Supplement to:

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.

1 Appendix

1.1 Full simulation results

Tables 1 and 2 present averages from 100 simulation runs for all cases examined.

The simulation conditions correspond to the following parameter sets: no homophily: $\beta = 0, \gamma = 0$; low homophily: $\beta = 0.4, \gamma = 0$; main simulations: $\beta = 0.7, \gamma = 0$; homophily on X: $\beta = 0, \gamma = 0.35$; high homophily: $\beta = 1.2, \gamma = 0$; low heterophily: $\beta = -0.4, \gamma = 0$; medium heterophily: $\beta = -0.7, \gamma = 0$; heterophily on X: $\beta = 0, \gamma = -0.35$; high heterophily: $\beta = -1.2, \gamma = 0$. β determines homophily on the Y variable, while γ determines homophily on the X variable.

Homophily is measured in relative terms. We compute the average fraction of ingroup nodes for the majority group $H_{Y=1}$, and then normalize by group size $P_{Y=1}$:

$$M_{Y=1} = \frac{H_{Y=1} - P_{Y=1}}{1 - P_{Y=1}} \tag{1}$$

Since all homophily values are zero or positive by construction this number ranges from (approximately) 0 to 1, with 0 indicating "no additional ingroup association over random mixing" and 1 indicating "complete ingroup association".

1.2 Example extension: incorporating other factors into the estimand

Researchers often care about actions in addition to node characteristics. For instance, what is the fraction of content seen broken down by gender of the content author? Addressing this question is important for examining visibility (Karimi et al. 2018) by gender online, and requires combining information about node gender and action.

In this case, Equation (2) in the main text is modified with a variable A_j indicating some action of alter j. Assume A_j represents number of messages sent by j, and $Y_i = a$ indicates that i is a woman.

$$H_{\text{extended}} = T[Y_i]^{-1} \sum_{(i,j)\in E} \frac{A_j}{d_i} Y_i Y_j. \tag{2}$$

In words, Equation 2 represents the average fraction of messages seen by members of group a which come from members of group a. When incorporating the additional variable A_j into the equation, we can apply the same logic as discussed in Equations (3)–(5) in the main text to obtain an unbiased estimate: incorporate $\frac{A_j}{d_i}$ into the predictive model, instead of $\frac{1}{d_i}$ alone.

1.3 Example extensions: Coleman's homophily index

We studied average egonet composition in the main text, but another popular measure of homophily is Coleman's homophily index (Coleman 1958). This measure studies the fraction of within group links from the perspective of a certain group, relative to the proportion expected by chance.

The challenge is estimating the proportion of within-group links from the perspective of a given group a. This can be done in a manner similar to Equation (2) in the main text,

$$H_{\text{Coleman}} = T[Y_i]^{-1} \sum_{(i,j)\in E} Y_i Y_j. \tag{3}$$

This turns out to be a simpler version of the egonet estimand considered in the main text, and can be addressed with similar modeling strategies.

Sampling	Simulation	Maj. grp. homophily	No model	Dyad	Dyad (no network)	Dyad (more covariates)	Node (no network)	Node	Node (more covariates)	Ego-alter	Ego-alter (more covariates)
Dyad	No homophily	-0.002	5.5%	4.0%	3.8%	4.0%	3.0%	3.1%	3.0%	3.3%	3.0%
	Low homophily	0.247	5.5%	3.5%	3.1%	3.6%	4.4%	5.2%	5.2%	3.5%	3.2%
	Main simulations	0.392	4.3%	3.1%	3.1%	3.1%	6.4%	7.3%	7.3%	4.1%	3.2%
	Homophily on X	0.378	4.4%	3.3%	3.2%	3.4%	2.5%	2.7%	2.7%	2.9%	2.7%
	High homophily	0.574	4.1%	2.5%	2.5%	2.4%	11.4%	12.5%	12.5%	6.2%	4.9%
	Low heterophily	-0.138	5.6%	4.6%	4.6%	4.6%	6.3%	7.6%	7.5%	5.6%	4.1%
	Medium heterophily	-0.258	7.1%	5.3%	5.4%	5.1%	12.8%	14.7%	14.5%	9.1%	4.8%
	Heterophily on X	-0.306	8.3%	5.2%	5.3%	4.9%	3.3%	3.3%	3.3%	3.3%	3.4%
	High heterophily	-0.463	11.0%	8.1%	8.5%	7.9%	31.8%	34.1%	33.2%	20.2%	8.9%
Node	No homophily	-0.002	7.8%	5.3%	6.0%	5.2%	2.9%	3.2%	3.2%	3.3%	3.3%
	Low homophily	0.247	6.7%	4.6%	4.6%	4.6%	5.9%	5.3%	5.0%	3.1%	3.1%
	Main simulations	0.392	6.9%	4.2%	4.7%	4.2%	8.8%	7.5%	7.1%	3.5%	3.4%
	Homophily on X	0.378	5.9%	4.5%	4.7%	4.3%	2.3%	2.4%	2.5%	2.5%	2.5%
	High homophily	0.574	6.4%	3.9%	4.0%	3.9%	13.1%	11.9%	11.4%	4.7%	4.5%
	Low heterophily	-0.138	8.9%	5.6%	6.2%	5.5%	9.2%	7.6%	7.0%	4.6%	4.3%
	Medium heterophily	-0.258	9.4%	6.6%	7.6%	6.9%	20.7%	16.2%	14.6%	7.3%	6.0%
	Heterophily on X	-0.306	12.8%	6.7%	7.2%	6.6%	3.7%	3.9%	4.1%	4.3%	4.4%
	High heterophily	-0.463	10.9%	7.7%	9.3%	8.0%	51.5%	38.7%	33.7%	15.6%	10.6%
Prop. to degree	No homophily	-0.002	36.7%	4.5%	5.4%	5.0%	3.0%	3.2%	3.2%	3.3%	3.2%
	Low homophily	0.247	34.8%	3.3%	5.0%	3.9%	4.9%	5.6%	5.5%	4.0%	3.3%
	Main simulations	0.392	33.6%	3.5%	5.0%	4.0%	6.8%	7.9%	7.9%	4.4%	3.5%
	Homophily on X	0.378	29.7%	6.3%	4.8%	6.1%	2.4%	2.5%	2.5%	2.6%	2.6%
	High homophily	0.574	33.1%	3.3%	5.7%	3.7%	10.7%	11.9%	11.8%	5.7%	4.4%
	Low heterophily	-0.138	42.3%	5.9%	8.9%	6.5%	5.9%	7.0%	7.0%	4.8%	3.9%
	Medium heterophily	-0.258	47.9%	7.5%	14.7%	7.6%	12.8%	15.2%	15.1%	9.9%	5.5%
	Heterophily on X	-0.306	62.6%	9.1%	8.5%	9.3%	3.4%	3.4%	3.4%	3.3%	3.3%
	High heterophily	-0.463	55.4%	15.7%	30.0%	11.4%	31.1%	33.3%	32.4%	19.1%	8.7%

Table 1: Mean absolute error by simulation

Sampling	Simulation	Maj. grp. homophily	No model	Dyad	Dyad (no network)	Dyad (more covariates)	Node (no network)	Node	Node (more covariates)	Ego-alter	Ego-alter (more covariates)
Dyad	No homophily	-0.002	0.6%	0.1%	-0.1%	0.0%	0.2%	0.2%	0.1%	0.1%	0.1%
	Low homophily	0.247	0.0%	-0.3%	0.7%	-0.3%	-4.2%	-5.1%	-5.1%	-2.9%	-2.0%
	Main simulations	0.392	1.9%	0.5%	1.5%	0.4%	-6.4%	-7.3%	-7.3%	-3.7%	-2.4%
	Homophily on X	0.378	0.6%	0.2%	-0.3%	0.3%	-0.1%	0.0%	-0.1%	0.0%	0.1%
	High homophily	0.574	-0.7%	-0.6%	0.5%	-0.6%	-11.4%	-12.5%	-12.5%	-6.2%	-4.9%
	Low heterophily	-0.138	0.8%	0.5%	-1.6%	0.5%	6.1%	7.5%	7.4%	4.7%	2.4%
	Medium heterophily	-0.258	0.8%	0.6%	-2.8%	0.6%	12.8%	14.7%	14.5%	9.0%	4.0%
	Heterophily on X	-0.306	-1.0%	-1.2%	-1.8%	-1.2%	0.1%	0.1%	0.2%	0.2%	0.2%
	High heterophily	-0.463	1.2%	0.2%	-6.2%	0.3%	31.8%	34.1%	33.2%	20.2%	8.3%
Node	No homophily	-0.002	0.0%	0.2%	0.4%	0.2%	-0.7%	-0.6%	-0.6%	-0.6%	-0.6%
	Low homophily	0.247	0.7%	-0.1%	0.8%	-0.1%	-5.8%	-5.0%	-4.7%	-2.0%	-1.7%
	Main simulations	0.392	1.4%	1.0%	2.2%	1.0%	-8.8%	-7.5%	-7.1%	-2.7%	-2.5%
	Homophily on X	0.378	-1.1%	-1.0%	-1.8%	-0.7%	0.1%	-0.1%	-0.2%	-0.2%	-0.1%
	High homophily	0.574	0.1%	0.7%	1.9%	0.6%	-13.1%	-11.9%	-11.4%	-4.6%	-4.3%
	Low heterophily	-0.138	-1.1%	-1.0%	-2.5%	-1.0%	9.0%	7.2%	6.5%	2.7%	2.0%
	Medium heterophily	-0.258	0.8%	0.0%	-4.0%	0.2%	20.7%	16.2%	14.6%	6.6%	4.7%
	Heterophily on X	-0.306	-1.6%	-1.0%	-1.7%	-1.4%	-0.4%	-1.2%	-1.0%	-0.9%	-0.9%
	High heterophily	-0.463	1.9%	0.9%	-6.0%	1.1%	51.5%	38.7%	33.7%	15.4%	10.2%
Prop. to degree	No homophily	-0.002	-36.7%	0.5%	0.1%	0.6%	0.1%	0.0%	0.0%	-0.1%	0.0%
	Low homophily	0.247	-34.8%	0.3%	3.2%	-0.3%	-4.5%	-5.2%	-5.2%	-3.1%	-2.0%
	Main simulations	0.392	-33.6%	0.6%	4.1%	-0.4%	-6.8%	-7.9%	-7.9%	-4.2%	-2.9%
	Homophily on X	0.378	-29.7%	-5.9%	-3.5%	-5.4%	0.4%	0.6%	0.6%	0.6%	0.7%
	High homophily	0.574	-33.1%	1.2%	5.2%	-0.1%	-10.7%	-11.9%	-11.8%	-5.7%	-4.3%
	Low heterophily	-0.138	-42.3%	-3.2%	-7.7%	-2.8%	5.6%	6.9%	6.9%	4.1%	1.8%
	Medium heterophily	-0.258	-47.9%	-6.3%	-14.6%	-4.9%	12.8%	15.2%	15.1%	9.8%	4.5%
	Heterophily on X	-0.306	-62.6%	-7.0%	-5.1%	-6.2%	0.1%	0.1%	0.1%	0.1%	0.1%
	High heterophily	-0.463	-55.4%	-15.3%	-30.0%	-8.4%	31.1%	33.3%	32.4%	19.1%	7.9%

Table 2: Bias by simulation