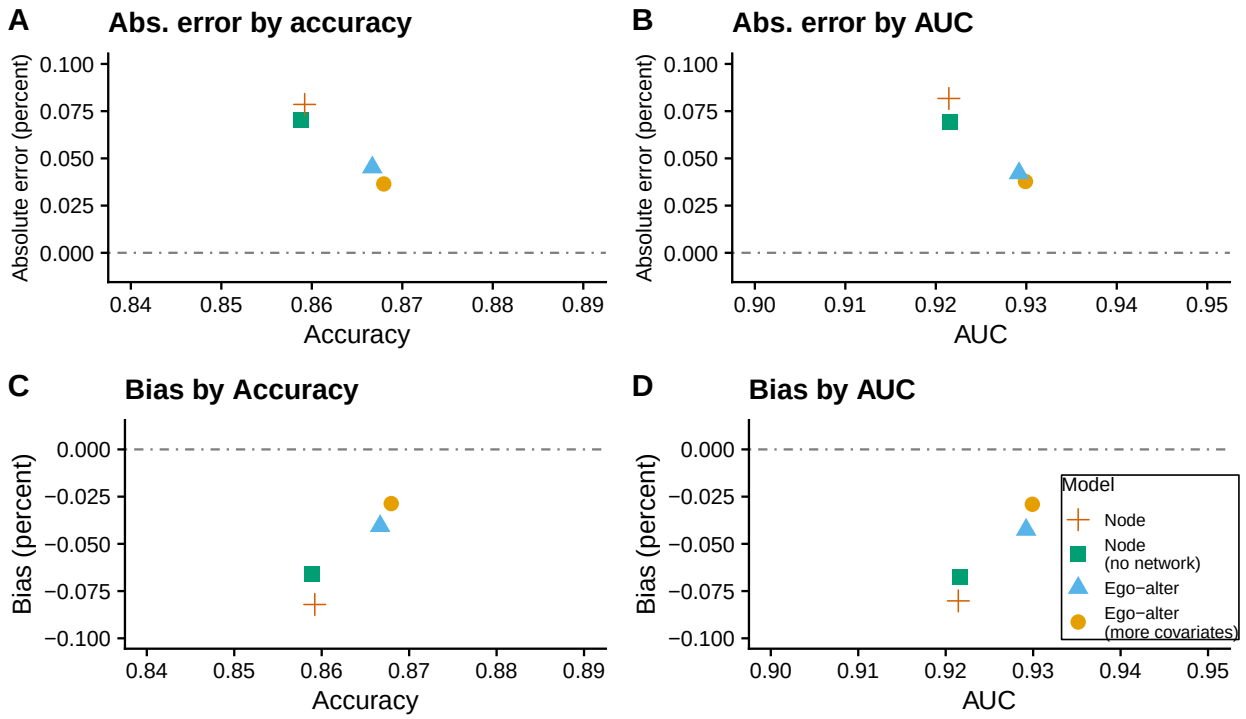


Figure 6: Models with similar node-level classification performance produce different levels of absolute error and bias when estimating homophily using proportional-to-degree sampling. In (*C*), the best model in terms of homophily bias (ego–alter with additional controls; gold) has an average bias of about –2.5 percent and a node-level classification accuracy of 86.7 percent. The worst model in terms of homophily bias (node model; red) has an average bias of about –7.5 percent and a node-level classification accuracy of 85.9 percent. A degradation in classification accuracy of 0.8 percent is associated with tripling the bias when estimating homophily. Although this plot shows results for sampling proportional to degree, results are similar for node and edge sampling and extend to the additional performance metrics of precision and recall.

estimation.[5] This can be seen clearly in Figure 6, which plots node-level model performance against bias in estimating homophily. Models differ only slightly on traditional performance measures yet produce large differences in homophily bias. The best model's accuracy is 0.8 percent better than the worst model yet reduces bias by two-thirds.

A meta-analysis of research on demographic classification on social media (Cesare et al. 2017a) found a median accuracy of 0.81 for predicting race/ethnicity, whereas simulations presented here have an average accuracy of around 0.86. This suggests that similar levels of homophily estimation bias may be found in real-world tasks.

Varying Sources of Homophily and Extra Covariates

Homophily may be generated along the lines of the attribute of interest Y, the predictors of the attribute X, or some combination of both. For example, if wealth is predictive of political affiliation, political homophily may stem from homophily

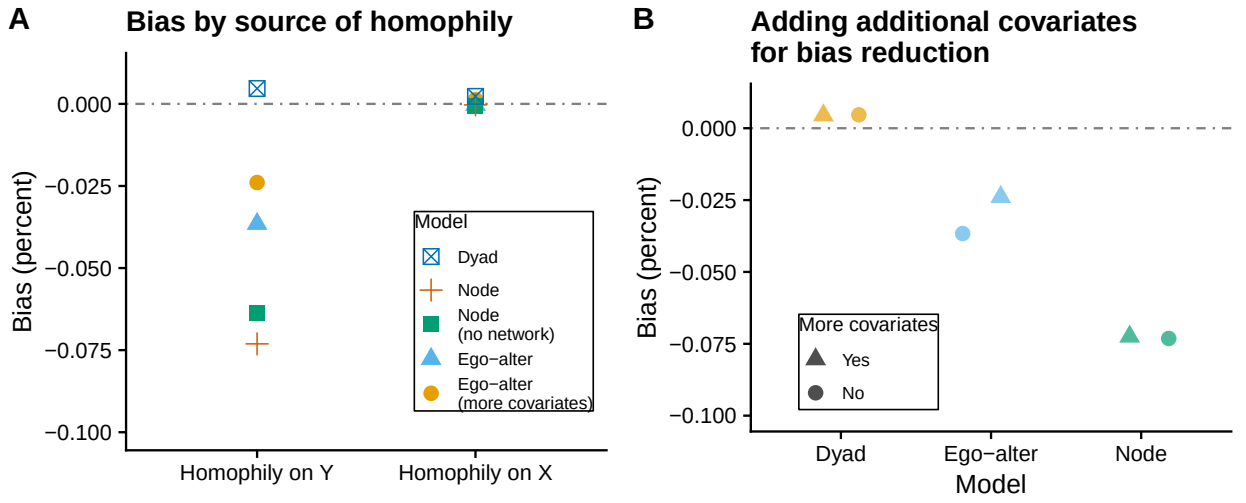


Figure 7: (*A*) shows that the source of homophily determines the level of bias when using a model-based approach. When homophily happens directly on the Y variable (egos choose alters based on Y) and the X variable does not fully capture the variation in Y associated with homophily, bias occurs unless an edge model is used. However, when homophily happens based on the X variable (egos choose alters based on X), no such bias occurs for any of the models examined. Note that the overall level of homophily in the graph is equal for these two cases. (*B*) demonstrates that different modeling strategies respond differently to the inclusion of additional network information (here, log degree). The ego–alter modeling strategy benefits from this additional information. Bias is displayed here, but the pattern is similar for absolute error: the ego–alter model shows gains from additional information where node and edge models do not. These plots show edge sampling, but results are similar for node and proportional-to-degree sampling.

along the lines of either political affiliation, wealth, or both. We examine this by altering a fundamental assumption of the simulation by modifying Equation (8) so that homophilious links are chosen on X rather than Y:

$$\mathbb{P}((i,j)=1) = \frac{\exp(\alpha \log d_j + \gamma x_j)}{\sum_k \exp(\alpha \log d_k + \gamma x_j)}, \tag{9}$$

where $\gamma = 0.35$ produces a homophily value similar to a simulation with $\beta = 0.7$. The results of this are presented in Figure 7(A). Homophily bias is zero in all cases when homophily is completely determined by observable covariates X. This demonstrates that the source of bias in estimating homophily comes from variation in Y, which is not explained by X but remains predictive of homophily. Unless the homophily observed along the attribute Y is completely explained by X, bias will arise. In practical terms, researchers are often interested in homophily along attributes such as political affiliation where research indicates that people associate directly on the attribute (Mosleh et al. 2021). In this case, bias is a concern in practice.

The primary simulation case we study, where $\beta > 0, \gamma = 0$, is an extreme case where homophily is completely determined by the Y value itself. In reality, there is likely a mix of both factors.

Finally, we examine whether additional network-based control variables reduce overall error when estimating homophily. This is operationalized by including the

log of ego degree in the node model and both the log of ego degree and the log of neighbor degree in the edge and ego–alter models. Figure 7(B) indicates that only the ego–alter model benefits from this additional network information, suggesting an interaction between model form and included information.

Practical Advice

When studying homophily using predictions, researchers can adopt several practices to improve estimates: include network information in models, sample edges and predict dyad categories directly (or where this is infeasible, use the ego–alter modeling strategy), and use cross-validation to check for the presence of network residual correlation.

The task of homophily estimation can be viewed as predicting the categories of edges in a graph. When feasible, research design should match this goal: a random sample of edges and a dyad-level model including $\frac{1}{d_i}$ as a covariate produces an unbiased estimate. This strategy also has among the lowest overall errors when homophily is moderate or greater (0.2 or higher in simulations studied here) and performs the best at high homophily levels.

When it is infeasible to sample and label dyads directly, the ego–alter approach provides a better strategy than a node-level model. It's able to account for some information at the dyadic level directly without requiring a set of labeled dyads. Although biased, in all cases studied here the ego–alter approach provides a large bias reduction over a standard node-level model and provides among the lowest overall errors of any approach at low to moderate homophily levels. Modeling strategies that propagate residuals in the network (Jia and Benson 2020) may provide additional error reduction.

When using node-level models, researchers can obtain an estimate of network residual correlation by using cross-validation (see the discussion in Molina and Garip [2019] for a brief introduction to cross-validation; see Chapter 7 of Hastie, Tibshirani, and Friedman [2008] for a more extensive discussion). Cross-validation splits the training data into a number of folds (usually five or 10) and uses all but one fold to train a model, with the held-out fold used to evaluate the model. This proceeds in a round-robin fashion so that the entire training set is scored in a way that approximates out-of-sample prediction. In the context of homophily estimation, estimating the residual term in Equation (7) can provide important information about network residual correlation. This can be accomplished in a cross-validation setting by dividing up all dyads in the training set into folds and examining the three bias terms in Equation (7). This strategy does not ensure unbiased homophily estimates, particularly in the presence of nonrandom ground truth sampling, but it does provide a potentially useful diagnostic.

Throughout, we have focused on the dyadic part of homophily in Equation (2). There is also the task of estimating the total number of nodes in a given group, represented by the $T[Y_i]$ term. The straightforward way to do this is to draw a node sample and fit a node-level model predicting node categories. If proportional-to-degree sampling is done, nodes are drawn with probability $\frac{d_i}{|E|}$, and standard inverse probability weighting can be used to construct a node-level model that

generalizes to the population of nodes. If edge sampling is performed, standard respondent driven sampling results can be used to estimate the proportion of nodes in each group (see Eq. [29] of Salganik and Heckathorn [2004]).

Conclusion

We have examined the problem of estimating homophily when predictions must be used for node attributes. Although the problem is challenging, the results we present indicate that homophily can be studied in online networks when careful attention is given to sampling and modeling.

The strategies outlined here also provide a pathway for the measurement of other network-level properties. Examples are other measures of homophily such as Coleman's homophily index (Coleman 1958) (see section 1.3 of the online supplement for discussion) or measures of negative experiences such as the amount of hate speech viewed by a particular demographic group (Davidson, Bhattacharya, and Weber 2019). In studies of dynamic network processes such as contagion, models to reduce measurement error (Berry et al. 2019) may benefit from the results here. In the case of signed or multiplex networks, the distribution of different types of edges across groups may be important. Similarly to homophily estimation, consideration of how model errors intersect with graph properties is important for reliable use of predictions in network contexts.

Results here intersect with machine learning research. Machine learning strategies may provide methods to further reduce bias and overall error relative to those we have explored. Node embeddings (Perozzi and Skiena 2014; Grover and Leskovec 2016; Kipf and Welling 2016; Hamilton, Ying, and Leskovec 2017) may provide a modeling framework to improve the quality of models predicting node or edge categories. It is an open question if advanced modeling techniques such as graph neural networks can provide better estimates than the standard regression-based methods we have explored. One strategy that may provide benefits is joint node and edge prediction (Gong et al. 2014). Jia and Benson (2020) have examined a strategy for propagating node-level classification errors along edges, which could also provide a methodology for reducing bias. As examined by Stamm et al. (2020), improving model predictions can reduce edge uncertainty and in turn clarify important network questions.

Finally, results presented here concerning model errors correlated along dyads clarify the importance of focusing on dyad-level classification when dyad-level outcomes are of interest. When predictions are aggregated in a network, it is not enough to predict nodes well. Advanced modeling strategies predicting node-level attributes do not solve the fundamental problem with node models identified in this article: errors correlated along dyads threaten the validity of aggregates. The importance of sampling strategy is also clarified in our examination of sampling bias. The results presented here suggest that additional research on these topics could benefit both our understanding of social relationships and the use of prediction methods in networks.

Notes

1 We choose this measure of homophily instead of Coleman's homophily measure (Coleman 1958) because it is not dominated by high degree nodes, although the online supplement shows that results for the average ego network measure apply to Coleman's measure as well.

2 See, for instance, Angrist and Pischke (2009:25) for details.

3 If i is sampled proportional to $\frac{d_i}{|E|}$ without replacement, then the probability that any link is sampled, $p(i,j)$, is the sum of the mutually exclusive events that any given link from i is sampled: $p(i,j) = \sum_j \frac{d_i}{|E|}\frac{1}{d_i} = d_i \cdot \frac{d_i}{|E|}\frac{1}{d_i} = \frac{1}{|E|}$.

4 We normalize homophily between -1 and 1 using the process described in section 1.1 of the online supplement.

5 Only models that produce node-level category predictions can be evaluated this way, meaning the edge model cannot be used here.

References

Aggarwal, Charu, Gewen He, and Peixiang Zhao. 2016. "Edge Classification in Networks." In *Proceedings of the 32nd International Conference on Data Engineering*, pp. 1038–49. New York: IEEE.

Al Zamal, Faiyaz, Wendy Liu, and Derek Ruths. 2012. "Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors." In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*. Palo Alto, CA: AAAI.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Bakshy, Eytan, Solomon Messing, and Lada A. Adamic. 2015. "Exposure to Ideologically Diverse News and Opinion on Facebook." *Science* 348(6239):1130–32. https://doi.org/10.1126/science.aaa1160.

Barabási, Albert-László, and Réka Albert. 1999. "Emergence of Scaling in Random Networks." *Science* 286(5439):509–12. https://doi.org/10.1126/science.286.5439.509.

Barberá, Pablo. 2016. "Less Is More? How Demographic Sample Weights Can Improve Public Opinion Estimates Based on Twitter Data." Working Paper, Center for Data Science, New York University. http://www.pablobarbera.com/static/less-is-more.pdf.

Barberá, Pablo, John T. Jost, Jonathan Nagler, Joshua A. Tucker, and Richard Bonneau. 2015. "Tweeting from Left to Right: Is Online Political Communication More Than an Echo Chamber?" *Psychological Science* 26(10):1531–42. https://doi.org/10.1177%2F0956797615594620.

Berry, George, Christopher J. Cameron, Patrick Park, and Michael W. Macy. 2019. "The Opacity Problem in Social Contagion." *Social Networks* 56:93–101. https://doi.org/10.1016/j.socnet.2018.09.001.

Berry, George, Antonio Sirianni, Nathan High, Agrippa Kellum, Ingmar Weber, and Michael Macy. 2018. "Estimating Group Properties in Online Social Networks with a Classifier." In *Social Informatics: 10th International Conference, SocInfo 2018, St. Petersburg, Russia, September 25–28, 2018, Proceedings, Part I*, edited by Steffen Staab, Olessia Koltsova, and Dmitry I. Ignatov, pp. 67–85. Cham, Switzerland: Springer.

Blau, Peter M. 1977. "A Macrosociological Theory of Social Structure." *American Journal of Sociology* 83(1):26–54. https://doi.org/10.1086/226505.

Boutyline, Andrei, and Robb Willer. 2017. "The Social Structure of Political Echo Chambers: Variation in Ideological Homophily in Online Networks." *Political Psychology* 38(3):551–69. http://doi.wiley.com/10.1111/pops.12337.

Cesare, Nina, Christan Grant, Quynh Nguyen, Hedwig Lee, and Elaine O. Nsoesie. 2017a. "How Well Can Machine Learning Predict Demographics of Social Media Users?" Preprint, arXiv:1702.01807 [cs.SI]. http://arxiv.org/abs/1702.01807.

Cesare, Nina, Hedwig Lee, Tyler McCormick, and Emma S. Spiro. 2017b. "Redrawing the 'Color Line': Examining Racial Segregation in Associative Networks on Twitter." Preprint, arXiv:1705.04401 [cs.SI]. http://arxiv.org/abs/1705.04401.

Coleman, James S. 1958. "Relational Analysis: The Study of Social Organizations with Survey Methods." *Human Organization* 17(4):28–36.

Colleoni, Elanor, Alessandro Rozza, and Adam Arvidsson. 2014. "Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data." *Journal of Communication* 64(2):317–32. https://doi.org/10.1111/jcom.12084.

Davidson, Thomas, Debasmita Bhattacharya, and Ingmar Weber. 2019. "Racial Bias in Hate Speech and Abusive Language Detection Datasets." In *Proceedings of the Third Workshop on Abusive Language Online*, pp. 25–35. Stroudsburg, PA: Association for Computational Linguistics.

De Choudhury, Munmun. 2011. "Tie Formation on Twitter: Homophily and Structure of Egocentric Networks." In *Proceedings of the Third International Conference on Privacy, Security, Risk and Trust and Third International Conference on Social Computing*, pp. 465–70. New York: IEEE.

DiPrete, Thomas A., Andrew Gelman, Tyler McCormick, Julien Teitler, and Tian Zheng. 2011. "Segregation in Social Networks Based on Acquaintanceship and Trust." *American Journal of Sociology* 116(4):1234–83. https://doi.org/10.1086/659100.

Forman, George. 2005. "Counting Positives Accurately despite Inaccurate Classification." In *European Conference on Machine Learning: ECML 2005*, pp. 564–75. Berlin: Springer.

Forman, George. 2008. "Quantifying Counts and Costs via Classification." *Data Mining and Knowledge Discovery* 17:164–206. https://doi.org/10.1007/s10618-008-0097-y.

Gao, Wei, and Fabrizio Sebastiani. 2016. "From Classification to Quantification in Tweet Sentiment Analysis." *Social Network Analysis and Mining* 6:19. https://doi.org/10.1007/s13278-016-0327-z.

Gong, Neil Zhenqiang, Ameet Talwalkar, Lester Mackey, Ling Huang, Eui Chul Richard Shin, Emil Stefanov, Elaine (Runting) Shi, and Dawn Song. 2014. "Joint Link Prediction and Attribute Inference Using a Social-Attribute Network." *ACM Transactions on Intelligent Systems and Technology* 5(2):27. https://doi.org/10.1145/2594455.

González, Pablo, Alberto Castaño, Nitesh V. Chawla, and Juan José Del Coz. 2017. "A Review on Quantification Learning." *ACM Computing Surveys* 50(5):74. https://doi.org/10.1145/3117807.

Grover, Aditya, and Jure Leskovec. 2016. "node2vec: Scalable Feature Learning for Networks." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 855–64. New York: Association for Computer Machinery.

Halberstam, Yosh, and Brian Knight. 2016. "Homophily, Group Size, and the Diffusion of Political Information in Social Networks: Evidence from Twitter." *Journal of Public Economics* 143:73–88. https://doi.org/10.1016/j.jpubeco.2016.08.011.

Hamilton, Will, Zhitao Ying, and Jure Leskovec. 2017. "Inductive Representation Learning on Large Graphs." In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 1025–35. New York: Association for Computing Machinery.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2008. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer. https://doi.org/10.1007/978-0-387-84858-7.

Himelboim, Itai, Kaye D. Sweetser, Spencer F. Tinkham, Kristen Cameron, Matthew Danelo, and Kate West. 2016. "Valence-Based Homophily on Twitter: Network Analysis of Emotions and Political Talk in the 2012 Presidential Election." *New Media & Society* 18(7):1382–1400. https://doi.org/10.1177%2F1461444814555096.

Hobbs, William R., Moira Burke, Nicholas A. Christakis, and James H. Fowler. 2016. "Online Social Integration Is Associated with Reduced Mortality Risk." *Proceedings of the National Academy of Sciences* 113(46):12980–84. https://doi.org/10.1073/pnas.1605554113.

Hofstra, Bas, Rense Corten, Frank van Tubergen, and Nicole B. Ellison. 2017. "Sources of Segregation in Social Networks: A Novel Approach Using Facebook." *American Sociological Review* 82(3):625–56. https://doi.org/10.1177%2F0003122417705656.

Hofstra, Bas, and Niek C. de Schipper. 2018. "Predicting Ethnicity with First Names in Online Social Media Networks." *Big Data & Society* 5(1). https://doi.org/10.1177%2F2053951718761141.

Jia, Junteng, and Austion R. Benson. 2020. "Residual Correlation in Graph Neural Network Regression." In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 588–98. New York: Association for Computing Machinery.

Karimi, Fariba, Mathieu Génois, Claudia Wagner, Philipp Singer, and Markus Strohmaier. 2018. "Homophily Influences Ranking of Minorities in Social Networks." *Scientific Reports* 8:11077. https://doi.org/10.1038/s41598-018-29405-7.

Kipf, Thomas N., and Max Welling. 2016. "Semi-supervised Classification with Graph Convolutional Networks." Preprint, arXiv:1609.02907 [cs.LG]. http://arxiv.org/abs/1609.02907.

Kossinets, Gueorgi, and Duncan J. Watts. 2009. "Origins of Homophily in an Evolving Social Network." *American Journal of Sociology* 115(2):405–50. https://doi.org/10.1086/599247.

Lee, Eun, Fariba Karimi, Claudia Wagner, Hang-Hyun Jo, Markus Strohmaier, and Mirta Galesic. 2019. "Homophily and Minority-Group Size Explain Perception Biases in Social Networks." *Nature Human Behaviour* 3:1078–87. https://doi.org/10.1038/s41562-019-0677-4.

Lewis, Kevin, Marco Gonzalez, and Jason Kaufman. 2012. "Social Selection and Peer Influence in an Online Social Network." *Proceedings of the National Academy of Sciences* 109:68–72. https://doi.org/10.1073/pnas.1109739109.

Marsden, Peter V. 1987. "Core Discussion Networks of Americans." *American Sociological Review* 52(1):122–31. https://doi.org/10.2307/2095397.

McPherson, Miller, Lynn Smith-Lovin, and Matthew E. Brashears. 2006. "Social Isolation in America: Changes in Core Discussion Networks over Two Decades." *American Sociological Review* 71(3):353–75. https://doi.org/10.1177%2F000312240607100301.

McPherson, Miller, Lynn Smith-Lovin, and James M. Cook. 2001. "Birds of a Feather: Homophily in Social Networks." *Annual Review of Sociology* 27:415–44. https://doi.org/10.1146/annurev.soc.27.1.415.

Messias, Johnnatan, Pantelis Vikatos, and Fabrício Benevenuto. 2017. "White, Man, and Highly Followed: Gender and Race Inequalities in Twitter." In *Proceedings of the International Conference on Web Intelligence*, pp. 266–74. New York: Association for Computing Machinery.

Molina, Mario, and Filiz Garip. 2019. "Machine Learning for Sociology." *Annual Review of Sociology* 45:27–45. https://doi.org/10.1146/annurev-soc-073117-041106.

Mollica, Kelly A., Barbara Gray, and Linda K. Treviño. 2003. "Racial Homophily and Its Persistence in Newcomers' Social Networks." *Organization Science* 14(2):123–36.

Mosleh, Mohsen, Cameron Martel, Dean Eckles, and David Gertler Rand. 2021. "Shared Partisanship Dramatically Increases Social Tie Formation in a Twitter Field Experiment." *Proceedings of the National Academy of Sciences* 118(7):e2022761118. https://doi.org/10.1073/pnas.2022761118.

Overgoor, Jan, Austin R. Benson, and Johan Ugander. 2019. "Choosing to Grow a Graph: Modeling Network Formation as Discrete Choice." In *WWW '19: The World Wide Web Conference*, pp. 1409–20. New York: Association for Computing Machinery.

Perozzi, Bryan, and Steven Skiena. 2014. "DeepWalk: Online Learning of Social Representations." In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 701–10. New York: Association for Computating Machinery.

Salganik, Matthew J., and Douglas D. Heckathorn. 2004. "Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling." *Sociological Methodology* 34(1):193–240. https://doi.org/10.1111%2Fj.0081-1750.2004.00152.x.

Smith, Jeffrey A., Miller McPherson, and Lynn Smith-Lovin. 2014. "Social Distance in the United States: Sex, Race, Religion, Age, and Education Homophily among Confidants, 1985 to 2004." *American Sociological Review* 79(3):432–56. https://doi.org/10.1177%2F0003122414531776.

Stamm, Felix I., Leonie Neuhäuser, Florian Lemmerich, Michael T. Schaub, and Markus Strohmaier. 2020. "Systematic Edge Uncertainty in Attributed Social Networks and Its Effects on Rankings of Minority Nodes." Preprint, arXiv:2010.11546 [cs.SI]. http://arxiv.org/abs/2010.11546.

Thelwall, Mike. 2009. "Homophily in MySpace." *Journal of the American Society for Information Science and Technology* 60(2):219–31. https://doi.org/10.1002/asi.20978.

Wimmer, Andreas, and Kevin Lewis. 2010. "Beyond and below Racial Homophily: ERG Models of a Friendship Network Documented on Facebook." *American Journal of Sociology* 116(2):583–642. https://doi.org/10.1086/653658.

Acknowledgments: We thank Thomas Davidson, Mario Molina, Pablo Barberá, Christopher Cameron, Rebecca A. Johnson, Benjamin Cornwell, and Steven Strogatz; participants in the 2020 American Sociological Association section on Mathematical Sociology; the members of the Cornell Social Dynamics Lab; and the members of the Dartmouth Junior Faculty Writing Group for helpful comments and discussions.

George Berry: Department of Sociology, Cornell University. E-mail: geb97@cornell.edu.

Antonio Sirianni: Department of Sociology, Dartmouth College.
E-mail: antonio.d.sirianni@dartmouth.edu.

Ingmar Weber: Qatar Computing Research Institute. E-mail: iweber@hbku.edu.qa.

Jisun An: School of Computer and Information Systems, Singapore Management University. E-mail: jisun.an@acm.org.

Michael Macy: Department of Sociology, Cornell University. E-mail: mwm14@cornell.edu.