

A Robust Inference Method for Decision Making in Networks

Aaron Schechter¹, Omid Nohadani², and Noshir Contractor²

¹Department of Management Information Systems, University of Georgia, Athens, GA 30602

²Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL 60208

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 likely perceive the network's structure differently. 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. Using robust maximum likelihood, we derive an estimation technique which immunizes inference to errors, without knowing a priori the source or realized magnitude of the error. We demonstrate the efficacy of our methodology on both a classic social network dataset and simulated data. Our results indicate that the approach successfully assigns zero weight to measures which are not perceived by actors in the network, while identifying effects that were perceived. Further, tests on real data indicate that we can improve model fit by a factor of up to five in the presence of one-percent measurement error.

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

A Robust Inference Method for Decision Making in Networks

INTRODUCTION

Social network analysis is an increasingly popular tool which is grounded upon the investigation of diverse relationships between various social entities (Monge and Contractor 2003; Wasserman and Faust 1994). The application of social network analysis to understand organizational phenomena has become increasingly viable thanks in part to the proliferation of online network data in all facets of life (Contractor, 2018; Kane et al. 2014; Lazer et al. 2009). 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 social network sites like Facebook or Twitter, posts to online forums such as Reddit and GitHub, clickstream data, electronic transactions, and mobile app usage. Many of these 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 connected (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; Wasko and Faraj 2005). The advantage of online observable network 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, online networks are cheaper to collect than surveys, 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 exist and those likely to work well in a team (Leonardi and Contractor, 2018).

However, there is a caveat to the use of social network data collected from online sources when our goal is to determine why people behave in certain ways. When we observe links through online data collection, we assume that they represent a form of “ground truth.” However, social network research has long been aware that in some cases the “ground truth” is not the actual observed network but the networks as they are perceived by actors in the network (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 W. I. Thomas, more recently network scholars such as Pattison (1994) and Krackhardt (1987) observed that network “perceptions are real in their consequences even if they do not map one-to-one onto observed behaviors” (p. 128). For researchers drawing on such theories, a discrepancy between observed and self-reported measures would suggest that perceived (or self-report) measures are the “ground truth” and there is a measurement error in using observed communication network data.

Much of recent empirical research using digital trace network data assumes, incorrectly, that these data are equivalent to individuals' perception of the network. Indeed, individuals' perceived ties with others in the network, as well as their perceptions of ties among others in the network, comprise a cognitive social structure (CSS, Krackhardt 1987). Recent research comparing email networks (Johnson et al. 2012; Quintane and Kleinbaum 2011; Wuchty and Uzzi 2011), mobile phone proximity data (Eagle et al. 2009), and mobile phone call records (Onnela et al. 2007) find that while these reconstructed networks based on digital data correlate

with the underlying perceived social networks – as reported by the participants – they are at best approximations. Accordingly, there is a definitive gap – what we call, pardon the pun, a *perception gap* – between network data which is collectively observable and treated as truth, and the networks that are individually perceived, i.e., CSSs.

This perception gap is a particularly salient problem when network inference is conducted. 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, Menon and Thompson, 2012), the cause is derived from an individual’s goals, preferences, and their perspective on the network’s current state (Kilduff and Brass 2010). This view contrasts with the assumption that the observable network data is indeed ground truth when it comes to explaining individuals’ behaviors based on their perceptions of the network.

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? And, if they are not, can we develop robust inference techniques that have the capacity to identify variables which are significantly non-zero, 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? To address this question, we first review studies which make

use of digital network data and discuss their assumptions with regards to perception. In doing so we establish the pervasiveness of the identified gap. Second, we determine the relevant sources of discrepancy between these networks. In particular, we identify two factors: (1) the technical features of online networks which can cause errors, which include size, rate of change, and visibility; and (2) the cognitive factors which create individual bias, such as tendencies towards compression, personality traits, and network position. Finally, we propose a robust reformulation of inferential network models to address these various sources of error. Our methodology leverages recent advances in the field of robust optimization, specifically robust maximum likelihood estimation, to common models such as exponential random graph modeling (ERGM), stochastic actor-oriented models (SAOM), and relational event models (REM). We derive a general form of a predictive model for social networks which is an extension of Bertsimas and Nohadani (2018) applied to social network analysis. As a proof of concept, we test our model on the CSS collected by Krackhardt (1987) for his foundational study. Additionally, we conduct simulation experiments to more rigorously test the efficacy of the method.

Our contributions are thus three-fold: first, we identify a problem in the way that social network data is analyzed, given a variety of technical and psychological factors; second, we propose a method that remedies this issue by immunizing parameters from data errors using robust optimization principles; and third, we demonstrate the potential for new theoretical and practical insights, particularly in scenarios where individual decision-making is predicated on network structure.

BACKGROUND & MOTIVATION

Despite the potential limitations of online social network data matching perceptions, these forms of data are frequently used to test hypotheses about human attitudes and behavior. In

order to demonstrate the pervasiveness of online network data in management studies, we conducted a survey of recent publications in top journals, including but not limited to MIS Quarterly, Information Systems Research, and Management Science. Appendix 1 provides a representative sample of studies over the past 10 to 15 years using digital network data.

Some of these studies rely on network data that is *explicitly visible* to individuals in the population. For example, threaded replies in community forums enable individuals to see exactly who has replied to whom, and how often someone has posted (Faraj and Johnson 2011; Johnson et al. 2014). Networks are also more likely to be visible on email listservs, where the entire group sees every message (Kudaravalli and Faraj 2008; Singh and Tan 2010). A second set of studies rely on network data that is *implicitly visible* to individuals in the population. These are platforms where network links are inferred between individuals by virtue of their affiliation with, or joint contribution(s) to, some entity, such as a software project. For example, networks constructed from revisions to software, e.g. open source projects, are visible due to the implementation of version control systems (Grewal et al. 2006; Singh et al. 2011). Because every activity is logged, a developer could determine who interacted with what section of code over time (Conaldi and Lomi 2013). However, perceptions of these third-party ties are less likely to be accurate, given that an individual would have to determine if there are any shared projects between every other pair of individuals in the network. Accordingly, these flawed perceptions of networks may not accurately explain behavior. As we discuss later, the extent to which online interactions are explicitly or implicitly visible to participants in the network is related to the visibility affordance of the technology (Treem and Leonardi, 2012).

Indeed, most network studies which employ digital trace data assume that the observable networks match the perceived networks when testing their hypotheses and interpreting their

results, regardless of whether the networks are explicitly or implicitly visible. Consider for example analyses of brokerage strategies (Quintane and Carnabuci 2016; Spiro et al. 2013) and structural holes (Zaheer and Soda 2009). These studies are predicated on the fact that individuals are perfectly aware of third-party relationships and potentially alter the network structure to take advantage of brokerage opportunities. Similarly, studies of interactions on MOOC forums (Vu et al. 2015) and technical question and answer communities (Stadtfeld and Geyer-Schulz 2011) assume that actors are able to make decisions based on the prior activity of other individuals, as well as third-party interactions. Their interpretation of these hypothesized behaviors assumes accurate perception of relations, both structurally and temporally.

In order to address this perception gap, it is important to determine what factors may cause a misalignment between observable and perceived networks. In the following sections, we detail two sources of discrepancies – technical and cognitive – which cause perceived networks to diverge from observable networks. We then briefly describe how these deviations can impact statistical inference.

Technical Sources of Discrepancies

Online digital interactions may be technically visible to participants. However, as we discuss in this section, 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. For instance, 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). 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 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). Thus, even ties built on trust or frequent interaction can fade – at least temporarily – in an individual's consciousness. Networks based on affiliation, such as those connecting open source software developers, also tend to be quite large, with hundreds or thousands of projects and developers. In this case, developers may have indirect connections to hundreds of other individuals (Singh et al., 2011); this scale can be problematic when considering mechanisms such as triadic closure or preferential attachment. A large-scale study of mobile phone usage indicated that people generally only communicate with a small handful of other individuals (Onnela et al., 2007), suggesting that most individuals make use of only a small portion of their potential connections. Indeed, there is significant empirical evidence indicating that people apply a variety of compression heuristics to make sense of their networks (Brashears and Quintane 2015; Gigerenzer and Brighton 2009; Quintane and Kleinbaum 2011).

A second feature of online network data which impacts 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. Of course, there is a degree of variability here; some interactions may be long and recur frequently during a time period, while others may be short, intense periods of interaction. Indeed, 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 (Vu et al., 2015). Thus, making inference on behavior based on the aggregated network structure can misrepresent the relationships in the network (Quintane et al. 2014). Further, there

is an issue of memory in networks. While many studies focus on the network at one point in time, individuals tend to react differently to long-term trends and recent occurrences. People are embedded in long-term patterns, which constitute regular behaviors and routines; on the other hand, more ephemeral short-term interactions may cause deviations from those trends (Quintane et al. 2013; Quintane and Carnabuci 2016).

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 and the meta-knowledge embedded in it (Treem and Leonardi, 2012). Treem and Leonardi (2012) identify a number of features which can improve the visibility of social media networks, including lists of friend connections, recommendation algorithms, status updates, and personal profiles. 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, 2012). For instance, social media networks such as Facebook or Twitter allow users to see “who is friends with whom” or “who follows whom,” etc. These platforms also provide recommendations on who to add as a friend based on existing connections. 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). On an online forum, users are able to scroll through posts and determine relationships vis-à-vis replies to comments (Faraj and Johnson, 2011). Yet, as is the case in social media networks, users enact these affordances to varying degrees. Further, this information can be presented in different ways across different sites. Thus, the fidelity of network perceptions will in part be dependent on the affordances of the digital platform from which information is collected.

Cognitive Sources of Discrepancies

In addition to the technical sources discussed in the previous section, there are cognitive sources that contribute to discrepancies between observed networks and individuals' perceptions of these networks. These include errors in recalling individual relations, a tendency to simplify relational patterns, and a bias towards closed triads. This tendency stems from a variety of factors, such as a tendency to believe the world is a "small-world" (Kilduff et al. 2008). Likewise, humans have demonstrated a knack for remembering clusters of relations, but perform poorly when asked to recall specific relationships (Brashears and Quintane, 2015; Quintane and Kleinbaum, 2011). Accordingly, even in the case of perfectly observable data, people will cognitively perceive networks that diverge from those data.

Early studies on informant accuracy and recall focused on the ability of individuals to correctly report what they had witnessed or experienced (Bernard and Killworth, 1977; Bernard, Killworth and Sailer, 1980, 1982; Bernard, Killworth and Cronenfeld, 1984; Freeman, Romney and Freeman, 1987; Krackhardt, 1987; Heald, Contractor, Koehly, and Wasserman, 1998). In general, individuals had a difficult time recalling their own interactions, as well as observed interactions among others. This research helped spawn the notion that what one perceives, and subsequently reports, can diverge significantly from external observation. One prominent set of theories offered to explain this discrepancy points to the presence of a systematic bias towards structures that reduce cognitive stress. Freeman et al. (1987) found that informants tend to default to the long-term patterns they observed; for instance, they are more likely to think someone was present at a meeting if they generally had good attendance. Further, individuals tend to view themselves as being more central, and that there are more ties, more reciprocation,

and more transitivity among their reported friends (Krackhardt and Kilduff 1999; Kumbasar et al. 1994).

Individuals tend to use basic schema or categorizations to classify the relations around them, particularly those of which they are unsure (Freeman 1992; Freeman and Webster 1994). Indeed, the theory of foci (Feld, 1981) argues that individuals form relationships through a shared focus on something external to the network itself; i.e., the foci act as collecting forces. Further, Feld (1981) argues that in the absence of a shared foci, individuals will construct one to categorize their relationships. Corman and Scott (1994) expound upon this idea, arguing that foci create the underlying cognitive processes which lead to interactions, which subsequently change or reinforce perceptions of the network. Essentially, humans utilize aggregating mechanisms to make sense of their social relationships. However, as the complexity of the network grows, it becomes increasingly difficult for a person to accurately recall linkages (Burt et al. 2013), and thus they rely on their higher-level affiliations. Accordingly, a consistent finding among these studies is the tendency towards simplicity, and a rejection of structures that are dissonant with the mental models of the individual. For instance, people believe their networks of friends display balance (Heider, 1958), and they tend to view their network as having a small-world, clustered structure (Kilduff et al., 2008). In general, individuals who communicate with one another are more likely to have similar – even if inaccurate perceptions – of the overall network (Heald et al. 1998).

Impact on Inference

In order to determine how statistical inference is affected by discrepancies in perceived versus observable 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*. In other words, there is a high false-positive rate or Type I error in the cognitive social structure. Alternatively, individuals may make the error of ignoring ties that do indeed exist, i.e., an *error of commission*. The prevalence of this mistake is equal to the rate of Type II or false positive errors. Both errors can cause biased inference results.

To illustrate how a misalignment between the CSS and an observable digital network can impact inference, we present two commonly used measures from social network analysis. For example, we may test the tendency for individuals to initiate communication towards individuals who are highly active in the network, a principle known as preferential attachment (Barabási and Albert 1999). However, in practice it is unlikely that the potential initiator of communication is accurately aware of how active other individuals are in the network. They may over- or underestimate that individual's centrality or popularity. We draw another example based on balance theory (Heider 1958), which suggests that an actor i is more likely to have a network tie with an actor k if actor i has a network tie with actor j and actor j has a network tie with actor k . Implicit in this argument is the assumption that actor i is cognitively – and accurately - aware of the presence or absence of a network link from actor j to actor k . Such a correspondence between observable and perceived network ties cannot be assumed.

Technical features of online networks such as scale, rate of change, and varying visibility can lead to both types of errors. In an electronic network, linkages between individuals may be difficult to perceive, and the large scale makes accurate recall difficult. Consequently, individuals will be likely to have a higher false negative rate (Brashears 2013; Brashears and Quintane 2015). On the other hand, individuals embedded in a network derived from an online forum, for example, may assume that two individuals are linked because they contribute to the

same thread. However, those two people may never directly communicate. Thus, the perceived link may not actually exist, or its strength may be overestimated.

Cognitive factors have commonly been linked to false positive errors. Prior work has demonstrated a tendency to perceive closed triads and a locally clustered network structure (Kilduff et al. 2008). Indeed, people tend to believe that their friends are linked to one another, and that there is a shorter path connecting themselves to anyone else in the network. On the other hand, individuals also show a propensity to recall long-term trends while forgetting more short-term or unusual linkages (Freeman et al. 1987). The tendency to use compression heuristics also leads individuals to perceive group structures relatively well, while misattributing the state of individual ties (Brashears 2013; Brashears and Quintane 2015; Gigerenzer and Brighton 2009; Quintane and Kleinbaum 2011). Accordingly, humans are also prone to false negative errors.

THE ROBUST INFERENCE APPROACH

Given that both technical and cognitive features can contribute to false positive (errors of commission) as well as false negative errors (errors of omission), 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 able to correct for 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 which the data input is noisy, incomplete, or erroneous (for an overview, see Ben-Tal et al. 2009; Ben-Tal and Nemirovski 2002; Bertsimas et al. 2011). This branch of research provides computationally tractable methods for solving real world problems in which data or parameters are inherently uncertain. In the case of models for social network inference, data uncertainty takes the form of individual perceptions of the surrounding network. Specifically, the model assumes that the data accurately reflects the information that is used by each actor to make decisions. However, perception errors, hidden information, or poor memory could lead individuals to make choices based on an erroneous set of data which is not observable to the researcher. As a result, the estimated parameters from the base model may not accurately reflect the impact of each network structure on decision making. The elements of a robust optimization problem include: nominal or original data; an uncertainty set, which specifies the range over which the nominal data may vary; an objective function, which is a deterministic combination of model data and parameters that the researcher would like to maximize or minimize.

Notation and Definitions

Throughout the rest of this paper we will denote vectors with a lowercase bold letter, and matrices with a capitalized 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, and the cardinality of the set is $|\mathcal{A}_t| = N_t$. 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 $\mathbf{X} \in \mathbb{R}^{M \times N \times P}$. 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 $\mathbf{x}_{t\bullet} \in \mathbb{R}^P$ 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 “true” information perceived by individuals and our observable network. Thus, while we observe \mathbf{X}^{obs} , that may not be the true value perceived by the actor. We thus represent our data as

$$\mathbf{X}^{\text{true}} = \mathbf{X}^{\text{obs}} - \Delta \mathbf{X}, \quad (1)$$

where \mathbf{X}^{true} is the matrix of features perceived by the individuals within the network. Networks and changes to the networks are implicitly modeled as reactions to the underlying *true* network structure, i.e., the network that individuals perceive. However, we consider here the case where only have access to an *observable* network which is collected through digital trace methods such as email or mobile data. Features derived from this network are represented by \mathbf{X}^{obs} . A consequence of this is an inherent bias in the variables we use to conduct inference. This

discrepancy is captured by $\Delta\mathbf{X} \in \mathbb{R}^{M \times N \times P}$, which represents the difference between the features we observe and the features that are perceived by individuals in the network. 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.

Given that information about individual-level biases and error distributions are typically not available, we take a conservative approach to modeling the discrepancies. We assume that, in general, we have no information about the nature of the errors and 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

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

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.

Computing Robust Estimators

Methods for social network inference generally rely on the assumption that the relative likelihood of a given tie existing is log-linear by functional form (Holland and Leinhardt 1977,

1981). The authors proposed an exponential family of distribution as an ideal fit to the problem at hand; sufficient statistics and parameters should be linked together and determine the probability of a particular configuration (Holland and Leinhardt, 1981 pg. 36). Extending this notion, we may also consider the relative likelihood of an agent making a network change as having log-linear form. This theory does not distinguish between perceived versus observable networks, but rather provides a general density for dichotomous graph structure with parameters $\boldsymbol{\beta} \in \mathbb{R}^P$. 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:

$$f_t(\boldsymbol{\beta}; \mathbf{x}_{ty\bullet}^{true}, \mathcal{A}_t) = \frac{\exp(\boldsymbol{\beta}' \mathbf{x}_{ty\bullet}^{true})}{\sum_{z \in \mathcal{A}_t} \exp(\boldsymbol{\beta}' \mathbf{x}_{tz\bullet}^{true})}. \quad (3)$$

This parameterization is consistent with common social network inference models such as ERGMs (Robins et al., 2007), stochastic actor-oriented models (Snijders, 1996), and relational event models (Butts 2008). 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). Inference based on the MLE procedure or derivations of it are a common method for fitting models and performing hypothesis testing for social network data (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, e.g. 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 \mathbf{X}^{obs} reflecting the changing network. We assume the presence of some error in our dataset, i.e., $\mathbf{X}^{obs} \neq \mathbf{X}^{true}$, and so we recast the likelihood of the full sequence of network events as:

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

Following the formulation of (Bertsimas and Nohadani, 2018), we define the log-likelihood function

$$\psi(\boldsymbol{\beta}; \mathbf{X}^{obs} - \Delta \mathbf{X}) = \log \left(\prod_{t=1}^M f_t(\boldsymbol{\beta}; \mathbf{x}_{ty\bullet}^{obs} - \Delta \mathbf{x}_{ty\bullet}, \mathcal{A}_t) \right). \tag{5}$$

Now, the maximum likelihood problem we considered previously becomes the following constrained optimization problem:

$$\max_{\boldsymbol{\beta}} \{\psi(\boldsymbol{\beta}; \mathbf{X}^{obs} - \Delta \mathbf{X}) : \Delta \mathbf{X} \in \mathcal{N}\}. \tag{6}$$

It follows that the solution to the MLE problem (6) must be a valid solution for any of the errors deemed possible by the choice of uncertainty set \mathcal{N} . Consequently, a solution that satisfies this constraint is the parameter vector which also fits to the log-likelihood function under the worst-case errors. Hence, the robust estimator is also the solution to the following robust optimization problem:

$$\max_{\boldsymbol{\beta}} \min_{\Delta \mathbf{X} \in \mathcal{N}} \psi(\boldsymbol{\beta}; \mathbf{X}^{obs} - \Delta \mathbf{X}). \tag{7}$$

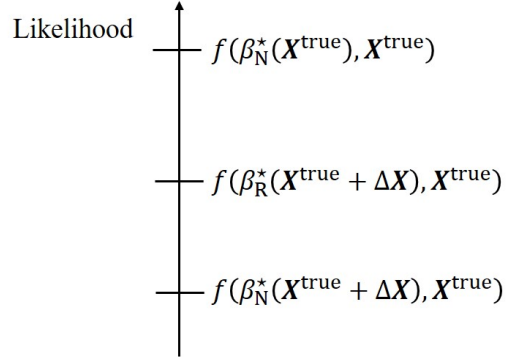
For details on how we compute the solution to this problem, as well as relevant proofs and algorithms, please refer to Appendices 2, 3, and 4.

Interpretation of Robust Estimators

The goal of robust inference is to estimate a set of parameters to the nominal data that will still provide a reasonable estimate of data which differs from what we used to tune the model. In other words, if we assume that our observable data does not in fact match the reality as

perceived by individuals in the network, the nominal estimators could change significantly, whereas the robust estimators will remain stable. We present Figure 1 to illustrate how robust estimators compare to nominal estimators.

Figure 1. Relative likelihoods of robust and nominal parameters



Notes. The likelihood function of the true data as perceived by the individuals in the network as a function of (a) the nominal parameters (β_N^) fit to the true data, (b) parameters (β_R^*) fit to observed data, and (c) nominal parameters (β_N^*) fit to the observed data.*

To show the benefits of the robust estimators β_R^* , we compare its likelihood $f(\beta_R^*)$ to that of standard estimators $f(\beta_N^*)$, which assumes data to follow the nominal distribution. As input, we use true data X^{true} and observable data X^{obs} , as discussed earlier. In general, it is expected that $f(\beta_N^*(X^{\text{obs}})) < f(\beta_N^*(X^{\text{true}}))$ because the assumptions are no longer met for $f(\beta_N^*(X^{\text{obs}}))$.

However, robust estimators are immune to assumption deviations, hence we observe

$f(\beta_N^*(X^{\text{obs}})) < f(\beta_R^*(X^{\text{obs}})) < f(\beta_N^*(X^{\text{true}}))$, i.e., the robust estimators outperform nominal estimators for observed data and cannot reach the optimality of $\beta_N^*(X^{\text{true}})$ due to lack of accurate information. In summary, robust parameters will better predict behaviors in a network than nominal parameters if the perceived network and observable network differ, but will always perform worse than the hypothetical optimum.

MODEL DEMONSTRATION & TESTING

To assess the viability of our robust approach, we conduct a series of tests on both existing and simulated data. We first apply our method to data derived from an aggregation of surveys that were collected by Krackhardt (1987) for his original CSS study. A network created by combining sociometric surveys represents a reasonable approximation for how a single observable network compares to individual perceptions; while related, the overall network almost certainly is distinct from how one person views it. An additional motivation for using this dataset is that we are able to compare the aggregate network to individual perceptions and can assess performance on deviations in a real scenario. Second, we conduct a simulation study to test our method more generally. In that step, we create sequences of networks where changes are motivated by individual network perceptions. Thus, we can generate “observable” data that is intentionally different from the “perceived” data and test our method’s performance compared to traditional maximum likelihood.

Empirical Illustration

Dataset & Method

We illustrate the efficacy of the robust approach on a classic network dataset taken from Krackhardt’s (1987) study on CSSs from managers in a small manufacturing firm. In these data, a sociometric survey was conducted for each of 21 employees, and each was asked to report their perceptions of the friendship as well as the advice network 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 $Y^{(i)}$, $i = 1, \dots, 21$. On average, the density of these networks is 0.1247 with a standard error of 0.0782. 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 $Y_{ij}^{\text{LAS}} = 1$ only if $Y_{ij}^{(i)} = 1$ and $Y_{ij}^{(j)} =$

1. The LAS graph can be thought of as the “true” data for our analysis; this aggregate network is a representation of the overall structure. This method of aggregation, as well as others, can be found in multiple reviews (Batchelder et al. 1997; Siciliano et al. 2012). Our aggregate network Y^{LAS} has 35 edges – a density of 0.1667 – and 12 total triangles.

For this analysis we fit an exponential random graph model to the LAS network using the approach described by Hunter et al. (2008). The authors demonstrated that instead of maximizing the likelihood directly, they can obtain a solution by maximizing the log-ratio of likelihood values, where the baseline likelihood is computed from an arbitrary parameter vector β_0 . We thus use the following expression:

$$\ell(\beta) - \ell(\beta_0) \approx (\beta - \beta_0)' \mathbf{g}(\mathbf{y}_{\text{obs}}) - \log \left(\frac{1}{m} \sum_{i=1}^m \exp((\beta - \beta_0)' \mathbf{g}(\mathbf{Y}_i)) \right) \quad (8)$$

Here, $\ell(\beta)$ is the likelihood of the network under the parameters β . The quantity $\mathbf{g}(\mathbf{y})$ is a vector of sufficient statistics computed from the graph \mathbf{y} . We include two statistics in our model, edges and triangles. Though the triangle statistic often leads to degeneracy and in practice can be replaced with a shared-partner statistic (c.f. Hunter and Handcock 2006), we use it for simplicity of demonstration. In this expression, we are taking the sum over the weights for a large sample of m networks simulated from the distribution of β_0 . Using the `ergm` (Hunter et al., 2008) and `statnet` (Handcock et al. 2008) packages, we obtain the maximum pseudo-likelihood estimator (MPLE) β_0 and subsequent likelihood $\ell(\beta_0)$ for the edge-triangle model of the LAS network. We then apply our subgradient descent approach as outlined in Algorithm 1 to find the best value of β in the nominal and robust case.

Results

Using the `ergm` package, we determined the MPLE for the LAS network. We then applied our method using uncertainty sets of size $\rho = 0$, i.e., the nominal case, $\rho = 0.1$, $\rho = 0.5$, and $\rho = 1.0$ on this aggregate network. The varying values of ρ correspond to an increasing tolerance for error that is built into the parameter estimates. The results for each model, as well as the log-likelihoods, are presented in Table 1.

Table 1. Parameter estimates and model fits for the LAS network

Variable	Nominal MLE ($\rho = 0$)	Robust MLE ($\rho = 0.1$)	Robust MLE ($\rho = 0.5$)	Robust MLE ($\rho = 1.0$)
Edges	-2.89	-2.86	-2.50	-2.14
Triangles	3.23	3.28	3.39	3.47
Log-Likelihood	-78.57	-78.78	-79.37	-80.76

Notes. The values provided are estimates of β computed by applying Algorithm 1 to Eq. 8, and reporting the subsequent likelihood. The value of ρ corresponds to the size of the global uncertainty set. Larger values of the log-likelihood indicate a better fit to the observed network.

We observe that as the size of the uncertainty set increases, the values of the parameters for both edges and triangles increase. Substantively, this finding suggests that CSSs which diverge significantly from the LAS are likely to have more edges and triangles than the overall aggregate network. Such a conclusion is generally in agreement with prior research on cognitive biases, such as the tendency to believe that your friends are also friends with one another (Krackhardt and Kilduff, 1999; Kumbasar et al., 1994). Further, we note that the log-likelihood values decrease (become more negative) as the value of ρ increases. Put simply, as we increase the robustness of the parameters, the quality of the fit on the LAS network suffers. This result is consistent with the relationship we presented in Figure 1.

The second step of our analysis is to apply both the nominal and robust estimators from the aggregate network to the cognitive social structures provided by each of the individuals in the

survey. However, each CSS is an “observable” dataset that in principle could differ from the true structure. Following the logic from Figure 1, the robust parameters should provide a better fit to these individual networks than the nominal MLE. To assess the relative performance of robust parameters against nominal parameters, we compute a likelihood ratio. If the ratio is greater than 1, we may conclude that the robust parameters better reflect the structure of the CSS than the nominal MLE parameters.

To demonstrate how our method works on a network which is as different from the LAS as possible, we first select the network which was provided by manager number 7. That individual provided 39 false positives and 13 false negatives, which were by far the most in our sample. To demonstrate the effectiveness of the robust estimators on an average network, we consider the response of individual 16, whose errors are similar to the median. We then compute the likelihood ratio comparing the nominal estimator to our three robust estimators for both the worst and average-case networks and present the values in Table 2.

Our findings indicate that the robust parameters are a significantly better fit to individual 7’s network than the MLE derived from the aggregate network ($LR_{\rho=0.1} = 5.39$), and the performance improves as the error tolerance increases ($LR_{\rho=0.5} = 92.70$; $LR_{\rho=1.0} = 309.61$). For the average case, the robust parameters are also a better fit than the nominal parameters. Similar to the worst-case, our analysis of individual 16’s network revealed that the likelihood ratio for the lowest error tolerance is greater than 1 ($LR_{\rho=0.1} = 1.18$), and the performance improves as the error tolerance increases ($LR_{\rho=0.5} = 2.00$; $LR_{\rho=1.0} = 2.63$). This finding shows that the robust estimators are less sensitive to which individual is selected than the standard MLE, and therefore better represent the overall behaviors.

Table 2. Likelihood ratio values for the worst-case and average-case CSS

Uncertainty Level	Likelihood Ratio (Worst Case)	Likelihood Ratio (Average Case)
$\rho = 0.0$	1.00	1.00
$\rho = 0.1$	5.39	1.18
$\rho = 0.5$	92.70	2.00
$\rho = 1.0$	309.60	2.63

Notes. The likelihood ratio is given for the robust estimator at all three levels of error. Values greater than 1 indicate that the robust model is a better fit to the CSS data.

Simulation Experiments

Data Generation & Method

To more rigorously evaluate the effectiveness of our robust estimators and demonstrate their utility, we perform experiments on a set of computer-generated data. We generate a set of simulated sequences of networks which 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 is built to incorporate errors, 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 . Then, we test the performance of the robust estimators against the nominal estimators on the contaminated data. The overarching goal of our experiments is to determine four elements of performance: (1) if a cognitive effect is not present, the model should correctly eliminate it; (2) if a cognitive effect is present, then the model should find a non-zero estimator, even in the presence of data contamination; (3) the overall likelihood for a sequence should be higher for the robust model when errors are present; (4) the robust model should assign a higher probability to actual events when errors are present.

We generate 5,000 changes between 50 actors who are randomly assigned to one of two groups for purposes of determining homophily. This process is repeated for 10 different

groupings. In each network, the mechanisms of change are varying individual activity rates, reciprocity, and a preference for homophily. For every synthetic network, we randomly assign each actor to one of two groups. Each individual is also given an activity rate drawn from a power-law distribution with parameter $\gamma = 2$. To simulate an action sequence, we randomly draw a sender from the network per the relative rates. The target is then randomly selected from the remaining actors. The influence of homophily is set so that 75% of events occur between individuals who share a group.

From each synthetic sequence, we compute our matrix of network features \mathbf{X}^{true} . We choose to include four commonly used characteristics: activity rate, reciprocity, homophily, and transitivity. For an illustration of these network statistics, please refer to the Appendix 5. 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 that individuals will not be able to properly determine the frequency with which other actors communicate with one another. Therefore, the network statistic transitivity will be subject to error. For each synthetic dataset, we also generated a matrix of errors $\Delta\mathbf{X}$. We computed errors in two ways. Each value of Δx_{tz4} was drawn randomly from a Uniform distribution on the range $[-1, 1]$. The values of $\Delta x_{tz1} = \Delta x_{tz2} = \Delta x_{tz3} = 0$ to represent the accurate network knowledge. We will determine the efficiency of the robust estimators by scaling up or down the magnitude of these error values and adding them to the true data.

To determine if the model can detect an influence of transitivity *if actors are cognizant of the effect*, we also generate a second set of sequences following the same logic, except in the

second case we include a positive influence of transitivity. In other words, in the second condition individuals are *aware of the value of the two-paths* connecting them to other actors, and are positively influenced by stronger relationships. We include a parameter of $\beta_4 = 1$ for transitivity. The value of this statistic is still subject to the same errors. We calculate the robust estimators for the true data \mathbf{X}^{true} using a tolerance parameter of $\rho = 0$ through $\rho = 1$ with a step size of 0.1. These robust estimators are denoted $\boldsymbol{\beta}_R^*(\mathbf{X}^{\text{true}}, \rho)$. Note that when $\rho = 0$, we have the nominal parameters of the standard MLE approach. We evaluate performance in three ways. First, we compare the values of $\boldsymbol{\beta}_R^*(\mathbf{X}^{\text{true}}, \rho)$ across different error thresholds. Second, we compute the log-likelihood of the observed sample $\psi(\boldsymbol{\beta}_R^*(\mathbf{X}^{\text{true}}, \rho), \mathbf{X}^{\text{true}} + \alpha\rho\Delta\mathbf{X})$, where the observed sample $\mathbf{X}^{\text{obs}} = \mathbf{X}^{\text{true}} + \alpha\rho\Delta\mathbf{X}$ is equal to the true data, plus the error. We use the parameter α to vary the magnitude of the error for each threshold ρ . We use $\alpha = 0.5$, $\alpha = 1.0$, and $\alpha = 1.5$; i.e., 50%, 100%, and 150% of the anticipated error magnitude (i.e., the presumed miscalculation of the social network statistic). Finally, we compute the deviance residuals for each sequence. A deviance residual for a network event is proportional to the log-likelihood for that specific event to occur, given the prior sequence of events. The formula for a deviance residual is

$$r_t(\boldsymbol{\beta}_R^*(\mathbf{X}^{\text{true}}, \rho), \mathbf{X}^{\text{obs}}) = -2 \left(\boldsymbol{\beta}_R^{*'} \mathbf{x}_{ty}^{\text{obs}} - \log \sum_{z \in \mathcal{A}_t} \exp(\boldsymbol{\beta}_R^{*'} \mathbf{x}_{tz}^{\text{obs}}) \right), \quad (9)$$

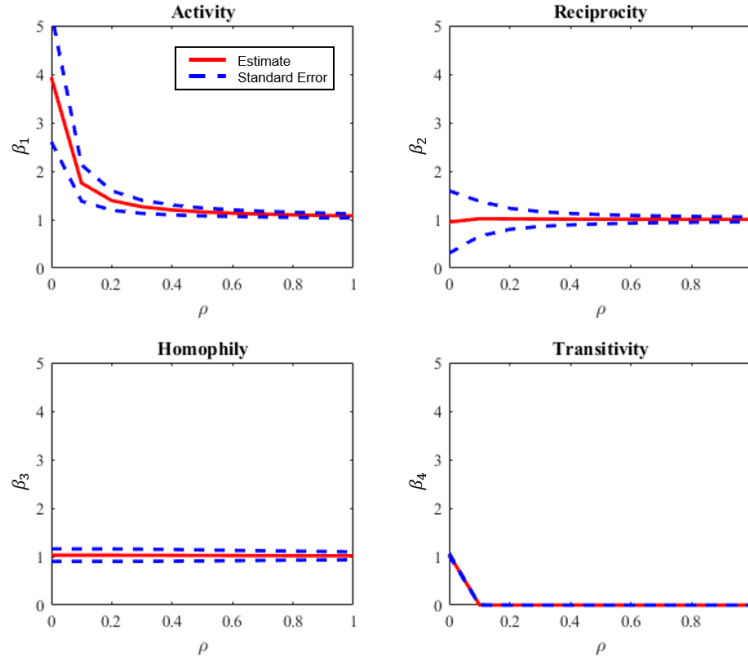
where a smaller value indicates that the model assigned a higher probability to the observed event.

Results

We first explore the results for our parameter estimates across different error values. The plots represented in Figure 2 are the mean values of $\boldsymbol{\beta}$ for each of our four network statistics

across experimental replications, with the dashed lines indicating the standard error of these estimates.

Figure 2. Parameter Estimates for Network Statistics across Error Thresholds

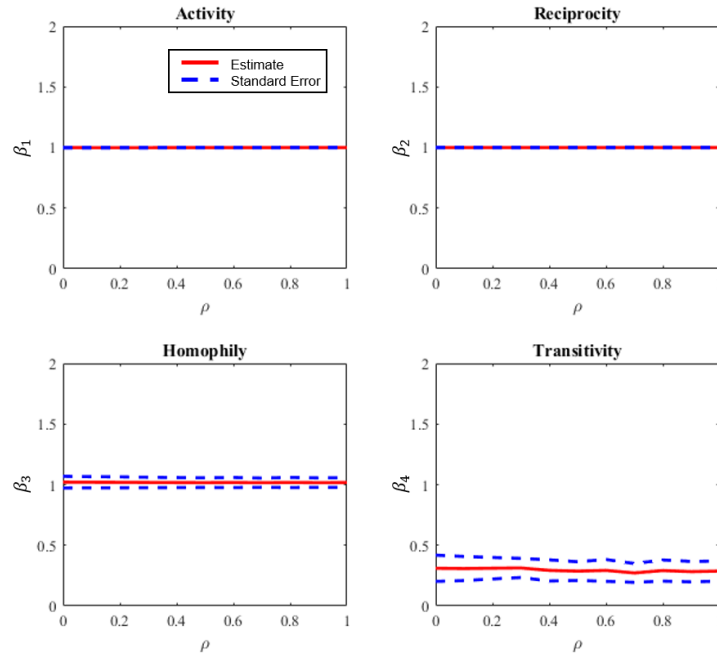


Notes: The vertical axis is the parameter value, while the horizontal axis is the error threshold. The solid red line is the fitted parameter, and the blue dashed line is one standard error.

For each of the three nominal parameters – activity, reciprocity, and homophily – we see relatively stable estimates. As a check of our simulation, the value of the homophily parameter, β_3 was 1.02. This result may be interpreted as the rate of communication between members of the same group relative to those outside the group is $\exp(\beta_3) = 2.77$, which is close to the ratio of 3 we used to generate the networks. Most importantly, we note that the nominal value of β_4 corresponding to transitivity was 1.05. However, upon introducing even a small amount of error, this term quickly drops off to zero. We thus see immediate evidence that the robust formulation will eliminate statistics which are not reliable when subject to perturbation.

An important element of robust inference is the capacity to identify variables which are significantly non-zero, even with contaminated data. In our first step – Figure 2 – our model correctly eliminated the transitivity term. This result corresponds to an effect that is not perceived even though it was “observed” under the assumption of no contamination. However, we would also like to test the scenario in which individuals do perceive third-party ties to some extent. In other words, a decision to form a tie will be motivated by transitivity. As before, we generate a set of synthetic data sequences, but for this exploration we set the true parameter for transitivity to $\beta_4 = 1$ and simulate accordingly. Following the same procedure for fitting models, we present the new parameter estimates in Figure 3.

Figure 3. Parameter Estimates for Network Statistics across Various Error Thresholds with a Transitivity Effect

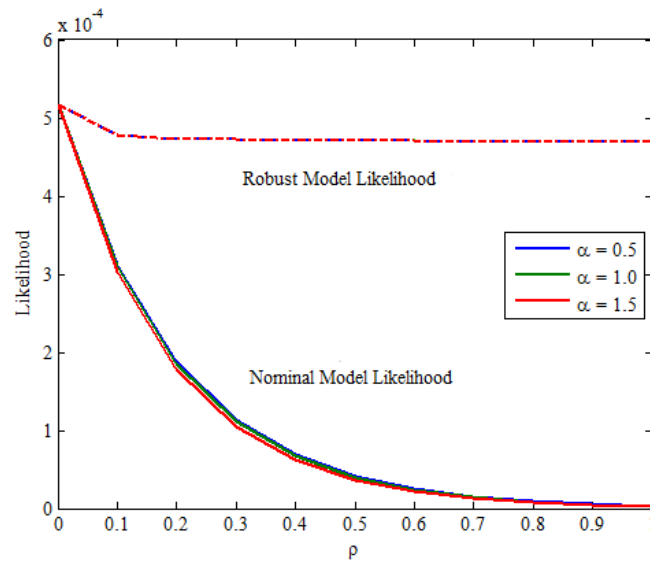


Notes: The vertical axis is the parameter value, while the horizontal axis is the error threshold. The solid red line is the fitted parameter, and the blue dashed line is one standard error.

We observe that as before, the parameter estimates for activity, reciprocity, and the homophily effect are roughly stable around 1. However, in this case the robust model *does find a*

positive impact for transitivity, indicating that even in the presence of data contamination we would conclude that transitivity does play a role in decision making over time. This value is less than 1 (the parameter used in the simulation) however, indicating that the robust estimator *did not capture the full effect*. Referring to our schematic in Figure 1, this is evidence of how the robust estimator will perform better than the nominal data, but not perform as well as the fit to the true data. Figures 2 and 3 together support the conclusion that a robust inference model will correctly eliminate variables which have no effect, and conservatively evaluate variables which do.

Figure 4. Likelihood Function Values for Nominal and Robust Estimates



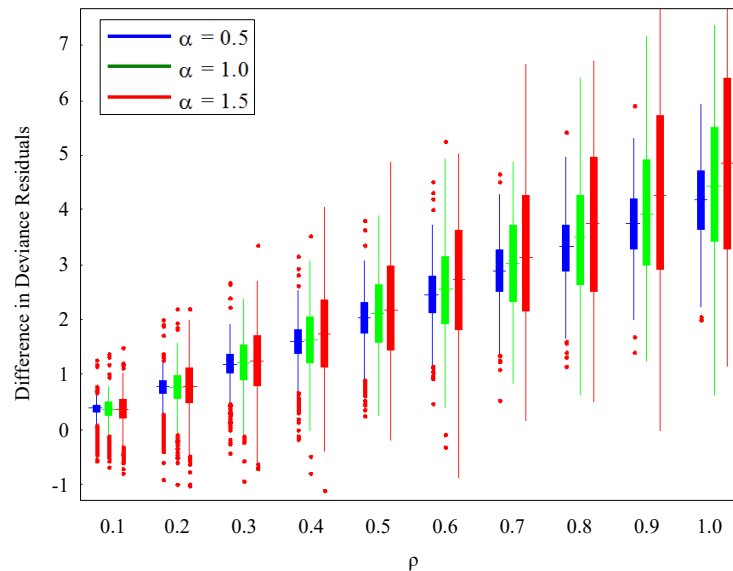
Notes: Blue lines indicate a multiplier of $\alpha = 0.5$, green lines indicate a multiplier of $\alpha = 1.0$, and red lines indicate a multiplier of $\alpha = 1.5$. The vertical axis is the likelihood of the simulated sequence, and the horizontal axis is the error threshold. Dashed lines represent robust models and solid lines indicate nominal models.

We next determine the effect of robust MLE on the likelihood of the observed change event sequence. A higher likelihood for the same sequence of events indicates a better overall model fit, and thus a better representation of the underlying behavioral mechanisms. In Figure 4 we present the likelihood function computed across all error thresholds, and with various

inflation factors α as described previously. From Figure 4, there are two immediate conclusions. First, the robust models – indicated by the dashed lines – outperform the nominal models in all cases where error is present. The difference between nominal and robust is enhanced as more error is introduced into the model. Second, we note that regardless of the value of α , the overall robust likelihood value is the same, and thus the lines overlap. This result suggests that even if errors are larger than anticipated, the robust model still provides a good fit to the data. Overall, we find a significant improvement in model fit with robust estimates in the presence of any error. Finally, we explore the deviance residuals for both nominal and robust models. The deviance residuals characterize how well each individual change is predicted. For all 5,000 events in our simulated dataset, we computed deviance residuals from (5) for the nominal and robust model at each error threshold. We subtracted the residual for the robust model from the residual of the nominal model. A positive value for the difference indicates that the robust model gave the event a higher probability of occurring than the nominal model and vice versa. We collected these differences and represent them in boxplots for each value of the error threshold in Figure 5. From Figure 5, we first observe that for the vast majority of data points, the robust model is more accurate in determining the likelihood of the actual event. Further, the average discrepancy increases as more error is introduced to the dataset. Consequently, in the presence of data contamination, robust parameter estimates are much better suited to predict individual network events. As expected, the variance in performance generally increases as the error multiplier increases. Additionally, while there are cases in which the nominal model outperforms the robust model, these instances constitute less than 50% of cases in the absence error, and less than 25% of error in the presence of significant contamination. We thus conclude that in the presence of

data error, robust models are more effective at predicting single events than nominal parameters, as well as constituting better fits to the data overall.

Figure 5. Difference in Deviance Residuals between Nominal and Robust Models



Notes: Blue bars indicate a multiplier of $\alpha = 0.5$, green bars indicate a multiplier of $\alpha = 1.0$, and red bars indicate a multiplier of $\alpha = 1.5$. The vertical axis is the difference between the nominal and robust residual for the same data-point, with larger values indicating a better fit for the robust model. The horizontal axis is the value of the error threshold. The case of zero error was excluded because the models are equivalent. Each boxplot represents 5,000 observations.

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 data sources, including the size and magnitude of online networks, the rate at which they evolve, and the varying degrees of network visibility across sources. The other source of discrepancy stems from natural cognitive tendencies of the individuals within the network, such as compression, personality traits, and position or power. Despite this body of literature, social network inference methodologies generally treat observable data, such as networks collected from online sources, as proxy for the perceived network. Inference about human behavior is derived from measures of

this observable information, which may deviate from the internal schema which individuals believe to be true. And, as discussed earlier, many social science theories 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 which accounts for these sources of error when conducting statistical inference and hypothesis testing. Following the framework advanced in Bertsimas and Nohadani (2018), we derived a robust framework which preserves the techniques of classic network analysis, while making the estimates resilient to the discrepancies we acknowledge are present in our data. Our approach makes an incremental methodological contribution by incorporating the exponential distribution, commonly applied in a network context, whereas prior research in robust MLE has been focused on normal distributions (Bertsimas and Nohadani, 2018). Consequently, our derivation may also be relevant to applications outside social network analysis in which exponential distribution arise, e.g. survival analysis.

As a proof of concept, we illustrated how our robust methodology could provide quantitative and qualitative advantages over standard network inference on a classic social networks dataset. Using individually reported friendship networks from 21 networks collected by Krackhardt (1987), we fit an ERG model and a robust model to a locally aggregated network. As we increased the global error tolerance – i.e., anticipating greater deviance from the aggregate network – the network parameters reflected a shift towards denser, more clustered networks. However, the added robustness did come with a price in the form of decreased model fit for the LAS network. When applied to the individually reported networks, our robust parameters outperformed the standard MLE for the majority of the CSSs in the sample. Further, the robust

parameters were a significantly better fit to the CSS network which deviated the most from the LAS. Our analysis suggests that robust MLE for networks has the ability to account for the discrepancies between collected or aggregated networks and the networks perceived by individuals.

Our experiments on synthetic data demonstrate the advantages of utilizing robust estimators. As error is introduced into the dataset, our model quickly brings to zero the parameter associated with the biased network statistic, transitivity. Consequently, when we account for the fact that individuals do not know the exact value of the second-degree connections (between their alters) in their network, we find no evidence that they are motivated to act based on transitivity. Our experiments also show that robust estimators lead to a better overall model fit, and more accurately predict the large majority of events in the simulated sequences. Thus, we may conclude that applying a robust approach will allow us to more accurately explain individual behaviors that are motivated by a network context, if we assume that perceptions deviate from our observations.

Beyond the quantitative advantages afforded 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 which 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 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. For instance, a robust formulation of a network model would account for the tendency of individuals to view themselves as more central, or to assume their friends to be friends with each other.

One caveat is that in the case of network inference, particularly on large digital networks, 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. This type of analysis would suggest there is some other mechanism at work, or that the relationship between the network structure and behavior is more subconscious in nature. However, if the social network data is derived from a relatively small sample with good visibility, then many of the sources of discrepancy in the observable network data are likely negligible. In these circumstances, the robust method may be unnecessary.

While the robust approach does produce results that are more sensitive to cognitive issues, there are still limitations to our interpretations. As in all methods of statistical inference, causality is a general limitation. 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. Beyond this limitation, there are alternative methods for correcting biased observable networks. First, 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 unwieldy if not impossible. Additionally, these reports would have to be collected *when the tie was formed*, otherwise the data is still subject to errors in recollection. The second approach is to infer an individual's perceived network, based on systematic biases found empirically. Essentially, a CSS could be estimated for an actor based on a probability distribution that is theoretically informed. The challenge with this approach is the difficulty 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.

Another 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, 2018). On the other hand, when only

maximum errors are known, a polyhedral set – as used in our implementation – offers a good description.

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 affects the computation and interpretation of hypothesized mechanisms in social network analysis conducted using digital trace data which offer varying degrees of technological affordances. These affordances impact the cognitive tendencies of individuals to discern who knows whom. Hence, explanations that assume individuals act and interact based on the observed network in the digital trace data are not always warranted. Just because a link is observed by someone, this does not always mean that the linked individuals are aware of – and make decisions based on – this observed link. In such cases, we propose a novel method which 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 actual self-report data as well as simulations, we illustrated the efficacy of our technique in adjusting the effects of variables impacted by perception and providing overall better predictive capabilities. Thus, this study highlights potential theoretical and methodological issues in the analysis of digital social networks, and provides a methodological solution.

APPENDIX 1

Summary of Studies

In order to demonstrate the pervasiveness of online network data in management studies, we conducted a survey of recent publications in top journals, including but not limited to MIS Quarterly, Information Systems Research, and Management Science. While the preceding list is not exhaustive, it is representative of the types of network studies commonly found in the last 10 to 15 years. Along with each study, we note the specific type of network data that is collected. We excluded publications in which the networks were derived from sociometric surveys as well as purely theoretical or methodological papers.

Study	Network Data Source
Ahuja et al. (2003)	Email exchanges
Aral and Van Alstyne (2011)	Email exchanges
Aral and Walker (2012)	Online social networks
Aral and Walker (2014)	Online social networks
Aral et al. (2009)	Instant messaging
Bapna and Umyarov (2015)	Online social networks
Bapna et al. (2017)	Online social networks
Bapna, Gupta, et al. (2017)	Online social networks
Centola (2010)	Online social networks
Chen et al. (2017)	Software project affiliation
Conaldi and Lomi (2013)	Software project affiliation
Fang and Hu (2018)	Online social networks
Faraj and Johnson (2011)	Forum posts
Grewal et al. (2006)	Software project affiliation

Johnson et al. (2014)	Computational model/forum posts
Kim et al. (2018)	Corporate social network
Kudaravalli and Faraj (2008)	Email exchanges
Lazer and Friedman (2007)	Computational model
Leonardi (2015)	Corporate social network
Lewis et al. (2012)	Online social networks
Lu et al. (2017)	Customer support forum
Quintane and Carnabuci (2016)	Email exchanges
Quintane and Kleinbaum (2011)	Email exchanges
Quintane et al. (2013)	Email exchanges
Quintane et al. (2014)	Software project affiliation
Shore et al. (2018)	Twitter links
Singh and Tan (2010)	Software project affiliation
Singh et al. (2011)	Software project affiliation
Soda et al. (2004)	Project affiliation
Spiro et al. (2013)	Archival records
Srivastava (2015)	Email exchanges
Stadtfeld and Geyer-Schulz (2011)	Forum posts
Susarla et al. (2011)	Online social networks
Tucker (2008)	Archival messaging networks
Vu et al. (2015)	Forum posts
Xue et al. (2018)	Interfirm managerial social ties
Zaheer and Soda (2009)	Project affiliation

APPENDIX 2

Derivation of Robust Estimator

We solve the core optimization problem (given in Equation 7) 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 which minimize the log-likelihood function. The inner problem is equal to:

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

In (1a) 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

$$\min_{\|\Delta\mathbf{x}_{zt}\| \leq \rho} \left(\boldsymbol{\beta}'(\mathbf{x}_{ty\bullet} - \Delta\mathbf{x}_{ty\bullet}) - \log \sum_{z \in \mathcal{A}_t} \exp(\boldsymbol{\beta}'(\mathbf{x}_{tz\bullet} - \Delta\mathbf{x}_{tz\bullet})) \right) \tag{2a}$$

for each instance t . We note that the objective function for each component of the inner optimization problem is decreasing in $\boldsymbol{\beta}'\Delta\mathbf{x}_{ty\bullet}$ and increasing in $\boldsymbol{\beta}'\Delta\mathbf{x}_{tz\bullet}$ for all $z \neq y$; for proof, see Appendix 3. As a result, the optimal value can be found by determining the maximum feasible value of $\boldsymbol{\beta}'\Delta\mathbf{x}_{ty\bullet}$ and the minimum value of $\boldsymbol{\beta}'\Delta\mathbf{x}_{tz\bullet}$ for all $z \neq y$. By applying Hölder's inequality, we know that

$$-\|\boldsymbol{\beta}\|\|\Delta\mathbf{x}_{zt}\| \leq \boldsymbol{\beta}'\Delta\mathbf{x}_{zt} \leq \|\boldsymbol{\beta}\|\|\Delta\mathbf{x}_{zt}\|. \quad (3a)$$

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

$$\begin{aligned} \min_{\|\Delta\mathbf{x}_{zt}\| \leq \rho} & \left(\boldsymbol{\beta}'(\mathbf{x}_{ty\cdot} - \Delta\mathbf{x}_{ty\cdot}) - \log \sum_{z \in \mathcal{A}_t} \exp(\boldsymbol{\beta}'(\mathbf{x}_{tz\cdot} - \Delta\mathbf{x}_{tz\cdot})) \right) \\ & = \boldsymbol{\beta}'\mathbf{x}_{ty\cdot} - \|\boldsymbol{\beta}\|\rho - \log \sum_{z \in \mathcal{A}_t} \exp(\boldsymbol{\beta}'\mathbf{x}_{tz\cdot} + (-1)^{u_z}\|\boldsymbol{\beta}\|\rho). \end{aligned} \quad (4a)$$

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 $\Delta\mathbf{x}_{tz\cdot}$ is given by:

$$\Delta\mathbf{x}_{tz\cdot}^*(\boldsymbol{\beta}) = (-1)^{1-u_z}\rho \frac{\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|} \times \mathbf{1}\{\boldsymbol{\beta} \neq \mathbf{0}\}. \quad (5a)$$

Thus, we combine all of our elements into a single solution for the inner problem:

$$\phi(\boldsymbol{\beta}; \mathbf{X}^{obs}) = \sum_{t=1}^M \left(\boldsymbol{\beta}'\mathbf{x}_{ty\cdot} - \|\boldsymbol{\beta}\|\rho - \log \sum_{z \in \mathcal{A}_t} \exp(\boldsymbol{\beta}'\mathbf{x}_{tz\cdot} + (-1)^{u_z}\|\boldsymbol{\beta}\|\rho) \right). \quad (6a)$$

We can now solve the robust outer MLE problem

$$\max_{\boldsymbol{\beta}} \phi(\boldsymbol{\beta}; \mathbf{X}^{obs}) = \max_{\boldsymbol{\beta}} \sum_{t=1}^M \left(\boldsymbol{\beta}'\mathbf{x}_{ty\cdot} - \|\boldsymbol{\beta}\|\rho - \log \sum_{z \in \mathcal{A}_t} \exp(\boldsymbol{\beta}'\mathbf{x}_{tz\cdot} + (-1)^{u_z}\|\boldsymbol{\beta}\|\rho) \right). \quad (7a)$$

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 Appendix 4.

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 $\alpha_k \rightarrow 0$ as $k \rightarrow \infty$, then the outlined procedure will converge to a locally optimal solution $\boldsymbol{\beta}^*$, and will do so in polynomial time (Bertsimas et al. 2010). We implemented a step length of $\alpha_k = \frac{\|\nabla \phi(\boldsymbol{\beta}^{(0)})\|}{k}$.

In practice, with a tolerance criterion of $\epsilon = 10^{-6}$ our algorithm converged to a local optimum in around 10 iterations.

Algorithm 1:

-
1. Initialize with an estimator $\boldsymbol{\beta}^{(0)}$. Set $k = 0$.
 2. Solve the inner problem for all $t = 1, \dots, M$ to obtain the optimal errors $\Delta \mathbf{x}_{tz}^*(\boldsymbol{\beta}^{(k)})$ for each $z \in \mathcal{A}_t$.
 3. Using the worst case errors $\Delta \mathbf{X}^*(\boldsymbol{\beta}^{(k)})$, calculate $\phi(\boldsymbol{\beta}^{(k)}; \mathbf{X}^{obs}) = \psi(\boldsymbol{\beta}^{(k)}; \mathbf{X}^{obs} - \Delta \mathbf{X}^*(\boldsymbol{\beta}^{(k)}))$. By applying Danskin's theorem, we know that computing $\nabla \psi(\boldsymbol{\beta}^{(k)})$ is equivalent to computing $\nabla \phi(\boldsymbol{\beta}^{(k)})$ (Bertsimas and Nohadani, 2018). 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.
-

APPENDIX 3

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 $\boldsymbol{\beta}'\Delta\mathbf{x}_{tq\bullet}$. For notational purposes, let

$$h_t = \boldsymbol{\beta}'(\mathbf{x}_{ty\bullet} - \Delta\mathbf{x}_{ty\bullet}) - \log \sum_{z \in \mathcal{A}_t} \exp(\boldsymbol{\beta}'(\mathbf{x}_{tz\bullet} - \Delta\mathbf{x}_{tz\bullet})).$$

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

Here, $\mathbf{1}\{q = y\}$ is an indicator function, taking a value of 1 if $q = y$ and 0 otherwise. Because

the value of $\frac{\exp(\boldsymbol{\beta}'(\mathbf{x}_{tq\bullet} - \Delta\mathbf{x}_{tq\bullet}))}{\sum_{z \in \mathcal{A}_t} \exp(\boldsymbol{\beta}'(\mathbf{x}_{tz\bullet} - \Delta\mathbf{x}_{tz\bullet}))}$ is non-negative and at most 1, we can conclude that

$$\frac{\partial h_t}{\partial(\boldsymbol{\beta}'\Delta\mathbf{x}_{tq\bullet})} \leq 0 \text{ for } q = y, \text{ and } \frac{\partial h_t}{\partial(\boldsymbol{\beta}'\Delta\mathbf{x}_{tq\bullet})} \geq 0 \text{ for all other } q. \text{ Thus, } h_t \text{ is a monotonically}$$

decreasing function of $\boldsymbol{\beta}'\Delta\mathbf{x}_{ty\bullet}$ and a monotonically increasing function of $\boldsymbol{\beta}'\Delta\mathbf{x}_{tz\bullet}$ for all $z \neq y$.

APPENDIX 4

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 $\Delta \mathbf{x}^*(\boldsymbol{\beta})$ – must be computed. We assume that the norm $\|\bullet\|$ refers to the Euclidean norm. For our specified likelihood function, the gradient is as follows:

$$\begin{aligned} \nabla \phi_p(\boldsymbol{\beta}) &= \frac{\partial}{\partial \beta_p} \left(\sum_{t=1}^M \boldsymbol{\beta}' \mathbf{x}_{ty\bullet} - \|\boldsymbol{\beta}\| \rho - \log \sum_{z \in \mathcal{A}_t} \exp(\boldsymbol{\beta}' \mathbf{x}_{tz\bullet} + (-1)^{u_z} \|\boldsymbol{\beta}\| \rho) \right) \\ &= \sum_{t=1}^M x_{typ} - \frac{\beta_p}{\|\boldsymbol{\beta}\|} \rho - \frac{\sum_{z \in \mathcal{A}_t} \left(x_{t zp} + (-1)^{u_z} \frac{\beta_p}{\|\boldsymbol{\beta}\|} \rho \right) \exp(\boldsymbol{\beta}' \mathbf{x}_{tz\bullet} + (-1)^{u_z} \|\boldsymbol{\beta}\| \rho)}{\sum_{z \in \mathcal{A}_t} \exp(\boldsymbol{\beta}' \mathbf{x}_{tz\bullet} + (-1)^{u_z} \|\boldsymbol{\beta}\| \rho)} \end{aligned}$$

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 \mathbf{v} that satisfies the following inequality

$$\psi(\boldsymbol{\beta}_2; \mathbf{X}^{\text{obs}} - \Delta \mathbf{X}^*(\boldsymbol{\beta}_2)) - \psi(\boldsymbol{\beta}_1; \mathbf{X}^{\text{obs}} - \Delta \mathbf{X}^*(\boldsymbol{\beta}_1)) \leq \mathbf{v} \cdot (\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1),$$

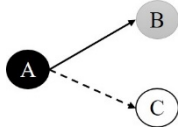
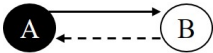
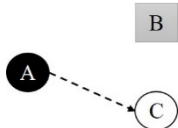
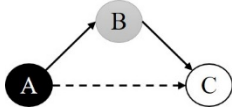
where \cdot is the dot product. Thus, the subgradient used in the optimization of the outer objective function is as follows:

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

APPENDIX 5

Summary of network statistics

The number of links directed from a to b up to time t is denoted n_{abt} , and y_{ab} is an indicator variable, which takes a value of 1 when a and b share a group, and 0 otherwise.

Variable	Illustration	Formula	Interpretation
Activity		$x_1(a, c, t) = \frac{\sum_b n_{abt}}{\sum_l \sum_b n_{lbt}}$	As A sends more messages relative to the rest of the network, A is more likely to send a new message.
Reciprocity		$x_2(a, b, t) = \frac{n_{bat}}{\sum_k n_{kat}}$	As B increasingly sends A messages, A becomes more likely to send a message to B .
Homophily		$x_3(a, c, t) = y_{ac}$	A is more likely to send a message to C and not B if they are members of the same group and B is not.
Transitivity		$x_4(a, c, t) = \frac{\sum_b n_{abt} n_{bct}}{\sum_l \sum_b n_{abt} n_{blt}}$	A is more likely to send a message to C if there are frequent messages from A to other actors B and from those actors to target C .

Notes: The sender of the message is in black, while the receiver is in white. Other individuals are grey. The shape of a node is used to indicate which group they belong to. Arrows represent direction of a tie, where solid lines indicate prior communication and dashed lines indicate future interaction.

REFERENCES

- Ahuja, M. K., Galletta, D. F., and Carley, K. M. 2003. "Individual Centrality and Performance in Virtual R&D Groups: An Empirical Study," *Management Science* (49:1), pp. 21–38.
- Aral, S., Muchnik, L., and Sundararajan, A. 2009. "Distinguishing Influence-Based Contagion from Homophily-Driven Diffusion in Dynamic Networks," *Proceedings of the National Academy of Sciences* (106:51), pp. 21544–21549. (<https://doi.org/10.1073/pnas.0908800106>).
- 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., and Albert, R. 1999. "Emergence of Scaling in Random Networks," *Science* (286:5439), pp. 509–512.
- 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.
- Bernard, H., Killworth, P., & Sailer, L. (1980). Informant accuracy in social network data IV. *Social Networks*, 2, 191-218.

- Bernard, H., Killworth, P., & Sailer, L. (1982). Informant accuracy in social network data V., *Social Science Research*, 11, 30-66.
- Bernard, H.R., Killworth, P., & Cronenfeld, D. (1984). The problem of informant accuracy: The validity of retrospective data. *Annual Review of Anthropology*, 13, 495-517.
- Bertsimas, D., Brown, D. B., and Caramanis, C. 2011. "Theory and Applications of Robust Optimization," *SIAM Review* (53:3), pp. 464–501.
- Bertsimas, D., Nohadani, O., and Nohadani, O. 2018. "Robust Maximum Likelihood Estimation," *INFORMS Journal on Computing*.
- Bertsimas, D., Nohadani, O., and Teo, K. M. 2010. "Robust Optimization for Unconstrained Simulation-Based Problems," *Operations Research* (58:1), pp. 161–178.
- 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. 2013. "Humans Use Compression Heuristics to Improve the Recall of Social Networks," *Scientific Reports* (3).
- 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.
- 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.
- 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.
- Centola, D. 2010. "The Spread of Behavior in an Online Social Network Experiment," *Science* (329:5996), pp. 1194–1197. (<https://doi.org/10.1126/science.1185231>).
- Chen, W., Wei, X., and Zhu, K. X. 2017. "Engaging Voluntary Contributions in Online Communities: A Hidden Markov Model," *MIS Quarterly* (42:1).
- Conaldi, G., and Lomi, A. 2013. "The Dual Network Structure of Organizational Problem Solving: A Case Study on Open Source Software Development," *Social Networks* (35:2), pp. 237–250.

- 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 B. Foucault Welles & S. Gonzalez-Bailon (Eds.), *Oxford Handbook of Communication in the Networked Age*. Oxford University Press.
- 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.
- Fang, X., and Hu, P. J. 2018. "Top Persuader Prediction for Social Networks," *MISQ Quarterly* (42:1), pp. 63-82.
- 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., Amanatullah, E. T., and Ames, D. R. 2006. "Helping One's Way to the Top: Self-Monitors Achieve Status by Helping Others and Knowing Who Helps Whom," *Journal of Personality and Social Psychology* (91:6), p. 1123.
- 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.
- 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.
- Freeman, L. C., and Webster, C. M. 1994. "Interpersonal Proximity in Social and Cognitive Space," *Social Cognition* (12:3), p. 223.
- Gigerenzer, G., and Brighton, H. 2009. "Homo Heuristicus: Why Biased Minds Make Better Inferences," *Topics in Cognitive Science* (1:1), pp. 107–143.
- Grewal, R., Lilien, G. L., and Mallapragada, G. 2006. "Location, Location, Location: How Network Embeddedness Affects Project Success in Open Source Systems," *Management Science* (52:7), pp. 1043–1056.
- Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., and Morris, M. 2008. "Statnet: Software Tools for the Representation, Visualization, Analysis and Simulation of Network Data," *Journal of Statistical Software* (24:1), p. 1548.
- 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.
- Heider, F. 1958. *The Psychology of Interpersonal Relations*, Psychology Press.

- 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.
- Hunter, D. R., and Handcock, M. S. 2006. "Inference in Curved Exponential Family Models for Networks," *Journal of Computational and Graphical Statistics* (15:3), pp. 565–583. (<https://doi.org/10.1198/106186006X133069>).
- 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.
- 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.
- Kilduff, M., Crossland, C., Tsai, W., and Krackhardt, D. 2008. "Organizational Network Perceptions versus Reality: A Small World after All?," *Organizational Behavior and Human Decision Processes* (107:1), pp. 15–28.
- Kim, Y., Jarvenpaa, S., and Gu, B. 2018. "External Bridging and Internal Bonding: Unlocking the Generative Resources of Member Time and Attention Spent in Online Communities," *MIS Quarterly* (42), pp. 265–283. (<https://doi.org/10.25300/MISQ/2018/13278>).
- Krackhardt, D. 1987. "Cognitive Social Structures," *Social Networks* (9:2), pp. 109–134.
- Krackhardt, D. 1990. "Assessing the Political Landscape: Structure, Cognition, and Power in Organizations," *Administrative Science Quarterly*, pp. 342–369.
- 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.

- Kudaravalli, S., and Faraj, S. 2008. "The Structure of Collaboration in Electronic Networks," *Journal of the Association for Information Systems* (9:10/11), pp. 706–726.
- Kumbasar, E., Romney, A. K., and Batchelder, W. H. 1994. "Systematic Biases in Social Perception," *American Journal of Sociology*, pp. 477–505.
- Lazer, D., and Friedman, A. 2007. "The Network Structure of Exploration and Exploitation," *Administrative Science Quarterly* (52:4), pp. 667–694. (<https://doi.org/10.2189/asqu.52.4.667>).
- 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>).
- Leonardi, P. M. 2015. "Ambient Awareness and Knowledge Acquisition: Using Social Media to Learn" Who Knows What" and" Who Knows Whom"," *MIS Quarterly* (39:4), pp. 747–762.
- Leonardi, P., & Contractor, N. S. (2018, November 25). Better People Analytics: Measure who they know, not just who they are. *Harvard Business Review*, 70–81.
- Levin, D. Z., Walter, J., Walter, J., & Murnighan, J. K. (2011). Dormant Ties - The Value Of Reconnecting. *Organization Science*, 22(4), 923–939.
- Lewis, K., Gonzalez, M., and Kaufman, J. 2012. "Social Selection and Peer Influence in an Online Social Network," *Proceedings of the National Academy of Sciences* (109:1), pp. 68–72.
- 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>).
- Mariotti, F., & Delbridge, R. (2012). Overcoming Network Overload and Redundancy in Interorganizational Networks: The Roles of Potential and Latent Ties. *Organization Science*, 23(2), 511–528.
- Monge, P. R., and Contractor, N. S. 2003. *Theories of Communication Networks*, Oxford University Press.
- 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.

- Pattison, P. (1994). Social cognition in context: Some applications of social network analysis. In S. Wasserman & J. Galaskiewicz (Eds.), *Advances in social network analysis: Research in the social and behavioral sciences* (pp. 79-109). Thousand Oaks, CA: Sage.
- 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.
- Quintane, E., Pattison, P. E., Robins, G. L., and Mol, J. M. 2013. "Short-and Long-Term Stability in Organizational Networks: Temporal Structures of Project Teams," *Social Networks* (35:4), pp. 528–540.
- Richards, W. D. (1985). Data, models, and assumptions in network analysis. In R. D. McPhee & P. K. Tompkins (Eds.), *Organizational communication: Themes and new directions* (pp. 109-147). Newbury Park, CA: Sage.
- Shore, J., Baek, J., and Dellarocas, C. 2018. "Network structure and patterns of information diversity on Twitter," *MIS Quarterly* (42:3), pp. 849-872.
- Siciliano, M. D., Yenigun, D., and Ertan, G. 2012. "Estimating Network Structure via Random Sampling: Cognitive Social Structures and the Adaptive Threshold Method," *Social Networks* (34:4), pp. 585–600.
- Simpson, B., and Borch, C. 2005. "Does Power Affect Perception in Social Networks? Two Arguments and an Experimental Test," *Social Psychology Quarterly* (68:3), pp. 278–287.
- Simpson, B., Markovsky, B., and Steketee, M. 2011. "Power and the Perception of Social Networks," *Social Networks* (33:2), pp. 166–171.
- Singh, P. V., and Tan, Y. 2010. "Developer Heterogeneity and Formation of Communication Networks in Open Source Software Projects," *Journal of Management Information Systems* (27:3), pp. 179–210.
- 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., 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. 1996. "Stochastic Actor-oriented Models for Network Change," *Journal of Mathematical Sociology* (21:1–2), pp. 149–172.

- 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.
- Soda, G., Usai, A., and Zaheer, A. 2004. "Network Memory: The Influence of Past and Current Networks on Performance," *Academy of Management Journal* (47:6), pp. 893–906. (<https://doi.org/10.2307/20159629>).
- Spiro, E. S., Acton, R. M., and Butts, C. T. 2013. "Extended Structures of Mediation: Re-Examining Brokerage in Dynamic Networks," *Social Networks* (35:1), pp. 130–143.
- Srivastava, S. B. 2015. "Intraorganizational Network Dynamics in Times of Ambiguity," *Organization Science* (26:5), pp. 1365–1380.
- Stadtfeld, C. 2012. *Events in Social Networks: A Stochastic Actor-Oriented Framework for Dynamic Event Processes in Social Networks*, KIT Scientific Publishing.
- Stadtfeld, C., and Geyer-Schulz, A. 2011. "Analyzing Event Stream Dynamics in Two-Mode Networks: An Exploratory Analysis of Private Communication in a Question and Answer Community," *Social Networks* (33:4), pp. 258–272.
- 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>).
- Thomas, W. I. & Thomas, D. S. (1928). The methodology of behavior study. In W. I. Thomas & D. S. Thomas, *The Child in America: Behavior Problems and Programs* (pp. 553–576). New York: Knopf.
- Treem, J. W., & Leonardi, P. M. (2012). Social media use in organizations: Exploring the affordances of visibility, editability, persistence and association. *Communication Yearbook*, (36), 143–189.
- Tucker, C. 2008. "Identifying Formal and Informal Influence in Technology Adoption with Network Externalities," *Management Science* (54:12), pp. 2024–2038. (<https://doi.org/10.1287/mnsc.1080.0897>).
- Vu, D., Pattison, P., and Robins, G. 2015. "Relational Event Models for Social Learning in MOOCs," *Social Networks* (43), pp. 121–135.
- Walter, J., Levin, D. Z., & Murnighan, J. K. (2015). Reconnection Choices: Selecting the Most Valuable (vs. Most Preferred) Dormant Ties. *Organization Science*, 26(5), 1447–1465. <http://doi.org/10.1287/orsc.2015.0996>
- Wasko, M. M., and Faraj, S. 2005. "Why Should I Share? Examining Social Capital and Knowledge Contribution in Electronic Networks of Practice," *MIS Quarterly*, pp. 35–57.
- Wasserman, S., and Faust, K. 1994. *Social Network Analysis: Methods and Applications*, Cambridge University Press.

- 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.
- Xue, L., Yang, K., and Yao, Y. 2018. "Examining the Effects of Interfirm Managerial Social Ties on IT Components Diversity: An Agency Perspective," *MIS Quarterly* (42:2), pp. 679-694.
- Yenigun, D., Ertan, G., & Siciliano, M. (2017). Omission and commission errors in network cognition and network estimation using ROC curve. *Social Networks*, 50, 26–34.
<http://doi.org/10.1016/j.socnet.2017.03.007>
- Zaheer, A., and Soda, G. 2009. "Network Evolution: The Origins of Structural Holes," *Administrative Science Quarterly* (54:1), pp. 1–31.