REFINING THE TELESCOPE: ADDRESSING TWO CHALLENGES TO SOCIOLOGICAL RESEARCH IN ONLINE SOCIAL NETWORKS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

George Edward Berry

August 2020

REFINING THE TELESCOPE: ADDRESSING TWO CHALLENGES TO SOCIOLOGICAL RESEARCH IN ONLINE SOCIAL NETWORKS

George Edward Berry, Ph.D.

Cornell University 2020

The promise of digital trace data for sociological research is still partly unfulfilled due to the complexities of trace data. Such data are behavioral byproducts rather than collected to directly answer sociological questions. In this dissertation, I address two fundamental challenges to working with networked digital trace data: the opacity problem and the demographic prediction problem. The opacity problem presents a fundamental and unrecognized limitation to our ability to understand the dynamics of social contagion using behavioral data. I examine the problems it creates and demonstrate that careful statistical methodology can reduce error and provide meaningful insight into contagion dynamics. Then, I address the well-known demographic prediction problem in the context of networks. Network-correlated errors present a formidable challenge for our ability to use any method with measurement error (such as predictions or recall of alter group) to quantify intergroup contact. I formalize the problem and present a solution, employing the methodology I develop to derive estimates of interaction within and between groups on Twitter. These estimates are compared to the General Social Survey, indicating that homophily by race and gender are somewhat less than is present in offline discussion networks. Finally, I employ evolutionary game theory to examine conditions

under which homophily can facilitate or inhibit cooperation in pairwise games.

BIOGRAPHICAL SKETCH

George Berry grew up in suburban New Jersey. His first exposure to social science was in college, when he became interested in economics. This lead to an interest in networks, which in turn lead to network sociology. He has been fortunate to learn from and develop his knowledge in several academic and industry environments. When not thinking about networks, he spends his time reading about urban and welfare policy, biking around cities, and participating in politics.

This document is dedicated to my mother and father, who have been supportive throughout my life.

ACKNOWLEDGEMENTS

I'm grateful first to my family. I would like to thank for their companionship throughout graduate school: Milena Tsvetkova, Youngmin Yi, Thomas Davidson, Antonio Sirianni, Yuqi Lu, Lucas Drouhot, Mauricio Bucca, Mario Molina, Rediet Abebe, Xiao Ma, Christopher Cameron, Patrick Park, Yongren Shi, Daniel DellaPosta, Daniel Citron, Isabel Kloumann, Veronica Smith, Krystal Cannon, Lauren Griffin. I would like to thank for their practical mentorship: Sean Taylor, Solomon Messing, Lada Adamic, Alex Dow, Annie Franco, Alex Peysakhovich, Brent Cohn, Lukas Vermeer, Eytan Bakshy, Vladimir Barash, Erich Owens. I would like to thank for wonderful and eye-opening course experiences: Lionel Levine, Jon Kleinberg, Claire Cardie, Richard Swedberg, Anil Nerode. I would like to thank for mentorship and patience: Michael Macy, Benjamin Cornwell, and Steven Strogatz.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The rise of digital trace network data (139) provides rich insight into social and communication networks. The benefits of digital trace data have frequently been heralded as providing a "telescope" (180) for the social sciences. Studies of mass behavior (62; 81) have provided an unprecedented picture of large-scale human social and economic behavior.

Networks are complex, however, and present researchers with fundamental challenges to understanding. These challenges arise both from the complexity of networks, and the interplay of networks and behavioral trace data. This dissertation studies two such fundamental problems at the intersection of sociology and online social networks.

First: how can we estimate individual sensitivity to peer behavior in networked diffusion processes. While this question has long been addressed in sociological research (44; 170), I discover that data generated by networked diffusion processes poses previously unrecognized challenges for diffusion studies. Substantively, this finding has implications for our ability to estimate whether a contagion is complex—or whether contagion is present at all. By extension, it has implications for our ability to understand broader contagion dynamics, particularly in environments of heterogeneous agents as is common in online social networks.

Second: I address the incorporation of demographics into studies of on-

line social network data. Demographics are generally missing from digital trace data, making much otherwise rich data unable to answer core sociological questions about race, class, and gender. I develop an estimation strategy using supervised machine learning to predict demographics in networks which addresses the difficulties of making prediction in networked spaces. Importantly, this strategy makes it more likely that we obtain valid estimates of network properties like intergroup contact.

Using this methodology, I then study homophily along race and gender lines in two major metropolitan areas in the United States drawing on data from Twitter, finding that homophily on Twitter is generally less pronounced than in available data on offline social interaction. There is substantial nuance to this pattern, however, and these findings suggest further investigation is necessary to understand changing patterns of social interaction in both online and offline spaces.

As a final note, I develop a game theoretic model of behavior within given homophily arrangements which explores the potential causal impact of homophily on resolving pairwise social dilemmas. This indicates that homophily itself can be a factor in facilitating cooperation in coordination games, and it can reduce the strength of ingroup bias needed to overcome the Prisoner's Dilemma.

It is my hope that findings here will be useful to both sociologists and social scientists in other disciplines.

CHAPTER 2

THE OPACITY PROBLEM: A HIDDEN LIMITATION FOR CONTAGION RESEARCH

Many social phenomena can be modeled as cascades in networks, where nodes adopt a behavior in response to peers adopting. When studying cascades, researchers typically assume that the number of active peers when a node adopts is equivalent to the node's threshold for adoption. This assumption is rarely justified due to the "opacity problem": networked cascades reveal intervals which contain thresholds, rather than point estimates. Existing approaches take the maximum of each node's threshold interval, which biases models of social influence. Opacity is inevitable in many small graphs when using the threshold model, resulting from the networked process itself rather than data collection techniques. Using simulation, we extend this finding to the probabilistic SI (independent cascade) model. We confirm these theoretical results by studying 50 large hashtag cascades among 3.2 million Twitter users, finding that 20% of adoptions suffer from opacity. Different assumptions in response to opacity qualitatively change conclusions about peer influence. While opacity is a far-reaching problem, it can be addressed. Using information from nodes who have tightly bounded intervals allows building models to reduce error in estimating node thresholds.

2.1 Introduction

Like epidemic diseases (110; 172; 32), social contagions are ubiquitous, highly consequential, and widely studied (85; 164; 165; 34; 119; 169; 170; 40; 171). Unlike nearly all diseases, social contagions often require multiple sources of infection, especially if adoption is costly, risky, motivated by affect, or entails positive network externalities (34).

In this paper, we explore a challenge for studies of contagion: data generated by networked processes typically does not provide all of the information needed to recover peer influence, even when cascades are tracked precisely. This occurs due to the "opacity problem": cascades provide intervals in which node thresholds lie, but often do not indicate the threshold itself. Researchers typically infer the amount of social reinforcement required for adoption by recording node exposure (number of active neighbors) when the node activates. We refer to this practice as the exposure-at-activation (EAA) rule, which is equivalent to taking the maximum of the threshold interval for each node. This leads to an upward bias in estimates of peer reinforcement needed for adoption at the node level, and can bias models of peer effects.

The EAA rule can be found in diverse studies across disciplines, including sociology (71; 170; 58), economics (56; 28; 12), medicine (40), and information science (3; 150; 163; 83; 168; 25; 183). It is even implicitly present in "snapshot" studies (8; 50) and in networks surveyed in waves such as Add

Health (17).

Simulated cascades using both deterministic and probabilistic activation rules indicate that the opacity problem creates substantial bias. An empirical study of hashtag cascades among 3.2 million users on Twitter confirms this intuition: about 20% of hashtag first usages have uncertain adoption thresholds. For 2 in 5 hashtags, different responses to the opacity problem produce qualitatively different conclusions about whether peer reinforcement promotes diffusion.

We provide several tools for addressing the opacity problem. First, a simple condition can be applied to recover node threshold intervals. Nodes with small threshold intervals can be used to estimate a threshold model, which can then be applied to all nodes. In simulations, this process substantially reduces error in estimating thresholds. In addition, we expect future research to develop novel ways to further reduce measurement error introduced by opacity and the EAA rule.

While the opacity problem may seem complex, the core intuition is simple and is summarized in Figure 2.1: for some network-threshold configurations, some nodes will always activate with exposure greater than threshold. This occurs because all nodes cannot check the status of neighbors all of the time. To use a familiar example: suppose a person "looks away" from her phone while many friends adopt a behavior on a social media platform. When she comes back, she sees that several friends have adopted and she promptly adopts. Ascertaining which friend was pivotal is a complex

counterfactual question which the data does not directly answer. A similar analogy can be made about states adopting policies (176), college students selecting majors (56), or any number of other behaviors.

The paper proceeds as follows: In Section 2, we introduce the threshold model and show that for some specific network configurations, the opacity problem is inevitable. An exhaustive survey of all small graphs with all threshold assignments is conducted, which indicates that even in small networks, the opacity problem creates substantial uncertainty. We propose a condition for bounding node-level uncertainty which is the basis of many of our subsequent results. In Section 3, we conduct a variety of simulations using the threshold model from Section 2, finding that the application of the EAA rule creates substantial upward bias in estimates of thresholds. We also simulate the susceptible-infected (SI) (96) or independent cascade model (ICM) (107), finding that the opacity problem arises in the probabilistic case as well. We show that a modeling procedure using node threshold intervals with low uncertainty can substantially reduce the upward bias of the EAA rule. In Section 4 we turn to an empirical case: hashtag cascades on Twitter. This analysis indicates that the opacity problem is present in a data-rich case from social media. We conclude with a discussion of implications and suggestion for future work.

To our knowledge, we are the first to highlight the full scope of this problem, although aspects of it have been discussed in (170; 115; 150).

2.2 Model

2.2.1 Model definition

We choose a threshold model similar to previous work (85; 179; 107) to illustrate the opacity problem. We study the threshold model formally, and show using simulation below that our results apply to probabilistic models of contagion as well.

A graph $G = (V, E)$ has nodes V and edges E. We assume that G is undirected and connected. Nodes are indexed by i and j, with $N(i)$ indicating all nodes in the neighborhood of i. A diffusion process plays out on this graph over times $t \geq 0$. Nodes have activation statuses at each t indicated by $y_i(t) \in \{0, 1\}$, with $y_i(t) = 0$ indicating inactive and $y_i(t) = 1$ indicating active. Once a node activates, it remains active for the rest of the diffusion process. Edges may have weights w_{ij}, although weights are set to 1 in results presented here to reduce the complexity of the model space[1]. Nodes have integer-valued thresholds $h_i \geq 0$, indicating the minimum level of peer reinforcement required for adoption[2]. A node adopts when it updates at time t and exposure is greater than threshold, $k_i(t) = \sum_{j \in N(i)} w_{ij} y_j^t \geq h_i$, where j ranges over neighbors of i.

[1]Constructing a weighted example where opacity happens is straightforward. For example, change one of the edge weights to 2 in Figure 2.1.

[2]Results do not change with fractional thresholds, as demonstrated in simulations below.

The update step proceeds as follows: at time t, an inactive node i is chosen at random to update. This update process produces a well-defined update ordering contained in vector $\mathbf{u}$. The vector $\mathbf{u}_i$ indicates all times at which i updates. When i updates, i checks the status of all neighbors $j \in N(i)$, and if $\sum_{j \in N(i)} y_j \geq h_i$, i activates immediately and $y_i(t) = 1$. While random updating and instantaneous activation are not realistic modeling choices, results are not sensitive to these simplifications[3].

In an important distinction from past work, whether or not i activates at t, i's exposure $k_i(t)$ is recorded. Recording i's exposure when i *does not* update is crucial to addressing the opacity problem. Throughout the paper, we will refer to exposure-at-activation (EAA) and the "EAA rule", which we define here.

Definition 1. *Assume node i first activates at time t. Then i's* ***exposure-at-activation (EAA)*** *is $k_i(t)$.*

Definition 2. *The* ***exposure-at-activation (EAA) rule*** *estimates threshold h_i by the exposure-at-activation. If i first activates at time t, the EAA is $k_i(t)$ and the EAA rule assumes $h_i = k_i(t)$.*

The use of the EAA rule is straightforward when adoptions are timestamped. However, many studies of diffusion use networks surveyed in

[3]Figure 2.1 shows a case which demonstrates this: the opacity problem occurs regardless of the order in which nodes update, meaning that random updating is not a critical assumption. Giving nodes an activation delay after updating would not substantively change results, since j or k would have to activate first, leading to the other node being measured with error.

waves, such as Add Health (e.g. (17)) or network snapshots (8). The activation decision is only observed after a (potentially unknown) delay. In this case, the time a survey is administered is functionally the "adoption time", since it is the first time a node is observed as active. If "survey time" or "collection time" is substituted for "activation time", our results fully apply to network snapshots as well.

2.2.2 Model remarks

We have chosen a simple threshold model which abstracts away many aspects of reality. While past work has made similar simplifying assumptions (85; 107; 108; 34; 179), this paper fundamentally concerns empirical data. We use a model to motivate a claim about the empirical study of social contagion.

Because of this, it is crucial that results presented here are not artifacts of particular modeling assumptions. We have attempted to examine our results under as many assumptions as possible. Cases we have considered are: integer and fractional thresholds, weighted networks, non-random update orderings, various distributions of thresholds (normal, uniform, exponential), allowing nodes to update simultaneously, probabilistic activation models (independent probabilities), allowing nodes to have activation delays, allowing nodes to have transmission delays, and allowing nodes to notify neighbors of activation. In each of these cases, the opacity problem

remains a problem, and can be demonstrated with simple examples akin to Figure 2.1.

On the other hand, the threshold model considered here has a surprising amount of flexibility. It makes few assumptions about G or the distributions of h_i and $\mathbf{u}$. This is by design: the opacity problem occurs in small graphs that are very likely to occur for a wide range of specific assumptions.

While we consider several extensions to this model, four assumptions are fixed throughout. The most important of these is the static nature of the graph G. In reality, edges are created and decay in networks all the time (111). While a dynamic graph does not eliminate the opacity problem, it may reduce its impact by sparsifying the graph. A second simplification is the assumption that diffusion is entirely endogenous: once the threshold is known, no other node-level information is relevant for the unfolding of the diffusion process. In practice, changing circumstances may dynamically alter node thresholds, creating additional challenges for estimating peer effects. Third, we assume throughout that peer influence is driving adoption, rather than a process orthogonal to the social network. Fourth, we assume that people are aware of the activation statuses of their neighbors and do not suffer from limited attention (182).

2.2.3 The opacity problem in the threshold model

The opacity problem is inevitable in certain network configurations using the threshold model defined above. Consider the simple network in Figure 2.1, which is a triad among nodes i, j, k. Node i is an innovator with threshold 0, while nodes j and k have threshold 1. Any update ordering for this graph will produce an over-estimate of at least one node's threshold. Since node i must adopt first, there are only two possible orderings in which nodes can activate: (i, j, k) or (i, k, j). In the former case, k adopts with 2 active neighbors despite having threshold 1, while in the latter case j adopts with 2 active neighbors despite having threshold 1. Applying the exposure-at-activation (EAA) rule would produce an over-estimate of the node threshold in these cases.

In practice, how can we determine if node thresholds are measured precisely or not? For instance, in Figure 2.1-A, node j could activate with 1 active neighbor but have a true threshold of zero. We need to place a lower bound on the threshold in addition to an upper bound. Recording exposure $k_i(t)$ when nodes remain inactive allows us to do this.

Consider the diffusion process in Figure 2.2, which replicates the graph and threshold assignment of Figure 2.1 but changes the update ordering to $\mathbf{u} = (j, k, i, j, k)$. This update ordering does not affect the order in which nodes activate (which remains (i, j, k)), but provides lower bounds on the thresholds of nodes j and k. The lower bound on node i arises by assum-

ing there are no negative thresholds[4]. Note that node j has an interval of $[0, 1]$, indicating that 0 active neighbors were insufficient to trigger adoption, while 1 active neighbor was sufficient. j's threshold is therefore known with certainty to be 1 (recall we assume integer thresholds). For node k, 0 neighbors was insufficient while 2 neighbors were sufficient, indicating uncertainty about whether k's threshold is 1 or 2. Therefore, j's threshold is measured precisely while k's threshold suffers from opacity.

This suggests a formal precise measurement condition. Consider a node i which updates at t and some later t'. Assume i is inactive at t ($y_i(t) = 0$) and active at t' ($y_i(t') = 1$), giving threshold interval $[k_i(t), k_i(t')]$.

Condition 1. *The threshold for i is **precisely measured** if i is inactive at t ($y_i(t) = 0$), active at t' ($y_i(t') = 1$), and the width of the interval $[k_i(t), k(t')] = 1$.*

In other words, if exposure changed by exactly one and this triggered i's adoption, then we know the last neighbor to activate provided the "final push" for adoption.

Condition 1 has two edge cases. First, innovators who adopt with 0 active neighbors are measured precisely only by assumption. These nodes technically have no lower bound on their threshold intervals. Second, some nodes may have no lower bound, for instance a node who first updates and activates at exposure 2. Condition 1 categorizes these nodes as imprecisely measured.

[4]The appropriateness of this assumption has to be investigated in each individual case. Negative thresholds may indicate "super-eager" nodes.

In Figure 2.3 we apply Condition 1 to all small graphs of orders 2 to 4, with all threshold assignments that support diffusion. These graph-threshold combinations are the building blocks of large cascades. Figure 2.3 indicates that for 75% of these graph-threshold combinations, at least one node always has an interval which is not precisely measured according to Condition 1. This plot also shows the proportion of precisely measured nodes given nodes update randomly, indicating that measurement is difficult in many small graphs.

2.3 Simulation evidence

Simulating the threshold model in the previous section provides a method to examine the severity of the opacity problem under different assumptions (121). In this case, thresholds are known with certainty which allows exact quantification of error.

We use simulation to perform two analyses. First, the impact of opacity on observational threshold distributions is assessed. Then, we demonstrate that combining Condition 1 with a predictive model such as linear regression can reduce the error in estimated thresholds.

Graphs are generated either with a power law with clustering (PLC) algorithm (101) with clustering parameter 0.1, or with a Barabàsi-Albert (BA) procedure (13). All graphs have 1000 nodes, and have degrees ranging be-

tween 12 and 20. Each parameter set is replicated 1000 times to minimize sampling variance[5].

The threshold model developed above is simulated in the following way. At each time step t', choose an inactive i. Node i checks its exposure $k_i(t')$, and we record this exposure regardless of whether i activates. If $k_i(t') \geq h_i$, $y_i(t')$ is set to 1. If i activates, its exposure at t' is differenced from its exposure at i's previous update at t, $k_i(t') - k_i(t)$. We apply Condition 1, and call i "precisely measured" if $k_i(t') - k_i(t) = 1$. By assumption, nodes that activate with 0 active neighbors are precisely measured.

In addition to the threshold model, we also simulate the susceptible-infected (SI), or independent cascade model (ICM) (96; 107; 140). We provide the simulation algorithms in the Appendix. Since the SI model is probabilistic, this analysis provides a robustness check against the deterministic nature of the threshold model. In the SI model, each active node provides an independent chance for its neighbors to adopt. We set this *transmission probability* to $p = 0.2$. Two types of contagion mechanics are simulated for the SI model. First "pull" dynamics: when a node is selected to update, it checks how many newly active neighbors it has, flips that many coins with $P(\text{heads}) = p$, and activates if at least one comes up heads. Second, "push" dynamics: an active node is selected at random, and a single coin with P(heads) = p is flipped for each of its neighbors. Any neighbors that flip a heads are activated immediately. Each active node may only "push"

[5]We also examined Watts-Strogatz graphs but omit them for brevity since they did not meaningfully change the results.

the contagion once.

For the SI model, the true "threshold" or "critical value" for any node is unknown in advance. The simulation process reveals, for each node, a series of coin flips. We record the first such coin flip that comes up heads as the "critical value" or "threshold" for the node.

2.3.1 Effect of opacity on threshold distributions

The distribution of active neighbors recovered using the EAA rule systematically differs from the true threshold distribution. Figure 2.5 shows the divergence using the Barabàsi-Albert graph with mean degree 12. Distributional divergence is the case for integer thresholds, fractional thresholds, and the SI model. The case in Figure 2.5-D is surprising: all nodes have a fractional threshold of 0.2, yet the distribution recovered using the EAA rule is roughly uniform. In this case, the EAA rule records about 10% of nodes as having "threshold" 1.0.

The EAA rule performs best in Figure 2.5-B, where the threshold distribution is normal. In this case, the EAA rule still produces a long tail of "high threshold" nodes, which gives the impression that a heroic level of peer reinforcement is required for some nodes to adopt.

These results indicate that opacity presents substantial challenges for estimating threshold distributions. Regression-based approaches relying on a

biased distribution can have an influence coefficient biased in either direction. When the EAA rule moves observations from zero to nonzero thresholds, the strength of influence can be overestimated. On the other hand, when the EAA rule moves low threshold nodes to higher values, the effect of peer influence can be under-estimated. The particular bias is based on the relative strength of these two forces.

2.3.2 Reducing measurement error with predictive models

Bias in thresholds can be meaningfully reduced using predictive models. The strategy is straightforward: apply Condition 1 to all nodes, determine which nodes are measured precisely, and estimate a model on these nodes only. We use Ordinary Least Squares (OLS) regression, although in principle any machine learning regressor can be used. Then, we predict thresholds for nodes with measurement error. As shown in Figure 2.4 Results indicate this approach reduces root mean squared error (RMSE) dramatically, from 3.1 to 8.1 times baseline to 1.15 to 1.35 times baseline.

We use a power law with clustering graph and threshold function $h_i = 5+3x_i+\epsilon_i$, where $x_i, \epsilon_i \sim \mathcal{N}(0,1)$. Thresholds below 0 are set to 0. Cascade dynamics are simulated as described above. The covariate x_i represents some feature of i (e.g. age, wealth) that provides information about how much social reinforcement i needs before activation.

Treating thresholds as a function of node-level attributes (170) allows

Mean degree	Nodes	Activations	Precisely measured
12	1000	720.7	113.4
16	1000	924.1	43.0
20	1000	983.1	13.0

Table 2.1: Descriptive statistics on simulations, averaged over 1000 runs. The average number of activated nodes increases as degree increases. However, it becomes more difficult to precisely measure nodes in higher mean degree graphs.

estimating a threshold model

$$h_i = \alpha + \beta x_i + \epsilon_i \tag{2.1}$$

This model can be estimated using only the precisely measured set of nodes to reduce bias in thresholds. Results of this procedure compared to the EAA rule are displayed in Figure 2.4. We note that this model predicts thresholds based on node attributes, rather than on endogenous network characteristics.

The error variance represents the inherent unpredictability in thresholds and provides a natural baseline error for comparing error-reduction methods. An ordinary least squares (OLS) model estimated on the true data will have a root-mean-square error (RMSE) equal to the error variance, which here is set to 1.

2.4 Hashtag cascades

Studying hashtag cascades on Twitter, we find that the opacity problem can qualitatively change conclusions about the effects of peer reinforcement on adoption. Twitter provides an ideal setting for this analysis: tweets (and therefore adoption events) are timestamped and the contagion (a hashtag) can be clearly tracked. For about 2 in 5 hashtags, different responses to opacity yield opposite conclusions about the effectiveness of social reinforcement. At the adoption level, about 1 in 5 hashtag adoptions show uncertainty about the amount of peer reinforcement required for adoption.

We tracked cascades for 50 hashtags (see Appendix for list and selection procedure) with between 40,000 and 360,000 unique adopters and 41,000 to 1.3 million total usages per tag. Retweets were filtered out. A total of 3.2 million users ("active users") tweeted using one or more of these hashtags. These active users are connected by 45 million bidirected @mention links, which is a common measure of ties on Twitter (150). An additional 105 million bidirected @mention edges connect active users to users who were exposed but never adopted ("inactive users"). These data contain a total of 7 billion tweets.

We choose to construct the network from bidirected @mentions because they represent a relatively strong connection that can facilitate diffusion. This type of link is sparser than other types of edges on Twitter, such as follows. Combined with our findings above about the positive correlation

between network density and severity of the opacity problem, using the sparser bidirected @mention graph can be considered a conservative test of the impact of opacity.

We examine the effect of peer reinforcement with $p(k)$ curves (50; 150; 114), which plot the probability of first activation with exposure k, given ever being k-exposed. Work using this methodology has found that higher exposure levels substantially increase the probability of hashtag adoption, particularly for controversial issues like politics (50; 150). More recent work has found that hashtags spread more like simple contagion where nodes are immunized after the first exposure (114). Since complex contagions have different cascade dynamics than simple contagions (33; 14), understanding the effects of reinforcement is consequential for predictions and interventions in online networks.

2.4.1 Constructing threshold intervals

To construct threshold intervals, we assume updates correspond with tweet times. While this is not a perfect assumption, a user is engaged with the Twitter platform when tweeting, and likely checks tweets in the seconds before or after sending a tweet. Let $\mathbf{u}_i$ be the times i tweeted in increasing order. For hashtag g, let $t_i^g \in \mathbf{u}_i$ be i's first usage of g, and $t_i^{g-1} \in \mathbf{u}_i$ be the tweet immediately prior to i's first usage of g.

We construct descending time intervals $(t_i^{g-1}, t_i^g], (t_i^{g-2}, t_i^{g-1}]$, etc., which

give the periods of time between i's updates. Iterating through these intervals, we find the most recent interval for which one or more neighbor adopted g, and count the total number T^* of neighbor adoptions in that interval. If $T^* = 1$, then Condition 1 is satisfied, since i's exposure changed by exactly one and then i adopted. Otherwise if $T^* > 1$, i's threshold is uncertain within an interval of size greater than 1. For instance if 2 neighbors activated in $(t_i^{g-1}, t_i^g]$ and 4 total neighbors activated before t_i^{g-1}, then i's threshold interval is $[4, 6]$. This interval is considered "imprecisely measured" since it has size greater than 1.

2.4.2 Empirical measurement rates

Measurement uncertainty affects 21% of activations, although there is substantial variation both by hashtag and exposure level. We exclude innovators (0 active neighbors at activation) from this analysis both to focus on the effects of opacity on estimates of peer reinforcement. Nodes which adopt with higher exposure are more likely to suffer from measurement uncertainty 2.2. Among nodes that activate with exposure 1, 91% are measured precisely. Among nodes that activate with exposure 2, only 66% are precisely measured. For each hashtag, we compute the clustering coefficient (178) among the induced subgraph of adopters, and display measurement rates for the top and bottom quartile of hashtags.

	All	Low clustering	High clustering
Exposure	Proportion precisely measured		
1	0.91	0.90	0.97
2	0.66	0.74	0.46
3	0.64	0.70	0.43
4	0.63	0.68	0.41
5	0.63	0.66	0.42
N hashtags	50	13	13
N adoptions	1,625,759	548,420	262,890

Table 2.2: Descriptive statistics for Twitter hashtag cascades. There are 1.6 million adoption events where exposure is between 1 and 5 at the time of adoption. Precise measurement rates are higher for adoptions with exposure 1 than higher exposure levels. Splitting hashtags into quartiles of clustering coefficient among adopters, low and high clustering tags display substantially different precise measurement rates.

2.4.3 Social reinforcement and $p(k)$ curves

We construct two $p(k)$ curves for each hashtag based on the threshold intervals for each node. The "upper" $p(k)$ curve, notated $p_U(k)$, takes the maximum of each threshold interval for each activation. This is equivalent to applying the EAA rule found in much work on social contagion. The "lower" $p(k)$ curve, notated $p_L(k)$, takes the minimum of each threshold interval for each activation. In other words, it assumes that the EAA rule is maximally wrong. While taking the minimum of each threshold interval may seem like an unrealistic assumption, it has two advantages: 1) the lower curve does not have the long right tail present when using the EAA rule (as seen in Figure 2.5); 2) assuming the EAA rule is maximally wrong allows us to bound the true value in an interval, assuming that peer influence was the

actual reason for adoption.

This analysis is presented in Figure 2.6. We find that nearly all hashtags had initially increasing upper curves, $p_U(2) > p_U(1)$, corresponding to social reinforcement facilitating adoption and qualitatively agreeing with past work (50; 150). When turning to lower curves, we noticed via visual inspection that many hashtags had the opposite pattern, $p_L(2) < p_L(1)$, where the probability of adoption drops after the first exposure. This pattern indicates that contagion spreads more like an independent cascade, with exposure leading to immunization (as proposed in (114)). We find that for 19 out of 50 hashtags (full list in Appendix), the upper curve displays the usual pattern of increasing adoption probability in k, while the lower curve displays the decreasing pattern. For the remaining 31 hashtags, both upper and lower curves indicate adoption probability increasing in k, with a smaller increase for the lower curve.

We examine three characteristics of cascades which could be correlated with discrepancies between the upper and lower curves: 1) the total number of hashtag adopters; 2) the clustering coefficient among the induced subgraph of adopters; 3) the temporal burstiness of the cascade, measured by the Gini coefficient of the number of adoptions across days between the first and last usages of a particular hashtag. Of these three factors, clustering coefficient and burstiness have correlations with a decreasing lower curve (Pearson correlation for clustering coefficient: 0.26, $p = 0.063$, for Gini coefficient: 0.34, $p = 0.017$). Hashtags with high clustering coefficients tend to

have particularly pronounced differences between upper and lower curves, as seen in Figure 2.6-B.

Taken together, these results indicate that the opacity problem creates more uncertainty about cascade dynamics when adopters are clustered in dense networks and cascades happen in short time spans. This paradoxically makes it difficult to determine the effects of social reinforcement when adopters have high clustering, even though contagion which requires social reinforcement is most likely to spread in clustered networks (34).

2.5 Implications and future work

The opacity problem and the application of the EAA rule present several challenges for the study of contagion and social influence. These can be summarized into the following categories: 1) estimating individual thresholds and the distribution of thresholds; 2) assessing impacts on empirical models of contagion; 3) using data to better assess uncertainty.

The first challenge we have substantially addressed in this paper. Future work can build on this by assessing the information contained in low-error observations. We have only considered observations measured precisely, but an observation with a threshold interval of size 2 is much more certain than one with size 20. Bayesian and machine learning techniques may prove useful in this direction. For instance, formulating the labeling problem as a

semi-supervised task could lead to lower error threshold predictions than the method we have proposed. In addition, the predictive value of various signals remains to be assessed empirically. For instance, in the Twitter case, using text embeddings (130) representing cultural taste and graph embeddings (86) indicating attention may prove useful for predicting thresholds.

The second category poses a series of issues that we have not touched upon in this paper: the specific implications of using exposure as an independent variable in a model estimating peer effects on adoption. A simple version of this model can be written $y_i(t) = \beta k_i(t) + \epsilon_i$. In this case, the opacity problem plus EAA rule presents a difficult errors-in-variables problem: $k_i(t)$ is measured with error *conditional on* $y_i(t) = 1$, with non-negative error for all nodes i. Depending on the specific form measurement error takes, $\hat{\beta}$ may be over- or under-estimated. Future work can assess the impacts of opacity in a variety of empirical cases.

The third challenge takes two forms. When data about adoption times is relatively granular (as in the case of social media), methodology proposed here can be applied to assess the impacts of opacity. This is straightforward, and doing so will provide insight both about the substantive social process and the difficulty of measuring various diffusion processes. When data is not granular, as in the case of surveys which may be administered yearly, care needs to be taken to assess opacity. To address the problem, a yearly survey may consider asking about the adoption of a behavior and about the specific time that behavior was adopted, allowing a partial reconstruction

of the adoption ordering among network neighbors. This could reduce the severity of the opacity problem and lead to better inferences about social influence.

We are aware of one case where the opacity problem can be avoided entirely: ego-network randomization(10; 4). In this experimental design, the researcher can control the number of active neighbors displayed, leading to valid estimates of adoption probabilities at various exposure levels.

Recognizing and addressing the opacity problem has the potential to yield more precise answers to important questions. For instance, models of node critical values can be used to better study complex contagion (34) empirically. Results from such an analysis can then be applied to problems such as influence maximization (107; 108) and the cascade prediction problem (38).

2.6 Figures

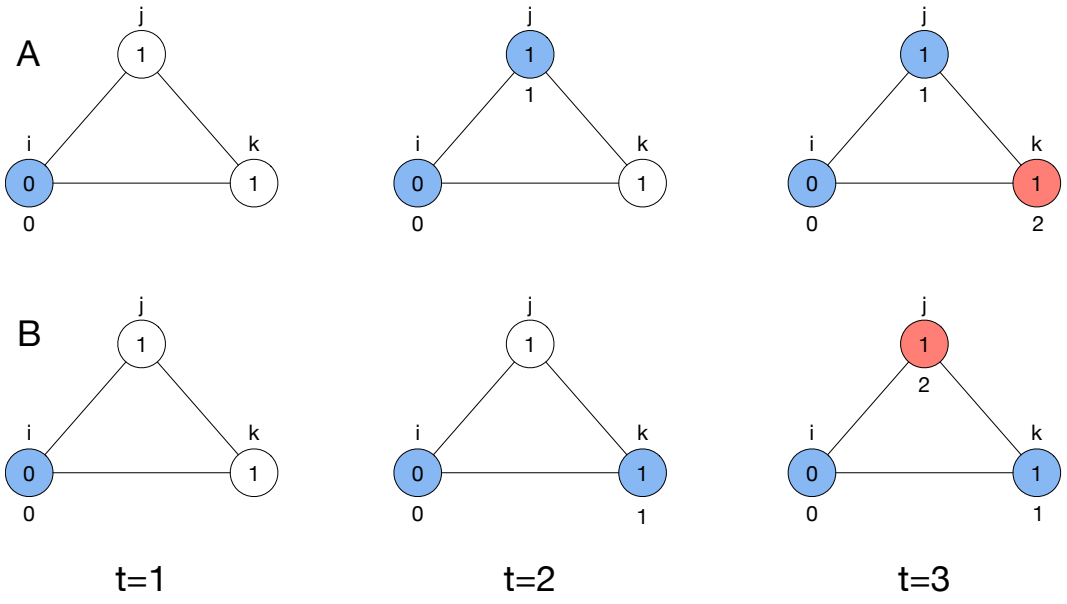

Figure 2.1: A triad between nodes (i, j, k) with thresholds $(0, 1, 1)$ over three time periods. A) and B) display the only possible activation orderings: (i, j, k) or (i, k, j), respectively. Node thresholds are displayed on the node, while the exposure at activation (EAA) is displayed below the node. Blue indicates a node activates with EAA equal to threshold, while red indicates that EAA is greater than threshold. Since these are the only two possible activation orderings, either j or k will always have EAA greater than threshold, leading to measurement error.

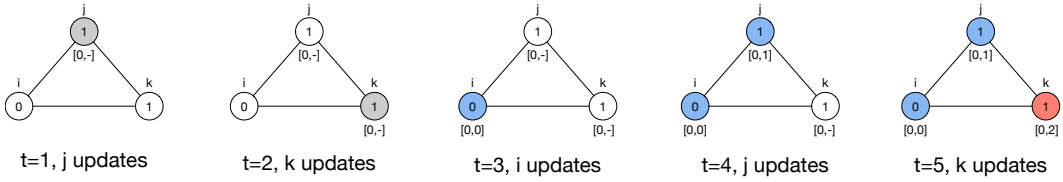

Figure 2.2: A triad between nodes (i, j, k) with thresholds $(0, 1, 1)$ over five time periods. Nodes update in order (j, k, i, j, k). This ordering allows placing bounds on thresholds for each node, indicated by the intervals below each node (where "-" indicates no measurement). Gray indicates a node updates but does not activate, blue indicates a node activates with exposure-at-activation (EAA) equal to threshold, while red indicates a node activates with EAA greater than threshold. When nodes j and k update at $t = 1, 2$, respectively, their exposure is recorded even though neither activates. This allows placing a lower bound on their thresholds. As the process plays out, k activates with EAA greater than threshold. The lower bound of 0 from k's update at $t = 2$ means k has a threshold interval of $[0, 2]$.

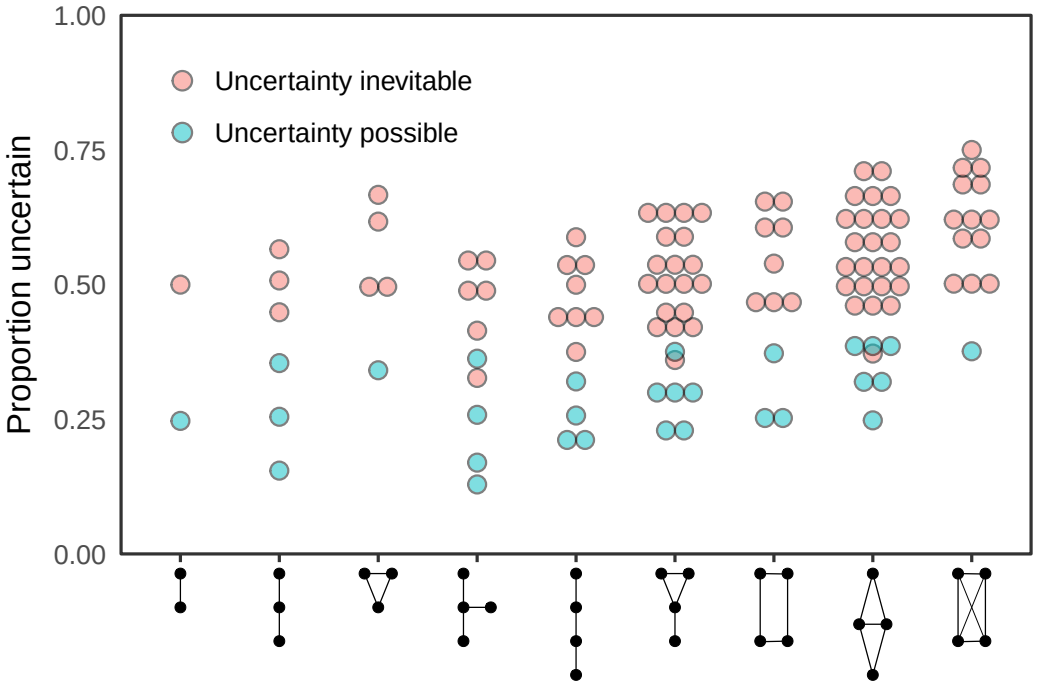

Figure 2.3: The severity of the opacity problem in all connected graphs of order 2 to 4, with all threshold assignments that support diffusion. Each point represents a unique threshold assignment for the graph on the x-axis. The y-axis displays the mean number of nodes whose thresholds are uncertain when applying Condition 1. For 86/115 (75%) of graph-threshold combinations, the EAA rule always produces at least one uncertain node (red dots). Since these small graph-threshold combinations are the beginnings of large cascades, we should expect measurement error when studying social contagion observationally.

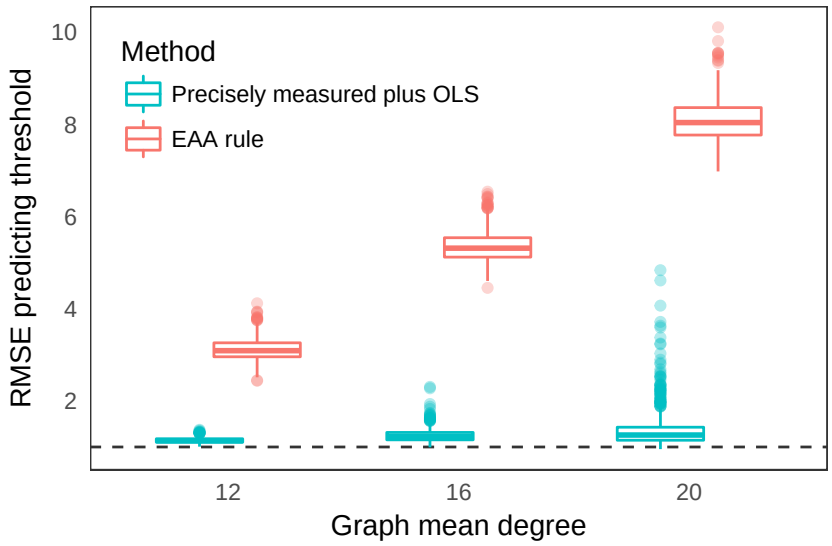

Figure 2.4: Since the exposure-at-activation (EAA) rule takes the maximum of threshold intervals, it leads to over-estimates of thresholds (shown in red). The dashed black line gives the baseline RMSE, which is 1 in these simulations. The EAA rule leads to an average error between 3.1 and 8.1 times this baseline. A model estimated on the precisely measured subset (< 15% of nodes) and then used to predict all thresholds (blue) produces much lower error of 1.14-1.35 times baseline.

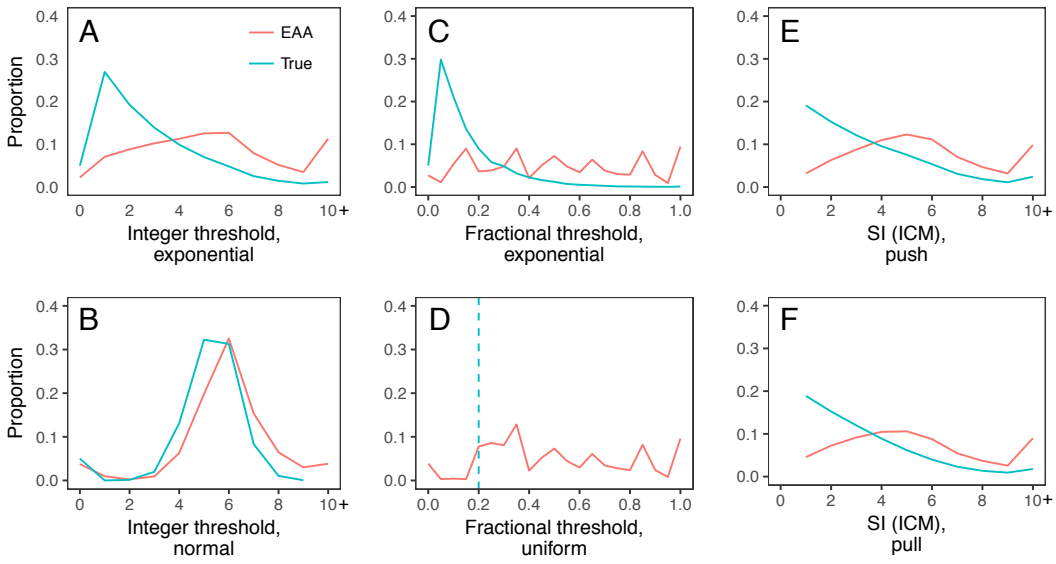

Figure 2.5: True thresholds (blue) compared to measurements using the exposure-at-activation (EAA) rule (red) for a variety of contagion processes. A) and B) show the threshold model with integer thresholds drawn from an $\exp(\beta = 3)$ and $\mathcal{N}(5, 1)$, respectively. C) and D) use fractional thresholds (note the different x-axis scale) with 5% seed nodes. C) uses an $\exp(\beta = 3)$ normalized to the $[0, 1]$ interval, while D) sets all non-seed nodes to have threshold 0.2. E) and F) use a susceptible-infected (SI) model, also called the independent cascade model (ICM), where each node has an independent chance $p = 0.2$ to activate peers. In this case, the "threshold" is the first coin flip to come up heads. In E), when a node activates it "pushes" contagion to its neighbors by causing them to activate with probability p in a random order. In F), when nodes update they check all neighbor statuses and flip a probability p coin for each newly activated neighbor.

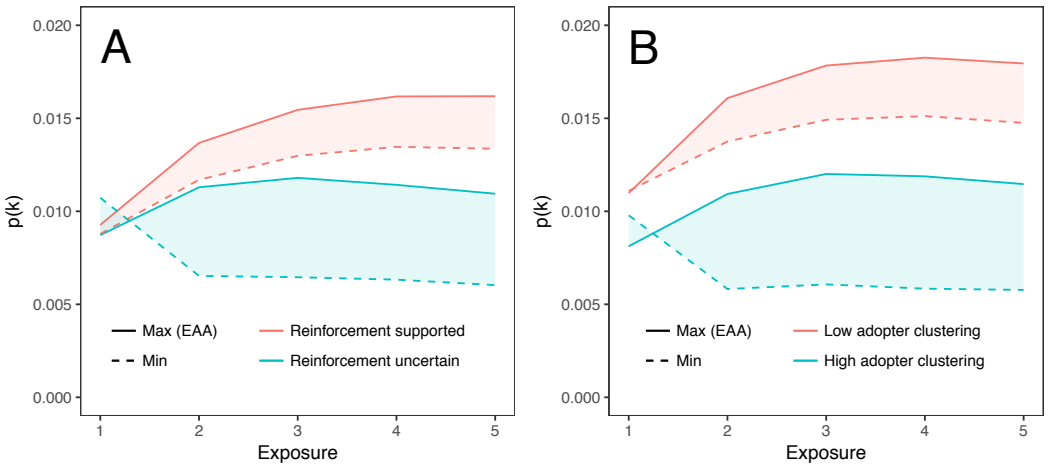

Figure 2.6: Different responses to the opacity problem support different conclusions of the importance of peer reinforcement for adoption. This can be seen with $p(k)$ curves, which plot the probability of first activation at k-exposure given ever being k-exposed. Solid lines indicate using the exposure-at-activation (EAA) rule (the maximum of threshold intervals). Dashed lines indicate using the minimum of threshold intervals. The shaded region is the discrepancy between the two methods. A) shows that for 19/50 hashtags (blue), the upper and lower curves lead to different conclusions about the importance of reinforcement: use of the EAA rule supports a complex contagion hypothesis, while taking the minimum of threshold intervals supports an independent contagion hypothesis. B) shows that this pattern can be replicated by taking hashtags with high and low clustering in the bidirected @mention graph among adopters, indicating that the dynamics of hashtags in highly clustered networks are particularly uncertain.

2.7 Appendix

2.7.1 Simulation details

We simulate cascades using the NetworkX package in Python 3 (87). All code is available at https://github.com/georgeberry/thresholds. Statistical analyses are done using Scikit Learn (141) in Python 3, and in R.

We use 4 different graph generation routines at various points: Barabàsi-Albert (13), power-law with clustering (101), Watts-Strogatz (181), and an "atlas" of all small graphs (146). All four are built-in to the NetworkX package.

Threshold model

The algorithm for simulating diffusion is as follows.

At each time step t',

1. An inactive node i is chosen at random
2. Node i checks how many active neighbors it has, $k_i(t')$
3. The exposure at t' is recorded
4. If $k_i(t') \geq h_i$, set $y_i(t') = 1$
5. If i activates at t'

(a) Apply Condition 1 by differencing the exposure between t' and i's previous update at t, $k_i(t') - k_i(t)$

(b) If i did not update at any previous t, we set node i to "imprecisely measured" unless node i activated with exposure 0, in which case we set it to "precisely measured"

(c) If $k_i(t') - k_i(t) = 1$, set node i to "precisely measured", otherwise "imprecisely measured"

End iteration when A) all nodes have activated; B) each inactive node has been checked consecutively without activating.

2.7.2 SI/ICM "pull" model

This model is similar to the threshold model, except activations are decided by coin flips instead of deterministically.

At each time step t',

1. An inactive node i is chosen at random
2. Node i checks how many active neighbors it has, $k_i(t')$
3. i checks the difference between its exposure at last update t and present, $d_i = k_i(t') - k_i(t)$
4. i flips d_i coins, each of which has probability $P(\text{heads}) = 0.2$

5. If the jth coin comes up heads, activate i and set i's "threshold" to $k_i(t)+j$ (previous exposure plus coins flipped at t' before getting heads)
6. If i activates at t'
   (a) Apply Condition 1 by differencing the exposure between t' and i's previous update at t, $k_i(t') - k_i(t)$
   (b) If i did not update at any previous t, we set node i to "imprecisely measured" unless node i activated with exposure 0, in which case we set it to "precisely measured"
   (c) If $k_i(t') - k_i(t) = 1$, set node i to "precisely measured", otherwise "imprecisely measured"

End iteration when A) all nodes have activated; B) each inactive node has been checked consecutively without activating.

2.7.3 SI/ICM "push" model

In this model active nodes are selected to update, and they get a single chance to activate each one of their neighbors.

Initialize an empty array of active nodes that have updated, a.

At each time step t',

1. An active node i not in a is chosen at random and appended to a

2. The neighbors of i are randomized, shuffle($N(i)$)

3. For each j in shuffle($N(i)$)

   (a) Node j's exposure $k_j(t')$ is recorded

   (b) Node j flips a coin with P(heads) = 0.2 and activates if it comes up heads

   (c) If j activates at t'

      i. Apply Condition 1 by differencing the exposure between t' and j's previous update at t, $k_j(t') - k_j(t)$

      ii. If i did not update at any previous t, we set node j to "imprecisely measured" unless node j activated with exposure 0, in which case we set it to "precisely measured"

      iii. If $k_j(t') - k_j(t) = 1$, set node j to "precisely measured", otherwise "imprecisely measured"

End iteration when A) all nodes have activated; B) all active nodes are in a.

2.7.4 Small graphs

NetworkX provides a full enumeration of all small graphs ("graph atlas"). We take all such graphs with between two and four nodes. Then, we filter out graphs which have more than one component, giving a set of all con-

nected graphs with between two and four vertices. Call this set of graphs G.

For each $g \in G$, we generate all possible critical value assignments given that the critical value is less than or equal to node degree. For node i in graph G with degree d_i, i has the critical value set $H_i = \{h_i : h_i \leq d_i\}$. Taking the product of H_i for all $i \in g$ gives the set of critical value assignments for the graph $H(g)$.

For each critical value assignment, we simulate cascades by updating nodes randomly and activating nodes immediately if exposure is greater than critical value, $k_i \geq h_i$. We filter out graphs where at least one node never activates, which occurs when each inactive node has updated yet none activates.

For these set of "admitted" graphs G^* where all nodes eventually activate, we simulate 100 cascades per graph. For each simulated cascade, we record exposure at each node update. This allows applying the exposure-at-activation (EAA) rule to determine if nodes are precisely measured or not. If node i updates at t and some later t', and $k_i^t + 1 = k_i^{t'}$, then the node is precisely measured. If there is no lower bound (e.g. no update before activation) and the exposure at activation is greater than zero, we call the node imprecisely measured. Innovators which adopt with 0 active neighbors are considered precisely measured here. As we discuss in the main text, this assumption about innovators may not be appropriate in all cases.

2.8 Twitter analysis

The Twitter data used for this analysis was collected for another project via the Twitter REST API between November 2013 and October 2014. This was supported by an NSF grant (SES 1226483). Tweets were localized to a country using the method described in (48). Users in Anglophone countries (US, UK, CA, AU, NZ, SG) were extracted and retweets were filtered out. We selected users who had used one or more of 50 hashtags with moderate-to-high usage. We analyzed the entire timeline of these users which was collected during the data collection process.

2.8.1 Hashtag selection

Hashtags were selected in the following way. The first occurrence of the hashtag in the data was between 2012 and 2014. Two authors (G.B. and C.C) manually examined the top 1000 hashtags by usage, and independently nominated tags which they believed were related to 1) specific offline events, such as #riprobinwilliams; 2) social media phenomena, such as #windows8. The nominated tags were pooled and the authors discussed disagreements.

The list of lowercased tags which are classified as "reinforcement supported" in Figure 2.6 is: benghazi, bringbackourgirls, cantbreathe, drawsomething, election2012, euro2012, firstvine, gop2012, harlemshake, ios6,

jodiarias, justicefortrayvon, linsanity, marriageequality, miley, nfldraft2014, nobama, obama2012, romney, romney2012, romneyryan2012, samelove, snowden, springbreak2014, teamobama, trayvon, trayvonmartin, voteobama, whatdoesthefoxsay, windows8, zimmerman

The list of lowercased tags which are classified as "reinforcement uncertain" in Figure 2.6 is: betawards2014, debate2012, ferguson, goodbyebreakingbad, governmentshutdown, hurricanesandy, inaug2013, ivoted, kony2012, mentionsomebodyyourethankfulfor, newtown, olympics2014, prayersforboston, prayfornewtown, replaceashowtitlewithtwerk, rippaulwalker, riprobinwilliams, sharknado2, worldcupfinal

CHAPTER 3

GRAPH WALKING AND POST-HOC CORRECTIONS FOR ESTIMATING GROUP PROPERTIES IN NETWORKS

We consider the problem of obtaining unbiased estimates of group properties in social networks when using a classifier for node labels. Inference for this problem is complicated by two factors: the network is not known and must be crawled, and even high-performance classifiers provide biased estimates of group proportions. We propose and evaluate AdjustedWalk for addressing this problem. This is a three step procedure which entails: 1) walking the graph starting from an arbitrary node; 2) learning a classifier on the nodes in the walk; and 3) applying a post-hoc adjustment to classification labels. The walk step provides the information necessary to make inferences over the nodes and edges, while the adjustment step corrects for classifier bias in estimating group proportions. This process provides debiased estimates at the cost of additional variance. We evaluate AdjustedWalk on four tasks: the proportion of nodes belonging to a minority group, the proportion of the minority group among high degree nodes, the proportion of within-group edges, and Coleman's homophily index. Simulated and empirical graphs show that this procedure performs well compared to optimal baselines in a variety of circumstances, while indicating that variance increases can be large for low-recall classifiers.

3.1 Introduction

When seeking to understand social interaction online, researchers are commonly faced with a paradox: online data is behaviorally rich but lacks even basic demographic annotation. The lack of demographic information frustrates seemingly straightforward questions: for instance, what is the gender breakdown on an online platform? Since many important social science questions require demographic data, classifiers are commonly used to predict node-level attributes such as gender (144; 41; 118; 65), education (59), age (134), race (131; 129), income (122; 52), or political affiliation (1; 15). However, classifiers introduce error which can bias estimates of group properties, from demographic distributions to cultural homophily.

Adding to the challenge posed by limited demographic information, studies often rely on convenience samples of the underlying social network. The combination of classification error and non-representative sampling poses substantial challenges for obtaining valid estimates of group properties.

We propose a framework, which we term AdjustedWalk, to address these dual problems in order to obtain unbiased estimates of group properties in networks. The idea is to combine re-weighted random walk sampling (also called respondent driven sampling) (77; 152) with quantification learning (68; 69). The sampling method provides the information necessary for inference over nodes and edges. A subset of the nodes is labeled and

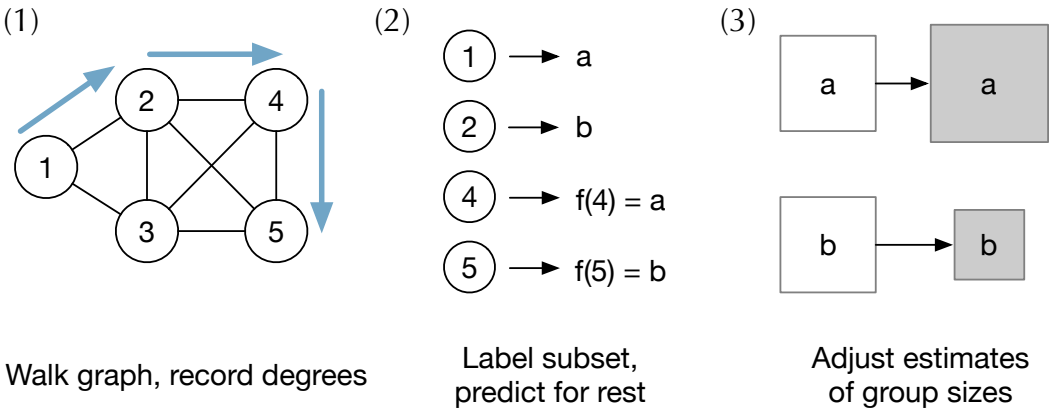

Figure 3.1: A demonstration of the steps in the AdjustedWalk process. First, the graph is randomly walked and node degrees are recorded. Second, a subset of the walked nodes are given ground-truth labels (nodes 1 and 2) and a machine learning model is used to predict for the remaining nodes in the sample (nodes 4 and 5). Third, the mean of the relevant quantity (e.g. proportion in b) is estimated using the RWRW estimator and then adjusted to remove bias.

classified, and a post-hoc correction is applied to correct for classification bias at the group level.

We assume that the sampling procedure starts with an arbitrary node in an undirected graph, and that node labels are predicted by a possibly biased classifier with a known error rate. We show that the framework proposed performs well relative to baselines in four estimation tasks: group proportions, within-group edge proportions, group visibility (the proportion of the minority group in the top 20% of the degree distribution) (104), and Coleman's homophily index (45).

The problem we study has three parts: drawing a representative sample from a network, building a classifier to predict node attributes, and correcting classification error to obtain unbiased estimates of group proportions. Many previous studies have examined each of these parts individually. Research on applications of respondent driven sampling (RDS) to online social networks has provided a robust toolkit for sampling online social networks when demographics are freely available from the site itself (152; 80; 77; 78; 79). On the other hand, many studies have addressed the task of predicting demographics with high observation-level accuracy, with less attention paid to the representativeness of the sample or group-level estimates (52; 65; 53; 173; 15; 122; 41; 134; 118; 59; 82; 1; 177; 129). Finally, a literature on quantification learning (68; 69; 73) has addressed the problem of using a classifier to make population inferences. We demonstrate that combining these separate literatures allows social scientists to make better inferences about social groups in online social networks. This has applications for digital demography and can be useful for online-offline comparisons of social interaction, homophily, and representation (104).

3.2 Summary of contributions

- This paper proposes AdjustedWalk, a framework for estimating group properties in online networks when neither the network nor group labels are known in advance. It contains the following steps: 1) a reweighted random walk (RWRW) of the graph; 2) training a classifier

by labeling a subset of the nodes walked; 3) adjusting group-level error introduced by the classifier to remove bias. The process is visualized in Figure 3.1. Importantly, walking the graph before labeling and classification makes it more likely that population-level inferences will be valid.

- The performance of RWRW sampling is compared to three other plausible sampling procedures: node sampling, edge sampling, and snowball sampling (175). RWRW performs well relative to the optimal sampling method for each task, and performs only slightly worse than node sampling overall despite the absence of a sampling frame.
- An analytical expression for the increased variance of the adjusted estimate is provided, which can be expressed as a function of classifier recall.
- AdjustedWalk is evaluated in a variety of conditions. We examine both simulated and empirical graphs, across a range of sampling fractions and classification accuracies. These analyses demonstrate that relatively small samples perform quite well. When considering classification accuracy, recall scores greater than 0.8 produce reasonable estimates, while recall scores lower than 0.8 quickly increase variance.
- We discuss the conditions under which the results presented here apply to directed graphs in addition to undirected ones.

3.3 Problem setup

3.3.1 Graph and groups

We sample from a graph $G = (V, E)$, where V indicates vertices and E indicates edges. N is the number of nodes in G. Call a node i and an edge from i to j e_{ij}. Let d_i be the degree of i, $D = \sum_i d_i$ the total degree of the graph, and $\bar{d} = D/N$ the average degree of the graph. We assume G is undirected, so that $e_{ij} \in E \implies e_{ji} \in E$. We also assume there are no self links or multiedges, that the graph has at least one triangle, that G is connected, and that all edges have weight 1.

Nodes belong to one of two *social groups*, denoted a and b. By convention, b is the minority group. These groups represent characteristics of individuals such as age, race, ethnicity, gender, or wealth status. p_a is the proportion of nodes belonging to group a, and $\mathbf{p} = (p_a, p_b)$ is the vector of *population proportions*. Since $p_a + p_b = 1$, we will frequently reason about one group with the implication that the same analysis holds for the other one. s_{ab} is the proportion of edges from group a to group b, with $\mathbf{s} = (s_{aa}, s_{ab}, s_{bb})$ the vector of *edge proportions*. Since the graph is undirected, s_{ab} represents all edges with one end in a and the other in b.

When a classifier is used to categorize nodes, we do not observe $\mathbf{p}$ or $\mathbf{s}$ directly. We instead obtain estimates of these quantities after classification error. $\mathbf{m} = (m_a, m_b)$ is $\mathbf{p}$ after classification error, and $\mathbf{t} = (t_{aa}, t_{ab}, t_{bb})$ is $\mathbf{s}$ after

classification error. We use hat notation (e.g. $\hat{\mathbf{m}}$) for estimates resulting from sampling part of the graph and then classifying the sampled nodes.

3.3.2 Classification for quantification

Assume we have a relationship between a set of features x_i and an outcome $y_i \in \{a, b\}, y_i = f(x_i)$. A classifier approximates f, $y_i = \hat{f}(x_i) + \epsilon_i$. The model $\hat{f}$ makes classification errors, which are counted and stored in a confusion matrix

$$F = \begin{bmatrix} \text{count}(\hat{a} \mid a) & \text{count}(\hat{a} \mid b) \\ \text{count}(\hat{b} \mid a) & \text{count}(\hat{b} \mid b) \end{bmatrix}.$$

We use the notation $(\hat{b} \mid a)$ to represent "a true member of a classified as a member b". We will work with the column-stochastic *misclassification matrix*, which we get by column-normalizing $\hat{f}$,

$$C = \begin{bmatrix} c_{\hat{a}|a} & c_{\hat{a}|b} \\ c_{\hat{b}|a} & c_{\hat{b}|b} \end{bmatrix},$$

where $c_{\hat{b}|a} = \text{count}(\hat{b} \mid a)/\text{count}(\hat{a} \mid a) + \text{count}(\hat{b} \mid a))$ represents the probability of a true member of a being classified as b. In practice $\hat{f}$ and C are constructed via cross-validation and holdout sets.

The matrix C is important for removing bias from estimates. We refer

to the off-diagonal elements of C as the "misclassification rate", and note that the diagonals are equivalent to both recall and accuracy. Precision depends on the relative sizes of groups. The misclassification rate represents the probability that a true member of group a is classified as b, and vice versa.

If C is known and only $\mathbf{m}$ is observed, $\mathbf{p}$ can be recovered. C maps $\mathbf{p}$ to $\mathbf{m}$,

$$\begin{bmatrix} c_{\hat{a}|a} & c_{\hat{a}|b} \\ c_{\hat{b}|a} & c_{\hat{b}|b} \end{bmatrix} \begin{bmatrix} p_a \\ p_b \end{bmatrix} = \begin{bmatrix} m_a \\ m_b \end{bmatrix},$$

which can be written compactly as

$$C\mathbf{p} = \mathbf{m}. \tag{3.1}$$

This implies that inverting C provides a way to recover the true population proportions $\mathbf{p}$,

$$\mathbf{p} = C^{-1}\mathbf{m}. \tag{3.2}$$

This procedure is referred to as "adjusted classify and count" in the machine learning literature on *quantification* (68; 69), or recovering population proportions. In general, even high performance classifiers produce biased estimates of group proportions (73), particularly for small groups. While

group proportions may be both over- and under-estimated, classifiers tend to favor larger groups since loss functions are optimized at the observation level. This means that the size of large groups is often overstated by classifier predictions. Models which directly try to estimate group proportions (73) usually under-perform models trained at the observation level and then corrected as in Equation 3.2.

For quantification, an important assumption is required for inference to be valid from the individual to group level (69).

Assumption 1. *Stable conditional feature distribution*

$$P_{train}(X = x_i|Y = y_i) = P_{population}(X = x_i|Y = y_i) \tag{3.3}$$

Assumption 1 states that the feature distribution within each class is the same in training set and population. This can fail for some (but not all) types of sample selection bias (187; 117). For instance, Assumption 1 would fail to hold in a case where one group (e.g. men) tended to be sampled only if they had a certain feature (e.g. owned a car), and this feature was used in the model $\hat{f}$. In this case, car owners would be over-represented in the training set relative to the population. Assumption 1 allows sampling different groups at different rates, as long as the samples drawn from within each group preserve the within-group feature distribution.

If these assumptions hold, then $C_{\text{train}} = C_{\text{population}}$. This implies that bias can be corrected in the entire sample. If the sample is drawn from the population, this provides an inference about the population quantity.

We propose conducting the network walk *before* labeling and classifying cases so that Assumption 1 is more likely to be satisfied. In this case, if the walk draws a valid sample from the population, then Assumption 1 holds. In cases where a classifier is trained on cases not from the sample, Assumption 1 cannot be tested since y_i is unknown in the sample. We also assume that the misclassification matrix C_{train} learned from the labeled set is known exactly. In practice, there is some uncertainty around the confusion matrix, although this can be addressed both through cross validation and the use of holdout set.

3.3.3 Matrix adjustment vs. calibration

We work with the class labels, but an alternative to the matrix method discussed here is to train a second model which calibrates $\hat{f}$. A calibration model takes predicted scores $\hat{s}_i \in [0, 1]$ from $\hat{f}$ and fits a model so that $\hat{s}_i \approx P(y_i = a)$. Gao and Sebastiani (73) examine a variety of quantification methods on many natural language processing tasks and find that the "adjusted classify and count" method used in this paper is statistically indistinguishable from using a calibrated model. However, the calibrated model has somewhat better average performance.

When considering the increased variance of adjustment methods, a closed-form can be derived for the matrix method (shown in the Appendix). A calibration method has the potential to provide lower variance estimates

because it incorporates more information. However, it can also be more difficult to assess the increased variance because of correlated errors between observations which enter into the variance. For simplicity, we study the matrix method and note that variance may be reduced by instead fitting a calibration model.

3.3.4 Graph walking

Re-weighted random walking (RWRW) is a Markov Chain Monte Carlo (MCMC) sampling procedure (80). It is also commonly referred to as Respondent Driven Sampling (RDS) (152), which is a process for applying RWRW to offline social networks such as those of jazz musicians (90) or injection drug users (143).

RWRW allows randomly walking a connected, undirected graph and obtaining a valid estimate of node properties through reweighting by node degree. The basic intuition is to conduct a random walk of the undirected graph G, recording node degrees along the walk. The walk itself provides a random sample of edges, while degree information can be used to approximate a random sample of the nodes.

We present only the RWRW estimator here. A derivation may be found in (80). A more general introduction to the method may be found in (152; 174). For extensions to online network applications, see (148; 112).

Assume we wish to take a mean of a function g over the nodes i of graph G, where g is an indicator function for i being a member of minority group b. Choose a seed node with probability $\pi(i) = d_i/D$, or proportional to the node's degree. Then randomly walk the graph starting from the seed for n steps. Each jump samples node i with probability $\pi(i) = d_i/D$. For each node j along the walk, we record the node's degree d_j and $g(X_j)$, where X_j represents the jth node encountered while walking. The RWRW estimator for estimating the proportion of nodes belonging to b (assume no classification error) is given by

$$\hat{p}_b = \frac{1}{\sum_{j=0}^{n-1} 1/d_j} \sum_{j=0}^{n-1} \frac{g(X_j)}{d_j}. \tag{3.4}$$

Since the random walk naturally samples nodes proportional to degree, we correct for this through the weighting procedure in Equation 3.4. This estimator provides an asymptotically unbiased estimate of p_b.

Note that g may be a continuous valued function as well, allowing estimates of mean degree and the degree distribution (see (148) for an example).

In addition to a mean of g over the nodes of the graph, the RWRW process records information which may be used to estimate the distribution of g. We discuss the details in the Appendix, and use this fact to estimate quantiles of the degree distribution for estimating node visibility.

3.3.5 Walking in a directed graph

The RWRW procedure relies on an important property of walks in undirected graphs: the probability of being at a given node at any time is proportional to that node's degree. However, many online networks are directed, such as Twitter follow relationships. If the directed graph can be redefined as an undirected one, than a walk can still be conducted. For instance, if both inlinks and outlinks can be accessed when a node is visited, then the combined inlink-outlink graph can be walked as if it were undirected. If this is not possible, additional approaches are possible but beyond the scope of this paper (see (79) for suggestions).

3.4 Results

We study four outcomes: proportion of nodes in the minority group, the fraction of in-group edges in the minority group, the percentage of minority group members in the top 20% of the degree distribution (visibility), and Coleman's homophily index.

Node and edge proportions are straightforward measures. Visibility has been studied in recent work (104; 175). It can indicate lack of status among minority group members. For instance, if the minority group comprises 20% of the population but only 10% of the top quintile of the degree distribution, minority group members are systematically underrepresented in

the highest status positions.

Coleman's homophily index (45) has long been employed by social scientists. It treats random mixing as a baseline, with a value of 1 indicating perfect homophily and -1 indicating perfect heterophily. For group a, this measure is defined as

$$H_a = \left\{ \begin{array}{ll} \frac{s_a - p_a}{1 - p_a} & \text{for } s_a - p_a \geq 0 \\ \frac{s_a - p_a}{p_a} & \text{for } s_a - p_a < 0 \end{array} \right\}. \tag{3.5}$$

3.4.1 Simulation parameters

We follow the procedure in (175) to generate homophilous power-law graphs for the simple case of two groups, a and b. Graphs have 10,000 nodes, mean degree of 8, minority group fraction of 0.2, ingroup preference parameter of 0.8 (strong ingroup preference). The graph generation procedure balances ingroup preference with preference for links to high degree nodes. Note that the preference parameter is not the Coleman homophily index, but is correlated with the homophily index.

These graph parameters present a challenge for the technique presented in this paper, since RWRW is known to have higher variance in homophilous networks.[1]

[1]We also conducted simulations with minority group sizes of 0.35 and 0.5, and ingroup preferences of 0.2 (heterophily) and 0.5. The case we present is on balance the most challenging, although heterophilous graphs can present difficulties as well. We omit these ad-

Misclassification rates (off-diagonals of C) considered are 0%, 10%, 20% and 30%. We discuss the 20% case most frequently because it is high enough to present a challenge for our method and low enough to correspond to much published research. We draw samples of size 1000 to 3000, in increments of 500.

We compare the proposed RWRW method with three other sampling methods that have been used in recent literature: random node sampling, random edge sampling, and snowball sampling (175). These methods provide reasonable comparisons: we expect node sampling to outperform RWRW for node-level tasks and edge sampling to outperform RWRW for edge-level tasks. Snowball sampling is included since it is a convenience sampling method that was commonly used before RWRW was developed.

3.4.2 Group proportions and visibility

Figure 3.2 shows node proportion and visibility estimates across the four sampling methods for a 20% misclassification rate and a sample size of 3000[2]. Performance for each sampling method is presented for three cases: no classification error, classification error with no correction, and classification error with correction. We present error distributions rather than the standard normalized root mean squared error (148; 112) in order to assess

ditional cases for brevity, and because homophilous graphs are the case we are most often faced with empirically.

[2]The process for estimating the degree distribution for visibility is described in the Appendix.

bias.

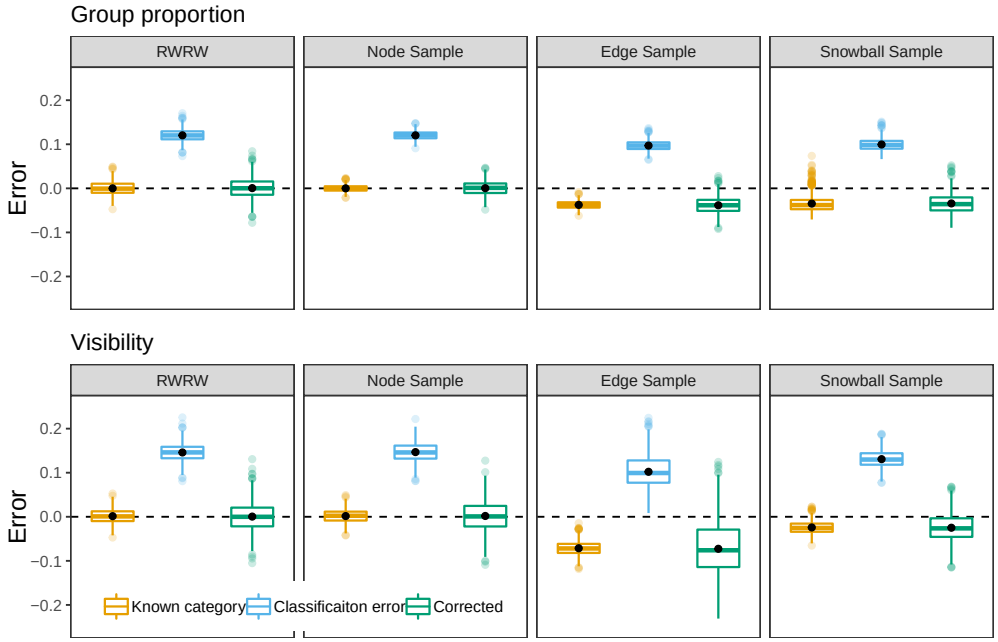

Figure 3.2: Estimates of group proportions and minority group visibility across sampling methods for 20% classification error and sample size of 3000. For each plot, estimates are presented for three cases: no classification error, classification error, and corrected classification error. The dashed line at 0 indicates no error, and the black dot for each estimate indicates the mean. Whiskers represent 95% of the distribution. RWRW performs well in both cases, while node sampling has the lowest variance as expected. Correcting for classification error removes bias only for RWRW and node sampling, while edge sampling and snowball sampling demonstrate bias even after correction.

For group proportions, node sampling and RWRW are both unbiased, and node sampling has a lower variance as expected. Edge and snowball sampling perform poorly. For visibility, RWRW performs the best since it does not have bias in the corrected case, while applying the correction to

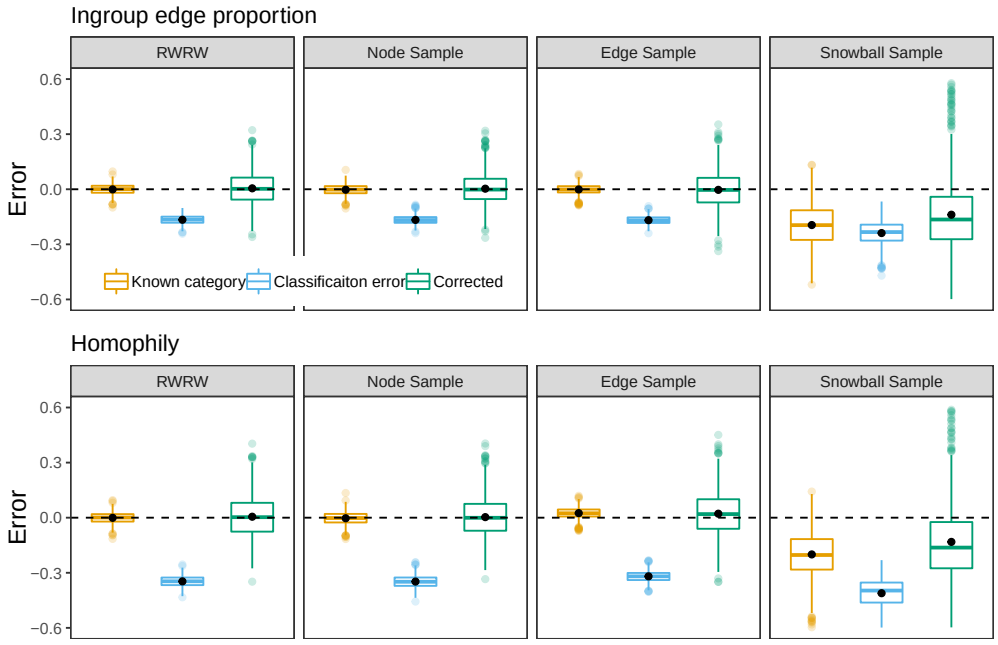

Figure 3.3: Estimates of ingroup edge proportions and group homophily across sampling methods for 20% classification error and sample size of 3000. For each plot, estimates are presented for three cases: no classification error, classification error, and corrected classification error. The dashed line at 0 indicates the true value, and the black dot for each estimate indicates the mean. Whiskers represent 95% of the distribution. Ingroup edge proportions are sampled well by RWRW, node sampling, and edge sampling. Homophily is more difficult but is well captured by both RWRW and node sampling.

node sampling does not remove bias.

A 20% misclassification rate without correction introduces a large magnitude error. While p_b is 20% in the population, misclassification causes this to rise to around 30% in all cases, an inflation of 50%. Since uncorrected estimates are low variance, a draw from an uncorrected distribution is likely to

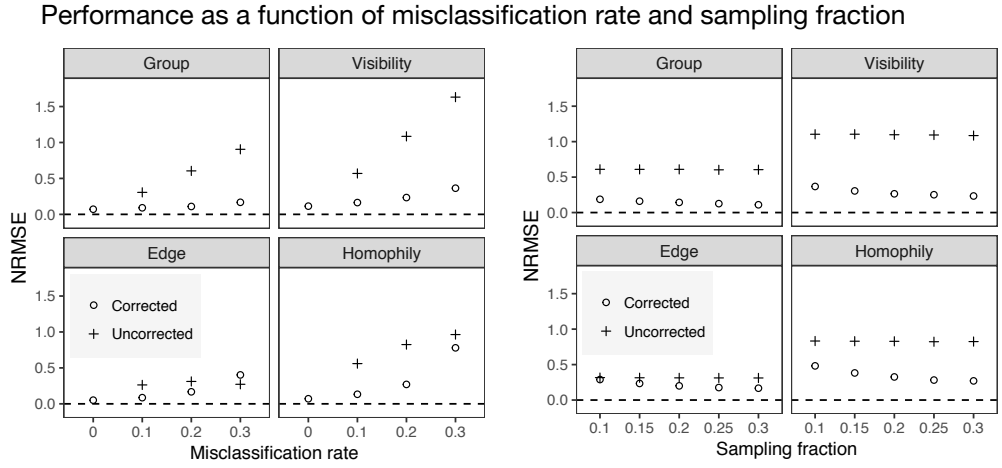

Figure 3.4: Performance of RWRW as a function of misclassification rate and sampling fraction. For misclassification rate, a fixed walk size of 3000 is chosen. for a sample size of 3000. Circles indicate that bias correction has been applied, while plus signs indicate no correction. For group proportions and visibility, bias correction reduces NRMSE by a large amount. For edge proportions, the NRMSE reduction is smaller and disappears completely for misclassification rates of 0.3. Homophily shows large gains at misclassification rates of 0.1 and 0.2, and smaller gains at 0.3. In all cases, increasing sample size produces better after-correction estimates while barely changing before-correction estimates.

produce a poor estimate of the true value. In cases where assessing the size of a minority group is of importance, the upwardly biased estimates can lead to conclusions that understate the differences between groups. Smaller groups are likely to incur larger relative upward biases, since classifier loss functions penalize errors at the observation level and therefore tend to make errors at greater rates on smaller groups.

3.4.3 Edge proportions and homophily

Figure 3.3 shows error for estimates for edge proportions and homophily. The variance for these measures is larger than for the node-level measures (note the rescaled axes compared to Figure 3.2).

Homophily in particular is a difficult inference task since it combines estimates of both group sizes and interaction rates. The error magnitude for uncorrected homophily estimates is large, averaging -0.35 on a -1 to 1 scale. Crossing 0 for the homophily score causes a qualitative difference in interpretation. This error can easily turn an insular minority into a gregarious plurality. In the graphs studied here, the true average homophily value is 0.42, meaning that the average error introduced by misclassification is about 83% of the true value.

The corrected estimates remove bias in exchange for an increase in estimate variance. For this level of misclassification (20%), we argue researchers should take the variance in exchange for bias reduction. After correcting for classification error, RWRW produces estimates with error drawn from $\mathcal{N}(0.005, 0.136)$ while the 95th percentile of the uncorrected error distribution is -0.282 (this is the "top" closest to the no-error line). This indicates that nearly all draws (about 98%) from the corrected distribution are lower-error than even the best draws from the uncorrected one.

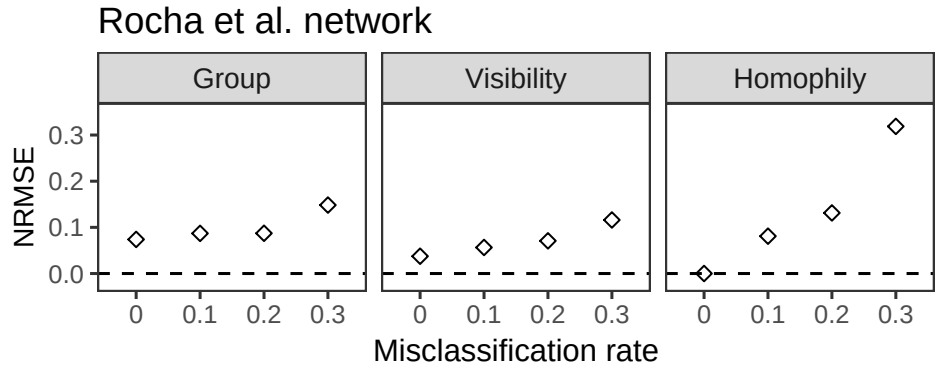

Figure 3.5: Performance of RWRW after bias correction on the sexual contact network in (149) as misclassification rate is increased, with sample size given in Table 3.1. The denominator of the NRMSE is 0 for edge proportions, so we omit that measure. Akin to other studied graphs, this network shows reasonable performance at low misclassification rates, although group proportion NRMSE is somewhat higher than expected. This is potentially caused by the perfectly heterophilous nature of the graph. At high misclassification rates, NRMSE becomes quite large for homophily and begins to rapidly increase for other measures.

3.4.4 Sample size and misclassification rates

Both increasing sample size and building better classifiers have costs. Figure 3.4 reports the sensitivity of RWRW estimates to sample sizes and classification error rates. Corrected RWRW estimates are unbiased, and are compared to uncorrected estimates, which are biased but have lower variance. The metric of performance used is the normalized root mean squared error (NRMSE), given by

$$NRMSE = \frac{\sqrt{E[(\hat{\theta} - \theta)^2]}}{\theta}.$$

When increasing classification error (Figure 3.4, first panel), two patterns emerge. For group proportions and visibility, bias correction substantially reduces NRMSE at all misclassification rates. On the other hand, the NRMSE reduction is modest for edge proportions at misclassification rates of 0.1 and 0.2, while the uncorrected estimate has lower NRMSE at a misclassification rate of 0.3. For homophily, there is substantial reduction in NRMSE at misclassification rates of 0.1 and 0.2, with a more modest correction at misclassification rate of 0.3.

When examining NRMSE response to sample size (Figure 3.4, second panel) while holding misclassification at 20%, we find an interesting pattern: increasing sampling fraction has little effect on the NRMSE of the uncorrected estimate for all measures, but reduces NRMSE for the corrected estimates. This indicates that increasing sampling fraction without correcting for classifier error may not bring estimates closer to the truth. In the simulated graphs here, tripling the sample size hardly changes the error without correction.

3.5 Empirical graphs

We study two empirical graphs, the social network Pokec (166) and the sexual contact network studied in (149). The Pokec graph (166) was crawled from an online social network in Slovakia and comes labeled with gender and age attributes. Age was binarized into "over 28 years old" or not, corre-

	Pokec gender	Pokec age	Sexual contact
Number of nodes	1,198,235	764,845	15,810
Number of edges	8,312,749	4,172,385	38,540
Group studied	Female	> 28 years old	Female
Group proportion	51.3%	22.2%	39%
Group homophily	0.034	0.344	-1
Sample size	25,000	25,000	3,000

Table 3.1: Summary statistics for empirical graphs

sponding roughly to the top quintile of the distribution. The sexual contact network (149) was collected in Brazil and represents links between high end escorts and their clients. We study gender as our category of interest, and note that this network is perfectly heterophilous (all ties are between men and women). Summary statistics for both networks can be found in Table 3.1. These graphs were preprocessed by limiting the graph to mutual ties in the largest connected component, and then deleting nodes which had missing values for the relevant demographic. In the case of the Pokec graph, where two different demographics are examined, we created separate graphs for each demographic.

Figures 3.5 and 3.6 present results for these empirical graphs using RWRW and bias correction as misclassification rate is increased. Performance is comparable to the simulated graphs. For the low-homophily graph (Pokec gender), sampling is quite easy at all misclassification levels. The sexual contact network presents a possible challenge since it is perfectly heterophilous, but performance is in line with the other graphs considered. The Pokec age graph has a moderate amount of homophily, although sam-

pling a relatively small proportion of the network (25,000 out of 764,000 nodes) produces reasonable results.

3.6 Limitations

There are several limitations to the work presented here, which researchers should consider carefully. Most importantly, we have assumed that the classifier error rate is known. Cross validation and holdout sets can be used to estimate this quantity, but empirically some uncertainty will still exist. An alternate route may be to train a calibrated model, although then the challenge becomes estimating the calibration accuracy. In some cases, either with small training sets or difficult classification problems, there may be substantial uncertainty about classifier performance.

Beyond this issue, online networks present data quality issues which may impact the validity of estimates. For instance, bots or inactive accounts which appear to belong to a certain demographic group can bias estimates. Accounting for bias from these factors can be difficult.

3.7 Conclusion

Classifiers offer an opportunity for social scientists to study demographics and other attributes online, in combination with rich behavioral data

often absent from offline surveys. However, care must be taken in order for classification to yield accurate population-level estimates, particularly in networked online settings where the sampling frame is unknown. The framework we study in this paper provides a method to obtain unbiased estimates of group proportions in this setting. Additionally, it can be applied in many domains and requires relatively small adjustments to existing practice. This offers new opportunities for digital demography and comparisons of online settings with offline survey data.

3.8 Appendix

3.8.1 Variance of corrected estimates

Consider the simple case of classifying group proportions with two groups. We obtain a sample from the population with true proportions $\hat{p}$ and estimated proportions $\hat{m}$. Multiplying out Equation 3.2 gives expressions for for the estimated $\hat{p}_a$ and $\hat{p}_b$,

$$\hat{p}_a = \frac{\hat{m}_a c_{\hat{b}|b} - \hat{m}_b c_{\hat{a}|b}}{\det(C)}, \quad \hat{p}_b = \frac{\hat{m}_b c_{\hat{a}|a} - \hat{m}_a c_{\hat{b}|a}}{\det(C)}. \tag{3.6}$$

The variance of the mean is given by,

$$\text{Var}(E[\hat{p}_a]) = \text{Var}(\frac{E[\hat{m}_a] - c_{\hat{a}|b}}{\det(C)}) \tag{3.7}$$

$$= \frac{1}{\det(C)^2} \text{Var}(E[\hat{m}_a]), \tag{3.8}$$

where we use the assumption that C is constant to pull it out of the variance expression.

When there is no classification error, $\det(C) = 1$, and when the classifier guesses randomly (.5 in every cell), $\det(C) = 0$ and the variance is undefined. $\det^2(C)$ provides a clear quantification of the variance increase we expect for group proportions. For instance, if $C = [0.8, 0.2; 0.2, 0.8]$ with $\det^2(C) = 0.6$, we expect a variance increase of $1/0.6^2 = 2.78$. If classifier performance improves to $C = [0.9, 0.1; 0.1, 0.9]$, the variance increase is $1/0.8^2 = 1.56$.

$\text{Var}(E[\hat{m}_a])$ comes from the random walking procedure itself and is generally not known in closed form. Two methods for closed-form variance have been proposed (174; 80). The Volz-Heckathorn (174) estimator is biased but provides reasonable estimates in practice. The Goel-Salganik (80) variance estimator relies on knowing the homophily of the network. Bootstrap resampling methods based on creating "synthetic chains" from the estimated transition matrix between groups have also often been used (92).

Simulations of various RWRW estimators (75) show that factors such as a non-equilibrium seed selection, group homophily, and number of waves from each seed affect both the bias and variance of RWRW estimates. Gen-

erally, one long chain provides the best results, rather than many shorter chains. It is easier to sample from lower homophily networks, and equilibrium seed selection (proportional to degree) is useful if one must use relatively short chains. Otherwise, if chains may be long, a burn-in period can be used to simulate equilibrium seed selection.

3.8.2 Correcting visibility

While RWRW gives the mean of g over the population of nodes, the distribution of g is often an object of interest. For instance, if we wish to estimate the proportion of minority group members in the top 20% of the degree distribution, we need to estimate the joint distribution of $(g(i), d_i)$ and take nodes in the top 20% of the distribution of d_i.

Fortunately, importance resampling (151; 125) based on the data obtained during an RWRW walk provides a method to do this. If we know node i with degree d_i is sampled with probability $\pi(i)$ and we want to sample it with probability $1/N$ (a uniform distribution over the nodes), then we construct an importance weight using the ratio of desired over actual distributions

$$\frac{1/N}{\pi(i)} = \frac{D}{Nd_i} = \frac{\bar{d}}{d_i} = w_i. \tag{3.9}$$

w_i provides a resampling weight for node i. We then normalize $w_i / \sum_j w_j$ and

resample data $(f(i), d_i)$ according to this probability to approximate draws from the desired distribution $1/N$.

An importance resample produces a distribution of $(d_i, g(i))$ which mirrors the distribution in the population. We then sort the resampled nodes by degree d_i and take the proportion in the top 20% of degree where $g(i) = b$, or where i is a member of the minority group. In the case with no classification error, this procedure produces an unbiased estimate of the fraction of minority group members in the top 20% of the degree distribution.

With classification error, we need to add an additional step to correct the importance resample. Call $\hat{m}_b^{\mathcal{I}}(20)$ the measured proportion of group b in the top 20% of the degree distribution in importance resample $\mathcal{I}$. Likewise, there is a vector that contains measures for all groups $\hat{\mathbf{m}}^{\mathcal{I}}(20)$. Then we can use a procedure similar to Equation 3.2 to correct the importance resample proportions:

$$\hat{\mathbf{p}}^{\mathcal{I}}(20) = C^{-1}\hat{\mathbf{m}}^{\mathcal{I}}(20). \tag{3.10}$$

To see when $\hat{\mathbf{p}}^{\mathcal{I}}(20)$ is unbiased, repeat the reasoning for estimating the population proportion $\hat{\mathbf{p}}$ above. This shows that $\hat{\mathbf{p}}^{\mathcal{I}}(20)$ is unbiased when the importance resample provides an unbiased estimate of $\mathbf{m}^{\mathcal{I}}(20)$. A similar argument applies for the variance, and the determinant of C may be used to estimate the increase in variance.

3.8.3 Correcting edge proportions

Correcting estimates of ties between groups presents a more substantial challenge than correcting group proportions. Akin to C, there is a dyadic misclassification matrix M which maps $\mathbf{s}$ to $\mathbf{t}$,

$$M\mathbf{s} = \mathbf{t}, \tag{3.11}$$

where

$$M = \begin{bmatrix} c_{\hat{a}|a}^2 & c_{\hat{a}|a}c_{\hat{a}|b} & c_{\hat{a}|b}^2 \\ 2 * c_{\hat{a}|a}c_{\hat{b}|a} & c_{\hat{a}|a}c_{\hat{b}|b} + c_{\hat{a}|b}c_{\hat{b}|a} & 2 * c_{\hat{a}|b}c_{\hat{b}|b} \\ c_{\hat{b}|a}^2 & c_{\hat{b}|a}c_{\hat{b}|b} & c_{\hat{b}|b}^2 \end{bmatrix},$$

which implies that we can use a technique similar to Equation 3.2 at the dyad level

$$\mathbf{s} = M^{-1}\mathbf{t}. \tag{3.12}$$

In practice, we obtain a sample $\hat{\mathbf{t}}$ rather than $\mathbf{t}$ for the entire graph, which is then used to estimate true edge proportions $\hat{\mathbf{s}}$. $\hat{\mathbf{s}}$ is unbiased when the sampling method employed produces unbiased estimates of $\mathbf{t}$. If $B = M^{-1}$, the expectation for $\hat{s}_a$ is given by

$$E[\hat{s}_{aa}] = b_{00}E[\hat{t}_{aa}] + b_{01}E[\hat{t}_{ab}] + b_{02}E[\hat{t}_{bb}]. \tag{3.13}$$

As in the node case, we can expect the variance of $E[\hat{s}_{aa}]$ and $E[\hat{s}_a]$ to increase when applying classification bias correction. Simulations below indicate that variance inflation for $E[\hat{s}_a]$ is larger than for $E[\hat{p}_a]$. Note that $\hat{s}_a = 2\hat{s}_{aa}/(2\hat{s}_{aa} + \hat{s}_{ab})$ is unbiased under the same conditions as $\hat{s}_{aa}$.

If $B = M^{-1}$, then the variance for $\hat{s}_{aa}$ is

$$\mathrm{Var}(E[\hat{s}_{aa}]) = b_{00}^2 \,\mathrm{Var}(E[\hat{t}_{aa}]) + b_{01}^2 \,\mathrm{Var}(E[\hat{t}_{ab}]) + b_{02}^2 \,\mathrm{Var}(E[\hat{t}_{bb}]). \tag{3.14}$$

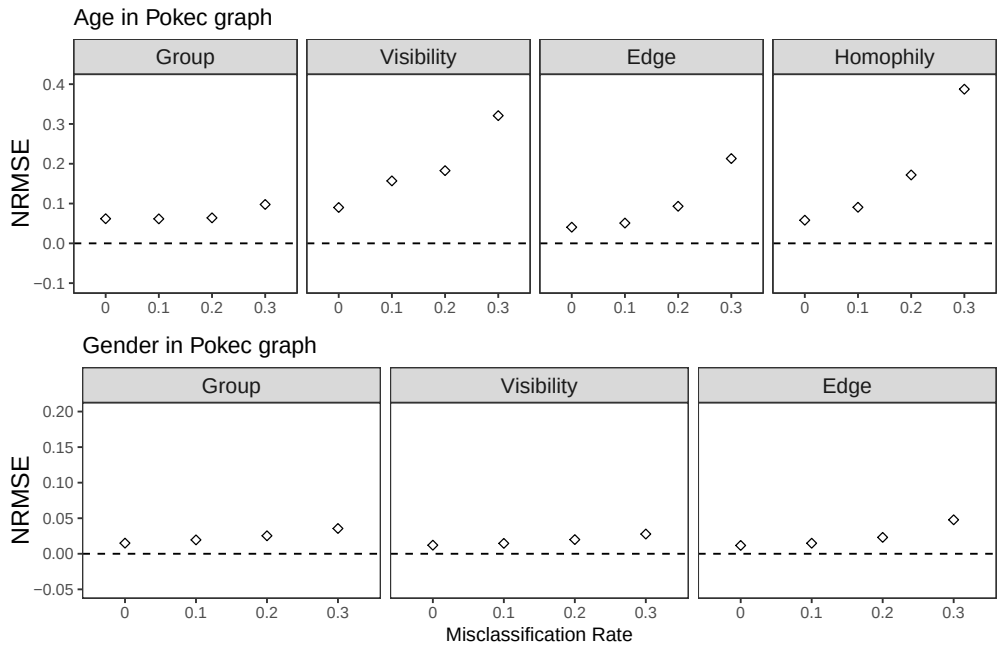

Figure 3.6: RWRW performance after bias correction on two graphs derived from the Pokec network (166), with sample size given in Table 3.1. The top panel presents results for age, which is moderately homophilous. The pattern of quickly accelerating error at misclassification rates of 0.3 appears again. The bottom panel shows performance on gender in the Pokec graph, where we omit the homophily measure since its true value is very close to 0 which inflates NRMSE. This is the homophily value closest to 0 that we study in the paper, and as expected performance is quite good (note the y-axis on the bottom panel). Ingroup edge proportion is typically difficult to sample, but even with a misclassification rate of 0.3, performance is strong. This suggests that when graphs do not display large amounts of homophily, less accurate classifiers can be used or smaller samples drawn.

CHAPTER 4

USING NETWORK-AWARE PREDICTIONS TO ESTIMATE HOMOPHILY

In online social networks, it is common to use predictions of node categories to estimate measures of homophily and other relational properties. However, online social network data often lacks basic demographic information about the nodes. Researchers must rely on predicted node attributes to estimate measures of homophily, but little is known about the validity of these measures. We show that estimating homophily in a network can be viewed as a dyadic prediction problem, and that homophily estimates are unbiased when dyad-level residuals sum to zero in the network. Node-level prediction models, such as the use of names to classify ethnicity or gender, do not generally have this property and can introduce large biases into homophily estimates. Bias occurs due to error autocorrelation along dyads. Importantly, node-level classification performance is not a reliable indicator of estimation accuracy for homophily. We compare estimation strategies that make predictions at the node and dyad levels, evaluating performance in different settings. We propose a novel "ego-alter" modeling approach that outperforms standard node and dyad classification strategies. While this paper focuses on homophily, results generalize to other relational measures which aggregate predictions along the dyads in a network. We conclude with suggestions for research designs to study homophily in online networks.

4.1 Introduction

Researchers have long sought to understand the pattern (124; 127), causes (111), and consequences (23) of *homophily* (45), or the tendency for like to associate with like. Measuring the similarity of nodes along along racial (124; 127; 162; 37; 133; 185), ethnic (99), gender (129; 167; 39), social status (111), cultural (116), emotional (97), political (88; 11; 47), and socioeconomic (60) lines is a core area of research in network science (128). In online networks, understanding the structure of homophily is crucial for understanding echo chambers (16), access to information, and network integration (105).

Online social networks present a particular challenge for understanding this fundamental aspect of networks: demographic and attitudinal information is often absent. The common strategy for addressing this is to predict demographics or attitudinal attributes (37; 129) based on publicly available information such as names, profile photos, text, or other information (15; 1; 100; 39; 129). This information is combined with *ground truth* labels (known values of the category of interest for a set of individuals), and a supervised learning classifier (132) is then used to predict the node category for all nodes in the network.

Although predicted node attributes are widely used to empirically measure homophily and other relational properties (37; 129; 97; 26; 98; 11; 47; 39), there is a lack of theoretical methodological work investigating when and

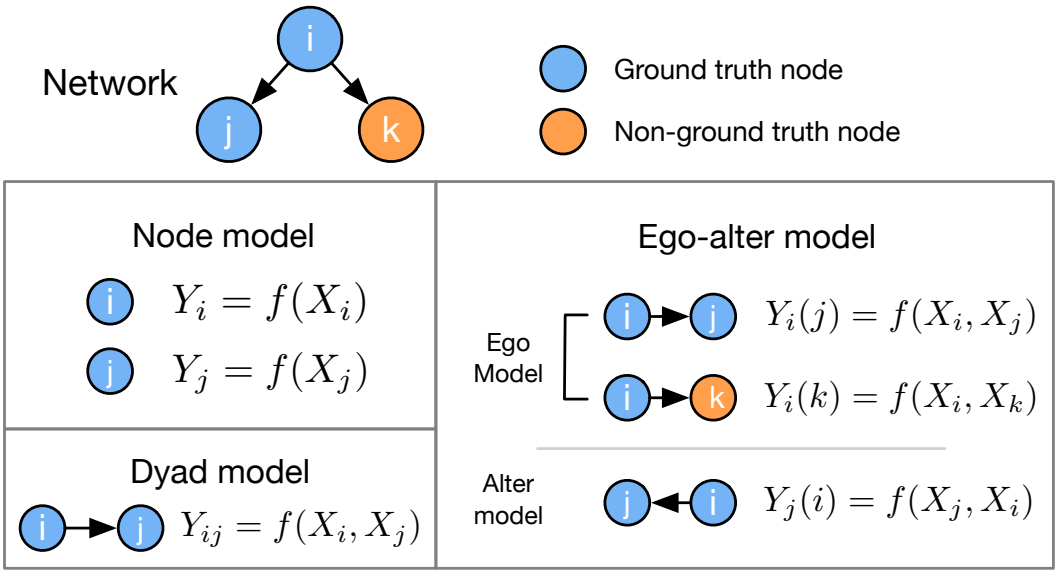

Figure 4.1: A depiction of how the models studied in this paper (node, dyad, and ego-alter) turn a simple network into a prediction task. The node model trains on each ground truth node using only ego features X_i. The dyad model trains on each ground truth dyad using features from both ego and alter, X_i, X_j. The ego-alter model fits an "ego model" predicting ego's category for each link from ego, and a second "alter model" for each link incoming to each alter. Both ego and alter models incorporate features of ego and alter, X_i and X_j, for each prediction, producing a "family of predictions" for each node.

to what extent the predictions produce reasonable estimates (20). The most common strategy is to choose a model which maximizes a node-level measure of classification performance. Because of the complexities of networks, this criterion is not sufficient for reliable estimation of relational measures such as homophily.

In this paper, we formalize the homophily estimation problem as a

dyadic prediction problem. This expression clarifies the difficulty in using node-level predictions to draw larger inferences: node residuals are multiplied along edges, magnifying a node's residual in proportion to its degree, and opening the door to residual autocorrelation along dyads. Theoretically, we should expect correlated errors along edges due to latent homophily (57), or the correlation of unobserved factors in the network. For example, name-based classifiers (100; 39; 98) have frequently been used for gender classification. If a name-based method codes "Leslie" as "woman" because this is more common in the population, yet for a specific network community the name "Leslie" tends to indicate "man", model errors will be correlated with the network and can bias overall gender homophily estimates. This issue is compounded by the highly skewed degree distributions of online networks (106), which introduces the additional possibility that misclassification for high degree nodes will bias the overall estimate.

We show that dyad-level predictions produce unbiased homophily estimates. However, such estimates are often high-variance for a given ground truth labeling budget[1]. This motivates a two-step modeling procedure (called "ego-alter") which predicts the category of a node from the perspective of each one of its network neighbors. This allows incorporating network information beyond the ego level, while still using standard modeling tools such as logistic regression. This ego-alter approach is theoretically less biased than a node-level model, and across a range of simulations

[1]This fact suggests that even when possessing the "ideal" random edge sample with labels, modeling should be employed as a variance reduction technique.

outperforms both node-level and dyad-level models in overall error. Figure 4.1 visually compares these three approaches. While we primarily study homophily with node demographics in mind, results here apply to a wide range of networked outcomes, such as estimating the fraction of people belonging to a certain race/ethnicity experience hate speech in their social media feeds (55).

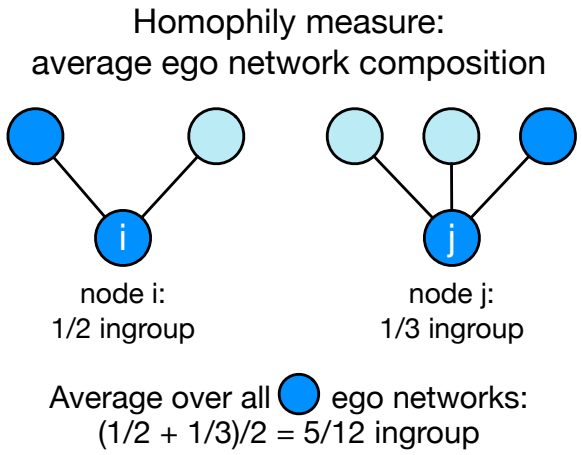

Figure 4.2: We use average ego network composition to capture homophily from the perspective of the dark blue nodes. We assume node categories (light and dark blue) must be predicted with a model. This estimand is expressed analytically in Equation 4.2.

4.2 Homophily Measure

We study the average fraction of ego networks composed of ingroup members (visualized in Figure 4.2). Average ego network composition has been extensively studied in sociology, primarily in research concerning the Gen-

eral Social Survey network module (124; 127). 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[2] (124).

Average egonet composition can be written as a sum over ego networks, taking into account the category of both ego and alter. Let Y_i indicate the category of i, for instance in the case of racial homophily, category a may indicate Black, category b may indicate White, and so on. Without loss of generality, assume that we are studying a binary outcome where $Y_i \in \{a, b\}$. For compactness, we write Y_i^a to denote $\mathbb{1}_{Y_i=a}$. Then the average fraction of group a's ego networks which are composed of alters in group a (Figure 4.2) can be written,

$$H^{aa} = T[Y_i^a]^{-1} \sum_i Y_i^a \frac{1}{D_i} \sum_{j \in \mathrm{N}(i)} Y_j^a, \tag{4.1}$$

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

Note that Equation 4.1 can be re-written as a sum over dyads in the net-

[2]We choose this measure of homophily instead of Coleman's homophily measure (45) because it is not dominated by high degree nodes, although the Appendix shows that results for the average egonet measure apply to Coleman's measure as well.

work by rearranging the summation,

$$H^{aa} = T[Y_i^a]^{-1} \sum_{(i,j)\in E} \frac{1}{D_i} Y_i^a Y_j^a = T[Y_i^a]^{-1} \sum_{(i,j)\in E} \frac{1}{D_i} Y_{ij}^{aa}, \tag{4.2}$$

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

4.3 Dyadic regression as an unbiased estimator

Equation 4.2 shows how homophily can be estimated with knowledge of edge categories Y_{ij}^{aa}. Assume edge categories Y_{ij}^{aa} are obtained for a random sample of the edges S, and features correlated with edge categories X_{ij} are available both for the sample S and the population of edges P. Assuming random sampling simplifies the argument, and we discuss deviations from this assumption in the Appendix.

Assume a model predicting Y_{ij}^{aa} is chosen which has the property that the residuals sum to zero in the population and are uncorrelated with features[3] $\sum_{ij} e_{ij} = 0$, $\text{Cov}(e_{ij}, X_{ij}) = 0$. Then this model trained on the sample S provides an unbiased estimate of homophily in the population given that $\frac{1}{D_i}$ is included in X_{ij} as a feature.

[3]For instance, ordinary least squares and logistic regression both have this property, as does any model with mean squared error loss.

The reason for including $\frac{1}{D_i}$ as a variable can be seen by the following argument. First, recall the conditional expectation (CEF) function expression (2): $Y_{ij}^{aa} = \mathbb{E}[Y_{ij}^{aa}|X_{ij}]+e_{ij}$, where $\mathbb{E}[Y_{ij}^{aa}|X_{ij}]$ can be estimated with a model such as OLS. An estimator for Equation 4.2 can be written in terms of predictions,

$$\hat{H}^{aa} = T[Y_i^a]^{-1} \sum_{(i,j)\in E} \frac{1}{D_i} \mathbb{E}[Y_{ij}^{aa}|X_{ij}]. \tag{4.3}$$

We now need to examine when using model predictions in Equation 4.3 provides an answer equal to Equation 4.2 in expectation. This can be done by substituting the CEF into Equation 4.3 to obtain,

$$\hat{H}^{aa} = \underbrace{T[Y_i^a]^{-1} \sum_{(i,j)\in E} \frac{1}{D_i} Y_{ij}^{aa}}_{\text{True value}} - \underbrace{T[Y_i^a]^{-1} \sum_{(i,j)\in E} \frac{e_{ij}}{D_i}}_{\text{Sum of residuals}}. \tag{4.4}$$

This indicates that the homophily estimate will be unbiased when the sum of residuals term is zero.

Condition 2. *When* $\mathbb{E}[\sum_{(i,j)\in E} \frac{e_{ij}}{D_i}] = 0$, *the expectation of the estimate equals the true value,* $E[\hat{H}^{aa}] = H^{aa}$.

Since we assumed a model is used where $\sum_{ij} e_{ij} = 0$ and $\text{Cov}(e_{ij}, X_{ij}) = 0$, if $\frac{1}{D_i}$ is included in X_{ij} then $\mathbb{E}[\sum_{ij} \frac{1}{D_i} e_{ij}] = 0$.

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 OLS, $\frac{1}{D_i}$ is the only required variable in for an unbiased estimate, although this can produce a high variance

estimate. Including robust predictive features is therefore still important for variance reduction and addressing cases of non-random sampling.

4.4 Approximating dyadic regression with node-level data

Sampling and labeling limitations often make collecting a large set of ground-truth dyads infeasible. In this situation, node-level ground truth information can be employed to estimate homophily. We present a two-step modeling strategy which we term "ego-alter" which uses only node-level ground truth information, reduces bias over a standard node-level model, and reduces variance compared to the edge model in Section 4.3. The ego-alter approach is biased, although the magnitude of bias in simulations we examine below is generally substantially less than a node-level model.

The ego-alter model is a hybrid approach: it uses dyadic features X_{ij} to predict the node-level ground truth Y_i and Y_j separately, producing one prediction per edge for both ego and alter (see Figure 4.1 for a visual representation). This has the benefit of reducing bias in two ways: first, predictions for Y_i and Y_j are improved by including a richer set of features which improves prediction accuracy; second, it reduces bias by reducing dyadic residual autocorrelation.

Since $Y_{ij}^{aa} = Y_i^a Y_j^a$, H^{aa} can be estimated with the product of node predic-

tions,

$$\begin{aligned}\hat{H}^{aa} &= T[Y_i^a]^{-1} \sum_{(i,j)\in E} \frac{1}{D_i} \mathbb{E}[Y_i^a | X_{ij}] \mathbb{E}[Y_j^a | X_{ij}] \\ &= T[Y_i^a]^{-1} \sum_{(i,j)\in E} \frac{1}{D_i} (Y_i^a - e_i^a)(Y_j^a - e_j^a),\end{aligned}$$

where the second line follows by substituting the CEF. This can be expressed as the true homophily value plus two bias terms,

$$\hat{H}^{aa} = \underbrace{\frac{\sum_{(i,j)} \frac{1}{D_i} Y_i^a Y_j^a}{T[Y_i^a]}}_{\text{True value}} - \underbrace{\frac{\sum_{(i,j)} \frac{1}{D_i} e_i^a Y_j^a}{T[Y_i^a]}}_{R_1} - \underbrace{\frac{\sum_{(i,j)} \frac{1}{D_i} \mathbb{E}[Y_i^a | X_{ij}] e_j^a}{T[Y_i^a]}}_{R_2}. \tag{4.5}$$

The bias terms R_1 and R_2 both indicate dyadic correlation of the model residuals with neighbor outcomes. Assuming $\frac{1}{D_i}$ is included as a model feature, R_1 and R_2 can be thought of similarly: when model residuals are correlated with neighbor outcomes, the terms will be non-zero. This can happen when models produce pockets of similar errors in the network due to unobserved, network correlated features. When inverse ego degree $\frac{1}{D_i}$ is not included as a model feature, the bias terms become substantially more complex because of the interaction between degree, errors by degree, and errors along dyads.

Equation 4.5 indicates that the estimate equals its true value when $R_1 = R_2 = 0$ or when $R_1 = -R_2$. Note that $R_1 = -R_2$ is unlikely, this is because residuals for e_i and e_j have similar correlations with neighbor true outcomes Y_j and Y_i since both ego and alter models score the entire network[4].

Note that R_2 is the result of combining two terms, since $E[Y_i^a | X_{ij}] e_j^a = Y_i^a e_j^a + e_i^a e_j^a$. This suggests an "augmented" ego-alter model: first, fit an ego

[4]This is confirmed by simulations, where R_1 and R_2 tend to have similar values.

model for i, and then fit an alter model for j which includes the ego predictions $E[Y_i^a|X_{ij}]$ as a feature. This, in expectation, eliminates the R_2 term and reduces bias assuming R_1 and R_2 have the same signs.

4.5 Simulation

We use simulations to evaluate the effectiveness of dyadic regression and ego-alter regression for estimating average egonet composition (Equation 4.2; Figure 4.2). For an outcome Y_i which takes on values $\{a, b\}$, the probability of $Y_i = a$ is simulated as follows:

$$Y_i^a \sim \text{Bernoulli}(p_i), \;\; p_i = \text{logistic}(2X_i + Z_i), \tag{4.6}$$

where X_i and Z_i represent individual and network-correlated features, respectively. The individual-level feature is drawn from a normal distribution, $X_i \sim \mathcal{N}(0, 1)$, while the network feature is the maximum of the individual feature among the neighbors of each node i: $Z_i = \max_{j \in \text{N}(i)}(X_j)$. Z_i is then standardized to follow a normal distribution. This creates outcomes correlated along some dyads in the network, where nodes with large values of X_i "influence" neighbors. If Z_i is omitted, model residuals will be correlated along dyads and bias homophily estimates. The level of homophily generated by these parameters is moderate: the average ego network for group a contains 59% of nodes in group a ($H^{aa} = 0.59$), while Coleman's homophily index for group a is 0.14.

We choose Z_i to be the maximum X_j value among the alters j of ego i to provide a challenging setting for models: the true response is determined at the ego-network level yet models operate at the node or dyad levels, meaning a dyadic regression cannot capture the true functional form of the data generating process. This both approximates real-world scenarios where the data generating process is unknown, and demonstrates the argument about bias in Section 4.3. Five alternative simulation specifications are examined in the Appendix, with qualitatively similar results to this simulation.

Networks with 4000 nodes are generated according to a preferential attachment graph (13) with five links per node and a powerlaw exponent $k = 0.8$. Links are considered bidirected. Preferential attachment graphs have high degree disparities, providing a challenging setting for the estimation task considered here, since model errors on individual high degree nodes can bias estimates. We conduct simulations for both node and edge sampling, selecting 20% of nodes or 2.5% of edges randomly as ground truth cases. This produces roughly 800 ground truth nodes for both types of sampling. Note that sampling nodes into the ground truth set provides some ground truth dyads (and vice versa), meaning both node and dyad models can be fit with either type of sampling.

Using the ground truth sample to estimate a model, we classify all edges and estimate homophily across 500 simulation runs. Model performance is estimated in two ways: bias and absolute error. Bias is the average of $(\hat{H}^{aa} - H^{aa})/H^{aa}$ 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 $|\hat{H}^{aa} - H^{aa}|/H^{aa}$ across all simulation runs. It captures how far estimates tend to be from the true value.

Since both absolute error and bias are normalized, they have the interpretation of "percent error." The bias-variance tradeoff means that we should not expect the method with the lowest bias to also have the lowest absolute error.

We evaluate three types of models: node, dyad, and ego-alter (see Figure 4.1 for a depiction), all fit with logistic regression. For the node model, we examine models with and without network features. For the ego-alter model, we examine both the basic version and the "augmented" version. This gives a total of 5 models, which are described here in terms of their regression equations, where $f(X_i, X_j)$ indicates a main-effects linear model $\beta_0 + \beta_1 X_i + \beta_2 X_j$.

$$
\begin{aligned}
\text{Node (no network)} \quad & Y_i^a = f(X_i) + e_i \\
\text{Node} \quad & Y_i^a = f(X_i, \frac{1}{D_i}, D_i) + e_i \\
\text{Dyad} \quad & Y_{ij}^{aa} = f(X_i, X_j, \frac{1}{D_i}, D_i, D_j) + e_{ij} \\
\text{Ego-alter} \quad & Y_i^a(j) = f(X_i, X_j, \frac{1}{D_i}, D_i, D_j) + e_{i(j)} \\
& Y_j^a(i) = f(X_i, X_j, \frac{1}{D_i}, D_i, D_j) + e_{j(i)} \\
\text{Ego-alter (augmented)} \quad & Y_i^a(j) = f(X_i, X_j, \frac{1}{D_i}, D_i, D_j) + e_{i(j)} \\
& Y_j^a(i) = f(X_i, X_j, \frac{1}{D_i}, D_i, D_j, \hat{Y}_i^a(j)) + e_{j(i)}
\end{aligned}
$$

The notation $Y_i^a(j)$ indicates that we predict i's category for each neighbor j separately, using features of both i and j in the prediction. Ego and alter degree are included in models because they tend to reduce the bias and variance of estimates and are available to researchers conducting network studies.

4.5.1 Simulation results

As shown in Figure 4.3, the default approach of using node-level classifier with no network features performs poorly. Homophily is underestimated by between 10% and 20%, 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 still underestimates homophily by around 3%. This large reduction in bias indicates the importance of including network in-

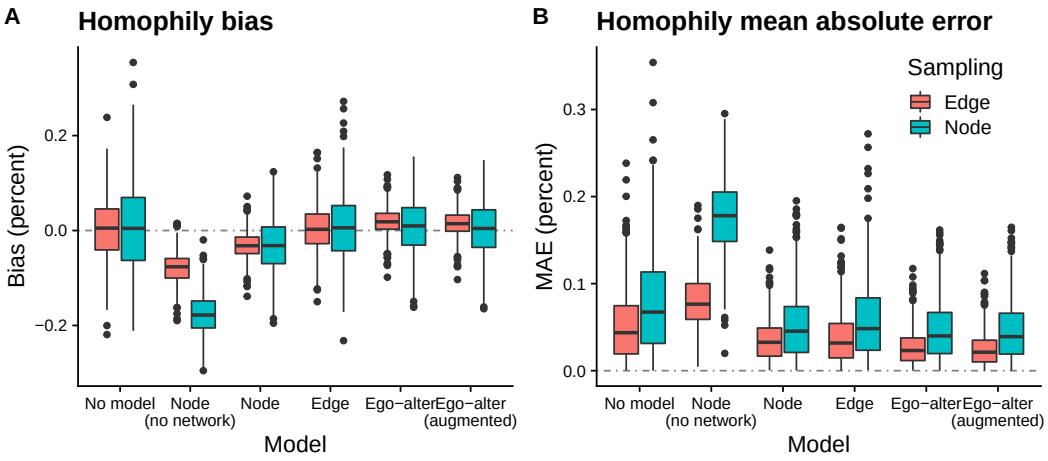

Figure 4.3: The bias and absolute error of homophily estimates using five different models, for random node and random edge sampling. Node level models without network variables display large biases in the presence of network-correlated unobserved features. Including network information reduces this bias, and using edge or ego-alter models reduces this bias further. Note that while dyadic regression is unbiased, it does not provide the lowest error estimates. Since roughly similar numbers of nodes are sampled in both edge and node sampling, edge sampling is more efficient.

formation in the model predicting node categories, while the remaining bias indicates the limitations of a node-level approach in the presence of network-correlated outcomes.

In this simulation, homophily is under-estimated. This indicates that errors tend to be positively correlated along dyads, increasing the sum of residuals term in Equation 4.4 and reducing the overall homophily estimate. In other words, there are pockets of the network where the model errors are similar. An alternative scenario exists where a model produces negatively

correlated dyadic errors and an over-estimate of homophily. An instance where this happens is residual-degree correlation in the network. When high degree nodes have positive residuals and low degree nodes have negative residuals, the overall residual term in Equation 4.4 can be negative and cause an over-estimate of homophily[5].

A dyadic model produces an unbiased estimate of homophily, according with the argument in Section 4.3. However, the dyadic approach does not produce the lowest absolute error. Despite a small amount of bias (around 1%), the ego-alter approach produces lower absolute error on average than the dyadic approach. In alignment with the theoretical argument in Section 4.4, including ego predictions in the alter model reduces bias about 20% on average.

While this simulation uses random sampling, the ego-alter model is also more robust to deviations from random sampling compared to other methods, as shown in the Appendix. In the presence of non-random edge samples, an edge model can be brittle. One potential corrective is weighting the ground truth data, but the often high-dimensional nature of edge features risks large design effects due to the curse of dimensionality (102). Additionally, a "meta-network-correlation" problem can arise, where errors in weights are network correlated.

A more practical approach is to employ a modeling strategy such as ego-

[5]An instance of this phenomenon can be seen in the Section 4.8.1, with the simulation called "degree."

alter which can more flexibly learn class probabilities in a network-aware way. While we intentionally restrict models here to logistic regression with only main effects, more flexible functional forms can also be employed to better approximate class probabilities within subgroups.

4.5.2 Node-level performance and network-level estimands

Machine learning models are usually evaluated on observation-level performance metrics such as precision, recall, and area under the curve (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 dyad-autocorrelated errors that bias homophily estimates.

This can be seen clearly in Figure 4.4, which plots model performance against bias in estimating homophily[6]. Models differ only slightly on traditional performance measures, yet produce large differences in homophily bias. The best model's AUC is 0.8% better than the worst model, yet has a bias reduction of 95% (worst: 17.6% bias; best 0.8% bias).

Note that a meta-analysis of research on demographic classification

[6]Only models which produce node-level category predictions can be evaluated this way, which necessitates excluding the edge model.

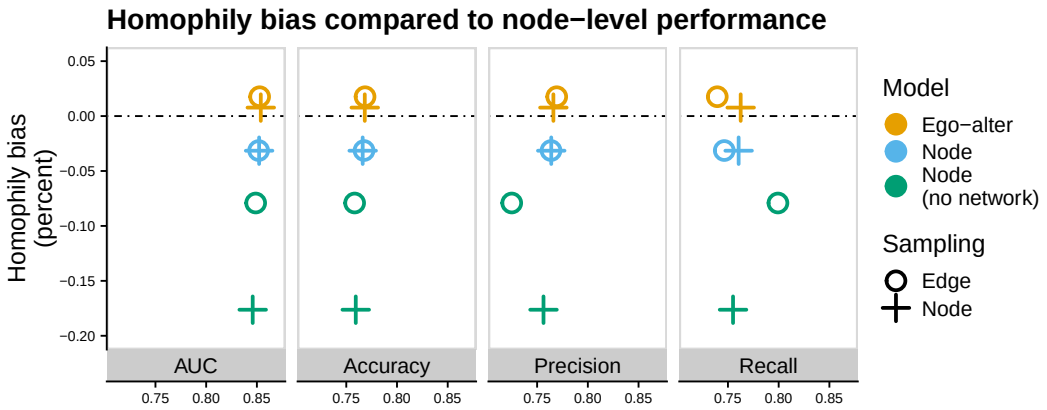

Figure 4.4: Models with similar node-level classification performance produce different levels of bias when estimating homophily. The three models in Figure 4 have average AUCs between 0.846 and 0.853, yet produce average biases ranging from 0.8% to 17.6%. This demonstrates that observation-level classification performance and estimation of relational measures are distinct tasks.

on social media (36) found a median accuracy of 0.81 for predicting race/ethnicity, while simulations presented here have an average accuracy of around 0.77. This indicates that similar biases may be present with the type of classification performance found in real-world tasks.

4.6 Practical advice

When studying homophily in online communities, researchers can potentially improve the quality of estimates in five ways: including network information in models, using the ego-alter modeling strategy, improving model flexibility, sampling edges, and using cross-validation to check for

the presence of network-residual correlation.

First and most importantly, network information should be incorporated into prediction models. Evidence from Sections 4.5 and 4.8.1 indicates the single largest improvement in model performance comes from including degree information ($\frac{1}{D_i}$) in models. The specific information to include is dependent on the estimand, and can even extend to behavioral information when outcomes such as political affiliation are studied in networks. We give an example of applying the process from Section 4.3 to new estimands in the Appendix (Sections 4.8.2 and 4.8.3).

Second, the simulation results consistently demonstrate that the ego-alter modeling strategy performs well both in terms of bias and absolute error. This is true even in the presence of a non-random ground truth sample. Since the ego-alter strategy is new, we recommend that researchers present results from both a node-level model and an ego-alter model, with network information included for both models.

Third, and closely related to the choice of modeling strategy is the choice of model itself: a logistic regression with only main effects in the presence of a non-random sample can produce large biases, as seen with the edge model and non-random edge sampling in the Appendix (Table 4.1). A more flexible model can mitigate this by better learning conditional class probabilities, although the performance will depend on having sufficient ground truth data.

Fourth, edges should be sampled instead of nodes when possible. A consistent finding of our simulations is that for the same labeling budget, a random edge sample outperforms a random node sample in terms of absolute error. In practical settings, such as Twitter, it is often much easier to randomly sample nodes than edges. One strategy for edge sampling is to use a result from the respondent driven sampling literature (152) that a random walk through an undirected network approximates an edge sample (see (20) for a discussion in the context of online networks). While this may not be feasible in some research settings, researchers may want to consider edge sampling if a random walk approach is possible.

Finally, researchers can obtain an estimate of network residual correlation by using cross-validation (see the discussion in (132) for a brief introduction to cross validation; see Chapter 7 of (89) for a more extensive discussion). Cross validation splits the training data into a number of folds (usually 5 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 approximating out-of-sample prediction. In the context of homophily estimation, estimating the residual term in Equation 4.4 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 performing cross validation on the ground truth dyads. If $\sum_{(i,j)\in S_{\text{train}}} \frac{1}{D_i} e_{ij} \neq 0$, where S_{train} is the training set, then models may need adjustment before providing reliable estimates of homophily. This strategy does not ensure unbiased homophily

estimates, particularly in the presence of non-random ground truth sampling, but it does provide a useful diagnostic.

4.7 Conclusion

We have examined the problem of estimating homophily when predictions must be used for node attributes. While the problem is challenging, the results we present indicate that homophily can be studied in online networks when classification performance is strong and network information is incorporated directly into models.

The strategies outlined here also provide a pathway for the measurement of other network-level properties. Examples are triadic properties, such as social closure by demographic group. In studies of dynamic network processes such as contagion, models to reduce measurement error (19) 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 networks.

4.8 Appendix

4.8.1 Five simulations

In addition to the simulation presented in the main text, we examine four additional simulations. These simulations further demonstrate the strengths and weaknesses of the approaches considered in the main text. We describe these simulations by the data generating process for Y_i^a.

1. **Independent**: $Y_i^a = 2X_i$, where $X_i \sim \mathcal{N}(0, 1)$. In the independent simulation, node categories depend only on a node-level feature X_i which is uncorrelated with the network.
2. **Degree**: $Y_i^a = 2X_i + Z_i$, where $X_i \sim \mathcal{N}(0, 1)$ and $Z_i = \text{mean}_{j \in \text{N}(i)}(X_j)$, Z_i standardized. In this simulation, node categories are a function of neighbor degree, meaning nodes with high-degree neighbors are more likely to have $Y_i^a = 1$.
3. **Main**: the simulation described in Section 4.5 of the main text.
4. **Unobserved**: $Y_i^a = 2X_i + Z_i$, where $X_i \sim \mathcal{N}(0, 1)$, $Z'_j \sim \mathcal{N}(0, 1)$ and $Z_i = \max_{j \in \text{N}(i)}(Z'_j)$, Z_i standardized. Z'_j is a standard normal unobserved variable which causes outcomes to be correlated in the network.
5. **Sampled**: This simulation samples nodes into the ground truth set according to degree and node features. Y_i^a is generated identically to the main simulation. For the edge simulation, dyads are sampled into

the ground truth set by,

$$p((i, j) \in \text{ ground truth}) = \text{logistic}(\alpha_{\text{dyad}} + 0.05D_i + 0.05D_j + 0.2X_i + 0.2X_j).$$

For the node simulation, nodes are sampled into the ground truth set by,

$$p(i \in \text{ ground truth}) = \text{logistic}(\alpha_{\text{node}} + 0.05D_i + 0.2X_i).$$

The α variables are constants chosen to make the total number of ground truth nodes or dyads equivalent to the sampling fractions for the other simulations.

Tables 4.1 and 4.2 present all simulations by all models, including a "no model" category using just the ground truth observations for comparison. The best performing model for each simulation (each row) is bolded.

4.8.2 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 (104) by gender online, and requires combining information about node gender and action.

Table 4.1: Bias by simulation, with the best model in each row bolded.

Sampling level	Simulation	No model	Node (no network)	Node	Edge	Ego-alter	Ego-alter (augmented)
Edge	Independent	0.003	0.001	**-0.0003**	0.001	-0.0005	-0.001
	Degree	0.005	0.089	0.058	**0.006**	0.019	0.019
	Unobserved	0.001	-0.053	-0.006	**-0.001**	0.018	0.018
	Main	0.002	-0.079	-0.031	**0.003**	0.018	0.014
	Sampled	-0.947	-0.064	-0.031	0.623	0.025	**0.021**
Node	Independent	0.004	-0.004	0.001	0.001	0.0003	**0.0002**
	Degree	0.006	0.114	0.033	0.002	**0.001**	**0.001**
	Unobserved	0.004	-0.145	-0.003	0.005	**-0.001**	**-0.001**
	Main	0.007	-0.176	-0.032	0.006	0.008	**0.003**
	Sampled	-0.057	-0.116	-0.032	0.030	0.015	**0.011**

In this case, Equation 4.2 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.

$$H^{aa}_{\text{extended}} = T[Y_i^a]^{-1} \sum_{(i,j)\in E} \frac{A_j}{D_i} Y_i^a Y_j^a. \tag{4.7}$$

In words, Equation 4.7 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 Section 4.3 to obtain an unbiased estimate: incorporate $\frac{A_j}{D_i}$ into the

Table 4.2: Mean absolute error by simulation, with the best model in each row bolded.

Sampling level	Simulation	No model	Node (no network)	Node	Edge	Ego-alter	Ego-alter (augmented)
Edge	Independent	0.054	0.030	0.024	0.045	**0.023**	**0.023**
	Degree	0.063	0.090	0.060	0.051	**0.030**	**0.030**
	Unobserved	0.051	0.055	**0.026**	0.042	0.029	0.029
	Main	0.051	0.079	0.035	0.038	0.027	**0.025**
	Sampled	0.947	0.081	0.058	1.585	0.051	**0.049**
Node	Independent	0.093	0.044	0.060	0.076	**0.059**	0.060
	Degree	0.107	0.116	**0.072**	0.088	**0.072**	**0.072**
	Unobserved	0.076	0.145	**0.049**	0.069	0.050	0.050
	Main	0.079	0.176	0.052	0.059	0.047	**0.046**
	Sampled	0.060	0.116	0.037	0.040	0.029	**0.028**

predictive model, instead of $\frac{1}{D_i}$ alone.

4.8.3 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 (45). 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 4.2,

$$H^{aa}_{\text{Coleman}} = T[Y^a_i]^{-1} \sum_{(i,j)\in E} Y^a_i Y^a_j. \tag{4.8}$$

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.

CHAPTER 5

DEMOGRAPHIC HOMOPHILY IN TWO LARGE AMERICAN CITIES

Social media platforms have reduced communication costs across social and geographic distance, yet can create echo chambers leading to increased insularity. To better understand social segregation in online spaces, we study the pattern of gender and racial social segregation in the ego networks of 470,000 Twitter users, bounded to the New York City and Dallas-Forth Worth metro areas. For both race and gender, we find that the tendency towards social segregation along demographic lines is substantially less online than in both strong-tie and acquaintanceship offline networks. The exception to this rule is White Twitter users, who do not appear to associate across demographic lines with increased frequency relative to offline. In order to study social segregation at scale, we introduce a novel machine learning approach which addresses bias found in previous studies. This modeling approach can be combined with either survey or rater data, facilitating more accurate studies of social segregation in online spaces.

5.1 Introduction

The rise of social media was predicted to reduce barriers to communication, with the expected effect that social media would facilitate discourse across traditional boundaries. Scholars have recently focused on the extent of cross-boundary communication, focusing on online ideological echo

chambers and demographic interaction.

Although a picture of cross-demographic interaction online has begun to emerge for weak ties (99; 129; 37), substantial uncertainty exists about whether online social networks facilitate cross-demographic interaction and information diffusion.

In this study, we examine the race and gender ego network composition of Twitter users in the New York and Dallas-Fort Worth metropolitain areas. Unlike Facebook, where network acquaintainces ("friends") are frequently real-life acquaintences, Twitter functions as both a social and information network where many ties are online-only. This dual informational and social use of Twitter means that studying Twitter allows examination of a range of tie types, from passive information consumption (follows) to mutual interaction (bidirected mentions). Follows in particular can be expected to contain a substantial fraction of alters that ego does not know personally.

This study therefore concerns ties that tend to be more online, more informational, and more instrumental instrumental than studies of Facebook. These properties of the Twitter network lend themselves to an online-offline comparison.

We primarily study ego network composition. The ego network represents an individual's experience in a given social space. Other measures of social segregration, such as Coleman's homophily measure, place emphasis on ties and weight high-degree nodes in proportion to their degree. Instead,

we focus on the experience from the perspective of individuals on Twitter: Who is authoring tweets seen? Who is ego conversing with? Whose information is ego retweeting to their own followers?

A growing literature addresses online social segregation along demographic and other lines, often using predictive methods (such as machine learning classifiers or dictionary lookups) for the characteristic of interest. Common examples are gender prediction based on first names (100), race prediction based on raters (37), and political leanings based on attention given to accounts (26).

Studies using prediction-based methodologies face an unrecognized challenge: network residual correlation. When classifiers make errors in ways correlated with the network, both over and under-estimates of social segregation are possible. We formalize the the network residual correlation problem and show that it can be phrased only in terms of dyads. We then propose a modeling and dyadic cross-validation approach which allows researchers to empirically determine the network residual correlation of a given model. When network residual correlation is low, estimates of social segregation are more credible. To our knowledge, we are the first to formalize this problem and propose a solution.

Using this methodology, we find that Twitter networks of all kinds tend to have lower levels of race and gender social segregation than found in self-report offline networks. This is true for both strong-tie and acquaintanceship networks. The findings in this paper also contribute directly to

an ongoing debate on the level of racial social segregation on Twitter.

5.2 Core questions and hypotheses

The primary goal of this study is to quantify the pattern of racial and gender homophily on Twitter, and contextualize these patterns by comparison with the General Social Survey. While previous work has addressed this question (129; 37; 99), in the Twitter context there is disagreement about the extent of homophily. As shown in (21), using a model to predict demographics in a network and then studying demographic homophily poses substantial challenges. This study is the first to examine homophily accounting for these issues, which is likely to provide more precise homophily estimates.

We develop the following hypotheses.

H1: Homophily is lower across follow, mention, and retweet ties on Twitter than it is among both offline discussion and acquaintanceship networks for both race and gender.

H2: Homophily online is higher for reciprocated ties than unidirectional ties.

5.3 Defining homophily

Homophily can be studied as an existing pattern of attribute similarity in networks which structures further social interaction (128; 24), or as the process which creates a particular structure, either due to another structure such as school (111) or as the result of individual choices (54). (128) define the presence of association at rates above random mixing as "inbreeding homophily", while distinguishing homophily "induced" by other structures from "choice" homophily which is a product of choices within a given structure. For instance, a school may induce homophily by grouping individuals with similar interests together in a class, yet within that class students make choices about whom to befriend. The overall pattern of association above baseline for some attribute, whatever the cause, may be described as inbreeding homophily and has its own implication for subsequent social interaction (18; 74).

In this study, we focus primarily on the pattern of homophily ("inbreeding homophily" in McPherson et al.'s terminology) and its implications. On a site like Twitter—where news and culture are topics of real-time discussion—the pattern of association along racial and gender lines has implications for information diets (11) and perception of outgroup members.

Additionally, we present evidence about choice homophily through the mechanism of tie reciprocation. The choice to reciprocate indicates which incoming links an ego prefers to engage with. While this does not pro-

vide direct evidence for discrimination, when reciprocation is higher for in-group members it suggests barriers to communication for out-group members.

5.3.1 Homophily measure

We study the composition of ego networks: are in-group members under-represented or over-represented (124)? This focuses the analysis on the experiences of individuals in the network, treating individuals with different degrees as equally important. The alternative is an edge-based measure of homophily such as Coleman's homophily index (45), which takes edges to be the primary unit of analysis and gives higher degree nodes more weight.

Formally, if Y_i^a is an indicator variable which is 1 when i is a member of group a and 0 otherwise, the ego network composition measure can be expressed,

$$H^{aa} = T[Y_i^a]^{-1} \sum_{(i,j)\in E} \frac{1}{D_i} Y_i^a Y_j^a. \tag{5.1}$$

Equation 5.1 expresses the average proportion of group a's ego networks which are comprised of in-group members. $T[Y_i^a]$ indicates the total number of people in group a, while D_i is the degree of node i and E is the set of edges in the network.

Variants of this measure can be estimated depending on the question,

for instance H^{ab} indicates the average fraction of group a's ego networks comprised of members of group b. Throughout the paper, this estimate is used in a way normalized for the size of the group in the population,

$$M^{aa} = \left\{ \begin{array}{ll} \frac{H^{aa}-p_a}{1-p_a} & \text{for } (H^{aa} - p_a) \geq 0 \\ \frac{H^{aa}-p_a}{p_a} & \text{for } (H^{aa} - p_a) < 0 \end{array} \right\}. \tag{5.2}$$

This measure ranges between -1 and 1, and takes random mixing as a baseline. If group a is 70% of the population, and 70% of group a's ego networks are comprised of members of a, then M^{aa} is 0.

5.4 Previous work

5.4.1 Discussion networks

We build most directly on research in sociology on core discussion networks (124; 127). These studies use the General Social Survey "name generator" question, which ego networks of discussion partners, up to size 5. (124) describes findings from the 1985 GSS and (127) describe the 2004 GSS. (162) describe the changes in demographic interaction among core discussion networks in the US between 1985 and 2004, finding that both baseline and observed racial heterophily increased, and that baseline gender homophily remained the same while observed gender heterophily increased.

Recently, sociologists have turned to studying intergroup contact in online networks. (99) use an innovative method to study the contact and discussion networks of Dutch Facebook users, finding that core networks are more homophilous than extended (Facebook friend) networks, except in the case of members of the ethnic majority, where extended and core networks are equally segregated.

This paper studies Twitter, which is a research site that functions as both discussion and information network. In an early crawl of the entire Twitter network, (113) find that macroscale properties of the Twitter interaction graph do not resemble most social networks, and argue that the service provides fast information diffusion more closely resembling an information or communication network. Yet conversations do happen on Twitter. (42) studies Black Twitter, focusing on the processes that constitute Black Twitter as a space. In this analysis, Clark finds both that conversations and information play important roles. (35) finds using a network analysis that Black Twitter users have denser reiciprocal networks, which she argues indicates that Black users may use Twitter more as a social network, while White users employ Twitter more as an informational network.

While sociological studies of ego network composition have often focused on the name generator question in the GSS, (159; 160) finds that important matters are frequently discussed with "unimportant alters", defined as alters that ego does not feel emotionally close to. Small focuses on two contexts in which egos activate alters: when alter is knowledgeable ("tar-

geted mobilization") or when alter is available at the right time ("opportune mobilization"). It is likely that, for some, Twitter functions as an expert mobilization site—particularly by other experts, in contexts such as #AcademicTwitter. Twitter is also used as a support-seeking service during moments of crisis,

5.4.2 Previous studies of online homophiliy

To our knowledge, two studies have directly addressed the question of ego network composition on Twitter. These studies have produced different conclusions about interaction rates, particularly for White Twitter users.

(129) use F++ facial recognition to study follow and interaction relationships among Twitter users in the United States. They find that the outbound follow networks of White Twitter users are 79% White, 11% Black, and 10% Asian; and that the outbound follow networks of Black Twitter are 55% White, 32% Black, and 13% Asian. For combined measures of outbound interaction (retweets plus mentions), they find that the network of White users is 74% White, 11% Black, and 15% Asian, while the network of Black users is 48% White, 38% Black, and 14% Asian.

On the other hand, (37) use a combination of face classification and human labeling to study same-race connectedness for inbound, outbound, and mutual follower networks among White and Black Twitter users in the United States. They find that the outbound follower network of White egos

is 44.6% White; the inbound follower network is 46.1% White; and the mutual follower network is 59.7% White. For Black egos, they find that the outbound follow network is 34.0% Black; the inbound follow network is 35.8% Black; and the mutual follow network is 49.8% Black.

The implications of these two studies are different. In an enviornment where rougly 60% of the people are White and 11% are Black (186) ((129) estimates 68% and 14%, respectively), the difference between White egos having 60% ingroup interaction rates and 80% ingroup interaction rates is the difference between homophily and heterophily. From an intergroup contact perspective, the majority group having about half of its contact with outgroups, as Cesare et al. estimate, would mean majority group members routinely interact with minority group members and are routinely exposed to concerns and interests of outgroup members.

5.4.3 Dataset description

We study Twitter, a popular microblogging service for the one year period between 2014/11/01 and 2015/10/30. Twitter is an open platform which defaults to public communication, which allows researchers to study rich behavioral traces in a large online network.

We limited our collection to the New York City and Dallas-Fort Worth metro areas (see Appendix for filtering details). This yielded 346,050 NYC users and 122,584 DFW users which had a profile picture of a human and

had active and non-protected accounts (which we used as criteria for distinguishing humans from other types of accounts). Data were collected via Twitter's REST API, and resulted in 156,719,072 tweets for NYC users and 56,684,443 tweets for DFW users.

We collected follower and followee relationships for these users, which effectively induces a subgraph within each metro area. This decision restricts results to links within-city, which plausibly increases the likelihood of offline social relationships in the data.

Tie type	Edge type	Edges
Follow	Bidir	4,341,561
	Directed	9,547,646
Retweet	Bidir	441,065
	Directed	1,246,104
Mention	Bidir	166,411
	Directed	1,250,580

Table 5.1: Number of edges per network type in the subgraph induced on nodes in New York City or Dallas-Forth Worth.

5.4.4 Classification methodology

Demographics on Twitter are not known and must be detected. Common methodologies are to use publicly available signals present on profiles, such as names (100) or profile pictures (129).

We adopt a strategy of combining human profile annotation with a rich set of signals. Using CrowdFlower (CF), we have 3 raters classify the race,

and collapse categories for the purposes of prediction into “Black”, “White”, and “other”. A total of 5,588 profiles are labeled with race.

For predictor variables, we collect Face++ (faceplusplus.com) labels for profile pictures, use word2vec (130) embeddings of tweet text, use DeepWalk (142) embeddings of the follow graph, and use additional profile signals generated from Twitter metadata. These include profile language, name-based ethnicity using Census first names, Tweets sent and follower/followee counts.

This rich set of signals provides strong classification performance as displayed in Table 5.2. In addition to standard classification metrics, we incorporate the inverse degree term $\frac{1}{D_i}$ as a covariate which addresses residual correlation with degree as discussed in (21). Network-residual correlation in predictions (which biases homophily estimates) is mitigated by both strong classification performance and by the explicit inclusion of network information in the from of node embeddings and degree information.

Category	Accuracy	AUC	F1
White	0.85	0.91	0.81
Black	0.92	0.96	0.95

Table 5.2: One-vs-rest classification performance using 5-fold cross-validation.

For gender, we combine first names with F++ profile picture predictions, building a ground truth set of nodes where both F++ and name agree. Then we predict using the same covariates as for race. Additional information about the dataset and classification process can be found in the Appendix.

5.5 Homophily on Twitter

Tie type	Demographic	Group fraction	Ego network frac. ingroup
Follow	Black	0.19	0.52
	White	0.64	0.77
Mention	Black	0.18	0.46
	White	0.66	0.78
Retweet	Black	0.19	0.53
	White	0.65	0.76
Follow	Woman	0.49	0.54
	Man	0.51	0.57
Mention	Woman	0.48	0.57
	Man	0.52	0.60
Retweet	Woman	0.49	0.61
	Man	0.51	0.58

Table 5.3: Group fractions and raw-ingroup association rate in bidirected networks. For instance, in the bidirected follower graph, Black Twitter users are about 19% of the overall network, and the average ego network in the follow graph for a Black user comprises about 52% Black users.

5.6 Comparison with offline homophily measures

The General Social Survey provides two modules which provide context for the Twitter results in this paper. The first is the well-studied GSS network module, which was administered in 1985 and 2004 (124; 127). The network module used the "name generator" question, asking about people the respondent discusses "important matters" with. The types of ties it elicits are themselves an area of study (9), but connections elicited this way tend to be stronger ties than are present on Twitter. For instance, 55% of the ties

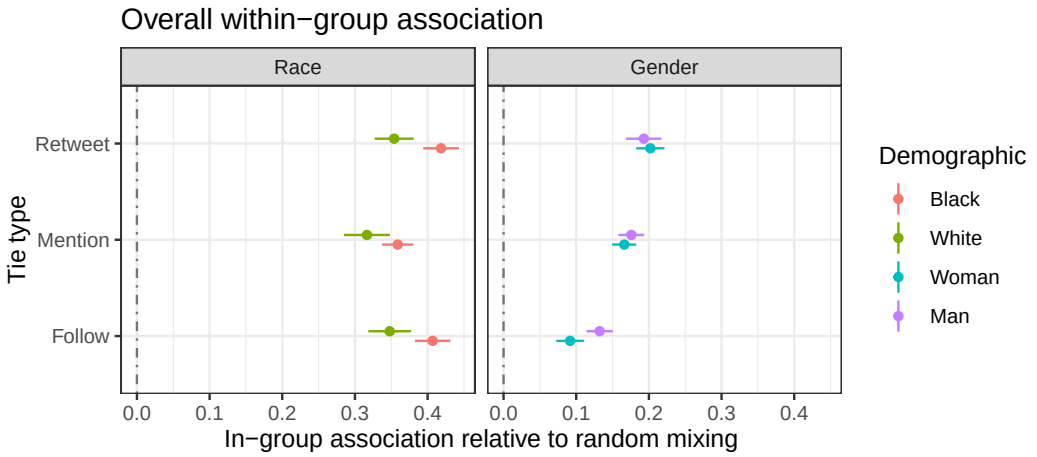

Figure 5.1: Overall association relative to random mixing for bidirected networks of each tie type. Zero indicates that the average ego network for a given group contains the number of same-group alters which would be expected by random chance. All four demographic groups studied exhibit homophily, with distinctly different patterns for gender and race. For race, Black Twitter users have higher homophily than White users, with mention ties having the lowest homophily. For gender, men and women have similar levels of homophily except along follow ties, where men are slightly more homophilous. For gender, follow networks have the lowest homophily, with mentions and retweets exhibiting more preferential association.

elicited are kin ties. The social network module contains the respondent's self-report of the race and gender of alters. For a comparison to Twitter networks, we exclude kin ties.

The second relevant GSS module is the 2006 social connectedness module, which has been employed in past research on homophily in online (37) and acquaintanceship (60) networks. This module asked a question about whether the people the responded is "acquainted" with are the same race

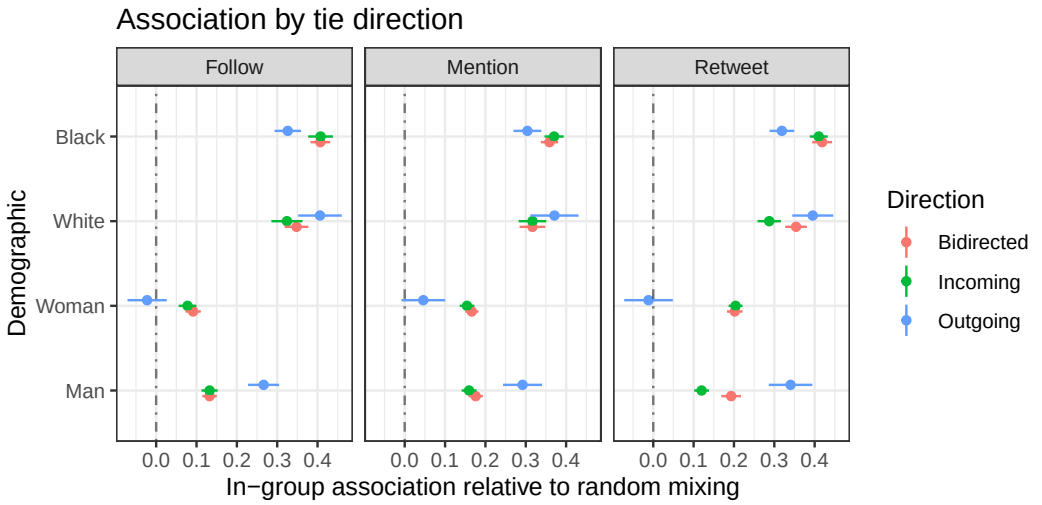

Figure 5.2: Association by tie direction. Outgoing ties—which individuals have control over—exhibit a distinct pattern for both race and gender. Black outgoing ties are less homophilous while White outgoing ties are more homophilous; likewise for gender, outgoing ties by women are less homophilous while those by men are more homophilous. This suggests that members of traditionally disadvantaged groups make more of an effort to link to outgroup members in online settings.

as respondent. The answers are an ordinal scale varying from "almost all a different race" to "almost all the same race". This question captures an individual's broad impression of the types of contacts they have.

Both of these instruments are imperfect comparisons for Twitter, but both provide important context. The social network module provides an idea of the homophily expected in an individual's close contact network. However, online interaction on a site like Twitter is generally thought to consist of weaker ties (although these weaker ties may still support dis-

cussion of important matters (159)). There is therefore a potential tie-type mismatch between the network module and Twitter.

The social connectedness module asks a question which is likely closer to the Twitter domain: it concerns acquaintances. However, the question relies on a general impression and does not ask respondents to provide estimates for the fraction of alters who are the same race. To address this, (37) impose a fraction on each ordinal element to generate "percent equivalents". In addition, it does not ask about groups besides the respondent's, which means that it's impossible to answer the question: "what fraction of a black respondent's acquaintance network is white?" Additionally, the same question is asked about only race, not gender.

Our view is that both of these questions provide useful comparison cases, but that we must be careful when drawing conclusions about the differential rates of online and offline homophily. In particular with the social connectedness module, the need to convert an ordinal scale to percentages creates uncertainty. Our replication of Cesare et al.'s work indicates that different choices of scale have meaningfully different homophily estimates. The scale Cesare et al. choose is 10/30/50/70/90 for a 1-5 Likert variable, which we refer to as the "10-90 scale." This scale implies that White Americans have almost no homophily above baseline in their acquaintanceship networks, which is a surprising finding that does not accord with other evidence on acquaintanceship networks (61).

The 10-90 scale scale assumes that everyone's acquaintanceship network

is at least 10% members of a different race. Considering that many localities in the United States (for instance, Vermont) are well over 90% White, this assumption may cause an under-estimate of homophily. We propose an alternative scale which ranges from 0/20/50/80/100 and evaluate using this scale.

5.6.1 Network module comparison

Non-kin results from the network module are presented in Table 5.6.1. From 1985 to 2004, ego network segregation among both White and Black Americans fell considerably, with the exception that Blacks were not more likely to associate with Whites (M of -0.83 in both cases). The tendency of Whites to associate with other Whites fell considerably, from M of 0.62 to 0.44.

For sex[1], we find that in-group preference is slightly stronger among men than women. From 1985 to 2004, same-gender preference among women rose substantially (0.36 to 0.52), and fell slightly among men (0.62 to 0.57).

[1]The GSS asks about sex with a binary male/female response option. In 2018, the GSS began asking about gender with the $SEXNOW$ question.

5.6.2 Social connectedness module comparison

Results from the 2006 social connectedness module are presented in Table 5.6.2. We use a percent equivalent strategy of: Almost all the same race as you (100%); Mostly the same race as you (80%); About evenly divided (50%); Mostly a different race than you (20%); Almost all a different race than you (0%). This scale accords more closely with the text of the questions. However, it still may underestimate homophily: the large overdispersion coefficients for race reported in (61) suggest that many acquaintanceship networks may lie in the 80% to 100% range. Even with this potential underestimate, we estimate $M = 0.53$ for Whites and $M = 0.66$ for Blacks[2].

5.7 Limitations

Estimating social segregation in any setting is a difficult task. The present study is complicated by the use of a classifier for race and gender, which introduces complications not present when individuals are asked to report their own demographics. Because of this concern, we have created several new approaches for the study of interactions between groups in online networks, and adapted a specification curve approach to a new domain.

However, these strategies do not completely remove uncertainty about our conclusions. It is possible that unmeasured variables are correlated both

[2] An evenly spaced scale of 100/75/50/25/0 gives $M = 0.32$ for White and $M = 0.63$ for Black.

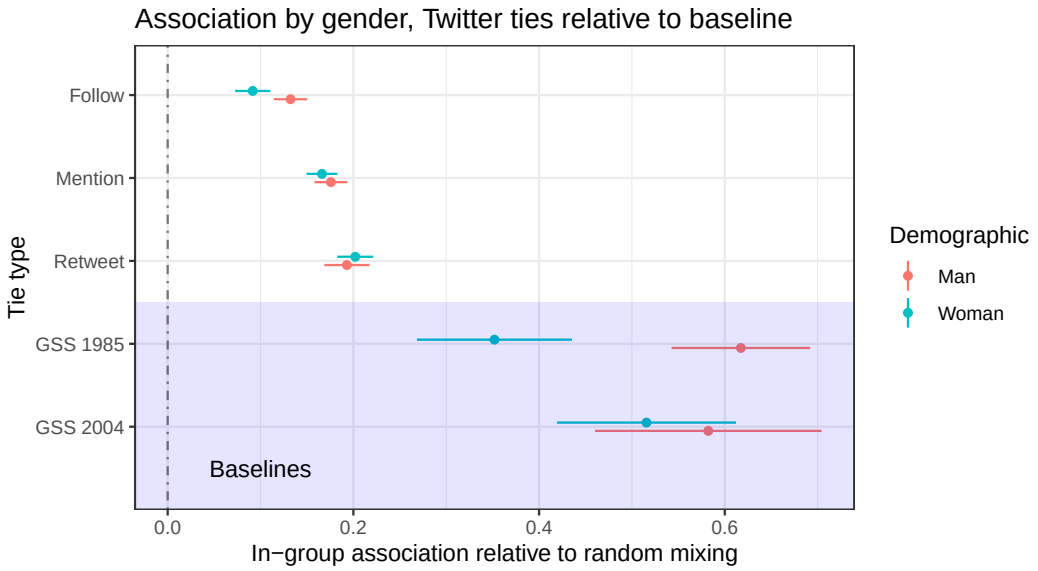

Figure 5.3: In-group association relative to random mixing for gender, compared to a baseline of GSS non-kin ties (shaded blue area). Homophily is substantially lower online. Ties on Twitter are bidirected, which provides a natural comparison for discussion network ties studied in the GSS.

with inclusion in the training set and with demographics, which has the potential to bias our models.

A second limitation is the comparison with offline surveys. While offline surveys have error just as classifiers do (individuals report their perception of their network, and may report incorrectly), it is also not clear which offline network is the correct comparison for Twitter. We have compared both a non-kin network of close confidants and an acquaintanceship network.

A third limitation is the time difference between the collection of the offline networks and the collection of the online networks. This is perhaps

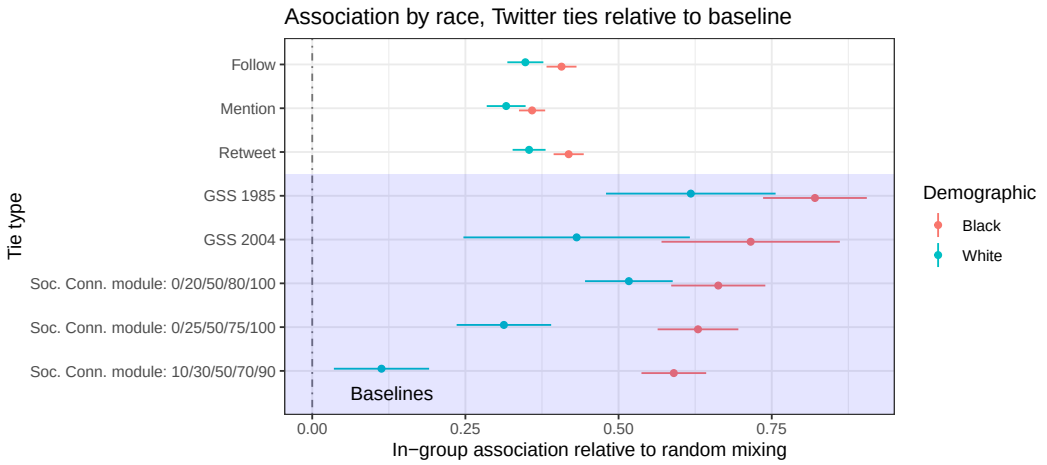

Figure 5.4: In-group association relative to random mixing for race, compared to a baseline of GSS non-kin ties and the GSS social connectedness module (shaded blue area). For Black Twitter users, homophily is lower online than offline. For White Twitter users, the picture is ambiguous depending on the choice of baseline.

the largest limitation of our study: if offline social segregation declined significantly from the survey period (2004 or 2006) to the online data collection period (2015), then it is possible that the online network we study is simply reflecting underlying social changes, rather than deviating from the offline world. This may be particularly plausible in the case of White people: offline social segregation among non-kin confidants fell from $M = 0.62$ to $M = 0.44$ between 1985 and 2004. This is about a point per year. If this trend continued linearly we would expect $M = 0.34$ in 2015, or right within the range of estimates from the online networks we study. In the case of Black Americans, the trend toward lower segregation is present but less pronounced, proceeding at about half the rate of that for Whites. Even

if social segregation were to decline at double the rate observed between 1985 and 2004, M would equal about 0.58 in 2015, far higher than the rates observed in the online networks we study.

5.8 Conclusion

The challenges of estimating demographic homophily in online social and communication networks are formidable. However, this information is necessary to assess the visibility of minority groups (104), the access of disadvantaged groups to information and majority networks, and to understand differences between online and offline social integration.

This paper has shown that from the perspective of cross-demographic interaction, sites such as Twitter can promote substantial rates of communication and information flow across demographic lines. Much like previous work on exposure to cross-cutting sources of news on Facebook (11), this paper suggests that the presence of demographic filter bubbles (138) is less pronounced than those occurring in offline social relationships.

Additionally, it has provided additional evidence on an empirical point where uncertainty currently exists. Estimates here suggest that homophily is present on Twitter, but that it is less pronounced than offline discussion networks and, at least for Black Twitter users, less pronounced than offline acquaintanceship networks. Substantial uncertainty exists about the rate

of offline discussion network and acquaintanceship network homophily for Whites in particular, suggesting additional research is needed on this topic.

This study has also uncovered an important type of asymmetry in homophily on Twitter: outgoing ties are less homophilous when originating from members of a disadvantaged group (Blacks, Women) than from members of an advantaged group (Whites, Men).

Variables	Year	Mean	SE	N
Ego demographic				
White	1985	86.8 %	0.86	1534
	2004	77.3	0.79	2812
	2006	70.1	0.68	4510
Black	1985	10.0	0.76	1534
	2004	11.5	0.60	2812
	2006	11.9	0.48	4510
Male	1985	47.0	1.28	1534
	2004	47.8	0.94	2812
	2006	47.1	0.74	4510
Female	1985	53.0	1.28	1534
	2004	52.2	0.94	2812
	2006	52.9	0.74	4510
Network module				
Number of alters	1985	3.06^{α}	0.04	1531
	2004	2.18	0.05	1426
Number of kin	1985	1.58	0.03	1531
	2004	1.18	0.03	1426
Number of non-kin	1985	1.48	0.04	1531
	2004	1.01	0.04	1426
Soc. conn. module				
Same-race acquaintances	2006	1.99^{β}	0.03	1322^{γ}

Table 5.4: Descriptive statistics of race and gender in the the American population from the 1985, 2004, and 2006 GSS. Main variables from the 1985 and 2004 network module and 2006 social connectedness module is also presented. $^{\alpha}$: Figures in this column for the network module section are the raw number of contacts in each category. A maximum of 6 was allowed; following (124), cases reporting 6 are recoded to 6.5. $^{\beta}$: Same-race acquaintances is measured on a 1-5 Likert scale; this is the mean of the scale, according to acquaintances being "mostly the same race as you". $^{\gamma}$: $N = 1322$ for people of all races and $N = 1206$ for people who identify as Black or White.

Year	Ego	Strategy	Mean	SE	M	M_se
2006	White	10/30/50/70/90	0.74	0.01	0.12	0.04
2006	Black		0.64	0.03	0.59	0.03
2006	White	0/25/50/75/100	0.80	0.01	0.32	0.04
2006	Black		0.67	0.03	0.63	0.04
2006	White	0/20/50/80/100	0.86	0.01	0.53	0.03
2006	Black		0.70	0.04	0.66	0.04

Table 5.5: Percent-equivalents, self-report same-race acquaintanceship networks from the 2006 GSS.

5.9 Appendix

5.9.1 Machine learning for demographic classification

Many studies have developed machine learning methods for race and gender classification in online settings without the explicit goal of quantifying cross-group interaction rates. Studies commonly use name, picture, profile, content (e.g. Tweets), and network information. (36) review this work, finding that accuracy of gender predictions have the highest mean accuracy, but are high variance; the average accuracy of race predictions are somewhat lower than gender but lower variance. The bulk of such studies work within a *maximum performance* framework, where the goal is to maximize an accuracy measure such as precision or recall.

When the goal is to estimate group properties, this is not always appropriate as studies of *quantification* recognize (69; 20; 21). We employ a quantification framework here, where the primary goal is to estimate a property of groups from individual-level predictions. It is the opposite of ecological inference (109), where the goal is to go from the group to the individual. Although performance measures are not dispositive for quantification tasks, we note that we achieve among the highest reported performance for classification of Twitter profiles.

5.10 Network residual correlation

A primary challenge to obtaining reliable estimates of intergroup contact in networks is the **network residual correlation** problem (21). When a method with error (such as a survey question or machine learning classifier) is used in a network, there is the possibility that errors will be correlated along the network.

For example, if a machine learning classifier predicts ego is "White" but ego is actually "Black", it is then more likely that the neighbors of ego both will actually be "Black" (due to network correlation) and that they will be classified as "White" (due to network residual correlation).

This formalization of this problem provided in (21) shows that the problem can be written entirely in terms of dyads.

5.11 Normalized egonet composition

An estimate of average egonet composition is more useful if normalized by expected egonet composition. This facilitates cross-network comparisons: without normalization by expected composition, it is difficult to draw conclusions about whether an observed ingroup association indicates high or low segregation.

Previous work comparing online and offline ego networks has not em-

ployed normalization by group size, making drawing conclusions about relative rates of segregation difficult. Few purely online studies report normalized figures, often working directly with group percentages in ego networks (e.g. (99)).

Inspired by Coleman's homophily measure (45), we propose a measure of normalized average egonet composition, which we refer to as M. This measure captures, from the perspective of the average person in a given group, the relative likelihood of associating with an individual in a target group.

Similarly to Coleman's homophily, M between groups a and b (where possibly $a = b$) is defined as

$$M^{aa} = \left\{ \begin{array}{ll} \frac{H^{aa}-p_a}{1-p_a} & \text{for } (H^{aa} - p_a) \geq 0 \\ \frac{H^{aa}-p_a}{p_a} & \text{for } (H^{aa} - p_a) < 0 \end{array} \right\}. \tag{5.3}$$

In contrast to Coleman's homophily, M^{aa} concerns average ego networks, not the fraction of links within and between groups. Coleman's homophily measure can be influenced by idiosyncratic patterns among high-degree nodes, and the Twitter networks we study exhibit several orders of magnitude of degree differences. M^{aa} focuses on the average over all people in a group, which is more appropriate for studying how participants in online communities perceive spaces than a link-based measure.

We use M^{aa} as a standardized measure of segregation. We present both

nominal (H^{aa}) and standardized (M^{aa}) measures. Nominal measures provide important information, but are difficult to interpret alone.

5.12 Uncertainty estimates

We use the bootstrap (63) to provide confidence intervals for estimates. For each bootstrap replicate, we begin by resampling the nodes in the network. Nodes may appear more than once. Then, we construct a replicate edgelist taking into account the multiplicity in the node replicate. For instance, if node a occurs twice in the node replicate, its edges to other nodes in the node replicate are doubled in the edge replicate.

5.13 Classifying race and gender

Race and gender classification are common tasks in machine learning and social science (36). For race, we build our ground truth set from a human-labels on the CrowdFlower labeling site. For gender, we use the intersection of two well-studied signals: F++ predictions based on profile photos and first name recognition based on Census data.

For prediction models, we use logistic regression which provides well-calibrated probabilities (135). We employ covariates for prediction as described in the next section.

5.13.1 Covariates

Models predicting race and gender use the following covariates:

1. Tweet text: 32-dimensional word2vec embeddings of timelines, with word embeddings averaged within nodes to create a "timeline embedding". This is done with Gensim (147), using skipgrams, negative sampling, 25 epochs, and a window size of 5.
2. Network data: for each tie type, we use inverse in-degree and inverse out-degree. We also embed the follow graph using 32-dimensional embeddings based on random walks (similar to the DeepWalk method (142)). We use the same settings as for word vectors. Notably, this means using negative sampling rather than DeepWalk's hierarchical softmax for negative examples.
3. Names: We use the race and gender breakdown of first names.
4. Language: We use language of the account detected by Twitter.
5. F++: We use classifications of race, gender, and age from F++.
6. Twitter activity information: the following covariates are provided by Twitter at the user level: days since joining, number of Tweets, number of followers, number of times listed, number of people user is following, number of favorites (likes), and days since active.

5.13.2 Differences with past GSS estimates

Small differences in estimates compared to prior work is due to two factors: different weighting schemes and a potential error by (127). This paper uses the GSS variables $ADULTS$ to weight the 1985 sample, and the combined $ADULTS * WTSSNR$ to weight the 2004 and 2006 samples.

Differences between Table 5.6.1 and (124) can be explained by Marsden using an unweighted sample. Here, the variable $ADULTS$ is used to weight the sample, giving these estimates an individual-level rather than household-level interpretation.

Differences between Table 5.6.1 and (127) can be explained by two factors. First, McPherson et al. use the inverse of the weights, e.g. $1/ADULTS$ rather than $ADULTS$. Using $1/ADULTS$ replicates the 1985 estimates presented in McPherson et al.

A second discrepancy of unknown origin leads to McPherson et al. claiming $N = 1467$ rather than $N = 1426$. This discrepancy in observations cannot be explained by the inclusion of category 9 for the $NUMGIVEN$ variable ("No answer"), which would change the number of observations to 1472.

Differences between Table 5.6.1 and (37) can be explained by Cesare et al's use of an unweighted sample. Cesare et al.'s $N = 1206$ is only accounting for egos who identify as White or Black.

CHAPTER 6

EXAMINING THE IMPLICATIONS OF HOMOPHILY ON SOCIAL DILEMMAS

Probabilistic social structural models in the tradition of Peter Blau provide useful theoretical description of group arrangements. However, behavioral dynamics with specific structural configurations are often discussed more informally. To address this, I propose an analytic evolutionary game theory model which examines pairwise game outcomes within social structures defined by intergroup contact rates. This model allows examining game dynamics given a level of homophily. This is a minimal model that allows exploration of the entire homophily-by-pairwise-game space in the case of two groups. The model yields several interesting results. First, the equilibrium outcome of the prisoner's dilemma is not changed by social structure. However, when group-based payoffs exist—as in a modified prisoner's dilemma with in-group preferences—some structural configurations can facilitate cooperation. In the case of coordination games such as the stag hunt, homophily can facilitate cooperation even when games do not have payoffs which vary with group status. The reason for this is that once a single group has converged on a high-payoff outcome, the strategy can then spread to the other group. Taken together, these results indicate that for pairwise games, homophily can both facilitate and inhibit cooperation.

6.1 Introduction

Sociologists have long argued for the causal importance of social structure on macrolevel outcomes (23; 30; 31). To examine the effects of social structure on resolving social dilemmas, I propose a model that I call the Social Evolutionary Game (SEG), which integrates a probabilistic model of macrosocial structure (145; 66; 156) with the theory of asymmetric evolutionary games (153; 76). In contrast with previous formally similar models (94; 93), SEGs are parameterized in a way that allows direct examination of the effects of important structural parameters, such as group size and homophily.

In Section 6.2, I lay out the two theoretical traditions I draw on: Blau's theory of social structure and asymmetric evolutionary games. In Section 6.3, I state the replicator dynamics of the SEG model formally and detail the model's specific assumptions and interpretation for sociologists. In section 6.4, I show that social structure is irrelevant for facilitating cooperation in the prisoner's dilemma. In section 6.5, I analyze two important classes of situations for which social structure can play a pivotal role in engendering cooperation: when there are *social roles* and *group-based payoff preferences*.

By allowing formal analysis of the effects of social structure on macrolevel outcomes, the SEG model provides scope conditions for the situations where structure matters. In Section 6.6, I provide simple heuristics for such cases. This paper uses a stylized model to illustrate how social

structure can both facilitate and inhibit cooperation.

6.2 Theory

6.2.1 Social Structure

Sociologists have long been concerned with the importance of social structure. Structure is commonly conceptualized as a network of contacts, who provide information (84), social support (46), brokerage opportunities (30), and the ability to transmit behaviors (72). In addition to specific network positions, sociologists have developed a variety of methods for inferring hierarchies (70; 49; 123; 22) and characterizing types of positions based on network structure (184; 29).

In contrast to inductively determining positions and roles from network structure, class analysts and demographers have been concerned with interaction rates among socially salient groups, such as marriages across income or educational levels (103; 154; 155). Such analyses are deductive: they begin with knowledge of the socially salient groups (e.g. income category), and seek to understand the important interaction patterns among individuals with such socially salient characteristics. Such analyses can be said to be *probabilistic*, in the sense that an important part of the analysis resides in determining the level of homophily (45; 126) for each group.

Blau's theory of social structure (23) falls into the second category: we are interested in deriving properties of the social world from interaction rates among groups. For this type of macrostructural theory, specific network structures and the position of any given individual are less important than the interaction rates among socially salient groups. The inductive and deductive conceptualizations of social structure are not incompatible: a given network may have a stochastic blockmodeling procedure applied to inductively find similar individuals, while also containing information on interactions between various social groups.

However, the two procedures emphasize different levels: an inductive approach seeks to characterize microlevel specifics, while a deductive approach seeks to summarize larger patterns of interaction based on our prior knowledge about important groups in the population.

In this article, I take a deductive, probabilistic approach to social structure. I assume that socially salient characteristics are known in advance, and that interactions between individuals may be affected by such characteristics. This choice is motivated by a desire to model macrolevel outcomes, which can be difficult when considering the vast array of network topologies that may be consistent with any model of probabilistic interaction.

Work in this deductive, probabilistic often leaves a model of action implicit (23; 145; 24; 156; 66; 67; 157). The main action studied has been whether or not to associate with another individual based on characteristics. Considerations of behavioral outcomes, such as cooperation and co-

ordination, in such structures have only recently become objects of study (93; 94; 95). I argue that the usefulness of such probabilistic models is limited without a behavioral component.

Blau's (23) analysis stops after determining that group sizes and correlation between individual characteristics matter for rates of intergroup contact. But why is intergroup contact a worthy object of study? I argue that we should care about intergroup interaction rates because different interaction patterns have different implications for macrolevel outcomes (e.g. cooperation). In this sense, the importance of structural theories is greatly limited when not accompanied by a theory of action.

Formalization

I will use a simple mathematical object to represent social structure: the interaction matrix. If there are b social groups, the interaction matrix B is the $b \times b$ matrix where entry b^{tu} represents group t's interaction with group u. B must be stochastic, such that $\sum_u b^{tu} = 1$. However, it does not have to be symmetric, so b^{tu} does not necessarily equal b^{ut}.

Interaction probabilities can be decomposed further to explicitly incorporate sociological parameters such as group size, mean degree, and homophily. For group t, call group fraction N^t and fraction of edges to group u e^{tu}. Then any cell of B is just $b^{tu} = e^{tu}$. This is the fraction of t's edges originating in t that end in u. Homophily can be computed straightforwardly for

groups:

$$h^t = \frac{e^{tt} - N^t}{1 - N^t}$$

B models interaction between groups. In cases where groups along more than one dimension (e.g. race and gender) are analyzed, analysis can proceed by creating a new set of groups of the form (race, gender) and a new, larger interaction matrix that represents one race-gender group.

6.2.2 Evolutionary Games

The game theoretical models most familiar to sociologists are in the form of the prisoner's dilemma (91; 6; 120; 119). The PD models a situation where social benefit (the sum of all player payoffs) is maximized by a strategy that nobody has an individual incentive to play. Evolutionary games (76) expand upon this type of one-shot game by modeling how a population players change strategies over time to reach an equilibrium state.

For instance, in an evolutionary one-shot PD, nearly everyone may start by cooperating, but over time players receive more benefit from defection, and so switch strategies at some rate towards the asymptotically stable state of everyone defecting.

Motivation

Evolutionary games are traditionally motivated by modeling a process of genetic selection (161; 137). An animal has a strategy (e.g. aggressive or passive) hard-wired. In a strategic environment, this strategy confers a certain *fitness*, or number of offspring. More successful strategies are those which *replicate* themselves at greater rates by having higher fitness. Replicator dynamic (51) equations are employed to analyze how the population of players changes strategies over time, with *asymptotic stability* as the key solution concept.

An asymptotically stable equilibrium is a refinement of a Nash equilibrium, and is any equilibrium where all trajectories of the replicator dynamics in some area around the equilibrium converge to the equilibrium as time goes to infinity. This means that an asymptotically stable point is robust to invasions by small numbers of mutants, much as an evolutionary stable point is. However, evolutionary stability is often defined as a point that is robust to invasion by one alternative strategy, whereas an asymptotically stable point is robust to invasions by mixed strategies (76). If an ESS is defined as robustness against mixed strategy mutants, then ESS and asymptotic stability coincide.

Evolutionary games have been applied to explaining human culture and cooperation across generations (7; 27). In these models, the strategy played by a human is assumed to be fixed within individuals, and certain behav-

iors (e.g. cooperation) give higher or lower fitness, and then replicate accordingly across generations.

This multi-generational interpretation of the replicator dynamic is common, but it is usually not appropriate for sociology. We study changes that take place on shorter timescales. For instance, we would not try to explain the collapse of a state in terms of a multi-generational process of differential reproduction of certain genes. However, the replicator dynamic can also be interpreted as a *fixed population* of individuals who change strategies based on *information*.

Consider the the following scenario (76): in the morning, individuals are paired with another member of the population at random and play a pairwise game (the *stage game*). In the afternoon, individuals are paired with another random individual, and exchange information about the strategies they played and the payoffs they received. In the evening, players reflect by comparing the payoff they received in the morning versus the payoff they learned about in the afternoon, and decide to try out a new strategy the following morning with some probability if they can receive a higher payoff by switching.

This type of interpretation is consistent with the replicator dynamics, and with the sociological goal of analyzing behavioral changes within a fixed population. Here, strategies replicate based on the *utility* they confer to individuals. In this scenario, the replicator dynamics are a result of individuals learning new information about payoffs they can receive for

playing different strategies in a given environment.

Asymmetric Evolutionary Games

In a standard evolutionary game, there is one population, and pairs of players are paired to play the stage game. In an asymmetric game, there are two populations that are paired both within and between groups. Populations may receive different payoffs when paired with the ingroup and outgroup (5; 51; 93; 94).

As an example, consider the case of heterosexual marriage. We could model this social situation as an asymmetric evolutionary game. A simple asymmetric model would have women and men playing only between-groups, so that all pairings were between a woman and a man. In a richer model, we can allow women and men to play both within- and between-group games. For instance, women play other women for the opportunity to inhabit the "woman" space, and play between groups to select partners.

Asymmetric evolutionary games either assume random mixing or fixed group sizes (93; 94). However, in the social world, different groups have different interests and interact at different rates. I extend the previous work on asymmetric evolutionary games by incorporating arbitrary probabilistic social interactions into such models.

Such a model yields a phase space that may be compared with empirical data or tested in the laboratory.

6.3 Model

In this section, I present the assumptions and formal statement of the SEG model. I present a two-group, two-strategy version of the model for simplicity. Such a model may be easily solved numerically, along with facilitating analytical treatment similar to (93).

6.3.1 Assumptions

Here are four important assumptions that are encoded in the SEG model.

Assumption 2. ***Fixed individual characteristics****: Individuals cannot change their group.*

Assumption 3. ***Fixed group sizes****: While a simplification of reality, this assumption means that interaction and information drive the evolution of behavior.*

Assumption 4. ***Behavior separate from characteristics****: Within-individuals we assume that characteristics are fixed, but that individuals still have choices about behavior.*

Assumption 5. ***Heterogeneous reference groups****: Individuals look to specific others based on characteristics. I assume that individuals compare their payoffs with ingroup members.*

These assumptions straightforwardly outline the scope of the model. We do not allow certain elements, such as individual characteristics and the

interaction pattern, to change. The result of the model then yields a set of ESS which can be analyzed. One parameter can then be varied (e.g. group interaction rates) to examine how the ESS set changes.

I leave endogenizing parts fixed parts of the model for future research. For instance, allowing group interaction rates to change endogenously could produce insights into how payoffs affect group interaction rates.

6.3.2 Formal Statement

I assume two groups and two strategies. I'll call the strategies c and d, which could represent actions such as cooperate/defect, hunt stag/hunt rabbit, aggressive/passive, and so forth. I'll call the groups t and u, which could represent sex, ethnic groups, nationalities, fans of two different baseball teams, and so forth. The *stage game* is assumed to be two-player, so that pairs of players are selected, play the game, and receive a payoff at each time period.

Assume we have a payoff matrix for the stage game for each group against each other group. Since we have two groups, this yields four payoff matrices[1]. I will denote the payoff matrices for t by A^{tt} and A^{tu} and the payoff matrices for u by A^{uu} and A^{uu}. I will use the convention that superscripts refer to groups and subscripts refer to strategies. This gives,

[1] t against t, t against u, u against u, and u against t.

$$A^{tt} = \begin{bmatrix} a^{tt}_{cc} & a^{tt}_{cd} \\ a^{tt}_{dc} & a^{tt}_{dd} \end{bmatrix}$$

$$A^{tu} = \begin{bmatrix} a^{tu}_{cc} & a^{tu}_{cd} \\ a^{tu}_{dc} & a^{tu}_{dd} \end{bmatrix}$$

$$A^{uu} = \begin{bmatrix} a^{uu}_{cc} & a^{uu}_{cd} \\ a^{uu}_{dc} & a^{uu}_{dd} \end{bmatrix}$$

$$A^{ut} = \begin{bmatrix} a^{ut}_{cc} & a^{ut}_{cd} \\ a^{ut}_{dc} & a^{ut}_{dd} \end{bmatrix}$$

If a member of t plays c against a player of u group playing c, they receive a^{tu}_{cc}, while a member of u playing c against a member of t playing c receives a^{ut}_{cc}. These payoff matrices represent the payoffs to individuals from a one-shot game.

To incorporate temporal dynamics, we focus on two quantities, the proportion of t playing c and the proportion of u playing c, which I'll denote p^t and p^u, respectively, while p is the overall state of the game given by p^t and p^u. Note that the proportions of each group can be derived from these quantities straightforwardly: the proportion of t playing d is $(1 - p^t)$. This means our dynamics need only two equations, modeling the change in the proportion of cooperators in t and u over time. Denote these quantities $\dot{p}^t$ and $\dot{p}^u$, respectively.

Recall from Section 6.2.1 that the interaction matrix is given by B. I use π to denote payoffs. $\pi_c^t(B, p)$ is the payoff to playing c for group t with interaction matrix B, when the game is at state p. $\bar{\pi}^t(B, p)$ is the average payoff for group t with interaction matrix B at game state p. We then have a system of two differential equations:

$$\dot{p}^t = p^t[\pi_c^t(B, p) - \bar{\pi}^t(B, p)] \tag{6.1}$$

$$\dot{p}^u = p^u[\pi_c^u(B, p) - \bar{\pi}^u(B, p)] \tag{6.2}$$

Equivalently,

$$\dot{p}^t = p^t(1 - p^t)[\pi_c^t(B, p) - \pi_d^t(B, p)] \tag{6.3}$$

$$\dot{p}^u = p^u(1 - p^u)[\pi_c^u(B, p) - \pi_d^u(B, p)] \tag{6.4}$$

Take Equation 6.1, which is the rate of change for cooperators in group t at game state p. This equation says that the proportion of cooperators increases in group t when cooperators do better than the average member of t, and decreases when cooperators do worse than the average member of t. This is not a forward-looking *best response* dynamic, but encodes the assumption that individuals change to better strategies with some positive probability. Adjusting the rates of switching does not change the trajectories of the replicator dynamic, but does change the speed with which convergence occurs (76).

Consider the payoff term $\bar{\pi}^t(B, p)$. To understand the dynamics of the game, it's helpful to write down the components of this term explicitly:

$$\pi_c^t(B, p) = b^{tt}(p^t a_{cc}^{tt} + (1 - p^t)a_{cd}^{tt}) + b^{tu}(p^u a_{cc}^{tu} + (1 - p^u)a_{cd}^{tu}) \tag{6.5}$$

$$\pi_d^t(B, p) = b^{tt}(p^t a_{dc}^{tt} + (1 - p^t)a_{dd}^{tt}) + b^{tu}(p^u a_{dc}^{tu} + (1 - p^u)a_{dd}^{tu}) \tag{6.6}$$

This form can be simplified further, but this expression clearly demonstrates how social structure enters into payoffs: b^{tt} multiplies the in-group payoff and b^{tu} multiplies the out-group payoff. In a case where playing c yields a high in-group payoff and low out-group payoff, a high in-group interaction rate can make c beneficial.

6.4 Social Structure as Correlation

When analyzing the evolutionary prisoner's dilemma, scholars have argued that *correlation* (also "clustering") provides a potential explanation for the presence of cooperation (6; 7; 158). Correlation occurs when individuals who play similar strategies are paired at higher-than-average rates. For instance, a group of cooperative animals may by chance inhabit the same environment. This pairing of cooperators then leads to cooperative animals having high payoffs and successfully reproducing.

This argument relies on a reliable mechanism to pair individuals based

on *strategies*. If organisms have hard-to-mimic signs of cooperativity and strategies are fixed within organisms, correlation provides a plausible mechanism for cooperation to evolve. Alternatively, if an omniscient and costless mechanism[2] pairs individuals based on strategies, then defectors can be consistently punished by interacting with other defectors.

In the social world, two problems with this argument arise. First, strategies are often different from characteristics. Second, individuals are paired on characteristics and not strategies, since characteristics are publicly observable, while strategies are only revealed at the time of action.

Together, these two features make it unlikely that correlation provides a reliable mechanism for the evolution of cooperation without other factors, such as difficult-to-change strategies, reputation, or punishment (136). However, this discussion does reveal an often overlooked theoretical point about social structure: it can be seen as a source of correlation in the social world, although the interpretation of this correlation for social behavior is less straightforward than for biological behavior since strategies are divorced from characteristics.

[2]Perhaps individuals would be willing to pay some cost, but assume that individuals do not have the ability to agree on such a mechanism in advance.

6.4.1 Irrelevant Structure in the Prisoner's Dilemma

If we accept that pairing individuals based on the strategies they *will* play is infeasible, then we rely on pairing individuals based on characteristics. For instance, assume brown-eyed people are be paired to play together, either through choice homophily or some outside structure that encourages this interaction, leading to inbreeding homophily $h_{\text{blue eyes}} > N_{\text{blue eyes}}$ (128).

Consider the prisoner's dilemma game given by the following payoff matrix:

$$A = \begin{bmatrix} 3 & 0 \\ 5 & 1 \end{bmatrix}$$

Then we have the following result: if $A = A^{tt} = A^{tu} = A^{uu} = A^{ut}$, then for any values of B, defection is still the only asymptotically stable strategy. Further, the basin of attraction for defection remains the entire range of p values. In other words, if all groups play a prisoner's dilemma against all other groups, no social structure B changes the substantive outcome of the game.

The reason for this is easy to see. Note that in Equation 6.3, group t's fraction of c players increases precisely when[3] $\pi_c^t > \pi_d^t$, or when strategy c has a higher payoff than strategy d. For social structure B to facilitate cooperation, there must be some B such that $\pi_c^t > \pi_d^t$, for some value of p. In

[3]For this discussion I simplify $\pi_c^t(B, p)$ to π_c^t.

other words, if there is any hope of B facilitating cooperation, at some p, c must do better than d. If at all points d does better than c, then the replicator dynamics will never lead to a larger proportion of c.

Using the "column operation", we can get zeroes on the off-diagonals of A by subtracting a constant from a column. This does not change the replicator dynamics. We use this to get A',

$$A' = \begin{bmatrix} -2 & 0 \\ 0 & 1 \end{bmatrix}$$

We then write the payoffs for group t playing strategy c and d by plugging in for Equations 6.5 and 6.6 (the zeroes in A' make these expressions simpler),

$$\pi_c^t = -2(b^{tt}p^t + b^{tu}p^u) \tag{6.7}$$

$$\pi_d^t = (1 - p^t)b^{tt} + (1 - p^u)b^{tu} \tag{6.8}$$

Since p^t, p^u, b^{tt}, b^{tu} are always non-negative, $\pi_c^t \leq 0$ and $\pi_d^t \geq 0$. These are both zero in the case where $b^{tt} = 0 = p^u$ or $b^{tu} = p^t = 0$. As long as we have a *mixed trajectory*, where all strategies in all groups appear with positive probability, $\pi_d^t > \pi_c^t$ since $(1 - p^t)b^{tt} + (1 - p^u)b^{tu} > 0$ and $-2(b^{tt}p^t + b^{tu}p^u) < 0$.

This proves that in the case of the prisoner's dilemma given by A for any mixed trajectory, the replicator equations 6.1 and 6.2 are always negative. In other words, cooperation is always declining, for any values of B and p.

In an arbitrary version of the Prisoner's dilemma given by

$$A = \begin{bmatrix} R & S \\ T & P \end{bmatrix}$$

where $T > R > P > S$, we can rewrite Equations 6.7 and 6.8 as:

$$\pi_c^t = (R - T)(b^{tt}p^t + b^{tu}p^u) \tag{6.9}$$

$$\pi_d^t = (P - S)((1 - p^t)b^{tt} + (1 - p^u)b^{tu}) \tag{6.10}$$

Since we can always add a constant to a column of A without changing the replicator dynamics, without loss of generality we can assume that $T, P, R, S > 0$. Note that $R - T < 0$ and $P - S > 0$. Using the fact that p^t, p^u, b^{tt}, b^{tu} are non-negative, again we have the result that for all mixed trajectories, $\pi_d^t > \pi_c^t$.

In this case, social structure given by B can be said to be *irrelevant*. This should not be surprising, given that all individuals in this case have a strictly dominant choice in the stage game: defection. No matter what the other groups are doing, individuals do best by defecting.

6.4.2 Numerical Evidence

We can see the above result from solving prisoner's dilemma models numerically and plotting the *phase space*. This shows the trajectories of the replicator dynamics equations 6.1 and 6.1. I assume groups are of equal size, and choose the following interaction matrices:

$$B_{\text{no}} = \begin{bmatrix} .5 & .5 \\ .5 & .5 \end{bmatrix}$$

$$B_{\text{high}} = \begin{bmatrix} .9 & .1 \\ .1 & .9 \end{bmatrix}$$

In B_{no}, there is no homophily. In B_{high}, homophily is 0.8 for both groups. We can see the results in Figures 6.4.2 and 6.4.2. Trajectories are nearly unchanged, and the bottom left corner (full defection for both groups) is the only basin of attraction.

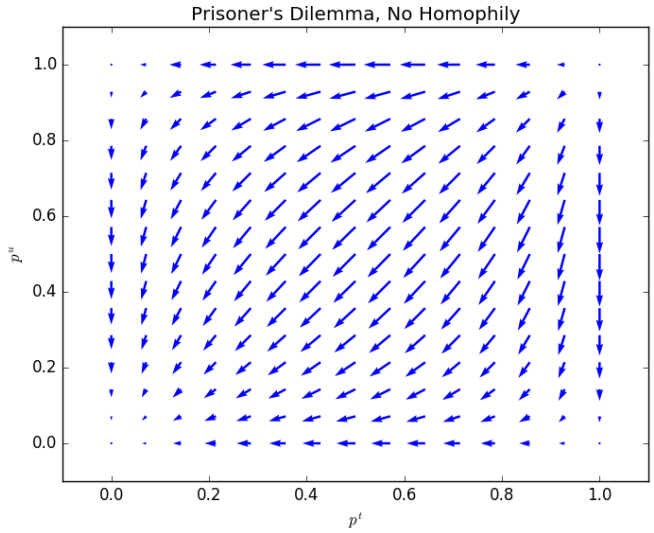

Figure 6.1: The Prisoner's Dilemma with two groups and no homophily. There is a single equilibrium at defect-defect (bottom left).

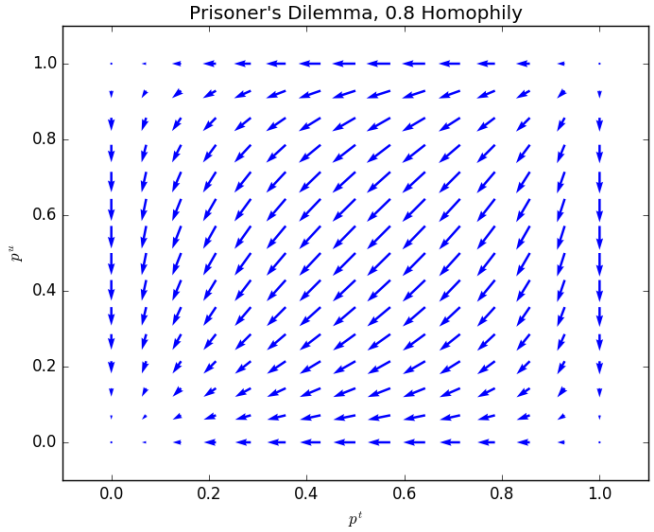

Figure 6.2: The Prisoner's Dilemma with two groups and a homophily value of 0.8. This leaves the single equilibrium (defect-defect) unchanged (bottom left).

6.5 When Social Structure Matters

The previous section can be interpreted pessimistically: no social structure B will change the outcome of a pure prisoner's dilemma. However, when *social roles*, *social norms*, or *group-based preferences* exist, social structure can play a pivotal role in facilitating cooperation (64; 43) . I present two such examples in this section using numerical methods.

6.5.1 Roles and Norms

Social roles and social norms are encoded the payoff matrices $A^{tt}, A^{tu}, A^{uu}, A^{ut}$. For instance, in a simple game where there is one leader and one follower, the leader could derive higher payoffs from an aggressive strategy, and the follower could derive higher payoffs from a passive strategy. In this case, roles and norms would be internalized, since they are reflected in individual payoffs and not enforced by an outside sanction.

Consider a simple situation where individuals play the stag hunt (assurance game) against ingroup members, and a prisoner's dilemma against outgroup members, must chose to play one strategy regardless of the identity of their opponents. This could represent a situation where a social norm of cooperation obtains within-group, but there is no norm prohibiting defection in between-group play. Payoff matrices are given by:

$$A_{\text{ingroup}} = \begin{bmatrix} 3 & 0 \\ 1 & 1 \end{bmatrix}$$

$$A_{\text{outgroup}} = \begin{bmatrix} 3 & 0 \\ 5 & 1 \end{bmatrix}$$

In this situation, social structure has a pivotal effect. Increasing levels of homophily make cooperation asymptotically stable. In this sense, social structure has the ability to change an in-group norm into a norm for the entire population, even though between-group interactions are governed by a mixed-motive game. In Figures 6.5.1, 6.5.1, and 6.5.1, we plot the phase space of the system with homophily values of 0, .2, .4, .8. Note that at homophily of .4, cooperation has an asymptotically stable basin of attraction of reasonable size. At homophily of .8, cooperation has the largest basin of attraction.

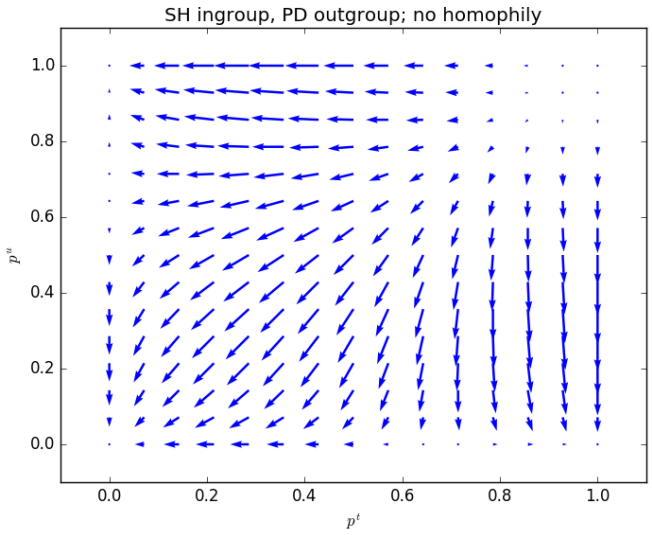

Figure 6.3: A game where individuals face a coordination game (Stag Hunt) with ingroup members and a Prisonner's Dilemma with outgroup members, and must decide on a strategy to play. The defect-defect outcome (bottom left) is the most probable outcome but other outcomes are possible depending on the starting state. The defect-cooperate (top left; bottom right) outcomes are possible, where one group is filled with cooperators and the other defectors. If the game starts near cooperate-cooperate (top right), it will stay there, although relatively small perturbations will push the game to a different equilibrium.

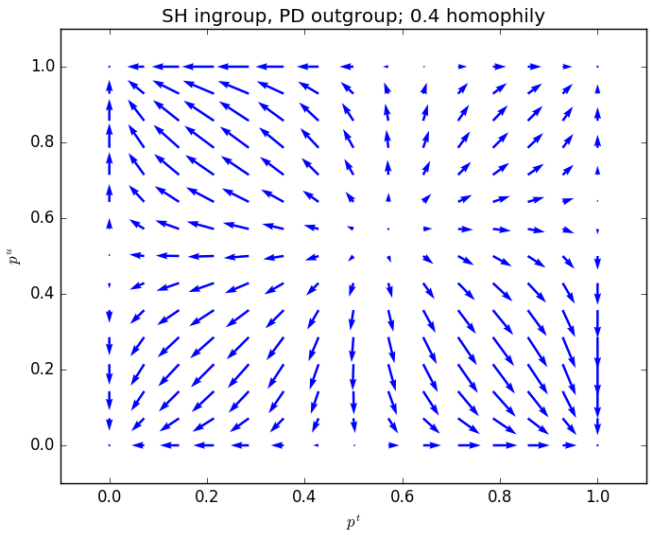

Figure 6.4: A game where individuals face a coordination game (Stag Hunt) with ingroup members and a Prisonner's Dilemma with outgroup members, and must decide on a strategy to play. As homophily is increased, the non defect-defect outcomes (bottom left) become more likely.

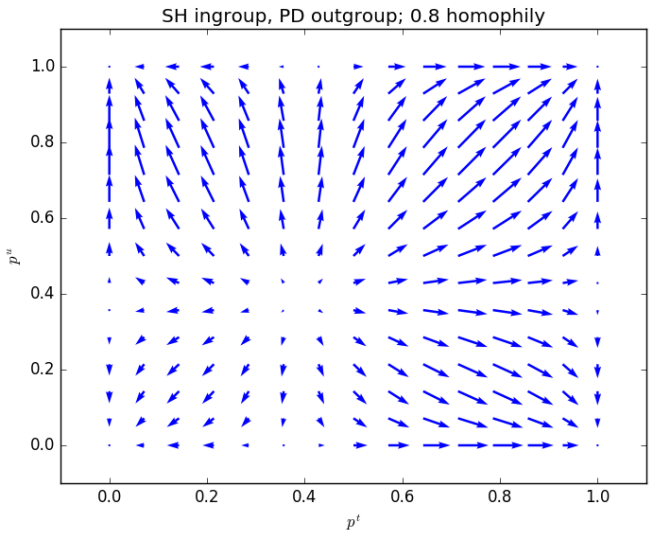

Figure 6.5: A game where individuals face a coordination game (Stag Hunt) with ingroup members and a Prisonner's Dilemma with outgroup members, and must decide on a strategy to play. With high levels of homophily (0.8 in this plot), the cooperative outcome becomes the most likely.

6.5.2 Group Preferences

Individuals often favor their own group over others. This could be represented by the following payoff matrices:

$$A_{\text{ingroup}} = \begin{bmatrix} 6 & 0 \\ 5 & 1 \end{bmatrix}$$

$$A_{\text{outgroup}} = \begin{bmatrix} 3 & 0 \\ 5 & 1 \end{bmatrix}$$

In other words, individuals play a prisoner's dilemma game against the outgroup, but receive 6 points for cooperation in-group. In this case, high levels of homophily can facilitate cooperation, as seen in Figure 6.5.2.

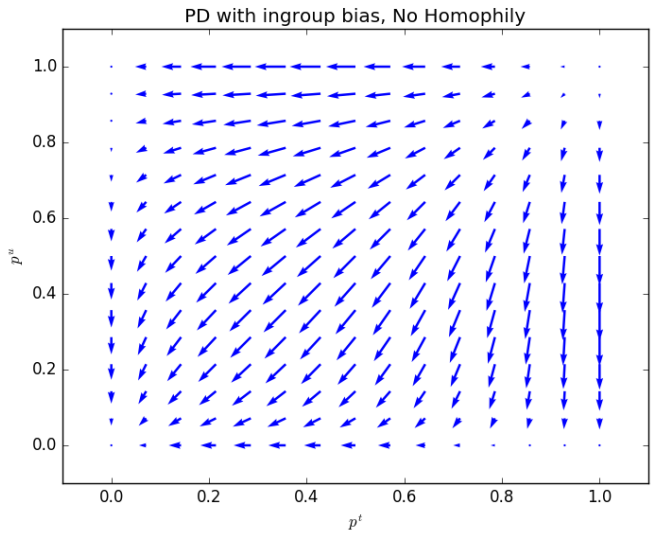

Figure 6.6: The Prisoner's Dilemma but with an ingroup bias and no homophily. Defection dominates (bottom left).

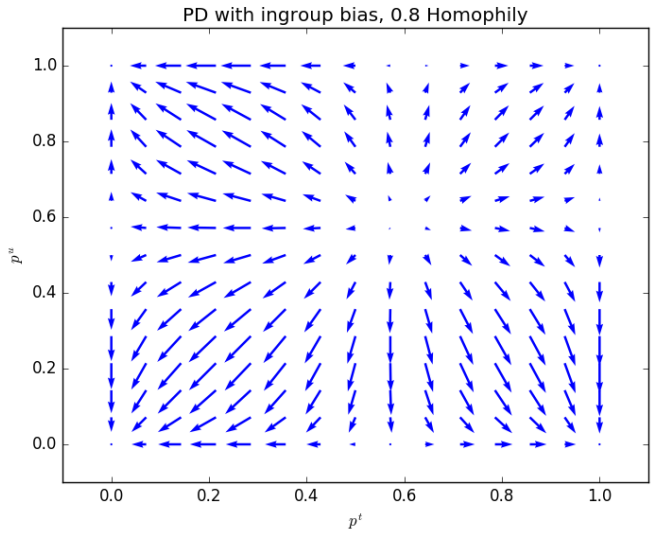

Figure 6.7: The Prisoner's Dilemma but with an ingroup bias and high homophily (0.8). Cooperation, as well as mixed cooperate-defect outcomes, are stable equilibria in addition to defect-defect.

6.6 Conclusion

The two main results of this paper appear contradictory: social structure can be irrelevant, and changes to social structure can qualitatively alter the set of stable outcomes. The tension in these results imply that social structure's effect on outcomes is situation-specific. Such a result aligns with the experience of sociologists. If we understand the utility that individuals derive from social interactions, and we understand information transmission procedures among individuals, then we can directly analyze how social structure affects macrolevel outcomes.

Sociologists have long argued for the causal importance of social structure. This paper has shown that in one theoretical formulation of this problem, social structure can be pivotal for facilitating or inhibiting cooperation.

This model has clear limitations that can be studied in future work. Here, agents play a simple pairwise game and must pick a single strategy. In reality, it is often easy for individuals to choose strategies based on the identity of their game partners. Future work can study the impact of strategy switching based on group identity. A particularly interesting case could be when individuals must pay a cost to switch strategies (which in some cases can be thought of as a cost of discrimination against outgroup members).

The game space can also be further explored. I have studied primarily well-known games like the Prisoner's Dilemma and the Stag Hunt. However, other pairwise games may provide fruitful for analysis.

CHAPTER 7

CONCLUSION

This dissertation has explored two challenges to sociological research using behavioral trace data, along with empirical findings about intergroup contact online and theoretical implications of homophily for cooperation. There are important future directions to this work, several of which I outline here.

Microlevel social contagion research must account for the opacity problem to properly estimate individual sensitivity to peer behavior. This increased precision can directly improve the precision of estimates of social reinforcement. It can also be employed to better study complex contagion. Accounting for opacity has the potential to facilitate a better typology of how behaviors, norms, fads, and social change spread through social network by facilitating understanding of microlevel dynamics. Better understanding of microlevel dynamics can also facilitate more accurate macrolevel modeling: how far will cascades reach? Which nodes will ultimately refuse to adopt and stall cascade progress? Because cascades are difficult to study experimentally, accounting for the opacity problem can facilitate better understanding of macrolevel dynamics as well via understanding the heterogeneity in behavior. This heterogeneity is fundamental to contagion research, being a core part of models dating at least to (85), but is difficult to study in practice. Subsequent research can more directly address the presence of and implications of heterogeneity in social contagion.

Research into social segregation in online spaces is frequently difficult because of data access, privacy, and data quality concerns. In public media such as Twitter, an important challenge is obtaining predictions for social interaction within and between groups. I have studied two approaches to the problem: correcting group proportions post-hoc, or adjusting predictions to account for the network. These approaches allow better inference of intergroup contact and can be employed to more accurately study social structure in online spaces. Future methodological work can focus on at least two primary directions. First, there is the issue of using novel methodology to improve prediction and inference quality. Methods for summarizing graph properties such as node embeddings may prove useful in this direction. Secondly, there is the application of these results to offline measures of social interaction, such as survey methodology found in the General Social Survey. When an individual reports from memory the education, race, gender, age, income, or occupation of a peer, there is the possibility for error. When error is correlated with the network, intergroup contact may be estimated with error much in the same way a prediction model would make an error. There is the need for additional research on both of these topics.

Substantively, studying social segregation in online spaces is important for many reasons. Among these are understanding the reproduction of offline social segregation online, the presence of unique cross-cutting foci, the presence or absence of informational echo chambers, the presence of in-group social support in online communities, and the process of insular communities generating harassing or other toxic behavior. The findings

presented here about race and gender social segregation on Twitter indicate that cross-gender interaction is more common in this online space than is reported among discussion partners in the General Social Survey, while cross-race interaction displays an asymmetric pattern: it is lower for Black Twitter users and may be lower for White Twitter users. These findings suggest that online spaces can facilitate greater cross-group interaction, but say nothing as to the particular qualities of those interactions. Further research is needed on this topic to better understand types of communication across groups.

It is my hope that the work in this dissertation will meaningfully facilitate addressing these important questions.

BIBLIOGRAPHY

[1] Faiyaz Al Zamal, Wendy Liu, and Derek Ruths. Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors. *ICWSM*, 270, 2012.

[2] Joshua D Angrist and Jrn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricists Companion*. Princeton University Press, 2009.

[3] Sinan Aral, Lev Muchnik, and Arun Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51):21544–21549, 2009.

[4] Sinan Aral and Dylan Walker. Identifying influential and susceptible members of social networks. *Science*, 337(6092):337–341, 2012.

[5] Pierre Auger, RB Parra, and E Sanchez. Behavioral Dynamics of Two Interacting HawkDove Populations. *Math. Model. Methods \ldots*, 11(4):645–661, 2001.

[6] Robert Axelrod. The Emergence of Cooperation among Egoists. *American Political Science Review*, 75(02):306–318, June 1981.

[7] Robert Axelrod and WD Hamilton. *The Evolution of Cooperation*, volume 211. 1981.

[8] Lars Backstrom. Group formation in large social networks: membership, growth, and evolution. *KDD*, pages 44–54, 2006.

[9] Stefanie Bailey and Peter V Marsden. Interpretation and interview context: examining the General Social Survey name generator using cognitive methods. *Social Networks*, 21(3):287–309, July 1999.

[10] Eytan Bakshy, Dean Eckles, Rong Yan, and Itamar Rosenn. Social Influence in Social Advertising: Evidence from Field Experiments. In *EC '12*, pages 146–161, 2012.

[11] Eytan Bakshy, Solomon Messing, and Lada Adamic. Exposure to ideologically diverse news and opinion on Facebook. *Science*, 10(1126):1–5, 2015.

[12] Abhijit Banerjee, Arun G. Chandrasekhar, Esther Duflo, and Matthew O. Jackson. The diffusion of microfinance. *Science*, 341(July):1236498, 2013.

[13] Albert-Laszlo Barabasi and Reka Albert. Emergence of Scaling in Random Networks. *Science*, 286(5439):509–512, October 1999.

[14] Vladimir Barash, Christopher Cameron, and Michael Macy. Critical phenomena in complex contagions. *Social Networks*, 34(4):451–461, October 2012.

[15] Pablo Barbera. Less is more ? How demographic sample weights can improve public opinion estimates based on Twitter data . *Work. Pap. para NYU*, 2016.

[16] Pablo Barbera, John T Jost, Jonathan Nagler, Joshua A Tucker, and Richard Bonneau. Tweeting From Left to Right: Is Online Politi-

cal Communication More Than an Echo Chamber? *Psychol. Sci.*, 26(10):1531–42, 2015.

[17] Peter S. Bearman and James Moody. Suicide and Friendships Among American Adolescents. *American Journal of Public Health*, 94(1):89–95, January 2004.

[18] Rachel Behler, Chan Suh, Matthew Brashears, and Yongren Shi. Familiar faces, familiar spaces: Social similarity and co-presence in non-relational behavioral convergence. *Network Science*, 6(3):396–429, September 2018.

[19] George Berry, Christopher J. Cameron, Patrick Park, and Michael W. Macy. The Opacity Problem in Social Contagion. *Social Networks*, 56:93–101, January 2019. arXiv: 1702.02700.

[20] George Berry, Antonio Sirianni, Nathan High, Agrippa Kellum, Ingmar Weber, and Michael Macy. Estimating Group Properties in Online Social Networks with a Classifier. In *Social Informatics*, volume 11185, pages 67–85. Springer International Publishing, Cham, 2018.

[21] George Berry, Antonio Sirianni, Ingmar Weber, Jisun An, and Michael Macy. Going beyond accuracy: estimating homophily in social networks using predictions. *arXiv:2001.11171 [cs]*, January 2020. arXiv: 2001.11171.

[22] Elisa Jayne Bienenstock and Phillip Bonacich. The core as a solution to exclusionary networks. *Social Networks*, 14(3-4):231–243, 1992.

[23] Peter M Blau. A Macrosociological Theory of Social Structure. *American Journal of Sociology*, 83(1):26–54, 1977.

[24] Peter M. Blau, Terry C. Blum, and Joseph E. Schwartz. Heterogeneity and Intermarriage. *American Sociological Review*, 47(1):45, 1982.

[25] Javier Borge-Holthoefer, Raquel A. Banos, Sandra Gonzalez-Bailon, and Yamir Moreno. Cascading behaviour in complex socio-technical networks. *Journal of Complex Networks*, 1(1):3–24, June 2013.

[26] Andrei Boutyline and Robb Willer. The Social Structure of Political Echo Chambers: Variation in Ideological Homophily in Online Networks: Political Echo Chambers. *Political Psychology*, 38(3):551–569, June 2017.

[27] Robert Boyd and Peter J. Richerson. *Culture and the Evolutionary Process*. University of Chicago Press, 1988.

[28] Yann Bramoulle, Habiba Djebbari, and Bernard Fortin. Identification of peer effects through social networks. *Journal of Econometrics*, 150(1):41–55, May 2009.

[29] Ronald S. Burt. Cohesion Versus Structural Equivalence as a Basis for Network Subgroups. *Sociol. Methods Res.*, 7(2):189–212, 1978.

[30] Ronald S. Burt. *Structural Holes: The Social Structure of Competition*. Harvard University Press, 1995.

[31] R.S. Burt. Structural Holes and Good Ideas. *American Journal of Sociology*, 110(2):349–399, 2004.

[32] Ellsworth Campbell and Marcel Salath. Complex social contagion makes networks more vulnerable to disease outbreaks. *Scientific Reports*, 3:1905, 2013.

[33] Damon Centola, Victor M. Eguiluz, and Michael W. Macy. Cascade dynamics of complex propagation. *Physica A: Statistical Mechanics and its Applications*, 374(1):449–456, 2007.

[34] Damon Centola and Michael Macy. Complex Contagions and the Weakness of Long Ties. *American Journal of Sociology*, 113(3):702–734, 2007.

[35] Nina Cesare. United We Tweet?: A Quantitative Analysis of Racial Differences in Twitter Use. page 181, 2017.

[36] Nina Cesare, Christan Grant, Quynh Nguyen, Hedwig Lee, and Elaine O. Nsoesie. How well can machine learning predict demographics of social media users? *arXiv:1702.01807 [cs]*, February 2017. arXiv: 1702.01807.

[37] Nina Cesare, Hedwig Lee, Tyler McCormick, and Emma S. Spiro. Redrawing the 'Color Line': Examining Racial Segregation in Associative Networks on Twitter. *arXiv:1705.04401 [cs]*, May 2017. arXiv: 1705.04401.

[38] Justin Cheng, Lada Adamic, P. Alex Dow, Jon M. Kleinberg, and Jure Leskovec. Can cascades be predicted? *WWW*, pages 925–936, 2014.

[39] M. De Choudhury. Tie Formation on Twitter: Homophily and Structure of Egocentric Networks. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 465–470, October 2011.

[40] Nicholas A. Christakis and James H. Fowler. The Spread of Obesity in a Large Social Network Over 32 Years. *New England Journal of Medicine*, 357(4):370–9, 2007.

[41] Morgane Ciot, Morgan Sonderegger, and Derek Ruths. Gender Inference of Twitter Users in Non-English Contexts. In *EMNLP*, pages 1136–1145, 2013.

[42] Meredith D. Clark. To Tweet Our Own Cause: A Mixed-Methods Study of the Online Phenomenon of "Black Twitter". 2014.

[43] James Coleman. *Foundations of Social Theory*. Harvard University Press, 1994.

[44] James Coleman, Elihu Katz, and Herbert Menzel. The Diffusion of an Innovation Among Physicians. *Sociometry*, 20(4):253–270, 1957.

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

[46] James S Coleman. Social Capital in the Creation of Human Capital. *American Journal of Sociology*, 94(May):S95–S120, 1988.

[47] Elanor Colleoni, Alessandro Rozza, and Adam Arvidsson. Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data: Political Homophily on Twitter. *Journal of Communication*, 64(2):317–332, April 2014.

[48] Ryan Compton, David Jurgens, and David Allen. Geotagging One Hundred Million Twitter Accounts with Total Variation Minimization. *arXiv Preprint. arXiv1404.7152*, 2014.

[49] Karen S. Cook and Richard M Emerson. Power, Equity and Commitment in Exchange Networks. *American Sociological Review*, 43(5):721–739, 1978.

[50] David Crandall, Dan Cosley, Daniel Huttenlocher, Jon Kleinberg, and Siddharth Suri. Feedback effects between similarity and social influence in online communities. *KDD*, page 160, 2008.

[51] Ross Cressman and Yi Tao. The replicator equation and other game dynamics. *Proceedings of the National Academy of Sciences*, 111(S3):10810–10817, 2014.

[52] Aron Culotta and Jennifer Cutler. Predicting Twitter User Demographics using Distant Supervision from Website Traffic Data. 55:389–408, 2016.

[53] Aron Culotta, Nirmal Ravi Kumar, and Jennifer Cutler. Predicting the Demographics of Twitter Users from Website Traffic Data. In *AAAI*, pages 72–78, 2015.

[54] Sergio Currarini, Matthew O. Jackson, and Paolo Pin. An Economic Model of Friendship: Homophily, Minorities, and Segregation. *Econometrica*, 77(4):1003–1045, 2009.

[55] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. Racial Bias in Hate Speech and Abusive Language Detection Datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy, 2019. Association for Computational Linguistics.

[56] Giacomo De Giorgi, Michele Pellizzari, and Silvia Redaelli. Identification of Social Interactions through Partially Overlapping Peer Groups. *American Economic Journal*, 2:241–275, 2010.

[57] Daniel DellaPosta, Shi Yongren, and Michael Macy. Why Do Liberals Drink Lattes? *American Journal of Sociology*, 120(5):1473–1511, 2015.

[58] Paul J. DiMaggio and Filiz Garip. Network Effects and Social Inequality. *Annual Review of Sociology*, 38(1):93–118, 2012.

[59] Yuxin Ding, Shengli Yan, YiBin Zhang, Wei Dai, and Li Dong. Predicting the attributes of social network users using a graph-based machine learning method. *Computer Communications*, 73:3–11, January 2016.

[60] Thomas A. DiPrete, Andrew Gelman, Tyler McCormick, Julien Teitler, and Tian Zheng. Segregation in Social Networks Based on Acquaintanceship and Trust. *American Journal of Sociology*, 116(4):1234–83, 2011.

[61] Thomas A. DiPrete, Andrew Gelman, Julien Teitler, Tian Zheng, and Tyler McCormick. Segregation in social networks based on acquaintanceship and trust. 2008.

[62] N. Eagle, A. Pentland, and D. Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36):15274–15278, September 2009.

[63] B. Efron. Bootstrap Methods: Another Look at the Jackknife. *Statistics (Ber).*, 10(2):1100–1120, 1982.

[64] Richard M Emerson. Power-Dependence Relations. 27(1):31–41, 1962.

[65] Quan Fang, Jitao Sang, Changsheng Xu, and M Hossain. *Relational User Attribute Inference in Social Media*, volume 17. July 2015.

[66] Thomas J. Fararo. Biased networks and social structure theorems. *Social Networks*, 3(2):137–159, 1981.

[67] Thomas J Fararo and John Skvoretz. Action and Institution , Network and Function: The Cybernetic Concept of Social Structure. *Sociol. Forum*, 1(2):219–250, 1986.

[68] George Forman. Counting positives accurately despite inaccurate classification. *Machine Learning: ECML 2005*, pages 564–575, 2005.

[69] George Forman. Quantifying counts and costs via classification. *Data Mining and Knowledge Discovery*, 17(2):164–206, October 2008.

[70] Linton C Freeman. Centrality in Social Networks Conceptual Clarification. *Social Networks*, 1(1968):215–239, 1978.

[71] Noah E. Friedkin and Eugene C. Johnsen. Social Influence and Opinions. *Journal of Mathematical Sociology*, 15(3-4):193–205, 1990.

[72] Noah E Friedkin and Eugene C Johnsen. *Social Influence Networks and Opinion Change*. 1999.

[73] Wei Gao and Fabrizio Sebastiani. From classification to quantification in tweet sentiment analysis. *Social Network Analysis and Mining*, 6(1), December 2016.

[74] Michael Genkin, Cheng Wang, George Berry, and Matthew E. Brashears. Blaunet: An R-based graphical user interface package to analyze Blau space. *PLOS ONE*, 13(10):e0204990, 2018.

[75] Krista J Gile and Mark S Handcock. Respondent-Driven Sampling: an Assessment of Current Methodology. *Sociol. Methodol.*, 40(1):285–327, 2010.

[76] Herbert Gintis. *Game Theory Evolving*. Princeton University Press, 2000.

[77] Minas Gjoka, Maciej Kurant, Carter T. Butts, and Athina Markopoulou. A Walk in Facebook: Uniform Sampling of Users in Online Social Networks. *arXiv:0906.0060 [physics, stat]*, May 2009. arXiv: 0906.0060.

[78] Minas Gjoka, Maciej Kurant, Carter T. Butts, and Athina Markopoulou. Walking in facebook: A case study of unbiased sampling of OSNs. *Proc. - IEEE INFOCOM*, 2010.

[79] Minas Gjoka, Maciej Kurant, Carter T. Butts, and Athina Markopoulou. Practical Recommendations on Crawling Online Social Networks. *IEEE Journal on Selected Areas in Communications*, 29(9):1872–1892, October 2011.

[80] S Goel and M J Salganik. Respondent-driven sampling as Markov chain Monte Carlo. *Stat Med*, 28(17):2202–2229, 2009.

[81] Scott A. Golder and Michael W. Macy. Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures. *Science*, 333(6051):1878–1881, 2011.

[82] Neil Zhenqiang Gong, Ameet Talwalkar, Lester Mackey, Ling Huang, Eui Chul Richard Shin, Emil Stefanov, Elaine (Runting) Shi, and Dawn Song. Joint Link Prediction and Attribute Inference Using a Social-Attribute Network. *ACM Transactions on Intelligent Systems and Technology*, 5(2):1–20, April 2014.

[83] Sandra Gonzalez-Bailon, Javier Borge-Holthoefer, Alejandro Rivero,

and Yamir Moreno. The Dynamics of Protest Recruitment through an Online Network. *Scientific Reports*, 1(1), December 2011.

[84] Mark Granovetter. The Strength of Weak Ties. *American Journal of Sociology*, 78(6):1360–1380, 1973.

[85] Mark Granovetter. Threshold models of collective behavior. *American Journal of Sociology*, 83(6):1420–1443, 1978.

[86] Aditya Grover and Jure Leskovec. node2vec: Scalable Feature Learning for Networks. pages 855–864. ACM Press, 2016.

[87] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring Network Structure, Dynamics, and Function using NetworkX. *SciPy*, pages 11–16, 2008.

[88] Yosh Halberstam and Brian Knight. Homophily, group size, and the diffusion of political information in social networks: Evidence from Twitter. *Journal of Public Economics*, 143:73–88, November 2016.

[89] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*, volume 1 of *Springer series in statistics New York*. 2008.

[90] Douglas Heckathorn and Joan Jeffrib. Finding the beat: Using respondent-driven sampling to study jazz musicians. *Poetics*, 28:307–329, 2001.

[91] Douglas D. Heckathorn. Collective Sanctions and the Creation of Prisoner's Dilemma Norms. *American Journal of Sociology*, 94(3):535, 1988.

[92] Douglas D. Heckathorn. Respondent-Driven Sampling II: Deriving Valid Population Estimates from Chain-Referral Samples of Hidden Populations. *Soc. Probl.*, 49(1):11–34, 2002.

[93] Dirk Helbing and Anders Johansson. Cooperation, norms, and revolutions: A unified game-theoretical approach. *PLoS One*, 5(10), 2010.

[94] Dirk Helbing and Anders Johansson. Evolutionary dynamics of populations with conflicting interactions: Classification and analytical treatment considering asymmetry and power. *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.*, 81(1):56–58, 2010.

[95] Dirk Helbing, Wenjian Yu, Karl Dieter Opp, and Heiko Rauhut. Conditions for the emergence of shared norms in populations with incompatible preferences. *PLoS One*, 9(8), 2014.

[96] Herbert W. Hethcote. The mathematics of infectious diseases. *SIAM review*, 42(4):599–653, 2000.

[97] Itai Himelboim, Kaye D Sweetser, Spencer F Tinkham, Kristen Cameron, Matthew Danelo, and Kate West. Valence-based homophily on Twitter: Network Analysis of Emotions and Political Talk in the 2012 Presidential Election. *New Media & Society*, 18(7):1382–1400, August 2016.

[98] William R. Hobbs, Moira Burke, Nicholas A. Christakis, and James H. Fowler. Online social integration is associated with reduced mortality risk. *Proceedings of the National Academy of Sciences*, 113(46):12980–12984, November 2016.

[99] Bas Hofstra, Rense Corten, Frank van Tubergen, and Nicole B. Ellison. Sources of Segregation in Social Networks: A Novel Approach Using Facebook. *American Sociological Review*, page 000312241770565, May 2017.

[100] Bas Hofstra and Niek C de Schipper. Predicting ethnicity with first names in online social media networks. *Big Data & Society*, pages 1–14, 2018.

[101] Petter Holme and Beom Jun Kim. Growing scale-free networks with tunable clustering. *Physical review E*, 65(2):026107, 2002.

[102] Stefano M. Iacus, Gary King, and Giuseppe Porro. Causal Inference without Balance Checking: Coarsened Exact Matching. *Political Analysis*, 20(1):1–24, 2012.

[103] Matthijs Kalmijn. Intermarriage and Homogamy: Causes, Patterns, Trends. *Annual Review of Sociology*, 24:395–421, 1998.

[104] Fariba Karimi, Mathieu Gnois, Claudia Wagner, Philipp Singer, and Markus Strohmaier. Visibility of minorities in social networks. *arXiv preprint arXiv:1702.00150*, 2017.

[105] Fariba Karimi, Mathieu Gnois, Claudia Wagner, Philipp Singer, and Markus Strohmaier. Homophily influences ranking of minorities in social networks. *Scientific Reports*, 8(1):11077, July 2018.

[106] Shoko Kato, Akihiro Koide, Takayasu Fushimi, Kazumi Saito, and Hiroshi Motoda. Network Analysis of Three Twitter Functions: Favorite, Follow and Mention. In *Knowledge Management and Acquisition for Intelligent Systems*, volume 7457, pages 298–312. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[107] David Kempe, Jon Kleinberg, and Eva Tardos. Maximizing the spread of influence through a social network. *KDD*, page 137, 2003.

[108] David Kempe, Jon Kleinberg, and Eva Tardos. Influential Nodes in a Diffusion Model for Social Networks. *ICALP*, 3580:1127–1138, 2005.

[109] Gary King, Ori Rosen, and Martin Abba Tanner, editors. *Ecological inference: new methodological strategies*. Analytical methods for social research. Cambridge University Press, Cambridge, UK ; New York, 2004.

[110] A.S. Klovdahl, J.J. Potterat, D.E. Woodhouse, J.B. Muth, S.Q. Muth, and W.W. Darrow. Social networks and infectious disease: The Colorado Springs study. *Social Science & Medicine*, 38(1):79–88, January 1994.

[111] Gueorgi Kossinets and Duncan J. Watts. Origins of Homophily in an

Evolving Social Network. *American Journal of Sociology*, 115(2):405–450, 2009.

[112] Maciej Kurant, Minas Gjoka, Carter T. Butts, and Athina Markopoulou. Walking on a Graph with a Magnifying Glass: Stratified Sampling via Weighted Random Walks. In *Proceedings of the ACM SIGMETRICS Joint International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '11, pages 281–292, New York, NY, USA, 2011. ACM.

[113] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is Twitter, a social network or a news media? page 591. ACM Press, 2010.

[114] Kristina Lerman. Information Is Not a Virus, and Other Consequences of Human Cognitive Limits. *Future Internet*, 8(2):21, May 2016.

[115] Jure Leskovec and Lada a Adamic. The Dynamics of Viral Marketing. 1(May 2007):1–46, 2008.

[116] K. Lewis, M. Gonzalez, and J. Kaufman. Social selection and peer influence in an online social network. *Proceedings of the National Academy of Sciences*, 109:68–72, 2012.

[117] Anqi Liu and Brian Ziebart. Robust classification under sample selection bias. In *Advances in neural information processing systems*, pages 37–45, 2014.

[118] Wendy Liu and Derek Ruths. What's in a Name? Using First Names as Features for Gender Inference in Twitter. In *AAAI spring symposium: Analyzing microtext*, volume 13, page 01, 2013.

[119] Michael W. Macy. Chains of Cooperation: Threshold Effects in Collective Action. *American Sociological Review*, 56(6):730–747, 1991.

[120] Michael W. Macy and John Skvoretz. The Evolution of Trust and Cooperation between Strangers: A Computational Model. *American Sociological Review*, 63(5):638, 1998.

[121] Michael W. Macy and Robert Willer. From Factors to Actors: Computational Sociology and Agent-Based Modeling. *Annual Review of Sociology*, 28(1):143–166, August 2002.

[122] Eric Malmi and Ingmar Weber. You Are What Apps You Use: Demographic Prediction Based on User's Apps. In *ICWSM*, pages 635–638, 2016.

[123] Barry Markovsky, John Skvoretz, David Willer, and Michael J Lovaglia. The Seeds of Weak Power: An Extension of Network Exchange Theory. *American Sociological Review*, 58(2):197–209, 1993.

[124] Peter V Marsden. Core Discussion Networks of Americans. *American Sociological Review*, 52(1):122–131, 1987.

[125] Murdoch K. McAllister and James N. Ianelli. Bayesian stock assessment using catch-age data and the sampling-importance resampling

algorithm. *Candian Journal of Fisheries and Aquatic Sciences*, 54(2):284–300, 1997.

[126] J Miller Mcpherson and Lynn Smith-lovin. Homophily in Voluntary Organizations: Status Distance and the Composition of Face-to-Face Groups. *American Sociological Review*, 52(3):370–379, 1987.

[127] M. McPherson, L. Smith-Lovin, and M. E. Brashears. Social Isolation in America: Changes in Core Discussion Networks over Two Decades. *American Sociological Review*, 71(3):353–375, 2006.

[128] Miller McPherson, Lynn Smith-Lovin, and James M. Cook. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, 27(May):415–444, 2001.

[129] Johnnatan Messias, Pantelis Vikatos, and Fabrcio Benevenuto. White, man, and highly followed: gender and race inequalities in Twitter. In *Proceedings of the International Conference on Web Intelligence - WI '17*, pages 266–274, Leipzig, Germany, 2017. ACM Press.

[130] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[131] Ehsan Mohammady and Aron Culotta. Using county demographics to infer attributes of Twitter users. *Acl 2014*, page 7, 2014.

[132] Mario Molina and Filiz Garip. Machine Learning for Sociology. *Annual Review of Sociology*, 45(1):27–45, 2019.

[133] Kelly A. Mollica, Barbara Gray, and Linda K. Trevino. Racial Homophily and Its Persistence in Newcomers' Social Networks. *Organization Science*, 14(2):123–136, 2003.

[134] Dong-Phuong Nguyen, Rilana Gravel, Rudolf Berend Trieschnigg, and Theo Meder. " How old do you think I am?" A study of language and age in Twitter. 2013.

[135] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning - ICML '05*, pages 625–632, Bonn, Germany, 2005. ACM Press.

[136] M. A. Nowak. Five Rules for the Evolution of Cooperation. *Science*, 314(5805):1560–1563, December 2006.

[137] Hisashi Ohtsuki and Martin A. Nowak. The Replicator Equation on Graphs. *Journal of theoretical biology*, 243(1):86–97, November 2006.

[138] Eli Pariser. *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*. Penguin, May 2011. Google-Books-ID: wcalrOI1YbQC.

[139] Patrick S Park, Joshua E Blumenstock, and Michael W Macy. The strength of long-range ties in population-scale social networks. page 5, 2018.

[140] Romualdo Pastor-Satorras, Claudio Castellano, Piet Van Mieghem, and Alessandro Vespignani. Epidemic processes in complex networks. *Reviews of modern physics*, 87(3):925, 2015.

[141] Fabian Pedregosa, Gal Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and others. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011.

[142] Bryan Perozzi and Steven Skiena. DeepWalk : Online Learning of Social Representations. *Sigkdd*, pages 701–710, 2014.

[143] Jesus Ramirez-Valles, Douglas D. Heckathorn, Raquel Vzquez, Rafael M. Diaz, and Richard T. Campbell. From Networks to Populations: The Development and Application of Respondent-Driven Sampling Among IDUs and Latino Gay Men. *AIDS and Behavior*, 9(4):387–402, December 2005.

[144] Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. Classifying Latent User Attributes in Twitter. pages 37–44, 2009.

[145] Anatol Rapoport. A Probabilistic Approach to Networks. *Soc. Networks*, 2(1):1–18, 1980.

[146] Ronald Read, C. and Robin J. Wilson. *An Atlas of Graphs*. Clarendon Press, 1998.

[147] Radim Rehurek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *In Proceedings of the Lrec 2010 Workshop on New Challenges for Nlp Frameworks*, pages 45–50, 2010.

[148] Bruno Ribeiro and Don Towsley. Estimating and Sampling Graphs with Multidimensional Random Walks. In *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement*, IMC '10, pages 390–403, New York, NY, USA, 2010. ACM.

[149] Luis E. C. Rocha, Fredrik Liljeros, and Petter Holme. Simulated Epidemics in an Empirical Spatiotemporal Network of 50,185 Sexual Contacts. *PLOS Computational Biology*, 7(3):e1001109, March 2011.

[150] Daniel M Romero, Brendan Meeder, and Jon Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. *WWW*, pages 695–704, 2011.

[151] Donald B. Rubin. The Calculation of Posterior Distributions by Data Augmentation: Comment: A Noniterative Sampling/Importance Resampling Alternative to the Data Augmentation Algorithm for Creating a Few Imputations When Fractions of Missing Information Are Modest: The SIR Algorithm. *Journal of the American Statistical Association*, 82(398):543–546, 1987.

[152] Matthew J Salganik and Douglas D. Heckathorn. Sampling and Esti-

mation in Hidden Populations Using Respondent-Driven Sampling. *Sociological Methodology*, 1:193–240, 2004.

[153] Larry Samuelson and Jianbo Zhang. Evolutionary stabilty in asymmetric games. *J. Econ. Theory*, 57:363–391, 1992.

[154] Christine R. Schwartz and Robert D. Mare. Trends in Educational Assortative Mating From 1940 to 2003. 2005.

[155] Christine R. Schwartz and Robert D. Mare. The Proximate Determinants of Educational Homogamy: The Effects of First Marriage, Marital Dissolution, Remarriage, and Educational Upgrading. *Demography*, 49(2):629–650, May 2012.

[156] John Skvoretz. Salience, Heterogenaeity and Consolidation of Parameters: Civilizing Blau's Primitive Theory. *American Sociological Review*, 48(3):360–375, 1983.

[157] John Skvoretz. Complexity theory and models for social networks. *Complexity*, 8(1):47–55, 2002.

[158] Brian Skyrms. *Evolution of the Social Contract*. Cambridge University Press, 2014.

[159] Mario Luis Small. Weak ties and the core discussion network: Why people regularly discuss important matters with unimportant alters. *Social Networks*, 35(3):470–483, July 2013.

[160] Mario Luis Small. *Someone To Talk To*. Oxford University Press, September 2017. Google-Books-ID: 7sQ2DwAAQBAJ.

[161] J. Maynard Smith and G. R. Price. Logic of animal conflict. *Nature*, 246(5499):15–18, 1973.

[162] Jeffrey A. Smith, Miller McPherson, and Lynn Smith-Lovin. Social Distance in the United States: Sex, Race, Religion, Age, and Education Homophily among Confidants, 1985 to 2004. *American Sociological Review*, 79(3):432–456, 2014.

[163] Greg Ver Steeg, Rumi Ghosh, and Kristina Lerman. What stops social epidemics? *arXiv preprint arXiv:1102.1985*, 2011.

[164] David Strang and Sarah A. Soule. Diffusion in Organizations and Social Movements: From Hybrid Corn to Poison Pills. *Annual Review of Sociology*, 24(1):265–290, 1998.

[165] Steven H. Strogatz. Exploring complex networks. *Nature*, 410(6825):268–276, 2001.

[166] Lubos Takac and Zabovsky. Data Analysis in Public Social Networks. Lomza, Poland, 2012.

[167] Mike Thelwall. Homophily in MySpace. *Journal of the American Society for Information Science and Technology*, 60(2):219–231, 2009.

[168] Johan Ugander, Lars Backstrom, Cameron Marlow, and Jon Klein-

berg. Structural diversity in social contagion. *Proceedings of the National Academy of Sciences*, 109(16):5962–5966, April 2012.

[169] Thomas W. Valente. *Network Models of the Diffusion of Innovations*. Cresskill New Jersey Hampton Press, 1995.

[170] Thomas W. Valente. Social network thresholds in the diffusion of innovations. *Social Networks*, 18(1):69–89, 1996.

[171] Thomas W. Valente. Network Interventions. *Science*, 337(6090):49–53, 2012.

[172] Susan van den Hof, Christine M. A. Meffre, Marina A. E. Conyn-van Spaendonck, Frits Woonink, Hester E. de Melker, and Rob S. van Binnendijk. Measles outbreak in a comunity with very low vaccine coverage, the Netherlands. *Emerging Infectious Diseases*, 7(3):593–597, 2001.

[173] Svitlana Volkova, Yoram Bachrach, Michael Armstrong, and Vijay Sharma. Inferring Latent User Properties from Texts Published in Social Media. In *AAAI*, pages 4296–4297, 2015.

[174] Erik Volz and Douglas D. Heckathorn. Probability based estimation theory for respondent driven sampling. *Journal of official statistics*, 24(1):79, 2008.

[175] Claudia Wagner, Philipp Singer, Fariba Karimi, J?rgen Pfeffer, and Markus Strohmaier. Sampling from Social Networks with Attributes. *WWW*, pages 1181–1190, 2017.

[176] Jack L. Walker. The Diffusion of Innovations among the American States. *American Political Science Review*, 63(03):880–899, September 1969.

[177] Pengfei Wang, Jiafeng Guo, Yanyan Lan, Jun Xu, and Xueqi Cheng. Your Cart tells You: Inferring Demographic Attributes from Purchase Data. pages 173–182. ACM Press, 2016.

[178] Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[179] Duncan J. Watts. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, 99(9):5766–5771, 2002.

[180] Duncan J. Watts. *Everything Is Obvious: *Once You Know the Answer*. Crown, March 2011. Google-Books-ID: kT_4AAAAQBAJ.

[181] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–2, June 1998.

[182] L. Weng, A. Flammini, A. Vespignani, and F. Menczer. Competition among memes in a world with limited attention. *Scientific Reports*, 2, March 2012.

[183] Lilian Weng, Filippo Menczer, and Yong-Yeol Ahn. Virality Prediction and Community Structure in Social Networks. *Scientific Reports*, 3, August 2013.

[184] Harrison C. White, Scott A. Boorman, and Ronald L. Breiger. Social Structure from Multiple Networks. I. Blockmodels of Roles and Positions. *American Journal of Sociology*, 81(4):730–780, 1976.

[185] Andreas Wimmer and Kevin Lewis. Beyond and below racial homophily: ERG models of a friendship network documented on Facebook. *American Journal of Sociology*, 116(2):583–642, 2010.

[186] Stefan Wojcik and Adam Hughes. How Twitter Users Compare to the General Public | Pew Research Center, April 2019.

[187] Bianca Zadrozny. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the twenty-first international conference on Machine learning*, page 114. ACM, 2004.