Misspecification, Sparsity, and Superpopulation Inference for Sparse Social Networks

Alexander D'Amour* and Edoardo Airoldi

Department of Statistics, Harvard University

*damour@fas.harvard.edu

Abstract

Recent interest in network data has driven a flurry of research into generative network models. However, despite impressive theoretical progress, these models have a mixed record in scientific application. In particular, there is a disconnect between two of the major use cases for network models. In the first case, which we call single-sample problems, investigators hope to understand the network dynamics *within* a fixed set of individuals. In the second case, which we call superpopulation problems, investigators hope to understand network dynamics that are common *between* network samples obtained from distinct sets of individuals, so that different network samples (for example, from different cities) can be compared and understood together. Despite the importance of both of these problems, most theoretical work and successful investigations have focused on single-sample rather than superpopulation problems. Unlike the classical case of independent data, for network data, the theories of estimation in large single-sample problems and in superpopulation problems are not equivalent.

In this paper, we develop a theoretical framework for the network superpopulation inference problem and use it to understand why many network models are ineffective at predicting, comparing, or sharing information across network samples. We tie these difficulties to two of the perennial complications in network modeling: model misspecification and network sparsity. Motivated by this characterization, we propose a modeling and inference framework that is robust to the sparse scaling of social netwoks. This framework avoids specifying the mechanism that generates the sparsity in the underlying social process by instead fully specifying the likelihood for the same data filtered through a different observation mechanism. The derived sparsity-robust estimator inherits the easy extensibility and theoretical guaranteed of MLE estimators, and has the added advantage of compuational efficiency. We demonstrate this framework on simulated data.

Contents

1 Introduction

In recent years, social network data have become available that catalogue social interactions between actors in a wide range of contexts, from coauthorship to personal relationships to email correspondence. These have sparked investigations about network structure in a variety of fields including organizational behavior, marketing, political science, and sociology. In response, the statistical and machine learning communities have offered a variety of modeling approaches that give intuitive quantitative summaries of networks in terms of generative parameters (see [33] for an overview).

Generally, we can divide investigators' inferential questions into one of two categories. The first is single-sample problems, where investigators wish to infer some properties of a social network defined on a fixed, finite set of vertices. For example, investigators may wish to infer links that are missing from the current dataset, or predict future interactions among a given set of actors. The second category is superpopulation problems, where investigators wish to infer properties that are shared between social networks defined on different actor sets. We call these superpopulation problems because they require an appeal to a superpopulation to justify generalizing inferences between heterogeneous samples. For example, investigators may wish to test whether two network samples were generated by the same process, or define a hierarchical model to borrow strength between network samples.

Generative network models have shown promising performance in answering single-sample questions, but have been less successful for superpopulation questions. In these contexts, parameter estimates are often unstable when investigators wish to compare networks of different size, a problem msot notably documented in [20]. We also see this problem in cases where only a single network is of interest, but models developed from small-sample intuition (e.g., by obseving 18 monks in Sampson's Monastery) are applied to large datasets (e.g., messaging behavior among Facebook users). In these cases, we see that parameter estimates land outside of the range of reasonable effect sizes, and are thus not easily interpreted and incorporated into social science theory (see simulation results in Section 8).

At first, this failure appears puzzling given the impressive array of theoretical work that has been developed to support many popular network models, e.g., [3, 8]. In actuality, this situation is unsurprising because single-sample and superpopulation questions interrogate different aspects of a data-generating process. Given that no simple network model can capture the full complexity of human social dynamics, there is little reason to believe that a model that is effective for answering single-sample questions should also be useful for

superpopulation questions. Indeed, these properties only coincide in the classical setting of independent data where large samples and superpopulations have the same stochastic process structure. With the dependence present in network data, separate arguments are necessary to show that a particular procedure captures single-sample or superpopulation properties of the date-generating process. So far, in extending notions of large-sample consistency to network models, authors in this literature have focused on single-sample arguments.

In this paper, we develop a theoretical framework for evaluating a model's suitability for network superpopulation investigations. Within this framework, it becomes clear that the poor performance of network models in generalization tasks is a symptom of model misspecification, specifically the aspect of the model that embeds the observed network sample into a superpopulation process. As expected, this misspecification is largely immaterial to answering single-sample questions, but has serious consequences for superpopulation investigations. In particular, we find that one particular type of misspecification, which we call *sparsity misspecification*, is sufficient to derail superpopulation analyses that hope to generalize inferences between network samples of different size. We say a model is sparsity misspecified if it does not precisely capture the sparsity of a social interaction process. Heuristically, sparsity refers to the tendency of social interaction networks to have vanishing network density – defined as the ratio of the number of observed interactions to the number of potential interactions in a network – as the network sample becomes large. Model misspecification and sparsity are thorny issues that are always lurking in the background in the statistical analysis of networks; one advantage of our theoretical approach is that it allows us to reason about these issues in one coherent framework.

Sparsity misspecification is a ubiquitous problem among popular network analysis models, most notably those that assume independence between pairs of actors, or dyads, in network samples conditional on observable or latent characteristics. On the other hand, several large questions of interest in social science require inferences that can be reliably generalized between network samples for out-of-sample prediction, between-sample comparison, and multilevel modeling. Model improvement is an attractive option, but information about the sparsity of a social process is difficult to obtain from a small number of network samples and models that have flexible sparsity patterns are difficult to specify and fit. To solve this impasse, we propose that sparsity invariance is a more realistic and robust modeling principle, and suggest a modeling and inference framework where the object of inference and the inferential procedure are invariant to the sparsity of the underlying population.

1.1 A running example: inventor collaboration network

Throughout the paper, we use the data analytic problem that motivated this work as a running example. We use an inventor-disambiguated version of the US patent record [21] to build a collaboration network among inventors who filed for patents in the United States between 1975 and 2010. In the network representation, inventors are represented as vertices, and coauthorships on patents are represented as pairwise outcomes between the actors. The data set contains the date of each coauthorship (which we define as the application date), and we often see repeated coauthorships between pairs of inventors. Thus, at full resolution, each pairwise outcome has a point-process structure, but over fixed time ranges the outcome can also be represented as counts.

The inventor data also contains side information that we can use as covariates to model collaboration behavior, including each inventor's firm and zipcode. Throughout the paper, we will consider three simple binary covariates that are available for each inventor-pair collaboration event: whether the inventors live in the same zipcode, whether the inventors work for the same firm (the "assignee") at the time of the patent application, and whether the inventors had a previous patent collaboration before the current patent application.

Some simple analyses based on these covariates showcase the problems we have described so far. Consider a point-process regression model, in the style of [27], where we specify the log-hazard of a collaboration event as a linear combination of the zipcode, assignee, and previous collaboration covariates described above (see Section 8.1.1 for additional details). We apply this model to regional collaboration networks constructed from a 6-year window of interaction data beginning in 1983. The results for each Census Bureau Statistical Area, corresponding to the metropolitan area surrounding each city, are shown on the left of Figure 1. These display a strong dependence between the parameter estimates and the size of the network sample, and extremely large effect estimates and extremely small uncertainy estimates. We also display the results of our sparsity-invariant methodology on the right.

Taking one region at a time, these extreme parameter estimates are not surprising. For example, when collaboration events are relatively rare compared to the total number of inventor-pairs, we would expect collaboration events between inventors who have already generated a patent together to be orders of magnitude more common than events occuring between any arbitrary pair of inventors. However, if we wish to distinguish between collaboration patterns in different regions of the country, it is unclear how we would use these parameter estimates to do so. Certainly some variation should be expected between regions,

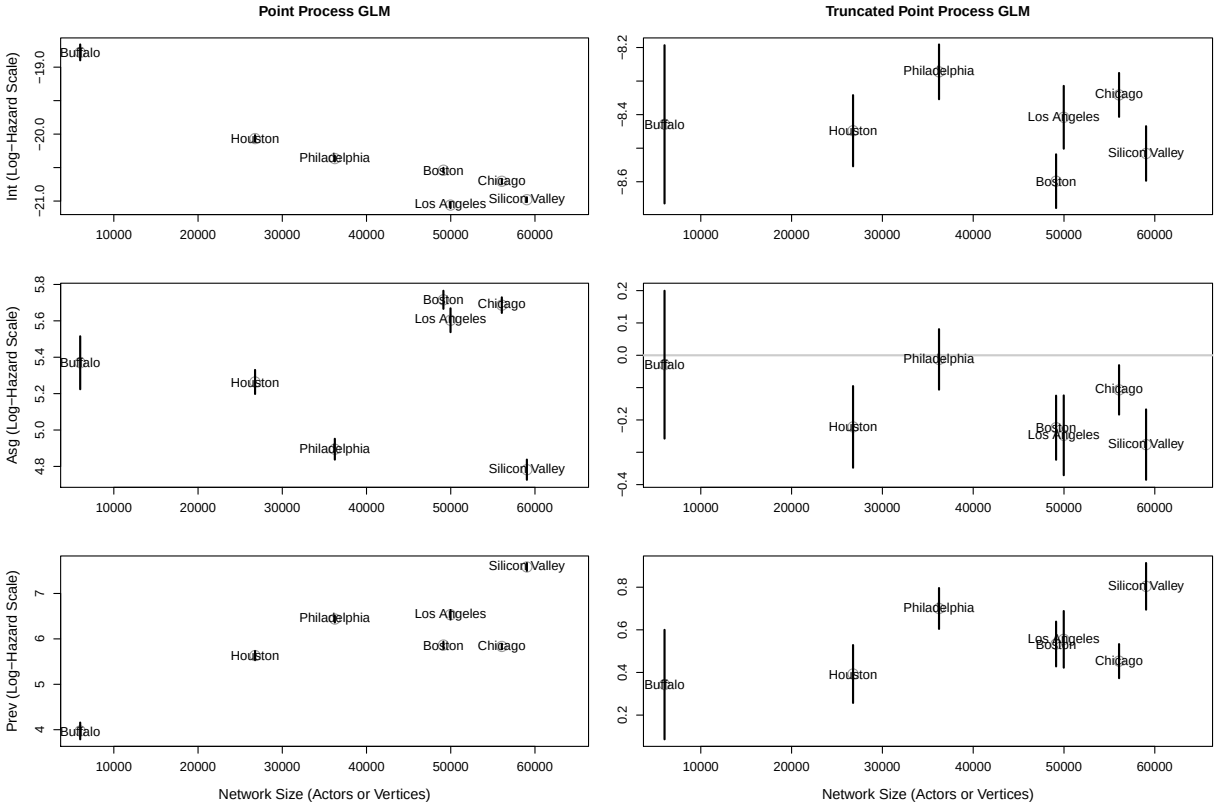

Figure 1: Inferred parameter values and asymptotic intervals from a simple point process regression model explaining patent collaboration events occurring in different regional inventor networks in the United States. (Left) parameter estimates from this standard conditionally independent dyad (see Section **??**) model show strong dependence on sample size, extremely large effect estimates, and very small error estimates. (Right) parameter estimates from our truncated methodology (see Section **??**) show stability across regions with more realistic effect and error estimates.

but this example makes clear that it is difficult to separate the sparsity effect in the parameter estimates from true differences in the data generating processes between these network samples. This difficulty and methods to avoid it are the main focus of this paper. We will return to a simulated version of this example in Section 8.

1.2 Contributions and related work

This paper has two main pieces. The first introduces the theory of sparsity misspecification, while the second proposes a modeling framework and corresponding inferential procedure that is sparsity-invariant.

The theory section is built around a statistical framework that defines the notions of sample and superpopulation in the context of networks. Our formulation extends Shalizi's work in [30], which defined a network superpopulation as a stochastic process indexed by actor-sets, and samples drawn from this network as realizations of finite dimensional distributions of this population process. [30] used this framework to characterize the properties of exponential random graph models (ERGMs), specifically to determine whether embedding an ERGM into a stochastic process is feasible at all, a property they call projectibility. In this paper, we use similar formalism but tackle a different question. Instead of asking whether a proposed *model* is projective – the models we consider have a natural extension to a stochastic process, and are thus trivially projective – we use the stochastic process framework to investigate whether the *inferences* obtained from samples drawn from the same superpopulation exhibit the coherence necessary to be suitable for generalization.

The question of whether a model gives *stable* inferences, in the sense that nominally similar samples yield similar inferences, is a critical question when we consider utility of simple parametric models in scientific arguments. Because we know that simple models for complex social phenomena must be misspecified in some way, stability is one of the only criteria by which we can judge whether the parameter estimates for a given model are capturing scientifically useful signal. Notions of stability have appeared in many areas of Statistics (see [38] for a summary). In network analysis, [29] investigated this idea in identifying instability in ERGM models that have particular degeneracies in their supports on the space of sufficient statistics, and a number of papers followed in a similar vein in the ERGM literature, e.g., [20]. These ERGM studies have treated stability of realized estimates with respect to small perturbations of the observed data. On the other hand, we are interested in a broader notion of stability, namely whether the *target* of estimation remains invariant

between samples from the same population that differ in an aspect that is ancillary to the underlying social process of interest. In this case, the size of the sample that the investigator chooses to analyze is the ancillary aspect. The stochastic process framework is a powerful tool for probing this type of instability, and represents a novel approach to this question within the networks literature.

With this framework, we show that the sparsity of social interaction networks induces an instability in inferences when the working model is sparsity-misspecified. We begin this discussion with a novel definition of sparsity, which we define as an asymptotic property of the network population process. This is in contrast to the single-sample networks literature, which has used a working definition of sparsity as a sample-wise property, saying that a given network sample is sparse if the fraction of nonzero interactions in the sample is small. Asymptotic arguments based on this definition do not appeal to a superpopulation, but instead reason by analogy about whether there is enough information in the small number of realized actions in a sample to reliably fit a model [3, 8]. Thus our superpopulation-oriented results about the instability of inferences from sparsity-misspecified models are qualitatively different from the consistency results that have appeared in the literature before.

Our instability result has major implications for the network modeling literature. A number of authors have shown that popular latent variable models for social interaction data do not capture network sparsity because their large-sample limits are dense [3, 23]. We show here that even under weaker misspecifications than these, generative network models will not produce model fits that are stable across samples sizes. There have been proposals for generative network processes that do achieve a population sparsity property. Many of these rely on additional information that makes actors non-exchangeable, for example the actors' order of entry into the network, and when this information is not available, require imputation in combinatorially large sample spaces [36]. In another vein, [7] present some novel work using a point process specifiction to achieve network samples that are sparse in some sense, but the mapping of this process to the conventional setting of having a network subsample with a known actor set is still not fully understood. In all of these process models, the specification of the underlying process places strong restrictions on the rate at which the network density falls to zero as samples become large, meaning that sparsity misspecification is still a major concern.

In the second half of the paper, instead of trying to model sparsity explicitly, we adopt a sparsity-invariant approach to modeling and inference. We propose dividing the network generating process into two stages, with one process that governs the sparsity of the network,

and a second process, defined conditioned on the first, that governs observable interactions. Given that the process that induces sparsity is difficult to model, we focus on drawing parametric inferences about the latter conditional interaction process. This approach was inspired by [27], in which the authors introduced the notion of a "risk set" to the networks literature, where the risk set defines the subset of dyads in a network sample that are "at risk" of producing observable interactions. There, the risk set was a vestigial piece of the authors' Cox proportional hazard model specification (in the original survival analysis context, the risk set is used to identify which patients in a study have not yet died or been lost to follow-up), and in their analysis, the authors chose to pre-specify the risk set as all dyads in the network sample, but referenced the possibility of inferring a non-trivial risk set instead. Here, we treat the risk set as an underlying social relationships that are pre-requisites to the generation of observable interactions.

To avoid the difficult question of modeling a sparse relationship structure, we propose an inferential approach that estimates the parameters of the conditional distribution of observed interactions without inferring or even specifying the marginal distribution of relationships on which they are conditioned. Instead, we condition on which dyads have produced nonzero interactions, and infer the parameters of the interaction process using the zero-truncated distribution. This approach is most generally a partial likelihood method [5, 37], although it can also be classified more specifically as a conditional likelihood [14] because we have chosen to condition on a statistic that isolates the parameters of the conditional distribution of interest. It is also possible, however, to view our zero-truncated approach as a *marginal* likelihood method, as introduced in [13], where we have chosen to ignore the actual sample size of the data and to marginalize over it instead. Both of these views are useful for characterizing the properties of our estimation procedure.

Proposals have appeared before in the networks literature to adjust network models to achieve inferential stability across sample size. [20] proposed an offset term that stabilizes change statistics in ERGMs, but did not attempt to justify this as a likelihood-based approach. [17] proposed generative models for the true observation in fixed rank nomination networks that the effect of removing sample-size dependent artifacts that appeared in previous naïve modeling approaches. Our approach here differs in that we use the truncated data model to create a likelihood-based adjustment that is completely agnostic to the process that induces sparsity in the network. Procedures similar to zero-truncation, including dyad subsampling and zero-inflation, have also been proposed in the literature before, but, rather than invariance to sparsity, these proposals have focused on single-sample fit [4], novel network representations [31], or approximate likelihood inference for computational efficiency

[15]. Notably, our proposed procedure is able to achieve similar computational efficiency using an exact likelihood function.

We demonstrate the stability and efficiency of our estimation procedure in analytical and simulated examples. Social scientific questions about the organizational behavior of inventors holding patents in the United States in, e.g., [24], were the original motivation of this work. In another series of papers, [10] and [12], we extend this modeling framework to the causal inference setting and use data from the US patent record made available by [21] to infer the causal effect of a policy change on the collaboration dynamics of inventors.

The paper is organized as follows. Section 2 outlines the theoretical framework for network superpopulation problems, and presents stability criterion for a model to be useful in a superpopulation investigation. Section 3 outlines our definition of sparsity, and introduces the notion of sparsity misspecification with examples of model classes that are misspecified in this way. Section 4 combines these ideas to show our main instability result. Section 5 and Section 6 present our sparsity-invariant modeling framework and inference procedures respectively. We present a analytical and computational examples in Section 7 and Section 8, and conclude with a discussion in Section 9.

1.3 Technical notes

Throughout, we assume that the investigator is employing maximum likelihood estimation, so we treat specifying a model and specifying an estimator as equivalent operations. We discuss potential generalizations of our results to other inference methods that map models to estimators differently in Section 9.

We focus exclusively on undirected network models. In the likelihoods of models of these networks, we simply write sums or products over ij but these can be taken to mean sums or products over $i < j < n$ if n is the size of the set.

2 Network Superpopulation Inference

The network analysis problems that we consider have the following form. We have a set of actors V, and a record Y_V of the pairwise interactions between them. In some cases, we also have an array of pairwise covariate vectors X. The outcome space of the interaction

records $\mathcal{Y}$ varies by the particular application – for example, the records may be binary, to indicate presence or absence of ties, count-valued, to indicate interaction counts, point-valued, to indicate timestamps of interactions, categorically-valued, to indicate relationship types, or some combination thereof. We indicate the outcome associated with a particular pair of individuals ij, where $i < j < n$, as Y_{ij}, and similarly the covariate vector associated with this pair as X_{ij}. We call Y_V an *interaction graph* and an observed instance of Y_V a *network sample*. We assume that the investigator proposes a generative model for Y_V and estimates the parameters of the model using maximum likelihood estimation (See Section 9 for discussion on generalizations to other estimation methodologies.).

We consider a social scientific application where superpopulation inference is the goal. That is, we consider the case whete the investigator's goal to obtain parameter estimates and predictive distributions from Y_V that can be used in downstream analyses that involve new actor sets V' drawn from a larger actor population. Such downstream analyses could include testing whether separate samples were drawn from a similar population by comparing parameter estimates, predicting interaction outcomes within a new actor set, or shrinking together estimates from separate samples in a hierarchical model. We call this "superpopulation inference" because any of these downstream analyses requires that notion that these exists some common, underlying structure that justifies propagating information from actor set to another. We define superpopulation inference in contrast to single-sample inference, where all downstream analyses are assumed to take place within the observed actor set V. These analyses might include imputing unobserved links within this actor set, or projecting the behavior of these actors forwward in time. These analyses only require a probabilty structure specific to the observed actor set V.

2.1 Network superpopulations

To formally characterise network superpopulation inference, we require a notion of a network superpopulation. In conventional i.i.d. settings, a superpopulation is defined as an infinite population from which a finite sample was drawn. Similarly, we define a network superpopulation as an infinite random graph from which we can obtain finite network samples by choosing finite subsets of actors and observing only those interactions that take place between them. Formally, we follow [30], and define a network superpopulation as an actor-indexed stochastic process. Here, we use slightly different notation to emphasize the relationship to the data analysis settting.

Let $\mathbb{V}$ be a countably infinite population of actors, so that each finite subset $V \subset \mathbb{V}$ corresponds to a set of actors whose interactions we could potentially observe. From this actor superpopulation, we define the interaction graph population as follows

Definition 1 (Random Interaction Process)**.** *A random interaction process $Y_{\mathbb{V}}$ is a stochastic process indexed by a countably infinite vertex set $\mathbb{V}$ whose finite-dimensional distribution for any finite subset $V \subset \mathbb{V}$ defines an interaction graph Y_V with vertex set V. Denote the law of $Y_{\mathbb{V}}$ as $\mathbb{P}_{\mathbb{V}}$ and the law of a finite-dimensional projection Y_V as $\mathbb{P}_V$.*

We emphasize that in a statistical argument, the network superpopulation plays a distinct role from sequences of increasingly large random networks that are often invoked in asymptotic analysis of network models for single-sample inference. In those cases, the limit of the sequence is meant to serve as an analogy for the structure within a single large network sample. Here, we invoke a network superpopulation to describe the structure *between* different network samples. Because the random interaction process is stochastic consistency across actor sets, the process simultaneously defines the distribution of interactions among any finite actor set $V \subset \mathbb{V}$ by projection. By defining coherent projections on differing actor sets, say V and V', the superpopulation specifies which aspects of the interaction graphs Y_V and $Y_{V'}$ remain invariant across actor sets.

We write the network superpopulation estimation problem as follows. Let $Y_{\mathbb{V}}$ be a random interaction process that is the superpopulation of interest, let $\mathbb{P}_{0,\mathbb{V}}$ be the law of the superpopulation process, and let $\mathbb{P}_{0,V}$ be the finite-dimensional distribution for the interaction graph Y_V of an actor set V. To estimate the law of the population process, we propose a model family $\mathcal{P}_{\Theta,\mathbb{V}} \equiv \{\mathbb{P}_{\theta,\mathbb{V}}\}_{\theta\in\Theta}$ indexed by (potentially infinite dimensional) parameter $\theta \in \Theta$, so that for each θ, $\mathbb{P}_{\theta,\mathbb{V}}$ is a population law. For any finite actor set $V \subset \mathbb{V}$, the population-level family implies a corresponding finite-dimensional model family. Let $\mathcal{P}_{\Theta,V} \equiv \{\mathbb{P}_{\theta,V}\}_{\theta\in\Theta,V\in\mathbb{V}}$ be the projected model family, where for each value of θ, $\mathbb{P}_{\theta,V}$ is a finite-dimensional distribution of $\mathbb{P}_{\theta,\mathbb{V}}$.

Operationally, maximum likelihood inference for superpopulation estimands proceeds identically to single-sample inference – to draw inferences from a particular observed interaction graph Y_V, we derive an estimator for θ from the projected model family $\mathcal{P}_{\theta,V}$ and we obtain an estimate $\hat{\theta}_V$ from Y_V. The superpopulation case only differs in that we specify and interpret the finite model for Y_V as a finite-dimensional projection of a superpopulation model, and thus interpret the estimate $\hat{\theta}_V$ as an estimate of the parameters of both a sample law $\mathbb{P}_{\hat{\theta},V}$ and a superpopulation law $\mathbb{P}_{\hat{\theta},\mathbb{V}}$. This interpretation translates practically into plugging $\mathbb{P}_{\hat{\theta}_V,\mathbb{V}}$ into downstream analyses (with accompanying uncertainty estimates), for example, testing

whether separate samples were drawn from a similar population by comparing parameter estimates, predicting interaction outcomes within a new actor set, or shrinking together estimates from separate samples in a hierarchical model.

2.2 Misspecification and superpopulation estimation

When the model we propose is misspecified, so that $\mathbb{P}_{\mathbb{V}} \notin \mathcal{P}_{\Theta}$, we cannot rely on the nominal interpretation of parameter estimates alone to draw scientific conclusions. Instead, we hope the estimator itself has some properties that can reveal some of the structure of the underlying process. One such property is *stability*, or an estimator's tendency to map similar generating processes to similar values in the parameter space.

Many notions of stability have been proposed in the Statistics literature, and it has played a particularly large role in the misspecification literature [38], although the idea has generally been presented in terms of consistency. Huber [18] most famously showed that the MLE converges to a point in the parameter space that Sawa [28] called the "pseudo-true" parameter, while [35] showed asymptotic normality. These results suggest that in large samples, while the MLE may not be directly interpretable, it is stable between replications of that sample. This notion underlies many of the asymptotic arguments made about the effectiveness of network models for single-sample estimation problems.

However, misspecification raises different concerns in superpopulation estimation problems than in single-sample estimation problems. In this case, we do not only require stability across replications of the same network sample Y_V; we also require stability across samples drawn from distinct actor populations. Thus, the convergence results for misspecified MLE's are not as useful for establishing the stability we require of the MLE in this case. Here, we will use different properties of the pseudo-true parameter to develop a notion of stability more appropriate for the superpopulation context.

For a given sample distribution $\mathbb{P}_{0,V}$ and sample model $\mathcal{P}_{\theta,V}$, the pseudo-true parameter $\bar{\theta}_V$ is given by

$$\bar{\theta} = \arg\max_{\theta \in \Theta} \mathbb{E}_{\mathbb{P}_0}[\log \mathbb{P}_{\theta,V}(Y)], \tag{1}$$

or the maximizer of the *expected* log-likelihood. Note that when the model is correctly specified, so that there is some parameter θ_0 such that $\mathbb{P}_{0,V} = \mathbb{P}\theta, V$, then $\bar{\theta} = \theta_0$. Even under misspecification, this quantity has many interpretations in the context of maximum

likelihood estimation. We can arrive at the pseudo-true parameter by inverting several desirable properties of an estimator. For example, recalling that Fisher consistency is one of the defining properties of the MLE, $\bar{\theta}$ is value in the parameter space Θ for which the MLE is Fisher consistent. Additionally, from the more general framework of estimating equations, $\bar{\theta}$ is the quantity for which the score equation defined by $\mathbb{P}_{\theta,V}$ is unbiased. Finally, it can be shown that the optimization in Equation 1 is equivalent to minimizing the KL divergence $KL(\mathbb{P}_0||\mathbb{P}_\theta)$ among all models in $\mathcal{P}_{\Theta,V}$ [28]. Thus, the effective estimand of the MLE can be interpreted as the KL projection of the true distribution of Y_V into the finite-dimensional model family $\mathcal{P}_{\Theta,V}$.

Most importantly, Spokoiny [32] established finite sample concentration inequalities for the MLE around the pseudo-true parameter $\bar{\theta}$ regardless of model misspecification or dependence in the distribution of Y_V. We can use this result to characterize the behavior of the MLE across finite samples indexed by different finite actor sets V. Heuristically, when we compute the MLE from a given sample Y_V, the estimator is effectively targeting the pseudo-true parameter $\bar{\theta}$. We find this intuition so compelling that we refer to $\bar{\theta}$ as the *effective estimand*.

Recall that the critical characteristic of the network superpopulation is that it encodes properties of the network generating process that are invariant across the set of actors V taht indexes the sample. It is therefore natural to expect that, if we are going to interpret a parameter estimate $\hat{\theta}_V$ as a superpopulation parameter estimate, the estimator should be estimating a quantity that does not depend on the index set V. This idea underlies the following superpopulation parameter stability criterion.

Proposition 1. *For parameter estimates to be interpreted as superpopulation parameters, for any finite sample Y_V, the effective estimand $\bar{\theta}_V$ of the estimator $\hat{\theta}_V$ should not depend on the indexing set V.*

Remark 1. *Note here that this criterion only puts a condition on the effective estimand $\bar{\theta}_V$, and not on $\hat{\theta}_V$ itself. For the purposes of this paper, we consider the semantic argument that an estimator that does not satisfy Proposition 1 is* estimating a different parameter *for different index sets V to be powerful enough to motivate an investigator to seek a different modeling strategy. In a separate paper [11] in which we treat the "effective estimand" idea in greater generality and detail, we push through the large deviation bounds in [32] to describe how violations Proposition 1 affect the distribution of $\hat{\theta}$ and the properties of downstream inferences. However, in our simulation study in Section 8 we do see that the parameter vector estimate $\hat{\theta}$ does concentrate around the effective estimand $\bar{\theta}$ as expected.*

The criterion in Proposition 1 is not always directly verifiable, because computing the effec-

tive estimand $\bar{\theta}_V$ requires computing an expectation over the true distribution $\mathbb{P}_{0,V}$. However, in the case of social network modeling, there are often known properties of the true social process that we were not able to encode directly into the model, but which we can use to check whether the effective estimand could remain stable if $\mathbb{P}_0$ had this property. If the effective estimand changes with V, this suggests that the model may not be useful for super-population inference. In the following sections, we show that this criterion is violated when the proposed network model does not account for the sparsity property of social networks.

3 Sparsity

Sparsity is one of the most salient features of social networks. In this section, we will formally define this property so that we can characterize the behavior of the effective estimand of the MLE when the true social process is sparse. As defined specifically in the context of networks, sparsity is the phenomenon that, in large social interaction network samples, an overwhelming proportion of actor-pairs engage in no interactions, and that the larger the network sample is, the more dominating this proportion of zeros becomes. Formally, we represent this by encoding the social interactions in an outcome space $\mathcal{Y}$ in which one particular value in this space that corresponds to "no interaction", which we will call 0. In the case of binary or count-valued outcomes, this is simply the number 0, while in the case of timeseries of point-valued outcomes, this may correspond to the timeseries that is identically 0 at every point in the observation interval.

Sparse graphs have been a common topic in both the Probability and Statistics literatures. Bickel and Chen [3] and Bollobás et al [19], among others have approached sparsity in terms of sequences of distributions over random graphs of growing size or expected size. Notably, these definitions do not constrain the sequence to be Kolmogorov consistent, and so elements of the sequence cannot be understood to be drawn from the same population process. Instead, the limits of these sequences are meant to serve as analogies for single network samples with a relatively small number of observed ties. In short, existing definitions of sparsity have been more amenable to describing the properties of single network samples rather than the properties of the underlying superpopulation process.

Because we wish to focus on superpopulation questions, we require a categorically different different definition of sparsity. For ease of discussion, we define a density operator, which corresponds to the proportion of dyads in an interaction graph with corresponding nonzero interactions.

Definition 2 (Density Operator)**.** *Let Y_V be an interaction graph with vertex set V. Fix an element of the outcome space $\mathcal{Y}$ to be zero, denoted by $\mathbf{0}$, and define the indicator random variables $A_{ij} = \mathbb{I}_{\{Y_{ij} \neq \mathbf{0}\}}$ for each $i < j \in V$.*

The density operator D with respect to the element $\mathbf{0}$ has the form

$$D(Y_V) = \frac{\sum_{ij} A_{ij}}{\binom{|V|}{2}},$$

giving the proportion of components of Y_V that are non-zero.

Intuitively, a population process is sparse if, as we sample additional vertices from the population process, the expected density of the sampled interaction subgraph converges to zero. Formally,

Definition 3 (Sparse Interaction Graph Process)**.** *Let $Y_{\mathbb{V}}$ be an interaction graph process on $\mathbb{V}$. $Y_{\mathbb{V}}$ is* sparse *if and only if for any $\epsilon > 0$ there exists an n such that for any subset of vertices $V \in \mathbb{V}$ with $|V| > n$ the corresponding finite dimensional interaction graph Y_V has the property $\mathbb{E}(D(Y_V)) < \epsilon$.*

A key consequence of this definition is that any increasing subgraph sequence defined with respect to a sparse random graph process has a sparse limit, i.e., for any increasing sequence of vertex sets (V_n) ordered by subset inclusion, $D(Y_{V_n}) \to 0$ as n grows large. This property is invariant to the scheme used to construct the subgraph sequence.

It is also useful to define the *sparsity rate* of a process, which characterizes how quickly the densities of growing samples drawn from a given population process converge to zero.

Definition 4 (Sparsity Rate)**.** *Let (V_n) be an increasing sequence of vertex sets ordered by subset inclusion. We say an interaction graph process $Y_{\mathbb{V}}$ has sparsity rate $\epsilon(n)$ iff there exists some finite positive constant C such that for any sequence (V_n),*

$$\frac{\mathbb{E}[D(Y_V)]}{\epsilon(n)} \to C$$

as $n \to \infty$. Similarly, we say interaction graph processes defined on the same index set $\mathbb{V}$, $Y_{\mathbb{V}}$ and $Y'_{\mathbb{V}}$, have the same sparsity rate iff there exists some finite positive constant C such that for any sequence (V_n),

$$\frac{\mathbb{E}[D(Y_V)]}{\mathbb{E}[D(Y'_V)]} \to C$$

as $n \to \infty$.

3.1 Sparsity misspecification

Sparsity is an attribute of real-world social networks that may or may not be well-represented by a generative network model. When the sparsity of the real process is not correctly represented by the inferential model, we say that the model is *sparsity misspecified*. Formally, sparsity misspecification occurs when there is no member of the inferential model family with the same sparsity rate as the true superpopulation process. Formally,

Definition 5 (Sparsity Misspecification)**.** *For an inferential family $\mathcal{P}_{\Theta,\mathbb{V}}$ and true population process $\mathbb{P}_{0,\mathbb{V}}$, we say that the inferential family is sparsity misspecified if*

$$\frac{\mathbb{E}_\theta[D(Y_V)]}{\mathbb{E}_0[D(Y_V)]} \to 0 \text{ or } \infty \quad \forall \beta \in B, \gamma \in \Gamma, \tag{2}$$

where $E_{\beta,\gamma}$ and E_0 are expectations taken with respect to $P_{\beta,\gamma,n}$ and $P_{0,n}$, respectively.

We give several examples of sparsity misspecification in the remainder of this section.

3.2 Exchangeable random graph models with covariates

Sparsity misspecification is particularly prominent in model families that are built on local assumptions about how individual actors make decisions to interact. Exchangeable random graph models form the most prominent class of such generative network models. These models assume that the likelihood for the interaction network is invariant to permutations of the actors in the network – this translates to joint exchangeability of the rows and columns of the adjacency matrix. We consider the extension of these models to the case where actors are exchangeable up to observed covariates. These models are appealing because they imply that an observed network sample can be decomposed into a set of conditionally independent replications given sets of observed and potentially unobserved covariates. They also imply that independent predictions can be made at the dyad level.

We pay special attention to the simplest subclass of exchangeable random graph model that treats all pairwise outcomes in the network as conditionally independent given observed pairwise covariates. These models reduce network generation problem to a regression problem on the vectorized adjacency matrix. Generally, these models are specified as a generalized linear model, and have been proposed with binary, count-valued, and point process-valued outcomes (see, for example, [27, 34, 16, 33]). These assign a particular observed social

interaction graph Y_V with covariates X_V a likelihood of the form:

$$P(Y_V \mid X_V) = \prod_{ij} P(Y_{ij} \mid X_{ij}). \tag{3}$$

We call models in this subclass *conditionally independent dyad* or CID models. This model class subsumes models that assume node-level covariates, as these can be encoded as dyad-level covariates.

Exchangeable random graph models also include more complex specifications that assume conditional independence between the dyads given *unobserved* covariates. These models have seen an explosion of interest with a wide variety of structures proposed for the latent covariate structure including latent single- and mixed-membership classes, latent positions, latent eigenspaces, and their infinite- dimensional counterparts [23]. This class of models has been unified under an array-exchangeability representation by Aldous and Hoover that, up to permutation, maps these latent covariate processes to a single surface X on the unit square. Given this surface, the network is generated by randomly assigning each actor a position C_i so that the pairwise covariate for dyad ij is generated by querying $X(C_i, C_j)$. Several recent works have been dedicated to estimating this latent surface, called the graphon, directly [6, 1]. Models with this structure induce the following likelihood on network samples

$$P(Y \mid X) = \int_{\mathcal{C}} \prod_{i<j<n} P(Y_{ij} \mid X(C_i, C_j)) dF(C). \tag{4}$$

Model specifications that mix latent and observed covariates have also been proposed in several places, e.g., [16].

Several authors have noted that exchangeable graph models without covariates cannot be extended to form sparse graph processes, or, in our current terminology, that these models are sparsity misspecified. Orbanz and Roy state this most clearly: "If a random graph is exchangeable, it is either dense or empty." [26] They justify this statement with a simiple law of large numbers argument. With appropriate conditions on observed covariates X, we can extend this result to exchangeable random graph models with covariates, including CID models.

Theorem 1. *For a an exchangeable interaction process model $\mathcal{P}_{\Theta,\mathbb{V}}$, and corresponding covariate set $X_{\mathbb{V}}$, denote by $\mathcal{N}_\theta \subset \mathcal{X}$ the set of covariate vectors so that for a given $\theta \in \Theta$, $\mathbb{P}_\theta(Y_{ij} \neq 0 \mid X_{ij} \in N_\theta) = 0$. Assume that for each $\theta \in \text{int}\Theta$, the limiting proportion of covariate vectors $X_{ij} \in \mathcal{N}_\theta$ is $1 - \epsilon$ for some nonzero ϵ. If this is the case, the model is*

sparsity misspecified.

The argument here is straightforward. The covariate vector X_{ij} simply parameterizes the surface W described by Aldous and Hoover, so that every X_{ij} defines a corresponding surface $W_{X_{ij}}$. The marginal probability $\mathbb{P}(Y_{ij} \neq 0 \mid X_{ij})$ is the integral of $W_{X_{ij}}$. Thus, if the limiting proportion of covariate vectors that define a zero-integral latent surface W does not converge to 1, then for some positive proportion of dyads, we will have latent surfaces with positive integrals so that $\mathbb{P}(Y_{ij} \neq 0 \mid X_{ij})$ for these dyads, resulting in a limiting positive network density by LLN.

Intuitively, unless the model is able to *a priori* exclude an arbitrarily high proportion of dyads from interaction on the basis of the observed covariates X_{ij}, it will be sparsity misspecified. In most social network analysis applications, such a highly informative set of covariates is not available – in fact, regression, latent variable, or combined modeling schemes are often proposed precisely because so little is known a priori about the network's structure.

In this case, it is possible to confirm sparsity misspecification *a priori* because these model families only incude social interaction processes whose densities converge to positive constants. Other non-exchangeable model families, for example preferential attachment, do allow for sparse extensions and in these cases, it may be difficult to determine from a finite sample whether a family is sparsity-misspecified. Families that include flexibility in parameters the determine sparsity rate, for example the power in a power law distribution, may not be sparsity misspecified, but checking this is important. Many families that allow sparse extensions lock in a particular functional form for the sparsity rate, which increases the risk of sparsity misspecification.

4 Main Result: Moving Target Theorem

In the last few sections, we have established a statistical framework for representing superpopulation inference, discussed conditions under which an estimate from a misspecified model can be interpreted as a superpopulation quantity, and identified sparsity misspecification as a common issue in network modeling. In this section, we bring these ideas together and show that most MLE's derived from sparsity-misspecified models do not admit a superpopulation interpretation because their effective estimand is not invariant across samples drawn from the same population.

We introduce one final definition before we proceed to the theorem.

Definition 6 (Responsiveness)**.** *Let* (V_n) *be an arbitrary increasing sequence of vertex sets from* $\mathbb{V}$*, ordered by subset inclusion. We say an estimator is* responsive *to a statistic* $T(Y_V)$ *under a true generating process* $\mathbb{P}_{0,V}$ *if and only if*

$$|\mathbb{E}_{\bar{\theta}_{V_n}}(T(Y_{V_n})) - \mathbb{E}_0(T(Y_{V_n}))| = o_p(1), \tag{5}$$

for any (V_n)*, or when the distribution indexed by the effective estimand gives an asymptotically unbiased prediction for the statistic* $T(Y_V)$*.*

Note that responsiveness is generally considered a minimum requirement for an estimator. It implies that the estimator's plug-in distribution yields an asymptotically unbiased prediction of the test statistic.

When a sparsity misspecified model is responsive to the network density $D(Y_{V_n})$, we can show that the the MLE does not estimate a population parameter because, while a population parameter remains invariant across samples from the same population, the effective estimand varies as a function of the size of V_n. In essence, if $\mathcal{P}_{\Theta,\mathbb{V}}$ is sparsity misspecified, but the members of $\mathcal{P}_{\Theta,\mathbb{V}}$ are able to provide good pointwise approximations to $\mathbb{P}_{0,V_n}$ for each n, then the fact that $\mathcal{P}_{\Theta,\mathbb{V}}$ is sparsity misspecified implies that the members that provide these appoximations at different sample sizes are necessarily different.

Theorem 2 (Moving target theorem)**.** *Let* (V_n) *be an arbitrary increasing sequence of vertex sets from* $\mathbb{V}$*, ordered by subset inclusion. Suppose that the following hold:*

1. *The inferential family* $\mathcal{P}_{\Theta,\mathbb{V}}$ *is sparsity misspecified for the true population process* $\mathbb{P}_{0,\mathbb{V}}$*.*
2. *The marginal distribution of the binarized data* A *identifies* θ *in* $\mathcal{P}_{\Theta,\mathbb{V}}$*.*
3. *The inferential model is responsive to the sample density* $D(Y_{V_n})$ *under the true population process and*

$$|\mathbb{E}_{\bar{\theta}_{V_n}}(D(Y_{V_n})) - \mathbb{E}_0(D(Y_{V_n}))| = \delta(n). \tag{6}$$

4. *The rate of the effective estimand's plug-in prediction bias* $\delta(n)$ *and the sparsity rate* $\epsilon_0(n)$ *of* $\mathbb{P}_{0,\mathbb{V}}$ *are such that, for some finite constant* C*,*

$$\frac{\delta(n) + \epsilon_0(n)}{\epsilon_0(n)} \to C. \tag{7}$$

Then, $\bar{\theta}_{V_n}$ varies with n in the sense that for any n, there exists an $n' > n$ such that $\bar{\theta}_{V_n} \neq \bar{\theta}_{V_{n'}}$

Proof. Because the effective estimand's plug-in prediction bias for the network density $|\mathbb{E}_{\bar{\theta}_{V_n}}(D(Y_{V_n})) - \mathbb{E}_0(D(Y_{V_n}))|$ is of equal or smaller order than the sparsity rate of $\mathbb{P}_{0,\mathbb{V}}$, $\mathbb{E}_{\bar{\theta}_{V_n}}(D(Y_{V_n}))$ conveges to zero at rate $\epsilon_0(n)$. But because the family $\mathcal{P}_{\Theta,\mathbb{V}}$ is sparsity misspecified, there is no single model $\mathbb{P}_{\theta,\mathbb{V}}$ whose expected densities can follow this rate. Thus, for any n, there exists an n' such that $\bar{\theta}_{V_n} \neq \bar{\theta}_{V_{n'}}$. □

This result implies that for network superpopulation inference problems, sparsity misspecified models violate Proposition 1. In particular, it shows that estimators computed from samples of differing size are effectively estimating distinct quantities even if they are drawn from the same network superpopulation. This, in turn, implies downstream analyses of these estimates that rely on a stable notion of a network superpopulation, for example, hypothesis tests or shrinkage schemes, are ill-defined. Even in cases where the desire is to simply interpret the parameter estimates for theoretical context, this inhomogeneity of interpretation with respect to size presents challenges when applying models that were developed for analysis of small networks (e.g., Sampson's monastery) to large-sacle social networks. Depending on the application, establishing a meaningful scale for such parameter estimates may not be possible.

We demonstrate these effects with the Poisson regression example introduced above.

4.1 Example: Poisson regression with binary covariate

Let there be a superpopulation of inventors, from which we have sampled n individuals. Let Y_{V_n} be a matrix recording the number of pairwise patent collaborations that have taken place between the n sampled inventors, so that Y_{ij} is the number of times inventor i and inventor j appeared together on the same patent application. Denote the true distribution of Y_{V_n} as $\mathbb{P}_{0,V_n}$. For each entry ij of Y_{V_n}, let X_{ij} be a binary covariate that indicates whether inventors i and j work for the same firm. The investigator is interested in how this covariate is related to the outcome Y_{V_n}, and in particular in comparing this outcome across networks, to make statements about whether within-firm collaborations are more prominent in one industry than another. The investigator is also aware that the true interaction process $\mathbb{P}_{0,\mathbb{V}}$ is sparse.

Despite this knowledge, the investigator proposes a finite-dimensional model family $\mathcal{P}_{\Theta,V_n}$

for the sample that has the form of a Poisson regression model

$$Y_{ij} \overset{\perp\!\!\!\perp}{\sim} \text{Pois}(\exp(\theta_0 + X_{ij}\theta_1)) \qquad \forall i < j < n, \tag{8}$$

where the parameter vector $\theta \equiv (\theta_0, \theta_1)$ can take values in $\Theta \equiv \mathbb{R}^2$. According to standard interpretations of GLM coefficients, θ_0 is the log of the interaction rate of any "between-firm" inventor pair, while θ_1 is the log ratio of interaction rates between any "within-firm" and any "between-firm" inventor pair.

Assume as we have throughout that we have a fixed increasing sequence of sets of actors (V_n), ordered by subset inclusion. For the purposes of this problem, we also assume we have a corresponding sequence of covariate arrays (X_n) associated with each actor set in (V_n).

Because this is an exponential family, the effective estimand has a particularly appealing form that mimics the form of the MLE:

$$\bar{\theta}_{0n} = \log\left(\frac{\sum_{ij}\mathbb{E}_0(Y_{ij} \mid X_{ij} = 0)(1 - X_{ij})}{\sum_{ij}(1 - X_{ij})}\right) \tag{9}$$

$$\bar{\theta}_{1n} = \log\left(\frac{\sum \mathbb{E}_0(Y_{ij} \mid X_{ij} = 1)X_{ij}}{\sum X_{ij}} \Bigg/ \frac{\sum \mathbb{E}_0(Y_{ij} \mid X_{ij} = 0)(1 - X_{ij})}{\sum(1 - X_{ij})}\right). \tag{10}$$

We maket he following assumptions about the true process $\mathbb{P}_{0,\mathbb{V}}$ to characterize the effective estimands. Consider the case where $\mathbb{E}_0(Y_{ij} \mid X_{ij})$ is finite for all ij, and that for some finite $n'' > n'$, the expected number of within-firm and between-firm interactions are nonzero.

Under these assumptions, consider how $\bar{\theta}_n$ behaves as we allow n to vary. As n becomes large, one or both of the between-firm and within-firm groups of dyads must grow at the same rate as the total number of dyads, that is, at least one of $\sum_{ij}(1 - X_{ij})/\binom{N}{2} \to c > 0$ and $\sum_{ij} X_{ij}/\binom{N}{2} \to c > 0$ must hold. WeLOG, assume that the number of between-firm dyads scales with the total number of dyads.

According to the sparsity assumption, the expected proportion of nonzero dyads among the between-firm dyads converges to zero with rate $\epsilon_0(n)$. It can also be shown that the ratio in Equation 9 converges to 0 with rate $\epsilon_0(n)$, and by Taylor expanding the expression for $\mathbb{E}_{\bar{\theta}}(D(Y_{V_n}))$, we can see that the difference with $E_0(D(Y_{V_n}))$ converges to zero with rate $\epsilon_0(n)$. Thus, this model is responsive to $D(Y_{V_n})$ with the appropriate rate, and the moving target theorem applies. Indeed, because $\bar{\theta}_{V_n}$ is a function of a ratio with rate $\epsilon_0(n)$, that $\bar{\theta}_{V_n}$ varies with n is clear even without this general result.

In particular, as the proportion in Equation 9 falls to zero, $\bar{\theta}_{0n} \to -\infty$. Indeed, the nested structure of the problem carries forward the sample averages in Equation 9 and Equation 10 as we increment n, so $\bar{\theta}_{0V_n}$ traverses all of the values between its finite value at n'' and $-\infty$ rather smoothly in n, so we expect that for all finite samples with $n > n''$, the value of $\bar{\theta}_{0V_n}$ will depend strongly on n.

This is enough to show that the effective estimand varies with sample size, but we can also examine $\bar{\theta}_{1V_n}$. Depending on whether the within-firm group of dyads grows at a slower rate or the same rate as the between-firm group of dyads, $\bar{\theta}_{1V_n}$ may either diverge or converge to an arbitrary constant depending on the exact growth rate of the between-dyad group. In the latter case, $\bar{\theta}_{1V_n}$ can show convergent behavior, so that the effective estimand could serve as a meaningful quantity for scientific comparison. However, this behavior relies on a strong assumption about covariate behavior that may be difficult to confirm in practice.

5 Conditionally Independent Relationship Models

So far, we have established that sparsity misspecification is difficult to avoid and that sparsity misspecified models have little hope of yielding scientifically meaningful parameter estimates for superpopulation inquiries. As a solution to this problem, we propose a model class that has a similarly intuitive structure to CID or exchangeable random graph models, but which explicitly separates some superpopulation quantities of interest from the sparsity of the network generating process. By redefining the problem in this way, we obtain a set of summaries of network samples that we can expect to remain stable between samples of different size, regardless of the sparsity of the true network generating process. We call this class of models *conditionally independent relationship*, or CIR models.

For our discussion of CIR models, we shift notation slightly. To make our notation more compact, WeLOG we fix a particular increasing sequence (V_n) of vertex sets, ordered by subset inclusion, and only write the subscript n instead of V_n.

In CIR models, we assume that the dyad-level observations are drawn from a zero-inflated process. This corresponds to the generative intuition that in order to support an observable interaction, two actors must first have an unobservable social relationship. Conditional on this set of relationships, observed interactions are assumed to have the same conditional independence structure as CID models, and to have a distribution goverened by the paramter β. See Figure 2.

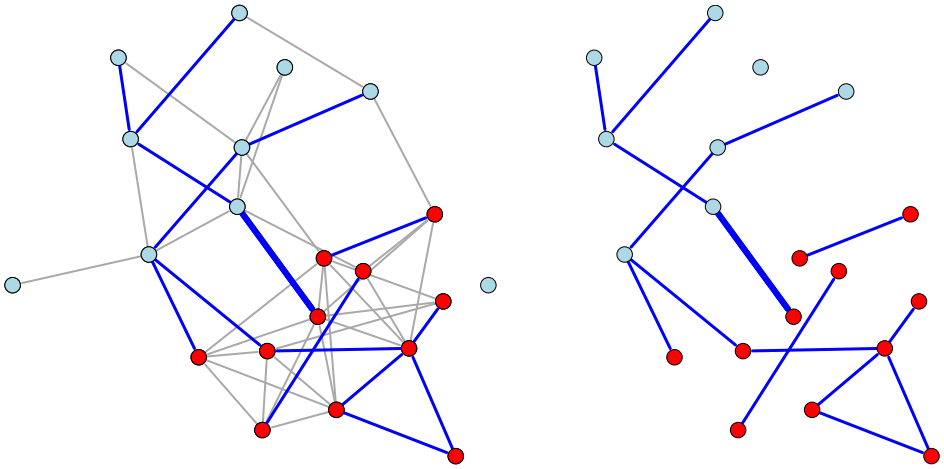

Figure 2: (Left) Diagram of CIR generation process, where gray ties are "relationships" and blue ties are observed interactions. To generate in observable interaction, a pair of actors much first have a relationship. (Right) The observed network sample, where relationships with no observed interactions are indistinguishable from dyads with no relationship.

Formally, we write the law of the observable interaction process Y in terms of an unobservable interaction process R, which we call the *relationship process*. R is itself a binary random interaction process, with finite-dimensional samples R_n, which we call *relationship graphs*. To generate a network sample Y_n from law $\mathbb{P}_n$, we follow two-stage generating process, where a relationship graph R_n is drawn first, and conditional on this graph, the observered interactions Y_{ij} are drawn independently. For each ij where $R_{ij} = 0$, Y_{ij} is set deterministically to 0. Thus, while R_n itself is unobserved, it is known that $R_{ij} = 1$ whenever $Y_{ij} > 0$.

We parameterize CIR inferential families with a parameter vector $\theta = (\beta, \gamma)$ that we divide into parameters of interest β and nuisance parameters γ. The modeling family assumes that the conditional process $Y \mid R$ has a simple structure parameterized *only* by the parameters of interest β, while the relationship process R has a complex structure that may be parameterized by the entire vector θ, including the potentially infinite-dimensionsal nuisance parameter vector γ. We write the marginal likelihood for Y_n as the joint likelihood of R_n and Y_n with the unobserved elements of the latent process integrated out. Thus, the likelihood for Y_n becomes:

$$\mathbb{P}_{\beta,\gamma}(Y_n \mid X_n) = \sum_{\boldsymbol{R}} \mathbb{P}_\theta(R_n \mid X_n)\mathbb{P}_\beta(Y_n \mid X_n, R_n). \tag{11}$$

$$= \sum_{\boldsymbol{R}} \mathbb{P}_\theta(R_n \mid X_n) \prod \mathbb{P}_\beta(Y_{ij} \mid R_{ij} = 1, X_{ij})^{R_{ij}}. \tag{12}$$

Note that CID models are a special case of CIR models that specify R as an independent

process, so that each entry of R is conditionally independent. The network regression models referenced above take this a step further and specify R as a trivial process where each entry of R is deterministically 1.

The CIR structure decouples the conditional interaction process $Y \mid R$ from the sparsity rate of Y process. In particular, because we assume the conditional process $Y \mid R$ is conditionally independent between dyads, it only contributes a constant factor to the sparsity rate of the marginal process Y, while the law of R controls the functional form of the sparsity rate. Thus, an estimate of the parameter vector β based on the conditional process $Y \mid R$ has the potential to define a summary of a network sample Y_n that is stable across sample sizes.

Writing down an estimator of β with the desired sparisty-invariant property is non-trivial. If R were observable, we could obtain an MLE for β based only on the conditional distribution $Y \mid R$ without having to model the relationship process. However, because the relationships status R_{ij} for a given dyda ij is unknown if $Y_{ij} = 0$, a model for R is necessary to marginalize over these missing components. Unfortunately, because the marginal model for Y has the same sparsity rate as the model for R, and defining any model with the correct sparsity rate is diffucult, defining the problem in terms of a CIR family alone does not circumvent the sparsity misspecification problem. We also require a different inferential method is able to isolate the information in the conditional process $Y \mid R$ from any specifiation of the relationship process R. We develop such a method in the next section.

6 Truncated Inference for CIR models

In the previous section, we identified the conditional interaction process $Y \mid R$ as a sparsity-invariant portion of a network generating process, but noted that inferring any parameters of this process using the MLE would require the correct specification of a sparse process R. Our desire for inference that is invariant to this nuisance process R motivates a partial likelihood approach. Partial likelihood, proposed by Cox [9, 5], and rigorously treated by Wong [37], is an approach to estimation problems where the full likelihood is parameterized by a potentially infinite-dimensional parameter vector, but the parameters of interest are a finite subvector. Partal likelihood approaches estimate the parameter of interest in a way that is invariant to the nuisance parameters.

In general, to use this approach, we apply a transformation to the sequence of data $Y \mapsto$

(W, V) so that the full likelihood can be factorized as

$$P_\theta(Y) = \left[\prod_{i=1}^{n} P_\theta(w_i \mid v_{1,\cdots,i-1}, w_{1,\cdots,i-1})\right]\left[\prod_{i=1}^{n} P_\beta(v_i \mid v_{1,\cdots,i-1}, w_{1,\cdots,n})\right]. \tag{13}$$

When the above factorization holds, we can use only the second factor in Equation 13 to construct a *partial likelihood* for β. Maximizing the partial likelihood yields a maximum partial-likelihood estimator for β.

$$\hat{\beta}_n^{MPLE} = \arg\max_\beta \prod_{i=1}^{n} P_\beta(v_i \mid v_{1,\cdots,i-1}, w_{1,\cdots,i}) \tag{14}$$

Because the partial likelihood has no dependence on the nuisance components of θ, this estimation procedure is invariant to the portion of the data-generating process depending on these components of θ, at the cost of some efficiency.

By constructing a partial likelihood, we aim to design an inference procedure for estimating β from a model for $Y \mid R$ without specifying a model for R. In Section 6.1, we show that the structure of CIR models makes them amenable to a particularly simple partial likelihood approach. We call this approach *truncation* because the factor of the likelihood is the zero-truncated conditional interaction distribution – i.e., the conditional distribution of interactions Y_{ij} given that they are nonzero. In Section 6.2 we give conditions under which this estimatior is well-behaved and in Section 6.3 we compare the efficiency of the partial likelihood estimator to a correctly specified, full-likelihood estimator.

6.1 Derivation

Because our goal in proposing CIR models was to decouple the process controlled by β and the process that controls the sparsity of the data, it would be natural to split the data along these lines. Because the density of the sample only depends on the indicators $A_{ij} = \mathbf{1}_{Y_{ij}>0}$ for each ij, we apply the transformation $Y \mapsto (A, \{Y_{ij} : A_{ij} = 1\})$. Writing the likelihood

Equation 11 in the form of Equation 13, we obtain

$$P_\theta(Y \mid X) = \left[\prod_{ij}\sum_{R_{ij}} P_Q(R_{ij} \mid A_{kl<ij}, X) P_\beta(A_{ij} \mid R_{ij}, X_{ij})^{R_{ij}}\right]\left[\prod_{ij} P_\beta(Y_{ij} \mid A_{ij}, X_{ij})^{A_{ij}}\right] \tag{15}$$

$$= \left[\prod_{ij} P_\theta(A_{ij} \mid A_{kl<ij}, X)\right]\left[\prod_{ij} P_\beta(Y_{ij} \mid A_{ij}, X_{ij})^{A_{ij}}\right] \tag{16}$$

Rewritten in this way, we see that the second factor depends only on β, with no dependence on the nuisance parameter Q. This follows because R_n, and thus Q, only affects the distribution of Y_n through the matrix A_n – that is, the existence of a relationship only affects the probability that the interactions along a dyad are nonzero, not the whole interaction distribution. Thus, conditional on A_n, the distribution of Y_n does not depend on Q, yielding the desired invariance to the relationship process.

For convenience, we define the following shorthand for the pieces of Equation 16:

$$f_\beta(Y_n \mid A_n) \equiv \prod_{ij} P_\beta(Y_{ij} \mid A_{ij}, X_{ij})^{A_{ij}} \tag{17}$$

$$g_\theta(A_n) \equiv \prod_{ij} P_\theta(A_{ij} \mid A_{kl<ij}, X_n) \tag{18}$$

These are, respectively, the conditional probability of Y given A, and the marginal probability of $\boldsymbol{A}$. We call $f_\beta(Y \mid A)$ the *truncated likelihood*. Using this notation, maximum partial likelihood estimator of β is given by

$$\hat{\beta}_n^{MPLE} = \arg\max_\beta f_\beta(Y_n \mid A_n), \tag{19}$$

or maximizing the truncated likelihood. We refer to this estimator as the *truncated estimator*.

6.2 Characteristics of the truncated estimator

To characterize the behavior of the truncated estimator, we focus on the case where the true population P_0 exhibits a CIR structure, with an arbitrary relationship process Q_0 and a conditional independence structure for the Y_{ij}'s conditional on a sample path R from the relationship process. As before, we assume that P_0 is sparse. We also assume that the inferential conditional model for $Y_{ij} \mid X_{ij}, R_{ij}$ is correctly specified. We find this to be a

reasonable assumption because under the conditional independence structure, the form of the conditional distribution would be straightforward to check and modify until a reasonable specification is achieved.

In this setting, we note first that by definition, the truncated likelihood cannot be sparsity misspecified. For any sample size n, upon conditioning on A_n, we have restricted our attention to a set of dyads where $R_{ij} = 1$ for each included dyad. While the model and the true generating process may disagree in some ways, because the model does not make a statement about the sparsity of the observable process Y (this is determined by the unspecified process for the relationships R), it cannot disagree on a sparsity rate.

Although we are concerend primarily with superpopulation inference in this paper, it is also useful to note that the truncated estimator has desirable single-sample properties – ideally an estimator with both of these properties in superpopulation inference problems so that we can establish that estimates that are stable across indexing sets also encode useful summaries of the individual samples. For the rest of this subsection, we discuss some single-sample properties of the truncated estimator. Note, however, that we assume the same stochastically consistent asymptotic frame that we introduced in Section 2, so all sequences in this section correspond to increasing subgraphs, indexed by n, drawn from the same network superpopulation..

Overall, the truncated likelihood corresponds to an especially simple case of partial likelihood first treated in the literature as "conditional likelihood" by [22, 2, 14], which correspond to full likelihoods of derived sub-experiments of the original data generation (Cox). In this case, consider an experiment with a sampling mechanism that only records dyads with observed interactions and ignores all other dyads. The full likelihood for this experiment would correspond exactly to the truncated likelihood derived from the original experiment.

Heuristically, because the conditioning on A_n induces mutual independence between each Y_{ij}, and also independence between Y_{ij} and $A_{jk\neq ij}$, the truncated likelihood has the form of a likelihood with independent observations, and can thus be shown to be well-behaved when the truncated likelihood f_β is shown to meet any standard set of sufficient conditions (e.g., Doob, Wald, Wong) with one modification. One major difference between this and the standard likelihood setting is that, if we treat the truncated likelihood as a conventional likelihood for a sub-experiment, the sample size of the sub-experiment is a random variable. To obtain standard consistency and asymptotic normality properties, we must add an extra condtion that this effective sample size goes to infinity.

The effective sample is made up dyads for which the conditional distribution $f_\beta(Y_{ij} \mid A_{ij}, X_{ij})$ is non-degenerate. The conditional distribution of Y_{ij} can be degenerate for one of two reasons – either $A_{ij} = 0$, in which case Y_{ij} is trivially 0, or the distribution of Y_{ij} given that it is nonzero is degenerate, as is the case when Y_{ij} is binary. Let B_{ij} be an indicator for dyads where this second kind of model-based degeneracy occurs. Then $M_n = \sum_{ij}(1 - A_{ij})B_{ij}$ is the effective sample size for the truncated likelihood.

To apply standard consistency or asymptotic normality conditions, we only need to add the condition that $M_n \to \infty$ where the type of convergence is dictated by whether the investigator wishes to have prove weak or strong consistency. Convergence rates for these results can be stated in terms of M_n. In standard cases where all Y_{ij} have non-degenerate nonzero distributions – for example, if the Y_{ij} are Poisson distributed – the rate at which M_n grows in n is dictated by the sparsity rate of the relationship process R_0.

To show how such an argument can put pushed through in a rigorous manner, we give some details of a consistency argument, modeled after Wong's treatment of the more general partial likelihood case.

6.2.1 Example: Consistency

In this section, we lay out the conditions for consistency of the MPLE in general, and provide some simplifications of these conditions for the truncated likelihood.

Let $\beta_0 \in B$ be the true value of the conditional interaction parameter. Following Wong [37], we define the following quantities:

$$R_N(\beta) = \sum_{ij<N} \log(f_{\beta_0}(Y_{ij} \mid A_{ij})/f_\beta(Y_{ij} \mid A_{ij})) \tag{20}$$

$$I_N(\beta) = \mathbb{E}_{g_{\beta_0}} R_N \tag{21}$$

$$J_N(\beta) = \mathbb{V}\mathrm{ar}_{g_{\beta_0}} R_N \tag{22}$$

$$M_N(\beta) = R_N(\beta) - I_N(\beta) \tag{23}$$

$R_N(\beta)$ is the observed log-likelihood ratio up to observation N, $I_N(\beta)$ and $J_N(\beta)$ are the expectation and variance of this quantity taken with respect to the complete data model, respectively, and $M_N(\beta)$ is the residual. Further, we define the lower-case quantities $r_{ij}(\beta), i_{ij}(\beta), j_{ij}(\beta), m_{ij}(\beta)$ to denote the individual terms of the corresponding sums above. We can now restate Wong's consistency theorm in the context of our problem.

Theorem 3 (Consistency of Partial Likelihood Estimator (Wong 1986))**.** *Suppose we have a partial likelihood given by $f_\beta(Y_{ij} \mid A_{ij})$, and true parameter β_0. Suppose for each $\beta \neq \beta_0$ there exists an open neighborhood G_β whose closure does not contain β_0 and that there are constants $\delta > 0$, $\zeta_N \uparrow \infty$ (which may depend on θ) such that*

1. $P(\inf_{\beta' \in G_\beta} I_N(\beta')/\zeta_N > \delta) \to 1$
2. $J_N(\beta')/\zeta_N^2 \to_P 0$ *for all* $\beta' \in G_\beta$
3. *The distribution of* $\zeta_N^{-1} M_N(\beta')$ *is tight in* $C(G_\beta)$*, where* $M_N = R_N - I_N$ *and* $C(G_\beta)$ *is the space of continuous functions on* G_β*.*
4. *There exists a compact subset* K *of* B *such that* $\beta_0 \in \text{int}K$*, and* $P(\inf_{\beta' \notin K} R_N(\beta') \leq 0) \to 0$*.*

Then $\hat{\beta}_{PL} \to \beta_0$*.*

The conditions for consistency in the partial likelihood context are quite intuitive. We need to establish that the log-partial likelihood ratio R_N evaluated with respect to the true paramter β_0 diverges in probability, and that the point of this divergence is unique in the closure of B. The first three conditions establish divergence; the last condition establishes uniqueness.

The simple conditioning structure in the truncated likelihood, where each observation Y_{ij} is conditioned on its own variable A_{ij} simplifies checking the conditions for log-likelihood ratio divergence. While the general proof must eliminate cases where the filtration associated with the sequence of observations provides less and less unique information per observation, in the case of the truncated likelihood, the sigma algebra that each observation conditions on is disjoint. Thus, we may simplify the conditions for the truncated MLE's consistency as follows:

Theorem 4 (Consistency for Truncated Estimator)**.** *Assume that*

1. $\sum \boldsymbol{R}_0 \xrightarrow{p} \infty$*.*
2. *With probability 1, a proportion* $\delta > 0$ *of* ij *where* $A_{ij} = 1$ *has* $i_{\beta_0}^{(ij)}(\beta) > 0$ *given* $A_{ij} = 1$*.*
3. *There exists a compact subset* K *of* B *such that* $\beta_0 \in \text{int}K$*, and* $P(\inf_{\beta' \notin K} R_N(\beta') \leq 0) \to 0$*.*

then, the truncated estimator for β is consistent.

Proof. We only need to show that the first two conditions here imply the first three conditions in Theorem 3. We can show this by LLN because of the independence of $r_{\beta_0}^{(ij)}(\beta)$, so long as $i_{\beta_0}^{(ij)}(\beta) > 0$ for some infinite subsequece of A_{ij}'s. The first two conditions guarantee this because having an infinite sequence of R_{ij}'s guarantees an infinite sequence of A_{ij}'s, and condition one ensures that the expected log-partial likelihood ratio will be positive at each of these. LLN implies tightness in the residuals, so this gives the first three conditions of Theorem 3. □

There are several simplifications that can be made here for common network models. For example, when the zero-truncated distribution is in the exponential familiy, the convexity condition automatically satisfied.

6.3 Efficiency

In principle, the robustness of the truncated estimator should come at the cost of estimation efficiency when compared to a full likelihood approach. However, quantifying this loss in general is difficult because the form of Q, which is required for the specification of the full likelihood, can be too complex to practically specify for realistic models. In this section, we consider the efficiency lost in the worst case to establish intuition about which information the truncated estimator leaves behind. In particular, we study the case where the Q is a known deterministic process that fixes exactly which $R_{ij} = 1$ and which $R_{ij} = 0$, i.e. in the setting where an oracle has told the investigator exactly which dyads are at risk of interacting. In this setting, there is no missing information, so this serves as an upper bound for the amount of information available for estimating β from the observed interactions Y_n. In this situation, we compare the information accrued by the full-data procedure to the information accrued by the truncated procedure.

For convenience, we define the following quantities:

$$p_{ij}^{(\beta)} = P_\beta(A_{ij} = 1 \mid R_{ij} = 1, X_{ij}) \tag{24}$$

$$l^{(\beta)}(Y_{ij}) = \log P_\beta(Y_{ij} \mid A_{ij} = 1, X_{ij}) \tag{25}$$

$$l_{tr}^{(\beta)}(Y_{ij}) = \log P_\beta(Y_{ij} \mid A_{ij} = 1, X_{ij}) \tag{26}$$

These are, respectively the probability that a given dyad has an observed nonzero interaction

value given that it is at risk, and the truncated log-likelihood for a single dyad.

Under the assumption that the relationship graph R is fully available, all dyads ij for which $R_{ij} = 0$ (i.e., that have no relationship) are deterministically zero, and therefore contribute nothing to either the full or truncated likelihood. We can then rewrite the full log-likelihood $l^{\beta)}$ for the whole dataset Y_n in terms of the truncated log-likelihood l^{β}_{tr}:

$$l^{(\beta)}_{tr}(Y_n) = \sum_{ij \in \mathcal{R}} A_{ij}(l^{(\beta)}(Y_{ij}) - \log(p^{(\beta)}_{ij})) \tag{27}$$

$$l^{(\beta)}(Y_n) = \sum_{ij \in \mathcal{R}} A_{ij} \log p^{(\beta)}_{ij} + (1 - A_{ij}) \log(1 - p^{(\beta)}_{ij}) + A_{ij}(l^{(\beta)}_{tr}(Y_{ij})). \tag{28}$$

First, we give an intuitive description of where the truncated procedure loses information compared to the oracle procedure. We can express the Fisher Information as the negative expectation of the Hessian of the log-likelihood. Partitioning the sample space according to the non-zero observation indicator A_{ij}, we obtain

$$\mathcal{I}^{(\beta)}_{tr}(Y_n) = -\left(\sum_{ij \in \mathcal{R}} p^{(0)}_{ij} \mathbb{E}_0(\nabla^2_\beta \, l^{(\beta)}(Y_{ij}) \mid A_{ij} = 1) - p^{(0)}_{ij} \, \nabla^2_\beta \log p^{(\beta)}_{ij} \right) \tag{29}$$

$$\mathcal{I}^{(\beta)}(Y_n) = -\left(\sum_{ij \in \mathcal{R}} p^{(0)}_{ij} \mathbb{E}_0(\nabla^2_\beta \, l^{(\beta)}(Y_{ij}) \mid A_{ij} = 1) + (1 - p^{(0)}_{ij}) \mathbb{E}_0(\beta^2_\beta l^{(\beta)}(Y_{ij}) \mid A_{ij} = 0) \right). \tag{30}$$

The expectation decomposition reveals that the truncated procedure loses information via two lost comparisons. The most obvious lost comparison results from the fact that the truncated procedure excludes all dyads ij that are observed to be zero so that $A_{ij} = 0$. Cleary, these zero dyads can contribute no information to the truncated procedure, so the information contributed by these dyads appears in Equation 30 but not in Equation 29. Because it corresponds directly to information lost by paring down the sample, we call this loss of information the *sample size effect*.

The second lost comparison is more subtle, and corresponds to the lost counterfactual comparison of observed non-zero dyads to themselvs if they were instead observed to be zero. In particular, because all dyads included in the truncated procedure have $A_{ij} = 1$, there is no information gained by observing that $Y_{ij} > 0$, so this must be subtracted off. This corresponds to the negative term in Equation 29. We call this loss of information the *identification effect* because it results from confounding the data inclusion mechanism in the

truncated procedure with the interaction generation process.

To bound the information loss of the truncated procedure, representing the Fisher Information as the variance of the score function is more convenient. Taking the gradient of Equation 28 with respect to β, we obtain

$$S_{tr}^{(\beta)}(Y_n) = \sum_{ij \in \mathcal{R}} A_{ij} S_{tr}^{(\beta)}(Y_{ij}) \tag{31}$$

$$S^{(\beta)}(Y_n) = \sum_{ij \in \mathcal{R}} \frac{A_{ij}}{p_{ij}^{(\beta)}} + \frac{1 - A_{ij}}{1 - p_{ij}^{(\beta)}} + A_{ij} S_{tr}^{(\beta)}(Y_{ij}). \tag{32}$$

Using this notation, we can compare the resulting expressions for the Fisher Information of each procedure:

$$\mathcal{I}_{tr}^{(\beta)}(Y_n) = \sum_{ij \in \mathcal{R}} p_{ij}^{(0)} \, \mathbb{V}\mathrm{ar}(S_{tr}^{(\beta)}(Y_{ij}) \mid A_{ij} = 1) \tag{33}$$

$$\mathcal{I}^{(\beta)}(Y_n) = \mathbb{V}\mathrm{ar}_0(S^{(\beta)}(Y_n)) \tag{34}$$

$$= \mathbb{E}_0(\mathbb{V}\mathrm{ar}_0(S^{(\beta)}(Y_n) \mid A_n)) + \mathbb{V}\mathrm{ar}_0(\mathbb{E}_0(S^{(\beta)}(Y_n) \mid A_n)) \tag{35}$$

$$= \sum_{ij \in \mathcal{R}} p_{ij}^{(0)} \, \mathbb{V}\mathrm{ar}_0(S^{(\beta)}(Y_{ij}) \mid A_{ij} = 1) + \sum_{ij \in \mathcal{R}} \mathbb{V}\mathrm{ar}_0(\mathbb{E}_0(S^{(\beta)}(Y_{ij}) \mid A_{ij})). \tag{36}$$

Note that the first term in Equation 36 is exactly equal to Equation 33, so the second term corresponds to the information lost by using the truncated procedure instead of the oracle procedure. When the variance is decomposed, as above, by partitioning the outcome space of each Y_{ij} into the non-zero part, where $A_{ij} = 1$ and the zero part where $A_{ij} = 0$, this term corresponds to the "between variance" of the score function between these two partitions. As intuition would suggest, the information lost by the truncated procedure results from being unable to compare zero and non-zero outcomes in aggregate.

By inspecting Equation 36, we can establish conditions under which the relative efficiency of the truncation procedure with respect to the oracle procedure (defined in terms of the ratio of any matrix norm of the information matrices) converges to a constant. In particular,

- $\mathbb{V}\mathrm{ar}_0(S^{(\beta)}(Y_{ij}))$ is bounded for all ij.
- $\binom{\sum_{ij} R_{ij}}{2}^{-1} \sum_{ij \in \mathcal{R}} p_{ij}^{(0)} \, \mathbb{V}\mathrm{ar}(S^{(\beta)}(Y_{ij}) \mid A_{ij} = 1) \to c_2$ where $0 < c_2 < \infty$.

These conditions establish that no single dyad, even in the limit, provides infinite information to the oracle or truncated procedures, and that the sample average of the information

contributed by each dyad to the truncated procedure converges to a positive constant.

6.4 Remark on computation

Computation of the truncated estimatior is highly efficient as it uses only dyads with nonzero outcomes as opposed to the full $\binom{n}{2}$ dimensional generalized adjacency matrix required by full likelihood methods. In cases where R_{ij} must be integrated out, the computation is on the order of $\mathcal{O}(n^2)$ for full inference, whereas the computation is on the order of $\mathcal{O}(\sum_{ij} A_{ij})$ for the truncated inference. Thus, when the rate of information accumulation is the same, the ratio information accumulated per unit of computation if unbounded in favor of the truncated procedure.

7 Analytical Examples

7.1 Analytical non-exchangeable sparse CID example

We consider a particularly simple case of a CIR model whose properties we can elicit analytically. Suppose that we have an interaction network population and an ordering for the vertex population $\mathbb{V}$ so that R_{ij} are independent indicators for each ij given the index of the smaller vertex i and the covariate vector X_{ij}. In addition, suppose that we know $P_\gamma(R_{ij} \mid X_{ij}) = f_\gamma(i, X_{ij})$ where f is a decreasing function of i if X_{ij} is held constant. For example, we may have

$$f(i, X_{ij}) = \text{logit}^{-1}(\log(i)/i + \gamma' X_{ij}).$$

This setting allows us to consider cases where we have a CID model that is sparse. This corresponds to having a covariate that sets the limiting marginal probability of seeing a nonzero edge $P(A_{ij} \mid X_{ij})$ to zero, representing a different regime from the well-behaved covariate regime considered earlier in which CID models were shown to produce a dense limits. In this setting, we may concretely study how sparsity misspecification of f affects inference.

For convenience, we define the following quantities:

$$q_{ij}^{(\gamma)} = P_\gamma(R_{ij} = 1 \mid X_{ij}) \qquad q_{ij}^{(0)} = P_0(R_{ij} = 1 \mid X_{ij}) \tag{37}$$

$$p_{ij}^{(\beta)} = P_\beta(A_{ij} = 1 \mid R_{ij} = 1, X_{ij}) \qquad p_{ij}^{(0)} = P_0(A_{ij} = 1 \mid R_{ij} = 1, X_{ij}) \tag{38}$$

$$l_{tr}^{(\beta)}(Y_{ij}) = \log P_\beta(Y_{ij} \mid A_{ij} = 1, X_{ij}) \tag{39}$$

Quantities in the left-hand column are probabilities associated with the inferential model parameterized by β and γ; quantities in the right-hand column are probabilities associated with the true generative process, denoted with superscript (0). The rows define, respectively, the probability that an at-risk dyad generates a nonzero interaction, the probability that a given dyad is at risk, and the truncated log-likelihood of an interaction given that it is nonzero.

Because of the independence structure of the problem, marginalizing out R^{mis} is simple. The observed data log-likelihood has the form

$$l_{\beta,\gamma}(Y \mid X) = \sum_{ij} (1 - A_{ij}) \log(1 - q_{ij}^{(\gamma)} p_{ij}^{(\beta)}) + A_{ij} \left(\log(p_{ij}^{(\beta)} q_{ij}^{(\gamma)}) + l_{tr}^{(\beta)}(Y_{ij}) \right). \tag{40}$$

We can now derive moving target and truncated efficiency results for this subclass of model.

Using this log-likelihood we can obtain the effective estimand that corresponds to β. Taking an expectation with respect to the true generating process, we obtain

$$\begin{aligned} \mathbb{E}_0 l_{\beta,\gamma}(Y \mid X) = \sum_{ij} \; & (1 - q_{ij}^{(0)} p_{ij}^{(0)}) \log(1 - p_{ij}^{(\beta)} q_{ij}^{(\gamma)}) + q_{ij}^{(0)} p_{ij}^{(0)} \log(p_{ij}^{(\beta)} q_{ij}^{(\gamma)}) + \\ & q_{ij}^{(0)} p_{ij}^{(0)} \mathbb{E}_0 \left[l_{tr}^{(\beta)}(Y_{ij}) \mid A_{ij} = 1 \right]. \end{aligned} \tag{41}$$

Note that the first line corresponds to the Bernoulli log-likelihood for A (a consequence of the independence assumption on R), while the second corresponds to the conditional log-likelihood of Y give A.

Taking the gradient with respect to β,

$$\nabla_\beta \mathbb{E}_0 l_{\beta,\gamma}(Y \mid X) = \sum_{ij} (p_{ij}^{(\beta)} q_{ij}^{(\gamma)} - p_{ij}^{(0)} q_{ij}^{(0)}) \left(\frac{\nabla_\beta p_{ij}^{(\beta)}}{p_{ij}^{(\beta)}(1 - p_{ij}^{(\beta)} q_{ij}^{(\gamma)})} \right) + p_{ij}^{(0)} q_{ij}^{(0)} \nabla_\beta \mathbb{E}_0 \left[l_{tr}^{(\beta)}(Y_{ij}) \mid A_{ij} = 1 \right]. \tag{42}$$

$$\nabla_\gamma \mathbb{E}_0 l_{\beta,\gamma}(Y \mid X) = \sum_{ij} (p_{ij}^{(\beta)} q_{ij}^{(\gamma)} - p_{ij}^{(0)} q_{ij}^{(0)}) \left(\frac{\nabla_\gamma q_{ij}^{(\gamma)}}{q_{ij}^{(\gamma)}(1 - p_{ij}^{(\beta)} q_{ij}^{(\gamma)})} \right) \tag{43}$$

We care about the case where the truth is sparse and has sparsity rate $\epsilon_0(n)$:

$$\binom{n}{2}^{-1} \sum_{ij} q_{ij}^{(0)} \in O(\epsilon_0(n)) \to 0 \tag{44}$$

In this context, sparsity misspecification manifests by definition as either

$$\frac{\sum_{ij} q_{ij}^{(0)}}{\sum_{ij} q_{ij}^{(\gamma)}} \to 0 \qquad \text{(Model for } R \text{ too dense), or} \tag{45}$$

$$\frac{\sum_{ij} q_{ij}^{(\gamma)}}{\sum_{ij} q_{ij}^{(0)}} \to 0 \qquad \text{(Model for } R \text{ too sparse).} \tag{46}$$

Under these circumstances, we are able to show the following corollary.

Corollary 1 (Moving target for CID)**.** *Let Y be a sparse interaction process with sparsity rate $\epsilon_0(n)$ and $\mathcal{P}_{\beta,\gamma}$ a CID model. Suppose that the following hold:*

1. *$\mathcal{P}_{\beta,\gamma}$ is sparsity misspecified.*
2. *A identifies β in the presence of γ.*
3. *The absolute values of the gradient vectors $\nabla_\gamma q_{ij}^{(\gamma)}$ are bounded for all ij, for all $\gamma \in \Gamma$.*
4. *The span of the gradient vectors $\nabla_\gamma q_{ij}^{(\gamma)}$ include an all-positive vector.*

Then for every n there exists an $n' > n$ such that $\bar{\beta}_n \neq \bar{\beta}_{n'}$.

We present the proof in the appendix.

7.2 Analytical Logistic-Poisson example

Here we show the consequences of moving target behavior in the context of a particular model. We also demonstrate the efficiency calculation comparing the truncated estimator to the an oracle estimator with full knowledge of the risk set.

Consider the following independent CIR model:

$$R_{ij} \mid X_{ij} \sim \text{Bern}(\text{logit}^{-1}((f(i), X_{ij})'\gamma)) \tag{47}$$

$$Y_{ij} \mid X_{ij}, R_{ij} \sim \text{Pois}(\lambda_{ij}^{(\beta)}) \tag{48}$$

$$\lambda_{ij}^{(\beta)} \equiv \exp(X_{ij}'\beta) \tag{49}$$

Assume that this model is sparsity misspecified for some true generative process. Denote the true sparsity rate $\epsilon_0(n)$ and the misspecified sparsity rate $\epsilon_\gamma(n)$.

Suppose that for this example, the covariates X meet the following criteria.

1. X_{ij} are bounded for each ij.
2. X_{ij} includes one component that corresponds to an intercept (i.e. is 1 for all ij).
3. X_{ij} cannot be reparameterized as dummy variables identifying $k < |\gamma|$ groups.

Given the model specification, the probability gradients are given by

$$\nabla_\gamma q_{ij}^{(\gamma)} = X_{ij} q_{ij}^{(\gamma)}(1 - q_{ij}^{(\gamma)}) \qquad \nabla_\beta p_{ij}^{(\beta)} = X_{ij}(1 - p_{ij}^{(\beta)})\lambda_{ij}^{(\beta)}. \tag{50}$$

7.2.1 Moving target

Based on these functional forms, it is simple to show that the conditions given for the covariates are sufficient to satifsfy conditions 2–4 in Corollary 1. By the corollary, this implies that the effective estimand changes with size.

Taking the effective estimand score equations that correspond to γ and β, respectively, we

have for each n,

$$\sum_{ij}(p_{ij}^{(\beta)}q_{ij}^{(\gamma)} - p_{ij}^{(0)}q_{ij}^{(0)})\left(\frac{(1-q_{ij}^{(\gamma)})}{1-p_{ij}^{(\beta)}q_{ij}^{(\gamma)}}\right)(f(i), X_{ij}) = 0 \tag{51}$$

$$\sum_{ij}(p_{ij}^{(\beta)}q_{ij}^{(\gamma)} - p_{ij}^{(0)}q_{ij}^{(0)})\left(\frac{(1-p_{ij}^{(\beta)})\lambda_{ij}^{(\beta)}}{p_{ij}^{(\beta)}(1-p_{ij}^{(\beta)}q_{ij}^{(\gamma)})}\right)X_{ij} = -\sum_{ij}p_{ij}^{(0)}q_{ij}^{(0)}\left(\frac{\lambda_{ij}^{(0)}}{p_{ij}^{(0)}} - \frac{\lambda_{ij}^{(\beta)}}{p_{ij}^{(\beta)}}\right)X_{ij} \tag{52}$$

To show that $\bar{\beta}_n$ is not fixed, we can proceed by contradiction and see that if $\bar{\beta}_n$ were fixed, the score system would be overdetermined. Note in particular that we can rewrite the system:

$$\sum_{ij}\frac{(p_{ij}^{(\beta)}q_{ij}^{(\gamma)} - p_{ij}^{(0)}q_{ij}^{(0)})}{(1-p_{ij}^{(\beta)}q_{ij}^{(\gamma)})}(1-q_{ij}^{(\gamma)})(f(i), X_{ij}) = 0 \tag{53}$$

$$\sum_{ij}\frac{(p_{ij}^{(\beta)}q_{ij}^{(\gamma)} - p_{ij}^{(0)}q_{ij}^{(0)})}{(1-p_{ij}^{(\beta)}q_{ij}^{(\gamma)})}\left(G_{ij}(\beta)\right)X_{ij} = H(\beta) \tag{54}$$

Where G_{ij} and H are functions of β alone, and thus constants with respect to γ. This formulation makes it clear that none of the score equations are redundant, and thus require more that $k = |\gamma|$ degrees of freedom to satisfy simultaneously.

In this particular case, we can go further and characterize exactly how $\bar{\beta}_n$ changes. Examining the score equation for the intercept of β, we obtain

$$\sum_{ij}\lambda_{ij}^{(\beta)}\left(q_{ij}^{(\gamma)}\left(\frac{(1-p_{ij}^{(\beta)})}{(1-p_{ij}^{(\beta)}q_{ij}^{(\gamma)})}\right) + p_{ij}^{(0)}q_{ij}^{(0)}\left(\frac{(1-q_{ij}^{(\gamma)})}{(1-p_{ij}^{(\beta)}q_{ij}^{(\gamma)})}\right)\right) = \sum_{ij}\lambda_{ij}^{(0)}q_{ij}^{(0)} \tag{55}$$

Assuming that $\bar{\gamma}_n$ converges to some value in Γ, denoted γ_∞, the model's sparsity misspecification implies that the ratio $\frac{\sum_{ij}q_{ij}^{(0)}}{\sum_{ij}q_{ij}^{(\bar{\gamma}_\infty)}}$ must diverge or go to zero. We can then divide the limiting behavior of $\bar{\beta}$ into two cases:

Model is too dense. In this case, $\frac{\sum_{ij}q_{ij}^{(0)}}{\sum_{ij}q_{ij}^{(\bar{\gamma}_\infty)}} \to 0$. To obtain the limit for $\bar{\beta}_n$, we can divide the score equation by $\sum_{ij}q_{ij}^{(\bar{\gamma}_n)}$. In the limit, the second term on the LHS and the term on the RHS converge to zero, and we obtain the limiting score equation:

$$\lim_{n\to\infty}\left(\sum_{ij}q_{ij}^{(\bar{\gamma}_n)}\right)^{-1}\sum_{ij}\lambda_{ij}^{(\bar{\beta}_n)}q_{ij}^{(\bar{\gamma}_n)}\left(\frac{(1-p_{ij}^{(\bar{\beta}_n)})}{(1-p_{ij}^{(\bar{\beta}_n)}q_{ij}^{(\bar{\gamma}_n)})}\right) = 0 \tag{56}$$

Normalizing by the sum of the factors that multiply $\lambda_{ij}^{(\bar{\beta}_n)}$, we see that $\lim_{n\to\infty} \sum_{ij} \lambda_{ij}^{(\bar{\beta}_n)} / \binom{n}{2} = 0$, or that the average mean expected number of pairwise interactions converges to zero. This implies by the functional form of λ_{ij} and the above assumptions about the covariates that at least one component of β must diverge.

Model is too sparse. In this case, $\frac{\sum_{ij} q_{ij}^{(\bar{\gamma}\infty)}}{\sum_{ij} q_{ij}^{(0)}} \to 0$. We apply a similar operation, this time dividing the score equation by $\sum_{ij} q_{ij}^{(0)}$ to obtain

$$\lim_{n\to\infty} \left(\sum_{ij} q_{ij}^{(0)} \right)^{-1} \left[\sum_{ij} \lambda_{ij}^{(\bar{\beta}_n)} q_{ij}^{(0)} \left(p_{ij}^{(0)} \frac{(1 - q_{ij}^{(\bar{\gamma}_n)})}{(1 - p_{ij}^{(\bar{\beta}_n)} q_{ij}^{(\bar{\gamma}_n)})} \right) - \lambda_{ij}^{(0)} q_{ij}^{(0)} \right] = 0 \tag{57}$$

In this case, the factors multiplying $\lambda_{ij}^{(\bar{\beta}_n)}$ and $\lambda_{ij}^{(0)}$ cannot be reconciled. In particular, the factor in parentheses is strictly less than one, so the limiting solution for $\bar{\beta}_n$ gives a set of expected interaction numbers $\lambda_{ij}^{(\bar{\beta}_n)}$ that are asymptotically biased upward – thus, if the model for the observable interactions $Y \mid R$ is correctly specified, the estimator $\hat{\beta}_n$ is inconsistent. Nonetheless, it appears that $\bar{\beta}_n$ has the potential to converge to a point on the interior of the parameter space, so for large samples the moving target behavior may be negligible for a particular application.

7.2.2 Efficiency

We can also derive the efficiency loss of the truncated estimator with respect to the oracle estimator in this example. The efficiency of the oracl estimator is easily obtained by replacing each $q_{ij}^{(0)}$ and $q_{ij}^{(\gamma)}$ with the known values of R_{ij} in Equation 52 and differentiating with respect to β. This yields:

$$\mathcal{I}_{or}(\beta) = \sum_{ij \in \mathcal{R}} \lambda_{ij}^{(\beta)} X_{ij} X_{ij}^\top \tag{58}$$

$$= \sum_{ij \in \mathcal{R}} \left(\lambda_{ij}^{(\beta)} - p_{ij}^{(0)} \frac{\lambda_{ij}^{(\beta)}}{p_{ij}^{(\beta)}} \left(1 - (1 - p_{ij}^{(\beta)}) \frac{\lambda_{ij}^{(\beta)}}{p_{ij}^{(\beta)}} \right) \right) X_{ij} X_{ij}^\top + \mathcal{I}_{tr}(\beta) \tag{59}$$

$$\mathcal{I}_{tr}(\beta) = \sum_{ij \in \mathcal{R}} \left(p_{ij}^{(0)} \frac{\lambda_{ij}^{(\beta)}}{p_{ij}^{(\beta)}} \left(1 - (1 - p_{ij}^{(\beta)}) \frac{\lambda_{ij}^{(\beta)}}{p_{ij}^{(\beta)}} \right) \right) X_{ij} X_{ij}^\top \tag{60}$$

Because both the oracle estimator and the truncated estimator are linearly parameterized natural exponential families, the information matrices have the form of a weighted inner product of the design matrix X [25]. Thus, the sum of the weights alone can be used to

quantify the amount of information accumulated by each procedure since both procedures employ the same design matrices X.

The weighting factor in Equation 60 has three main factors that have appealing interpretations. The first factor, $p_{ij}^{(0)}$, represents the sample size effect on information, and accounts for the fact that the truncated estimator only extracts information from those dyads that are actually observed to be nonzero. This factor appears regardless of the model form. The second factor corresponds to the "baseline" information contained in a nonzero observation. Heuristically, because the truncated procedure is derived from a Poisson procedure, we can specify the baseline information to be equal to its expectation, as is the case with the Poisson family. Under the truncated observation mechanism, this expectation is $\frac{\lambda_{ij}^{(\beta)}}{p_{ij}^{(\beta)}}$. The third factor is one minus the information lost due to identification, as discussed in Section 6.3. The fraction of lost information corresponds to the information that the procedure loses by not being able to compare the observed outcome to the counterfactual siutation where that outcome was zero instead. This lost information is given by the expected baseline information gained in such a situation if there were no truncation, $\frac{\lambda_{ij}^{(\beta)}}{p(\beta)_{ij}}$, multiplied by the probability of such a situation, as implied by β, or $1 - p_{ij}^{(\beta)}$.

We can observe how the information loss depends on the parameter β by plotting components of the information weights as a function of $\lambda_{ij}^{(\beta)}$. In Figure 3, we first plot the third identification factor in the truncated information weight. As expected, we see that when $\lambda_{ij}^{(\beta)}$ is large, this factor approaches one, indicating that there is little information lost due to identification. This is because the counterfactual outcome where $Y_{ij} = 0$ is highly unlikely for $\lambda_{ij}^{(\beta)}$ large, and so the lost comparison to this counterfactual had little information to contribute.

In Figure 3, we also plot the asymptotic truncated information weight, which we obtain by substituting $p_{ij}^{(\beta)}$ for $p_{ij}^{(0)}$ in the weight expression in Equation 60. This substitution is reasonable asymptotically when $p_{ij}^{(\hat{\beta})} \to p_{ij}^{(0)}$, since the information will be evaluated in the neighborhood of the MLE. Using this substitution, the truncated information weight is only a function of $\lambda_{ij}^{(\beta)}$. Here we see that for $\lambda_{ij}^{(\beta)}$ large, both sample size loss and identification loss become small, and the information weight in the truncated procedure approaches the information weight in the oracle procedure, given by $\lambda_{ij}^{(\beta)}$.

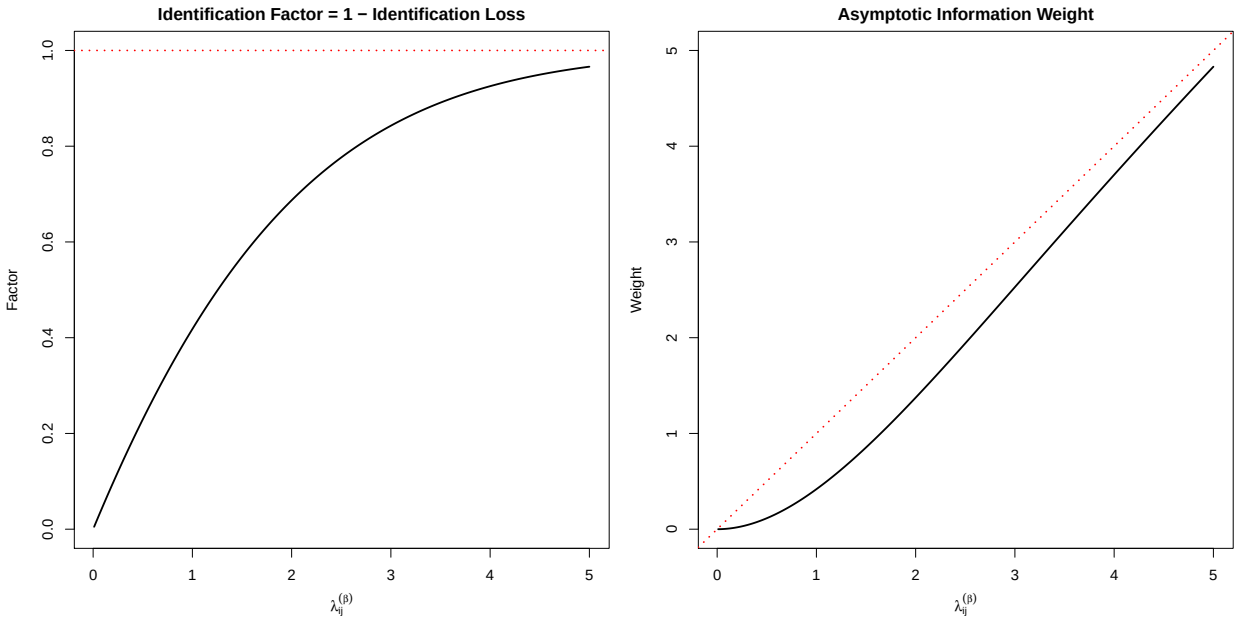

Figure 3: **Left**: Identification (black line) factor corresponding to how baseline information is scaled based on identification loss. For larger values of $\lambda_{ij}^{(\beta)}$ this factor approaches one (red dotted line), while for smaller values, the counterfactual observation of a zero for dyad ij carries information unavailable to the truncated procedure. **Right**: Asymptotic information weight obtained by substituting $p_{ij}^{(\beta)}$ for $p_{ij}^{(0)}$ in the weight expression in Equation 60. For large values of $\lambda_{ij}^{(\beta)}$, both sample size loss and identification loss become smaller, and the information weight for the truncated estimator (black line) approaches the information weight of the oracle estimator (red dotted line), implying that the truncated procedure is nearly fully informative for large $\lambda_{ij}^{(\beta)}$.

8 Simulated Examples

8.1 Simulated counting process examples

In this section, we conduct a simulation study designed to mimic the structure of the inventor collaboration network analysis presented in Section 1.1. We first demonstrate the moving target phenomenon from Theorem 2 under sparsity misspecification, characterizing the instability of the effective estimand, and the corresponding instability in the MLE. We then demonstrate the robustness of the truncated estimator to sparsity misspecification. Finally, we explore the properties of the truncated estimator more generally, using a full factorial design to explore how the efficiency and coverage properties of the truncated estimator and its corresponding asymptotic confidence interval depend on the underlying generative parameters. The results of the factorial experiment speak to the issues discussed in Section 6.2 and Section 6.3.

8.1.1 Model specification

For this problem, a natural choice of model is counting process regression [27]. This is a CID model, and is thus sensitive to sparsity misspecification under regularity conditions. According to this model, we represent the outcome $Y_{ij}(\cdot) \mid R_{ij} = 1$ for any pair of actors ij as a counting process with instantaneous hazard given by a GLM specification:

$$\log \lambda_{ij}(t) = \beta' X_{ij}(t). \tag{61}$$

In this case, $X_{ij}(t)$ represent covariates associated with each pair which may depend on time, and which may include aspects of the history of the counting process itself. Conditional on the relationship graph R, this model yields the log-likelihood for β:

$$L_{tr}(\beta) = \sum_{ij \in \mathcal{R}} \left(- \int_0^T \lambda_{ij}(s \mid \mathcal{F}_s) ds \right) + \sum_{k=1}^{Y_{ij}(T)} \log \lambda_{ij} \left(t_{ij}^{(k)} \mid \mathcal{F}_{t_{ij}^{(k)}} \right), \tag{62}$$

where $t_{ij}^{(k)}$ is the time of the kth observed interaction between actors i and j. Likewise, the truncated log-likelihood for β has the form

$$L_{tr}(\beta) = \sum_{ij:A_{ij}=1} \left(-\int_0^T \lambda_{ij}(s \mid \mathcal{F}_s) ds \right) + \sum_{k=1}^{Y_{ij}(T)} \log \lambda_{ij} \left(t_{ij}^{(k)} \mid \mathcal{F}_{t_{ij}^{(k)}} \right), \tag{63}$$
$$- \log \left(1 - \exp \left(- \int_0^T \lambda_{ij}(s \mid \mathcal{F}_s^0) ds \right) \right)$$

where $\mathcal{F}_s^0$ is the history that would have been induced if no interactions had taken place between actors i and j before time s.

To simulate from this model, we define a set of actors and assign each actor a zipcode and assignee (firm). Using these attributes, we define binary covariate vectors for each pair of actors that report whether the actors live in the same zipcode, or work for the same assignee. As we allow the process to unfold, we also keep track of whether at time t the actors have had previous collaborations. Using these covariates, we simulate from the following CIR model:

$$\begin{aligned} R_{ij} \mid X_{ij} &\sim \text{Bin}(\rho_{ij}) \\ \text{logit}\, \rho_{ij} &\equiv \gamma_0 \,\text{logit}(\alpha(i)) + \gamma_1 \cdot \text{Zip}_{\text{ij}} + \gamma_2 \cdot \text{Asg}_{\text{ij}} \\ Y_{ij}(t) \mid R_{ij}, X_{ij}, \mathcal{F}(t) &\sim \begin{cases} CP(\lambda_{ij}(t)) & \text{if } R_{ij} = 1 \\ \mathbf{0}(t) & \text{if } R_{ij} = 0 \end{cases} \\ \log \lambda_{ij}(t) &\equiv \beta_0 + \beta_1 \cdot \text{Zip}_{\text{ij}} + \beta_2 \cdot \text{Asg}_{\text{ij}} + \beta_3 \cdot \text{Ng0}_{\text{ij}}(\text{t}) \end{aligned} \tag{64}$$

In the above specification, Zip and Asg are indicators for whether actors I and J live in the same zipcode, or work for the same firm, repsectively, and Ng0 is an indicator for previous collaboration, i.e., $Y_{ij}(t) > 0$. γ is a vector of relationship process coefficients, while $\alpha(i)$ is a function of i that approaches 0 as $i \to \infty$, and controls the sparsity of the generating process.. Both γ and $\alpha(i)$ are considered nuisance parameters in this case. β is a vector of conditional interaction process coefficients, which are the parameters of interest. In these simulations, we test our ability to recover β by various estimators.

We generate a network of size $n = 2000$ in which we observe 2000 interactions. From this network, we draw subsamples by sampling groups of vertices that have the same assignee attribute – this is analogous to building a network sample drawing a firm randomly from the set of all firms and adding all employees to the network sample. Fixing this sample sequence, we regenerate the network 100 times to create 100 replications.

For each of the competing models, we assume that we have correctly specified the family of the conditional interaction process given the relationship graph. We compare the results of the misspecified CID approach that simply assumes the counting process specification for all dyads to the truncation approach, which makes no assumptions about the sparsity of the network.

8.1.2 Moving target sensitivity and robustness

To demonstrate the moving target behavior derived in Theorem 2, we focus on a single set of simulation parameters. Here, we set $\alpha i = \log(i)/i$, $\gamma = (0.02, 1, 2)$, and $\beta = (1e-5, 0, 0.2, 3)$. Thus, the expected *relationship* degree for person i goes as $\log(i)/i$, with relationships concentrated more heavily between individuals in the same zip code and working for the same assignee. Conditional on these relationships, we assume zip code has no effect on the frequency of interactions between individuals who have a relationship, while asignee has a small positive effect on this frequency and having at least one previous collaboration has a large positive effect on this frequency.

Model family is dense. In our first example, we consider a model family that assumes the risk process R is fully connected, corresponding to the popular GLM approach of network regression. For each subsample generated by the sequence above, we compute the effecive estimand of the misspecified model in addition to the MLE and MPLE from the dense and truncated models, respectively. We repeat this for each of the 100 replications. We plot these against the true values of β in Figure 4. The simulations highlight several results from the discussion above. The effective estimands of the misspecified models show the moving target behavior as they vary with n, and the estimators track closely with their effective estimands. The truncated estimator shows no sensitivity to the sparsity of the population process.

Model family is sparse, but rate is misspecified. The above example is an extreme case of sparsity misspecification because the proposed model family was dense. However, we can also demonstrate that sparsity misspecification damaging in cases where the model family is sparse, but the rate is misspecified. In the following plots, both the truth and the model family follow a CIR model defined above, but in this case the intercept of the logistic equation that defines $P(R_{ij} \mid X_{ij})$ goes as $\log(i)/i$ for the true model, but it goes as $1/i$ in the inferential model. The inferential model thus assumes a risk process whose rate is too sparse.

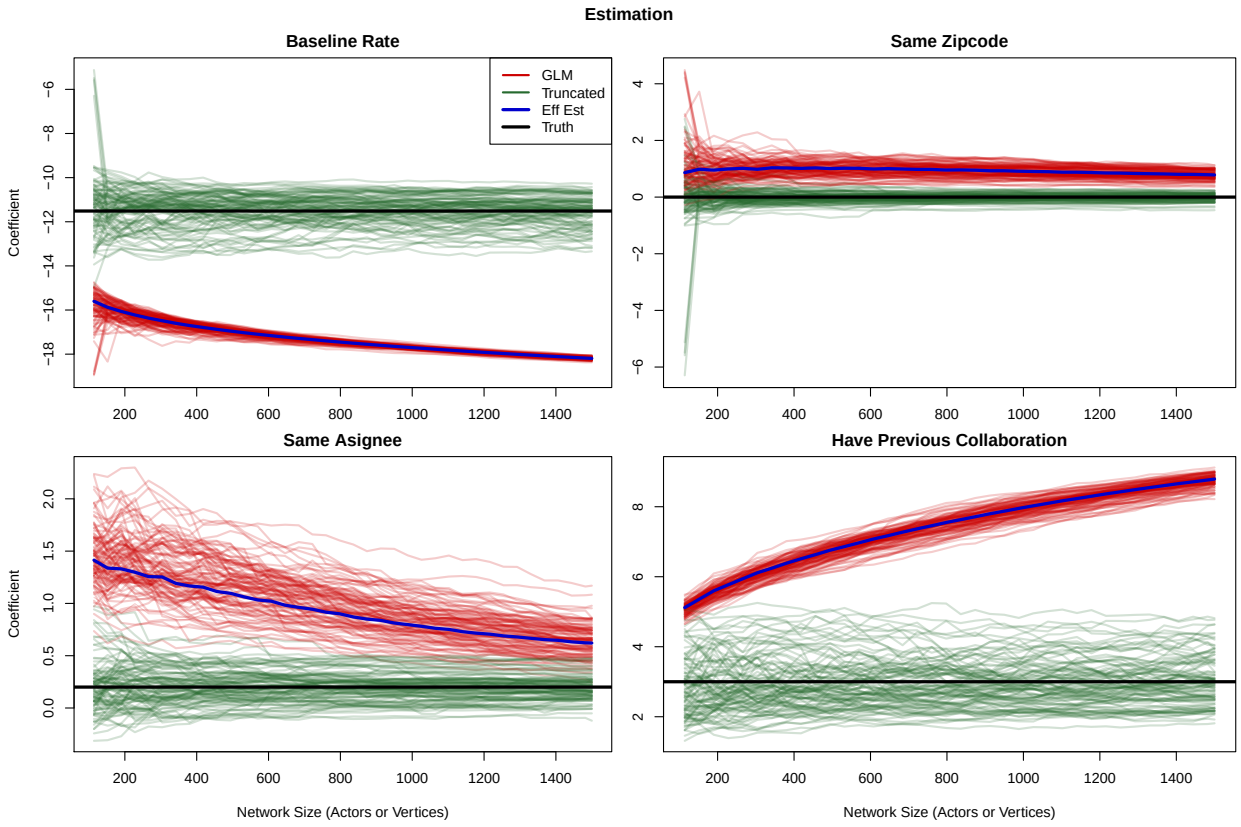

Figure 4: Plots of the sampling distribution of sequences $\hat{\beta}_n$ computed from the sparsity misspecified counting process model (red), and the truncated model (green) from samples of differing size. We also plot the effective estimand for the misspecified model (blue) and the true values of β (black).

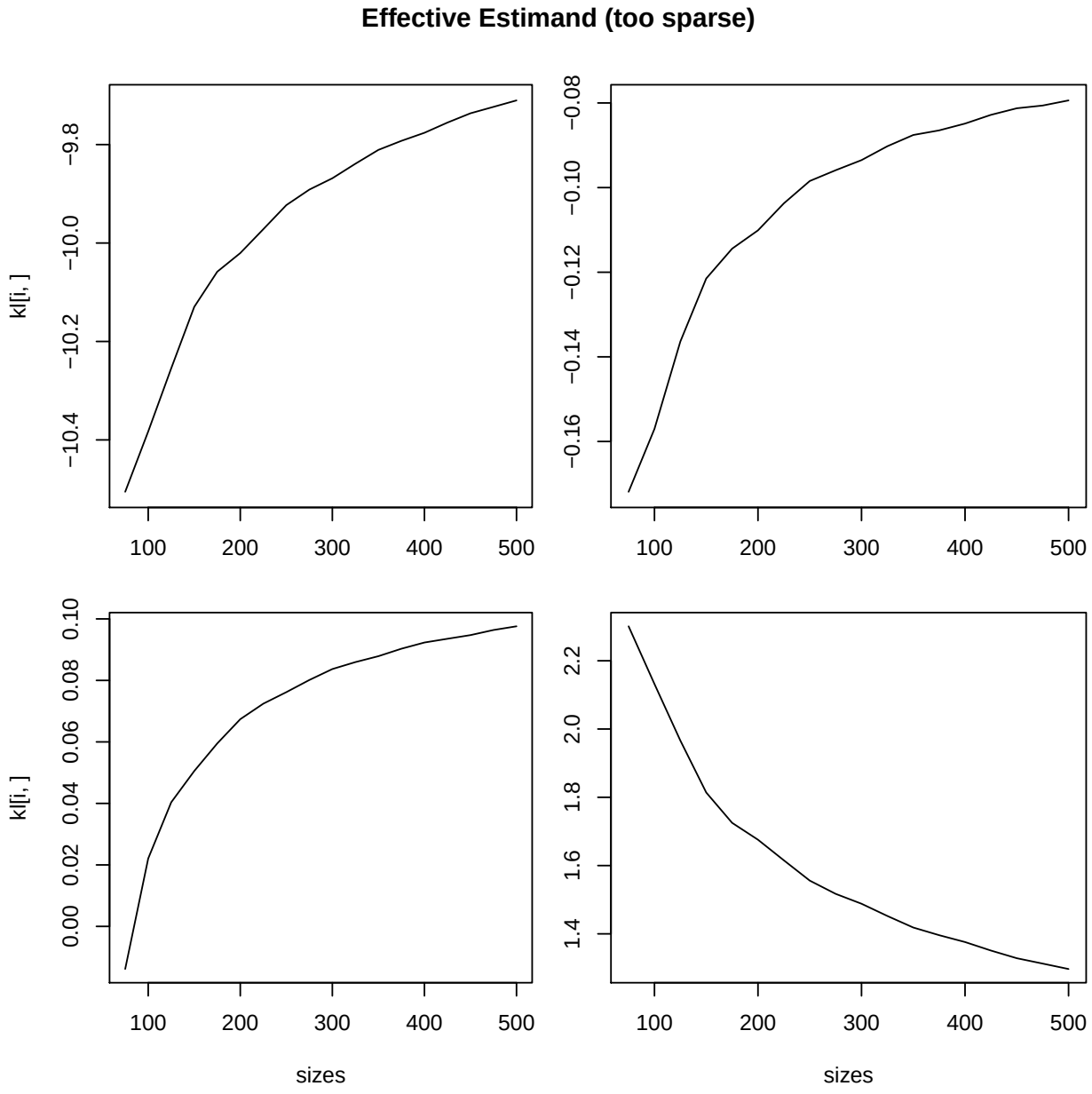

Figure 5: Plots of the effective estimand when the proposed model family is too sparse. Here the true logistic model for R_{ij} has intercept function $\log(i)/i$ and the inferential model assumes the intercept goes at $1/i$.

The same behavior may be seen when the inferential model is too dense. Consider switching the intercept functions above, so that the truth model goes as $1/i$ but the inferential model goes as $\log(i)/i$. This behavior is similar to the behavior when the investigator assumed a dense model. In large samples, these will behave qualitatively similarly.

8.1.3 Efficiency and coverage of truncated estimator

We also use this simulated example to demonstrate the efficiency and coverage properties of the truncated estimator and its corresponding asymptotic confidence interval in both the finite sample and large-sample limit. For this demonstration, we expand the above simulation to a full factorial design over the interaction parameter space B and the space of network sample sizes. Using the same simulation design as above, we fix each of the β coefficients corresponding Zip, Asg, and Ng0 at one of four levels while keeping the intercept coefficient fixed across all runs, yielding 64 design points. We generate 100 replicated datasets at each design point, and within each experimental run, we obtain estimates from 8 nested samples of increasing sample size. We assess the efficiency and coverage properties of the truncated estimator and its corresponding asymptotic confidence interval for each of the four components of β (Intercept, Zip, Asg, Ng0).

Efficiency. Following Section 6.3, we compute the variance inflation factor of the truncated estimator with respect to an oracle estimator given by the MLE when the risk set is fully known. For finite sample sizes, we compute this inflation factor from the outputs of the factorial experiment. The simulation yields draws from the sampling distributions of the truncated and oracle estimators for each component of β at each design point and sample size. To compute the variance inflation factor, we take the ratio of the sampling distribution variances of the two estimators at each design point and sample size. The full output of the simulation at one design point, $(0, 0.2, 3)$, is shown in Figure 7 as an example. As expected, the sampling distributions of estimmates from the oracle estimator are more concentrated than those of the truncated estimator at all sample sizes.

Because this example is analytically tractable, we also compute the large-sample limiting variance inflation factor for each parameter combination by computing the limit of the inverse Fisher information matrix. We assume that zipcode and assignee sizes remain fixed while the number of actors in the network grows to infinity, dyads that match on neither zipcode nor assignee (i.e. $\text{Zip}_{ij} = 0$ and $\text{Asg}_{ij} = 0$) dominate the limiting sample, yielding convenient simplifications. Details of this calculation, as well as a table of limiting variance inflation

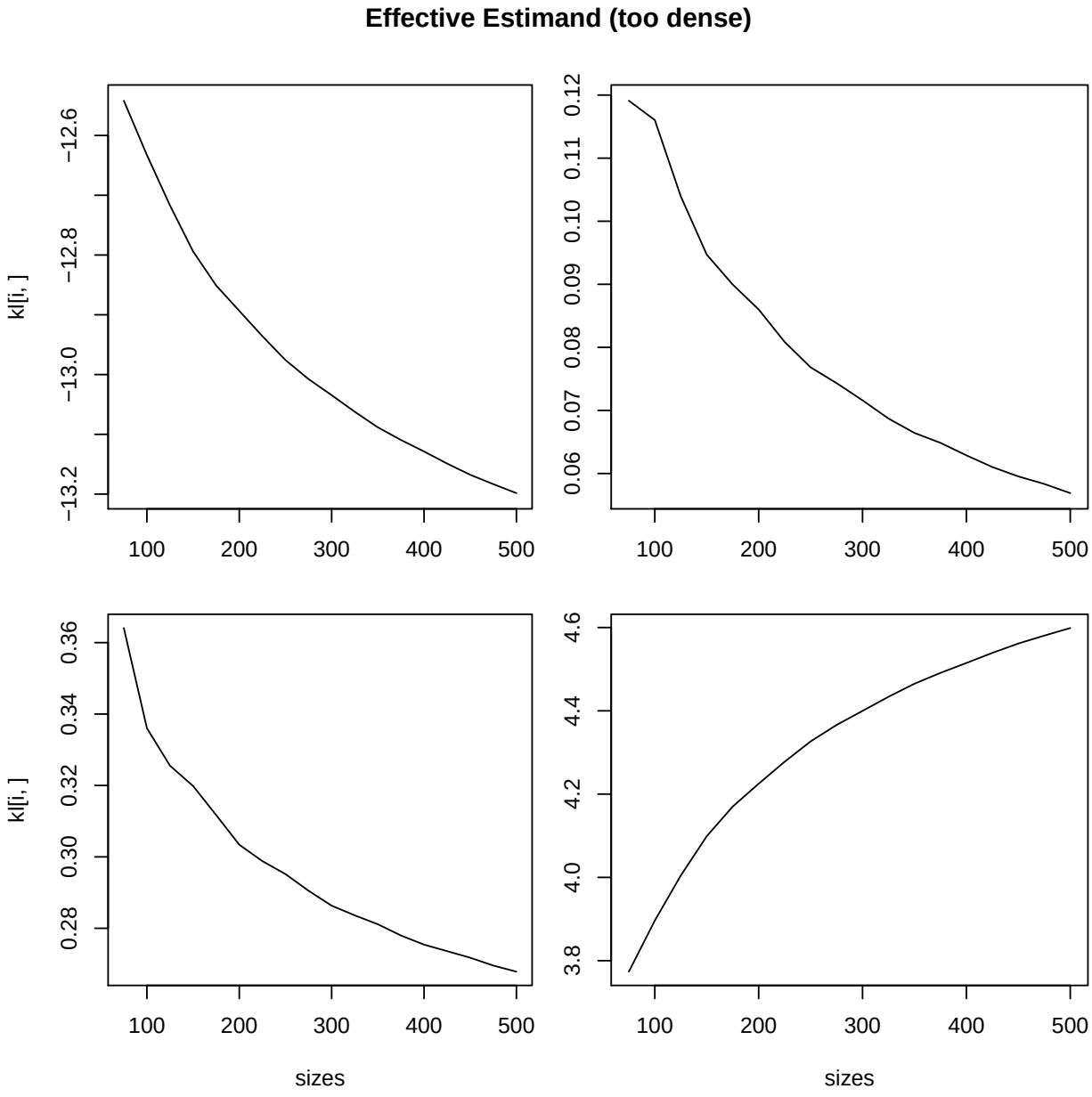

Figure 6: Plots of the effective estimand when the proposed model family is too dense. Here the true logistic model for R_{ij} has intercept function $1/i$ and the inferential model assumes the intercept goes at $\log(i)/i$.

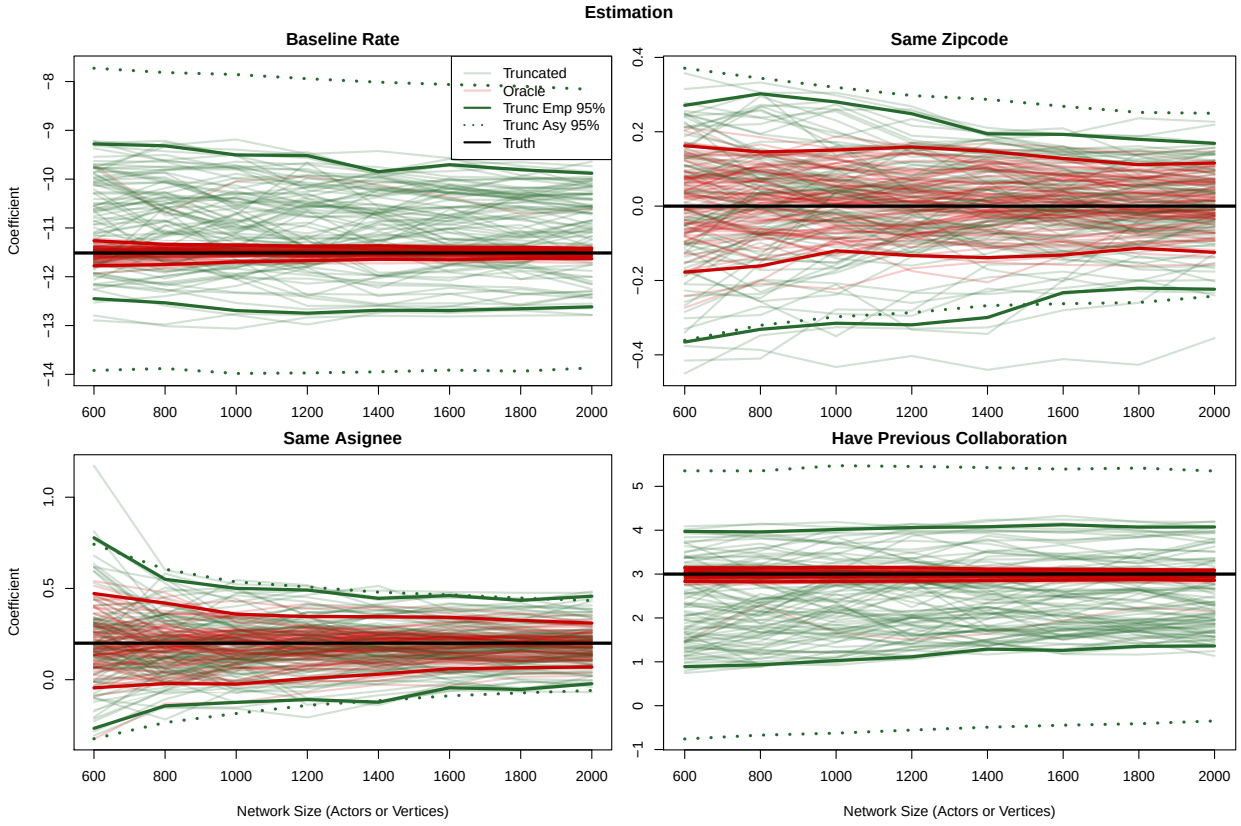

Figure 7: Plots of sampling distribution of sequences $\hat{\beta}_n$ computed from the truncated model (green) and the oracle model (red). The oracle model has full knowledge of the risk set R and is computed using the full likelihood on this subset of dyads.

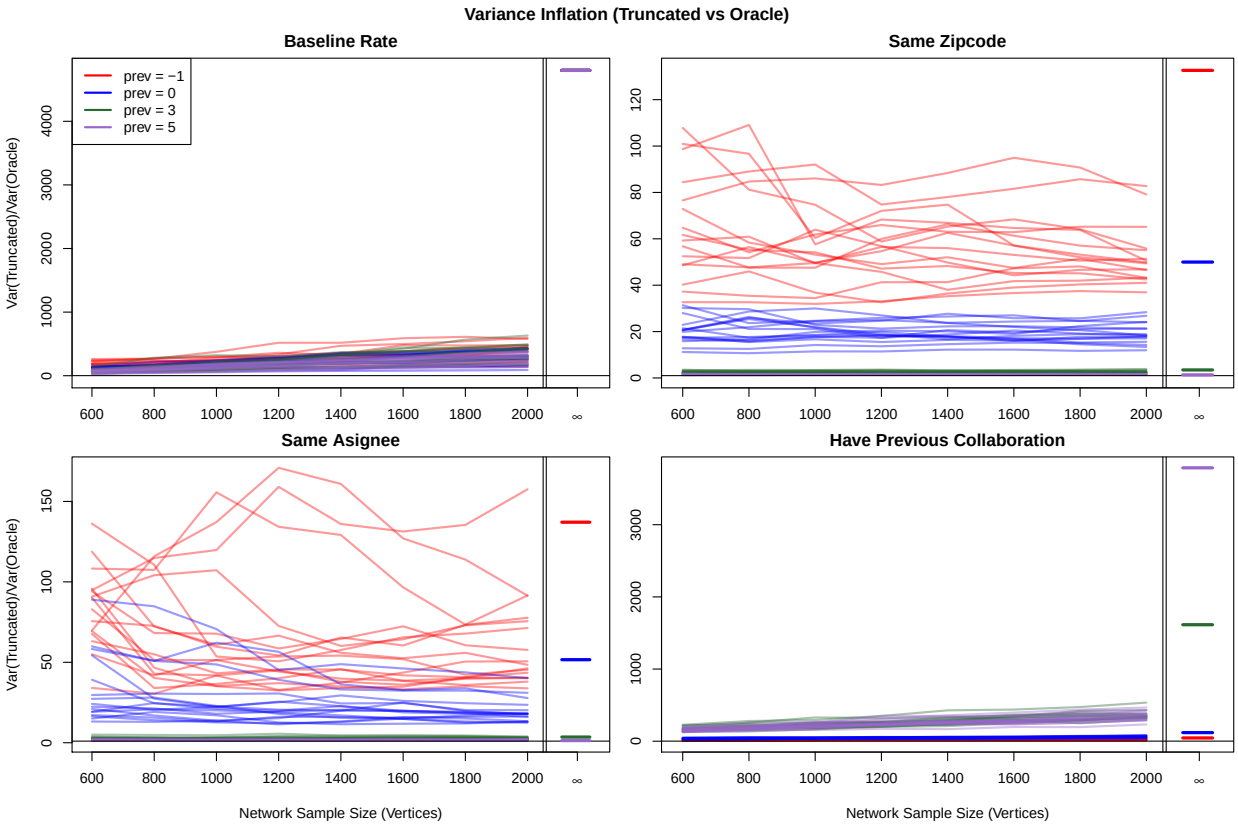

Figure 8: Variance inflation factors resulting from the comparison of the truncated estimator's sampling distribution to the oracle estimator. Lines are colored by the value of the "previous collaboration" coefficient, which shows the most influence on the efficiency of the non-intercept coefficients. For each value of "prev", the largest limiting variance inflation factor among all remaining parameter configurations is shown on the right. These are computed from the limit of the inverse of the Fisher information matrix. Note that the variance inflation of the intercept is the same for all parameter combinations.

factors at each design point are given in the appendix.

The results in Figure 8 confirm the theory in Section 6.3. First, while in many cases the variance inflation factor is relatively large, it is finite in the large sample limit in all cases. Secondly, the scale of the variance reduction factors confirm that information is lost through both a loss of sample size and a loss of identification. In this particular case, the intercept, Zip, and Asg coefficients all lose efficiency because the truncated procedure drops all at-risk dyads with zero observed interactions. However, there is a greater loss of efficiency for the intercept and "previous collaboration" coefficients because all of the dyads dropped by the truncated procedure provide the oracle procedure with information about the intercept coefficient that is unconfounded with the "previous collaboration" coefficient. With the truncated procedure, these two coefficients are much more weakly identified by the time

intervals before the first observed collaboration among the included dyads. This loss of identification is by far the larger effect, resulting in large variance reduction factors for the intercept and "previous collaboration" coefficients. Because the intercept is affected by both forms of information loss, it has the largest variance inflation factor.

The variance inflation factors computed with respect to the oracle estimator represent an upper bound on the variance inflation one would obtain from a realistic full likelihood estimator. In a realistic case, a full likelihood estimator would require summation over the missing relationship indicators R using a prior measure that is not sparsity misspecified. Assuming such a prior were available, the variance inflation of the truncated estimator with respect to the full-likelihood procedure would depend on the fraction of missing information implied by this prior measure, with variance reduction coming at the cost of potentially influential prior assumptions.

Coverage. Because the truncated estimator is itself the MLE of a derived sub-experiment, it has a corresponding asymptotic confidence interval, computed from the inverse of the observed Fisher information matrix $\mathcal{I}^{tr}(\hat{\beta}^{tr})$. This asymptotic interval is guaranteed to achieve nominal coverage in the large sample limit. Here we explore the finite sample properties of this interval using the factorial design described above. For each of the 100 replications at each design point and sample size we check whether the asymptotic 95% intervals for each of the four parameters cover the true value and use logistic regression to quantify the sensitivity of the coverage rate to the true parameter values.

Table 1 shows the example output coverage table for the design point $(0, 0.2, 3)$, which we have used as an example throughout this section. In the replications at this design point, the asymptotic confidence intervals show undercoverage for the baseline and "previous collaboration" coefficients, while the intervals for the Zip and Asg coefficients remain close to nominal coverage levels. We summarize the sensitivity of coverage rates to parameter values in analysis of deviance tables for each parameter estimator. These tables summarize how much of the deviance in the logistic regression fit can be explained by the levels of the underlying parameters and their interactions. They are used informally to highlight the relative magnitude of coverage variabilities across parameter values. The exact values in these tables, particularly the p-values, should not be taken at face value because the logistic regression analysis performed here did not account for the nesting of samples of different size into increasing sequences, and because the ordering of the covariates, which influences the deviance statistics associated with each parameter class, was chosen arbitrarily. We present the analysis of deviance table for the intercept coefficient estimator in Table 2 and

Table 1: Coverage rates using the 95% asymptotic confidence interval from the truncated procedure. Note that coefficients that are partially confounded under the truncation procedure show undercoverage.

	600	800	1000	1200	1400	1600	1800	2000
Base	0.74	0.76	0.77	0.83	0.79	0.84	0.85	0.87
Zip	0.96	0.96	0.97	0.95	0.95	0.95	0.94	0.93
Asg	0.99	0.93	0.95	0.95	0.97	0.96	0.98	0.97
Before	0.73	0.77	0.77	0.82	0.77	0.82	0.84	0.86

reserve the remaining three tables for the appendix. In Table 2 the "previous collaboration" coefficient explains substnatially more deviance than the other parameters or interactions. This pattern holds for the estimators for the remaining three coefficeints.

The coverage rates associated with each value of the "previous collaboration" coefficient for each of the four estimators is shown in Figure 9. As suggested from the analysis of deviance table, the variability within each true "previous collaboration" value (boxplot length) is relatively small compared to the variability between these values (boxplot position). While the coefficient estimators for the Zip and Asg covariates show little sensitivity to the true value of the previous collaboration coefficient, the estimators for the intercept and previous collaboration coefficients show strong sensitivity, with coverage decreasing significantly when the true previous collaboration coefficient becomes large. This phenomenon is related to the discussion of efficiency above. Under the trunated procedure, the information about the intercept and previous collaboration coefficients is largely confounded. The only information that separates these coefficients comes from the time intervals before collaborations are observed on each dyad included in the truncated estimator. For larger values of the true previous collaboration coefficient, the confounded post-collaboration information accumulates more quickly, narrowing the intervals for both estimators, while the rate of information accumulation that separates the two coefficient accumulates at the same rate, keeping the finite sample bias the same. See Figure 7 for an illustration of this confounding and finite sample bias. As the number of actors in the sample grows, this finite sample bias slowly dissipates and the asymptotic intervals approach nominal coverage in the limit. Figure 9 shows evidence of this slow dissipation as well.

Table 2: Analysis of deviance table for Int coefficient, summarizing deviance explained by the levels of parameter values and interactions when asymptotic confidence interval coverage was modeled using a logistic regression. The coverages rates show strong sensitivity to the level of the "prev" coefficient. This table is meant for informal analysis as the logistic regression model does not take into account the nested generation mechanism employed in the simulations and uses an arbitrary ordering of the covariates.

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			50039	34911.1	
asg	3	121.00	50036	34790.1	4.70E-26
zip	3	41.83	50033	34748.3	4.36E-09
prev	3	1740.57	50030	33007.7	0.00E+00
size	7	16.87	50023	32990.8	0.018
asg:zip	9	50.75	50014	32940.1	7.80E-08
asg:prev	9	55.12	50005	32885.0	1.15E-08
zip:prev	9	36.20	49996	32848.8	3.65E-05
asg:size	21	10.53	49975	32838.3	0.971
zip:size	21	4.36	49954	32833.9	1.000
prev:size	21	14.43	49933	32819.5	0.851
asg:zip:prev	27	77.30	49906	32742.2	9.62E-07
asg:zip:size	63	22.21	49843	32720.0	1.000
asg:prev:size	63	30.39	49780	32689.6	1.000
zip:prev:size	63	21.99	49717	32667.6	1.000
asg:zip:prev:size	189	73.31	49528	32594.3	1.000

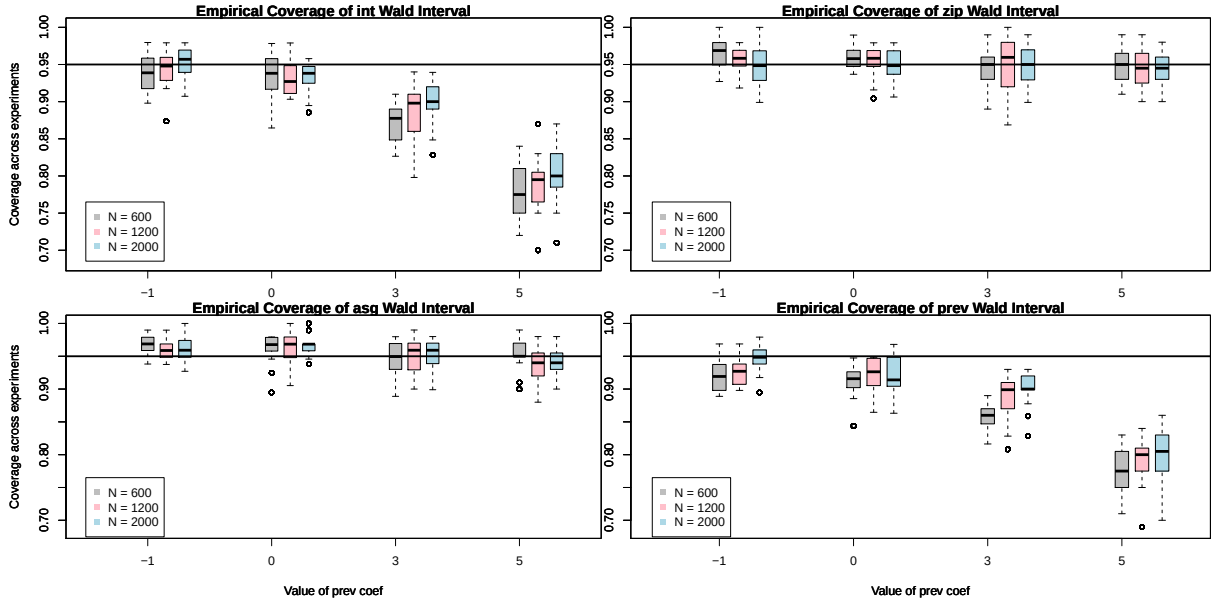

Figure 9: Coverage of 95% asymptotic confidence intervals computed using a full factorial design. Coverage was mostly sensitive to the level of the "prev" coefficient, which controls how much interaction frequency increases when a previous interaction has occurred. The truncation mechanism drops a portion of that data that uniquely informs the intercept coefficient without confounding this effect with the "prev" coefficient. For large values of "prev", confounded information for the intercept and "prev" coefficients accumulates more quickly but the finite sample bias from the portion of the truncated estimator that separates the coefficients descreases at the same rate, resulting in undercoverage. As sample size increases, this undercoverage slowly dissipates as the finite sample bias decreases.

9 Discussion

In the current era of "big data", we are encountering more and more datasets that do not fit neatly into the simple generative processes on which much of the classical theory of statistical estimation was built. For this reason, we should be careful to reconstruct the full scientific argument that we are making when we deploy a particular model in a given investigation, and make sure that the theoretical gurantees that we demand of our estimators are still relevant. In this paper, we considered the case of social network data, and followed one particular type of misspecification to show how a number of the social scientific arguments that we may wish to make with network models can fall apart when they are applied to superpopulation questions. Our investigation highlights the subtle differences between asymptotic arguments that can emerge when we study non-standard data – in this case, the non-equivalence of large-sample and superpopulation asymptotics. We hope that the thought process that we outlined here can spur on more theoretical investigations that are tailored toward the nature of the scientific question that the methodology in question is meant to answer.

Regarding the specific points of this paper, there are several loose ends that we wish to highlight.

- Although the theoretical results presented in this paper are specific to the MLE, they could be easily extended to more general model- or objective-function-based estimation procedures including GEE, M-estimation, and Bayesian approaches. In particular, several additional concentration results also due to Spokoiny allow us to generalize the notion of the effective estimand as defined in Section 2.2 to these other inference approaches.
- It may be the case that we took the "coward's way out" in pivoting out of the sparsity misspecification problem by shifting the question to sparsity-invariant estimands rather than tackling the problem of modeling sparsity structure head-on. We do hope that in ongoing research such as [26], more sophisticated probability models will be discovered that can address this nees. However, we do think that the CIR class of models can serve as a stopgap and that their computational properties make them an attractive option for asking social scientific questions of massive network data.
- We also hope that our ultimate solution to use a partial likelihood approach for eliminating the sparsity process can serve as an example for work pertaining to estimation in the presence of high-dimensional nuisance parameters. To our knowledge, this approach is not well-publicized in modeling circles where the invariance approach violates

the likelihood principle. However, in our experience here, we found it to offer an attractive level of robustness, and we will keep it as part of our modeling toolkit.

Acknowledgements

I would like to acknowledge my advisor Edo Airoldi for his valuable insight and near-infinite patience as I muddled (and continue to muddle) through this work. Thanks also to Lee Fleming for spurring this research with a fascinating social scientific question, and Edward Kao, Keli Liu, Alex Blocker, John Bischof, Alex Franks, Joe Blitzstein, and the Airoldi lab for discussion and feedback.

References

[1] Edoardo M Airoldi, Thiago B Costa, and Stanley H Chan. Stochastic blockmodel approximation of a graphon : Theory and consistent estimation. *Advances in Neural Information Processing Systems 26 (Proceedings of NIPS)*, pages 1–9, 2013.

[2] Erling Bernhard Andersen. Asymptotic Properties of Conditional Maximum Likelihood Estimators. *Journal of the Royal Statistical Society B*, 32(2):283–301, 1970.

[3] Peter J Bickel and Aiyou Chen. A nonparametric view of network models and Newman-Girvan and other modularities. *Proceedings of the National Academy of Sciences of the United States of America*, 106(50):21068–21073, December 2009.

[4] Michael Braun and André Bonfrer. Scalable Inference of Customer Similarities from Interactions Data using Dirichlet Processes. *Marketing Science*, 30(3):513–531, 2010.

[5] Göran Broström and Marie Lindkvist. Partial Partial Likelihood. *Communications in Statistics - Simulation and Computation*, 37(4):679–686, 2008.

[6] Diana Cai, Nathanael Ackerman, and Cameron Freer. An iterative step-function estimator for graphons. 2:1–27, 2014.

[7] Francois Caron and Emily B. Fox. Bayesian nonparametric models of sparse and exchangeable random graphs. *arXiv preprint*, pages 1–64, 2014.

[8] D. S. Choi, P. J. Wolfe, and E. M. Airoldi. Stochastic blockmodels with a growing number of classes. *Biometrika*, 99(2):273–284, April 2012.

[9] D. R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, January 1972.

[10] Alexander D'Amour and Edoardo Airoldi. Causal inference with social-interaction-valued outcomes. Ongoing work for publication and inclusion in dissertation.

[11] Alexander D'Amour and Edoardo Airoldi. Model misspecification and the effective estimand. Ongoing work for publicationa nd inclusion in dissertation.

[12] Alexander D'Amour, Edoardo Airoldi, and Lee Fleming. Measuring the causal effect of the Michigan Anti-trust Reform Act of 1986 on inventor collaboration dynamics in Michigan. Ongoing work for publication and inclusion in dissertation.

[13] Andrew Gelman. Parameterization and Bayesian Modeling. *Journal of the American Statistical Association*, 99(466):537–545, 2004.

[14] V P Godambe. Conditional Likelihood and Unconditional Optimum Estimating Equations. *Biometrika*, 63(2):277–284, August 1976.

[15] Prem K Gopalan and David M Blei. Efficient discovery of overlapping communities in massive networks. *Proceedings of the National Academy of Sciences of the United States of America*, 110(36):14534–9, 2013.

[16] Mark S. Handcock, Adrian E. Raftery, and Jeremy M. Tantrum. Model-based clustering for social networks. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 170(2):301–354, March 2007.

[17] Peter Hoff, Bailey Fosdick, Alex Volfovsky, and Katherine Stovel. Likelihoods for fixed rank nomination networks. *Network Science*, 1(03):253–277, 2013.

[18] Pj Huber. The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the fifth Berkeley symposium on ...*, pages 221–233, 1967.

[19] Adam Kirsch and Michael Mitzenmacher. Less Hashing , Same Performance : Building a Better Bloom Filter. *Building*, 33(2):456–467, 2006.

[20] Pavel N. Krivitsky and Mark S. Handcock. Adjusting for Network Size and Composition Effects in Exponential-Family Random Graph Models. *Statistical Methodology*, 8(4):319–339, 2011.

[21] Guan Cheng Li, Ronald Lai, Alexander D'Amour, David M. Doolin, Ye Sun, Vetle I. Torvik, Amy Z. Yu, and Fleming Lee. Disambiguation and co-authorship networks of the U.S. patent inventor database (1975-2010). *Research Policy*, 43(6):941–955, 2014.

[22] B. G. Lindsay. Nuisance Parameters, Mixture Models, and the Efficiency of Partial Likelihood Estimators. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 296(1427):639–662, 1980.

[23] James Lloyd, Peter Orbanz, Zoubin Ghahramani, and Daniel Roy. Random function priors for exchangeable arrays with applications to graphs and relational data. *Advances in Neural Information Processing Systems 26*, pages 1–9, 2013.

[24] M. Marx, D. Strumsky, and L. Fleming. Mobility, Skills, and the Michigan Non-Compete Experiment. *Management Science*, 55(6):875–889, 2009.

[25] P. McCullagh and John A. Nelder. *Generalized Linear Models, Second Edition*. CRC Press, 1989.

[26] Peter Orbanz and Daniel M. Roy. Bayesian Models of Graphs, Arrays and Other Exchangeable Random Structures. *arXiv*, 37(02):1–25, 2013.

[27] Patrick O. Perry and Patrick J. Wolfe. Point process modelling for directed interaction networks. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 75(5):821–849, 2013.

[28] Takamitsu Sawa. Information Criteria for Discriminating among Alternative Regression Models. *Econometrica*, 46(6):1273–1291, 1978.

[29] Michael Schweinberger. Instability, Sensitivity, and Degeneracy of Discrete Exponential Families. *Journal of the American Statistical Association*, 106(496):1361–1370, 2011.

[30] Cosma Rohilla Shalizi and Alessandro Rinaldo. Consistency under sampling of exponential random graph models. *Annals of Statistics*, 41(2):508–535, April 2013.

[31] Hossein Azari Soufiani and Em Airoldi. Graphlet decomposition of a weighted network. *Aistats*, 22:1–25, 2012.

[32] Vladimir Spokoiny. Parametric estimation. Finite sample theory. *The Annals of Statistics*, 40(6):2877–2909, 2012.

[33] Ulrike von Luxburg. Clustering Stability: An Overview. *Foundations and Trends ...*, pages 235–274, 2010.

[34] Dq Vu and Au Asuncion. Continuous-time regression models for longitudinal networks. *Advances in Neural ...*, pages 1–9, 2011.

[35] Halbert White. Maximum Likelihood Estimation of Misspecified Models. *Econometrica*, 50(1):1–25, 1982.

[36] Carsten Wiuf, Markus Brameier, Oskar Hagberg, and Michael P H Stumpf. A likelihood approach to analysis of network data. *Proceedings of the National Academy of Sciences of the United States of America*, 103(20):7566–7570, 2006.

[37] Wing Hung Wong. Theory of Partial Likelihood. *The Annals of Statistics*, 14(1):88–123, 1986.

[38] Bin Yu. Stability. *Bernoulli*, 19(4):1484–1500, 2013.

A Proof of Corollary 1

First, we establish that the model is responsive to the density of the network under the true, sparse process. We make use of the Silverman-Toeplitz theorem:

Theorem 5 (Silverman Toeplitz, 1911)**.** *An infinite matrix* $(a_{ij})_{i,j\in\mathbb{N}}$ *with complex-valued entries defines a regular summability method if and only if it satisfies the following three conditions:*

1. $\lim_{i\to\infty} a_{i,j} = 0 \forall j \in \mathbb{N}$
2. $\lim_{i\to\infty} \sum_{j=1}^{\infty} a_{i,j} = 1$
3. $\sup_i \sum_{j=1}^{\infty} |a_{ij}| < \infty$

A regular summability method is an infinite matrix transformation of a sequence such that, if the sequence is convergent, the limit is preserved.

Consider the sequence $(p_{ij}^{(0)} q_{ij}^{(0)})_{i<j<n}$, which are marginal probabilities that $A_{ij} = 1$ for each ij. We can rewrite the score system Equation 43 in terms of a set of regular transformations

of this sequence. Define

$$\varpi_{ij}^{(n)} = \left(\frac{\nabla_\gamma \, q_{ij}^{(\bar{\gamma}_n)}}{q_{ij}^{(\bar{\gamma}_n)}(1 - p_{ij}^{(\bar{\beta}_n)} q_{ij}^{(\bar{\gamma}_n)})} \right)$$
$$\omega_{ij}^{(n)} = \frac{\varpi_{ij}^{(n)}}{\sum_{ij} \varpi_{ij}^{(n)}}.$$

Under the boundedness and positivity conditions 3 and 4, we have that the three conditions of the Silverman-Toeplitz theorem are met. Thus, because the theorem gives that limits are preserved under this transformation, we have the following:

$$\lim_{n\to\infty} \binom{n}{2}^{-1} \sum p_{ij}^{(0)} q_{ij}^{(0)} = \lim_{n\to\infty} \binom{n}{2}^{-1} \sum p_{ij}^{(0)} q_{ij}^{(0)} \omega_{ij}^{(n)} \tag{65}$$

and similarly

$$\lim_{n\to\infty} \binom{n}{2}^{-1} \sum p_{ij}^{(\bar{\beta}_n)} q_{ij}^{(\bar{\gamma}_n)} = \lim_{n\to\infty} \binom{n}{2}^{-1} \sum p_{ij}^{(\bar{\beta}_n)} q_{ij}^{(\bar{\gamma}_n)} \omega_{ij}^{(n)} \tag{66}$$

Further, setting the score system for γ (Equation 43) equal to 0, we have that

$$\binom{n}{2}^{-1} \sum p_{ij}^{(0)} q_{ij}^{(0)} \omega_{ij}^{(n)} = \binom{n}{2}^{-1} \sum p_{ij}^{(\bar{\beta}_n)} q_{ij}^{(\bar{\gamma}_n)} \omega_{ij}^{(n)} \tag{67}$$

for each n. Combined with the previous result, we have

$$\binom{n}{2}^{-1} \sum p_{ij}^{(0)} q_{ij}^{(0)} - p_{ij}^{(\bar{\beta}_n)} q_{ij}^{(\bar{\gamma}_n)} \in o(1), \tag{68}$$

as n grows large, or that $|\mathbb{E}_0(D_n) - \mathbb{E}_{\bar{\beta}_n, \bar{\gamma}_n}(D_n)| \in o(1)$, i.e. that the model is responsive to sparsity when the weights $\omega_{ij}^{(n)}$ meet the conditions of the Silverman-Toeplitz theorem.

We now show that the asymptotic bias also has the proper rate. Define

$$\xi_{ij}^{(n)} = \frac{p_{ij}^{(\bar{\beta}_n)} q_{ij}^{(\bar{\gamma}_n)}}{\sum_{ij} p_{ij}^{(\bar{\beta}_n)} q_{ij}^{(\bar{\gamma}_n)}} \qquad \xi_{0,ij}^{(n)} = \frac{p_{ij}^{(0)} q_{ij}^{(0)}}{\sum_{ij} p_{ij}^{(0)} q_{ij}^{(0)}} \tag{69}$$

as weights. Under regularity conditions ($\sum_{ij} q_{ij}^{(\gamma)} \to \infty \forall \gamma \in \Gamma$ and $p_{ij}^{(\beta)} > 0 \forall \beta \in B$, and similar conditions on p_0, q_0), these weights satisfy the Silverman-Toeplitz conditions. We

may then check whether:

$$\lim_{n\to\infty} \frac{\sum_{ij} \xi_{ij}^{(n)} \omega_{ij}^{(n)}}{\sum_{ij} \xi_{ij}^{(n)}} \to c \qquad \lim_{n\to\infty} \frac{\sum_{ij} \xi_{0,ij}^{(n)} \omega_{ij}^{(n)}}{\sum_{ij} \xi_{ij}^{(n)}} \to c_0 \tag{70}$$

for some finite c and c_0. If so, by definition the numerators and the denominators have the same rates. Application of the Silverman-Toeplitz Theorem gives shows that this is the case, with $c = 1$. Thus, we have that:

$$\binom{n}{2}^{-1} \sum_{ij} p_{ij}^{(\bar{\beta}_n)} q_{ij}^{(\bar{\gamma}_n)} \sim \binom{n}{2}^{-1} \sum p_{ij}^{(0)} q_{ij}^{(0)} \sim \epsilon_0(n). \tag{71}$$

This implies that $|\mathbb{E}_0(D_n) - \mathbb{E}_{\bar{\beta}_n,\bar{\gamma}_n}(D_n)| = \delta(n) \in O(\epsilon_0(n))$, so the conditions for Theorem 2 are met. Under these conditions, by Theorem 2, sparsity-misspcified CID models do not define a coherent joint population estimand $(\bar{\gamma}, \bar{\beta})$. This may not be worrisome if the "moving target" behavior is confined to $\bar{\gamma}_n$. The identification condition ensures that changes in $\bar{\gamma}_n$ imply changes in $\bar{\beta}_n$ – otherwise, A would be S-ancillary for β, contradicting the assumption.

B Limiting variance inflation calculation from Section 8.1.3

In this example, β is four-dimensional, composed of the coefficients for the intercept, Zip, Asg, and previous collaboration coefficients, respectively. Let $\mathcal{I}^s(\beta)$ be the 4×4 Fisher information matrix for estimator s, written $\hat{\beta}^s$. Let $\mathcal{V}^s(\beta) = \mathcal{I}^s(\beta)^{-1}$ be the asymptotic covariance matrix of $\hat{\beta}^s$. We wish to compute the asymptotic variance ratios for each parameter estimate, given by $\frac{V_{kk}^{trunc}(\beta)}{V_{kk}^{full}(\beta)}$ for $k = 1, \cdots, 4$.

The information matrix for estimator s can be represented as follows:

$$\mathcal{I}_n^s(\beta) = \sum_{ij\in\mathcal{R}_n} \mathbb{E}\left[t_{ij}^{(1)}\right] w_{ij}^{pre,s} X_{ij}^{pre} X_{ij}^{pre\top} + \left(T - \mathbb{E}\left[t_{ij}^{(1)}\right]\right) w_{ij}^{post} X_{ij}^{post} X_{ij}^{post\top} \tag{72}$$

Here, $\mathbb{E}\left[t_{ij}^{(1)}\right]$ is the expected time of the first interaction to be observed on dyad ij, and can be used the divide the information matrix into expected information obtained from dyads before their first interactions and expected information obtained afterward. This deomposition is useful because within these time intervals the covariate vector for a dyad remains fixed. We use the superscripts *pre* and *post* to label those quantities relevant to the pre- and post-interaction periods, respectively. As is customary for generalized linear

models, we represent the information matrix contribution from each dyad ij as a weight w_{ij} and the outer product of the dyad's covariate vector X_{ij} with itself. Note that the oracle and truncated procedures only differ in the definition of w_{ij}^{pre}.

Note that because the covariates X_{ij} are discrete, the sums in Equation 72 can be collapsed into contributions by dyads with the same covariate values. In this case, because the intercept and "previous collaboration" covariates are fixed within the pre- and post-collaboration time intervals, there are only four unique covariate classes, corresponding to same/diffferent zip code, and same/different assignee. WeLOG, we fix the definitions of the covariate classes as follows:

$$\begin{aligned} X_1^{pre} &= (1,0,0,0)^\top & X_1^{post} &= (1,0,0,1)^\top \\ X_2^{pre} &= (1,0,1,0)^\top & X_2^{post} &= (1,0,1,1)^\top \\ X_3^{pre} &= (1,1,0,0)^\top & X_3^{post} &= (1,1,0,1)^\top \\ X_4^{pre} &= (1,1,1,0)^\top & X_4^{post} &= (1,1,1,1)^\top. \end{aligned}$$

Using c to index these covariate classes, and letting N_c be the number of at-risk dyads in class c so that $\sum_c N_c = \sum_{ij} R_{ij}$,

$$\mathcal{I}_n^s(\beta) = \sum_c N_c \left(\mathbb{E}\left[t_c^{(1)}\right] w_c^{pre,s} X_c^{pre} X_c^{pre\top} + \left(T - \mathbb{E}\left[t_c^{(1)}\right]\right) w_c^{post} X_c^{post} X_c^{post\top} \right). \tag{73}$$

Here $\mathbb{E}\left[t_c^{(1)}\right]$ is a slight abuse of notation, but is meant to emphasize that all dyads within a given class share the same expected time of first observed interaction.

Using Equation 73, we take the limit of the analytical inverse of $\mathcal{I}_n^s(\beta)$ for the truncated and full estimators. These limits depend on the limiting composition of N_c. For these simulations, we assume that both zip codes and assignees have fixed size as the network size grows to infinity. Combined with the generative assumption in Equation 64, this implies that asymptotically class 1, corresponding pairs of inventors with different zip codes and different assignees, grows at a faster rate than the other three covariate classes. In particular, $N_1 \in O(N_k^2)$ for $k = 2, 3, 4$.

We compute the analytic inverses using Cramer's rule, which gives $V_{kk}^s(\beta) = \frac{C_n^s(k,k)}{\det(\mathcal{I}_n^s(\beta))}$, where $C_n^s(l,m)$ is the cofactor of element l, m in $\mathcal{I}_n^s(\beta)$. Thus, the variance inflation factor can be

written

$$VI_k(\beta) = \lim_{n\to\infty} \frac{C_n^{tr}(k,k)}{C_n^{full}(k,k)} \frac{\det(\mathcal{I}_n^{full})}{\det(\mathcal{I}_n^{tr})}. \tag{74}$$

Beginning with the second factor of Equation 74, we note that these full determintants can be written as the diference of sums of four-way products of elements in $\mathcal{I}_n^s(\beta)$. The terms that grow fastest in this expression grow as N_1^2, so we can rewrite the determinant

$$\det(\mathcal{I}_n^s(\beta)) = (i_{n,22}^s i_{n,33}^s - (i_{n,23}^s)^2)(i_{n,11}^s i_{n,44}^s - (i_{n,14}^s)^2) + o(N_1^2). \tag{75}$$

Similarly, the cofactors can be written as the difference of sums of three-way products of elements in the corresponding information matrix. The relevant cofactors can also be written in terms of their fastest growing terms:

$$C_n^s(1,1) = (i_{n,22}^s i_{n,33}^s - (i_{n,23}^s)^2) i_{n,44}^s + o(N_1) \tag{76}$$
$$C_n^s(2,2) = (i_{n,11}^s i_{n,44}^s - (i_{n,14}^s)^2) i_{n,33}^s + o(N_1^2) \tag{77}$$
$$C_n^s(3,3) = (i_{n,11}^s i_{n,44}^s - (i_{n,14}^s)^2) i_{n,22}^s + o(N_1^2) \tag{78}$$
$$C_n^s(4,4) = (i_{n,22}^s i_{n,33}^s - (i_{n,23}^s)^2) i_{n,11}^s + o(N_1). \tag{79}$$

To write out the explicit forms of the elements of $\mathcal{I}_n^s(\beta)$, we define the following shorthand:

$$z_c^{pre,s} = \mathbb{E}\left[t_c^{(1)}\right] w_c^{pre,s} \qquad\qquad z_c^{post} = \left(T - \mathbb{E}\left[t_c^{(1)}\right]\right) w_c^{post}. \tag{80}$$

Evaluating Equation 73, the relevant elements of $\mathcal{I}_n^s(\beta)$ have the form

$$i_{n,11}^s = \sum_c N_c \left(z_c^{pre,s} + z_c^{post}\right) \tag{81}$$
$$i_{n,44}^s = i_{n,14}^s = \sum_c N_c z_c^{post} \tag{82}$$
$$i_{n,22}^s = N_3 \left(z_3^{pre,s} + z_3^{post}\right) + N_4 \left(z_4^{pre,s} + z_4^{post}\right) \tag{83}$$
$$i_{n,33}^s = N_2 \left(z_2^{pre,s} + z_2^{post}\right) + N_4 \left(z_4^{pre,s} + z_4^{post}\right) \tag{84}$$
$$i_{n,23}^s = N_4 \left(z_4^{pre,s} + z_4^{post}\right). \tag{85}$$

We compute the variance inflation factors by substitution. After simplification, we have

$$VI_1(\beta) = \frac{\sum_c N_c z_c^{pre,full}}{\sum_c N_c z_c^{pre,tr}} \tag{86}$$

$$VI_2(\beta) = \frac{N_3\left(z_3^{pre,tr} + z_3^{post}\right) + N_4\left(z_4^{pre,tr} + z_4^{post}\right)}{N_3\left(z_3^{pre,full} + z_3^{post}\right) + N_4\left(z_4^{pre,full} + z_4^{post}\right)} \frac{K^{full}}{K^{tr}} \tag{87}$$

$$VI_3(\beta) = \frac{N_2\left(z_2^{pre,tr} + z_2^{post}\right) + N_4\left(z_4^{pre,tr} + z_4^{post}\right)}{N_2\left(z_2^{pre,full} + z_2^{post}\right) + N_4\left(z_4^{pre,full} + z_4^{post}\right)} \frac{K^{full}}{K^{tr}} \tag{88}$$

$$VI_4(\beta) = \frac{\sum_c N_c\left(z_c^{pre,tr} + z_c^{post}\right)}{\sum_c N_c\left(z_c^{pre,full} + z_c^{post}\right)} \frac{\sum_c N_c z_c^{pre,full}}{\sum_c N_c z_c^{pre,tr}} \tag{89}$$

where

$$\begin{aligned} K^s = {} & N_2\left(z_2^{pre,s} + z_2^{post}\right) N_3\left(z_3^{pre,s} + z_3^{post}\right) + \\ & N_2\left(z_2^{pre,s} + z_2^{post}\right) N_4\left(z_4^{pre,s} + z_4^{post}\right) + \\ & N_3\left(z_3^{pre,s} + z_3^{post}\right) N_4\left(z_4^{pre,s} + z_4^{post}\right) \end{aligned} \tag{90}$$

To fix constants and ensure identification in the limit for the example in Section 8.1.3, we make additional assumptions about the sizes and ordering of the asignees and zip codes. We assume that each assignee has 200 people while each zipcode has 250 people, and that actors are assigned to these zipcodes and assignees sequentially. In this way, the adjacency matrix can be partitioned into sets of 4 zipcodes or 5 assignees such that there are no zipcode or assignee matches across these partitions. This implies that in the limit, $N_2 = 2N_3 = 3N_4$.

C Analysis of deviance tables

Table 3: Analysis of Deviance for Zip coefficient.

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			50039	19873.3	
asg	3.000	42.44	50036	19830.9	3.24E-09
zip	3.000	21.82	50033	19809.0	7.12E-05
prev	3.000	20.25	50030	19788.8	1.51E-04
size	7.000	6.18	50023	19782.6	0.518
asg:zip	9.000	32.60	50014	19750.0	1.57E-04
asg:prev	9.000	73.40	50005	19676.6	3.27E-12
zip:prev	9.000	18.16	49996	19658.5	0.033
asg:size	21.000	9.78	49975	19648.7	0.982
zip:size	21.000	7.86	49954	19640.8	0.996
prev:size	21.000	10.93	49933	19629.9	0.964
asg:zip:prev	27.000	114.86	49906	19515.0	8.30E-13
asg:zip:size	63.000	32.35	49843	19482.7	1.000
asg:prev:size	63.000	39.51	49780	19443.2	0.991
zip:prev:size	63.000	24.56	49717	19418.6	1.000
asg:zip:prev:size	189.000	141.74	49528	19276.9	0.996

Table 4: Analysis of Deviance for Asg coefficient.

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			50039	18617.7	
asg	3.000	3.30	50036	18614.4	0.347
zip	3.000	3.68	50033	18610.8	0.298
prev	3.000	47.54	50030	18563.2	2.67E-10
size	7.000	7.72	50023	18555.5	0.358
asg:zip	9.000	26.48	50014	18529.0	0.002
asg:prev	9.000	27.33	50005	18501.7	0.001
zip:prev	9.000	41.19	49996	18460.5	4.62E-06
asg:size	21.000	24.46	49975	18436.0	0.271
zip:size	21.000	14.08	49954	18422.0	0.866
prev:size	21.000	24.02	49933	18398.0	0.292
asg:zip:prev	27.000	135.41	49906	18262.5	2.15E-16
asg:zip:size	63.000	30.46	49843	18232.1	1.000
asg:prev:size	63.000	35.38	49780	18196.7	0.998
zip:prev:size	63.000	60.39	49717	18136.3	0.570
asg:zip:prev:size	189.000	139.69	49528	17996.6	0.997

Table 5: Analysis of Deviance for Prev coefficient.

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			50039	36518.9	
asg	3.000	67.19	50036	36451.7	1.70E-14
zip	3.000	22.12	50033	36429.6	6.16E-05
prev	3.000	1395.53	50030	35034.1	2.75E-302
size	7.000	46.68	50023	34987.4	6.44E-08
asg:zip	9.000	40.43	50014	34947.0	6.35E-06
asg:prev	9.000	26.53	50005	34920.4	0.002
zip:prev	9.000	20.05	49996	34900.4	0.018
asg:size	21.000	10.55	49975	34889.8	0.971
zip:size	21.000	6.42	49954	34883.4	0.999
prev:size	21.000	11.04	49933	34872.4	0.962
asg:zip:prev	27.000	101.10	49906	34771.3	1.70E-10
asg:zip:size	63.000	23.59	49843	34747.7	1.000
asg:prev:size	63.000	18.46	49780	34729.2	1.000
zip:prev:size	63.000	18.36	49717	34710.9	1.000
asg:zip:prev:size	189.000	71.78	49528	34639.1	1.000