Towards inference using dependent observations from a network

Anonymous Author 1
Unknown Institution 1

Anonymous Author 2
Unknown Institution 2

Anonymous Author 3
Unknown Institution 3

Abstract

Interest in and and opportunities for studying outcomes in networks are rapidly proliferating, but methods for analyzing data on outcomes sampled from a single network remain underdeveloped. When the relationships captured by network ties inform the dependence among observations sampled from the network, network data will exhibit complex forms of dependence. Network ties tend to be dense and erratic, and the resulting topology and dependence structure are very different from spatial or Euclidean topology and dependence. While methods abound in the statistics literature for dealing with spatial-temporal dependence, these methods are not immediately applicable to the network setting. In this paper we describe the challenges for statistical inference using outcomes sampled from a single network observed at one or only a few time points, and point towards some possible solutions.

1 Background

A network is a collection of units, or *nodes*, and the ties, or *edges*, between them. The presence of a tie between two nodes indicates that the nodes share some kind of a relationship; what types of relationships are encoded by network ties depends on the context. Some types of relationships are mutual, for example familial relatedness and shared place of work; others may go in only one direction. For simplicity we will assume throughout that edges are undirected (mutual) and indistinguishable from one another, but it is straightforward to extend the concepts and methods we discuss to handle networks with directed edges, weighted edges, or multiple categories of edges. A node whose characteristics we wish to explain or model is called an *ego*; nodes that share ties with the ego are its *alters*. If an ego's outcome may be affected by his contacts' outcomes then we say that the outcome exhibits *induction* or *contagion*.

Distance in a network is usually defined as the geodesic distance: the length of the shortest path of consecutive edges connecting two nodes, but other definitions are possible, for example taking into account the number of paths or the average path length between two nodes. In what follows we use the geodesic distance metric. In practice, the choice of distance metric is not arbitrary: successful inference using network data requires use of the distance metric(s) that truly describes the dependence structure of the network under analysis.

A small body of literature attempts to perform causal and statistical inference on data sampled from a single social network, with mixed success. There have been a number of high profile articles recently that use standard methods like generalized linear models (GLM) and generalized estimating equations (GEE) to attempt to infer causal relationships from network data (e.g. Christakis and Fowler, 2007, 2008, 2010), and a strong backlash from the statistical community (Cohen-Cole and Fletcher, 2008; Lyons, 2011; Shalizi and Thomas, 2011). These statistical models are not equipped to deal with network dependence and are rarely appropriate for estimating effects using network data. In some settings it may be possible to use them to test for the presence of network dependence, but some properties of such tests are unknown (VanderWeele et al., 2012; Shalizi, 2012).

Spatial autoregressive (SAR) models have been applied to the study of peer effects and induction in network settings (e.g. Goetzke, 2008; Lee, 2004; Lin, 2005; O'Malley and Marsden, 2008). The shortcoming of these models stems from the fact that, because the endogenous and exogenous variables are measured at the same time, they parameterize an equilibrium state rather than causal relationships. Few data generating processes give rise to true equilibrium states (Besag, 1974; Lauritzen and Richardson, 2002; Thomas, 2013);

Preliminary work. Under review by AISTATS 2014. Do not distribute.

therefore SAR models may often be misspecified or uninformative about causal relationships.

A hallmark of most of the work to date on outcomes sampled from a network is that it uses models, like GEE, GLM, and SAR models, that were developed for very different settings. Very recently, researchers have begun to develop methods designed specifically for the network setting. Exciting and innovative work by van der Laan (2012) harnesses independence assumptions that require observing the evolution of the network and outcomes over time. In many settings, however, we will only get a snap shot of the network, or we may observe it at multiple time points but not enough to capture the full evolution of the network. Many methods for interference, which is when one unit's exposure may affect another unit's outcome, are highly relevant to the analysis of network data (Aronow and Samii, 2013; Graham et al., 2010; Halloran and Struchiner, 1995; Halloran and Hudgens, 2011; Hong and Raudenbush, 2006, 2008; Hudgens and Halloran, 2008; Rosenbaum, 2007; Rubin, 1990; Sobel, 2006; Tchetgen Tchetgen and VanderWeele, 2012; VanderWeele, 2010; VanderWeele and Tchetgen Tchetgen, 2011a,b). However, the inferential methods developed in this context generally require observing multiple independent groups of units, which corresponds to observing multiple independent networks, or else they allow for testing but not for estimation of causal effects (Rosenbaum, 2007; Bowers et al., 2013).

Developing methods for statistical inference that can account for the dependence among observations sampled from a single network is a crucial step that has yet to be successfully completed.

2 Motivating example

Suppose that students attending the residential Faber College are weighed at the start and close of the school year, and we want to estimate and perform inference about the mean change in weight across all students. Students' weight changes are likely to covary, with a dependence structure that is informed by the social network according to which they live and socialize with each other, dine together, and influence one another.

We would like to be able to extrapolate from the Faber students' weight changes to weight changes in similar college populations across different colleges or across different years. Assume that the student body we observe at Faber College is representative of these other student populations, that is, that the true underlying mean weight change for the observed sample of Faber students is the same as the true underlying mean in the other college populations to which we want to extrapolate. Then one way to determine whether we are warranted in extrapolating from Faber students to the other similar groups of students is to calculate a confidence interval for the true mean weight change, based on a model of asymptotic growth of the sample. For example, if the sample is large enough that a central limit theorem approximately holds for the sample mean, then a Gaussian confidence interval around the sample mean is approximately valid. Under the assumption of the same true underlying mean, our confidence that this interval covers the true mean weight change among Faber College students is the same as our confidence that it covers the true mean weight change among students at a different college or in a different year. As in many settings for statistical inference, asymptotics are appropriate not because we care about an infinite population but because they shed light on finite samples.

In this paper we focus on the following very simple setting: we sample nodes from a single network at a single time point and observe an outcome, Y_i, at each node i. Our goal is valid frequentist inference about $E[Y] = \mu$. This expectation is taken with respect to a true underlying data-generating process; how to define the data generating process and the population over which it operates is not straightforward in the network setting. See Section 5.1 for further discussion of this issue.

The central question that we address in this paper is how to perform valid frequentist inference using a sample mean $\bar{Y}$ of dependent observations $\mathbf{Y} = (Y_1, ..., Y_n)$, where the dependence among observations is determined or informed by network structure. Our interest is in the true sample mean μ, for which $\bar{Y}$ is an unbiased estimator. If $\bar{Y}$ is consistent, then valid inference requires a consistent estimator of its standard error, or, equivalently, of the asymptotic variance $Avar(\bar{Y})$. If a central limit theorem holds for $\bar{Y}$ then valid inference can make use of a Gaussian confidence interval; otherwise it may be possible to construct exact intervals using moment inequalities or resampling methods. We will briefly discuss conditions under which a central limit theorem holds for network dependent data but will focus mainly on the problem of finding a consistent estimator $\widehat{\sigma}^2_{\bar{Y}}$ for $Avar(\bar{Y})$ in the general case.

In order not to presuppose stationarity of the distribution of $\mathbf{Y}$ or asymptotic normality of the distribution of $\bar{Y}$, define $Avar(\bar{Y})$ as follows: If $\bar{Y}$ converges in distribution to a random variable V with mean 0 and variance 1, that is if as $n \to \infty$

$$\frac{\sum Y_i - E\left[\sum Y_i\right]}{\sqrt{var\left(\sum Y_i\right)}} = \frac{\sqrt{n}\left(\bar{Y} - E\left[\bar{Y}\right]\right)}{\sqrt{var\left(\sum Y_i\right)/n}} \xrightarrow{d} V,$$

then $Avar(\bar{Y}) \equiv var\left(\sum Y_i\right)/n$ is the variance of the asymptotic distribution of $\sqrt{n}\left(\bar{Y} - E\left[\bar{Y}\right]\right)$. It would be natural to consider estimating $Avar(\bar{Y})$ with the sample estimator $\frac{1}{n}\sum_{i,j=1}^{n}\left(Y_i - \bar{Y}\right)\left(Y_j - \bar{Y}\right)$, however this is degenerate: $\sum_{i,j=1}^{n}\left(Y_i - \bar{Y}\right)\left(Y_j - \bar{Y}\right) = \left[\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)\right]^2 = \left[\sum_{i=1}^{n} Y_i - n\bar{Y}\right]^2 = 0$. One solution is to restrict the range of dependence so that the covariance terms only have to be calculated for some pairs (i, j), in which case a non-degenerate version of the estimator is available. Another is to use subsampling or resampling methods to approximate the sampling distribution of $\bar{Y}$ (see Sections 5.3 and 5.4). A third solution is to model the variance-covariance matrix of the entire vector $\mathbf{Y}$, from which it is easy to calculate the variance of $\bar{Y}$ (see Section 5.5). This setting is very simple, but it extends naturally to more sophisticated inferential problems, in particular to estimating the asymptotic variance of any M-estimator.

Throughout we make a series of simplifying assumptions, some of which can easily be relaxed. We assume that observations are sampled at a single point in time. If observations are sampled at multiple time points but the network topology remains constant in time, then methods developed for time series data could be used to account for dependence in time, in conjunction with the methods we discuss below for dealing with network dependence at each time point. Roughly, in this scenario we would have dependence along two dimensions (time and network distance) that are orthogonal to one another and can therefore be treated separately. However, if network topology changes between time periods then the two types of dependence may not be orthogonal and more sophisticated methods would be required. Other assumptions that we make throughout are that the network structure for the sample is entirely known and that ties between units are undirected and unweighted. It may be possible to relax the former assumption using methods to handle missing data in networks (e.g. Gile and Handcock 2006; Huisman 2009; Kossinets 2006; Stomakhin et al. 2011). The assumption that edges are binary can be relaxed easily.

3 Statistical inference for data with spatial dependence

The literature on statistics for dependent data is vast and multifaceted, but very little has been written on the dependence that arrises when observations are sampled from a single network. Most of the literature on dependent random variables assumes that the domain from which observations are sampled (e.g. time or geographic space) has an underlying Euclidean topology, and this renders most results at least prima facie incompatible with network dependence. There are two types of asymptotic frameworks in the dependence literature: increasing domain asymptotics, in which the minimum distance between any two observations is bounded away from 0 and new observations are added at at the boundaries of the sample domain, and infill asymptotics, in which the domain is bounded and the distance between observations converges to 0 as sample size grows (Cressie, 1993). Most existing theory and literature is concerned with increasing domain asymptotics; less has been written about infill asymptotics. In general results for the latter framework are weaker and they tend to rely more heavily on strong assumptions about the topology of the sample domain and on the data generating distribution (Anselin, 2001; Zhang and Zimmerman, 2005). These assumptions, e.g. that observations are sampled from a regular lattice in $\mathbb{R}^d$ or that observations are generated by a Gaussian random field (Wu et al., 2012; Lim and Stein, 2008), may be inappropriate for many network settings. We rely primarily on the increasing domain framework throughout.

In a slight abuse of terminology and for convenience, we refer to the literature that assumes an underlying Euclidean space and increasing domain asymptotics as the "Euclidean dependence" literature. The principles behind asymptotic results in the Euclidean dependence literature are simple and intuitive. They rely on a combination of stationarity assumptions, i.e. assumptions that certain features of the data generating process do not depend on an observation's location in the sample domain, and assumptions that bound the nature and the amount of dependence in the data. Most frequently these are mixing assumptions, which describe the decay of the correlation between observations as a function of the distance between them. Sometimes the stronger assumption of m-dependence is made, according to which two observations are independent if they are sampled from locations that are m or more units apart. Intuitively, in order to extract an increasing amount of information from a growing sample of dependent observations, old observations must be predictive of new observations, which is ensured by stationarity assumptions, and the amount of independence in the sample must grow faster than the amount of dependence, which is ensured by mixing conditions or m-dependence.

There is an extensive literature on Euclidean dependence, primarily in the fields of econometrics, spatial statistics, and time series. Recent years have seen an explosion of generalizations: to data sampled from irregularly shaped lattices, to Euclidean space with a large number of dimensions, and to far-ranging ("near-epoch") and complex dependence (e.g. Davidson, 1992; Dedecker et al., 2007; Jenish and Prucha, 2009, 2012). However, this literature is not immediately applicable to the network setting. Roughly, this is due to the difference between Euclidean and network topology. While it is possible to embed a network in $\mathbb{R}^d$, to do so is to allow d to increase at rate n^2, but Euclidean dependence results generally require d to be fixed, which implies that as new observations are sampled at the boundary of a Euclidean domain, the average and maximum pairwise distance between observations increases. Networks, on the other hand, often do not have a clear boundary to which we can add observations in such a way that ensures growth in the sample domain. In a large sample with Euclidean dependence, most observations will be distant from most other observations. This is not necessarily the case in networks. The maximum distance between two nodes can be small even in very large networks, and even if the maximum distance between two nodes is large, there may be many nodes that are close to one another. Even under m-dependence with small m, networks exist in which most observations remain less than m units apart as $n \to \infty$. Therefore, mixing conditions and m-dependence do not necessarily result in more independence than dependence in a large sample from a network. Research indicates that social networks generally have the small-world property (sometimes referred to as the "six degrees of separation" property), meaning that the average distance between two nodes is small (Watts and Strogatz, 1998). Therefore distances in many real-world networks likely grow slowly with sample size. Of course some types of networks, e.g. lattices, embed in $\mathbb{R}^d$ as n grows, but these are generally trivial cases that are not useful for naturally occurring networks like social networks.

4 Sources of network dependence

In the Euclidean dependence literature, dependence is often implicitly assumed to be the result of latent traits that are more similar for observations that are close in Euclidean distance than for observations that are distant. This type of dependence is likely to be present in many network contexts as well. In networks, edges present opportunities to transmit traits or information, and contagion or influence can be an important additional source of dependence.

Latent trait dependence will be present in data sampled from a network whenever observations from nodes that are close to one another are more likely to share unmeasured traits than are observations from distant nodes. In social networks homophily, or the tendency of similar people to befriend one another, is a paradigmatic example of latent trait dependence. Contagion or influence arises when the outcome under study is transmitted from node to node along edges in the network. Dependence due to contagion has known, though possibly unobserved, structures that can sometimes be harnessed to facilitate inference. Time and distance act as information barriers for dependence due to contagion, giving rise to many conditional independencies that can sometimes be used to make network dependence tractable. (See van der Laan, 2012 for a recent example.) If the network under a contagious process is observed at a single time point, then dependence due to contagion is indistinguishable from latent variable dependence. For this reason, the methods that I discuss below are agnostic about the source of dependence.

5 Some paths towards valid statistical inference in networks

5.1 Population growth in networks

There are many complex issues surrounding asymptotic growth of a network, and the idea that an observed network is a finite subsample of an infinite underlying network is not necessarily useful or coherent. (See Shalizi and Rinaldo, 2013; Diaconis and Janson, 2007 for related discussions of consistency and convergence of subgraphs.) When the goal is asymptotics in the service of finite sample inference (see Section 2), it is not necessary to be entirely beholden to meaningful or realistic models of network generation and growth. (For the interested reader, discussion and review of such models can be found in Goldenberg et al., 2010; Newman, 2009.) In what follows, let the process $n \to \infty$ be represented by a sequence of networks of increasing size that manifest the same outcome distribution. Certain key features of network topology must also be preserved in each network in the sequence, for example the distribution of number of ties per person. Precisely which features of network topology must be preserved by asymptotic growth depends on the method to be used to analyze the data and, in particular, on the assumptions on which it relies. In every method that we discuss below, stationarity is required to hold for some aspects of the outcome distribution and network topology, either at the individual level, across local groups of individuals, or across large communities. The simplest way to operationalize asymptotic growth for each method is to add one stationary unit, be it an individual, a small group, or a large community, at each step. This is imprecise and will not always be possible, but we hope it motivates the principles behind asymptotic growth in this setting.

Although it may not be reasonable to expect distances to grow quickly with sample size (see Section 3), a more reasonable requirement is that the diameter of the network, i.e. the maximum distance between two nodes, increases with sample size. This requirement rules out some network configurations, like the hub-and-spoke network in which one node is the hub and every other node is connected only to the central hub, but is consistent with some models of network generation and with some observed social networks.

In Section 5.2 we describe a simple and intuitive central limit theorem for observations sampled from a network. We also discuss the limitations of this CLT and suggest future work that would strengthen the result and increase its applicability. In Sections 5.3 - 5.5 we describe methods for consistently estimating $Avar(\bar{Y})$ using network data, and again discuss the limitations and open problems related to each method.

5.2 Limit theorems and local dependence

Most central limit theorems for Euclidean dependence rely on Bernstein's blocking method for proof. This method divides the sum $\sum_{i=1}^{n} Y_i$ into components corresponding to big blocks of consecutive observations separated from one another by smaller blocks of consecutive observations. The small blocks grow with n quickly enough to ensure the approximate independence of the big blocks, but slowly enough to be asymptotically negligible. Proofs of this nature rely on features that tend to be incompatible with network topology, such as observations sampled from a regular lattice in $\mathbb{R}^d$ and new observations sampled from the boundary of the sample space (Jenish, 2008).

Stein's method (Stein, 1972) is a more topologically flexible tool for proving central limit theorems that has been used to prove central limit theorems in a number of different dependent data settings (Barbour and Chen, 2005). We are unaware of its having been applied to the specific problem of sampling observations from a network, but this setting is analogous to other settings that have been studied through the lens of Stein's method, like dependency graphs and geometric features of random graphs (e.g. Chen and Shao, 2004; Ross, 2011).

Local dependence describes the following dependence structure: there is a dependency neighborhood $\mathcal{M}_i$ around each subject i such that $Y_i \perp Y_j$ for all $j \notin \mathcal{M}_i$. Using Stein's method, CLTs can be proved for locally dependent data; the results obviously depend on restrictions on the sizes of the neighborhoods but they are powerfully flexible in that they require no restrictions on the shape of the neighborhoods. I will informally review Stein's method of central limit theorem proof under local dependence and intuitively motivate assumptions under which it will hold in networks. For more rigorous treatments, see Chen and Shao (2004); Ross (2011).

Stein's method relies on two important results. The first result is that a random variable Z has the standard normal distribution if and only if $E\left[h'(Z) - Xh(X)\right] = 0$ for all absolutely continuous functions $h(\cdot)$ such that $E\left[|h'(Z)|\right] < \infty$. Building on this, the second result bounds the error in approximating the distribution of a random variable W by the distribution of Z. The error is measured by a metric of the form $d_H(W,Z) = sup_{h\in H}\left|E\left[h(W)\right] - E\left[h(Z)\right]\right|$. Stein showed that, for Z standard normal and for a particular class of functions H (functions with Lipchitz constant 1),

$$d_H(W,Z) \leq sup_{h\in H}\left|E\left[h'(W) - Wh(W)\right]\right|. \qquad (1)$$

When W is an appropriately normalized sample mean this gives a direct bound on the error of the central limit theorem. The right-hand side of (1) will depend on n. If it converges to 0 as $n \to \infty$ then the normalized sample mean converges in distribution to a standard normal. The order of the right-hand side gives the rate of convergence, with $n^{-1/2}$ being the best possible.

Suppose that $E[Y_i] = 0$. (This can be assumed without loss of generality if $\mathbf{Y}$ is mean stationary; if $\mathbf{Y}$ is not mean stationary then similar but wider bounds can be derived which do not rely on Y_i having mean 0.) Let $\bar{Y}_n = \frac{\sum_{i=1}^{n} Y_i}{var(\sum_{i=1}^{n} Y_i)}$. Under local dependence, an upper bound on the central limit theorem error is given by

$$d_H(\bar{Y}_n, Z) \leq \frac{M_n^2}{\left\{var\left(\sum_{i=1}^{n} Y_i\right)\right\}^{3/2}} \sum_{i=1}^{n} E\left[|Y_i|^3\right] + \frac{\sqrt{26}M_n^{3/2}}{\sqrt{\pi} var(\sum_{i=1}^{n} Y_i)} \sqrt{\sum_{i=1}^{n} E\left[Y_i^4\right]}$$

where $M_n = max_{i\in 1,\ldots,n}\left|\mathcal{M}_i\right|$ is the size of the largest dependence neighborhood for nodes 1 through n. The proof, which relies on little more than Taylor series expansions and the triangle and Cauchy-Schwartz inequalities, is given in Ross (2011). For refinements and generalizations of this result see Chen and Shao (2004). If M_n is bounded above as $n \to \infty$, and if $E[Y_i^4]$ is bounded for all i, then it is easy to show that $\bar{Y}_n$ converges in distribution to a standard normal at rate $n^{-1/2}$. If M_n is allowed to depend on n, a central limit theorem could still hold but convergence may be at a slower rate. Obviously, the right hand side of (**??**) will not go to 0 if M_n grows at a rate of $n^{1/4}$ or faster, though further restrictions on the overlap of dependence neighborhoods or on the amount of dependence allowed within neighborhoods could still recover a central limit theorem.

Local dependence requires fixed boundaries for the dependence between observations, but it generalizes m-dependence by allowing the size and shape of the dependency neighborhoods to depend on more than just distance between nodes. Local dependence might operate if dependence is due to latent traits that only induce dependence over certain groups of nodes. For example, if the latent traits are genetic then each node's dependency neighborhood would include his biological relatives. These dependency neighborhoods would likely be overlapping and could be inconsistent with m-dependence. In general, analysis under local dependence does not require identifying the neighborhoods, but asymptotic results require knowledge of the size and growth of the neighborhoods relative to the entire network.

5.3 Subsampling

Subsampling methods have been used with great success in many Euclidean dependence settings (e.g. Heagerty and Lumley, 2000; Lahiri, 1993, 2003; Politis and Romano, 1994a,b; Politis et al., 1999). Details and proofs can be complicated, but the principles underlying successful subsampling methods are simple. Stationarity is required to ensure that each subsample is representative of the underlying data generating distribution, and limits on the amount of dependence are required to ensure that most subsamples are approximately independent of most other subsamples. There is a tradeoff between stationarity and independence assumptions: roughly, as the size of the groups over which stationarity holds increases, the size of the subsamples must similarly increase (to ensure that all features of the data generating distribution have a chance of being selected into each subsample) and therefore stronger assumptions are required to ensure that most subsamples are independent from one another.

Neither the implementation nor the justifications for subsampling under Euclidean dependence translate immediately into the network setting. Consider the following two procedures for estimating $Avar(\bar{Y})$ by way of subsampling.

Subsampling Procedure 1

1. Select B (possibly overlapping) subsamples of l consecutive observations. Let I_b be the set of indices of the nodes included in subsample b for $b = 1, ..., B$.
2. For each subsample, compute the subsample mean $\bar{Y}_b = \frac{1}{l}\sum_{i\in I_b} Y_i$.
3. Estimate $Avar(\bar{Y})$ with $\widehat{\sigma}^2_{\bar{Y},sub1} = \frac{l}{B}\left[\sum_{b=1}^{B}\bar{Y}_b^2 - \left(\frac{1}{B}\sum_{b=1}^{B}\bar{Y}_b\right)^2\right]$.

For details on Procedure 1 see Lahiri (2003).

Subsampling Procedure 2

1. Select B (possibly overlapping) subsamples of consecutive observations. Let I_b be the set of indices of the nodes included in subsample b for $b = 1, ..., B$ and let n_b be the size of subsample b.
2. For each subsample, compute the subsample variance estimator $\widehat{\sigma}^2_b = \frac{1}{n_b}\sum_{i,j\in I_b}^{B}\left(Y_i - \bar{Y}\right)\left(Y_j - \bar{Y}\right)$.
3. Estimate $Avar(\bar{Y})$ with $\widehat{\sigma}^2_{\bar{Y},sub2} = \frac{1}{B}\sum_{b=1}^{B}\widehat{\sigma}^2_b$.

For details on procedure 2 see Heagerty and Lumley (2000). (Note that in Step 2 the mean $\bar{Y}$ is estimated from the entire sample, avoiding the problem of degeneracy mentioned in Section 2.)

Both procedures rely on stationarity of the mean and covariance of $\mathbf{Y}$ over groups of observations of a size that is of a smaller order than the size of the subsamples. This ensures that each subsample is representative of the underlying data generating distribution with respect to $Avar(\bar{Y})$. In both procedures, the consistency of the estimator in Step 3 relies on a limit to the amount of dependence among subsamples. In Procedure 1, which is the most common subsampling procedure for estimator of $Avar(\bar{Y})$, the estimator $\widehat{\sigma}^2_{\bar{Y},sub1}$ takes the form of the sample variance estimator for independent observations. This is a reasonable procedure if the covariance between subsamples is asymptotically negligible. Procedure 2 is reasonable if each subsample variance estimator is asymptotically unbiased for $Avar(\bar{Y})$, which is guaranteed by the weak stationarity described above, and if the variance of $\widehat{\sigma}^2_{\bar{Y},sub2}$ converges to 0 as $n \to \infty$. We can deconstruct the variance of $\widehat{\sigma}^2_{\bar{Y},sub2}$ as follows:

$$\begin{aligned}
Var\left(\widehat{\sigma}^2_{\bar{Y},sub2}\right) =& \frac{1}{B^2}\sum_{b=1}^{B} Var\left(\widehat{\sigma}^2_b\right) \\
&+ \frac{1}{B^2}\sum_{\|I_b,I_d\|\leq m} 2Cov\left(\widehat{\sigma}^2_b,\widehat{\sigma}^2_d\right) \\
&+ \frac{1}{B^2}\sum_{\|I_b,I_d\|> m} 2Cov\left(\widehat{\sigma}^2_b,\widehat{\sigma}^2_d\right)
\end{aligned}$$

where $\|I_b, I_d\| = max_{i\in I_b, j\in I_d}\|i,j\|$ is a measure of the distance between two subsamples. The first term on the right-hand side converges to 0 under mild conditions like uniform integrability of $\widehat{\sigma}^4_b$. The second and third terms are cross-subsample covariance terms for subsamples that are overlapping or close to one another and for subsamples that are more than m units apart, respectively. These two terms will converge to 0 if, for example, the number of subsamples that are close to one another grows at a rate slower than $1/B^2$ and m-dependence holds (under which the third term is exactly equal to 0). Alternatively, under some versions of mixing conditions the third term will converge to 0 even if m-dependence does not hold.

Step 1 poses a challenge for implementation in network settings. In Euclidean dependence settings, subsamples generally comprise all of the observations within a rectangle in $\mathbb{R}^d$; shifting the rectangle incrementally in the direction of each axis ensures that the subsamples cover the entire sample domain and that all points at a sufficient distance from the boundary of the sample space are selected into the same number of subsamples. The most intuitive way to sample consecutive observations from a network is to select a node to be the center of the subsample, then select all nodes that are 1 unit away from the central node, nodes that are 2 units away from the central node, and so on until the desired subsample size is achieved. In all but the most trivial of network configurations this will lead to nodes with many ties being selected into more subsamples than nodes with fewer ties. In some cases, this could result in bias in the sampling distribution of $\bar{Y}$, which may be remedied using inverse probability of sampling weights that are easy to calculate when the network is fully known. Another way to select consecutive subsamples is to divide the network into non-overlapping blocks such that the nodes in each block are consecutive (i.e. 1 unit away from at least one other node in the block). This may preserve the consistency of the subsampling estimator described above, but in a finite network it will often be difficult to do this while ensuring that each subsample preserves topological and distributional features of the original data.

5.4 Weakly dependent clusters

Suppose a network is comprised of distinct clusters or subcommunities. If observations in the same cluster have arbitrary dependence but observations between clusters are only weakly dependent, in particular if the clusters are asymptotically mean independent from one another, then two approaches from the Euclidean dependence literature may be appropriate for estimating $Avar(\bar{Y})$. Let $\bar{Y}_r$ be the sample mean in the r^{th} cluster, $r = 1, ..., R$.

If the cluster means are asymptotically normally distributed with a common mean μ, i.e.

$$\begin{pmatrix} \overline{Y}_1 \\ \vdots \\ \overline{Y}_R \end{pmatrix} \xrightarrow{d} N\left(\begin{bmatrix} \mu \\ \vdots \\ \mu \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_R^2 \end{bmatrix} \right),$$

then it follows from results due to Ibragimov and Müller (2010) that standard t-distribution based inference will be conservative for the sample mean. That is, the confidence interval $\overline{Y} \pm \widehat{\sigma} t_{(0.025,R-1)}$ has conservative coverage for μ, where $\widehat{\sigma}^2 = \frac{1}{R-1}\sum_{r=1}^{R}\left(\overline{Y}_r - \overline{Y}\right)^2$. This holds even if the number of clusters R is small and does not grow with n. Note that stationarity is required only at the cluster level, and only for the mean. The cluster variances can be heteroskedastic. It should be noted, though, that heteroskedasticity does not come for free: the more heteroskedastic the variances, the wider (more conservative) the confidence intervals.

If, instead, the cluster means converge to a distribution that is not necessarily normal but with common means μ, common variances σ^2, and covariances 0, then we can bootstrap the weakly dependent communities, treating the clusters as independent and identically distributed observations. For this procedure we require a larger number of clusters. If there can be an arbitrary amount of dependence among observations within a cluster then the number of clusters would have to grow with n in order to achieve consistency, but for a fixed number of clusters consistency could still be achieved if the information contained within each cluster increases with cluster size. Again, stationarity is required only at the cluster level, and only for the cluster means and their variances.

In the Euclidean dependence literature, the mean independence assumption on which these methods relies is justified by conditions on the relative size of the boundaries and interiors of the clusters as the clusters grow uniformly in d dimensions (along with mixing conditions or m-dependence). These types of conditions do not apply to the network setting, where clusters do not have clear boundaries or interiors, and growth does not occur in a fixed number of dimensions. The hypothesis of mean independence can be tested directly, but it is more informative to have low level conditions that justify the assumption. As analogues to the concepts of interior and boundary locations in Euclidean space, I propose ties that connect two nodes in the same cluster (interior ties) and ties that connect nodes in different clusters (boundary ties), respectively. Let e_{jl} be the total number of edges with on endpoint in cluster j and the other in cluster l. Define a mixing measure for cluster r, $\xi(r) = \frac{e_{rr}}{\sum_{j=1}^{R} e_{rj}}$, as the ratio of the number of interior edges to total number of edges with at least one endpoint in cluster r. Similarly, define a mixing measure for paths of length m to be the proportion of such paths that are entirely contained within a cluster. For example, $\xi^2(r) = \frac{p^2_{rrr}}{\sum_{l,j=1}^{R} p^2_{rjl}}$, where p^2_{rjl} is a path of length 2 that starts in cluster r, moves to cluster j, and ends in cluster l. We can summarize these cluster-level measures across the network by averaging, $\xi = \frac{1}{R}\sum_{r=1}^{R} \xi(r)$ and $\xi^m = \frac{1}{R}\sum_{r=1}^{R} \xi^m(r)$. Under m-dependence, clusters will be mean independent if $\xi^m \to 1$ as $n \to \infty$.

5.5 K-dependence

In some settings it may be easier to estimate the full variance-covariance matrix $Cov\left(\mathbf{Y}\right)$ directly, from which we can easily derive an estimate of $Avar(\bar{Y})$. We discuss one such setting in this section. Suppose that the covariance between two observations depends only on the distance between the nodes on which the observations were made: $Cov(Y_i, Y_j) = \sigma_k$, where $k = \|i, j\|$. This dependence structure, which we call k-dependence, is the network analogue of autocorrelation in time series data. Note that even if k-dependence is not plausible for the shortest path distance metric, it may be plausible under a definition of distance that includes types of paths or number of paths between two nodes.

Under k-dependence, we can estimate $Cov\left(\mathbf{Y}\right)$ by the following procedure. For each k from 1 to the maximum distance between two nodes observed in the network, let $\mathcal{A}_k$ be a set of H_k pairs of indices (i_h, j_h) such that $\|i_h, j_h\| = k$, $h = 1, ..., H_k$. Estimate σ_k with $\widehat{\sigma}_k = \frac{1}{H_k}\sum_{h=1}^{H_k}\left(Y_{i_h} - \bar{Y}\right)\left(Y_{j_h} - \bar{Y}\right)$, and estimate $Cov\left(\mathbf{Y}\right)$ with the plug-in estimator $\widehat{\Sigma} = \left[\widehat{\sigma}_{\|i,j\|}\right]$. If $\mathbf{Y}$ is mean stationary, then $\widehat{\sigma}_k$ is unbiased for σ_k and $\widehat{\Sigma}$ unbiased for $Cov\left(\mathbf{Y}\right)$.

Under m-dependence, a simple tweaking of the procedure above results in consistent estimation of $\widehat{\Sigma}$. As before, for each k from 1 to m we select H_k pairs of nodes (i_h, j_h) such that $\|i_h, j_h\| = k$, but we now add the requirement that for any two nodes i_h and j_g that are members of different pairs in $\mathcal{A}_k$, $\|i_h, j_g\| \geq m$. That is, the distance within pairs is k and the distance between pairs is sufficiently large to ensure that the pairs are mutually independent from one another. Now $\widehat{\sigma}_k$ is a sample mean of independent terms and will converge as $H_k \to \infty$.

In order to achieve consistency the number of mutually independent pairs must grow with sample size. This requires a model of population growth in which the number of pairwise independent observations increases with sample size. However, unlike other methods that rely on m-dependence or mixing conditions, it does not require the number of pairwise dependent observations to become vanishingly small as n grows to infinity. This breaks with one of the fundamental principles of Euclidean dependence, namely that as n goes to infinity the amount of independence in the data swamps the amount of dependence in the data. If the amount of dependence in the data remains significant as n increases, a central limit theorem is unlikely to hold. However, under mean stationarity, k-dependence, and m-dependence we are able to consistently estimate the first two moments of the distribution of $\bar{Y}$ and we may therefore be able to use moment inequalities to construct valid confidence intervals.

K-dependence might be a reasonable assumption if, for example, dependence is due to shared genetic material and network distance corresponds to degree of relatedness.

6 Conclusion

The statistical study of networks is in its infancy and an immense amount of work remains to be done. Perhaps most important for the purposes of valid statistical inference is the development of asymptotic theory tailored to network topology and to known and realistic models of network growth. A combination of increasing domain and infill asymptotics may be consistent with some models of network generation and growth. These two asymptotic frameworks have been successfully combined in the context of Euclidean dependence (Fazekas, 2003; Lahiri, 1999; Lahiri et al., 1999; Zhang and Zimmerman, 2005), and adapting these results to the network setting could prove powerful and useful. Another potentially fruitful direction to explore is models for asymptotic growth in which clustering and other topological features repeat at different scales as the network grows; we could call this *fractal asymptotics*. There is already considerable evidence that some types of networks may grow this way, though perhaps not social networks (Inaoka et al., 2004; Jung et al., 2002; Song et al., 2005, 2006). This type of asymptotic growth could be consistent with increasing domain asymptotics or a combination of increasing domain and infill asymptotics.

Substantive knowledge is required to determine when m-dependence or k-dependence might be reasonable assumptions, but network topology and growth models determine whether these assumptions will result in desirable limit theorems. Work is needed to identify low-level conditions on network generating models that are consistent with central limit theorems and laws of large numbers for network data. Similarly, further work is needed to characterize the types of clustering and mixing that are consistent with weakly dependent clusters.

References

Anselin, L. (2001). Spatial econometrics. *A companion to theoretical econometrics*, **310330**.

Aronow, P. M. and Samii, C. (2013). Estimating average causal effects under general interference. Tech. rep., Yale University.

Barbour, A. and Chen, L. H. Y. (2005). *Stein's method and applications*, vol. 5. World Scientific.

Besag, J. (1974). On spatial-temporal models and markov fields. In *Transactions of the Seventh Prague Conference on Information Theory, Statistical Decision Functions, and Random Processes*. Springer, 47–55.

Bowers, J., M, F. M. and C, P. (2013). Reasoning about interference between units: A general framework. *Political Analysis*, **21** 97–124.

Chen, L. H. and Shao, Q.-M. (2004). Normal approximation under local dependence. *The Annals of Probability*, **32** 1985–2028.

Christakis, N. and Fowler, J. (2007). The spread of obesity in a large social network over 32 years. *New England Journal of Medicine*, **357** 370–379.

Christakis, N. and Fowler, J. (2008). The collective dynamics of smoking in a large social network. *New England journal of medicine*, **358** 2249–2258.

Christakis, N. and Fowler, J. (2010). Social network sensors for early detection of contagious outbreaks. *PloS one*, **5** e12948.

Cohen-Cole, E. and Fletcher, J. (2008). Is obesity contagious? social networks vs. environmental factors in the obesity epidemic. *Journal of Health Economics*, **27** 1382–1387.

Cressie, N. A. (1993). *Statistics for Spatial Data*. Wiley, New York.

Davidson, J. (1992). A central limit theorem for globally nonstationary near-epoch dependent functions of mixing processes. *Econometric theory*, **8** 313–329.

Dedecker, J., Doukhan, P., Lang, G., Leon, J. R., Louhichi, S. and Prieur, C. (2007). Weak dependence, examples and applications. In *Lecture Notes in Statistics*, vol. 190. Springer.

Diaconis, P. and Janson, S. (2007). Graph limits and exchangeable random graphs. *arXiv preprint arXiv:0712.2749*.

Fazekas, I. (2003). Limit theorems for the empirical distribution function in the spatial case. *Statistics & Probability Letters*, **62** 251–262.

Gile, K. and Handcock, M. S. (2006). Model-based assessment of the impact of missing data on inference for networks. *Unpublished manuscript, University of Washington, Seattle.*

Goetzke, F. (2008). Network effects in public transit use: evidence from a spatially autoregressive mode choice model for new york. *Urban Studies*, **45** 407–417.

Goldenberg, A., Zheng, A. X., Fienberg, S. E. and Airoldi, E. M. (2010). A survey of statistical network models. *Foundations and Trends in Machine Learning*, **2** 129–233.

Graham, B., Imbens, G. and Ridder, G. (2010). Measuring the effects of segregation in the presence of social spillovers: A nonparametric approach. Tech. rep., National Bureau of Economic Research.

Halloran, M. and Hudgens, M. (2011). Causal inference for vaccine effects on infectiousness. *The University of North Carolina at Chapel Hill Department of Biostatistics Technical Report Series* 20.

Halloran, M. and Struchiner, C. (1995). Causal inference in infectious diseases. *Epidemiology* 142–151.

Heagerty, P. J. and Lumley, T. (2000). Window subsampling of estimating functions with application to regression models. *Journal of the American Statistical Association*, **95** 197–211.

Hong, G. and Raudenbush, S. (2006). Evaluating kindergarten retention policy. *Journal of the American Statistical Association*, **101** 901–910.

Hong, G. and Raudenbush, S. (2008). Causal inference for time-varying instructional treatments. *Journal of Educational and Behavioral Statistics*, **33** 333–362.

Hudgens, M. and Halloran, M. (2008). Toward causal inference with interference. *Journal of the American Statistical Association*, **103** 832–842.

Huisman, M. (2009). Imputation of missing network data: some simple procedures. *Journal of Social Structure*, **10** 1–29.

Ibragimov, R. and Müller, U. K. (2010). t-statistic based correlation and heterogeneity robust inference. *Journal of Business & Economic Statistics*, **28** 453–468.

Inaoka, H., Ninomiya, T., Taniguchi, K., Shimizu, T. and Takayasu, H. (2004). Fractal network derived from banking transaction–an analysis of network structures formed by financial institutions. *Bank of Japan, Working Paper* 04.

Jenish, N. (2008). *Asymptotic Theory for Spatial Processes*. Ph.D. thesis, University of Maryland, http://drum.lib.umd.edu/bitstream/1903/8535/1/umi-umd-5605.pdf.

Jenish, N. and Prucha, I. R. (2009). Central limit theorems and uniform laws of large numbers for arrays of random fields. *Journal of econometrics*, **150** 86–98.

Jenish, N. and Prucha, I. R. (2012). On spatial processes and asymptotic inference under near-epoch dependence. *Journal of Econometrics*, **170** 178–190.

Jung, S., Kim, S. and Kahng, B. (2002). Geometric fractal growth model for scale-free networks. *Physical Review E*, **65** 056101.

Kossinets, G. (2006). Effects of missing data in social networks. *Social networks*, **28** 247–268.

Lahiri, S. (1999). Asymptotic distribution of the empirical spatial cumulative distribution function predictor and prediction bands based on a subsampling method. *Probability theory and related fields*, **114** 55–84.

Lahiri, S. N. (1993). On the moving block bootstrap under long range dependence. *Statistics & Probability Letters*, **18** 405–413.

Lahiri, S. N. (2003). *Resampling methods for dependent data*. Springer, New York.

Lahiri, S. N., Kaiser, M. S., Cressie, N. and Hsu, N.-J. (1999). Prediction of spatial cumulative distribution functions using subsampling. *Journal of the American Statistical Association*, **94** 86–97.

Lauritzen, S. L. and Richardson, T. S. (2002). Chain graph models and their causal interpretations. *Journal of the Royal Statistical Society: Series B*, **64** 321–348.

Lee, L.-F. (2004). Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica*, **72** 1899–1925.

Lim, C. Y. and Stein, M. (2008). Properties of spatial cross-periodograms using fixed-domain asymptotics. *Journal of Multivariate Analysis*, **99** 1962–1984.

Lin, X. (2005). Peer effects and student academic achievement: an application of spatial autoregressive model with group unobservables. *Unpublished manuscript, Ohio State University*.

Lyons, R. (2011). The spread of evidence-poor medicine via flawed social-network analysis. *Statistics, Politics, and Policy*, **2**.

Newman, M. (2009). *Networks: an introduction*. Oxford University Press, Oxford.

O'Malley, J. A. and Marsden, P. V. (2008). The analysis of social networks. *Health services and outcomes research methodology*, **8** 222–269.

Politis, D., Romano, J. P. and Wolf, M. (1999). Weak convergence of dependent empirical measures with application to subsampling in function spaces. *Journal of statistical planning and inference*, **79** 179–190.

Politis, D. N. and Romano, J. P. (1994a). Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics*, **22** 2031–2050.

Politis, D. N. and Romano, J. P. (1994b). The stationary bootstrap. *Journal of the American Statistical Association*, **89** 1303–1313.

Rosenbaum, P. (2007). Interference between units in randomized experiments. *Journal of the American Statistical Association*, **102** 191–200.

Ross, N. F. (2011). Fundamentals of stein's method. *Probability Surveys*, **8** 210–293.

Rubin, D. (1990). Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, **5** 472–480.

Shalizi, C. and Thomas, A. (2011). Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research*, **40** 211–239.

Shalizi, C. R. (2012). Comment on "why and when 'flawed' social network analyses still yield valid tests of no contagion". *Statistics, Politics, and Policy*, **5**.

Shalizi, C. R. and Rinaldo, A. (2013). Consistency under sampling of exponential random graph models. *Annals of Statistics*, **41** 508–535.

Sobel, M. (2006). What do randomized studies of housing mobility demonstrate? *Journal of the American Statistical Association*, **101** 1398–1407.

Song, C., Havlin, S. and Makse, H. A. (2005). Self-similarity of complex networks. *Nature*, **433** 392–395.

Song, C., Havlin, S. and Makse, H. A. (2006). Origins of fractality in the growth of complex networks. *Nature Physics*, **2** 275–281.

Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proc. Sixth Berkeley Symp. Math. Stat. Prob.* 583–602.

Stomakhin, A., Short, M. B. and Bertozzi, A. L. (2011). Reconstruction of missing data in social networks based on temporal patterns of interactions. *Inverse Problems*, **27** 115013.

Tchetgen Tchetgen, E. J. and VanderWeele, T. (2012). On causal inference in the presence of interference. *Statistical Methods in Medical Research*, **21** 55–75.

Thomas, A. C. (2013). The social contagion hypothesis: Comment on 'social contagion theory: Examining dynamic social networks and human behavior'. *Statistical in Medicine*, **32** 581–590.

van der Laan, M. J. (2012). Causal inference for networks. *U.C. Berkeley Division of Biostatistics Working Paper Series*, **Working Paper 300**.

VanderWeele, T. (2010). Direct and indirect effects for neighborhood-based clustered and longitudinal data. *Sociological Methods & Research*, **38** 515–544.

VanderWeele, T. and Tchetgen Tchetgen, E. (2011a). Bounding the infectiousness effect in vaccine trials. *Epidemiology*, **22** 686.

VanderWeele, T. and Tchetgen Tchetgen, E. (2011b). Effect partitioning under interference in two-stage randomized vaccine trials. *Statistics & probability letters*, **81** 861–869.

VanderWeele, T. J., Ogburn, E. L. and Tchetgen, E. J. T. (2012). Why and when 'flawed' social network analyses still yield valid tests of no contagion. *Statistics, Politics, and Policy*, **3**.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of small-world networks. *Nature*, **393** 440–442.

Wu, W.-Y., Lim, C. Y. and Xiao, Y. (2012). Tail estimation of the spectral density for a stationary gaussian random field. *Journal of Multivariate Analysis*, **116** 74–91.

Zhang, H. and Zimmerman, D. L. (2005). Towards reconciling two asymptotic frameworks in spatial statistics. *Biometrika*, **92** 921–936.