A Robust Inference Method for Decision Making in Networks

Article *in* MIS Quarterly · June 2021

CITATIONS
0

READS
257

3 authors:

Aaron Schecter
University of Georgia
23 PUBLICATIONS **125** CITATIONS
SEE PROFILE

Omid Nohadani
Purdue University
46 PUBLICATIONS **759** CITATIONS
SEE PROFILE

Noshir Contractor
Northwestern University
277 PUBLICATIONS **11,039** CITATIONS
SEE PROFILE

A Robust Inference Method for Decision Making in Networks

AARON SCHECTER
University of Georgia
Department of Management Information Systems
Athens, GA
aschecter@uga.edu

OMID NOHADANI
Benefits Science Technology
Boston, MA
onohadani@gmail.com

NOSHIR CONTRACTOR
Northwestern University
Departments of Communication,
Industrial Engineering & Management Sciences,
and Management & Organizations
Evanston, IL
nosh@northwestern.edu

[ACCEPTED FOR PUBLICATION AT MANAGEMENT INFORMATION SYSTEMS QUARTERLY]

A Robust Inference Method for Decision Making in Networks

ABSTRACT

Social network data collected from digital sources is increasingly used to gain insights into human behavior. However, while these observable networks constitute an empirical ground truth, the individuals within the network can perceive the network's structure differently – and they often act on these perceptions. As such, we argue that there is a distinct gap between the data used to model behaviors in a network, and the data internalized by people when they actually engage in behaviors. We find that statistical analyses of observable network structure do not consistently take into account these discrepancies, and this omission may lead to inaccurate inferences about hypothesized network mechanisms. To remedy this issue, we apply techniques of robust optimization to statistical models for social network analysis. Using robust maximum likelihood, we derive an estimation technique that immunizes inference to errors such as false positives and false negatives, without knowing a priori the source or realized magnitude of the error. We demonstrate the efficacy of our methodology on real social network datasets and simulated data. Our contributions extend beyond the social network context, as perception gaps may exist in many other economic contexts.

Keywords. *Robust optimization; social network analysis; maximum likelihood estimation; network cognition; inferential models; online networks*

INTRODUCTION

Social network analysis is an increasingly popular tool for studying behavior that is grounded upon the investigation of diverse relationships between various social entities (Monge and Contractor 2003; Wasserman and Faust 1994). Social network data, like most behavioral data, has traditionally been obtained through surveys or indirect observation of people's choices. The empirical analysis of organizational phenomena such as social networks has become increasingly viable thanks in part to the proliferation of online data in all facets of life (Kane et al. 2014; Lazer et al. 2009; Lazer et al. 2020; Leonardi and Contractor 2018). For each of the activities in which we engage online, electronic footprints or "digital traces" are created. Digital records include email exchanges, links between people's social network sites like Facebook or Twitter, posts to online forums such as Reddit and GitHub, clickstream data, electronic transactions, and mobile app usage.

There are numerous examples in the IS field in which these digital traces are leveraged to examine human behavior and decision making. In a study of content producers on social networking sites, Bhattacharya et al. (2019) found that individuals tend to form connections with others who produce similar content, but over time alter their posting behavior to be distinct from their contacts. Here, the authors leverage the friendship relations between users of the platform. Social networks have also been used to study software development and innovation. For instance, Singh et al. (2011) predict the success of open source projects as a function of their internal and external ties. In this study, network ties are formed between individuals who edit the same projects. Finally, several studies have explored the role of social networks in the adoption of new technologies (Aral and Walker 2014; de Matos et al. 2014) and the spread of user preferences (Dewan et al. 2017; Susarla et al. 2011). Underpinning each of these studies is a

reliance on observable digital trace data to infer some type of relationship, such as friendship or knowledge sharing. These inferred relationships are then leveraged to understand how people behave, what their preferences are, or what expertise they contribute.

However, there is a caveat to the use of observational data, particularly those collected from online networks. Social science research has long been aware that in some cases, the "ground truth" is not the data observable by researchers, but the world as it is perceived by the actors themselves (Richards, 1985). Many social and psychological theories of human behavior are based on individuals' perceptions -- an assertion well captured by the observation that "if men (sic!) define situations as real, they are real in their consequences" (Thomas and Thomas, 1928, p. 572). Inspired by the work of Thomas and Thomas (1928), network scholars such as Pattison (1994) and Krackhardt (1987; p. 128) observed that network "perceptions are real in their consequences even if they do not map one-to-one onto observed behaviors."

Following this logic, we distinguish between two distinct realities. The first, an ***individual belief***, is one that a specific actor perceives. The individual belief system is composed of all entities, states, or events that the person believes exist. In the social network literature, the collection of individuals' perceived networks is referred to as a cognitive social structure (Krackhardt 1987). The second is an ***empirical instantiation*** that a third party such as a researcher witnesses or a computer server logs. An empirical instantiation is built from observed records. such as digital trace data, with each data point representing actual events. While this "view from above" is reality for those studying the social system, there is no guarantee that the individual beliefs and empirical instantiations are the same. Yet, much of recent empirical research using digital trace data assumes, incorrectly, that these data are equivalent to individuals' perceptions, and in particular their perceptions of the network. Accordingly, there is

a definitive gap – what we call a *perception gap* – between the individual beliefs and the empirical instantiations. In network parlance, the perception gap manifests in a discrepancy between the collection of *individual networks* or CSS and the *empirical network* collected by researchers (Corman 1990).

This perception gap is a particularly salient problem when conducting behavioral inference, a key focus of network research. As Brashears and Quintane (2015) argue, "[b]ecause individual, preference-driven decisions will be based not on the actual state of the network, but on the perceived state of the network, the manner in which social networks are encoded and represented in memory can have a profound impact on the ultimate structure of a network and the behavior of network members" (p. 113). Thus, when the network changes or when individuals take advantage of their network, either by activating it or mobilizing it (Smith et al. 2011, 2020), the rationale for such actions is derived from an individual's goals, preferences, and their perspective on the network's current state (Corman and Scott 1994; Kilduff and Brass 2010). This view contrasts with the assumption that the empirical network data is unequivocally ground truth when it comes to explaining individuals' behaviors based on their perceptions of the network. More generally, individuals often take actions in any context based upon what they *believe*, not purely the empirical instantiations that researchers observe.

Our key research goal is to address the theoretical and methodological issues stemming from this perception gap. Thus, we ask: To what extent are inferences based on models that assume individuals have perfect knowledge of others' network ties robust to the vagaries of individuals' accurate perceptions of these ties? In addition, if they are not, can we develop robust inference techniques that have the capacity to identify variables that are statistically significant, even with contaminated data, while at the same time ruling out variables that might

appear significant with the observed data but turn non-significant under plausible assumptions of the observed data being contaminated? In this paper, we make three contributions.

1. First, we identify a problem in the way that social networks constructed from digital traces are analyzed, leading to a *perception gap*. We then formally model this perception gap and consider its impact on inference.
2. Second, we propose a method that remedies this issue by immunizing parameters from data errors. We present a novel extension of recent work on *robust maximum likelihood estimation* that applies to the exponential family of probability distributions.
3. Third, we derive a *robust test statistic* and demonstrate the potential for stronger statistical inference.

This paper is organized as follows. We first review studies that make use of digital network data and discuss their assumptions concerning perception. Further, we propose a number of relevant sources of discrepancy between these networks. Next, we discuss the potential impact these errors have on statistical inference. Finally, we propose a robust reformulation of inferential network models to address these various sources of error. To test our method, we conduct three studies: a simulation study, a laboratory example, and an empirical example.

BACKGROUND & MOTIVATION

Interaction Data and Social Networks

Traditionally, empirical instantiations of networks were captured through individuals' self-report of their interactions (Corman and Scott 1994; Krackhardt 1987). Network ties were in essence a measure of how frequently two individuals reported communicating with one another. Increasingly, interactions between two actors are captured through digital traces, rather than self-reports or direct observation. The broad availability of digital trace data has helped drive the

growth of social network analysis within the IS field (Agarwal et al. 2008; Cao et al. 2015; Oinas-Kukkonen et al. 2010). We define *digital trace data* as electronic records of interactions captured by an information system (Berente et al. 2018; Contractor, 2018; Lazer et al. 2009). These data come in a variety of forms: clickstreams, messaging logs, forum posts, and software contributions are all records that information systems capture and store. Like other forms of observable interactions, digital traces can be converted into pairwise relations that connect two participants. Email messages can be treated as directed links from the sender to recipient (e.g. Quintane and Carnabuci 2016). Relations on social media sites such as "following" someone, "commenting" on a post, or "liking" a message provide data on who is engaging with whom (Kane et al., 2014). Online message forums can also be transformed into social networks by finding "who replies to whom" in a thread (Faraj and Johnson 2011; Johnson et al. 2014).

The advantage of interaction data as compared to traditional sociometric survey data is that links between individuals represent actual connections, and generally correspond to actions taken by people. Additionally, interaction data – particular those that are digital – are cheaper to collect than surveys, are dynamically updateable and do not suffer from lack of responses or missing participants. Accordingly, much larger and complex networks can be modeled and analyzed using online data (Lazer et al., 2009). More recently, there have been calls to use digital trace data generated in organizations to help HR leverage people analytics – to help identify influences, innovators, those likely to quit and those likely to work well in a team (Leonardi and Contractor, 2018).

Consequently, more IS research has leveraged digital traces to understand different aspects of human behavior (Berger et al. 2014; Howison et al. 2011). In order to demonstrate the pervasiveness of online network data in management studies, we conducted a survey of recent

literature relevant to our study and summarize these observations in Table 1. However, relying solely upon digital trace data – or any interaction data for that matter – to conduct inference may lead to validity concerns (Howison et al. 2011; Vial 2019).

Table 1. Summary of digital network studies

Data Source	*Research Topics*	*Exemplar Studies*
Social networks, messaging, and emails	Peer-to-peer influence Information diffusion Product virality	Aral and Van Alstyne (2011) Aral and Walker (2012) Aral and Walker (2014) Bampo et al. (2008) Bapna and Umyarov (2015) Bapna et al. (2017a) Bapna et al. (2017b) Dewan et al. (2017) de Matos et al. (2014) Quintane and Carnabuci (2016) Susarla et al. (2011)
Software project affiliation	Developer learning Developer collaboration Problem solving Core-periphery emergence	Brunswicker and Schecter (2019) Dahlander and O'Mahony (2010) Foss et al. (2016) Quintane et al. (2014) Singh and Phelps (2013) Singh and Tan (2010) Singh et al. (2011)
Online forums and communities	Patterns of contribution Emergence of structure Emergence of leadership Participant collaboration Knowledge sharing	Chen et al. (2017) Dahlander and Frederiksen (2011) Faraj and Johnson (2011) Johnson et al. (2014) Johnson et al. (2015) Kudaravalli and Faraj (2008) Lu et al. (2017) Bhattacharya et al. (2019)

Namely, digital engagement or interactions do not necessarily signify a social relationship in the same way that a survey or interview might (Corman 1990). For instance, the relation "friendship" could be determined by asking individuals who they consider their friends. With event data, one can only observe online engagements such as "liking" or "tagging" on a platform such as Facebook or YouTube.

Consequently, the network constructed from trace data is at best a proxy for the underlying pattern of social relations. Why is this problematic? As Howison et al. (2011) point out, digital traces are *events* and thus represent *instances* of a relation. They alone do not make a social tie. However, "when working with trace data, it seems there is a tendency to take evidence

of instances (what was) and transmute that uncritically into evidence of topology (what could be/have been)" (Howison et al. 2011, pg. 790). Accordingly, network measures constructed from *any form of event data* are prone to misalignment between the measurement and the underlying construct, even at the aggregate (i.e., population) level. Because of this misalignment, there is a gap between the observable network – constructed vis-à-vis digital trace data – and the relations as they are *perceived* by the individuals in the network. As a result, conducting inference on the observable network will not accurately capture certain patterns of social interaction such as trust, friendship, or leadership (Brashears and Quintane 2015; Kilduff and Brass 2010). However, despite the potential limitations of online social network data reflecting perceptions, these forms of data are frequently used to test hypotheses about human attitudes and behavior. In the following section, we describe a variety of potential causes – both technical and cognitive – for this perception gap.

Perception Errors in Social Networks

Prevalence of Perception Errors

Extant empirical research, while somewhat limited, has consistently shown that individuals' beliefs do not align with empirical reality. Studies comparing email logs (Johnson et al. 2012; Quintane and Kleinbaum 2011; Wuchty and Uzzi 2011) show positive correlations between message exchanges and self-reported ties, but there are significant misalignments. In Johnson et al.'s (2012) study of email exchanges in a bank, the authors found correlations of 0.20 to 0.30 between email links and self-reported friendship, advice, and information ties. Wuchty and Uzzi (2011) similarly studied email logs and self-reported networks in a professional services organization. Their model attempted to predict self-reported ties from email exchanges; their optimally-tuned models achieved a true positive rate of at most 83.6% and a false positive

rate of at minimum 12.2%. In other words, even the best models still incorrectly predicted the presence (or lack thereof) of a tie more than 12% of the time.

Using Facebook interaction data, Gilbert and Karahalios (2009) built a predictive model to identify strong and weak friendship ties, as self-reported by participants in their sample. Their model was able to predict tie strength with a mean absolute error of approximately 10%. Other studies have used mobile phone proximity data (Eagle et al. 2009) and mobile phone call records (Onnela et al. 2007) to reconstruct networks and compare them to self-reported ties. Eagle et al. (2009) found that even among friends, there was only a correlation of 0.412 between mobile proximity records and reported proximity. Further, the authors found a strong effect of recency, and determined that after approximately one week, recollection of interactions significantly degraded. Finally, Brashears and Quintane (2015) conducted an experiment to determine how well individuals are able to recall their communication interaction patterns in surveys. Using ERGMs, the authors found that participants were able to identify clusters of ties, but were unable to identify individual links with any regularity. Collectively, these studies lend credence to the notion of the perception gap in networks across a variety of mediums.

Thus, while these reconstructed networks based on trace data correlate with the underlying perceived social networks – as reported by the participants – they are at best approximations. Given that these errors are present in empirical settings, we now review the literature on *why* these errors might arise.

Errors inherent to Online Networks

We argue that in an online environment, individuals' ability to infer the network of interactions is difficult for three reasons: scale, rate of change, and the potentially translucent nature of online networks. First, a key feature of networks generated from digital trace data is

their scale. Studies of networks in online communities often encompass thousands of messages among hundreds of actors (see for example Faraj and Johnson, 2011; Johnson et al., 2014) and can, conceivably in the future, be conducted on billions of messages among billions of actors. On social media platforms, actors are able to create a large number of ties, even though the number of connections may far exceed a person's capacity to manage them (Kane et al., 2014), causing a sense of overload (Mariotti and Delbridge, 2012). Indeed, natural limitations on peoples' time and cognitive capacity dictate that some of the thousands of links they form become forgotten or overlooked, causing these relationships to become "latent" (Mariotti and Delbridge, 2012) or "dormant" (Levin et al., 2011; Walter et al., 2015). The variability in the scale contributes to potential sources of errors in online networks.

A second feature of online network data that affects perception is the rate at which these networks evolve. Digital network data is often collected over a period of months or even years (e.g. Zaheer and Soda, 2009). Links represent the presence of interactions during some portion of the observed interval. Some interactions may be long and recur frequently during a time period, while others may be short, intense periods of interaction. Clickstream data – such as forum posts or edits to an online repository – tend to be "bursty," i.e., it is characterized by periods of high activity followed by lulls (Barabasi 2010; Vu et al., 2015). The variability in the rate of messaging contributes to potential sources of errors in online social networks.

Finally, online social networks vary greatly in the technological affordances that users can enact. One such affordance is the degree of visibility, or the amount of effort individuals must expend to assess the state of the network (Treem and Leonardi, 2013). Further, digital networks vary on a second technological affordance, association, or the ability to determine which individuals and or content are related (Treem and Leonardi, 2013). For instance, social

media networks such as Facebook allow users to see "who is friends with whom". However, while individuals can view this information, they may vary greatly in how they use it, or if they use it at all (Kane et al., 2014). This variability in individuals' perceptions contributes to potential sources of error in online social networks.

Errors of Cognition

In addition to the technological sources discussed, there are cognitive sources that contribute to discrepancies between observed networks and individuals' perceptions of these networks, whether the data are digital or not. Early studies on informant accuracy and recall focused on the ability of individuals to correctly report with whom they had interacted or what they had witnessed (Bernard et al. 1982; Freeman et al. 1987; Heald et al. 1998). In general, individuals had a difficult time recalling their own interactions, as well as observed interactions among others. Recollection can be biased by recency or regularity (Freeman 1992), or can be triggered by engagement in a specific foci (Corman and Scott 1994).

Further, individuals tend to view themselves as being more central, and that there are more ties, more reciprocation, and more transitivity among those they report as friends (Krackhardt and Kilduff 1999). Humans have also demonstrated a tendency for remembering clusters of relations, but perform poorly when asked to recall specific relationships (Brashears and Quintane 2015). A consistent finding among these studies is the tendency towards simplicity (Burt et al. 2013), and a rejection of structures that are dissonant with the mental models of the individual. Alternatively, a person's ability to accurately comprehend the structure around them may be impacted by their personality and affective state such as feelings of low power (Casciaro 1998; Casciaro et al. 1999, 2014), or even their gender (Brashears et al., 2016). Personality traits such as a need for closure (Flynn et al. 2010) can bias individuals toward making errors in their

perceptions of networks. Janicik and Larrick (2005) demonstrated that individuals who can effectively recall missing links are more accurate in comprehending an incomplete network structure and recognizing brokerage opportunities.

IMPACT OF ERRORS ON INFERENCE

In order to determine how statistical inference is affected by discrepancies in individual's perceptions versus observed networks, we consider what types of errors of commission and omission may be caused by the factors previously detailed (Yenigun, Ertan, and Siciliano, 2017). The first error occurs when individuals mistakenly perceive ties that do not exist, i.e., an *error of omission*. Alternatively, individuals may make the error of ignoring ties that do indeed exist, i.e., an *error of commission*. Both errors can cause biased inference results. We proceed to compare these issues to the notion of measurement errors in econometrics.

Perception Errors versus Measurement Errors

In econometric models, an inaccurate operationalization of a construct would result in a type of measurement error, or variance not accounted for by the statistics included in the model (Wooldridge 2009). There are two types of measurement error models: the classical error model and the non-classical error model. The classical error model – or *classical errors* – assumes the measurement error to be additive and independent of the true measure and the residuals in the second stage estimation, whereas the non-classical model considers the measurement error as non-additive or correlated with the true measure or residuals (Carroll et al. 2006). By this definition, errors in social networks are *non-classical* for three reasons[1]. First, the errors can be asymmetric, i.e., they are not evenly distributed around zero because of consistent over- or

[1] It is worth noting that these three characteristics of errors apply beyond social networks. Indeed, in any context where data represent individual opinions or beliefs, measurement errors will likely be non-classical.

under-estimation. Second, the magnitude of the errors may depend on the value of the statistic. Finally, the prevalence and magnitude of errors can differ significantly across individuals.

Measurement errors compromise regression models by attenuating or amplifying the influence of the corresponding coefficient (Yang et al. 2018). Further, measurement errors can create bias in both the coefficient of the erroneous measurement and the other coefficients (Greene 2003). To account for these misspecifications, correction techniques such as method-of-moments estimation (Carroll et al. 2006), instrumental variables (Carroll and Stefanski 1994), or simulation extrapolation (Yang et al. 2018) can be applied. However, there are certain limitations to these techniques that are relevant when analyzing social network data. Many existing correction techniques such as method-of-moments or simulation extrapolation require measurement error variance to be known *a priori* and assume that the error variance is constant across the sample. With social networks, the error variance is a product of latent human perceptions that can only be captured through sociometric surveys. Implementing such a survey is typically infeasible for networks of even moderate size and surveying a subset of the network may not be representative of the true variance.

Conducting inference on social network data faces a variety of challenges. Coefficients may be attenuated when the error variance is large (Wooldridge 2009). Alternatively, systematic perception biases could amplify coefficients in one direction or the other (Carroll and Stefanski 1994; Yang et al. 2018). Because the magnitude and direction of perception errors can be difficult to identify, existing correction techniques are ill-suited to fix these issues. Of course, the degree to which measurement errors affect social network models is not known, particularly for techniques such as exponential random graph modeling (Lusher et al. 2012) which operates at the aggregate network level.

Illustration: Krackhardt's Office Managers

To demonstrate the potential influence of perception errors, we analyze a classic network dataset taken from Krackhardt's (1987) study of CSS data collected from managers in a small manufacturing firm. Though they are not constructed through digital traces, these networks provide an empirical reality (the self-reported network) as well as a set of individually perceived networks for comparison. In these data, a sociometric survey was conducted for each of 21 employees, and each was asked to report their perceptions of the friendship amongst all 21 managers. For our purposes, we will use the symmetric, binary friendship networks. These networks are the CSSs for each manager and represent the raw observable data which we notate as $Y^{(i)}$, $i = 1, \ldots, 21$. Following Krackhardt (1987), we compute the locally aggregated structure (LAS) from these CSSs, which we notate as Y^{LAS}. We apply the intersection rule, whereby $Y_{ij}^{\mathrm{LAS}} = 1$ only if $Y_{ij}^{(i)} = 1$ and $Y_{ij}^{(j)} = 1$ (Batchelder et al. 1997). The LAS graph can be thought of as the "true" self-reported data for our analysis. Our aggregate network Y^{LAS} has 35 edges – a density of 0.1667 – and 12 total triangles. For each of the twenty-one managers we count the number of edges and the number of triangles reported, their accuracy compared to the "true" network by counting the number of false positives and false negatives, as well as computing the Jaccard index as a measure of similarity. The results are presented in Table 2.

Relative to the aggregated "true" self-reported network, most managers make errors perceiving the network. False positives range from 0 to 78 with a median of 22, while false negatives range from 18 to 62 with a median of 44. The Jaccard index, which assesses overall agreement between two networks, ranges from 0.10 to 0.45 with a mean of 0.27. To place these numbers in context, the *most accurate* managers in the sample were still wrong about the state of a tie more often than they were correct. Interestingly, there is a strong and negative correlation

between the errors ($\rho = -0.84$). This observation indicates that people tend to systematically over- or under-estimate the density of the overall network. These would correspond to systematically making more errors of commission or omission respectively. We also find that most actors *overestimate* the number of triangles.

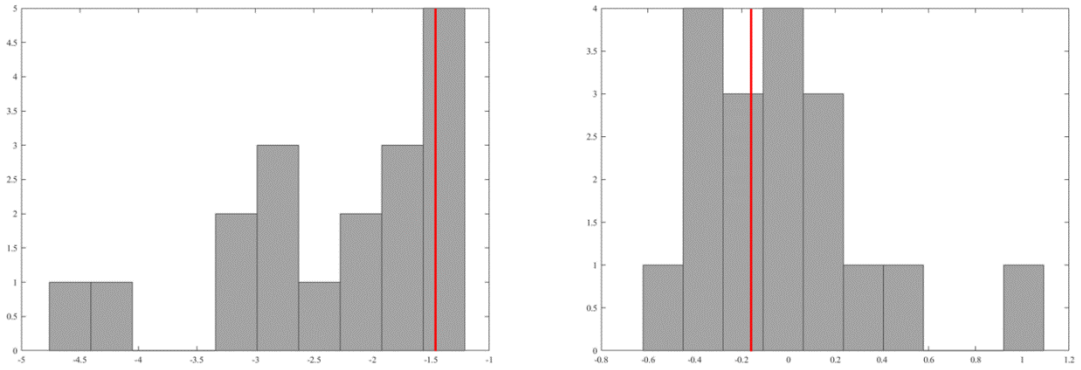
Table 2. Descriptive statistics and error rates for Krackhardt CSS data

Manager	*Edges*	*Triangles*	*False Positives*	*False Negatives*	*Jaccard Index*
1	36	22	40	38	0.29
2	14	1	6	48	0.29
3	5	1	2	62	0.11
4	25	10	22	42	0.30
5	49	48	46	18	0.45
6	20	7	16	46	0.28
7	61	80	78	26	0.30
8	4	0	0	62	0.11
9	4	0	0	62	0.11
10	28	12	32	46	0.24
11	49	39	56	28	0.33
12	18	4	10	44	0.33
13	25	2	24	44	0.28
14	38	27	32	26	0.43
15	20	13	22	52	0.20
16	20	4	18	48	0.25
17	30	13	28	38	0.33
18	14	1	8	50	0.26
19	51	45	52	20	0.41
20	8	0	8	62	0.10
21	31	21	24	32	0.40

We demonstrate the impact of the bias in the data by running a series of ERGMs on the CSS as well as the aggregate data using the approach described by Hunter et al. (2008). We include two statistics in our model, edges and directed dyadwise shared partners. This statistic accounts for the prevalence of "two-paths" in the network, i.e., a path from *A* to *B* is more likely if there is a path from *A* to *C* and *C* to *B*. We present the results in Figure 1. For the aggregate model, the coefficient for edges was -1.304 with a standard error of 0.096 ($p < 0.001$). The dyadwise shared partner statistic had a coefficient of -0.186 with a standard error of 0.069 ($p < 0.01$). For seventeen out of twenty-one managers, the aggregated model significantly overestimates the edge statistic. The average raw bias $\left(\hat{\beta}_{LAS} - \hat{\beta}_i\right)$ in the coefficient was 1.05, which is equivalent to an 80.9% overestimation of the individual values. The shared partner statistic had mixed results; for twelve managers, the aggregate model underestimated the

statistic, and in nine cases, the model overestimated the statistic. The overall raw bias for the coefficient was -0.176, which is equal to a 94.3% underestimation of the individual values.

Figure 1. Parameter estimates for Krackhardt CSS data



(a) (b)

Notes. Histograms of ERGM coefficients for (a) edge statistic and (b) shared partner statistic for 21 managers. Red line indicates aggregate estimate.

This illustration demonstrates that errors caused by individual differences in perception can lead to systematic biases in estimates of the model parameters. Further, the bias is not strictly an attenuation or amplification. As such, we are not able to determine a priori how well the results will reflect perceived reality. Thus, hypothesis testing using network data is subject to potentially significant errors.

THE ROBUST INFERENCE APPROACH

Given that network data are subject to errors of omission as well as errors of commission, it is not appropriate to simply correct for one problem or the other. Rather, network inference methods should be immunized to errors in a general sense, so that a preponderance of bias in either direction can be handled. We therefore advance that a robust approach is the most appropriate; robust optimization methods do not rely on a priori information or any distributional assumptions. Instead, a solution is found that is the best *in the worst case*, i.e., the discrepancies are as egregious as possible. Thus, the method we proceed to outline will be insular to both false positive *and* false negative errors, regardless of their source. We propose a conservative model

that uses the observable network data but allows the perceptions of individuals to vary randomly within a pre-specified range. The inferred parameter then holds for any variability within the range, i.e., the parameter is robust to cognitive errors.

Robust Optimization

Robust optimization broadly refers to the collection of techniques devised for finding optimal solutions to problems in the presence of uncertainties (for an overview, see Ben-Tal et al. 2009; Ben-Tal and Nemirovski 2002; Bertsimas et al. 2011). For network models, data uncertainty takes the form of individual perceptions of the surrounding network. Specifically, the model assumes that an empirical network accurately reflects each actor's individual network. However, perception errors, hidden information, or poor memory such as those discussed earlier under the categories of errors inherent to online networks and errors of cognition could lead individuals to make choices based on perceptions of the network to which the researcher is not privy. As a result, the estimated parameters from the model from observed data may not accurately reflect the impact of the perceived network structure on decision-making. The elements of a robust optimization problem include: nominal or original data (from the empirical network), an uncertainty set, and an objective function.

Notation and Definitions

Throughout the rest of this paper, we will denote vectors with a lowercase bold letter, and matrices with an uppercase bold letter. We consider an ordered series of M social network actions or events, which constitute the addition, removal, or alteration of a tie or node in the network. At each point in time $t = 1, \dots, M$, there are a set of these possible actions contained in the set $\mathcal{A}_t$. We assume that the set is finite with the cardinality $|\mathcal{A}_t| = N_t$. It is possible that there are varying numbers of available actions at any given time. Let $N = \max_t N_t$ be the largest

number of possible actions. The network at each point in time can be described by a collection of P characteristics, which we refer to as sufficient statistics of the graph. These statistics correspond to some structural element of the social network, i.e., degree of transitivity or number of edges, or some exogenous covariate. Given these dimensions, our network data may be represented as a matrix $\boldsymbol{X} \in \mathbb{R}^{M \times N \times P}$. This matrix represents a series of social network actions at M points in time, each of which can entail as many as N actions and the network at each point in time is represented by P characteristics or sufficient statistics. From this representation, it follows that $x_{typ} \in \mathbb{R}$ is a scalar corresponding to the value of statistic p of action y at time t. We may also define a slice of the matrix $\boldsymbol{x}_{ty\bullet} \in \mathbb{R}^P$ as a vector of all sufficient statistics for action y at time t.

Now, we assume that our network data is inaccurate due to a lack of coherence between the individual networks and the empirical network. Thus, while we observe $\boldsymbol{X}^{\text{net}}$, that may not be the true value perceived by the actor. We thus represent our data as

$$\boldsymbol{X}^{\text{ind}} = \boldsymbol{X}^{\text{net}} - \Delta\boldsymbol{X}\,, \tag{1}$$

where $\boldsymbol{X}^{\text{ind}}$ is the matrix of features perceived by the individuals in the network. We model networks and changes to them as reactions to the underlying *individual* structure, i.e., the network that individuals perceive. However, we consider here the case where we only have access to an *empirical* network which is collected through digital trace methods such as email or mobile data. Features derived from this network are represented by $\boldsymbol{X}^{\text{net}}$. A consequence of this is an inherent bias in the variables we use to conduct inference. This discrepancy is captured by $\Delta\boldsymbol{X} \in \mathbb{R}^{M \times N \times P}$. By modeling bias this way, we allow for the incorporation of internal, external, or data collection errors into our models without making any specific assumptions about their value, sign, or distribution.

We assume that, in general, we have no information about the nature of the errors and instead model them to reside in some bounded uncertainty set. In other words, an error may take on any value within this set. We describe this *uncertainty set* as

$$\mathcal{N} = \{ \Delta\mathbf{X} \mid \|\Delta \boldsymbol{x}_{tz\bullet}\| \leq \rho, z \in \mathcal{A}_t, t = 1, \dots, M \}. \quad (2)$$

Here and throughout the remainder of the paper, we will use the Euclidian norm. An uncertainty set $\mathcal{N}$ is a collection of all possible errors that meet a basic criterion. In our case, we restrict the errors for each vector of statistics to be limited in magnitude by a tolerance parameter ρ. This value corresponds to a level of discrepancy between our collected data and the true perceived information. A larger ρ will allow for greater deviance from the observed values, while a $\rho = 0$ will imply equality of observable and true data. Essentially, an uncertainty set is the collection of all possible differences between the true and observable data. While we refer to a single parameter ρ for the sake of simplicity, it is possible to have a large number of these values. For instance, each sufficient statistic may be constrained by a unique parameter, or each individual may have an independent error tolerance.

Our definition of the uncertainty set raises two questions: first, how does one interpret the value of ρ? Second, how does one select an appropriate ρ? To answer these questions, we adopt a probabilistic view of $\mathcal{U}$. We construct the set such that it contains all errors we believe may occur with a non-negligible probability. Put another way, errors of magnitude greater than ρ are rare enough that we do not need to immunize our estimates from them. Thus, the value of ρ should represent the threshold at which errors are no longer likely. To actually select a ρ, the underlying distribution of the data should be considered. Suppose that we knew the errors followed a standard normal distribution and the uncertainty set is modeled with a Euclidean norm. Then, for an m-dimensional error vector, the tolerance parameter is given as $\rho = \chi^2_m(\alpha)$,

i.e., the Chi-squared test statistic at confidence level α (Bertsimas et al. 2007). It follows that for a single variable modeled with error, $\rho = 1,2,3$ corresponds to errors falling within one, two, or three standard deviations, respectively. Of course, in practice we do not always know the distributions of the errors. In those cases, ρ can be estimated from the empirical data using a measure of spread, such as a standard deviation or quantile.

Computing Robust Estimators

To compute a robust maximum likelihood estimate, we first need to specify a probability density function f for the data. Generally speaking, the robust methodology we present will hold for any differentiable density function f. Traditionally, the probability density function for a social network has been considered part of the exponential family of probability distributions (Holland and Leinhardt 1977, 1981). The exponential family broadly describes a variety of distributions, including the normal, gamma, Weibull, and multinomial. Common social network inference models such as ERGMs (Robins et al., 2007), stochastic actor-oriented models (Snijders, 1996), and relational event models (Butts 2008) all employ an exponential probability distribution from this family. Other models including the Cox proportional hazards model (Cox 1972) and the conditional logit model (McFadden 1974) use the same specification. However, it is important to note that the method we are proposing is valid for *any choice of probability density.* Bertsimas and Nohadani (2019) have focused on the multivariate normal distribution, and in this paper, we extend the robust MLEs to the broader exponential family, which is central to social network analysis.

Social network models typically assume an exponential probability distribution parameterized by the vectore $\boldsymbol{\beta} \in \mathbb{R}^P$. Though these models typically use linear combinations of parameters and statistics, any differentiable function is valid for our method. Each element of $\boldsymbol{\beta}$

is interpreted as an intensity parameter which contributes to the likelihood of the observed network. The probability density for a single observation y at time t is thus:

$$f_t\left(\boldsymbol{\beta}; \boldsymbol{x}_{ty\bullet}^{\text{ind}}, \mathcal{A}_t\right) = \frac{\exp\left(\boldsymbol{\beta}' \boldsymbol{x}_{ty\bullet}^{\text{ind}}\right)}{\sum_{z \in \mathcal{A}_t} \exp\left(\boldsymbol{\beta}' \boldsymbol{x}_{tz\bullet}^{\text{ind}}\right)}. \tag{3}$$

Given that we want to make inferences on the values of $\boldsymbol{\beta}$ corresponding to the true network structure, the typical methodology would be to perform maximum likelihood estimation (MLE), i.e., finding the distribution parameters that maximize the probability density function and hence best fit the data. Inference based on the MLE procedure or derivations of it is a common method for social network analysis (Butts 2008; Snijders et al. 2010; Stadtfeld 2012).

In the remainder of this section, we generalize to multiple time-slices for longitudinal data, though our method holds true for a single instance such as a traditional ERGM analysis. The likelihood of a sequence of network alterations or events is equivalent to a product of (3) across all observations with the features $\boldsymbol{X}^{\text{net}}$ reflecting the changing network. We assume the presence of some error in our dataset, i.e., $\boldsymbol{X}^{\text{net}} \neq \boldsymbol{X}^{\text{ind}}$, and so we recast the likelihood of the full sequence of network events as:

$$\begin{aligned} \prod_{t=1}^{M} f_t\left(\boldsymbol{\beta}; \boldsymbol{x}_{ty\bullet}^{\text{ind}}, \mathcal{A}_t\right) &= \prod_{t=1}^{M} f_t\left(\boldsymbol{\beta}; \boldsymbol{x}_{ty\bullet}^{\text{net}} - \Delta \boldsymbol{x}_{ty\bullet}, \mathcal{A}_t\right) \\ &= \prod_{t=1}^{M} \frac{\exp\left(\boldsymbol{\beta}'\left(\boldsymbol{x}_{ty\bullet}^{\text{net}} - \Delta \boldsymbol{x}_{ty\bullet}\right)\right)}{\sum_{z \in \mathcal{A}_t} \exp\left(\boldsymbol{\beta}'\left(\boldsymbol{x}_{tz\bullet}^{\text{net}} - \Delta \boldsymbol{x}_{tz\bullet}\right)\right)}. \end{aligned} \tag{4}$$

As discussed by Bertsimas and Nohadani (2019), maximizing the likelihood yields the same solution as maximizing the log-likelihood function, which is computationally more manageable. We define the log-likelihood function $\psi(\boldsymbol{\beta}; \mathbf{X}^{\text{net}} - \Delta\mathbf{X}) = \log\left(\prod_{t=1}^{M} f_t\left(\boldsymbol{\beta}; \boldsymbol{x}_{ty\bullet}^{\text{net}} - \Delta \boldsymbol{x}_{ty\bullet}, \mathcal{A}_t\right)\right)$. Now, the maximum likelihood problem becomes the following constrained optimization problem:

$$\max_{\boldsymbol{\beta}} \{\psi(\boldsymbol{\beta}; \mathbf{X}^{\text{net}} - \Delta\mathbf{X}) : \Delta\mathbf{X} \in \mathcal{N}\}. \quad (5)$$

It follows that the solution to the MLE problem (5) must be a valid solution for any of the errors that may reside within the uncertainty set $\mathcal{N}$. Consequently, a solution that satisfies this constraint must also fit the log-likelihood function under the worst-case errors. Hence, the robust estimator is also the solution to the following *robust optimization problem*:

$$\max_{\boldsymbol{\beta}} \min_{\Delta\mathbf{X}\in\mathcal{N}} \psi(\boldsymbol{\beta}; \mathbf{X}^{\text{net}} - \Delta\mathbf{X}). \quad (6)$$

We solve the inner optimization problem (given in Equation 6) by decomposing it into an outer problem – maximization over parameters $\boldsymbol{\beta}$ – and inner problem – minimization over errors $\Delta\mathbf{X}$. We focus first on the inner problem, which finds the set of feasible errors that minimize the log-likelihood function. The inner problem is equal to:

$$\begin{aligned}\phi(\boldsymbol{\beta}; \mathbf{X}^{\text{net}}) &= \min_{\Delta\mathbf{X}\in\mathcal{N}} \psi(\boldsymbol{\beta}; \mathbf{X}^{\text{net}} - \Delta\mathbf{X}) \\ &= \min_{\Delta\mathbf{X}\in\mathcal{N}} \log\left(\prod_{t=1}^{M} f_t\left(\boldsymbol{\beta}; \boldsymbol{x}_{ty\bullet}^{\text{net}} - \Delta\boldsymbol{x}_{ty\bullet}, \mathcal{A}_t\right)\right) \\ &= \min_{\|\Delta\boldsymbol{x}_{tz\bullet}\|\le\rho} \sum_{t=1}^{M} \left[\log f_t\left(\boldsymbol{\beta}; \boldsymbol{x}_{ty\bullet}^{\text{net}} - \Delta\boldsymbol{x}_{ty\bullet}, \mathcal{A}_t\right)\right]. \end{aligned} \quad (7)$$

In (7) we are taking the sum of the logarithm of a probability density function, which guarantees that we are taking a sum of strictly non-positive numbers. Consequently, this problem is separable across time points $t = 1, \dots, M$. Applying our known density function, the inner problem reduces to solving

$$\min_{\|\Delta\boldsymbol{x}_{zt}\|\le\rho} \left(\boldsymbol{\beta}'\left(\boldsymbol{x}_{ty\bullet}^{\text{net}} - \Delta\boldsymbol{x}_{ty\bullet}\right) - \log \sum_{z\in\mathcal{A}_t} \exp\left(\boldsymbol{\beta}'\left(\boldsymbol{x}_{tz\bullet}^{\text{net}} - \Delta\boldsymbol{x}_{tz\bullet}\right)\right)\right) \quad (8)$$

for each instance t. We note that the objective function for each component of the inner optimization problem is decreasing in $\boldsymbol{\beta}'\Delta\boldsymbol{x}_{ty\bullet}$ and increasing in $\boldsymbol{\beta}'\Delta\boldsymbol{x}_{tz\bullet}$ for all $z \neq y$; for proof, see the Appendix. As a result, the optimal value can be found by determining the maximum

feasible value of $\boldsymbol{\beta}'\Delta\boldsymbol{x}_{ty\bullet}$ and the minimum value of $\boldsymbol{\beta}'\Delta\boldsymbol{x}_{tz\bullet}$ for all $z \neq y$. By applying Hölder's inequality, we know that

$$-\|\boldsymbol{\beta}\|\|\Delta\boldsymbol{x}_{zt}\| \leq \boldsymbol{\beta}'\Delta\boldsymbol{x}_{zt} \leq \|\boldsymbol{\beta}\|\|\Delta\boldsymbol{x}_{zt}\|. \quad (9)$$

Applying the extreme limits and the known bounds of our uncertainty set, we can solve the inner problem for each event as:

$$\begin{aligned}\min_{\|\Delta\boldsymbol{x}_{zt}\|\leq\rho}\left(\boldsymbol{\beta}'\left(\boldsymbol{x}_{ty\bullet}^{\text{net}} - \Delta\boldsymbol{x}_{ty\bullet}\right) - \log\sum_{z\in\mathcal{A}_t}\exp\left(\boldsymbol{\beta}'\left(\boldsymbol{x}_{tz\bullet}^{\text{net}} - \Delta\boldsymbol{x}_{tz\bullet}\right)\right)\right) \\ = \boldsymbol{\beta}'\boldsymbol{x}_{ty\bullet}^{\text{net}} - \|\boldsymbol{\beta}\|\rho - \log\sum_{z\in\mathcal{A}_t}\exp(\boldsymbol{\beta}'\boldsymbol{x}_{tz\bullet}^{\text{net}} + (-1)^{u_z}\|\boldsymbol{\beta}\|\rho).\end{aligned} \quad (10)$$

Here we define u_z as an indicator variable that is equal to 1 if $z = y_t$ and 0 otherwise. The specific solution for $\Delta\boldsymbol{x}_{tz\bullet}$ is given by $\Delta\boldsymbol{x}_{tz\bullet}^{\star}(\boldsymbol{\beta}) = (-1)^{1-u_z}\rho\frac{\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|} \times \mathbf{1}\{\boldsymbol{\beta} \neq \mathbf{0}\}$. Thus, we combine all of our elements into a single solution for the inner problem:

$$\phi(\boldsymbol{\beta}; \boldsymbol{X}^{obs}) = \sum_{t=1}^{M}\left(\boldsymbol{\beta}'\boldsymbol{x}_{ty\bullet}^{\text{net}} - \|\boldsymbol{\beta}\|\rho - \log\sum_{z\in\mathcal{A}_t}\exp(\boldsymbol{\beta}'\boldsymbol{x}_{tz\bullet}^{\text{net}} + (-1)^{u_z}\|\boldsymbol{\beta}\|\rho)\right). \quad (11)$$

We can now solve the robust outer MLE problem, $\max_{\boldsymbol{\beta}}\phi(\boldsymbol{\beta}; \boldsymbol{X}^{\text{net}})$:

$$\max_{\boldsymbol{\beta}}\sum_{t=1}^{M}\left(\boldsymbol{\beta}'\boldsymbol{x}_{ty\bullet}^{\text{net}} - \|\boldsymbol{\beta}\|\rho - \log\sum_{z\in\mathcal{A}_t}\exp(\boldsymbol{\beta}'\boldsymbol{x}_{tz\bullet}^{\text{net}} + (-1)^{u_z}\|\boldsymbol{\beta}\|\rho)\right) \quad (12)$$

directly using a subgradient method. The subgradient is required because the gradient does not exist at the point $\boldsymbol{\beta} = \mathbf{0}$. For a derivation of the gradient, see the Appendix.

Computing Standard Errors

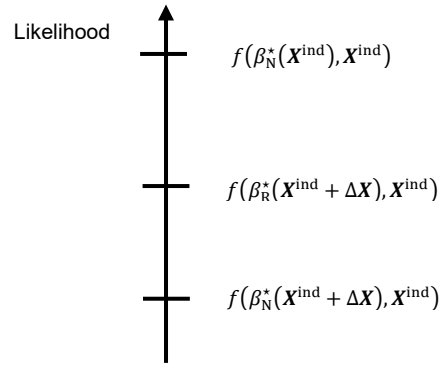
In order to conduct hypothesis testing with the robust estimators, we need to calculate the standard errors of the estimates. Because we are conducting maximum likelihood estimation, we can approximate standard errors using the Fisher information matrix. The Fisher information $\mathcal{I}$ is defined as $\mathcal{I}(\beta) = \mathbb{E}\left[-\frac{\partial^2}{\partial\beta^2}\log f(X \mid \beta)\right]$. This value is equal to the negative expected value of

the Hessian matrix for the likelihood function f, i.e., matrix of second derivatives, evaluated at β. Then, the variance of the optimal estimator, $\beta_p^\star$, is defined as the pth diagonal element of the inverse information matrix: $var(\beta_p^\star) = [\mathcal{I}(\beta^\star)]_{p,p}^{-1}$. By plugging in the likelihood function from equation (5) and the solution $\boldsymbol{\beta}_{\mathrm{R}}^\star$, we can obtain estimates for the variance and subsequently the standard errors. The exact expression for the Hessian is given in the Appendix. Finally, the test statistic $z_\rho = \beta_R^\star(\rho)/SE\big(\beta_R^\star(\rho)\big)$ is asymptotically normal with mean zero and standard deviation one. Using this fact, we can conduct hypothesis testing with the robust estimator.

Interpretation of Robust Estimators

The goal of robust inference is to estimate a set of parameters from the empirical network data that will still provide a reasonable estimate of data that differs from what we used to tune the model. In other words, if we assume that our empirical network does not in fact match the individual networks, the nominal estimators could change significantly, whereas the robust estimators will remain stable, both in terms of likelihood and in terms of bias. We present Figure 2 to illustrate how robust estimators compare to nominal estimators.

Figure 2. Impact of Robust Parameters on Likelihood



Notes. The likelihood function of the data as perceived by the individuals in the network as a function of the nominal parameters ($\boldsymbol{\beta}_{\mathrm{N}}^\star$) fit to the individual network data, parameters ($\boldsymbol{\beta}_{\mathrm{R}}^\star$) fit to empirical network data, and nominal parameters ($\boldsymbol{\beta}_{\mathrm{N}}^\star$) fit to the empirical network data.

To show the benefits of the robust estimators $\boldsymbol{\beta}_{\mathrm{R}}^\star$, we compare its likelihood $f(\boldsymbol{\beta}_{\mathrm{R}}^\star)$ to that of standard estimators $f(\boldsymbol{\beta}_{\mathrm{N}}^\star)$, which assumes data follows a distribution described by the nominal

estimators. As input, we use true data $\boldsymbol{X}^{\text{ind}}$ and observable data $\boldsymbol{X}^{\text{net}}$, as discussed earlier. In general, it is expected that $f\left(\boldsymbol{\beta}_{\text{N}}^{\star}(\boldsymbol{X}^{\text{net}})\right) < f\left(\boldsymbol{\beta}_{\text{N}}^{\star}\left(\boldsymbol{X}^{\text{ind}}\right)\right)$ because the assumptions are no longer met for $f\left(\boldsymbol{\beta}_{\text{N}}^{\star}(\boldsymbol{X}^{\text{net}})\right)$. However, robust estimators are immune to assumption deviations, hence we observe $f\left(\boldsymbol{\beta}_{\text{N}}^{\star}(\boldsymbol{X}^{\text{net}})\right) < f\left(\boldsymbol{\beta}_{\text{R}}^{\star}(\boldsymbol{X}^{\text{net}})\right) < f\left(\boldsymbol{\beta}_{\text{N}}^{\star}\left(\boldsymbol{X}^{\text{ind}}\right)\right)$, i.e., the robust estimators outperform nominal estimators for observed data and cannot reach the optimality of $\boldsymbol{\beta}_{\text{N}}^{\star}\left(\boldsymbol{X}^{\text{ind}}\right)$ due to lack of accurate information. In summary, robust parameters are expected to better predict behaviors in a network than nominal parameters if the individual networks and empirical network differ, but will always perform worse than the hypothetical optimum. This phenomenon is referred to as "the price of robustness" (Bertsimas and Sim 2004).

MODEL DEMONSTRATION & TESTING

Study 1: Simulation Experiments

Data Generation & Method

We first perform experiments on a set of computer-generated data. We generate a set of simulated sequences of networks that replicate common behaviors. By creating synthetic data, we can control ground-truth values for characteristics of the network at any point in the sequence. Because our method incorporates errors by design, we will also randomly generate perturbations in the values of the statistics. For each simulated dataset, we fit a nominal network model and determine the parameters. We then fit a set of robust estimators for varying levels of ρ. Finally, we test the performance of the robust estimators against the nominal estimators on the contaminated data. We generate 100 sequences of network events, where each event is a link (i, j) between two nodes in the network. Each sequence is composed of 5,000 events between 50 actors. We randomly assign the fifty individuals to one of two groups for purposes of determining homophily. To generate a random sequence, we specify four mechanisms which

drive the occurrence of a link: activity rate, reciprocity, homophily, and transitivity. The process for creating a sequence is as follows:

0. Initialize with a sequence $E = \{(i_1, j_1)\}$ where (i_1, j_1) is chosen randomly
1. For $t = 2, \ldots, 5000$ do:
    a. For all $(i, j) \in D$, compute $p_{ij}(t) = \lambda_{ij}(t) / \sum_{(u,v)\in D} \lambda_{uv}(t)$ where D is the set of all possible links
    b. Draw an event (i_t, j_t) from the multinomial probability distribution $p(t)$
    c. Add the event to the sequence: $E = \{E, (i_t, j_t)\}$
2. Return sequence E

In order to carry out these steps, we need to define the rate for each dyad at a given step t. Following our prior definition, $\log\left(\lambda_{ij}(t)\right) = \beta_1 x_{ij}^A(t) + \beta_2 x_{ij}^R(t) + \beta_3 x_{ij}^H(t) + \beta_4 x_{ij}^T(t)$. For simplicity, we set all parameters to one, $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 1$. In Table 3, we provide descriptions of the four statistics.

Table 3. Statistics for Generating Sequences

Variable	*Formula*	*Interpretation*
Activity	$x_{ij}^A(t) = \sum_k n_{ikt}$	As i sends more messages, i is more likely to send a new message.
Reciprocity	$x_{ij}^R(t) = n_{jit}$	As j increasingly sends i messages, i becomes more likely to send a message to j.
Homophily	$x_{ij}^H(t) = y_{ij}$	i is more likely to send a message to j and not k if they are members if the same group and k is not.
Transitivity	$x_{ij}^T(t) = \sum_k \sqrt{n_{ikt} n_{kjt}}$	i is more likely to send a message to j if there are frequent messages from i to other actors k and from those actors to target j.

The count of times an event on dyad (i, j) has occurred up to, but not including, step t is n_{ijt}, and y_{ij} is a dummy variable taking a value of one if i and j share a group.

We assume that an actor is cognizant of their own rate of communication, who they received links from, and who is in the same group. However, an individual who uses network

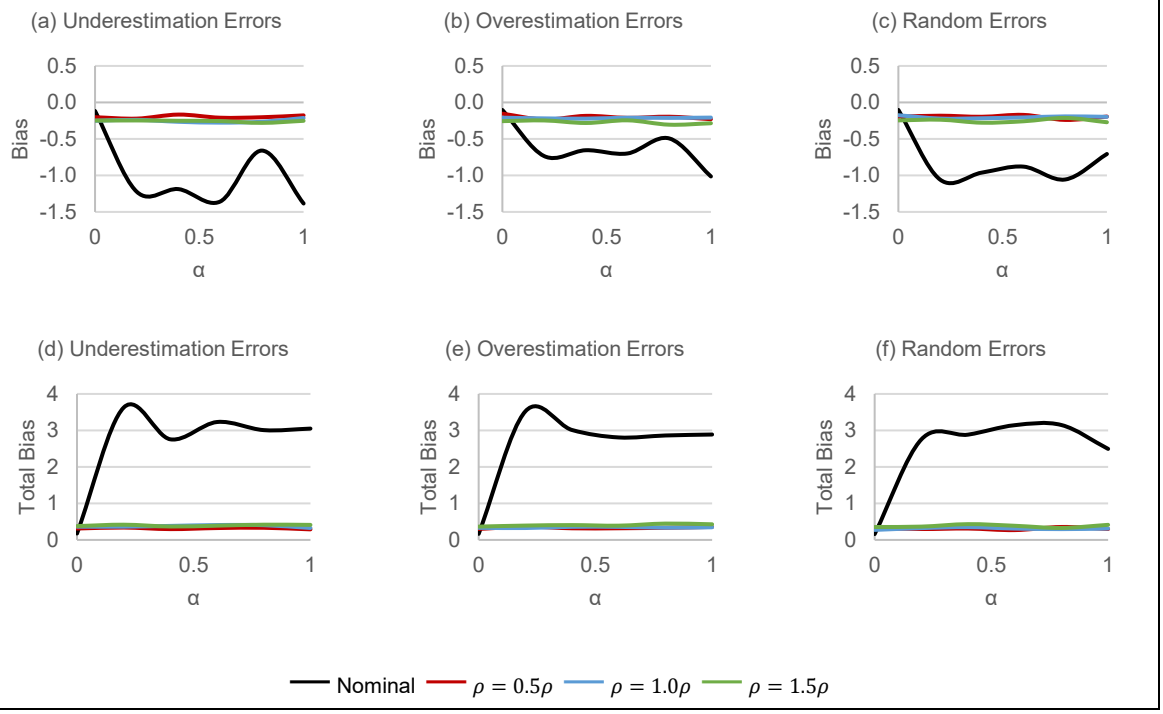
transitivity as a decision criterion will tend to create links with second-degree connections, or "friends of friends." Due to incomplete network awareness, we assume the network statistic transitivity will be subject to error. Essentially, we assume that there is some value, $z_{ij}^{T}(t) = x_{ij}^{T}(t) + \varepsilon_{ijt}$, that is equal to the observable transitivity statistic, plus some perception error. We specify the error term in three ways. First, we consider the case where an individual *overestimates* the strength of third-party connections to the target. Then, we draw the errors from a Uniform distribution: $\varepsilon_{ijt} \sim U\left(0, \alpha x_{ij}^{T}(t)\right)$. Here, α is a parameter we specify that dictates the extent of possible errors; for instance, setting $\alpha = 1$ means an individual can perceive the value of transitivity as up to *twice* as larger as it actually is. Second, we consider the case where an individual *underestimates* the strength of third-party connections. Then, we draw the errors from a Uniform distribution: $\varepsilon_{ijt} \sim U\left(-\alpha x_{ij}^{T}(t), 0\right)$. Finally, we consider the case of *random* perception errors, and draw these from a Uniform distribution: $\varepsilon_{ijt} \sim U\left(-\alpha x_{ij}^{T}(t), \alpha x_{ij}^{T}(t)\right)$. To account for a range of error magnitudes, we vary α from 0 (i.e., perception matches observable reality) to 1.

We calculate the robust estimators using observed data and using a tolerance parameter of $\rho = 0$, $\rho = 0.5 * \sigma$, $\rho = 1.0 * \sigma$, and $\rho = 1.5 * \sigma$, where σ is the observed standard error of the transitivity statistic. These robust estimators are denoted $\boldsymbol{\beta}^{\star}(\rho)$. Note that when $\rho = 0$, we have the nominal parameters of the standard MLE approach. After computing the robust estimators, we can compare them to the true values for the parameters. Specifically, there are two values of interest: 1) the bias in the transitivity parameter, ${\beta^{\star}}_{4}(\rho) - \beta_4$, and 2) the overall bias in the parameter vectors $\|\boldsymbol{\beta}^{\star}(\rho) - \boldsymbol{\beta}\|$. If the robust method we are proposing is superior, we would expect the bias to be closer to zero for $\rho > 0$ compared to $\rho = 0$ when some error is present, i.e., $\alpha > 0$.

Results

We first examine the level of bias in our estimation of the transitivity parameter at different error levels. The true value of the parameter is $\beta_4 = 1$; when computing bias, values less than one indicate underestimation and vice versa. A value of zero for bias indicates a consistent estimator. In Figure 3, we illustrate the average bias across all simulations for underestimation errors, overestimation errors, and random errors. We first note the nominal case with $\rho = 0$, i.e., the standard social network method. When any error is introduced ($\alpha > 0$), the estimated parameter quickly approaches zero, indicated by a bias of approximately negative one. This effect is an example of the attenuation bias in the measurement error literature. We also observe that when individuals *underestimate* transitivity, the nominal model tends to return a negative value for transitivity, i.e., bias of less than negative one.

Figure 3. Bias in transitivity parameter and parameter vector



Conversely, when individuals *overestimate* transitivity, the nominal model tends to yield a somewhat lesser bias, indicating that the estimated parameter is small and positive. For random

errors, the nominal model yields an estimated parameter of approximately zero. What these results reveal is that even at small magnitudes, non-classical errors do not necessarily lead to an estimate of zero. In fact, the direction of the bias in the estimator is consistent with the direction of the errors. Thus, in larger datasets these effects could be magnified, leading to an amplification of the parameter estimates (e.g., Yang et al. 2018).

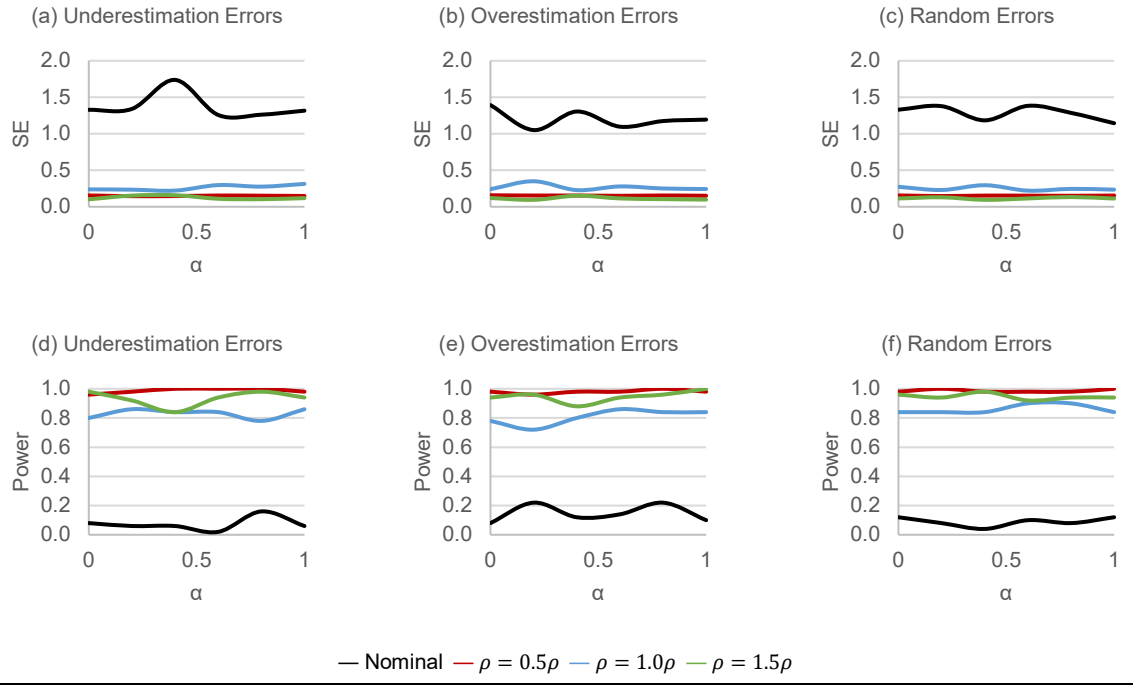
Turning to the robust models, we find significantly less bias in the estimated parameters when there is error. Across all three levels of robustness ($0.5\sigma, 1.0\sigma, 1.5\sigma$), the model produces relatively stable estimates, regardless of the magnitude of errors. We do observe that at lower levels of robustness, there is understandably less bias (i.e., estimate is closer to one), but the effect is small. Further, the value of the robust estimator is similar regardless of whether the errors represent underestimation, overestimation, or randomness. This finding is a feature of the robust approach; regardless of the direction of the errors, the model will return similar estimates of the parameters.

We further explore the effect of robustness and error on the overall bias in the estimated parameters. We compute the norm of the difference between the real parameters ($\beta_1 = \beta_2 = \beta_3 = \beta_4 = 1$) and the estimated parameters. Smaller values indicate that the estimate, $\boldsymbol{\beta}^{\star}(\rho)$, is closer to the real value. Again, we vary the magnitude of error α for underestimation, overestimation, and random errors. We find that when we introduce perception errors ($\alpha > 0$), the nominal model exhibits significantly greater bias compared to the robust model. This observation holds across error types and error magnitudes. Essentially, introducing error to the transitivity term not only impacts our estimation of the transitivity parameter, but also our estimation of each of the other terms as well. By contrast, the robust model consistently yields a small total bias indicating minimal deviation across estimators. In sum, these results indicate that

not only does the robust model provide a less biased estimator of the focal parameter, it provides a less biased estimator of *the entire parameter vector* when errors are present.

Next, we consider the standard errors of the proposed robust coefficients, as well as the statistical power of the estimates. The results are presented in Figure 4. In Figures 4a,b,c we observe that the robust estimators have much smaller standard errors on average relative to the uncorrected model. This holds true across error magnitudes and robustness level. Turning to Figures 4d,e,f we examine the power of the estimated coefficients. We compute statistical power by taking the percentage of the time the network model correctly identified the underlying effect at the 0.05 confidence level. Larger values indicate better power, i.e., greater ability to detect the effect.

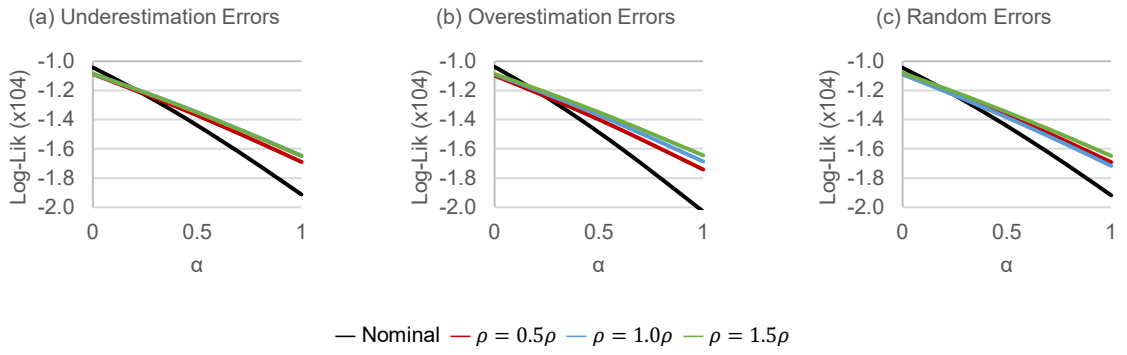
Figure 4. Standard errors and statistical power of estimates



We find that the robust model has significantly more power to detect transitivity, even in the presence of small deviations. This finding has clear implications for statistical inference: when errors are present, a robust model is more likely to correctly reject the null hypothesis. Finally, we compare the model fit of the different estimators. In this step, we fit a standard social network

model to data with no errors ($\alpha = 0$). We also fit our robust models to the same datasets. Then, we apply our estimated parameters to different sequences with progressively greater perception errors and calculate the likelihood of the sequences. By computing the likelihood using parameter estimates from an error-free model, we are testing how well the different models could predict events *out of sample*. We present our results in Figure 5.

Figure 5. Log-likelihood of nominal and robust estimators



We observe that the robust models outperform the nominal models in all cases where error is present. The difference between the nominal and robust fit grows as more error is introduced into the model, i.e., α increases. Further, while all models fare worse when more error is added, the model with the most robustness ($\rho = 1.5\sigma$) degrades slowest. In summary, if we fit a robust social network model to a dataset, then it will have greater predictive power when applied to observations with more error.

Study 2: Laboratory Example

Dataset & Analyses

As a further illustration of the robust network model, we apply our methodology to data collected from experiments on team decision-making. Our sample is composed of twenty, twenty-person multiteam systems (MTS, or teams of teams), (N = 400) engaged in a military-style strategic coordination task, collected as part of a larger research project. Each twenty-

person MTS comprised four 5-person teams. Participants were recruited at a Midwestern US university and participated in this study in exchange for either research credit or $35. Participants reported to a laboratory in groups of 20 and were randomly assigned into four teams. Each MTS session was divided into three phases: training, practice mission, and performance mission. During the training phase, the entire 20-person group was trained together, in a large room where they all watched a video explaining the enemy occupation and the nature of their mission. When the video concluded, participants completed a brief survey including demographic and prior familiarity items. The participants then performed a 15-minute practice mission during which they familiarized themselves with the game functionality, the communication channels, and their role responsibilities. After the practice mission, the main observation period began. The overarching goal of the group was to guide a convoy through a region comprised of four equally-sized sub-regions. Each team worked in a distinct sub-region to clear obstacles in the convoy's path. The conditions of the sub-region was a local team goal not shared by the other teams.

We collected data in a variety of formats. Surveys were given at the beginning and end of the session. At the beginning, participants answered questions regarding demographics, personality traits, and experience with video games. At the end, participants were asked a variety of team process questions. We also asked each individual to answer the following question: "With whom did you communicate during the mission?" Participants then checked a box for each of the other members of the group with whom they recall communicating. Their responses thus constitute an *individual network*, i.e., they represent the links that the members of the population believe exist. In addition to the survey measures, we collected digital trace data in the form of communication transcripts. Participants communicated with one another solely through Skype and could choose chat or audio messages. All audio messages were transcribed using

audio and video files to ensure accuracy. Chat messages were downloaded from our game server. We then combined these two data streams to form a complete timestamped transcript, with each observation formatted as <Timestamp, Sender, Receiver, Message, Channel>.

For each of the twenty sessions, we created two networks, which we denote z_k and y_k. The network z_k represents the individual network for session k. For every pair of individuals (i, j), then a link is present, i.e., $z_{ijk} = 1$, if individual i reports that they communicated with j. Otherwise, each link has zero value. The network y_k represents the *empirical network* for session k. Here, we create a dichotomous network by setting $y_{ijk} = 1$ if there were any messages sent from i to j during the session, and zero otherwise. Thus, y_k represents the typical "digital trace network" that can readily be collected through technology (i.e. Skype). To determine whether the digital trace operationalization leads to biased inference, we estimated ERGMs on both networks. We included four common network statistics: edges, mutual ties, in-degree centralization, and out-degree centralization. The edge statistic represents the network density, while the mutual statistic represents the frequency with which ties are reciprocated. In- and out-degree centralization represent the variance in the degree distribution, and can be loose proxies for power law tendencies. The parameter estimates from the ERGMs are notated $\boldsymbol{\beta}^{\star z}$ and $\boldsymbol{\beta}^{\star y}$.

Findings

Descriptive statistics for the twenty networks are presented in Table 4. We calculated the parameter estimates $\boldsymbol{\beta}^{\star z}$ and $\boldsymbol{\beta}^{\star y}$ for every session, as well as robust estimates $\boldsymbol{\beta}^{\star}(\rho)$ for different levels of robustness. We set three levels of ρ based on the standard deviation of the underlying statistic. For instance, the in-degree centralization statistic had a standard deviation of approximately 0.20 across the twenty session; we set $\sigma = 0.2$ for that statistic.

Table 4. Descriptive statistics for networks		
Variable	*Mean*	*SD*
Survey Density	0.416	0.066
Survey In-Centralization	6.735	2.914
Survey Out-Centralization	15.324	7.981
Number of Messages	948	130
Observed Density	0.192	0.009
Observed In-Centralization	4.554	2.098
Observed Out-Centralization	5.496	2.713

After fitting the models, we computed the bias in the individual parameters, as well as the parameter vector as a whole. For the parameters, we calculated a standardized bias, then squared it to compare absolute magnitudes (i.e., the mean square error). The formula for the bias in parameter p at robustness level ρ is: $bias(\beta_p)_\rho^2 = \left(\frac{\left(\hat{\beta}_p^z - \hat{\beta}_p^y(\rho)\right)}{\hat{\beta}_p^z}\right)^2$. This value is comparable across sessions as well as across statistics. Larger numbers indicate greater bias, while a value of zero would be a consistent estimate. We also computed total bias by taking the norm difference between the parameter vectors, $\|\boldsymbol{\beta}^{\star z} - \boldsymbol{\beta}^{\star y}(\rho)\| / \|\boldsymbol{\beta}^{\star z}\|$. We report the results in Table 5.

Table 5. Model bias for ERGM parameters				
	Squared Parameter Bias (Standardized)			
Variable	$\rho = 0$	$\rho = 0.5\sigma$	$\rho = 1.0\sigma$	$\rho = 1.5\sigma$
Edges	0.144	0.173	***0.096***	0.111
Mutual	4.427	3.736	***3.036***	3.780
In-Centralization	0.667	***0.456***	0.789	1.228
Out-Centralization	0.309	0.330	0.202	***0.160***
Standardized Norm Bias	0.464	0.457	***0.434***	0.494

Overall, our findings indicate that the robust model is significantly less biased with respect to the individual network as compared to the nominal digital trace empirical network. As indicated by the italics in Table 5, a non-zero level of robustness consistently resulted in the lowest bias for the individual statistics and for the parameter vector as a whole. Using a tolerance of one standard deviation overall yielded the best performance, with all but one parameter being less biased than the nominal model.

In addition to analyzing bias, we also examined the error rates of each model with respect to statistical inference. We coded a model as giving a *false positive* if it yielded a significant estimate of a coefficient, while the coefficient for the survey network was non-significant.

Similarly, a *false negative* occurs when the model determines a coefficient is non-significant, while in the survey network that coefficient was significant. We report the error rates, as well as the total error rate, in Table 6.

Table 6. Error rates for standard and robust models

Error Type	$\rho = 0$	$\rho = 0.5\sigma$	$\rho = 1.0\sigma$	$\rho = 1.5\sigma$
False Positive (%)	8.75%	10.00%	10.00%	10.00%
False Negative (%)	10.00%	2.50%	3.75%	6.25%
Total (%)	18.75%	12.50%	13.75%	16.25%

We find that when conducting hypotheses tests, models with robustness ($\rho > 0$) make fewer inference errors than a nominal model ($\rho = 0$). In particular, we find that the driving factor is the false negative rate. On average, running a model on the network from digital trace data yielded a false negative rate of 10%, while the robust model yielded a false negative rate of at most 6.25%. This finding indicates that analyzing a network using raw digital trace data will cause researchers to miss important effects about one in ten times. However, accounting for errors through robustness significantly reduces this problem. Further, all models have comparable false positive rates, which suggests that reducing false negatives does not mean increasing false positives.

There are three key takeaways from this analysis. First, a robust model will return parameter estimates that are closer in value to the perceived network, as determined by the bias in the estimation. Second, a nominal model tends to yield parameter estimates that are biased relative to the underlying perceived network, and the problem is worse for more complex statistics (e.g., centralization). Third, a robust model makes fewer inference errors, particularly false negatives. These findings have direct implications for statistical inference and hypothesis testing. Given that the robust model is less biased and less prone to inference errors relative to the perceived network, we posit that the corresponding parameter estimates are *better representations of the network structure as it relates to individual decision making*. Thus, when

hypotheses are formulated at an individual level, the robust model may be more appropriate. Finally, because the nominal model yields biased estimates, we argue that not using the robust model increases the risk of incorrectly rejecting (or failing to reject) the null hypothesis.

Study 3: Empirical Example

Dataset & Analyses

In our final study, we present evidence using real world data to support the contention that measurement errors are prevalent, and can have significant effects, when using digital trace data. Further, we sought a context that enables us to generalize beyond a strictly social network. We analyzed data from the online encyclopedia Wikipedia. We accessed data used in a recently published study (Lerner and Lomi 2019) that has been made publicly available for research[2]. Our sample is composed of all recorded edits to articles during October and November of 2017. In total, we analyzed 141,364 edit events made by 2,665 unique users on 49,914 unique pages. Every observation was recorded in the format (t, u, v) which represents <Time, User, Article>. The full dataset is the ordered sequence $E = \{e_1, e_2, \dots, e_{141364}\}$ where $e_i = (t_i, u_i, v_i)$ and $t_{i+1} > t_i$ for all $i = 1, \dots, 141{,}364$.

We conducted an analysis of contributor behavior, namely, the probability that a user u would contribute to article v at time t. As predictors, we used four explanatory mechanisms used in prior studies: inertia, activity, popularity, and four-cycle (Brunswicker and Schecter 2019; Lerner and Lomi 2019; Quintane et al. 2014). To calculate each of these statistics, we used the weight $\omega_{uv}(t)$ which captured the frequency of u editing v up to time t, weighted by recency. Specifically, the formula for $\omega(t)$ we used was: $\omega_{uv}(t) = \sum_{i:t_i<t} 1\{t_i = \tau, u_i = u, v_i = v\} \times$

[2] The dataset and description can be found at the following DOI: 10.5281/zenodo.1626322

$\exp\left(-\frac{(t-\tau)\log 2}{T_{1/2}}\right)$. We can interpret $\omega_{uv}(t)$ as the instances of u editing v up to time t, with the weight of each prior event decayed according to a half-life $T_{1/2}$. By using a half-life, we are able to account for the relative salience of more recent events compared to events in the distant past. For our analyses we used a half-life of $T_{1/2}$ = 7 days.

The variable inertia measures the frequency with which u has edited v prior to the present time t. The formula for inertia is $x_{uv}^{I}(t) = \omega_{uv}(t)$. Activity represents the frequency with which user u has made any edits in the past, and popularity represents the frequency with which article v has received any edits in the past. We calculate activity as $x_{uv}^{A}(t) = \sum_h \omega_{uh}(t)$ and popularity as $x_{uv}^{P}(t) = \sum_k \omega_{kv}(t)$. Finally, the four-cycle captures the extent to which u will edit v, as a function of how frequently u has jointly edited other articles h with other users k and how frequently those users have contributed to v. The formula for a four-cycle is $x_{uv}^{C}(t) = \sum_k \omega_{kv}(t) \sum_h \omega_{uh}(t)\, \omega_{kh}(t)$. Essentially, the four-cycle measures the tendency for editors to work in local clusters on the same subset of articles.

Given these four statistics, we estimate the following model predicting the probability of an edit event at a particular time:

$$\text{logit}(p_{uvt}) = \beta_1 x_{uv}^{I}(t) + \beta_2 x_{uv}^{A}(t) + \beta_3 x_{uv}^{P}(t) + \beta_4 x_{uv}^{C}(t) + \varepsilon_{uvt}.$$

The solution to this model, $\boldsymbol{\beta}^{\star}$, is the nominal estimator for our dataset. We then calculate our robust estimator, $\boldsymbol{\beta}^{R}(\rho)$, at three levels of robustness. These levels of robustness are premised on the assumption that editors might be inaccurate in their recollections of their own prior editing activities or those with whom they coedited. The values we used are $\rho = 0.1, 0.3, 0.5$, which are roughly equal to 0.5x, 1.0x, 1.5x the standard errors of the four variables.

Finally, we conducted experiments on the sampled data to determine the fit and bias of the robust estimator. We began by generating artificial "measurement error" in our dataset by

randomly perturbing each statistic by a small amount. Namely, each statistic $x_{uv}(t)$ was replaced by $\bar{\boldsymbol{x}}_{uv}(t) = \boldsymbol{x}_{uv}(t) + \delta$, where δ was a random error drawn from a Uniform$(-\alpha\sigma, \alpha\sigma)$ distribution. The value σ is the standard error of the statistic and α is a parameter controlling the magnitude of the errors; we test α = 0.5 to 1.5 in increments of 0.25. By adding random errors of various magnitudes, we are preserving the average values of all four statistics, while increasing the amount of variance in our data. By adding errors from the uniform distribution, we allow for any error within that range to be equally likely[3]. Essentially, we are replicating a scenario where our data contains some degree of measurement error made by editors, that we do not capture but might have influenced their actions. The statistics $\bar{\boldsymbol{x}}_{uv}(t)$ can be thought of as the individual belief, while $\boldsymbol{x}_{uv}(t)$ is the empirical instantiation we observe. Now, we can refit the prior logit regression which yields the estimated parameters $\boldsymbol{\beta}^E$, or parameters under error. We compare $\boldsymbol{\beta}^\star$ and $\boldsymbol{\beta}^R(\rho)$ to $\boldsymbol{\beta}^E$ in order to determine which estimates are less biased with regards to the perturbed data. The model with the least bias should be the one which best approximates the parameters related to individual beliefs.

Results

We report the nominal and robust estimators in Table 7. In the nominal case, all four variables are positive and significant, indicating that each of the mechanisms influence an individual's decision regarding which article to contribute to. Likewise, at all three levels of robustness we find consistent results; thus, we would reach the same qualitative conclusions from an explanatory perspective. As expected, the log-likelihood for the robust models is worse

[3] By contrast, normally distributed errors would make larger magnitude errors less likely than smaller errors, effectively reducing the noise in the data. We did however conduct tests with normally distributed errors and found the same results.

than the nominal case. Interestingly, we note that the robust model attenuates some variables but amplifies others.

Table 7. Nominal and Robust Estimates for Wikipedia Data

	Nominal	$\rho = 0.1$	$\rho = 0.3$	$\rho = 0.5$
Inertia (β_1)	11.210*	7.092*	7.738*	5.978*
	(0.213)	(0.003)	(0.003)	(0.004)
Activity (β_2)	1.131*	1.288*	2.108*	2.404*
	(0.004)	(0.001)	(0.001)	(0.001)
Popularity (β_3)	0.570*	1.112*	0.973*	1.449*
	(0.011)	(0.002)	(0.002)	(0.004)
Four-Cycle (β_4)	0.325*	1.296*	0.740*	1.015*
	(0.017)	(0.002)	(0.001)	(0.002)
Log Likelihood	-90,103	-165,490	-386,300	-317,830

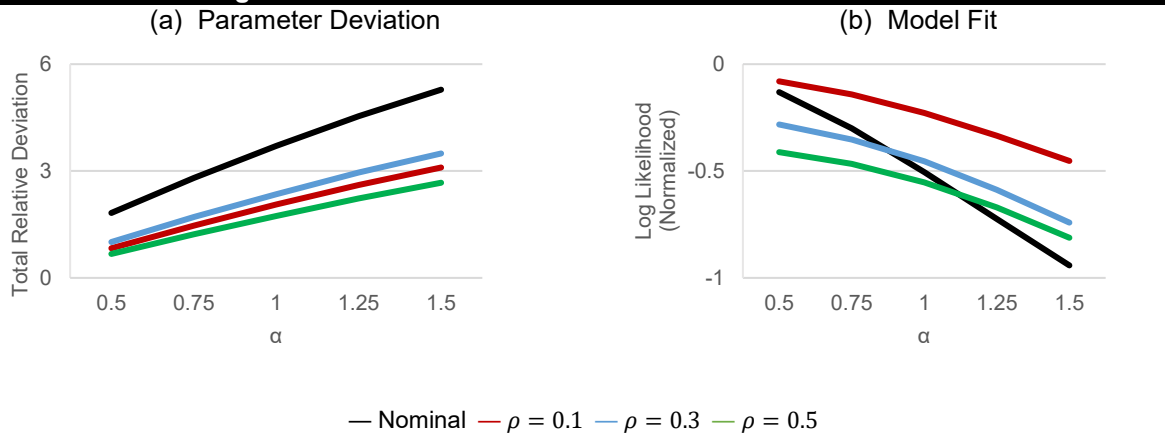
*Notes. Standard errors in parentheses. * p < 0.001*

In particular, the effect of inertia is attenuated in each of the robust models, indicating that the models are reducing the effect of inertia. On the other hand, activity, popularity, and the four-cycle are all *amplified* by the robust models. We also find that the parameter estimates do not increase or decrease linearly with the robustness parameter ρ. This finding demonstrates the capacity of robust optimization to find different local solutions over nonlinear data. Finally, the standard errors of the robust estimates are smaller than the nominal standard errors, often by an order of magnitude. Further, the standard errors for the robust estimates are relatively consistent over values of ρ. This finding reflects an important characteristic of the robust method; estimates are significantly more precise, even for small error tolerances.

We next compare the performance of the robust and nominal models when measurement error is added to the dataset. In Figure 6a we present the normalized deviation of the nominal $\frac{\|\boldsymbol{\beta}^E - \boldsymbol{\beta}^\star\|}{\|\boldsymbol{\beta}^E\|}$ and robust parameters $\frac{\|\boldsymbol{\beta}^E - \boldsymbol{\beta}^R\|}{\|\boldsymbol{\beta}^R\|}$ (from Table 7), and in Figure 6b we present the model fit of the estimates to the perturbed dataset. We find that the robust estimators $\boldsymbol{\beta}^R(\rho)$ deviate significantly less relative to the estimates under measurement error $\boldsymbol{\beta}^E$ (see Figure 6a). Essentially, if measurement error were present in our dataset, the typical estimator $\boldsymbol{\beta}^\star$ would be

significantly different than the true population parameters. The robust estimator on the other hand would be closer in value to those true parameters.

Figure 6. Deviation and Fit to Data with Measurement Error



Further, from Figure 6b we find that the robust estimators exhibit a better fit to the data under measurement error, indicating that the robust estimator has greater predictive power. It's worth noting that if the measurement error is significantly smaller than the robust tolerance, the nominal model still has a better overall fit (for example, $\rho = 0.5$ and $\alpha = 0.5$). However, with large errors, all robust models outperform the nominal model. Overall, our findings indicate that if empirical data contains some unobserved measurement error, a robust estimator will be less biased with respect to the parameters and exhibit better model fit than the nominal approach. Further, the robust model is able to achieve these performance gains while preserving the explanatory conclusions.

DISCUSSION

In this study, we explore how discrepancies between observable network data and perceived network data can bias statistical analyses. We delineate a number of sources of contamination in social network data. These include technical features of the nature of online data sources, including the size and magnitude of online networks, the rate at which they evolve, and the varying degrees of people's perceptions of the network (its translucency or visibility)

across online platforms. The second source of discrepancy stems from natural cognitive tendencies of the individuals within the network, such as compression, personality traits, and positions such as power. Despite this evidentiary body of literature, social network inference methodologies generally treat observable data, such as networks collected from online sources, as accurate proxies for the perceived network on the bases of which people often act. Inference about human behavior is derived from measures of this empirical information, which may deviate from the internal schema that individuals believe to be true. And, as discussed earlier, many social science theories, and more specifically IS studies, offer explanations based on people's perceptions of actions and interactions rather than their objective occurrences as captured, say, by digital trace data.

Our primary contribution is to introduce a methodology that significantly reduces the bias in estimates of error laden social network patterns, which subsequently leads to more accurate statistical inference. We extend the work of Bertsimas and Nohadani (2019) by deriving a robust maximum likelihood estimator for the exponential family of probability distributions. Further, we introduce a robust test statistic that allows researchers to conduct hypothesis testing after correcting for errors. Our framework preserves the techniques of classic network analysis, while making the estimates resilient to the discrepancies we recognize are present in our data. In both our simulation experiments and empirical examples, the robust MLE produced estimates that were less biased relative to ground-truth values. Further, our method leads to more accurate hypothesis testing; we found that the robust method had greater statistical power and resulted in fewer incorrect conclusions.

Beyond the quantitative advantages enabled by our approach, there are also qualitative advantages to the robust method. Robust estimation acts as a type of filter for cognitive effects,

essentially imposing a larger burden of proof on structures that may be difficult for individuals to detect. Researchers can now incorporate explanations based on people's agency into their network hypotheses knowing that the robust parameters broadly incorporate the potential sources of bias. By accounting for cognition, the standard and robust models ask inherently different questions. Standard network inference asks, "what is the effect of structure x on behavior y, assuming that the perceived structure matches the truth?" whereas robust network inference asks, "what is the effect of structure x on behavior y, if the actor responsible for that behavior perceives structure x somewhat differently than what is assumed to be the truth?"

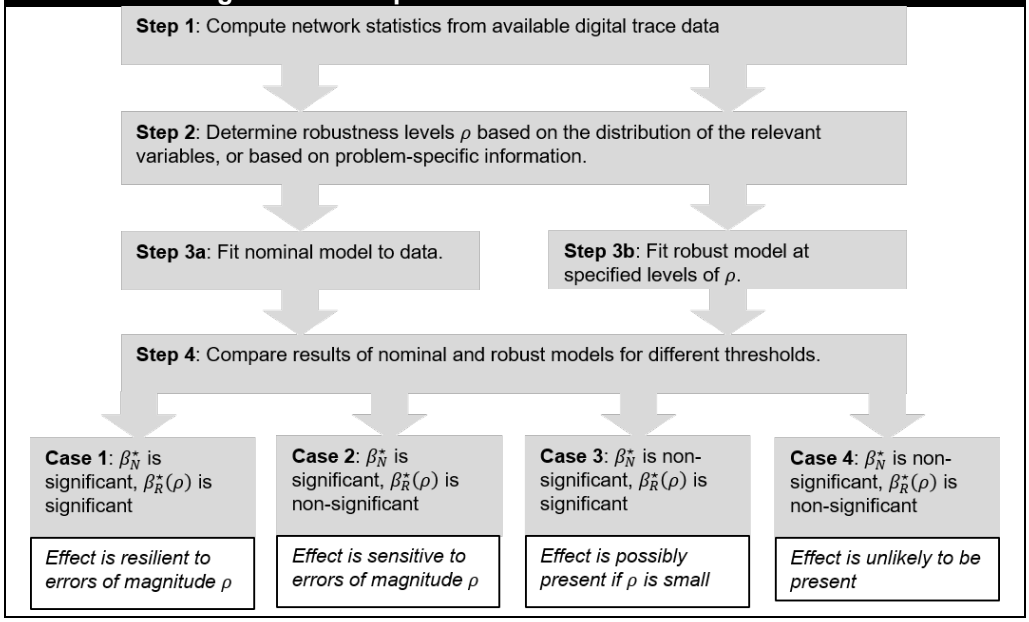
The differences between a nominal approach to network inference and a robust approach are subtle, but they highlight an important deviation from the standard method of analyzing networks. Current models of agentic behavior model cognition - e.g. recency effects incorporated in Butts' (2008) relational event framework – but only to the extent that empirical observation matches perceived reality. Taking a robust approach to analysis makes the interpretation of true agency more realistic, since it directly incorporates the systematic errors that are likely to occur.

Applying the Robust Method

Considering the findings of our study, we provide a general template for conducting statistical inference on digital trace network data with robust maximum likelihood. We identify four key steps in the process, which we illustrate in Figure 7. First, researchers should construct the network and compute the necessary statistics (e.g., transitivity or preferential attachment). The choice of statistics should be informed by the relevant theory or theories being tested. Second, the researcher should determine the appropriate levels of robustness, with the robustness level corresponding to confidence bounds on the error magnitudes. In this paper, we used multiples of the standard deviations of the network statistics. This approach would be appropriate

when considering multiple observations of networks (e.g., Faraj and Johnson 2011) or multiple individual-level actions (e.g., Quintane and Carnabuci 2016). In situations where only one network is being considered, uncertainty bounds could be selected based on percentages of the statistic value (e.g., 10% variation). Ideally, a range of robustness levels are selected for analysis.Third, multiple models are fit: the nominal model, as well as a robust model for each of the specified tolerance levels. This step will yield the nominal parameters $\boldsymbol{\beta}_N^\star$, robust parameters $\boldsymbol{\beta}_R^\star(\rho)$, and the standard errors for each coefficient using the procedure described. Finally, the research should conduct statistical inference by computing the test statistic $\beta/SE(\beta)$ for each coefficient across the robustness levels. The researcher will reach one of four conclusions for each level of robustness tested.

Figure 7. A Template for Robust Network Inference



Case 1: The nominal coefficient is significant (indicating a strong effect) and the robust coefficient at level ρ is also significant. Here, we conclude that even if the underlying data is not accurate (within a range given by ρ), the effect is still strong enough to be

detected. This case would lead the researcher to conclude their effect is *resilient to measurement errors of magnitude* ρ.

Case 2: The nominal coefficient is significant, but the robust coefficient at level ρ is not. Essentially, the hypothesized effect disappears if bias of magnitude ρ is present in the observed data, suggesting an amplification bias. The researcher would then conclude that the measure is *sensitive to measurement errors of magnitude* ρ; in other words, there is a greater probability of false positive errors.

Case 3: The nominal coefficient is not significant, but the robust estimate is significant at level ρ. When ρ is small (e.g., a fraction of one standard deviation), then this scenario could represent attenuation, i.e., measurement error leading to a bias towards zero. The robust estimator would account for that error and would thus *potentially remedy a false negative error*. However, the larger the value of ρ, the less likely it is that the underlying effect is present.

Case 4: Neither the nominal nor robust coefficients are significant. In this scenario, the researcher can conclude that the lack of observed effect is not due to simple errors.

As the above cases make clear, robust optimization should be used in addition to standard methods, not in place of them. The results obtained from the nominal data are important because they identify key patterns in the data, but they may be misleading with regards to network perception and individual decision making. If the nominal model identifies an independent variable as significant, but the robust version does not, then the effect should be interpreted through a more conservative lens, particularly with regards to cognition. Conversely, when the robust model identifies an effect that the nominal model does not, then that measure may be

suffering from attenuation bias. In both cases, the robust approach improves our ability to make accurate inferences, and subsequently improves our ability to conduct hypothesis testing.

Limitations

While the robust approach produces results that are more sensitive to cognitive issues, there are still limitations to our interpretations. First, the method introduced in this study reduces the likelihood of falsely concluding an association between two variables, but still does not imply what an individual does or does not perceive. Second, there are alternative methods for correcting biased observable networks. We could obtain CSS information from every individual at every time-point being considered. For a single network this may be reasonable, but for longitudinal networks this process becomes increasingly impractical. Additionally, these reports would have to be collected *when the tie was formed*, otherwise the data is still subject to errors in recollection. Alternatively, we could infer an individual's perceived network, based on systematic biases found empirically. The challenge with this approach is the impracticality in empirically identifying a specific underlying distribution for cognition that takes into account all the theoretical sources of bias. This approach could potentially provide a less conservative alternative to robust estimation, but further research is needed.

A third general limitation stems from the modeling decisions of the practitioner. Because we are explicitly assuming to have no a priori knowledge of the errors, our choice in the geometry of the uncertainty sets in fact represents our implicit assumptions about the errors. To mitigate the bias that could be injected by this process, we recommend testing a variety of uncertainty sets and carefully documenting the discrepancies between various models. When we have reasons to believe that the errors follow some distribution, then an ellipsoidal geometric set becomes a natural choice (see Bertsimas and Nohadani, 2019). On the other hand, when only

maximum errors are known, a polyhedral set – as used in our implementation – offers a good description. Likewise, the analyses presented in this paper assume an exponential likelihood function for the network data. Though this assumption is common in network analysis, there are other ways to model the data that we do not consider. Examples include multilevel models (Sweet et al. 2013), quadratic assignment (Krackardt 1987), and graph embedding (Cui et al. 2017). Robust optimization could be used to find estimators that are immunized to error in these models. However, future research is needed.

Finally, given the difficulty of collecting both trace data and self-reports for large networks, there are limited studies directly measuring the extent of perception errors, particularly in IS settings. However, evidence from large-scale studies of email exchanges, social media, and proximity data from wearable or mobile devices suggest that while trace information is relatively consistent with sociometric surveys, there is a persistent misalignment. Future research should more thoroughly examine the prevalence and magnitude of this perception gap.

CONCLUSION

The increasing availability of digital trace data is celebrated as a bonanza for computational social science approaches, including social network analytics. The problem of perception limits the interpretation of hypothesized mechanisms in social network analysis conducted using digital trace data that offer varying degrees of technological affordances. Hence, explanations that assume individuals act and interact based on the observed network in the digital trace data are not always warranted. We propose a novel method that looks for inferences that are robust to differences between the observed network data and individuals' perceptions of those data. Our proposed method applies advances in robust optimization to the field of social networks by integrating robust techniques into common inferential models. Using data from

computer simulations, laboratory experiments, digital sources and field settings, we illustrated the efficacy of our technique in adjusting the effects of variables impacted by perception and providing overall better predictive capabilities.

APPENDIX: TECHNICAL DETAILS

Proof of Solution to Inner Problem

In order to determine the exact solution to the inner problem, we first compute the gradient of the objective function with respect to the value $\boldsymbol{\beta}'\Delta\boldsymbol{x}_{tq\bullet}$. For notational purposes, let $h_t = \boldsymbol{\beta}'\left(\boldsymbol{x}_{ty\bullet} - \Delta\boldsymbol{x}_{ty\bullet}\right) - \log\sum_{z\in\mathcal{A}_t}\exp\left(\boldsymbol{\beta}'(\boldsymbol{x}_{tz\bullet} - \Delta\boldsymbol{x}_{tz\bullet})\right)$.

$$\frac{\partial}{\partial\left(\boldsymbol{\beta}'\Delta\boldsymbol{x}_{tq\bullet}\right)}(h_t) = -\mathbf{1}\{q = y_t\} + \frac{\exp\left(\boldsymbol{\beta}'\left(\boldsymbol{x}_{tq\bullet} - \Delta\boldsymbol{x}_{tq\bullet}\right)\right)}{\sum_{z\in\mathcal{A}_t}\exp\left(\boldsymbol{\beta}'(\boldsymbol{x}_{tz\bullet} - \Delta\boldsymbol{x}_{tz\bullet})\right)}$$

$$= \begin{cases} -1 + \dfrac{\exp\left(\boldsymbol{\beta}'\left(\boldsymbol{x}_{tq\bullet} - \Delta\boldsymbol{x}_{tq\bullet}\right)\right)}{\sum_{z\in\mathcal{A}_t}\exp\left(\boldsymbol{\beta}'(\boldsymbol{x}_{tz\bullet} - \Delta\boldsymbol{x}_{tz\bullet})\right)}, & q = y_t \\ \dfrac{\exp\left(\boldsymbol{\beta}'\left(\boldsymbol{x}_{tq\bullet} - \Delta\boldsymbol{x}_{tq\bullet}\right)\right)}{\sum_{z\in\mathcal{A}_t}\exp\left(\boldsymbol{\beta}'(\boldsymbol{x}_{tz\bullet} - \Delta\boldsymbol{x}_{tz\bullet})\right)}, & q \neq y_t \end{cases}$$

Here, $\mathbf{1}\{q = y\}$ is an indicator function, taking a value of 1 if $q = y$ and 0 otherwise. Because the value of $\frac{\exp\left(\boldsymbol{\beta}'(\boldsymbol{x}_{tq\bullet} - \Delta\boldsymbol{x}_{tq\bullet})\right)}{\sum_{z\in\mathcal{A}_t}\exp\left(\boldsymbol{\beta}'(\boldsymbol{x}_{tz\bullet} - \Delta\boldsymbol{x}_{tz\bullet})\right)}$ is non-negative and at most 1, we can conclude that $\frac{\partial h_t}{\partial\left(\boldsymbol{\beta}'\Delta\boldsymbol{x}_{tq\bullet}\right)} \leq 0$ for $q = y$, and $\frac{\partial h_t}{\partial\left(\boldsymbol{\beta}'\,\Delta\boldsymbol{x}_{tq\bullet}\right)} \geq 0$ for all other q. Thus, h_t is a monotonically decreasing function of $\boldsymbol{\beta}'\Delta\boldsymbol{x}_{ty\bullet}$ and a monotonically increasing function of $\boldsymbol{\beta}'\Delta\boldsymbol{x}_{tz\bullet}$ for all $z \neq y$.

Solving the Robust Estimator

We summarize the process of determining the robust estimators in the Algorithm 1. If the step-length parameter is chosen such that it has diminishing size, i.e., $\sum_k \alpha_k = \infty$, and $\alpha_k \to 0$ as $k \to \infty$, then the outlined procedure will converge to a locally optimal solution $\boldsymbol{\beta}^\star$ in polynomial time (Bertsimas et al. 2010). We used a step length of $\alpha_k = \frac{\|\nabla\phi(\boldsymbol{\beta}^{(0)})\|}{k}$.

Algorithm 1:

1. Initialize with an estimator $\boldsymbol{\beta}^{(0)}$. Set $k = 0$.
2. Solve the inner problem for all $t = 1, \ldots, M$ to obtain the optimal errors $\Delta \boldsymbol{x}_{tz\bullet}^{\star}\left(\boldsymbol{\beta}^{(k)}\right)$ for each $z \in \mathcal{A}_t$.
3. Using the worst case errors $\Delta \mathbf{X}^{\star}\left(\boldsymbol{\beta}^{(k)}\right)$, calculate $\phi\left(\boldsymbol{\beta}^{(k)}; \mathbf{X}^{obs}\right) = \psi\left(\boldsymbol{\beta}^{(k)}; \mathbf{X}^{obs} - \Delta \mathbf{X}^{\star}\left(\boldsymbol{\beta}^{(k)}\right)\right)$. By applying Danskin's theorem, we know that computing $\nabla\psi\left(\boldsymbol{\beta}^{(k)}\right)$ is equivalent to computing $\nabla\phi\left(\boldsymbol{\beta}^{(k)}\right)$ (Bertsimas and Nohadani, 2019). If the gradient does not exist, compute a subgradient. Denote the gradient or subgradient as g.
4. Update $\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + \alpha_k g$, where α_k is a step-length parameter.
5. Stop when the relative change in objective function is less than ϵ, $\epsilon > 0$ is a stopping criterion. Otherwise, $k = k + 1$ and return to Step 2.

Computation of the Subgradient

To solve the robust maximum likelihood problem, the gradient of the outer problem – assuming a known solution to the error terms $\Delta \boldsymbol{x}^{\star}(\boldsymbol{\beta})$ – must be computed. We assume that the norm $\|\bullet\|$ refers to the Euclidean norm. For our specified likelihood function, the gradient is as follows:

$$\nabla\phi_p(\boldsymbol{\beta}) = \frac{\partial}{\partial\beta_p}\left(\sum_{t=1}^{M} \boldsymbol{\beta}'\boldsymbol{x}_{ty\bullet}^{\text{net}} - \|\boldsymbol{\beta}\|\rho - \log\sum_{z\in\mathcal{A}_t}\exp(\boldsymbol{\beta}'\boldsymbol{x}_{tz\bullet}^{\text{net}} + (-1)^{u_z}\|\boldsymbol{\beta}\|\rho)\right)$$

$$= \sum_{t=1}^{M} x_{typ}^{\text{net}} - \frac{\beta_p}{\|\boldsymbol{\beta}\|}\rho - \frac{\sum_{z\in\mathcal{A}_t}\left(x_{tzp}^{\text{net}} + (-1)^{u_z}\frac{\beta_p}{\|\boldsymbol{\beta}\|}\rho\right)\exp(\boldsymbol{\beta}'\boldsymbol{x}_{tz\bullet}^{\text{net}} + (-1)^{u_z}\|\boldsymbol{\beta}\|\rho)}{\sum_{z\in\mathcal{A}_t}\exp(\boldsymbol{\beta}'\boldsymbol{x}_{tz\bullet}^{\text{net}} + (-1)^{u_z}\|\boldsymbol{\beta}\|\rho)}$$

In the case that the vector $\boldsymbol{\beta} = \mathbf{0}$, then the gradient cannot be directly computed. Instead, we may compute a subgradient. Given the convexity of the norm, the logarithmic function, and the exponential function, and that the objective function is the negation of these functions, we may conclude that the log-likelihood function for the robust problem is concave. As such, the subgradient is a vector $\boldsymbol{v}$ that satisfies the following inequality

$$\psi\left(\boldsymbol{\beta}_2; \mathbf{X}^{\text{net}} - \Delta\mathbf{X}^{\star}(\boldsymbol{\beta}_2)\right) - \psi\left(\boldsymbol{\beta}_1; \mathbf{X}^{\text{net}} - \Delta\mathbf{X}^{\star}(\boldsymbol{\beta}_1)\right) \leq \boldsymbol{v}\cdot(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1).$$

Thus, the subgradient used in the optimization of the outer objective function is as follows:

$$g(\boldsymbol{\beta}) = \begin{cases} \nabla\phi(\boldsymbol{\beta}), & \boldsymbol{\beta} \neq \mathbf{0} \\ \boldsymbol{v}, & \boldsymbol{\beta} = \mathbf{0} \end{cases}.$$

Computation of the Hessian Matrix

The robust standard errors for the estimator are derived from the inverse information matrix at the optimal solution. To obtain this matrix, we require the matrix of second derivatives of the likelihood function, i.e., the Hessian. We proceed to take the derivative of the subgradient as defined previously.

$$\begin{aligned}\nabla^2\phi_{pq}(\boldsymbol{\beta}) &= \frac{\partial^2}{\partial\beta_p\partial\beta_q}\left(\sum_{t=1}^{M}\boldsymbol{\beta}'\boldsymbol{x}_{ty\bullet}^{\text{net}} - \|\boldsymbol{\beta}\|\rho - \log\sum_{z\in\mathcal{A}_t}\exp(\boldsymbol{\beta}'\boldsymbol{x}_{tz\bullet}^{\text{net}} + (-1)^{u_z}\|\boldsymbol{\beta}\|\rho)\right)\\ &= \frac{\partial}{\partial\beta_q}\Big(\nabla\phi_p(\boldsymbol{\beta})\Big)\\ &= \sum_{t=1}^{M} -\mathbb{I}[p=q] + \frac{\beta_p\beta_q}{\|\boldsymbol{\beta}\|^3} - \frac{A}{C} + \frac{B}{C^2}\end{aligned}$$

The values A, B, and C above are simply placeholders for more complex expressions. Below, we provide the equations for each. Note that $\mathbb{I}[\dots]$ is the indicator function.

$$\begin{aligned}A &= \sum_{z\in\mathcal{A}_t}\left[\exp(\boldsymbol{\beta}'\boldsymbol{x}_{tz\bullet}^{\text{net}} + (-1)^{u_z}\|\boldsymbol{\beta}\|\rho)\left(x_{tzp}^{\text{net}}x_{tzq}^{\text{net}} + \frac{\beta_q}{\|\boldsymbol{\beta}\|}\rho\right.\right.\\ &\qquad\left.\left. + (-1)^{u_z}\left[x_{tzp}^{\text{net}}\frac{\beta_q}{\|\boldsymbol{\beta}\|}\rho + x_{tzq}^{\text{net}}\frac{\beta_p}{\|\boldsymbol{\beta}\|}\rho + \mathbb{I}[p=q] - \frac{\beta_p\beta_q}{\|\boldsymbol{\beta}\|^3}\right]\right)\right]\\ B &= \sum_{z\in\mathcal{A}_t}\left[\exp(\boldsymbol{\beta}'\boldsymbol{x}_{tz\bullet}^{\text{net}} + (-1)^{u_z}\|\boldsymbol{\beta}\|\rho)\left(x_{tzq}^{\text{net}} + (-1)^{u_z}\frac{\beta_q}{\|\boldsymbol{\beta}\|}\rho\right)\right]\\ C &= \sum_{z\in\mathcal{A}_t}\exp(\boldsymbol{\beta}'\boldsymbol{x}_{tz\bullet}^{\text{net}} + (-1)^{u_z}\|\boldsymbol{\beta}\|\rho)\end{aligned}$$

REFERENCES

Agarwal, R., Gupta, A. K., and Kraut, R. 2008. "**Editorial Overview** —The Interplay Between Digital and Social Networks," *Information Systems Research* (19:3), pp. 243–252. (https://doi.org/10.1287/isre.1080.0200).

Aral, S., and Van Alstyne, M. 2011. "The Diversity-Bandwidth Trade-Off," *American Journal of Sociology* (117:1), pp. 90–171. (https://doi.org/10.1086/661238).

Aral, S., and Walker, D. 2012. "Identifying Influential and Susceptible Members of Social Networks," *Science* (337:6092), pp. 337–341.

Aral, S., and Walker, D. 2014. "Tie Strength, Embeddedness, and Social Influence: A Large-Scale Networked Experiment," *Management Science* (60:6), pp. 1352–1370. (https://doi.org/10.1287/mnsc.2014.1936).

Bapna, R., Gupta, A., Rice, S., and Sundararajan, A. 2017. "Trust and the Strength of Ties in Online Social Networks: An Exploratory Field Experiment," *MIS Quarterly* (41:1), pp. 115–130. (https://doi.org/10.25300/MISQ/2017/41.1.06).

Bapna, R., Qiu, L., and Rice, S. 2017. "Repeated Interactions versus Social Ties: Quantifying the Economic Value of Trust, Forgiveness, and Reputation Using a Field Experiment," *MIS Quarterly* (41:3), pp. 841–866.

Bapna, R., and Umyarov, A. 2015. "Do Your Online Friends Make You Pay? A Randomized Field Experiment on Peer Influence in Online Social Networks," *Management Science* (61:8), pp. 1902–1920. (https://doi.org/10.1287/mnsc.2014.2081).

Barabási, A. L. (2010). *Bursts: the hidden patterns behind everything we do, from your e-mail to bloody crusades*. Penguin.

Batchelder, W. H., Kumbasar, E., and Boyd, J. P. 1997. "Consensus Analysis of Three-way Social Network Data," *Journal of Mathematical Sociology* (22:1), pp. 29–58.

Ben-Tal, A., El Ghaoui, L., and Nemirovski, A. 2009. *Robust Optimization*, Princeton University Press.

Ben-Tal, A., and Nemirovski, A. 2002. "Robust Optimization–Methodology and Applications," *Mathematical Programming* (92:3), pp. 453–480.

Berente, N., Seidel, S., and Safadi, H. 2018. "Data-Driven Computationally-Intensive Theory Development," *Information Systems Research* (forthcoming).

Berger, K., Klier, J., Klier, M., and Probst, F. 2014. "A Review of Information Systems Research on Online Social Networks," *CAIS* (35), p. 8. (https://doi.org/10.17705/1cais.03508).

Bernard, H. R., Killworth, P. D., and Sailer, L. 1982. "Informant Accuracy in Social-Network Data V. An Experimental Attempt to Predict Actual Communication from Recall Data," *Social Science Research* (11:1), pp. 30–66.

Bertsimas, D., Brown, D. B., and Caramanis, C. 2011. "Theory and Applications of Robust Optimization," *SIAM Review* (53:3), pp. 464–501.

Bertsimas, D., and Nohadani, O. 2019. "Robust Maximum Likelihood Estimation," *INFORMS Journal on Computing*, Ijoc.2018.0834. (https://doi.org/10.1287/ijoc.2018.0834).

Bertsimas, D., Nohadani, O., and Teo, K. M. 2007. "Robust Optimization in Electromagnetic Scattering Problems," *Journal of Applied Physics* (101:7), p. 074507. (https://doi.org/10.1063/1.2715540).

Bertsimas, D., Nohadani, O., and Teo, K. M. 2010. "Robust Optimization for Unconstrained Simulation-Based Problems," *Operations Research* (58:1), pp. 161–178.

Bertsimas, D., and Sim, M. 2004. "The Price of Robustness," *Operations Research* (52:1), pp. 35–53.
Bhattacharya, P., Phan, T. Q., Bai, X., and Airoldi, E. M. 2019. "A Coevolution Model of Network Structure and User Behavior: The Case of Content Generation in Online Social Networks," *Information Systems Research*. (https://doi.org/10.1287/isre.2018.0790).
Brands, R. A. 2013. "Cognitive Social Structures in Social Network Research: A Review," *Journal of Organizational Behavior* (34:S1), pp. S82–S103.
Brashears, M. E., and Quintane, E. 2015. "The Microstructures of Network Recall: How Social Networks Are Encoded and Represented in Human Memory," *Social Networks* (41), pp. 113–126.
Brunswicker, S., and Schecter, A. 2019. "Coherence or Flexibility? The Paradox of Change for Developers' Digital Innovation Trajectory on Open Platforms," *Research Policy*. (https://doi.org/10.1016/j.respol.2019.03.016).
Burt, R. S., Kilduff, M., and Tasselli, S. 2013. "Social Network Analysis: Foundations and Frontiers on Advantage," *Annual Review of Psychology* (64), pp. 527–547.
Butts, C. T. 2008. "A Relational Event Framework for Social Action," *Sociological Methodology* (38:1), pp. 155–200.
Cao, J., Basoglu, K. A., Sheng, H., and Lowry, P. B. 2015. "A Systematic Review of Social Networks Research in Information Systems: Building a Foundation for Exciting Future Research," *CAIS* (36), p. 37. (https://doi.org/10.17705/1cais.03637).
Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. 2006. *Measurement Error in Nonlinear Models: A Modern Perspective*, Chapman and Hall/CRC.
Carroll, R. J., and Stefanski, L. A. 1994. "Measurement Error, Instrumental Variables and Corrections for Attenuation with Applications to Meta-Analyses," *Statistics in Medicine* (13:12), pp. 1265–1282.
Casciaro, T. 1998. "Seeing Things Clearly: Social Structure, Personality, and Accuracy in Social Network Perception," *Social Networks* (20:4), pp. 331–351.
Casciaro, T., Carley, K. M., and Krackhardt, D. 1999. "Positive Affectivity and Accuracy in Social Network Perception," *Motivation and Emotion* (23:4), pp. 285–306.
Casciaro, T., Gino, F., and Kouchaki, M. 2014. "The Contaminating Effects of Building Instrumental Ties: How Networking Can Make Us Feel Dirty," *Administrative Science Quarterly* (59:4), SAGE Publications Inc, pp. 705–735. (https://doi.org/10.1177/0001839214554990).
Chen, W., Wei, X., and Zhu, K. X. 2017. "Engaging Voluntary Contributions in Online Communities: A Hidden Markov Model," *MIS Quarterly* (42:1).
Contractor, N. (2018). How Can Computational Social Science Motivate the Development of Theories, Data, and Methods to Advance Our Understanding of Communication and Organizational Dynamics? In Brooke Foucault Welles and Sandra González-Bailón (Eds.), *The Oxford Handbook of Networked Communication*. Oxford University Press.
Corman, S. R. 1990. "A Model of Perceived Communication in Collective Networks," *Human Communication Research* (16:4), pp. 582–602. (https://doi.org/10.1111/j.1468-2958.1990.tb00223.x).
Corman, S. R., and Scott, C. R. 1994. "Perceived Networks, Activity Foci, and Observable Communication in Social Collectives," *Communication Theory* (4:3), pp. 171–190. (https://doi.org/10.1111/j.1468-2885.1994.tb00089.x).

Cox, D. R. 1972. "Regression Models and Life-Tables," *Journal of the Royal Statistical Society. Series B (Methodological)* (34:2), pp. 187–220.
Cui, P., Wang, X., Pei, J., and Zhu, W. 2017. "A Survey on Network Embedding," *ArXiv:1711.08752 [Cs]*. (http://arxiv.org/abs/1711.08752).
Dahlander, L., and Frederiksen, L. 2011. "The Core and Cosmopolitans: A Relational View of Innovation in User Communities," *Organization Science* (23:4), pp. 988–1007. (https://doi.org/10.1287/orsc.1110.0673).
Dahlander, L., and O'Mahony, S. 2010. "Progressing to the Center: Coordinating Project Work," *Organization Science* (22:4), pp. 961–979. (https://doi.org/10.1287/orsc.1100.0571).
Dewan, S., Ho, Y.-J. (Ian), and Ramaprasad, J. 2017. "Popularity or Proximity: Characterizing the Nature of Social Influence in an Online Music Community," *Information Systems Research* (28:1), pp. 117–136. (https://doi.org/10.1287/isre.2016.0654).
Eagle, N., Pentland, A. S., and Lazer, D. 2009. "Inferring Friendship Network Structure by Using Mobile Phone Data," *Proceedings of the National Academy of Sciences* (106:36), pp. 15274–15278.
Faraj, S., and Johnson, S. L. 2011. "Network Exchange Patterns in Online Communities," *Organization Science* (22:6), pp. 1464–1480.
Flynn, F. J., Reagans, R. E., and Guillory, L. 2010. "Do You Two Know Each Other? Transitivity, Homophily, and the Need for (Network) Closure," *Journal of Personality and Social Psychology* (99:5), p. 855.
Foss, N. j., Frederiksen, L., and Rullani, F. 2016. "Problem-Formulation and Problem-Solving in Self-Organized Communities: How Modes of Communication Shape Project Behaviors in the Free Open-Source Software Community," *Strategic Management Journal* (37:13), pp. 2589–2610. (https://doi.org/10.1002/smj.2439).
Freeman, L. C. 1992. "Filling in the Blanks: A Theory of Cognitive Categories and the Structure of Social Affiliation," *Social Psychology Quarterly*, pp. 118–127.
Freeman, L. C., Romney, A. K., and Freeman, S. C. 1987. "Cognitive Structure and Informant Accuracy," *American Anthropologist* (89:2), pp. 310–325.
Gilbert, E., and Karahalios, K. 2009. *Predicting Tie Strength with Social Media*, presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, pp. 211–220.
Greene, W. H. 2003. *Econometric Analysis*, Pearson Education India.
Heald, M. R., Contractor, N. S., Koehly, L. M., and Wasserman, S. 1998. "Formal and Emergent Predictors of Coworkers' Perceptual Congruence on an Organization's Social Structure," *Human Communication Research* (24:4), pp. 536–563.
Holland, P. W., and Leinhardt, S. 1977. "A Dynamic Model for Social Networks," *Journal of Mathematical Sociology* (5:1), pp. 5–20.
Holland, P. W., and Leinhardt, S. 1981. "An Exponential Family of Probability Distributions for Directed Graphs," *Journal of the American Statistical Association* (76:373), pp. 33–50.
Howison, J., Wiggins, A., and Crowston, K. 2011. "Validity Issues in the Use of Social Network Analysis with Digital Trace Data," *Journal of the Association for Information Systems; Atlanta* (12:12), pp. 767–797.
Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M., and Morris, M. 2008. "Ergm: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks," *Journal of Statistical Software* (24:3), p. nihpa54860.

Janicik, G. A., and Larrick, R. P. 2005. "Social Network Schemas and the Learning of Incomplete Networks," *Journal of Personality and Social Psychology* (88:2), p. 348.
Johnson, R., Kovács, B., and Vicsek, A. 2012. "A Comparison of Email Networks and Off-Line Social Networks: A Study of a Medium-Sized Bank," *Social Networks* (34:4), pp. 462–469.
Johnson, S. L., Faraj, S., and Kudaravalli, S. 2014. "Emergence of Power Laws in Online Communities: The Role of Social Mechanisms and Preferential Attachment," *MIS Quarterly* (38:3), pp. 795–808.
Johnson, S. L., Safadi, H., and Faraj, S. 2015. "The Emergence of Online Community Leadership," *Information Systems Research* (26:1), pp. 165–187. (https://doi.org/10.1287/isre.2014.0562).
Kane, G. C., Alavi, M., Labianca, G. J., and Borgatti, S. 2014. "What's Different about Social Media Networks? A Framework and Research Agenda," *MIS Quarterly* (38:1), pp. 274–304.
Kilduff, M., and Brass, D. J. 2010. "Organizational Social Network Research: Core Ideas and Key Debates," *The Academy of Management Annals* (4:1), pp. 317–357.
Krackardt, D. 1987. "QAP Partialling as a Test of Spuriousness," *Social Networks* (9:2), pp. 171–186. (https://doi.org/10.1016/0378-8733(87)90012-8).
Krackhardt, D. 1987. "Cognitive Social Structures," *Social Networks* (9:2), pp. 109–134.
Krackhardt, D., and Kilduff, M. 1999. "Whether Close or Far: Social Distance Effects on Perceived Balance in Friendship Networks," *Journal of Personality and Social Psychology* (76:5), p. 770.
Lazer, D., Pentland, A. (Sandy), Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., and Van Alstyne, M. 2009. "Life in the Network: The Coming Age of Computational Social Science," *Science (New York, N.Y.)* (323:5915), pp. 721–723. (https://doi.org/10.1126/science.1167742).
Lazer, D. M. J., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., Freelon, D., Gonzalez-Bailon, S., King, G., Margetts, H., Nelson, A., Salganik, M. J., Strohmaier, M., Vespignani, A., & Wagner, C. (2020). Computational social science: Obstacles and opportunities. *Science*, *369*(6507), 1060–1062.
Leonardi, P., and Contractor, N. 2018. "Better People Analytics," *Harvard Business Review* (November–December 2018). (https://hbr.org/2018/11/better-people-analytics).
Lerner, J., and Lomi, A. 2019. "Reliability of Relational Event Model Estimates under Sampling: How to Fit a Relational Event Model to 360 Million Dyadic Events," *Network Science*, pp. 1–39. (https://doi.org/10.1017/nws.2019.57).
Lu, Y., Singh, P. V., and Sun, B. 2017. "Is a Core-Periphery Network Good for Knowledge Sharing? A Structural Model of Endogenous Network Formation on a Crowdsourced Customer Support Forum," *MIS Quarterly* (41:2), pp. 607–628. (https://doi.org/10.25300/MISQ/2017/41.2.12).
Lusher, D., Koskinen, J., and Robins, G. 2012. *Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications*, Cambridge University Press.
de Matos, M. G., Ferreira, P., and Krackhardt, D. 2014. "Peer Influence in the Diffusion of IPhone 3G over a Large Social Network," *MIS Quarterly* (38:4), pp. 1103–1134. (https://doi.org/10.2307/26627964).

McFadden, D. 1974. “Conditional Logit Analysis of Qualitative Choice Behavior,” *Frontiers in Econometrics*, pp. 105–142.
Monge, P. R., and Contractor, N. S. 2003. *Theories of Communication Networks*, Oxford University Press.
Oinas-Kukkonen, H., Lyytinen, K., and Yoo, Y. 2010. “Social Networks and Information Systems: Ongoing and Future Research Streams,” *Journal of the Association for Information Systems* (11:2), pp. 61–68. (https://doi.org/10.17705/1jais.00222).
Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J., and Barabási, A.-L. 2007. “Structure and Tie Strengths in Mobile Communication Networks,” *Proceedings of the National Academy of Sciences* (104:18), pp. 7332–7336.
Quintane, E., and Carnabuci, G. 2016. “How Do Brokers Broker? Tertius Gaudens, Tertius Iungens, and the Temporality of Structural Holes,” *Organization Science*.
Quintane, E., Conaldi, G., Tonellato, M., and Lomi, A. 2014. “Modeling Relational Events A Case Study on an Open Source Software Project,” *Organizational Research Methods* (17:1), pp. 23–50.
Quintane, E., and Kleinbaum, A. M. 2011. “Matter over Mind? E-Mail Data and the Measurement of Social Networks,” *Connections* (31:1), pp. 22–46.
Singh, P. V., Tan, Y., and Mookerjee, V. 2011. “Network Effects: The Influence of Structural Capital on Open Source Project Success,” *MIS Quarterly* (35:4), pp. 813-A7.
Smith, E. B., Brands, R. A., Brashears, M. E., and Kleinbaum, A. M. 2020. “Social Networks and Cognition,” *Annual Review of Sociology* (46:1), Null. (https://doi.org/10.1146/annurev-soc-121919-054736).
Smith, E. B., Menon, T., and Thompson, L. 2011. “Status Differences in the Cognitive Activation of Social Networks,” *Organization Science* (23:1), pp. 67–82. (https://doi.org/10.1287/orsc.1100.0643).
Snijders, T. A. B., Koskinen, J., and Schweinberger, M. 2010. “Maximum Likelihood Estimation for Social Network Dynamics,” *The Annals of Applied Statistics* (4:2), p. 567.
Stadtfeld, C. 2012. *Events in Social Networks: A Stochastic Actor-Oriented Framework for Dynamic Event Processes in Social Networks*, KIT Scientific Publishing.
Susarla, A., Oh, J.-H., and Tan, Y. 2011. “Social Networks and the Diffusion of User-Generated Content: Evidence from YouTube,” *Information Systems Research* (23:1), pp. 23–41. (https://doi.org/10.1287/isre.1100.0339).
Sweet, T. M., Thomas, A. C., and Junker, B. W. 2013. “Hierarchical Network Models for Education Research Hierarchical Latent Space Models,” *Journal of Educational and Behavioral Statistics* (38:3), pp. 295–318.
Treem, J. W., and Leonardi, P. M. 2013. “Social Media Use in Organizations: Exploring the Affordances of Visibility, Editability, Persistence, and Association,” *Annals of the International Communication Association* (36:1), pp. 143–189. (https://doi.org/10.1080/23808985.2013.11679130).
Vial, G. 2019. “Reflections on Quality Requirements for Digital Trace Data in IS Research,” *Decision Support Systems* (126), p. 113133. (https://doi.org/10.1016/j.dss.2019.113133).
Wasserman, S., and Faust, K. 1994. *Social Network Analysis: Methods and Applications*, Cambridge University Press.
Wooldridge, J. M. 2009. *Introductory Econometrics: A Modern Approach*, (4th ed.), Mason, OH: South Western, Cengage Learning.

Wuchty, S., and Uzzi, B. 2011. “Human Communication Dynamics in Digital Footsteps: A Study of the Agreement between Self-Reported Ties and Email Networks,” *PloS One* (6:11), p. e26972.
Yang, M., Adomavicius, G., Burtch, G., and Ren, Y. 2018. “Mind the Gap: Accounting for Measurement Error and Misclassification in Variables Generated via Data Mining,” *Information Systems Research* (29:1), pp. 4–24. (https://doi.org/10.1287/isre.2017.0727).