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 [34] for an overview).

The network data that we consider in this paper have the following form. There is a set of actors V, and a record Y_V of the pairwise interactions between the actors in V. We limit our discussion to *undirected* network data, so outcomes correspond to unique unordered actor-pairs, or dyads. Thus, Y_V contains $\binom{|V|}{2}$ outcomes. We denote an individual outcome in Y_V corresponding actors i and j, where $i < j < |V|$ as Y_V^{ij}. As we will discuss further in Section 2, the subscript V is used to indicate the index of the sample within an overarching stochastic process of which Y_V is a finite-dimensional projection, while the superscript ij indicates the index of an actor-pair or dyad within the specific sample Y_V.

Each outcome Y_V^{ij} lives in an outcome space of interaction records $\mathcal{Y}$, which 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. In addition to outcomes, there is often a corresponding covariate collection X_V, containing covariate information for each of the $\binom{|V|}{2}$ outcomes in Y_V. We denote the individual elements of X_V that correspond to a particular pair of actors i and j as X_V^{ij}.

Generally, we can divide the inferential questions that investigators seek to answer with this sort of data 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 V. In these cases, the probability law of interest is the replication distribution of a particular random graph Y_V. In examples of single-sample problems, investigators may wish to infer the presence or absence of links that are missing from the current dataset Y_V, or predict future interactions among the actors in V. The second category is superpopulation problems, where investigators wish to infer properties that are shared between social networks defined on different actor sets, say V and V'. We call these superpopulation problems because they require the notion of a superpopulation from which both Y_V and $Y_{V'}$ were drawn to

justify generalizing inferences between heterogeneous samples. In this cacse, the probability law of interest is governs the over-arching stochastic process from which both Y_V and $Y_{V'}$ were drawn. In examples of superpopulation problems, investigators may wish to test whether two network samples Y_V and $Y_{V'}$ were generated by the same stochastic 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 superpopulation contexts, parameter estimates are often unstable when investigators wish to compare networks of different size, a problem most notably documented in [21]. 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. We give an example of such a fit in Section 1.1.

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., [4, 9]. 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 (by characterizing the replication distribution of the observed sample Y_V) should also be useful for answering superpopulation questions about network samples defined on distinct actor sets (by characterizing the superpopulation from network samples defined on arbitrary actor sets V' are drawn). 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 data-generating process. So far, in extending notions of large-sample consistency to network models, authors in this literature have focused on arguments that are relevant to single-sample inference.

In this paper, we develop a theoretical framework for evaluating a network model's suitability for superpopulation investigations. Using this framework, we argue that the poor performance of network models in superpopulation inference tasks is a symptom of model misspecification, specifically the aspect of the model that implicitly embeds the observed

network sample into a superpopulation process. This misspecification is largely immaterial to answering single-sample questions, but is central to superpopulation investigations. We show that one particular type of embedding 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 nonzero interactions $\sum_{i<j<|V|} \mathbf{1}_{Y_V^{ij} \neq 0}$ to the number of potential interactions in a network sample $\binom{|V|}{2}$ – 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 that the dyad-wise outcomes in a network sample Y_V, are mutually independent conditional on observable or latent characteristics. Examples of these models include network regression models as in [28] or exchangeable random graph models [24], which include as special cases latent class [9] and latent space models [19]. This is make sparsity misspecification a major concern because 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 sparsity invariance as a realistic and robust modeling principle, and present 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 analysis problem that motivated this work as a running example. We use an inventor-disambiguated version of the US patent record [22] 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 V, and the pairwise outcomes Y_V record pairwise coauthorships on patents. 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, for each pair of inventors ij, the coauthorship record Y_V^{ij} has a point-process structure, but lower resolution representations are also possible. For example, we can define the outcome Y_V^{ij} as the number of collaborations between inventors i and j over some fixed observation interval.

The inventor data also contains side information that we can use as covariates X to model collaboration behavior, including each inventor's firm and zipcode. In examples throughout this paper, we 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. Thus, in this example we define X_V^{ij} to be a 3-component binary vector for each ij.

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 [28], where we specify the log-hazard of a collaboration event between inventors i and j as a linear combination of the zipcode, assignee, and previous collaboration covariates described above (we describe this specification in full detail in Section 6.1.1). We apply this model to regional collaboration networks constructed from a 6-year window of interaction data beginning in 1983, defining V for each model fit to be the set of inventors residing in a particular Census Bureau Statistical Area (CBSA) surrounding a major US city during the observation window. The results for each CBSA are shown on the left of Figure 1. These demonstrate that the parameter estimates for each fit depend strongly on the size of the network sample, and that the fits return extremely large effect estimates and extremely small uncertainy estimates. We also display the results of our sparsity-invariant methodology described in Section 5.1 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, but this example makes clear that it is difficult to separate the effect of network sparsity on the parameter estimates (manifested as large sample size effects) from true differences in

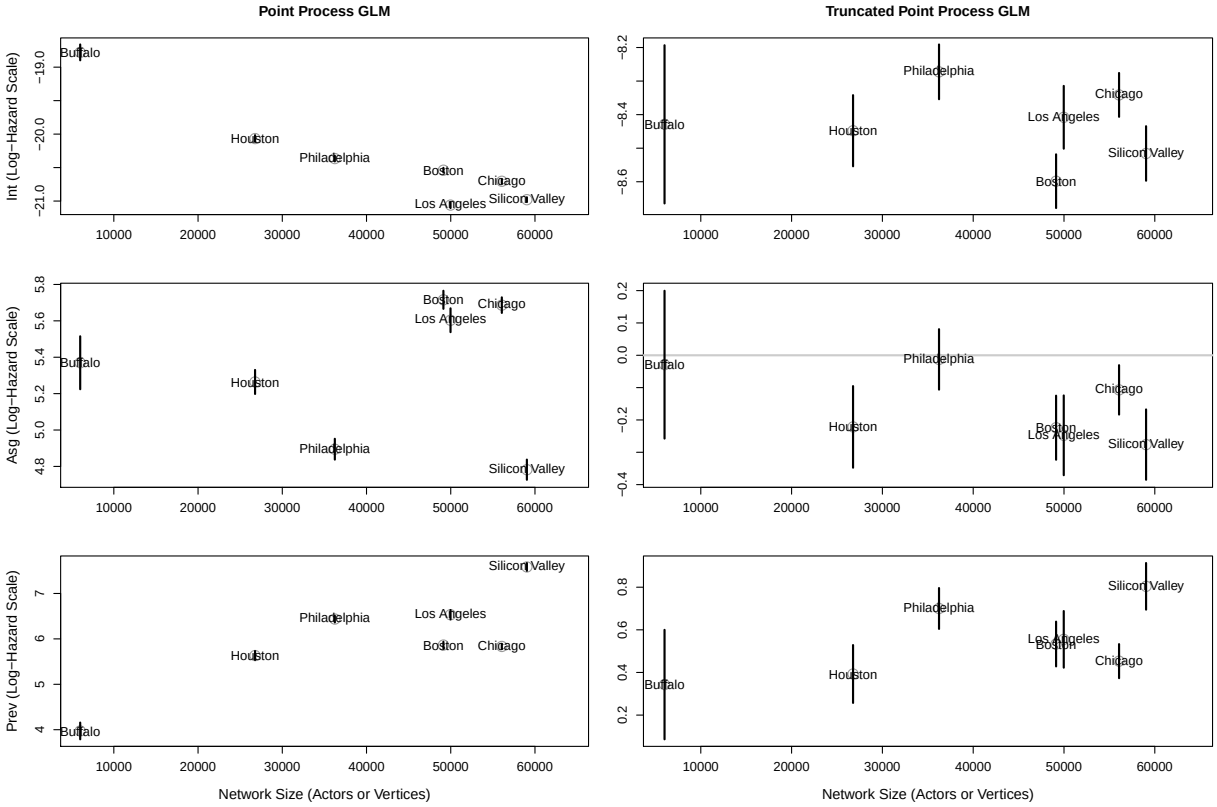

Figure 1: Inferred parameter values and asymptotic intervals from a simple point process regression model explaining patent collaboration events occurring in different regional inventor networks in the United States. (Left) parameter estimates from this standard conditionally independent dyad (see Section 3.3) 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 5.1) show stability across regions with more realistic effect and error estimates.

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 6.1.

1.2 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 and Rinaldo's work in [31], which defined a network superpopulation as a stochastic process indexed by actor-sets, and network samples drawn from this superpopulation as finite dimensional projections of this population process. Shalizi and Rinaldo 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 (we assume this is the case for all models we consider here) we use the stochastic process framework to investigate whether the *inferences* obtained from a model have the invariances necessary to be suitable for answering questions about network superpopulations. In particular, we require that inferences obtained from a model fit be stable across samples drawn from the same network superpopulation, regardless of the indices of those samples.

We devote a significant amount of effort to formalizing this notion of stability. The question of whether a model gives *stable* inferences, in the sense that nominally similar samples yield similar inferences (the meaning of "nominally similar" depends on the particular invariance that the investigator requires of the estimation procedure, as we describe below), is a critical question when we consider the 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 [39] for a summary). In network analysis, [30] 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., [21]. 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 on a dimension 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 dimension. 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.

In our main negative result, 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 within a sample to reliably fit a model [4, 9]. 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 network modeling. 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 under stochastically consistent extensions are dense [4, 24]. 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 [37]. In another vein, [8] 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, we develop a sparsity-invariant approach to modeling and inference which we present as an alternative to modeling sparsity explicitly. 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 [28], 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 specifying a non-trivial risk set instead. Here, we treat the risk set as a set of 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 [11, 38], although it can also be classified more specifically as a conditional likelihood [15] 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 [14], 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. [21] proposed an offset term that stabilizes change statistics in ERGMs, but did not attempt to justify this as a likelihood-based approach. [18] 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 in superpopulation inference, these proposals have focused on single-sample fit [6], novel network representations [32], or approximate likelihood inference for computational efficiency [16]. Notably, our proposed procedure is able to achieve similar computational efficiency using an exact likelihood function.

Social scientific questions about the organizational behavior of inventors holding patents in the United States in, e.g., [25], were the original motivation of this work. In another series of papers, [12] and [13], we extend this modeling framework to the causal inference setting and use data from the US patent record made available by [22] to infer the causal effect of a policy change on the collaboration dynamics of inventors.

1.3 Contributions

We have organized the contributions of this paper as follows. The theoretical contributions in the first half of the paper lay the groundwork for the main negative result presented in Section 4. This result requires three building blocks. First, in Section 2 we introduce formalism that defines superpopulation inference precisely in the context of social network analysis. Second, in Section 2.2, we state a stability criterion for superpopulation inferences to be scientifically useful. We state this criterion in terms of the "effective estimand" of a procedure, which describes the target of a procedure's estimation – the principle states that, to be useful, the effective estimand should remain stable across samples that are drawn from the same population. Third, in Section 3 we discuss a property of social network data that makes fulfilling the criterion in Section 2.2 difficult – in particular, we describe network sparsity in a superpopulation context. This section includes a result showing that many popular network models are "sparsity-misspecified", or fail to model this property correctly. Finally, we use these building blocks in Section 4 to establish the main negative result – that under mild conditions, the MLE of a sparsity-misspecified model violates our stability criterion because it targets a different effective estimand in different samples drawn from the same population.

In the second half of the paper, we propose methods based on a novel modeling framework that defines and stably measures properties of network generating processes that are conserved across samples, even when the generating process is sparse. In Section 5, we present a "Conditionally Independent Relationship" (CIR) class of graph processes that have a sparsity-independent component, and in Section 5.1 we present sparsity-invariant method-

ology for estimating properties of this sparsity-independent subprocess. Finally, we present simulated and real data examples in Section 6, and conclude with a discussion in Section 7.

1.4 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 7.

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

In this section, we present a formal characterization of network superpopulation inference problems, where the investigator's goal is to obtain parameter estimates and predictive distributions from a sample Y_V that can be used in downstream analyses that involve distinct actor sets $V' \neq V$. 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 the investigator specify some common, underlying probabilitstic structure that encodes the investigator's assumptions about how outcomes occurring among different sets of actors are relevant to each other. 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 that the investigator specify a probabilistic structure specific to the observed actor set V.

2.1 Network superpopulations

To formally characterise network superpopulation inference, we require a probabilistic object that can play the role of a network superpopulation in a statistical problem. 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 [31], and define an actor-indexed stochastic process that can serve as a network superpopulation. Here, we use slightly different notation from [31] to emphasize the relationship to the data analysis settting.

Let $\mathbb{V}$ be a countably infinite set 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 infinite actor set $\mathbb{V}$, we define the interaction graph population as follows

Definition 1 (Random Graph 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 $\mathbb{P}_V$.*

Using random graph processes as building blocks, we write the network superpopulation estimation problem as follows. Let $Y_{0,\mathbb{V}}$ be a random interaction process that is the true 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.

There are two points of our construction of the network superpopulation inference problem that we wish to emphasize.

First, despite being infinite objects, random graph processes have distinct properties and play a distinct role in our statistical arguments from sequences of increasingly large random networks that are often invoked in asymptotic analysis of network models for single-sample inference. The key mathematical difference between these objects is that the increasing random graph sequences that have been deployed before in large-sample consistency arguments are not required to be Kolmogorov consistent. For example, in [3], the authors define a sequence of ever-larger exchangeable random graphs whose expected degree, ρ_n, decreases with n so that the limit of the sequence has a vanishing network density, achieving a so-called sparse limit. As we discuss later in Section 3, it has been shown that no extension of a non-trivial exchangeable random graph process has a sparse limit [27], so this sequence cannot define a consistent stochastic process. This is not a problem for the purposes of a single-sample argument, where the limit of the infinite sequence serves as a deterministic analogy for a large but finite network sample Y_V – for example, such an analogy provides some guidance about how much information we can expect to recover about the internal structure of a large network sample with very few realized links.

On the other hand, a random graph process that is shared by different network samples is an essential element of superpopulation problems. In this case, the law being estimated must simultaneously define outcome distributions on differing actor sets to justify propagating inferences from one actor set to another. In the arguments that follow, we use the random graph process to test whether inferences obtained from distinct finite samples drawn from the same process maintain a particular type of invariance. Thus, instead of using this infinite object to generate a limit, we use it to interrogate relationships between analyses performed on its finite projections.

The second point we wish to emphasize is that the idea of a random graph process is not new; the critical part of Definition 1 is the representation we use for the random graph process. Following [31], we take a "top-down" view of this stochastic process rather than the "bottom-up" view that is commonly taken in analyses of the statistical properties of network models (note that we emphasize *statistical properties* here because top-down stochastic

process representations have appeared in the analysis of *generative properties* of network models, e.g., the point-process-based graph processes of [8]). In particular, previous treatments have represented an infinite stochastic process as the large-sample limit of a generating process, defining finite random networks as the primitive mathematical objects from which the stochastic process is derived by extension (e.g., Duplication-Attachment models in [37]). On the other hand, we represent a stochastic process here as the primitive mathematical object from which finite random graphs are derived by projection.

This difference in representation is important because, while all random interaction processes have both representations, the top-down representation allows an investigator to formally specify global properties of a network superpopulation *without* specifying a generative mechanism for achieving those properties. By allowing investigators to specify properties of a superpopulation that they are explicitly *unable* to capture in the generative working model, this top-down representation can serve as a powerful tool for assessing the impact of model misspecification in superpopulation inference.

2.2 Misspecification and superpopulation estimation

In superpopulation estimation for social systems, misspecification is a near-inevitability – there is little reason to believe that any parsimonious model can capture the full complexity of human social dynamics. This raises the question of what criteria a misspecified model must meet to play a useful role in a superpopulation inquiry. This question is not trivial, because when the proposed model is misspecified, so that $\mathbb{P}_{0,\mathbb{V}} \notin \mathcal{P}_\Theta$, we cannot rely on the nominal model-based interpretation of parameter estimates $\hat{\theta}_V$ alone to draw scientific conclusions. Instead, we think of the model-based estimator as a measurement of the underlying social system, and hope that this measurement reveals some of the relevant structure of the system, regardless of its model-based interpretation. One minimal property for such a measuremnt to be useful is *stability*, or an estimator's tendency to map similar generating processes to similar values in the parameter space, under appropriate deifnitions of "similar".

Many versions of stability (corresponding to different notions of "similar") have been proposed in the Statistics literature (see [39] for a review), although they are not always explicitly described as stability arguments. In the misspecification literature, stability arguments for the MLE have generally been presented in terms of large-sample consistency of the MLE for

a "pseudo-true" parameter [29], defined as the value in the parameter space that satisfies

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

or the maximizer of the *expected* log-likelihood. Huber [20] most famously showed that for data whose true generating process is iid and models that satisfy mild regularity conditions, the MLE converges to the pseudo-true parameter regardless of misspecification, while [36] 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: for a fixed underlying generating process $\mathbb{P}_{0,V}$, if samples $Y_V^{(1)} \sim \mathbb{P}_{0,V}$ and $Y_V^{(2)} \sim \mathbb{P}_{0,V}$, then $\hat{\theta}_V^{(1)} \approx \hat{\theta}_V^{(2)}$, where the "$\approx$" operator is defined in terms of a convergence rate tied to the size of the index set V. This type of stability is adequate if we think of the set of actors V as the extent of the social system we wish to characterize – it establishes that repeated measurements of this system will be internally consistent.

In superpopulation problems, investigators hope to measure social processes that are conserved between actor-sets. In this formulation, the choice of actor-set V is a design decision that is driven entirely by the investigator and not the system of interest, so estimators $\hat{\theta}_V$ and $\hat{\theta}_{V'}$ computed from distinct actor-sets V and V' are viewed as different measurements of the same underlying object. To be useful in this context, estimators $\hat{\theta}_V$ should be interpretable as measurements of the same superpopulation property regardless of the actor-set V used to make the measurement. This requires that stable between different choices of the actor-set V to include in the sample Y_V. Formally, for a fixed superpopulation process $\mathbb{P}_0$, we require that, for any V and V' such that $V \neq V'$, if $Y_V \sim \mathbb{P}_{0,V}$ and $Y_{V'} \sim \mathbb{P}_{0,V'}$, then $\hat{\theta}_V \approx \hat{\theta}_{V'}$, where the "$\approx$" operator is left to be defined.

Note that this criterion is defined in terms of finite-sample properties of an estimator – finite-sample differences between V and V', such as their relative sizes, are important here – so a definition of the "$\approx$" operator based on large-sample consistency is not generally applicable. Instead, we formalize our heuristic statement above by first defining the "target" of an estimation procedure (that is, the value the estimation procedure is measuring) when it is applied to a particular sample Y_V, and then defining $\hat{\theta}_V \approx \hat{\theta}_{V'}$ to mean that $\hat{\theta}_V$ and $\hat{\theta}_{V'}$ measure, or *effectively estimate*, the same quantity.

2.2.1 The effective estimand of the MLE

We make novel use of the pseudo-true parameter to define the target of an estimator in a finite sample. In particular, for any sample Y_V, we interpret the pseudo-true parameter $\bar{\theta}_V$ to be the deterministic target, or *effective estimand*, of the sample-specific MLE $\hat{\theta}_V$. This interpretation is most directly justified by finite-sample concentration results, showing that under mild conditions, the sampling distribution of the MLE is concentrated about the pseudo-true parameter, even in finite-samples (see, for example, Spokoiny [33] or Lemma 1). These results serve as finite-sample analogues to Huber's consistency result, as they admit a characterization of the MLE's sampling distribution in finite samples in terms of the pseudo-true parameter $\bar{\theta}_V$. Crucially for our argument about superpopulation inference, the effective estimand allows us to diagnose situations where the distribution of the MLE $\hat{\theta}_V$ depends systematically on the sample index V. For example, in the main result, we will use this to characterize the systematic effect of the sample size $|V|$ on MLE's computed from sparsity misspecified network models.

Note that the notion of an effective estimand is more general than the notion of a pseudo-true parameter – in this case, it happes that the pseudo-true parameter plays the role of the effective estimand for the MLE, but for other estimation procedures (e.g., Bayes rules or GEE), another functional would define the effective estimand. We discuss the more general notion of the effective estimand in terms of estimating functionals in a companion paper. Also note that it is possible that $\bar{\theta}_V$ is not a unique quantity, if the maximand in *Equation* 1 is not unique. In these case, we may also consider $\bar{\theta}_V$ to be set-valued – this does not change our results that characterize the effective estimand, although all of our examples will involve cases where the pseudo-true parameter is unique.

For the scope of this paper, we simply note several facts that justify our interpretation of the pseudo-true parameter as the effective estimand of the MLE. First, when the model $\mathcal{P}_{\mathbb{V},\Theta}$ is correctly specified, so that there is some parameter $\theta_0 \in \Theta$ such that $\mathbb{P}_{0,V} = \mathbb{P}_{\theta,V}$, then the pseudo-true parameter $\bar{\theta}_V = \theta_0$ for all $V \subset \mathbb{V}$ – by this argument, the MLE under misspecification is "estimating" the pseudo-true parameter in the same way that the MLE under a correct specification is estimating the true parameter θ_0. Furthermore, we can arrive at the pseudo-true parameter by inverting several desirable properties of an estimator, establishing that the pseudo-true parameter is "well-estimated" by the MLE. For example, recalling that Fisher consistency is one of the defining properties of the MLE, $\bar{\theta}_V$ is value in the parameter space Θ for which the MLE is Fisher consistent. Additionally, from the more general framework of estimating equations, $\bar{\theta}_V$ 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 $KL(\mathbb{P}_{0,V}||\mathbb{P}_{\theta,V})$ among all models in $\mathcal{P}_{\Theta,V}$ [29]. Thus, the pseudo-true parameter indexes the KL projection of the true distribution of Y_V into the finite-dimensional model family $\mathcal{P}_{\Theta,V}$. Given that the MLE $\hat{\theta}_V$ indexes the KL projection of the empirical distribution of Y_V into the model family $\mathcal{P}_{\Theta,V}$, we can interpret the MLE as a plug-in estimator of the pseudo-true parameter $\bar{\theta}_V$.

2.2.2 Stability criterion for superpopulation inference

Using our interpretation of the pseudo-true parameter $\bar{\theta}_V$ as the effective estimand of the MLE $\hat{\theta}_V$, we can establish a well-defined stability criterion for superpopulation inference.

Criterion 1. *A procedure is* superpopuatlion stable *for making inferences about a superopulation process $\mathbb{P}_{0,\mathbb{V}}$ only if, for any finite sample Y_V generated according to $\mathbb{P}_{0,V}$, the effective estimand $\bar{\theta}_V$ of the estimator $\hat{\theta}_V$ is invariant to the indexing set V.*

Criterion 1 is a common-sense, minimal bar to set for methods used in superpopulation inquiries – if we wish to interpret a sample-specific MLE $\hat{\theta}_V$ as an estimate of a superpopulation quantity, we should require that the target of estimation not depend on idiosyncratic properties of the sample encoded in V, or conversely, estimates computed from different samples drawn from the same source should be measurements of the same superpopulation quantities.

Note that Criterion 1 is not always directly verifiable, because computing the effective 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 the investigator was unable to encode directly in the model specification. When this is the case, we can use the representation of a network superpopulation presented in Section 2 to deduce how the effective estimand would behave if the true social process $\mathbb{P}_{0,\mathbb{V}}$ had this unmodeled property. In the sections below, we follow a line of inquiry of this style, and derive some properties of the effective estimand when the proposed model does not match the sparsity of the true data-generating process. In particular, we will show that the effective estimand $\bar{\theta}_V$ must depend on the sample size $|V|$, implying that sparsity misspecified models violate Criterion 1 and are therefore inappropriate tools for superpopulation inference.

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, the word "sparsity" is used to describe the following phenomenon: in large network samples, an overwhelming proportion of actor-pairs engage in no interactions, and the larger the network sample is, the more dominating this proportion of zeros becomes. Formally, we represent this by encoding pairwise social outcomes Y_V^{ij} 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 [5], among others have approached sparsity in terms of sequences of distributions over random graphs of growing size or expected size. Thus, in these discussions, "sparsity" is a property of a sequence of random graphs. Notably, these definitions do not constrain these random graph sequences 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 characterize the replication distributions of large network samples with fixed actor sets when the expected number of observed ties is relatively small. On the other hand, in this paper we wish to focus on superpopulation questions, so we define sparsity as a property of a random graph process instead of a random graph sequence.

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 0, and define the indicator random variables $A_V^{ij} = \mathbf{1}_{\{Y_V^{ij} \neq 0\}}$.*

The density operator D with respect to the element 0 has the form

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

giving the proportion of pairwise outcomes in 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 Graph Process)**.** *Let $Y_{\mathbb{V}}$ be a random 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 random graph Y_V has the property $\mathbb{E}(D(Y_V)) < \epsilon$.*

Note that, by our definition, a sparse graph process can be used to produce sparse graph sequences in the sense of [5]. In fact, 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. Note that a sparse graph sequence constructed in this way will be guaranteed to be Kolmogorov consistent.

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 a random 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),*

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

as $n \to \infty$. Similarly, we say random 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),

$$\frac{\mathbb{E}[D(Y_{V_n})]}{\mathbb{E}[D(Y'_{V_n})]} \to C$$

as $n \to \infty$.

3.1 Example: Empirically observed sparsity in patent collaboration network

In Figure 2 we show an example of an empirically observed "sparsity" phenomenon that maps cleanly onto the mathematical formalism presented in the previous section. From the dataset described in Section 1.1, we explore subsamples of the set of all patent coauthorships in the Boston area in a 6-year time interval beginning in 1983.We then obtain sequences of increasing, nested subgraphs from this regional collaboration network by randomly drawing a sequence of zipcodes and incrementally adding the batches of inventors who live in these zipcodes to the network subsample. In Figure 2, each line corresponds to one of these subgraph sequences, with the x-axis showing the number of inventors included in the subgraph and the y-axis showing the network density of that subgraph.

Even in this finite example, we see that the maximal density of the network clearly decreases with sample size. This justifies the "limiting to 0" notion presented in the Definition 3, despite the fact that the "limiting" network density in finite real-world networks is a positive constant. To obtain an empirical analogue of the sparsity rate $\epsilon_0(n)$, the figure would need to include all possible subgraph sequences (V_n).

3.2 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 $\mathbb{P}_{0,\mathbb{V}}$ is not correctly represented by the inferential model $\mathcal{P}_{\Theta,\mathbb{V}}$, we say that the model is *sparsity misspecified*. Intuitively, 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)**.** *Let (V_n) be an increasing sequence of vertex sets ordered by subset inclusion. 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, for any sequence (V_n),*

$$\frac{\mathbb{E}_\theta[D(Y_{V_n})]}{\mathbb{E}_0[D(Y_{V_n})]} \to 0 \text{ or } \infty \quad \forall \theta \in \Theta, \tag{2}$$

as $n \to \infty$, where $\mathbb{E}_\theta$ and $\mathbb{E}_0$ are expectations taken with respect to $\mathbb{P}_{\theta,\mathbb{V}}$ and $\mathbb{P}_{0,\mathbb{V}}$, respectively.

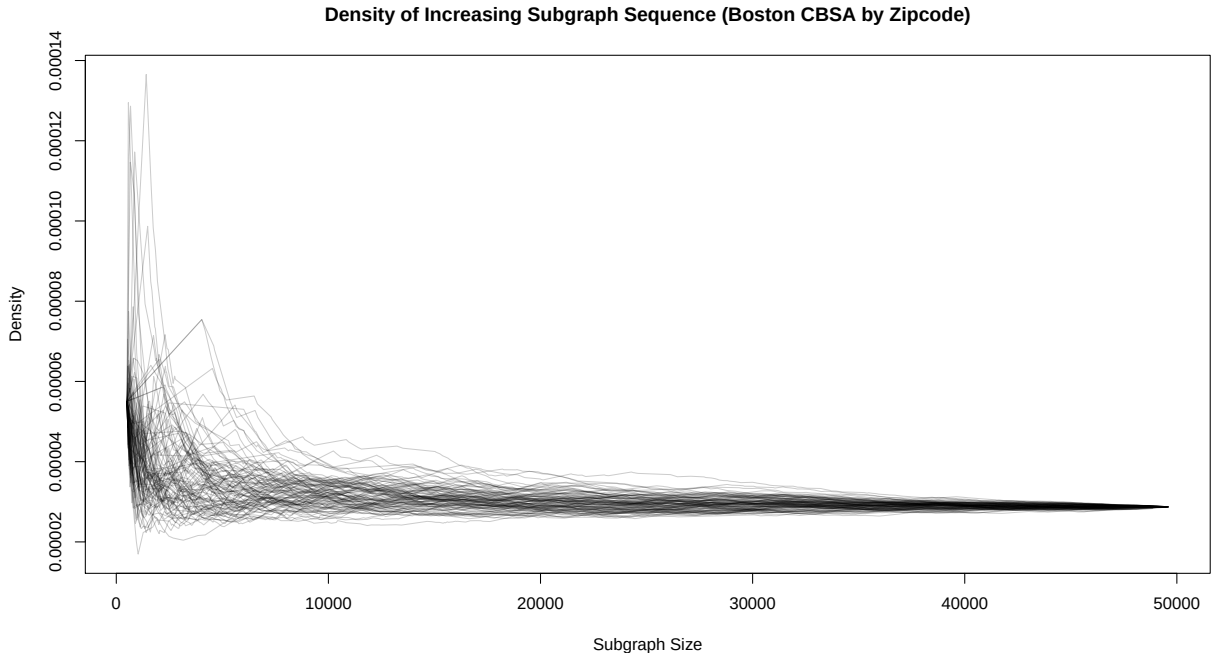

Figure 2: Sequences of random subgraphs drawn from the Boston-area inventor collaboration network obseved over a 6-year time interval beginning in 1983. Each line is a randomly generated subgraph sequence, generated by building up a subgraph zipcode-by-zipcode in a random order. A clear relationship between network size and density is visible here. This is the phenomenon meant that we model in Definition 3.

3.3 Example: Sparsity misspecificaiton in infinitely exchangeable random graph models

Sparsity misspecification is particularly prominent in model families that are built on local assumptions about how individual actors make decisions to interact. Infinitely exchangeable random graph models form the most prominent class of such generative network models. These models assume that the likelihood for any network sample Y_V drawn from a super-population process $Y_{\mathbb{V}}$ is invariant to permutations of the actor-set V – this translates to joint exchangeability of the rows and columns of any finite 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 Y_V can be treated as a set of $\binom{|V|}{2}$ pairwise conditionally independent replications, given observed and potentially unobserved covariates. They also generate simple predictions at the dyad level based only on local information. We describe several different classes of these models in turn.

The simplest subclass of infinitely exchangeable random graph model 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, [28, 35, 17, 34]). These models assign a particular observed network sample Y_V with covariates X_V a likelihood of the form:

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

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.

More general exchangeable random graph models include 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 [24] . This class of models has been unified under an array-exchangeability representation by Aldous and Hoover that, up to isomorphism, maps these latent covariate processes to a single probability surface W on the

Get better citation from Edo.

unit square. Given this surface, a network sample Y_V is generated by randomly assigning each actor in V a position C_V^i so that the pairwise covariate X_V^{ij} is generated by querying $W(C_V^i, C_V^j)$. Several recent works have been dedicated to estimating this latent surface, called the graphon, directly [7, 1]. Models with this structure induce the following likelihood on network samples

$$P(Y_V \mid X_V) = \int_{\mathcal{C}_V} \prod_{ij} P(Y_V^{ij} \mid W(C_V^i, C_V^j)) dF(C_V). \tag{4}$$

Model specifications that mix latent and observed covariates have also been proposed in several places, e.g., [17].

Several authors have noted that infinitely exchangeable graph models without covariates cannot be extended to form non-trivial sparse graph processes – that is, any infinitely exchangeable random graph process that is sparse can only generate empty network samples Y_V for any V. Orbanz and Roy show this most explicitly in [27], using a law of large numbers argument to show that any graph sequence constructed from an exchangeable random graph process would have a limiting network density $\lim_{n\to\infty} D(Y_{V_n}) = \frac{1}{2}\int_{[0,1]^2} W(x,y)dxdy$, which is 0 only if $W(\cdot,\cdot)$ is zero almost everywhere. In our current terminology, this result indicates that infinitely exchangeable random graph models are sparsity misspecified when they are applied to study sparse social networks. With appropriate conditions on observed covariates $X_{\mathbb{V}}$, we can extend this result to exchangeable random graph models with covariates, including CID models.

Theorem 1. *Let $\mathcal{P}_{\Theta,\mathbb{V}}$ be an infinitely exchangeable random graph process family.Let $X_{\mathbb{V}}$ be the population set of covariates, and denote by $\mathcal{N}_\theta \subset \mathcal{X}$ the set of covariate vectors so that for a given $\theta \in \Theta$, $\mathbb{P}_\theta(Y_{\mathbb{V}}^{ij} \neq 0 \mid X_{\mathbb{V}}^{ij} \in \mathcal{N}_\theta) = 0$. Assume that for each $\theta \in \text{int}\Theta$, the limiting proportion of covariate vectors $X_{V_n}^{ij} \in \mathcal{N}_\theta$ is $1-\nu$ for some nonzero ν. If this is the case, the model is sparsity misspecified.*

The argument here is straightforward. The covariate vector $X_{\mathbb{V}}^{ij}$ simply parameterizes the surface W described by Aldous and Hoover, so that every $X_{\mathbb{V}}^{ij}$ defines a corresponding surface $W_{X_V^{ij}}$. For each V_n, the marginal probability $\mathbb{P}(Y_{V_n}^{ij} \neq 0 \mid X_{V_n}^{ij})$ is the integral of $W_{X_{V_n}^{ij}}$. 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 $\mathbb{P}(Y_{V_n}^{ij} \neq 0 \mid X_{V_n}^{ij})$ 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 the case of infinitely exchangeable random graph models, it is possible to confirm sparsity misspecification for sparse social network applications *a priori* because all of the non-trivial random graph processes that these families include have limiting densities that converge to positive constants. Other non-exchangeable model families, for example the preferential attachment model , do include sparse graph processes, and in these cases it is not possible to judge sparsity misspecification *a priori*. However, many families that include sparse graph processes impose a particular functional form on the sparsity rate – in these cases, sparsity misspecification is still possible, but generally we do not have enough prior knowledge of the true generating process' sparsity rate to judge this misspecification until after the data have been examined.

citation

4 Main Result: Moving Target Theorem

In the last few sections, we have established a statistical framework for representing superpopulation inference, discussed a minimal condition under which the MLE from a misspecified model can be reasonably interpreted as a superpopulation quantity, and identified sparsity misspecification as a common issue in the study of sparse social networks. In this section, we bring these ideas together and show that MLE's derived from sparsity-misspecified models that meet a particular goodness-of-fit condition violate Criterion 1, and therefore do not admit a superpopulation interpretation.

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*

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

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

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 increasing sequence of vertex sets from $\mathbb{V}$. Suppose that the following hold:*

(M1) For some finite n, $\mathbb{E}_0(D(Y_{V_n})) > 0$.

(M2) The inferential family $\mathcal{P}_{\Theta,\mathbb{V}}$ is sparsity misspecified for the true population process $\mathbb{P}_{0,\mathbb{V}}$.

(M3) The inferential model is responsive to the sample density $D(Y_{V_n})$ under the true population process and

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

(M4) 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 $\bar{\theta}_{V_n} \neq \bar{\theta}_{V_{n'}}$, and the MLE of the model violates Criterion 1.

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 θ such that the law $\mathbb{P}_{\theta,\mathbb{V}}$ has rate $\epsilon_0(n)$, while also fulfilling the non-emptiness condition (M1). Thus, for any n, there exists an n' such that $\bar{\theta}_{V_n} \neq \bar{\theta}_{V_{n'}}$. □

This result establishes a fundamental tension between single-sample and superpopulation inference when a model is sparsity misspecified. In particular, if a sparsity misspecified model $\mathcal{P}_{\Theta,\mathbb{V}}$ fits individual network samples well, so that the targeted distribution of best fit $\mathbb{P}_{\bar{\theta},V}$ tends to capture the density of the network sample $D(Y_V)$, then this excludes the possibility that the model family can also be used for superpopulation inference. This resolves the seemingly paradoxical observation that popular sparsity misspecified models like CID models or exchangeable random graph models (described in Section 3.3) tend to give nonsensical results in superpopulation contexts despite having strong theoretical support for performance in single-sample inference – given that they are sparsity misspecified, these models fail as tools for superpopulation inference *precisely because* they are effective tools for single-sample inference.

The "moving target" problem identified here manifests in several ways in applied investigations. Because the inferential model's MLE is effectively estimating distinct quantities from network samples of different size, even if they are drawn from the same network superpopulation 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 illustrate these difficulties in parameter estimate interpretation in the next section.

4.1 Example: Poisson regression with binary covariate

We demonstrate some of the difficulties that result from the Theorem 2 in a simple example based on a hypothetical analysis of the patent collaboration network data presented in Section 1.1.

Let $\mathbb{V}$ be a superpopulation of inventors, from which we have sampled a set of individuals V of size n. Let Y_V be a matrix recording the number of pairwise patent collaborations that have taken place between the n sampled inventors, so that Y_V^{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 as $\mathbb{P}_{0,V}$. For each entry ij of Y_V, let X_V^{ij} be a binary covariate that indicates whether inventors i and j work for the same firm. The investigator is interested in

summarizing network samples Y_V so that they may be compared, e.g., to make statements about whether within-firm collaborations are more prominent in one industry than another.

The investigator also knows the following facts about the collaboration-generating process:

(A1) The true collaboration-generating process $Y_{0,\mathbb{V}}$ is sparse in the sense of Definition 3 with an unknown rate $\epsilon_0(n)$.

(A2) All firms have finite size.

(A3) A non-vanishing fraction of firms have a positive number of expected within-firm interactions.

However, the investigator is unable to encode all of these assumptions into a tractable modeling framework for network samples Y_V. Because it is intuitive and computationally convenient, the investigator proposes a model family $\mathcal{P}_{\Theta,\mathbb{V}}$ whose finite-dimensional distributions have the form of a Poisson regression model:

$$Y_V^{ij} \overset{\perp\!\!\!\perp}{\sim} \text{Pois}(\exp(\theta^{(1)} + X_V^{ij}\theta^{(2)})), \tag{8}$$

where the parameter vector $\theta \equiv (\theta^{(1)}, \theta^{(2)})$ can take values in $\Theta \equiv \mathbb{R}^2$. According to standard interpretations of GLM coefficients [26], $\theta^{(1)}$ is the log of the interaction rate of any "between-firm" inventor pair, while $\theta^{(2)}$ is the log ratio of interaction rates between any "within-firm" and any "between-firm" inventor pair. For a given sample Y_V, the investigator uses maximum likleihood estimation to obtain estimates $\hat{\theta}_V$, which will be used to compare different network samples.

We can now ask whether the analysis satisfies Criterion 1, which is a necessary condition for estimates $\hat{\theta}_V$ obtained from different samples to be comparable in general. We will show that under some simple conditions, Criterion 1 is indeed violated because the model's effective estimand depends on the size of the indexing set V.

We make the following assumptions to ensure that the analysis is identifiable

(B1) $\mathbb{E}_0(Y_V^{ij})$ is finite for all ij and V.

(B2) For some finite n', for every V such that $|V| > n'$, the expected number of within-firm

and between-firm interactions are nonzero:

$$\sum_{ij} \mathbb{E}_0(Y_V^{ij})(1 - X_V^{ij}) > 0 \quad \text{and} \quad \sum_{ij} \mathbb{E}_0(Y_V^{ij}) X_V^{ij} > 0$$

(B3) The variance of the total number of collaborations is proportional to its expectation, so that for all V, there exists a $d < \infty$ such that

$$\mathbb{V}\text{ar} \sum Y_V^{ij} \leq d\mathbb{E}_0 \sum Y_V^{ij}.$$

Because the model proposed in Equation 8 is an exponential family, the effective estimand has a particularly appealing analytical form that mimics the form of the MLE with expectations of sufficient statistics plugged in:

$$\bar{\theta}_V^{(1)} = \log \left(\frac{\sum_{ij} \mathbb{E}_0(Y_V^{ij} \mid X_V^{ij} = 0)(1 - X_V^{ij})}{\sum_{ij}(1 - X_V^{ij})} \right) \tag{9}$$

$$\bar{\theta}_V^{(2)} = \log \left(\frac{\sum_{ij} \mathbb{E}_0(Y_V^{ij} \mid X_V^{ij} = 1) X_V^{ij}}{\sum_{ij} X_V^{ij}} \Bigg/ \frac{\sum_{ij} \mathbb{E}_0(Y_V^{ij} \mid X_V^{ij} = 0)(1 - X_V^{ij})}{\sum_{ij}(1 - X_V^{ij})} \right). \tag{10}$$

Given this functional form, we can establish the following proposition

Proposition 1. *Fix a sequence of sets of actors (V_n), such that $|V_n| = n$. with a corresponding sequence of covariate arrays (X_{V_n}) associated with each actor set in (V_n). Under assumptions (A1), (A2), (B1), and (B2), the CID Poisson model in Equation 8 violates Criterion 1 for drawing superpopulation inferences about the collaboration-generating process $Y_{0,\mathbb{V}}$; in other words, the implied effective estimand depends on n.*

Proof. Given (A1), the true genrating process $Y_{0,\mathbb{V}}$ is sparse, so by Theorem 1, the CID Poisson model in Equation 8 is sparsity misspecified. Given (B1), all samples Y_{V_n} with $n > n'$ are expected to be non-empty. Now, we check that the model is responsive with respect to network density. Taking $g(x) = 1 - \exp(-\exp(x))$, or the c-log-log transformation, we can

write

$$\begin{aligned}
\mathbb{E}_{\bar{\theta}}(D(Y_{V_n})) &= \binom{n}{2}^{-1}\left[g(\bar{\theta}^{(1)}_{V_n}+\bar{\theta}^{(2)}_{V_n})\sum_{ij}X^{ij}_{V_n}+g(\bar{\theta}^{(1)}_{V_n})\sum_{ij}(1-X^{ij}_{V_n})\right] \\
&< \binom{n}{2}^{-1}\left[\exp(\bar{\theta}^{(1)}_{V_n}+\bar{\theta}^{(2)}_{V_n})\sum_{ij}X^{ij}_{V_n}+\exp(\bar{\theta}^{(1)}_{V_n})\sum_{ij}(1-X^{ij}_{V_n})\right] \\
&= \binom{n}{2}^{-1}\sum_{ij}\mathbb{E}_0(Y^{ij}_{V_n}) \\
&\sim O(\epsilon_o(n)),
\end{aligned}$$

where the second step follows from (A2) and the inequality $1-e^{-x} < x$ for $x > 0$, the third step follows from Equations 9 and 10, and the final step follows from assumptions (A1) and (B1). Thus, the model is responsive with respect to network density and the plug-in prediction bias decreases at the appropriate rate, so by Theorem 2, the model violates Criterion 1. □

In this particular investigation, Proposition 1 would manifest in a number of ways. We can show this directly by establishing that the MLE $\hat{\theta}_V$ concentrates around the effective estimand $\bar{\theta}_V$ for all finite samples Y_V, and then showing that the effective estimand can be manipulated arbitrarily by the choice of V.

Lemma 1. *The distribution of the MLE for the parameters of the model in Equation 8 concentrates around its effective estimand for all finite samples V, with probability bounds given by*

$$\begin{aligned}
\mathbb{P}(|\hat{\theta}^{(2)}_V-\bar{\theta}^{(1)}_V|\leq\log(1+\delta)) &\geq 1-\frac{d}{\delta^2\mathbb{E}_0\sum Y_i(1-X_i)} \\
\mathbb{P}(|\hat{\theta}^{(2)}_V-\bar{\theta}^{(2)}_V|\leq\log(1+\delta)) &\geq 1-\frac{4d}{\delta^2\mathbb{E}_0\sum Y_i(1-X_i)}-\frac{4d}{\delta^2\mathbb{E}_0\sum Y_iX_i}
\end{aligned}$$

The proof is included in the appendix.

Given this result, we can characterize the behavior of the MLE in terms of the effective estimands. First, we characterize the behavior of the effective estimand vector $\bar{\theta}_V$.

Proposition 2. *$\bar{\theta}^{(1)}_V$ can be made arbitrarily negative by selecting a large actor-set V.*

Proof. Given (A2), the proportion of between-firm dyads $\sum(1-X_{ij})/\binom{|V|}{2} \to c > 0$. Combined with the sparsity condition (A1) and the finite expectation condition (B1), the ratio

in Equation 9 must fall to zero as $|V| \to \infty$. □

Proposition 3. *The effective estimand $\bar{\theta}_V^{(2)}$ can be made arbitrarily positive and large by incorporating a larger number of firms in the study.*

Proof. Because of (A2), the ratio of within-firm to between-firm dayds falls to zero as $n \to \infty$, or formally

$$\frac{\sum_{V_n} X_{V_n}^{ij}}{\sum_{V_n}(1 - X_{V_n}^{ij})} \to 0.$$

Given the sparsity of the overall process (A1), and the scaling of between-firm dyads (B1), the denominator ratio in Equation 10 goes to zero as $n \to \infty$. Meanwhile, given (A2) and (A3), the numerator ratio in Equation 10 converges to a constant as $n \to \infty$. □

Combining Lemma 1 with these propositions, we have shown that the estimates $\hat{\theta}_V$ are strongly sensitive to the sizes and firm compositions of the samples that the investigator collects. Given this instability, it would be difficult to distill meaningful comparative conclusions from this analysis.

5 Conditionally Independent Relationship Processes

So far, we have established that sparsity misspecification is difficult to avoid and that sparsity misspecified models are poor tools for obtaining scientifically meaningful insights for superpopulation inquiries. This difficulty highlights a mismatch between the measurement tools (in the form of models, as in Section 3.3) that are currently available for describing social network generating processes, and the aspects of real social network processes that we hope to measure in superpopulation investigations. This motivates us to seek out aspects of network superpopulations that we can measure stably with modeling tools that are currently available.

As a solution to this problem, we describe a class of random graph processes that admit a particular factorization in their generating process that explicitly separates some of the process law from the sparsity of the process. For this class of processes, it is possible to make stable inferences about sparsity-invariant superpopulation properties, regardless of the sparsity rate of the process as a whole. We call this class of processes *conditionally independent relationship*, or CIR processes.

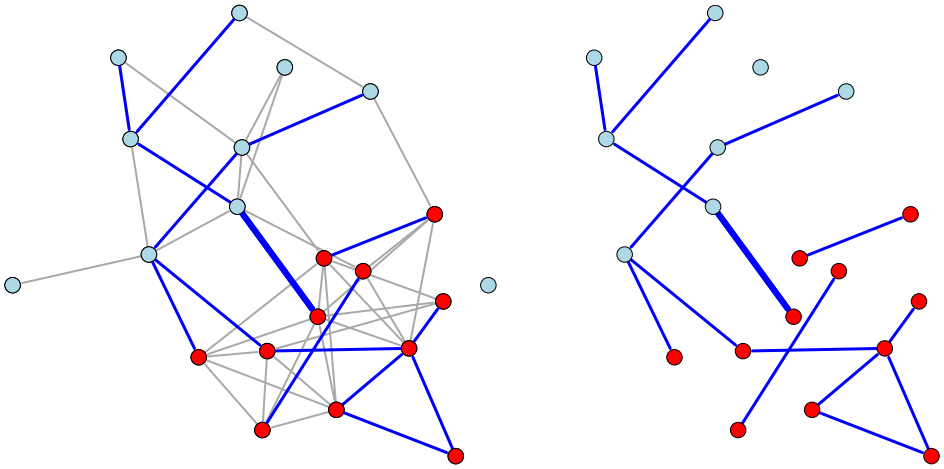

Figure 3: (Left) Diagram of CIR generation process, where gray ties are "relationships" and blue ties are observed interactions. To generate an observable interaction, a pair of actors much first have a relationship, or in the language of the diagram, blue ties can only appear on top of gray ties. (Right) The observed network sample, where relationships with no observed interactions are indistinguishable from dyads with no relationship.

In CIR processes, dyad-level observations Y_V^{ij} are drawn from a zero-inflated process in which only certain pairs of actors are capable of generating non-zero outcomes – we say these pairs of actors have a "relationship". This corresponds to the generative intuition that in order to generate an observable interaction where $Y_V^{ij} \neq 0$ (e.g., collaborate on a patent applications), two actors must first have an unobservable social relationship R_V^{ij} (e.g., they must be able to recognize each other on the street). Furthermore, conditional on these relationships and covariates X_V, pairwise outcomes Y_V^{ij} are independent – hence, the outcomes correspinding to each relationship in the actor-set are conditionally independent. Figure 3 provides a graphical description of this process.

Formally, for any actor-set index $V \subset \mathbb{V}$, we characterize the observed random graph sample Y_V jointly with unobservable random graph sample R_V, which we call the *relationship graph*. R_V is itself a binary random graph. Similarly to the observable outcome graph Y_V, we assume that the relationship graph is a subgraph from a superpopulation relationship process $R_{\mathbb{V}}$. According to this model, to generate a network sample Y_V from law $\mathbb{P}_V$, we follow two-stage generating process conditional on any covariates X_V: the relationship graph R_V is drawn first, then, for each ij where $R_V^{ij} = 0$, Y_V^{ij} is set deterministically to 0, while for each ij where $R_V^{ij} = 1$, Y_V^{ij} is drawn independently from its marginal distribution $\mathbb{P}(Y_V^{ij} \mid X_V^{ij})$. For each finite sample indexed by actor-set V, this induces a joint distribution function of the

following form:

$$\mathbb{P}_V(R_V, Y_V \mid X_V) = \mathbb{P}_V(R_V \mid X_V) \prod_{ij} \mathbb{P}_V(Y_V^{ij} \mid X_V^{ij}, R_V^{ij}). \tag{11}$$

$$= \mathbb{P}_V(R_V \mid X_V) \prod_{\{ij : R_V^{ij}=0\}} \mathbf{1}_{\{Y_V^{ij}=0\}}^{1-R_V^{ij}} \mathbb{P}_V^{(R)}(Y_V^{ij} \mid X_V^{ij})^{R_V^{ij}} \tag{12}$$

Given the factorization in Equation 11, we can make the following statement about the sparsity rate of a CIR process.

Proposition 4. *Let $Y_{\mathbb{V}}$ and $R_{\mathbb{V}}$ be the observable and unobservable components of a random graph process, whose finite dimensional distributions can be factorized according to Equation 11. Let $X_{\mathbb{V}}$ be the population set of covariates, and denote by $\mathcal{N}_\theta \subset \mathcal{X}$ the set of covariate vectors so that for a given $\theta \in \Theta$, $\mathbb{P}_{\mathbb{V}}(Y_{\mathbb{V}}^{ij} \neq 0 \mid R_{\mathbb{V}}^{ij} = 1, X_{\mathbb{V}}^{ij} \in \mathcal{N}_\theta) = 0$. Assume that for each $\theta \in \text{int}\Theta$, the limiting proportion of covariate vectors $X_{V_n}^{ij} \in \mathcal{N}_\theta$ is $1 - \nu$ for some nonzero ν. Then the sparsity rate of the marginal process $Y_{\mathbb{V}}$ is equal to the sparsity rate of the marginal process $R_{\mathbb{V}}$.*

This proposition can be shown using the same LLN argument as in Theorem 1. The independent structure of the observable process $Y_{\mathbb{V}}$ conditional on the relationship process $R_{\mathbb{V}}$ ensures that the marginal sparsity rates of $Y_{\mathbb{V}}$ and $R_{\mathbb{V}}$ can only differ by a constant factor. Thus the sparsity rate $\epsilon_0(n)$ of the observable process $Y_{\mathbb{V}}$ is not a function of the conditional distribution $\mathbb{P}_V(Y_V^{ij} \mid X_V^{ij}, R_V^{ij} = 1)$.

This fact implies that, if a true social process is in the CIR class, the conditional finite dimensional distributions $\mathbb{P}_V(Y_V \mid R_V, X_V)$ do not have the same sparsity-related inhomogeneities that characterize the marginal finite-dimensional distributions $\mathbb{P}_V(Y_V \mid X_V)$ and drive the result in Theorem 2. This makes the conditional outcome process law $\mathbb{P}_{\mathbb{V}}(Y_{\mathbb{V}} \mid R_{\mathbb{V}}, X_{\mathbb{V}})$ a promising object of measurement for superpopulation inquiries when it is infeasible to correctly model the sparsity of the social process of interest. In practical terms, if the true social process is sparse but allows a CIR factorizationa as in Equation 11, the answer to the general question "How does *any pair* of actors generate social interactions?" must explain why the social process is sparse, but the answer to the specific question "How do pairs of actors with an existing relationship generate social interactions?" does not include such an explanation. Thus, if our modeling tools are ill-equipped to correctly model sparsity, it is reasonable to switch focus to the latter question. We discuss a procedure for estimating the properties of the sparsity-invariant conditional outcome process in the next section.

5.1 Truncated estimator for CIR processes

We specify an inferential model family $\mathcal{P}_{\Theta,\mathbb{V}}$ composed of CIR processes for the observable process $Y_{\mathbb{V}}$ – the finite-dimensional distributions of these laws can be written as summations over the R_V component in the joint specification in Equation 11. For this model family, we divide the parameter space Θ into two components, so that $\theta = (\beta, \gamma)$ for $\beta \in B$ and $\gamma \in \Gamma$ and $\Theta \equiv B \times \Gamma$. We specify the laws contained in $\mathcal{P}_{\Theta,\mathbb{V}}$ to have finite dimensional distributions of the form

$$\mathbb{P}_{\theta,V}(Y_V \mid X_V) = \sum_{R_V \in \mathcal{R}_V} \left[\mathbb{P}_{\theta,V}(R_V \mid X_V) \mathbf{1}_{\{Y_V^{ij}=0\}}^{1-R_V^{ij}} \mathbb{P}_\beta^{(R)}(Y_V^{ij} \mid X_V^{ij})^{R_V^{ij}} \right], \tag{13}$$

where $\mathcal{R}_V$ is the space of all binary graphs on the actor-set V. Based on this specification and Proposition 4, the parameters β appear to represent the sparsity-invariant portions of a CIR process; thus, we defined these as the parameters of interest. The parameters γ (the remaining components of θ) are treated as nuisance parameters. Note that while the conditional distributions $\mathbb{P}_\beta(Y_V \mid R_V, X_V)$ are free of γ, the marginal distribution $\mathbb{P}_\theta(R_V \mid X_V)$ may depend on components of β.

Direct maximum likelihood estimation using Equation 13 would be the most straightforward option, but such an approach runs into the same sparsity misspecification problems described in Section 4. Because the relationship graph R_V corresponding to the sample is unobserved, the investigator is still required to specify a functional form for the marginal distribution $\mathbb{P}_{V,\theta}(R_V \mid X_V)$, running the same risk of sparsity misspecification-induced instability in parameter estimates. This is particularly undesirable given that the parameters of interest β are meant to characterize a sparsity-free component of the social process $Y_{\mathbb{V}}$. In particular, we can show the following corollary to Theorem 2.

Corollary 1 (Moving Target with Nuisance)**.** *In the setting of Theorem 2, assume in addition to (M2)–(M4) that*

(M5) The inferential family is specified such that parameters of interest β are identified by the binarized process A_{V_n}.

Then, the effective estimand for the parameters of interest $\bar{\beta}_{V_n}$ varies with n in the sense that for any n, there exists an $n' > n$ such that $\bar{\beta}_{V_n} \neq \bar{\beta}_{V_{n'}}$, and the MLE of the model violates Criterion 1.

Note that this corollary holds even when the true process law $\mathbb{P}_{0,\mathbb{V}}$ can be factorized according

to Equation 11 for all V_n, and the inferential family has specified the conditional process $\mathbb{P}_{\beta,\mathbb{V}}(Y_{\mathbb{V}} \mid R_{\mathbb{V}}, X_{\mathbb{V}})$ correctly. We discuss this further in Section 5.2.

To break this infinite regress, we develop a likelihood-based procedure that models *less* of the available data in order to obtain invariance to the sparsity of the social process $Y_{\mathbb{V}}$ – in paticular, instead of using the full likelihood approach that marginalizes over the nuisance process $R_{\mathbb{V}}$, we develop a *partial likelihood* approach that allows ignores the relationship process $R_{\mathbb{V}}$ entirely. Partial likelihood estimation – proposed by Cox [10, 11] and rigorously treated by Wong [38] – is a semiparametric estimation technique that uses a carefully designed factorization of the likelihood into a sequence of conditional probability functions, some of which are free of nuisance parameters. The parameters of interest are then estimated by maximizing a "partial likelihood" composed exclusively of these nuisance-free factors. Intuitively, one usually arrives at this factorization by only specifying a model for an incomplete subset of the observed data conditional on the rest. See [10, 11, 38] for examples.

In this case, we employ a convenient factorization of Equation 13 based on conditioning on intermediate observable indicators $A_V^{ij} = \mathbf{1}_{\{Y_V^{ij} \neq 0\}}$ for each ij:

$$\mathbb{P}_\theta(Y_V \mid X_V) = \sum_{R_V \in \mathcal{R}_V} \left[\mathbb{P}_\theta(R_V \mid X_V) \prod_{ij} \mathbf{1}_{\{Y_V^{ij}=0\}}^{1-R_V^{ij}} \mathbb{P}_\beta(A_V^{ij} \mid X_V^{ij}, R_V^{ij}=1)^{R_V^{ij}} \mathbb{P}_\beta(Y_V^{ij} \mid X_V^{ij}, A_V^{ij}=1)^{A_V^{ij}} \right] \tag{14}$$

$$= \left[\sum_{R_V \in \mathcal{R}_V} \mathbb{P}_\theta(A_V, R_V \mid X_V) \right] \left[\prod_{ij} \mathbb{P}_\beta(Y_V^{ij} \mid X_V^{ij}, A_V^{ij}=1)^{A_V^{ij}} \right]. \tag{15}$$

Notably, the second factor in Equation 15 has no dependence on the unobserved relationship graph R_V and can therefore be factored out of the summation over R_V, freeing it of any nuisance parameters. Intuitively, this reflects the fact that for each actor-pair ij for which a nonzero interaction is observed such that $A_V^{ij} = 1$, it is also known that there is an underlying relationship such that $R_V^{ij} = 1$. This nuisance-free factor corresponds to the zero-truncated distribution of the network sample Y_V implied by the model $\mathbb{P}_\theta(Y_V \mid X_V)$, and can be written

$$\mathbb{P}^{tr}_{\beta,V}(Y_V \mid X_V) \equiv \prod_{ij} \mathbb{P}_\beta(Y_V^{ij} \mid X_V^{ij}, A_V^{ij}=1)^{A_V^{ij}} = \prod_{ij} \left[\frac{\mathbb{P}_\beta(Y_V^{ij} \mid X_V^{ij}, R_V^{ij}=1)}{1 - \mathbb{P}_\beta(Y_V^{ij}=0 \mid X_V^{ij}, R_V^{ij}=1)} \right]^{A_V^{ij}} \tag{16}$$

We call this the *truncated likelihood*. We can maximize the log of these likelihood factors alone to obtain a partial estimator for β, which we call the *maximum truncated likelihood*

estimator, or MTLE, and write $\hat{\beta}_V^{tr}$.

Although we motivated the decomposition Equation 15 with partial likelihood theory, in this case, the MTLE $\hat{\beta}_V^{tr}$ is itself the MLE of a well-defined sub-experiment of the origianl observation model. In the original observation model, the investigator chose an actor-set V and observed all social interactions Y_V among those actors; in the derived sub-experiment the investigator only observed the non-zero outcomes in Y_V such that $A_V^{ij} = 1$ in the analysis, and ignores data pertaining to the rest of the actor-pairs, including the total number of actor-pairs $\binom{V}{2}$. Because it is a proper maximum likelihood estimator, we define the effective estimand of the maximum truncated likelihood estimator $\bar{\beta}_V^{tr}$ in the same way that we did for the MLE of the full data model. Formally, the maximum truncated likelihood estimator of $\hat{\beta}_V^{tr}$ and its effective estimand $\bar{\beta}_V^{tr}$ are given by

$$\hat{\beta}_V^{tr} = \arg\max_\beta \log \mathbb{P}^{tr}_{\beta,V}(Y_V \mid A_V) \qquad \text{and } \bar{\beta}_V^{tr} = \arg\max_\beta \mathbb{E}_0 \log \mathbb{P}^{tr}_{\beta,V}(Y_V \mid A_V) \tag{17}$$

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

5.2 Superpopulation stability of the truncated estimator

Here we show that the MTLE $\hat{\beta}_V^{tr}$ has an effective estimand that does not in general depend on the sparsity of the process $Y_\mathbb{V}$, making it a suitable candidate for drawing superpopulation inferences in investigations of social process that are sparse but allow factorization Equation 11. We show this in a special case, when the model family $\mathcal{P}_{\Theta,\mathbb{V}}$ includes a correct specification for the the conditional process $\mathbb{P}_\beta(Y_\mathbb{V} \mid X_\mathbb{V}, R_\mathbb{V})$.

Theorem 3 (Superpopulation Stability of Truncated Estimator)**.** *Let $Y_\mathbb{V}$ is a random graph process, $\mathbb{P}_{0,\mathbb{V}}$ be the true law governing this process, and $\mathcal{P}_{\Theta,\mathbb{V}}$ be a model family proposed by the investigator. Assume the following*

(T1) The finite-dimensional distributions of $Y_\mathbb{V}$ can be factorized as in Equation 11 for all sample indices V.

(T2) The model family $\mathcal{P}_{\Theta,\mathbb{V}}$ correctly specifies the conditional process $\mathbb{P}_{0,\mathbb{V}}(Y_\mathbb{V} \mid X_\mathbb{V}, R_\mathbb{V})$, so that there exists a $\beta_0 \in B$ such that $\mathbb{P}_{\beta_0,\mathbb{V}}(Y_V^{ij} \mid X_\mathbb{V}, R_\mathbb{V}) = \mathbb{P}_{0,\mathbb{V}}(Y_V^{ij} \mid X_\mathbb{V}, R_\mathbb{V})$.

(T3) The model family $\mathcal{P}_{\Theta,\mathbb{V}}$ is specified so that β is identified by the truncated data $\{Y_V^{ij} : A_V^{ij} = 1\}$.

Then the effective estimand of the MTLE does not depend on V and, in particular, $\bar{\beta}_V^{tr} = \beta_0$ for all V.

Proof. Expanding the effective estimand defined in Equation 17,

$$\bar{\beta}_V^{tr} = \arg\max_\beta \mathbb{E}_0 \sum_{ij} A_V^{ij} \left[\log \mathbb{P}_\beta(Y_V^{ij} \mid X_V^{ij}, A_V^{ij} = 1)\right] \tag{18}$$

$$= \arg\max_\beta \mathbb{E}_0 \left[\mathbb{E}_0 \left[\sum_{ij} A_V^{ij} \log \mathbb{P}_\beta^{(A)}(Y_V^{ij} \mid X_V^{ij}) \;\middle|\; A_V\right]\right]. \tag{19}$$

By the correct specification assumptions (T1) and (T2), the truncated likelihood derived derived from $\mathcal{P}_{\Theta,\mathbb{V}}$ is also correctly specified in situations where A_V is known. Thus, the argument of the inner conditional expection in Equation 19 is maximized by the same value β_0 for all values of A_V, so the entire expression in Equation 19 is maximized by β_0. If this were not the case, so that Equation 19 were maximized by some other value $\tilde{\beta} \neq \beta_0$, by (T3), all terms of the implicit sum in the outer expectation could be increased by switching the argument of the maximization to β_0, yielding a contradiction. Thus, the effective estimand is equal to β_0 for all V. □

Remark 1. *The identification assumption (T3) excludes several cases where the truncated estimator $\hat{\beta}_V^{tr}$ would be meaningless – for example, cases where the outcomes in Y_V are binary such that $Y_V^{ij} = A_V^{ij}$ for all ij. In this case, all parameter values β yields identical truncated likelihood functions for the data $\{Y_V^{ij} : A_V^{ij} = 1\}$ because the sample size $|V|$ is not included in the truncated likelihood.*

Although the correct specification conditions (T1) and (T2) in Theorem 3 are strong, this does not make the theorem trivial. The conditions isolate sparsity misspecification as a potential source of instabiity in the sense of Criterion 1 – note, for example, that even if this condition held in the full likelihood inference case, by Corollary 1, sparsity misspecification in the remaining components of the model would be sufficient to induce a violation of the superpopulation stability in Criterion 1. On the other hand, Theorem 3 puts no requirements on the sparsity of Y_V or the range of sparsity rates allowed by the model family $\mathcal{P}_{\Theta,\mathbb{V}}$. Thus, while there may be other reasons that a truncated estimator $\hat{\beta}_V^{tr}$ may violate Criterion 1 – for example, if the correct specification assumption is not met – an inadequate explanation for the sparsity of a social process Y_V is no longer a sufficient condition. We demonstrate the stability of the truncated estiator in simulation studies and real data analysis in Section 6.

5.3 Statistical efficiency of the truncated estimator

The truncated estimator $\hat{\beta}_V^{tr}$ achieves robustness to the sparsity of the social process Y_V by modeling less of the data than the investigator has available. Such a choice necessarily comes at the cost of statistical efficiency. In this section, we examine the worst-case efficiency loss that could be incurred from using the truncated estimator $\hat{\beta}_V^{tr}$ over an idealized full-likelihood "oracle" esimator $\hat{\beta}_V^{or}$. In particular, we study the case where the correct specification and identification assumptions (T1)–(T3) hold and we compute the oracle estimator $\hat{\beta}_V^{or}$ using the true the relationship graph R_V for all V, so that the only free parameters in the estimation problem are the parameters of interest β that characterize the conditional process $\mathbb{P}_{\mathbb{V}}(Y_{\mathbb{V}} \mid R_{\mathbb{V}}, X_{\mathbb{V}})$.

In this section, we evaluate the efficiency of the estimators $\hat{\beta}_V^{or}$ and $\hat{\beta}_V^{tr}$ in terms of Fisher Information, making use of asymptotic arguments. We supplement these arguments with with finite-sample simulation studies in Section 6.1.

For convenience, we define the following quantities:

$$p_{\beta,V}^{ij} = \mathbb{P}_{\beta,V}(A_V^{ij} = 1 \mid R_V^{ij} = 1, X_V^{ij}) \tag{20}$$

$$l_{\beta,V}^{ij}(Y_V^{ij}) = \log \mathbb{P}_{\beta,V}(Y_V^{ij} \mid R_V^{ij} = 1, X_V^{ij}) \tag{21}$$

These are, respectively, the probability that a given dyad has an observed nonzero interaction value, and the log-likelihood of the outcome of a single dyad, given that the dyad has an underlying relationship.As with previous notation, we write the true superpopulation analogues of these quantities with a subscript 0 instead of β.

Under the assumption that the relationship graph R_V 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 oracle or truncated likelihood. Taking $ij \in R_V$ to be shorthand for $\{ij : R_V^{ij} = 1\}$, we can then rewrite the oracle and truncated log-likelihoods for a whole sample Y_V:

$$L_{\beta,V}^{tr}(Y_V) = \sum_{ij \in R_V} A_V^{ij}(l_{\beta,V}^{ij}(Y_V^{ij}) - \log p_{\beta,V}^{ij}) \tag{22}$$

$$L_{\beta,V}(Y_V) = \sum_{ij \in R_V} \left[A_V^{ij} \log p_{\beta,V}^{ij} + (1 - A_V^{ij}) \log(1 - p_{\beta,V}^{ij})\right] + L_{\beta,V}^{tr}(Y_V). \tag{23}$$

The Fisher Information matrices for the truncated and oracle log-likelihoods are given by:

$$\mathcal{I}^{tr}_{\beta,V} = -\left[\sum_{ij\in R} p^{ij}_{0,V}\left(\mathbb{E}_0(\nabla^2_\beta\, l^{ij}_{\beta,V}(Y^{ij}_V) \mid A_{ij}=1) - \nabla^2_\beta \log p^{ij}_{\beta,V}\right)\right] \tag{24}$$

$$\mathcal{I}_{\beta,V} = -\left[\sum_{ij\in R} p^{ij}_{0,V}\,\nabla^2_\beta \log p^{ij}_{\beta,V} + (1-p^{ij}_{0,V})\,\nabla^2_\beta \log(1-p^{ij}_{\beta,V})\right] + \mathcal{I}^{tr}_{\beta,V}. \tag{25}$$

Quite intuitivley, the information ignored by the truncated procedure comes from the distribution of the binary variables $\mathbb{P}_V(A_V \mid R_V = 1)$. We note that the ignored information expression in Equation 25 scales as the number of relationships in the sample V, $\sum_{ij} R^{ij}_V$, whereas the information from the truncated likelihood $\mathcal{I}^{tr}_{\beta,V}(Y_V)$ scales as $\sum_{ij} p^{ij}_{0,V} R^{ij}_V$, or the expected number of nonzero outcomes in the sample Y_V. Thus, we can establish the following statement about the asymptotic fraction of ignored information, and thus lost efficiency, when using the truncated estimator $\hat{\beta}^{tr}_V$ over the oracle estimator $\hat{\beta}^{or}_V$ in this context.

Theorem 4 (Efficiency loss of the truncated estimator.)**.** *Assume the following conditions hold for all increasing sequences of actor-sets* (V_n) *from* $\mathbb{V}$*.*

(E1) *For all* V_n, $\left(\mathbb{E}_0(\nabla^2_\beta\, l^{ij}_{\beta_0,V}(Y^{ij}_V) \mid A_{ij}=1) - \nabla^2_\beta \log p^{ij}_{\beta_0,V}\right) > C_{tr}$ *for some constant positive definite* C_{tr} *for all* $ij \in R_{V_n}$ *.*

(E2) $\frac{\sum_{ij} R^{ij}_{V_n} p^{ij}_{0,V}}{\sum_{ij} R^{ij}_{V_n}} \to c_{size}$ *for some constant scalar* $c_{size} > 0$*.*

(E3) *The model family* $\mathcal{P}_{\Theta,\mathbb{V}}$ *is specified such that for all* ij *in all* V_n, $\mathbb{E}_0\,\nabla^2_\beta \log p^{ij}_{\beta_0,V_n}$ *and* $\mathbb{E}_0\,\nabla^2_\beta \log(1-p^{ij}_{\beta_0,V_n})$ *are both bounded from above by some finite constant positive definite matrix* C_{bin}*.*

Then the truncated and oracle estimators accumulate information at the same rate but differ by a constant factor. In particular, $\lim_{n\to\infty} \mathcal{I}^{tr}_{\beta_0,V_n}\left(\mathcal{I}_{\beta_0,V_n}\right)^{-1} \geq \left(I + (c_{size}C_{tr})^{-1}C_{bin}\right)^{-1} > 0$*.*

Under conditions (E1)–(E3) the result is straightforward. (*E*1) requires that each dyad provide some information under the truncated likelihood if it generates a nonzero outcome. (*E*2) requires that the expected number of nonzero outcomes grow proportionally to the number of relationships in R_{V_n}. (*E*3) requires that the information provided by the binary distribution $\mathbb{P}_{V_n}(A^{ij}_V \mid R^{ij}_V)$ not be too large. These conditions ensure that the information ignored by the truncated likelihood – that is, the sample size lost to observed 0's in Y_V, and the identification lost by not considering the binary model $\mathbb{P}_V(A_V \mid R_V = 1)$ – not dominate the total information available to the oracle likelihood.

Theorem 4 establishes a worst case scenario for efficiency loss from the truncated estimator. In real data anlytic contexts, the sparsity of a social process only presents difficulties if the relationship graph R_V is not known. Thus, in cases where an investigator would have reason to deploy the truncated estimator $\hat{\beta}_V^{tr}$, the relative efficiency of the truncated estimator with respect to a full-likelihood alternative that sums over a distribution for R_V would be strictly better than the limit established in Theorem 4. Taken together, the instability of full-likelihood estimators shown in Corollary 1, the stability of the truncated estimator shown in Theorem 3, and the constant relative efficiency bound of the truncated estimator shown in Theorem 4, make a compelling case for the robustness-efficiency tradeoff made by the truncated estimator in superpopulation investigations.

5.4 Other properties of the truncated estimator

5.4.1 Single-sample properties

Although the focus of this paper is the superpopulation stability of estimators, an estimator is only useful for a superpopulation inquirty if the property that it measures also characterizes individual samples – that is, a superpopulation estimator must still have good single-sample properties. Several parts of the statistical literature are relevant to establishing single-sample properties of the truncated estimator, including the partial-likelihood literature [10, 11, 38], the conditional likelihood literature [23, 2, 15], and more specific discussion of truncated data models, e.g., [14].

5.4.2 Computational properties

Computation of the truncated estimatior is highly efficient as computation of the likelihood in Equation 16 only requires the nonzero outcomes in Y_V as opposed to the full set of $\binom{|V|}{2}$ outcomes required by full likelihood methods. For sparse social processes, this implies that the computational cost of the truncated estimator grows at a slower rate than the computational cost of a full-likelihood estimator as we analyze larger and larger social network data – in fact, the ratio of computational cost rates here is exactly equal to the sparsity rate $\epsilon_0(n)$ of the social process Y_V. This make the truncated estimator a practical tool for analysis of modern massive social network data.

6 Simulated and Real Data Examples

In this section, we make the arguments of the paper concrete with real and simulated data. The examples here are meant to replicate aspects of the data analysis project that was the motivation for this work in the setting first described in Section 1.1. Originally, the goal of the project was the perform a comparative analysis of inventor collaboration networks across time periods and regions of the United States. Because the outcomes in Y_V were point process valued, we chose the counting process regression model described by Perry and Wolfe in [28], which had strong theoretical support for use in single-sample investigations. However, because the model is a member of the CID class described in Section 3.3, the estimates from this project showed strong signs of population instability that one would expect from a violation of Criterion 1.

6.1 Simulated counting process examples

We begin with simulated data. In this subsection, we first demonstrate the moving target phenomen from Theorem 2 under sparsity misspecification, by showing 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 applicability of the asymptotic results in Section 5.3 to finite sample data anlysis problems.

6.1.1 Model specification

According to the counting process regression model of Perry and Wolfe [28], we represent the pairwise outcomes from from a social process as counting processes $Y_V^{ij}(\cdot)$ with instantaneous hazard given by a GLM specification:

$$\log \lambda_V^{ij}(t) = \beta' X_V^{ij}(t). \tag{26}$$

In this case, $X_V^{ij}(t)$ 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_V, this model yields the log-likelihood for β:

$$L_{\beta,V}(Y_V) = \sum_{ij \in R_V} \left(- \int_0^T \lambda_V^{ij}(s \mid \mathcal{F}_s) ds \right) + \sum_{k=1}^{Y_V^{ij}(T)} \log \lambda_V^{ij} \left(t_{ij}^{(k)} \mid \mathcal{F}_{t_{ij}^{(k)}} \right), \tag{27}$$

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

$$\begin{aligned} L_{\beta,V}^{tr}(Y_V) = \sum_{ij : A_V^{ij}=1} & \left(- \int_0^T \lambda_V^{ij}(s \mid \mathcal{F}_s) ds \right) + \sum_{k=1}^{Y_{ij}(T)} \log \lambda_V^{ij} \left(t_{ij}^{(k)} \mid \mathcal{F}_{t_{ij}^{(k)}} \right), \\ & - \log \left(1 - \exp \left(- \int_0^T \lambda_V^{ij}(s \mid \mathcal{F}_s^0) ds \right) \right) \end{aligned} \tag{28}$$

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.

Recall that the patent database included inventor-specific information such as the zipcode of their residence or the firm that they worked for at the time of the patent application (called an "assignee"). In this simulation, we assign each actor a zipcode and assignee. 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 a CIR model.

We make the simulated CIR model sparse by introducing an ordering on all of the vertices in the actor population $\mathbb{V}$, and assuming that for any actor pair ij, the baseline probability of having a relationship is decreasing in the the population index of actor i. We denote dependence on this global population index by subscripting with the population $\mathbb{V}$.

The formal specification of the data simulation process is as follows:

$$\begin{aligned} R_V^{ij} \mid X_V^{ij} &\sim \text{Bin}(\rho_V^{ij}) \\ \text{logit}\, \rho_V^{ij} &\equiv \gamma_0 \,\text{logit}(\alpha_{\mathbb{V}}(i)) + \gamma_1 \cdot \text{Zip}_V^{ij} + \gamma_2 \cdot \text{Asg}_V^{ij} \\ Y_V^{ij}(t) \mid R_V^{ij}, X_V^{ij}, \mathcal{F}(t) &\sim \begin{cases} CP(\lambda_V^{ij}(t)) & \text{if } R_V^{ij} = 1 \\ 0(t) & \text{if } R_V^{ij} = 0 \end{cases} \\ \log \lambda_V^{ij}(t) &\equiv \beta_0 + \beta_1 \cdot \text{Zip}_V^{ij} + \beta_2 \cdot \text{Asg}_V^{ij} + \beta_3 \cdot \text{prev}_{ij}(t). \end{aligned} \tag{29}$$

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 "prev" is an indicator for previous collaboration, i.e., $Y_{ij}(t) > 0$. γ is a vector of relationship process coefficients, while $\alpha_{\mathbb{V}}(i)$ is a function of i that approaches 0 as the actor-population index $i \to \infty$, and controls the sparsity of the generating process by making the relationship graph ever sparser as individuals with higher actor-population indices are included in the sample.

Both γ and $\alpha_{\mathbb{V}}(i)$ 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 using full-likelihood estimator that make various assumptions about the generating process, and thus the sparsity rate, of $R_{\mathbb{V}}$ and the truncated likelihood estimator $\hat{\beta}_V^{tr}$. For each of the competing estimators, we have a correctly specified $\mathbb{P}_{\beta,V}(Y_V \mid R_V, X_V)$, or the outcome process given the relationship graph and covariates.

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.

6.1.2 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 actor i in the popultion $\mathbb{V}$ 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_V is fully connected for all V, corresponding to the popular GLM approach of vectorizing the data Y_V and treating each pair ij as conditionally independent given the covariates X_V. For each subsample generated by the sequence above, we compute the effecive estimand of the misspecified model in addition to the GLM MLE $\hat{\beta}_V$ and MTLE $\hat{\beta}_V^{tr}$ 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

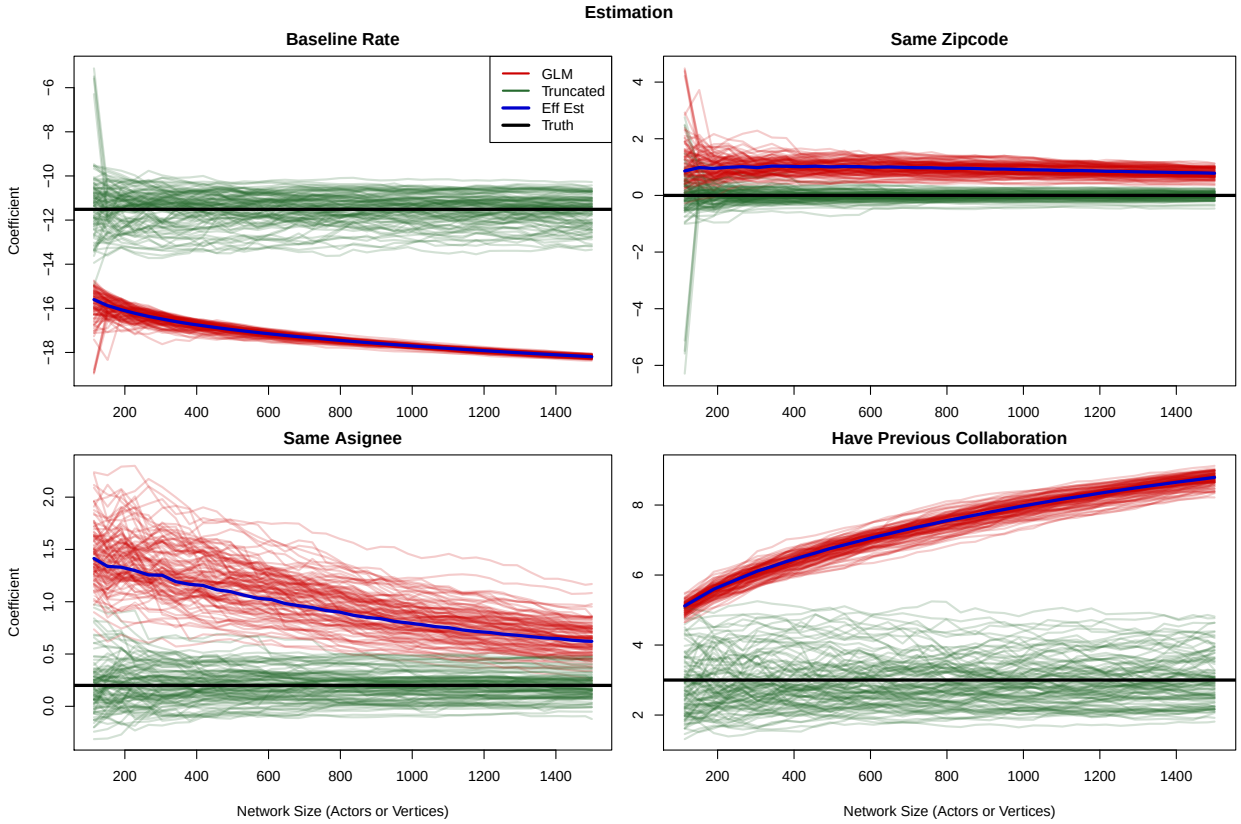

Figure 4: Plots of the sampling distribution of sequences the MLE $\hat{\beta}_V$ computed from the sparsity misspecified counting process model (red), and the MTLE $\hat{\beta}_V^{tr}$ computed from the truncated model (green) from samples of differing size. We also plot the effective estimand for the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

results from the discussion above. The effective estimands of the misspecified models show the moving target behavior as they vary with $|V|$, 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 $P(R_V^{ij} \mid X_V^{ij})$ 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.

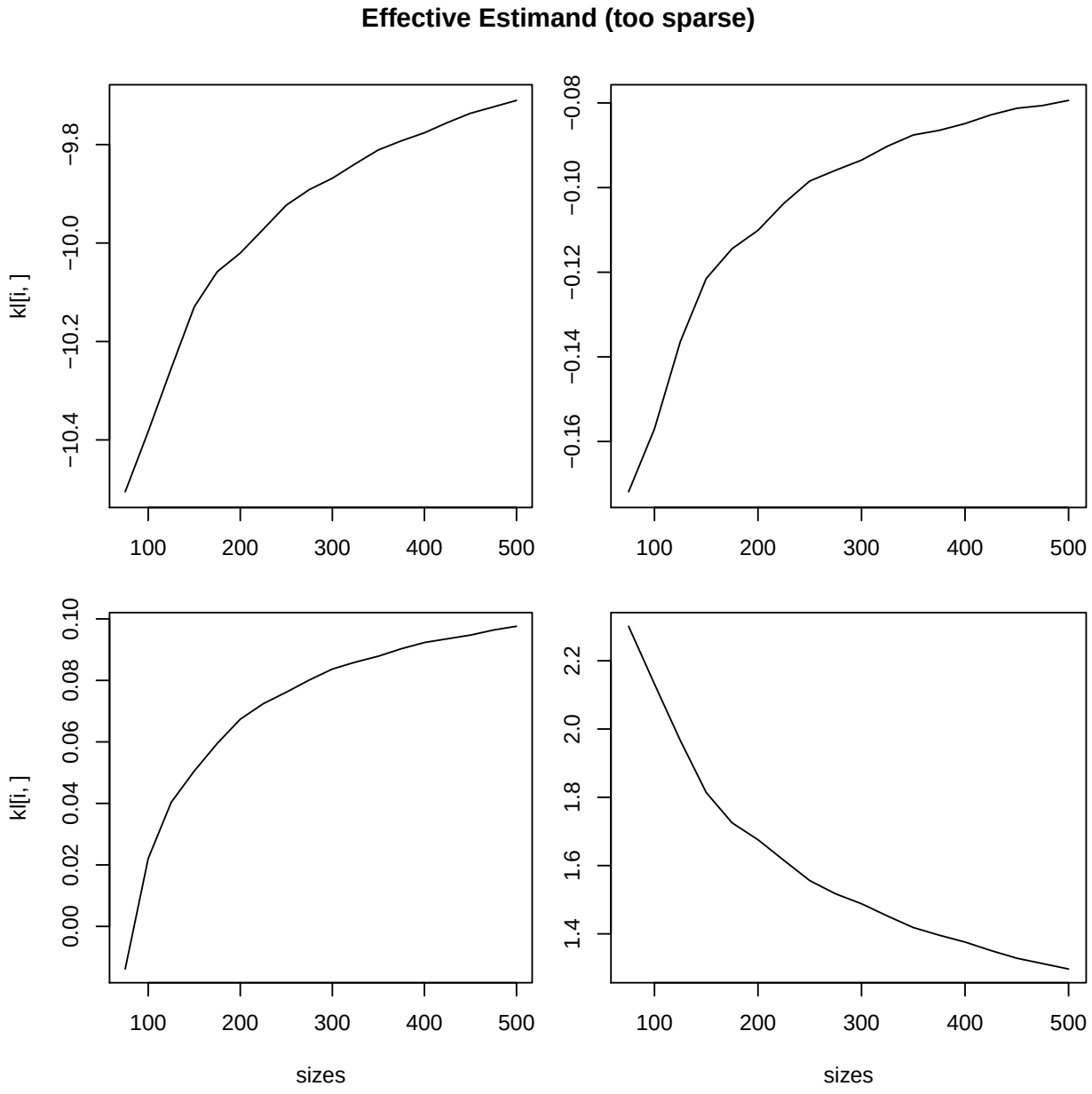

Figure 5: Plots of the effective estimand when the proposed model family is too sparse. Here the true logistic model for R_V^{ij} has intercept function $\log(i)/i$ and the inferential model assumes the intercept goes at $1/i$.

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.

6.1.3 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 "prev" 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, prev).

Efficiency. Following Section 5.3, 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

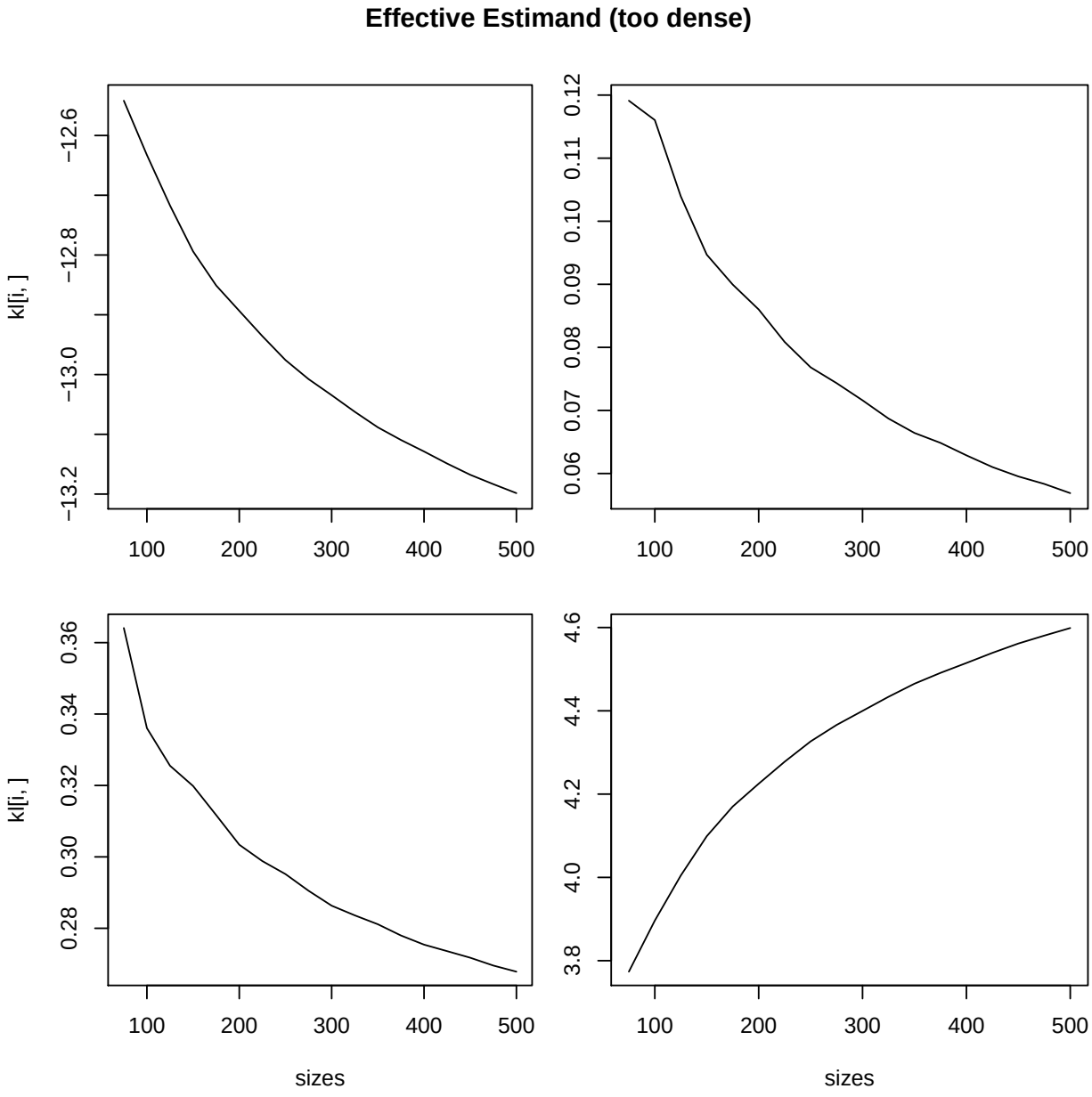

Figure 6: Plots of the effective estimand when the proposed model family is too dense. Here the true logistic model for R_{ij} has intercept function $1/i$ and the inferential model assumes the intercept goes at $\log(i)/i$.

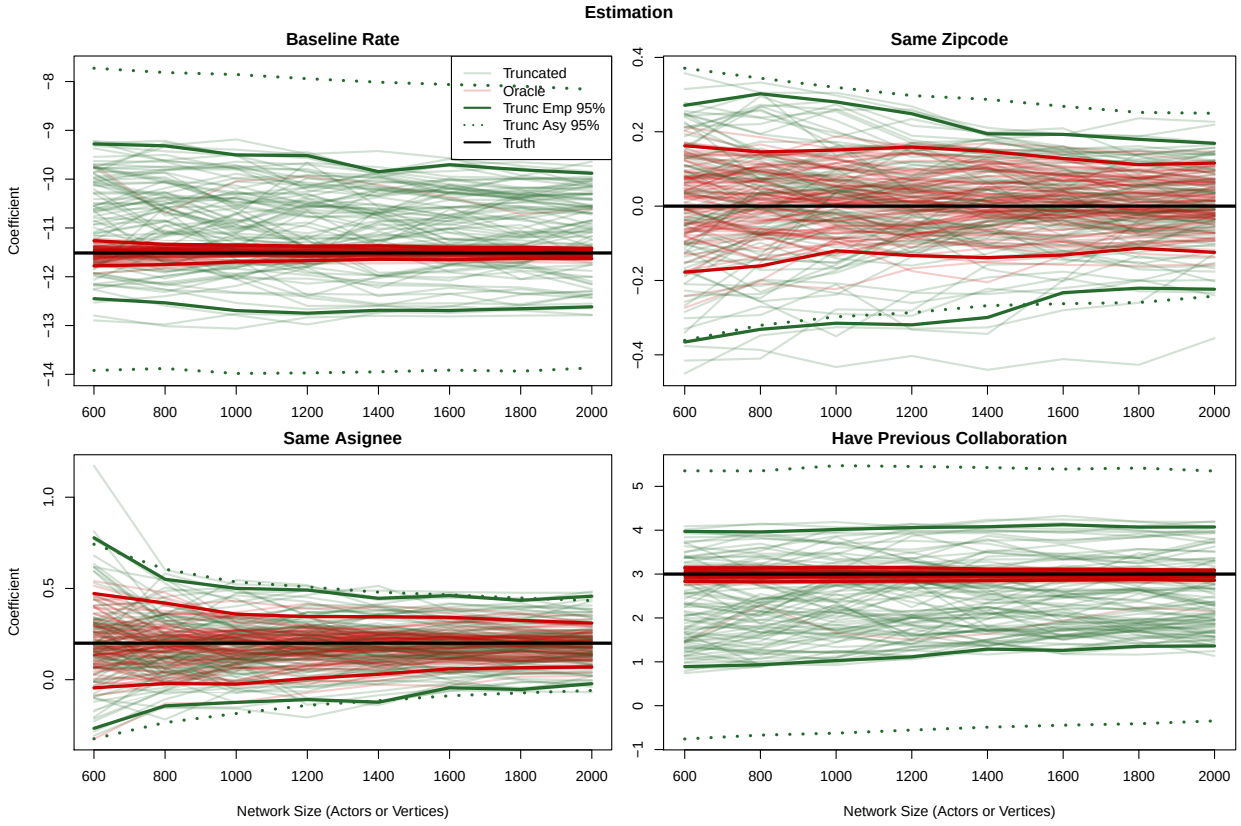

Figure 7: Plots of sampling distribution of sequences $\hat{\beta}_n$ computed from the truncated model (green) and the oracle model (red). The oracle model has full knowledge of the risk set R and is computed using the full likelihood on this subset of dyads.

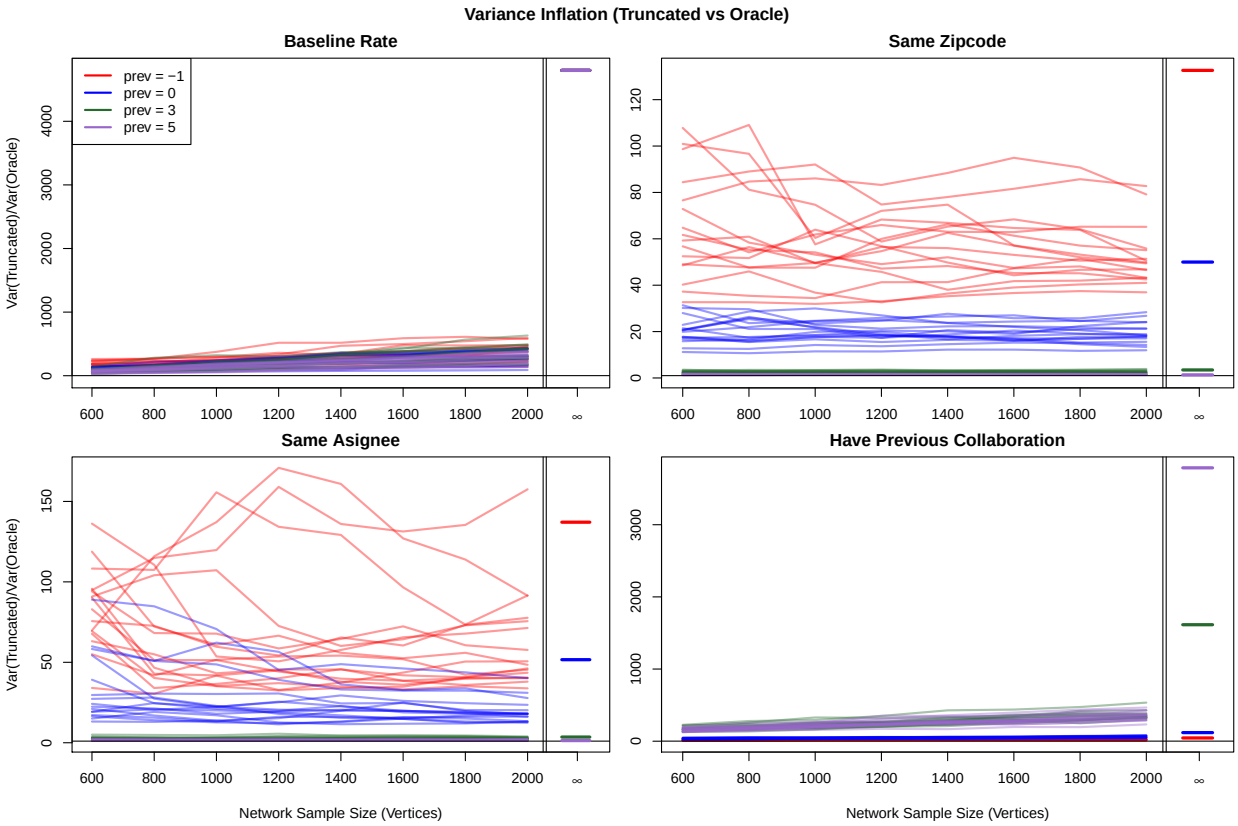

Figure 8: Variance inflation factors resulting from the comparison of the truncated estimator's sampling distribution to the oracle estimator. Lines are colored by the value of the "previous collaboration" coefficient, which shows the most influence on the efficiency of the non-intercept coefficients. 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 5.3. 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_V 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 $\mathcal{I}^{tr}_{\hat{\beta}^{tr}_V, V}$. 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

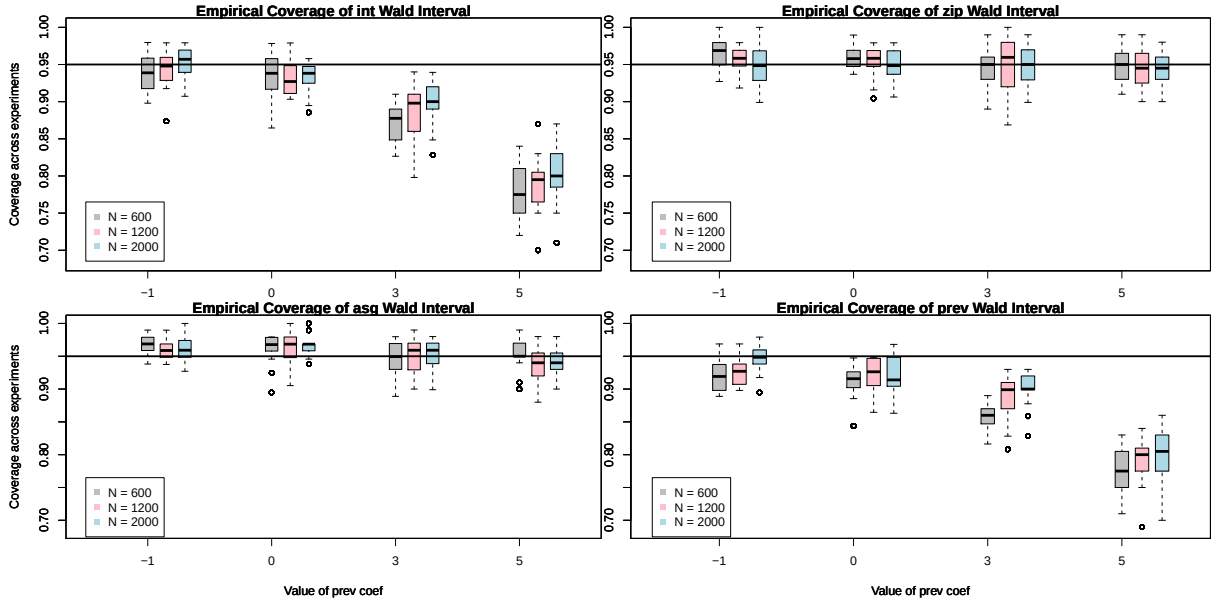

Figure 9: Coverage of 95% asymptotic confidence intervals computed using a full factorial design. Coverage was mostly sensitive to the level of the "prev" coefficient, which controls how much interaction frequency increases when a previous interaction has occurred. The truncation mechanism drops a portion 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.

6.2 Real data analysis

Finally, we return to the data analysis first presented in Section 1.1. The parameter estimates in Figure 1 show the results of fitting the [28] point process model (modified to include an intercept term) in a number of different metropolitan areas around the United States, using the real analogues of the "Asg" (indicator for i and j work for the same firm) and "prev" (indicator for i and j have collaborated before) covariates. As discussed in Section 1.1, the estimates from the naive GLM show a strong dependence on the size of the sample that is confounded with any true differences between regions.

The results from using the truncated estimator $\hat{\beta}_V^{tr}$ obtained by maximizing Equation 28 are shown on the right of the figure. These estimates appear on a realistic scale, where the relative rates of collaborations on patents in different regions do not exceed 1 on the log scale. The estimates also provide believeale standard errors. Most importantly, the estimates show no strong systematic dependence on the size of the sample. Put simply, the truncated estimator appears to be measuring an aspect of the patent collaboration process that is actually comparable across regions, and, as opposed to the highly sample-size-dependent, overconfident estimates on the left side of the figure, invite interpretation by social scientists.

7 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 [27], 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 Qiuyi Han, 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 Bickel, David Choi, Xiangyu Chang, and Hai Zhang. Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *Annals of Statistics*, 41(4):1922–1943, 2013.

[4] 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, dec 2009.

[5] Béla Bollobás and Oliver Riordan. Sparse graphs: Metrics and random models. *Random Structures & Algorithms*, 39(1):1–38, aug 2011.

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

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

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

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

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

[11] D R Cox. Partial likelihood. *Biometrika*, 62(2):269–276, aug 1975.

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

[13] 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.

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

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

[16] 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.

[17] 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, mar 2007.

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

[19] Peter D Hoff, Adrian E Raftery, and Mark S Handcock. Latent Space Approaches to Social Network Analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, dec 2002.

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

[21] 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.

[22] 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.

[23] 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.

[24] 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.

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

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

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

[28] 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.

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

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

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

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

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

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

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

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

[37] 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.

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

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

A Proof of Lemma 1

Given Equation 9 and Equation 10, bounding the probability of events in terms of the the log proportional difference between observed and expected within- and between-firm collaboration counts, is especially convenient. We derive the probability bound for $\hat{\theta}_0$ explicitly, and the same formulation can be followed for $\hat{\theta}_1$.

$$\begin{aligned}
\mathbb{P}(|\hat{\theta}_0 - \bar{\theta}_0| \leq \log(1+\delta)) &\geq \mathbb{P}\left((1-\delta) \leq \left(\frac{\sum Y_i(1-X_i)}{\sum \mathbb{E}_0 Y_i(1-X_i)}\right) \leq (1+\delta)\right) \\
&= \mathbb{P}\left(\left|\sum Y_i(1-X_i) - \sum \mathbb{E}_0 Y_i(1-X_i)\right| \leq \delta \sum \mathbb{E}_0 Y_i(1-X_i)\right) \\
&\geq 1 - \frac{\mathbb{V}\text{ar}_0\left(\sum Y_i(1-X_i)\right)}{\delta^2\left(\mathbb{E}_0 \sum Y_i(1-X_i)\right)^2} \\
&\geq 1 - \frac{d}{\delta^2 \mathbb{E}_0 \sum Y_i(1-X_i)},
\end{aligned}$$

where the penultimate step is an application of the Chebyshev inequality, and the final step applies assumption (B3) from Section 4.1.

For the other coefficient, we bound a similar deviation for the quantity

$$\widehat{(\theta_0 + \theta_1)} = \log\left(\frac{\sum \mathbb{E}_0 Y_i X_i}{\sum X_i}\right)$$

separately. This quantity has a rate related to the expected number of within-firm dyads:

$$\mathbb{P}(|(\widehat{(\theta_0 + \theta_1)}) - (\bar{\theta}_0 + \bar{\theta}_1)| \leq \log(1+\delta)) \leq 1 - \frac{d}{\delta^2 \mathbb{E}_0 \sum Y_i X_i}$$

Combining these bounds, we obtain a deviation bound for $|\hat{\theta}_1 - \bar{\theta}_1|$

$$\begin{aligned}
\mathbb{P}(|\hat{\theta}_1 - \bar{\theta}_1| \leq \delta) \geq\ & 1 - \mathbb{P}(|\hat{\theta}_0 - \bar{\theta}_0| \geq \delta/2) \\
& - \mathbb{P}(|(\hat{\theta}_0 + \hat{\theta}_1) - (\bar{\theta}_0 + \bar{\theta}_1)| \geq \delta/2) \\
\geq\ & 1 - \frac{4C_1}{\delta^2 \mathbb{E}_0 \sum Y_i(1-X_i)} - \frac{4C_2}{\delta^2 \mathbb{E}_0 \sum Y_i X_i}
\end{aligned}$$

B Limiting variance inflation calculation from Section 6.1.3

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:

$$\mathcal{I}_n^s(\beta) = \sum_{ij \in \mathcal{R}_n} \mathbb{E}\left[t_{ij}^{(1)}\right] w_{ij}^{pre,s} X_{ij}^{pre} X_{ij}^{pre\top} + \left(T - \mathbb{E}\left[t_{ij}^{(1)}\right]\right) w_{ij}^{post} X_{ij}^{post} X_{ij}^{post\top} \tag{30}$$

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 30 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 $\sum_c N_c = \sum_{ij} R_{ij}$,

$$\mathcal{I}_n^s(\beta) = \sum_c N_c \left(\mathbb{E}\left[t_c^{(1)}\right] w_c^{pre,s} X_c^{pre} X_c^{pre\top} + \left(T - \mathbb{E}\left[t_c^{(1)}\right]\right) w_c^{post} X_c^{post} X_c^{post\top} \right). \tag{31}$$

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 31, 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 29, 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

$$VI_k(\beta) = \lim_{n\to\infty} \frac{C_n^{tr}(k,k)}{C_n^{full}(k,k)} \frac{\det(\mathcal{I}_n^{full})}{\det(\mathcal{I}_n^{tr})}. \tag{32}$$

Beginning with the second factor of Equation 32, 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

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

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:

$$C_n^s(1,1) = (i_{n,22}^s i_{n,33}^s - (i_{n,23}^s)^2) i_{n,44}^s + o(N_1) \tag{34}$$
$$C_n^s(2,2) = (i_{n,11}^s i_{n,44}^s - (i_{n,14}^s)^2) i_{n,33}^s + o(N_1^2) \tag{35}$$
$$C_n^s(3,3) = (i_{n,11}^s i_{n,44}^s - (i_{n,14}^s)^2) i_{n,22}^s + o(N_1^2) \tag{36}$$
$$C_n^s(4,4) = (i_{n,22}^s i_{n,33}^s - (i_{n,23}^s)^2) i_{n,11}^s + o(N_1). \tag{37}$$

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

$$z_c^{pre,s} = \mathbb{E}\left[t_c^{(1)}\right] w_c^{pre,s} \qquad\qquad z_c^{post} = \left(T - \mathbb{E}\left[t_c^{(1)}\right]\right) w_c^{post}. \tag{38}$$

Evaluating Equation 31, the relevant elements of $\mathcal{I}_n^s(\beta)$ have the form

$$i_{n,11}^s = \sum_c N_c \left(z_c^{pre,s} + z_c^{post}\right) \tag{39}$$

$$i_{n,44}^s = i_{n,14}^s = \sum_c N_c z_c^{post} \tag{40}$$

$$i_{n,22}^s = N_3 \left(z_3^{pre,s} + z_3^{post}\right) + N_4 \left(z_4^{pre,s} + z_4^{post}\right) \tag{41}$$

$$i_{n,33}^s = N_2 \left(z_2^{pre,s} + z_2^{post}\right) + N_4 \left(z_4^{pre,s} + z_4^{post}\right) \tag{42}$$

$$i_{n,23}^s = N_4 \left(z_4^{pre,s} + z_4^{post}\right). \tag{43}$$

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

$$VI_1(\beta) = \frac{\sum_c N_c z_c^{pre,full}}{\sum_c N_c z_c^{pre,tr}} \tag{44}$$

$$VI_2(\beta) = \frac{N_3 \left(z_3^{pre,tr} + z_3^{post}\right) + N_4 \left(z_4^{pre,tr} + z_4^{post}\right)}{N_3 \left(z_3^{pre,full} + z_3^{post}\right) + N_4 \left(z_4^{pre,full} + z_4^{post}\right)} \frac{K^{full}}{K^{tr}} \tag{45}$$

$$VI_3(\beta) = \frac{N_2 \left(z_2^{pre,tr} + z_2^{post}\right) + N_4 \left(z_4^{pre,tr} + z_4^{post}\right)}{N_2 \left(z_2^{pre,full} + z_2^{post}\right) + N_4 \left(z_4^{pre,full} + z_4^{post}\right)} \frac{K^{full}}{K^{tr}} \tag{46}$$

$$VI_4(\beta) = \frac{\sum_c N_c \left(z_c^{pre,tr} + z_c^{post}\right)}{\sum_c N_c \left(z_c^{pre,full} + z_c^{post}\right)} \frac{\sum_c N_c z_c^{pre,full}}{\sum_c N_c z_c^{pre,tr}} \tag{47}$$

where

$$\begin{aligned} K^s = {} & N_2 \left(z_2^{pre,s} + z_2^{post}\right) N_3 \left(z_3^{pre,s} + z_3^{post}\right) + \\ & N_2 \left(z_2^{pre,s} + z_2^{post}\right) N_4 \left(z_4^{pre,s} + z_4^{post}\right) + \\ & N_3 \left(z_3^{pre,s} + z_3^{post}\right) N_4 \left(z_4^{pre,s} + z_4^{post}\right) \end{aligned} \tag{48}$$

To fix constants and ensure identification in the limit for the example in Section 6.1.3, 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

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

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 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